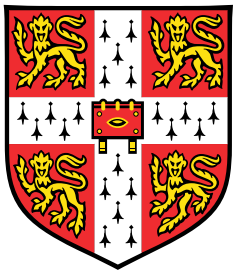
On Protein Rigidity and Quantum Effects in Biological Systems



Alexander Fokas

Department of Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

February 2017

I dedicate this thesis to my mother, whose selflessness, patience, and courage has given me a unique source of guidance

Declaration

This dissertation includes work carried out between October 2013 and October 2016 in the Theory of Condensed Matter Group at the Cavendish Laboratory, Cambridge, under the supervision of Dr. Alex Chin, Dr. Daniel Cole, and Professor Stephen Emmott. I hereby declare that, except where specific reference is made, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures. Parts of this dissertation have been published, as follows:

Alexander Fokas
February 2017

Acknowledgements

I would like to begin by thanking my supervisors, Dr Alex Chin, Dr Daniel Cole, and Professor Stephen Emmott, from whom I have learned enormously. Their clarity and direction has been an indispensable tool that I have called upon on many occasions. My time at TCM has provided me with the opportunity to have many fruitful discussion, from which I will highlight Greg Lever, Eric Franzosa, Stephen Wells, Nicholas Hine, and Sebastian Ahnert. Their advice and collaborative work allowed me to explore many far-reaching aspects of physics, both within and beyond my research question. I have to give a special mention to Michael Rutter, who has on numerous occasions filled gaps in my knowledge, in addition to providing those around him with entertaining reading regarding the world outside the walls of our lab. I would like to thank my friends for their encouragement, and in particular Tom Northey and Salvatore Tesoro, with whom I shared many influential discussions. Lastly, I would like to thank my family for their support, without which I would not be where I am today. This is particularly true of my father, whose wisdom, passion, and love for his family is a continual source of inspiration.

On Protein Rigidity and Quantum Effects in Biological Systems

A central principle in structural biology states that the function of a protein depends critically on its atomic structure. A protein's structure additionally determines the allowed conformal motions that are accessible in the native state, which also play a critical role in the observed function. In this thesis, multiple computational techniques, ranging from intra-protein contact networks to fully quantum mechanical simulations, have been employed to explore protein structures and, by analogy, their functions. Particular attention is given to the antenna complex of green sulfur bacteria, where the observation of optical quantum beats has stimulated extensive interdisciplinary work and debate regarding the stability and function of the excitonic coherences. To further understand the role played by the protein environment, the work herein combines linear-scaling electronic structure techniques with constrained geometric dynamical simulations to explore how large-scale structural dynamics in the Fenna-Matthews-Olson (FMO) complex are correlated to static disorder in its excited state energies. Remarkably, the slow conformal motion leads to large variations in the Q_y optical transition energies of pigments 3 and 4 across the ensemble of structures, while their energy difference is strongly conserved (303 ± 27 cm^{-1}). Furthermore, we identify the key secondary structure elements and motions that give rise to this correlated disorder, which is seen to arise from the highly constrained global structure of the FMO protein. The constrained dynamics input, that is a set of non-covalent and covalent interactions, has been explored more generally. Using these contacts, a network construction method is described that includes the effects of allowed conformal motions and local chemical environments. This novel amino acid network construction technique allows the identification of residues that regulate allostery in G-protein coupled receptor signaling.

Alexander Samuel Fokas

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

2DES	2-dimensional electron spectroscopy
AAN	Amino Acid Network
ATP	Adenosine-tri-phosphate
Bchls	Bacteriochlorophyll
CGN	Common Gα numbering system
Chl	Chlorophyll a
CNA	Constraint Network Analysis
CS	Clam Shell
DC	Distance constrant
DFT	Density Functional Theory
dN	Nonsynonomous mutations
DNA	Deoxyribonucleic acid
dS	Synonomous mutations
E_m	Site energy
EET	Excitation Energy Transfer
ENM	Elastic Network Model
EVT	Expected Visiting Time

FIRST Floppy Inclusion and Rigid Substructure Topography

FMO Fenna-Matthews-Olson

FRODA Framework Rigidity Optimized Dynamics Algorithm

FWHM Full width half maximum

GDP Guanosine diphosphate

GEF Guanine nucleotide exchange factor

GGPPS Geranyl-geranyl diphosphate synthase

GSB Green Sulphur Bacteria

GTP Guanosine triphosphate

H_{cut} Hydrogen bond energy cutoff

HFGFR1 Human fibroblast growth factor receptor 1

HK Hohenberg-Kohn

HOMO Highest occupied molecular orbital

LH1 Light harvesting complex 1

LH2 Light harvesting complex 2

LUMO Lowest unoccupied molecular orbital

MD Molecular Dynamics

MM Molecular Mechanics

MPI Message passing interface

NGWF Non-orthogonal generalised Wannier functions

NMA Normal Mode Analysis

NMR Nuclear magnetic resonance

ONETEP Order-N Electronic Total Energy Package

PBE Perdew–Burke–Ernzerhof

PC1	Principal component 1
PCA	Principal component analysis
PDB	Protein data bank
PPC	Pigment Protein Complex
QM	Quantum Mechanical
RC	Rigid Cluster
RGN	Residue Geometry Network
RMSD	Root-mean squared deviation
RMSF	Root-mean squared fluctuation
SNP	Single nucleotide polymorphism
unRGN	Unweighted Residue Geometry Network
wRGN	Weighted Residue Geometry Network
XC	Exchange-correlation

Chapter 1

Introduction

"For we cannot command nature except by obeying her"

-Francis Bacon, 1620

Proteins perception of functional states means that they can gather, integrate, and utilize sensory information, allowing them to form memories of the past and make predictions about the future [217]. Understanding the complex behaviour of proteins and the resulting functions can have dramatic consequences beyond health and medicine in fields such as computing and technology. Progress toward this goal has and will require scientists from different disciplines to focus on the common denominators of these problems. This non-traditional approach is demonstrated by the nascent field of Quantum Biology, where biologists and physicists gather skills from their respective disciplines to seek integrative solutions to complex problems.

1.1 Proteins as Molecular Machines

The complete sequencing of the human genome [232] granted a unique perspective into our hereditary history and genetic structure. Surprisingly, only ~24,000 protein-coding genes were identified, a fraction of the ~150,000 that had been estimated [43]. It became clear that the number of genes did not reflect the enormous complexity found in humans when compared, for example, to the flat worm *Caenorhabditis* elegans [224] which harbours ~19,000 genes of its own. One area of complexity is found in the proteome, where the number of expressed protein variants, at best estimates, exceeds ~90,000 [172, 71]. Perhaps more significant is the size of the human interactome, where we find more than ~650,000 protein-protein interactions [219]. These findings suggest that complexity is largely increased via added protein functionality. Understanding the fundamental principles of protein behaviour will

allow researchers to design and modify systems of increasing complexity, with widespread biotechnological and biomedical applications [142]. Ultimately, structural biology aims to identify tools that can manipulate biological machinery in the same way that engineers build and modify mechanical systems.

The functionality of a living organism, much like a computer, is information driven. So-called information dynamics have been studied by Frieden et al. [77]. They find that living systems maintain an extreme level of information by employing maximum entropic states. The importance of these dynamics are evidenced in cancer. In this disease, cells divide uncontrollably following crucial information losses that result from accumulated mutations, disordered morphology, and functional decline [77]. Proteins are employed by organisms to manipulate and interpret their environment, primarily through conformational changes, and bring about global structural and functional order. The information processing observed in protein networks raises an analogy with machines and computers. For example, proteins that participate in metabolic reactions have fine tuned substrate selection mechanisms that sense molecule shape. The selection process was described a century ago by the lock-and-key principle [143], and highlights the protein as an information processing unit. Seventy years later allosteric proteins, which exist in distinct functional states in response to their environment, were identified in pathways displaying negative feedback [155]. Allostery allows proteins to respond to the interactome causally and is born out of complex structure dynamics. Recently, researchers were able to control an enzyme's behaviour by incorporating this switchable, allosteric action [41]. To further build on these experiments will likely require a reductionist approach to identify generalisable mechanisms in allostery and, more generally, in any observed protein functions. This will, according to the central dogma of structural biology, follow a deep understanding of their unique structure and complex dynamics [18]. It is promising that in nature there exist universal features in protein structure, suggesting that common functional motifs exist and can be identified as well.

Proteins are constructed from amino acids, often referred to as *residues*, which share a common template. Each residue is composed of a central carbon atom that connects a hydrogen group, a carboxylic acid group, a variable group, and an amino group. The amine and carboxylic acid groups of adjacent residues undergo a condensation reaction to form a peptide bond. Enzymes are employed to sequentially carry out this reaction and form linear polypeptide chains, which fold into a 3-dimensional protein. Although there are only 20 variable groups, this small number of residues gives rise to enormous complexity; the volume of all the possible 100 residue proteins could not fit in the visible universe [162]. An infinitesimal subset of this sequence space has been explored by nature yet gives rise to essentially all cellular processes in living organisms.

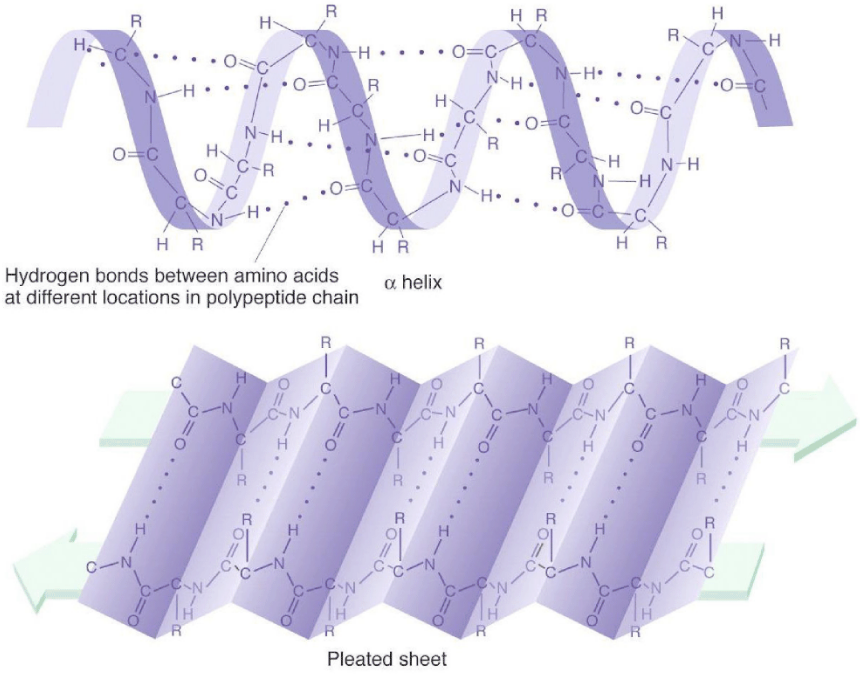


Fig. 1.1 A α-helix (top) and β-sheet (bottom). These structural motifs are formed using a regular array of hydrogen bonds (dots). The tertiary structure can be constructed from only α-helices (α proteins), only β-sheet (β proteins), or a mixture of the two ($\alpha\beta$ proteins). The figure has been taken from Ref. [17].

The linear sequence of residues that forms the polypeptide chain, known as the primary sequence, is encoded in the genome using deoxyribonucleic acid (DNA). The primary sequence undergoes a "folding reaction" to a unique native state, known as the tertiary structure of a protein. This final state is actually a collection of structures with similar free energies whose inter-conversions describe functionally relevant motions. This highly dynamic final state is composed of stable, ubiquitous secondary (sub)structures that are formed by a regular array of backbone hydrogen bonds (Figure 1.1). The two most common of these secondary structure elements are known as α-helices and β-sheets [31]. An α-helix forms when the polypeptide chain has a high propensity to turn back on itself, allowing stable hydrogen bonds to form between the carbonyl oxygen and the hydrogen atom of the amide group found four residues along in sequence. β-sheets differ from α-helices as they cannot hydrogen bond with nearby residues. Instead, they interact with a neighbouring β–strand to form β-sheets. The monomeric protein described by the tertiary structure can then fulfil a specific function. Alternatively, surface complementarity can give rise to the formation of multi-protein complexes, known as the quaternary structure (Figure 1.2). The quaternary structure can be used to impose higher levels of metabolic organisation and, critically, achieves greater complexity in protein frameworks required for certain functions [78].

A prime example of purpose-built quaternary structures can be found in the light harvesting systems of plants and bacteria, which support nearly all life on earth. These diverse pigment-protein complexes (PPCs) implement ultrafast photophysics to facilitate the capture, transport, and utilisation of photonic energy. To do so, photosynthetic proteins employ optically-active molecules known as pigments. Although the general principles of photosynthesis have been well characterised [25], the finer details of the energy transfer pathways remain unclear. For example, the quaternary structure of these photosystems is generally modelled as a static entity. However, these architectures can undergo functional rearrangements in response to environmental factors, such as light intensity [244]. Recent trends have placed the motions within the tertiary structure under scrutiny and, in particular, questions have been raised regarding the role of the protein matrix in the observed high quantum efficiencies. Understanding how proteins influence the embedded pigments for energy transport within a noisy thermal environment is also of interest in the field of optoelectronics, where biological solutions to problems of efficient and directed energy transfer could suggest new ways to break these otherwise limiting constraints in artificial devices [201, 58]. Critically, the observed quantum mechanical effects in a range of PPCs extracted from bacteria, plants and algae [138, 98, 47, 169, 64, 104] have broadened the scope of what may be learned from proteins participating in photosynthetic machinery.

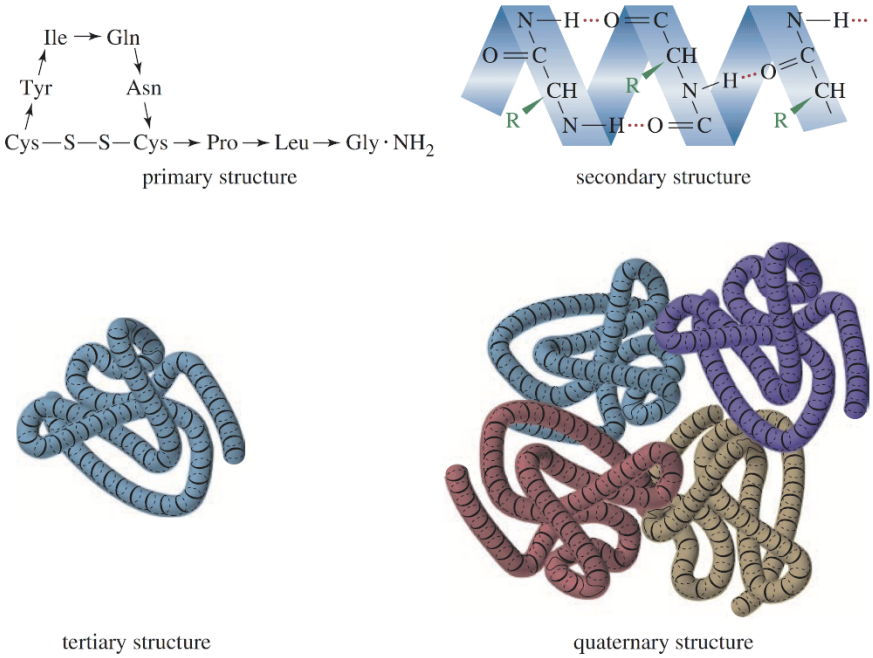


Fig. 1.2 The primary sequence includes the covalently bonded structure, including a disulphide bridge (-S-S-) that forms from the reduction of two thiol groups (the variable group of cysteine). Secondary structure elements, predominantly α-helices and β-sheets, form based on the dependencies of the primary structure. The final conformation of the molecule is known as the tertiary structure and the association of two or more peptide chains into multi-protein complexes is known as the quaternary structure. This figure has been adapted from Ref. [17].

1.2 Quantum Biology

In 1913, Niels Bohr showed that quantum mechanics is required to explain the stability of atoms and molecules [28]. He went on to define the principle of "complementarity" as the inherent indeterminism found in quantum mechanical systems. This followed experimental and theoretical observations that certain complement properties, such as position and momentum, could not be measured simultaneously [52]. Bohr went on to speculate that complement properties also existed in *biology*; one that describes subcellular mechanisms and another that describes the interactions at the level of organisms. In his 1932 lecture "Life and Light", Bohr's interest in biological systems inspired Max Delbrück to study the physical and chemical aspects of genetics [216]. This prompted Delbrück to formulate his physical model of the gene [253].

Delbrück established genetics as an information science, stimulating the interest of other physical scientists in biology. Perhaps most famously, Erwin Schrödinger discussed how and whether certain aspects of life could be accounted for by physics and chemistry in his 1944 book "What is life?" [203]. In this highly influential text, Schrödinger discussed how changes in molecular structures at an atomic scale governed by quantum mechanics could cause profound and visible change in the structure of living organisms. He went on to speculate the existence of an "aperiodic crystal" that could transmit this information. In 1953, James Watson and Francis Crick famously revealed the double-helix structure of DNA, affirming Schrödinger's hypothesis that genes are highly regular sequences in structure, composed of just 4 nucleic acids. This molecular model, which brought together the informational school led by Delbück with the structural school of crystallographers, in some sense led to the beginning of modern molecular biology.

Discussions in quantum physics formed the essential principles of modern molecular biology. However, these early discussions failed to uncover a *non-trivial* role of quantum mechanics in biology. This research actually showed that DNA could explain the basic phenomena of genetics, and by extension life, deterministically using classical mechanics. In fact, in the 1960's this reductionist view permeated through molecular biology, as is evident by Francis Crick's claim in *Of molecules of men* [50];

"The ultimate aim of the modern movement in biology is to explain all biology in terms of physics and chemistry"

This view has since been argued against [189], perhaps earliest by Bohr and Delbrück, who shared an anti-reductionist belief that certain functions, such as consciousness, would face limitations using purely reductionist methods. They hoped that a paradox - similar to complementarity - would arise in biology and need to be supplemented with a teleological

perspective [152]. Indeed, understanding the "hard problem of consciousness" [94] that is concerned with the subjectivity of perception cannot be accounted for under the current reductionist framework. Perhaps the occult nature of both quantum mechanics and consciousness naturally leads to their convergence. In particular, Sir Roger Penrose has speculated that coherent quantum processing can be found in collections of microtubules, the major structural support of cells, within brain neurons and thereby gives rise to consciousness [95, 96]. This 20-year-old theory was highly criticised as the brain was considered too warm, wet, and noisy and thus could not avoid decoherence from thermal vibrations. Although this theory was recently bolstered following the observation of quantum vibrations in the microtubules of brain neurons [193], it remains highly speculative example of a non-trivial effect in the brain [30].

The affect of general anaesthetics, whereby exposure to the drug is followed by the removal of all traces of perception, has been hypothesised to have a quantum mechanical action [229]. Despite its widespread use in medicine, anaesthetics remains poorly understood. Most peculiar is the fact that the nature of this interaction goes directly against the lock-and-key theory. General anaesthetics have a wide array of shapes, suggesting there is not one complementary pocket for their action. Furthermore, Claude Bernard discovered in 1878 that anaesthetics impacted plants and animals equivalently [91]. The smallest anaesthetic, the atom xenon, is chemically inert and therefore presents a particularly interesting puzzle for a chemical theory of action. But what xenon lacks in chemistry is made up for in physical reactivity. This led researcher Luca Turin to speculate that the pharmaceutical action of xenon, and anaesthetics generally, has roots in quantum mechanics. Turin et al. [229] provided evidence for this theory using fruit flies, where the total amount of free electron spins was found to increase when exposed to general anaesthetics [229]. Turnin suggests that the conductivity of anaesthetics may act disruptively, by throwing an "electronic spanner" in the electronic currents within proteins. Although this may not provide a quantum mechanical explanation for consciousness, the finding does suggest quantum mechanics could play a role in perturbing or disrupting consciousness. While fascinating and complex, the brain is not the first structure where quantum mechanics was hypothesised to play a non-trivial role in biology. Furthermore, these studies do not address the following fundamental question inferred by Delbrück, Bore, and Schrödinger: had organisms *evolved* a quantum mechanical function to gain a competitive edge over their classical counterpart? The earliest investigations focused on the action of enzymes following the speculation by Per-Olov Löwdin that tunnelling may be an important factor in DNA mutation [146]. In this 1963 paper, he termed this field of investigation "Quantum Biology".

Quantum biology investigates whether the non-trivial quantum effects are exploited by biological systems. One of the most iconic experiments, the so-called double slit experiment, challenged our understanding of the world by demonstrating that the structure of matter is exactly as ambiguous as that of light. Theorist and experimentalist would find a dualistic description of light and matter whereby, depending on the experimental circumstances, atoms could behave as particles or waves. This underlying uncertainty in the wavefunction allows the particle to penetrate a barrier, an effect known as tunnelling. Both electron [89] and hydrogen [157] tunnelling have been measured in biological systems. However, there is speculation that tunnelling is simply a byproduct of the enzyme environment. Although quantum corrections are found in these systems, enzymes do not necessarily provide an example of a function that has been competitively evolved to incorporate a quantum effect [137].

Spin is a property of elementary particles that describes an intrinsic form of angular momentum. This property was used by Wolfgang Pauli to explain how electrons fill orbitals [60]. Importantly, this purely quantum mechanical feature imparts sensitivity of atoms, and therefore chemical reactions, to magnetic fields [137]. This possibility has been explored in a series of behavioural studies involving the European robin [243]. During migratory season, these birds travel hundreds, or even thousands, of miles using a sensitivity to the earth's magnetic field. This function has recently been attributed to a class of proteins known as cryptochromes [29]. These proteins, according to the radical pair model, form magnetically-sensitive radical pairs upon photoexcitation. The differing reaction products are biologically detectable and the brain is informed through neuronal excitations in the retina. These radical-pair intermediates were shown to have millisecond lifetimes, suggesting that even the weak magnetic field of the earth could modulate singlet-triplet interconversions. As cryptochromes are found in a wide array of plants and animals, it may be that magnetoreception is not unique to migratory birds. Intriguingly, satellite images of grazing cattle find an alignment with the poles as well [21]. Spin effects have been found in other systems too. Experiments have found that the purple bacteria, in addition to possessing other quantum mechanical effects within their photosynthetic apparatus, have membranes that can behave as a light-dependent spin switch [63] .

Photosynthetic organisms appear to harness quantum coherence on physiologically important timescales. Among these is the Fenna-Matthews-Olson (FMO) complex, which facilitate the transport of electronic (excitation) energy from the antenna super-complexes to the reaction centres in green sulphur bacteria (GSB). Although it is only found in GSB, the FMO complex carries out a typical and important function that is found in all the light reactions termed excitation energy transfer (EET) [230, 26]. In 2007, Greg Engel et al. [64] employed

ultrafast femtosecond non-linear optical experiments to observe long-lasting quantum beating between electronic excited states in an ensemble of FMO complex structures. These studies verified the existence of delocalised excitation states and the efficient energy transfer pathways that give FMO its wire-like functionality. Unexpectedly, long-lasting oscillatory signals were observed at room temperature [169] that coincided with the timescales of energy transfer [59]. Their early assignment to long dephasing of energetic superpositions has raised profound and hotly debated questions about how electronic quantum coherence might be "protected" by PPC nanostructures, as well as its functional implications for photosynthesis and other biological processes. This problem requires employing tools in addition to those of experimental biochemistry. In particular, quantum mechanical simulations of the FMO complex have been proposed as an essential approach for understanding the impact of the protein environment.

The identification of delicate quantum effects in photosynthetic organisms is significant and unexpected, as the warm, wet, and noisy environment generally give rise to decoherence rates that inhibit any physiological impact. Furthermore, individual PPCs may retain electronic coherence for longer than has been identified in ensemble experiments [68]. *In silico* techniques thereby offer a promising approach for understanding the origin of quantum coherence, which will be the focus of Chapters 4 and 5.

1.3 Network Applications in Biology

Whereas the algorithms for network analysis are not familiar to the layman, networks themselves are so ubiquitous they feel instinctive. Interconnected structures are found everywhere; the "World Wide Web", electrical grids, neurons in the brain, gene regulation and, most famously, social networks, all have the same underlying network structure. The mathematical understanding of networks has thereby benefited engineers, neuroscientists, geneticists, and social scientists in understanding the diverse functions embedded in network systems. These so-called emergent properties of networks describe features that arise from the sum of its parts. For example, the "super-organism", or behaviour of a group of people, often differs from that of the individuals and has important implications for the morals of society [176]. Indeed, the ego may lead one to think they are separate from the influence of society yet research has shown that numerous aspects of our lives are determined by our surroundings; for better [73], worse [132], or simply bizarre [180].

In biology, networks are employed to simplify and understand the complex array of interactions found at the subcellular level. These techniques are commonly used to investigate gene networks [101] and protein-protein interaction networks [109]. More recently, intra-

protein networks have been constructed to understand the functionality of individual proteins, with applications found in protein design software [196]. As protein dynamics are pivotal to their function, these network construction techniques often employ expensive dynamical simulations to represent important conformal motions. This approach allows researchers to explore functional residues, predict folding reactions, analyse thermostability, and identify communication pathways [246]. Intra-protein networks, termed amino acid networks, will be the focus of Chapter 3.

1.4 Electronic Structure of Biological Systems

The ability of mathematics to explain essential qualities of physics has long been appreciated. Eugene Wigner highlighted this phenomenon in his lecture "The unreasonable effectiveness of mathematics in the physical sciences", where he emphasises the beauty of mathematical language in extending our knowledge of the physical universe [97]. In particular, the importance of quantum mechanics is emphasised by its ability to explain certain phenomena that classical laws cannot explain. Quantum mechanics describes an elegant theoretical framework that can be used by physicist, chemists, and now possibly biologists, to help describe an array of functionality. At the heart of quantum mechanics is a quantum state, which employs a wavefunction to describe an isolated quantum system. This function satisfies the Schrödinger equation [202] and can be used to describe all problems related to the electronic structure of matter. However, this equation can only be solved exactly for simple systems, such as the hydrogen atom. To describe more complex systems, such as those found in biology, requires approximations.

Density functional theory (DFT), formulated by Walter Kohn and Lu Jeu Sham [127], allows the many-electron problem to be expressed as a system of non-interacting electrons moving in an effective potential. In DFT, the electronic degrees of freedom are minimised through a self-consistent procedure which dramatically decreases the computational load associated with quantum mechanical calculations while maintaining a high degree of accuracy [128]. However, there is still a scaling problem that inhibits conventional DFT-like methods from tackling large systems, such as proteins. Indeed, the computational load of conventional DFT methods is of the order N^3 [126] where N is the number of particles in a system. Further approximations, such as the "nearsightedness" of single particle density matrices, yield a method of order N implemented in the Order-N Electronic Total Energy Package (ONETEP) code [210], where the computational cost increases only linearly with the number of atoms in the system. DFT is able to address some of the problems we find when modelling biological systems. In addition to their size, the small energy scales associated with the

accessible microstates requires very high accuracy to reach a precision under 1 kcal/mol. Furthermore, the widespread functionalities we find in biology requires advance electronic structure investigations, including solvation, strong correlations, dispersion, excited states, and spectroscopy [45]. This great advance in the field of *ab initio* electronic structure prediction allows multi-thousand atom systems to be simulated in their entirety via a DFT methodology. This powerful and accurate approach will be employed in Chapter 5 to explore the electronic structure of a light harvesting protein.

Outline

The above synopsis highlights important historical developments in structural biology and physics that preceded recent discoveries in the field of quantum biology. Chapter 2 provides a more detailed discussion of the observed quantum effects in photosynthetic organisms and an overview of network methods for studying protein functions. In Chapter 3, geometric simulation is reviewed and the the construction of the introduced "Residue Geometry Network" (RGN) is described. Applications of the RGN will be presented, which include an evolutionary analysis for a large protein dataset, identifying allosteric residues in G-protein coupled receptor signalling pathways, and predicting the destabilisation of protein structures. Chapters 4 and 5 describe the constrained dynamic simulations of the full trimeric FMO complex and the simulation of the optical properties using linear-scaling DFT, respectively. A brief summary and implications of the findings can be found in Chapter 6, as well as further applications of the RGN.

Chapter 2

Research Topics

2.1 Amino Acid Network

Networks are graphical representations of interconnected "bodies". Structural analysis using mathematical algorithms can then reveal patterns in the interacting elements that relate to so-called emergent properties, which refer to an associated behaviour or function. Perhaps the most useful aspect of networks are their transferablility to different settings, allowing many areas of science that are concerned with characterising how the components of a system interact to investigate their emergent properties. Networks have been highly successful in the analysis of social interactions where, despite their complex and variable structures, generalisable characteristics of society's behaviour have been uncovered regarding one's behaviour within a social network. Indeed, a significant sphere of influence on an individual's belief, health, career and emotional state can be identified in an their social network that extends 3 degrees [42], meaning that someone's friends' friends' friend has a measurable impact on key aspects of their life. This feature is related to the social evolution of humans [79]. Networks have also been used to understand biology at the systems level, allowing models of disease progression [103] (Figure 2.1), cell signalling pathways, metabolic pathways, and gene regulation to be conceived [121, 107].

In systems biology, the inter-protein networks are built in order to classify the emergence of some cellular behaviour. In structural biology, the structure-function dogma gives rise to the construction of intra-protein networks, referred to as amino acid networks (AANs), to model and/or investigate observed functions. This approach involves constructing an AAN using a portion or the whole amino acid and measuring the network properties. These measures can then provide insight into protein folding, protein-protein interactions, functionally important residues, intra- and inter-molecular communications, and thermostability [246]. In the current thesis, an AAN is constructed using an all-atom approach to interpret

chemical and geometric aspects of the environment with the aim of providing a static network construction method that provides insight into dynamical protein functions.

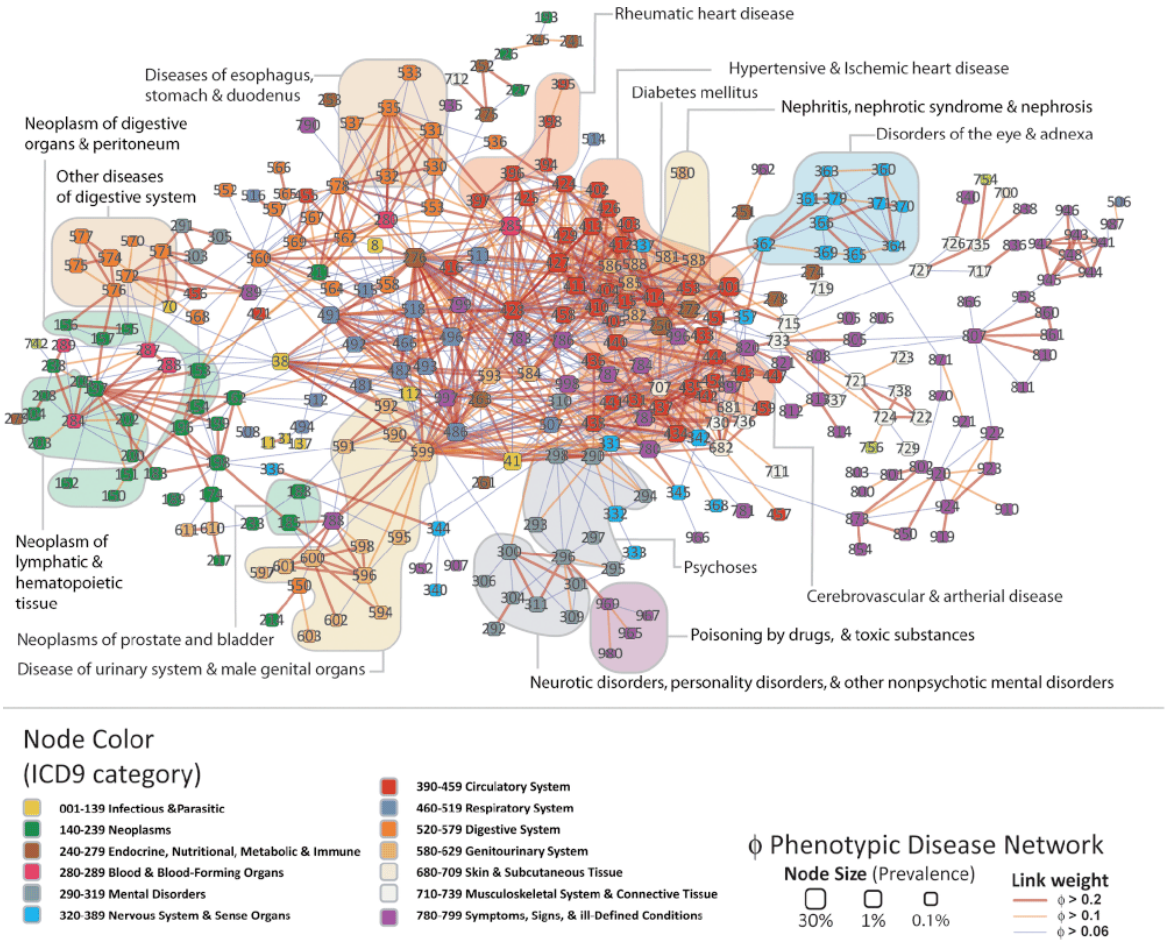


Fig. 2.1 Phenotype Disease Network. Disease phenotypes, according to the International Statistical Classification of Diseases categories, that have significant comorbidity are highlighted. Edges represent the the ϕ-correlation, namely the Pearson's correlation for binary variables, and gives information on the comorbidity of two diseases. This figure has been taken from Ref. [103].

AANs have gained popularity in the last decade as an approach for studying the structural properties of a protein. AANs contain bodies (nodes), each corresponding to a residue, connected by edges related to some measure of interaction. Early AANs were constructed using a physical distance-cutoff (DC) [90], whereby edges are placed between residues that are within a certain distance. This method previously showed that the AANs of proteins generally display small world properties, where few nodes are direct neighbours, but most nodes can be reached in few steps [16]. The benefit of such small world properties, which is likely to be employed by proteins, is the ability to effectively distribute information.

In addition, properties of such networks have been employed to score and subsequently discriminate between native and non-native structures [251]. While insightful, such AANs are considered coarse grained methods, as they only store information concerning the general protein shape. Therefore, although the DC construction technique requires low computational resources, it is at the expense of a failure to accurately model the chemical environment from which more advanced protein functions can potentially be inferred.

In recent models, the network-function relationship has evolved to account for motion, which is required to demonstrate functions such as allostery [87], recognition [27], and catalysis [208]. For an AAN to successfully provide insight into dynamical functions such as allostery, representation of the environment has to be extended beyond that of the DC method in an attempt to elucidate the cumulative effect of side-chain dynamics. The most commonly used dynamical techniques derive edge information from molecular dynamics (MD) simulations. In MD, the trajectories of atoms and molecules simulated by numerically solving Newton's equations of motion for atoms in the protein. From these expensive simulations, edges can be introduced based on the percentage of conformations in which two residues are in contact [208, 23]. Employing MD simulations certainly provides information on the chemistry of the environment, but at an added computational expense. While very insightful, a computationally cheap method that requires only the static structure without sacrificing detail of the protein environment is not only desirable but also achievable, and could have important applications in *de novo* protein design. A method with precisely these properties is presented in Chapter 4 and is used to identify allosteric communication in the G-protein coupled receptor signalling pathway.

2.2 Light Harvesting

The initial stages of photosynthesis, known as the light reactions, have the capacity for ultrafast photophysics by employing optically active molecules such as bacteriochlorophylls (Bchls), to capture, transfer, and utilise photonic energy. Photosynthetic organisms employ PPCs to carry out the elementary light reactions that begin the conversion of solar energy into stable chemical forms [26]. Light-harvesting proteins form quaternary structures and have been developed for the purpose of coordinating the embedded optically active molecules. PPCs are exquisite examples of purpose-built nanostructured organic optoelectronic "devices" and are ultimately employed for a variety of different functions, including energy transfer, charge separation, and photoprotection [26, 230]. The antenna complex is a critical component of the photosynthetic architecture model and acts to increase the cross section for light absorption. Understanding the static and dynamical relationships between their conformal

ensemble and electronic structure will offer a unique insight into how organic molecules can be organised into effective light harvesting arrays.

The photosynthetic apparatus of green sulphur bacteria (GSB) provide a prime example of a modular, interconnected system of PPCs that contains unique examples of PPCs suited to low-light environments [92]. GSB are anoxygenic photoautotrophs [124], allowing them to survive 80m below the surface of the Black Sea. At these depths, photosynthetic organisms receive over 10 orders of magnitude fewer photons than the average plant [201]. These organisms require specialised structures to aid solar energy capture to survive on this scarce energy reserve. The chlorosome, formed by the CsmA protein and a collection of Bchls, is one such structure. This antenna complex contains up to 10,000 molecules of Bchl c, d, and e arranged in higher-order structures with relatively little protein involvement [26]. In addition to the chlorosome, GSB are the sole host of the trimeric Fenna-Matthews-Olson (FMO) complex, whose strikingly robust optoelectonic structure under thermal motions are the subject of Chapters 4 and 5. The FMO complex forms an interface between the chlorosome and the membrane-embedded reaction centre, and is employed largely for the transport of excitation energy between them (Figure 2.2). The baseplate is a planar structure within the chlorosome that contacts the FMO complex directly [173]. The orientation of the FMO complex has been identified using mass spectrometry base foot-printing, where it was found to sit with its C3 symmetry axis perpendicular to the membrane, accepting excitation energy from the baseplate via a peripheral pigment [239, 200]. This work demonstrated that the FMO protein has evolved to drive the excitation close to the reaction centre where this energy initiates the primary photochemical event of photosynthesis; charge separation.

When a quantum of light is absorbed by optically active molecules in PPCs, two charged particles, namely an electron and a hole, are created simultaneously. The bound state of these charged particles gives rise to a neutral quasi-particle known as an exciton. The exciton state contains contributions from a number of pigment excited states. The process of energy migration under these conditions, known as excitonic energy transfer (EET), follows from energetic relaxation, transferring population between exciton states of different spatial extends. EET is a sensitive function of the inter-pigment orientation and distance, which determines their excitonic coupling [51]. Efficient EET therefore depends on several key electronic parameters that are *controlled* by the pigment-protein structure. As shown in Figure 2.2, each monomer of the FMO trimer positions eight pigments that facilitate EET from the giant chlorosome antenna to the reaction centre. The lowest energy transition, known as the Q_y transition, has a transition dipole moment along the y-axis of the pigment, which connects the N_B and N_D chemical groups of the porphyrin ring. As only one type of pigment (Bchl a) is found in the FMO complex, the protein environment uses local pigment-

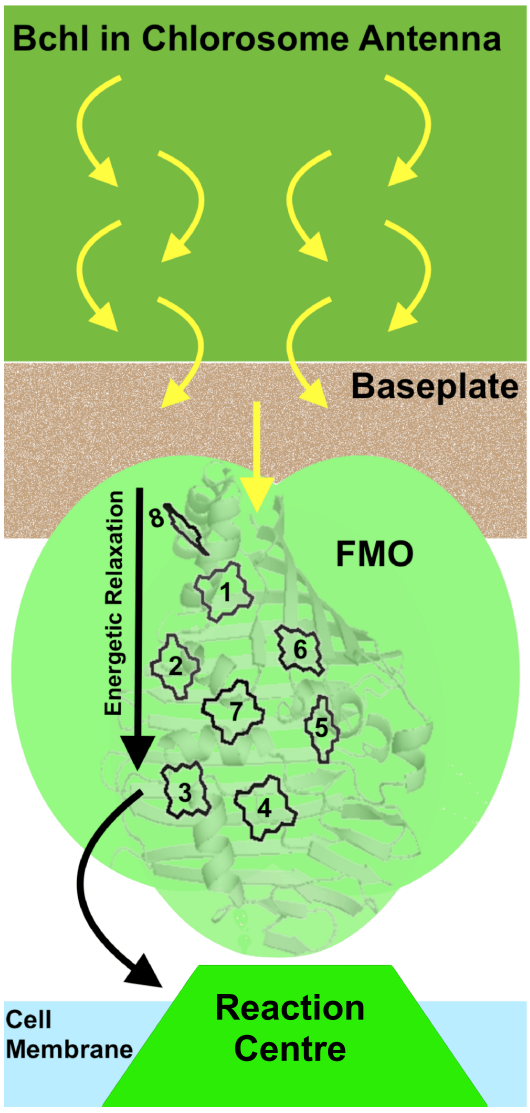


Fig. 2.2 Light absorbed by pigment molecules in the chlorosome migrate through the FMO complex to the reaction centre. This process is driven by dissipative exciton transfer (energetic relaxation) that forces excitations to migrate rapidly under the action of the pigment transition dipole-dipole coupling and emission of low energy environmental excitations, which include intra-pigment or protein vibrational quanta.

protein interactions to shift each pigment's local optical excitation energy (site energy), thereby creating an energetic funnel to direct energy transfer [156]. The modulation of the site energies and the excitonic couplings by the protein dynamics is described by the spectral density. This (excitonic) coupling arises from the Coulomb interaction that occurs between the pigment local excited states. The strengths of these couplings are dependent on the mutual orientation and separation of the molecular transition dipoles. The degree of delocalisation of the excitonic states is a sensitive function of the strength of dipole-dipole couplings and the differences in site energies of the pigments involved, such that coupling strengths significantly greater than the energy differences leads to fully delocalised excitons. Effective excitonic coupling is achieved by rigidly coordinating the pigments with sub-nm precision and fixed relative orientations. Despite the thermal environment, PPCs are so effective that the excited states are best described as delocalised excitons with contributions from $1-3$ pigments [230, 26]. Spatially directed EET is then driven by dissipative interactions with pigment and protein vibrational modes, causing rapid ($\sim$5 ps) energetic relaxation along the excitonic energy funnel to the lowest state localised around pigment three [5, 226]. In addition to the excitonic couplings, the protein determines the dynamical spectral density. This feature describes the strength and time scale of site energy modulations caused by the dynamic, thermal motion of the pigments' molecular surroundings and internal structure and provides the dissipative element required to drive relaxation along the excitation energy gradient. Indeed, energy that is dissipated during excitation relaxation can be absorbed and high-frequency dynamics have important implications in energy transfer [245, 66, 218]. Less understood are the slow dynamics that give rise to the so-called static disorder of the pigment site energies. The slow fluctuation around a mean value gives rise to inhomogeneous broadening in ensemble experiments, where pigments in many different positions are probed simultaneously. This broadening is generally described using a Gaussian distribution function and, as we will see, may have important implications for the optical observed EET dynamics observed in ensemble experiments.

Calculating Features of the Pigment-Protein Complex

The excitonic coupling V_{mn} between an excited electronic state and a pigment in the ground state is conveyed by Coulomb coupling. The respective Coulomb matrix elements of excitation energy transfer reads:

$$V_{mn} = < \phi_{m'}\, \phi_n | V_{Coulomb} | \phi_m\, \phi_{n'} >$$

where ψ_m/ψ_n denote the ground state, $\psi_{m'}/\psi_{n'}$ denote the the excited state wavefunctions of pigment m and n. $V_{Coulomb}$ represents the Coulomb coupling of the electrons and nuclei of

the two pigments. A point dipole can be estimated from the relative direction and strength of the transition dipole moments, which are themselves estimated using the effective dipole strength of the Q_y transition of BChla. The point dipole approximation can thereby be used to calculate the excitonic coupling as the dipole-dipole coupling between the optical transition dipoles. This technique has been employed to investigate the coupling found in the experimental structure of the FMO complex, which were used to test several dynamical theories of EET [5].

Quantum mechanical methods can be used to derive a general expression for the excitonic coupling as the coupling V_{mn} between the transition densities. For each pigment, the transition density is calculated by taking the integral over the product of ground and excited state wave functions using quantum chemistry calculations. The method of Madjet et al. [148] known as TrEsp can then be used to determine the atomic transition charges by fitting the electrostatic potential of point charges to the electrostatic potential of the quantum mechanical transition density, thus containing the full information of the *ab initio* transition density.

From classical theory, we know that the interaction between a dipole $\overrightarrow{\mu}$ and an electric field $\overrightarrow{E}$ can be calculated as:

$$\mathrm{W} = -\overrightarrow{\mu} \cdot \overrightarrow{E}$$

where the ground and excited states are described by their respective dipole moments and the electrochromical shift of the transition energy is obtained as the difference in interaction energies of the excited and the ground state $\Delta E = W_e - W_g$. The shift ΔE_i in the site energy can then be calculated for a pigment in a field of N point charges q_j:

$$\Delta E_i = \frac{1}{\varepsilon_{eff}} \sum_{j=1}^{N} q_j \cdot \frac{\Delta \overrightarrow{\mu_i} \cdot \overrightarrow{r}_{ij}}{r_{ij}^3}$$

where $\Delta\overrightarrow{\mu}_i = \overrightarrow{\mu}^i_{11} - \overrightarrow{\mu}^i_{00}$ is the difference of the permanent dipole moments of ground and excited state of the *i*th pigment, $\overrightarrow{r}_{ij}$ is the the vector connecting the centre of the *i*th pigment with the *k*th point charges q_k, and $\overrightarrow{r}_{ij} = \overrightarrow{r}_j - \overrightarrow{r}_i$ is the vector connecting the center of the *i*th pigment with the *j*th point charge. This so-called point charge density coupling method has been used to calculate site energy shifts due to charged amino acids in the FMO complex and demonstrated the energetic funnel that has evolved to concentrate energy near the reaction centre [4].

Optical Characteristics of the Fenna-Matthews-Olson complex

Structural studies allow the organisation of PPCs and photosynthetic architectures to be visualised. However, building the electronic Hamiltonian of photosynthetic systems requires an

understanding of the EET dynamics. Initial investigations using pump-probe measurements uncovered evidence for coherent EET dynamics in the FMO complex and the light harvesting complexes I and II (LH1, LH2) of purple bacteria [37, 197]. This technique irradiates the system using an initial laser to excite (pump) the system followed by a time-delayed pulse (probe) to monitor the excited state dynamics [230]. The insight from these measurements lies in the differential absorption spectrum; if one transition is bleached there will be a decrease in the absorption at that specific frequency, allowing the exciton dynamics to be seen by the movement of the (bleached) peak. Narrow-band pulse monitors the spectrum of the excitation as it migrates to a lower energy state while broadband measurements allow the entire excitation spectrum at once and a coherent superposition of multiple chromophore sites to be generated. However, as broadband measurements excite many sites simultaneously the resulting data is particularly noisy and loses information of the excitation and emission energies.

Broadband 2D electron spectroscopy (2DES) allows coherences to be probed whilst preserving information regarding the electronic structure of the system. Remarkably, oscillations in the 2DES spectra of the FMO complex were attributed to the coherent evolution of electronic excitations in photosynthetic proteins [64]. These studies were followed by a confirmation of these effects at room temperature [169], suggesting that electronic coherences play a role in EET *in vivo*. Previous theories of EET in FMO have employed models that incorporate the key electronic parameters to varying extents, allowing the prediction of kinetic and optical responses during EET. However, the necessity to include quantum coherence in the EET description introduces the need for new and sophisticated real-time simulations of density matrix evolution, higher precision prediction of key molecular parameters from *ab initio* techniques, and a reappraisal of the electronic and protein structure in terms of its capacity to host and coordinate quantum effects [136, 108, 11, 201].

Several recent theories have pointed to intra-pigment vibronic excited state effects and/or ground state vibrational wavepackets as the origin of the 2DES oscillations [245, 66, 218, 138]. However, relatively few studies have been able to address the much slower, long-range conformal dynamics of proteins that, through the dependence of the pigment site energies on the protein's global charge distribution, lead to an effective static disorder of energy gaps across the conformal ensembles probed in experiments. Positive inter-pigment correlations in static disorder have been suggested to be a characteristic feature of PPCs showing long-lasting signatures of coherence, as they greatly reduce the inhomogeneous dephasing of inter-excitonic beatings [138]. Furthermore, the efficiency of EET will not be affected if the electronic structure of the excited states is maintained across the ensemble, and reliability arising from controlled disorder would be a highly desirable and remarkable achievement in a

self-assembling system of excitonic wires. Fidler *et al*. have presented experimental evidence for such correlated disorder in the FMO complex [68], namely considerable variation in mean exciton energy levels and a constant beat frequency. However, these experiments are unable to distinguish between the intra-pigment vibronic excited state effects [245, 66, 218] and correlated disorder as the source of the electronic coherences.

Although direct, unambiguous evidence for the beneficial role of quantum effects in photosynthesis remains elusive, an experimental example of how transient coherence can be crucial in organic photophysics has recently been presented by Gélinas et al. [84], where ultrafast (< 80 fs) electron-hole separation at a bulk heterojunction were enhanced under well-ordered acceptor phases that support delocalised electron states. The ultrafast, coherent electron dynamics lead to sufficiently rapid charge separation that can overcome the strong Coulomb binding energy that otherwise traps charges close to the heterojunction interface and drastically reduces their internal quantum efficiency. As has been understood for a long time, the structure and motion of the protein environment is fundamentally important for understanding how quantum dynamics might impact on EET efficiency and has led quantum biologists to explore the surprising stability and role of quantum coherence in these structures. Acquiring a better understanding of these properties could provide interesting insights for other emerging types of quantum technology [136, 108, 11, 201].

2.3 Computational Methods

Theorists employ computational analysis to expand the power of purely theoretical models, which can only be solved exactly for simple systems, to more complex structures. Currently, an arsenal of computational techniques allows different aspects of experimental results to be interpreted and further explored through computer simulations. Experiments lie at the core of such theoretical investigations, as they establish the natural phenomenon from which generalisable models can be developed. In structural biology, x-ray crystallography is an important tool that is used to elucidate the atomic structure of a protein. This molecular structure alone can provide key insights into the observed behaviours, as is the case with the oxygen-transport protein, haemoglobin [192]. However, proteins are flexible objects and the crystal structure represent a static average of a continuum of energetically-similar conformations [242]. Information regarding this conformal ensemble, and the dynamics that are populated within it, are required to fully elucidate the link between a protein's structure and the observed function. Several techniques have been devised that can be applied to the experimentally derived structure to calculate these dynamics.

Elastic network models (ENMs) gained popularity following the experimental observation of functionally relevant collective and delocalised modes of motion in proteins [85]. While the motion of proteins in solution is highly complex, the success of NMA follows the observation that structural changes are dominated by the inter-residue contact topology and the most probable modes observed *in vivo* are those that require the smallest energy ascent in the energy landscape. In normal mode analysis (NMA), the atomic potentials are replaced with uniform harmonic potentials between interacting residues [227]. This approach is commonly coarse-grained whereby only the backbone of the polypeptide chain is used in the analysis. ENM has provided useful solutions for problems in structural biology, including structure prediction and investigations of large-amplitude and long-time scale functional dynamics [15]. However, ENM removes the high-frequency and anharmonic motions and protein function depends on these effects [147]. The breakdown of the harmonic approximation at low frequencies has been highlighted and suggests that the impact of very soft, slow conformal motions can only be accounted for phenomenologically [183].

All-atom simulations extract chemical details from the intra-protein environment to generate the conformal ensemble. Molecular mechanics (MM) use simple potential-energy functions, such as harmonic oscillators or Coulombic potentials, to model molecular systems. Such models are built using spheres as atoms and springs whose properties vary according to the bonds that they represent. In molecular dynamics (MD), Newton's equations of motion are solved numerically to generate the movements of atoms. This approach is preferable to NMA, as the time evolution of conformal motions can be investigated and kinetic and thermodynamic information can be obtained. However, NMA has the advantage of increased sampling efficiency, as MD is unable to completely sample the conformal ensemble on biologically relevant timescales [3]. An alternative, all-atom model is therefore desirable in particular cases where anharmonic, high-frequency motion of long-time scale dynamics is needed to elucidate protein function. This is true of PPCs, where the systems are large and function on timescales beyond the capacity of MD investigations.

Constrained geometric simulation [235] is an all-atom simulation technique that has roots in graph theory and percolation theory, allowing concepts of rigidity to be used to efficiently sample the conformal landscape associated with an experimental structure. The Lagrangian constraints-based approach flattens the potential energy surface, allowing the conformal ensemble to be efficiently explored through random perturbation whilst sustaining the fixed covalent and non-covalent constraints identified in the experimental structure. Critically, although there are no associated timescales of motion, a wide range of functions have been successfully investigated, suggesting that the dynamics associated with the experimental structure naturally emerge from the identified geometric structures. This approach can be

combined with principal component analysis to identify collective, long-time scale, large-amplitude motions that, while similar to those identified by NMA, are not limited by the harmonic approximation or coarse-grained abstraction.

The above models are able to simulate different aspects of protein motion. For example, global dynamics are generally used to identify large-scale allosteric changes but fail in the description of local environments, where MD simulations provide more accurate insight [14]. In general, their treatment of proteins as classical objects is limiting in certain cases. Indeed, to investigate the optical properties of pigments embedded in PPC, quantum chemistry calculations are required to investigate electronic structure. Density functional theory (DFT) has been developed in recent years to accommodate the large systems found in biology. Using this technique, the excited states of pigments embedded in PPCs can be investigated [44] allowing simulation of their optical absorption spectrum.

2.3.1 Investigations

The network of non-covalent interactions is central to constrained geometric simulation. This observation prompted the results in Chapter 3, where network construction using these geometric structures are used to investigate allosteric communication. The dynamical functions that are able to be investigated using only the amino acid network suggest that much of the information that can be drawn from the resulting constrained dynamics are encoded in the network of bonds.

Previous dynamical studies of the FMO complex using MD simulations [167, 205, 120, 116, 111] uncovered no evidence for correlated pigment spatial or site energy fluctuations [167]. It is important to note, however, that these studies focused on the high-frequency portion of the spectra due to limited time scales that can be accessed using MD. NMA has been employed to overcome these limitations and compute the low-frequency portion of the spectral density [183]. While the NMA model of a single monomer of the FMO complex does reveal significant correlation among the weak site-energy fluctuations arising from the lowest frequency modes (period ~ 2 ps), particularly between pigments 3–4 and 1–2, these spatial correlations do not play a role in energy transfer. It may be that even lower frequency modes that spread over the trimeric PPC structure are likely to induce correlated site energy variations that could be yet larger and classified as truly static disorder.

In Chapter 4, constrained geometric simulation is used to generate a conformal ensemble of the FMO complex. By employing this all-atom technique, dynamics beyond the time scales typically associated with MD simulations to be investigated, providing insight into the anharmonic, large-amplitude motions that are central to its function as a PPC. Correlation analysis is employed to study the motions involving the embedded pigments and protein

substructures that modify their optical properties. However, spatial correlations do not necessarily guarantee correlated static disorder in the energetic structure, as the pigment site energies are the result of a complex interplay of electric fields and pigment-protein interactions. Principal component analysis allows the identification of the collective and functionally relevant motions within the large conformal ensemble and has been used to identify the largest-variation in the conformal ensemble of the FMO complex. This method ensures that as large a dynamical range as possible is captured for the analysis of the excited state energies, which are calculated using fully quantum mechanical calculations of extracted 2000 atom clusters centred around pigments 3 and 4. Recall, the presence of inter-pigment correlations in the static disorder support the observed long-lasting signatures of coherences between these pigments [138, 169]. The excited state energies along the trajectory, which has been used to approximate the static disorder found in optical experiments, do indeed display positive inter-pigment correlations in their site energies in line with this theory. The results suggest that the novel combination of all-atom constrained dynamics with state-of-the-art *ab initio* electronic structure calculations are a powerful approach for calculating the static disorder in PPCs.

Chapter 3

Residue Geometry Network

Outline

Geometric simulation is an all-atom technique that uses the physico-chemical characteristics of proteins to identify a set of covalent and non-covalent interactions. These constraints can be used to simulate the conformal ensemble associated with the protein's structure, which describes internal degrees of freedom that are relevant to their function. Despite the lack of force fields, such as those used in Newtonian approaches, the Lagrangian constraints-based approach is able to recapitulate a broad range of functional dynamics [235, 236, 22]. Simulating the associated dynamics is trivial, as it involves simple random perturbations of the atoms that satisfy the fixed constraints. The relevant information is thereby largely encoded in the network of constraints.

Networks have become a popular method for studying the structure of proteins and are constructed using nodes and edges to represent residues and the specific interactions that occur between them, respectively. Computationally cheap methods often employ the static distance-cutoff (DC), where a residue is connected to all residues that fall within a user-identified sphere. Herein, a novel static amino acid network is proposed, termed the Residue Geometry Network (RGN), which is built using the fixed-constraints identified by geometric simulation [235]. As this network information can be used to simulate functionally-relevant dynamics, it is assumed that the RGN can be used to investigate dynamical functions, such as hinges and allosteric communication, which cannot be studied using the static DC methodology.

To verify this construction technique and build a version of the RGN that contains weighted edges, the evolutionary properties of residues belonging to 795 proteins are correlated with network measures of residue importance. Using the RGN, the expected visiting time, which is calculated using random walks through the network, is measured for each

residue. This measure of the transfer of information through the RGN is able to identify allosteric residues in several proteins involved in GPCR signalling. Comparisons are made with the DC method that show that the dynamic function of these residues remain hidden in the, also static, commonly used approach. The results presented here therefore suggest that, despite being a computationally cheap, static method, the RGN, which incorporates physico-chemical characteristics of the protein structure, can be employed to investigate protein functions born from dynamic behaviour.

3.1 FIRST

Graph theory involves the mathematical formulation of the pairwise relationship between objects [240]. In 2D, a lone body (vertex) has 3 trivial degrees of freedom (DOF), namely, one rotational and two translational. Multiple bodies connected by bars (edges) have internal, non-trivial DOF that describe dynamical motions within a structure. Depending on the connectivity of the edges, the body-bar structure can be described as either flexible or rigid with respect to the presence or absence of non-trivial DOF, respectively. This structural rigidity can be assigned to graphs using Laman's theorem [135], which states that any graph G with n vertices and m edges will have the following structural properties:

$m > 2n-3$	Flexible
$m = 2n-3$	Minimally rigid
$m > 2n-3$	Overconstrained

Minimally rigid describes the state of a graph whereby if one edge is removed from the graph it will become flexible, that is, having some internal DOF. Over-constrained graphs have redundant edges whose removal do not change the rigidity properties.

The rigidity of a body-bar graph can be calculated using an algorithmic implementation of the pebble game [214]. In the pebble game, graphs are built one bar at a time allowing the total number of floppy modes, or DOF, to be counted exactly. Using Laman's theorem, all of the rigid clusters (RCs) can be identified. A RC describes a subgraph of bodies within a graph that are rigid. This basic principle of the pebble game extends to 3D, described by Tay's theorem [222], where the minimally rigid basis is replaced by $m = 6n-6$ as there are now 6 trivial DOF (Figure 3.1).

Using the pebble algorithm, a rigidity analysis of the graph theoretical representation of the protein structure allows flexible and rigid regions to be quantified [235]. Knowledge of these structures has important applications in protein dynamics, as the internal DOF have

been found to play an important role in protein function [9]. Identifying RCs in the protein architecture is a central concept in geometric simulation. To do so requires projecting the protein's structure from experimental data, most commonly crystallographic, onto graphical space.

The 3D model of a protein is constructed using the body-bar-hinge framework, similar to the ball-and-stick model, where each atom is represented using a body (ball, vertex) and a link between the two interacting atoms is represented using a bar (stick, edge). In geometric simulation, each bar removes 1 DOF. A covalent bond is modeled as a hinge (5 bars). This hinge leaves 1 (dihedral) rotational DOF between 2 atoms, which have a fixed bond length. 6 bars connect bonds that cannot be rotated, such as peptide bonds and double bonds, which have higher energies of rotation. As rotation of terminal hydrogens does not generate a different conformation these are also connected using 6 bars [235]. In addition to covalent bonds, proteins employ non-covalent interactions to define secondary and tertiary structures. To calculate the internal DOF in protein systems therefore requires knowledge of the geometric and chemical factors in the protein.

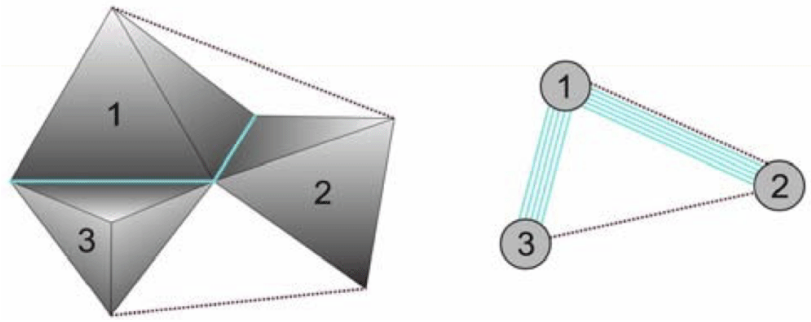


Fig. 3.1 Body-bar-hinge frameworks using the multi-graph representation. The tetrahedrons represent rigid bodies in the body-bar-hinge (left), which themselves can take different shapes, from portions of residues to several secondary structure elements. Bars (dashed lines) connect any points belonging to rigid bodies and remove 1 DOF, while hinges (solid lines) are located at common edges of tetrahedrons remove 5 DOF. In the associated multi-graph (right) the body, bar and hinge are represented by a node, (dashed) edge and five (solid) edges (connecting nodes belong to the interacting bodies), respectively. The above example of a body-bar-hinge framework is minimally rigid (having no redundant bars or hinges) corresponding to the fact that its associated multi-graph is has exactly 6n-6 edges, and its sub-graphs have maximum 6n'-6 edges, where n and n' are the numbers of nodes in the graph and sub-graphs, respectively. This image has been taken from Ref. [83].

Constrained geometric simulations are performed using the FIRST (Floppy Inclusions and Rigid Substructure Topography) software package [65, 236]. FIRST identifies hydrophobic tethers and quantifies the strengths of hydrogen bonds and salt bridges using a geometry-

based scoring scheme developed by Dahiyat et al. [53]. When calculating the energy of hydrogen bonds, the donor-hydrogen-acceptor geometry is used to calculate their energies, the strongest of which typically lie between -5 kcal/mol and -10 kcal/mol. The energy is computed using an equation developed by Mayo et al. [53]. This equation has a term that depends on distance and angle:

$$E_{HB} = D_0[\,5(\frac{R_0}{R})^{12} - 6(\frac{R_0}{R})^{10}]\,F(\theta,\Phi,\phi) \tag{3.1}$$

For sp^3 donor-sp^3 acceptor: $F = \cos^2\theta\cos^2(\Phi - 109.5)$
For sp^3 donor-sp^2 acceptor: $F = cos^2\theta cos^2\Phi$
For sp^2 donor-sp^3 acceptor: $F = cos^4\theta$
For sp^2 donor-sp^2 acceptor: $F = cos^2\theta cos^2(max[\Phi,\phi])$

where R_0 = 2.8 Å (hydrogen bond equilibrium distance), D_0 = 8 kcal/mol (well-depth), θ is the donor-hydrogen-acceptor angle, Φ is the acceptor-base angle (e.g. for carbonyl oxygen it is the main-chain carbon), and ϕ is the angle between the normal of the plane (defined by the 6 atoms) and the sp^2 centre (of the three sp^2 hybridised orbitals) (Figure 3.2). The greater bond strength and reduced directionality of the salt bridge are reflected in the use of a different energy function that includes a weak Coulombic term [53]. Hydrogen bonds and salt bridges are modeled as hinges that allow dihedral rotation about the bonding axis. FIRST places hydrophobic "tethers", for which the energy is not calculated, between aromatic or aliphatic sidechain carbon atoms that are within 4 Å of each other. This type of contact imposes 2 bars that represent a distance constraint. Interactions with metals and other ions are treated as covalent bonds.

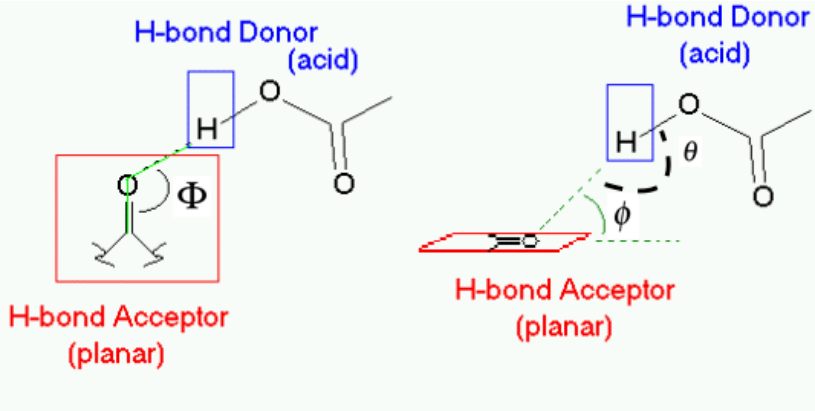


Fig. 3.2 Angles defined in Mayo et al. [53] used to calculate hydrogen bond energies.

Using the non-covalent and covalent interactions to build a graph, FIRST identifies flexible and rigid regions in the protein framework and thereby provides insight into the long-

range connectivity in the protein. As described above, the hydrogen bond energies are scored, which allows less influential bonds to be removed from the analysis. The hydrogen bond energy cutoff (H_{cut}) controls the strength of the hydrogen bonds and salt bridges involved in the rigidity analysis. As hydrophobic interaction energies are not explicitly calculated by FIRST, varying H_{cut} does not remove any of the hydrophobic tethers from the simulation. For example, if the H_{cut} = -2.0 kcal/mol, only hydrogen bonds and salt bridges with energies more negative than -2.0 kcal/mol would be included in the simulation, in addition to all of the hydrophobic and covalent edges.

Constraint Network Analysis

Including all of the interactions (H_{cut} = 0 kcal/mol) in the rigidity analysis of protein structures generally results in a single "giant" RC that dominates the majority of the protein. As the bonds are removed (by lowering the H_{cut}), a sharp phase transition is found in the decomposition of the network as the giant RC suddenly breaks. In proteins, in comparison to glasses, this percolation behaviour is more complex. Due to the highly regular array of stable hydrogen bonds, RCs at moderate values of H_{cut} tend to involve secondary structure elements [122].

As the value of H_{cut} is lowered and the RC decomposition becomes increasingly flexible an analogy with thermal unfolding naturally arises. Constraint network analysis (CNA) [133] behaves as a front and back end to the FIRST software. CNA can be used to sequentially remove (break) the constraints by lowering the H_{cut}, thereby simulating thermal unfolding. This analysis allows one to identify temperatures associated with global and local transitions. The H_{cut} is related to the temperature using an empirically determined linear function [179]. The increased functionality allows the quantification of macromolecular stability using FIRST generated networks via two main measures; (1) the percolation index (p_i) is the energy at which a residue separates from the giant RC and (2) the rigidity index (r_i) measures the energy when the residue segregates from a RC. Unfolding nuclei, namely sites during the thermal simulation that give rise to a phase transition, are found to correlate with structurally weak regions in experiment [241].

3.2 Rigid Clusters and Evolutionary Space

Mutations to the protein sequence perturb the non-covalent interaction network, impacting protein flexibility and dynamics. The dependence of protein structure on sequence gives rise to the observed weak, albeit statistically significant correlation between sequence entropy

and conformational mobility [144]. In line with these findings, the conformal space of a single protein domain explored in molecular dynamics (MD) simulations is found to strongly overlap with the conformal space explored by homologous, that is evolutionarily related, structures in a protein family [231]. An interesting difference observed between the former (single protein) MD space and the latter evolutionary space is the less complex behaviour of the evolutionary space, where often only a few deformable modes are explored. It follows that protein function restricts the access of some flexibility patterns to evolution. This prompted an investigation into the geometric structures and their involvement in evolutionary behaviour.

Regions of a protein that undergo large structural deformations coincide with evolutionary changes in structure [149]. This is true for the human fibroblast growth factor receptor 1 (HFGFR1) kinase domain. Using CNA, a positive correlation is found between evolutionary dynamics of regions in the HFGFR1 kinase domain and the percolation index (Figure 3.3). This is best exemplified by the flexible regions around residues 130 and 190 that correspond to regions with large fluctuations in evolutionary dynamics. The low evolutionary dynamics of the RCs suggests that geometric structures are able to constrain not just conformational space, but evolutionary space as well.

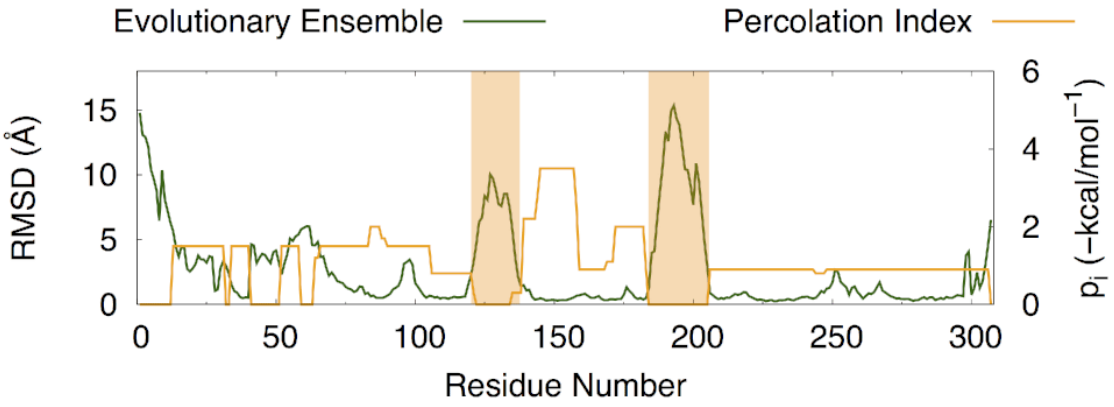


Fig. 3.3 RMSD of the evolutionary ensemble for HFGFR1 kinase domain [149], containing 12 non-redundant crystal structures. In the above graph, the percolation indices (p_i) have been overlayed. Highlighted examples of regions in HFGFR1 that display higher flexibility are found to have higher evolutionary dynamics are shaded orange. Note that a percolation index of 0 kcal/mol signifies that the region is never part of the giant percolating cluster.

To further investigate the relationship between sequence entropy and evolutionary dynamics a dataset of roughly 500 proteins (adopted from Ref. [76], trimming proteins with incomplete structures) has been analysed using CNA. In particular, the rigidity index (r_i) has been measured for all residues in the protein data set, allowing the residue to be binned into one of the 0.5 kcal/mol range bins extending from 0.0 kcal/mol to -6.0 kcal/mol. By

measuring the evolutionary rate of each bin, an estimation of the selection pressures imposed by this physical measure is made by building a signal using the large number of residues in each structural bin. The identified R value of 0.97 (linear fit) suggests that residues belonging to stronger RCs across the dataset have lower evolutionary rates. The above analysis of the rigidity and percolation indices provided the motivation to investigate the amino acid network (AAN) generated by FIRST in greater detail.

A Comment on Low-Frequency Motion

An important aspect of protein function is low frequency motion. These fluctuations play an important role in proteins since the large amplitude movements require a relatively small ascent of the energy landscape. The total energy of a protein is shared equally among the modes, meaning that these modes account for the largest fluctuations [15]. These collective motions are also notoriously associated with function [81]. Unsurprisingly, low-frequency normal modes, which emerge from a protein's structure, are found to be conserved in homologous structures and protein families. What *is* unexpected is the nature of their conservation. Namely, the strong conservation of normal modes arises from their high robustness to mutation [62]. This is not the expected trend; a biological explanation for their conservation would involve a discussion of environmental selection pressures.

Schaper and Louis [198] have investigated evolutionary behaviour in ribonucleic acid (RNA) structures, and concluded that the likelihood for a RNA structure, and therefore the associated phenotype, to appear in a population is biased by the frequency by which the RNA sequences map to particular structures. Namely, structures that are found to be coded by many sequences are more likely to be found in nature, even in cases where they do not correspond to the most "fit" phenotype. If this finding generalises from RNA structure to protein structure, then the robustness of low-energy protein normal modes to sequence variation holds further information. As low-energy motion supports the greatest amount of sequence variation [62], more sequences map to this, often functional [81], motion. This will have important implications for the appearance and evolution of protein function.

A speculation can be made in light of the above observations; the robustness of the low frequency modes to mutation will *bias* a function to appear in the conformational space it has come to consistently occupy. This protection against mutation would have a profound effect on the evolution of a protein (in a similar manner to which RNA phenotypes are biased [198]). For example, the incorporation of a function into a stable mode will safeguard that function to many random mutations. As a result, the so called "neutral space", or evolutionary space that does not affect the primary function, will be larger than if the function were located on higher energy modes that are less robust to mutation. The large neutral space can then

provide a platform to evolve secondary functions or improve aspects of the primary ones through dynamical correlations. Using these principles, a useful application of AANs may lie in *ab initio* protein design, where identification of the sparse network of residues that give rise to low frequency, functional motion, could in turn identify the large neutral space in proteins where new functionality or increased stability could be built.

3.3 Residue Geometry Network and Distance Constraint Network Construction

As discussed in chapter 2.1, amino acid networks (AANs) are a commonly employed technique for abstracting the protein structure and the application of traditional network analysis can provide a wealth of information regarding catalytic residues [252], allostery [24], protein design [251], and, more generally, the relationship between a protein's structure and its function [247]. However, construction of the AAN from the static crystal structure provides no information about dynamics and assigning edge weights using MD simulations, while providing useful dynamical information, is costly. The diverse range of functions [22, 115, 221, 145, 72] that can be investigated by FIRST suggests that the biologically relevant dynamics naturally emerge from the simple network of interactions that underlie the constraints-based approach. The specific advantages and disadvantages of the protein dynamics simulated using FIRST will be expanded upon in the following Chapter but, in brief, FIRST-generated conformal ensembles are able to show allosteric transitions [130] and enzyme functions on ms timescales [238]. This observation suggests that the construction of an AAN using FIRST geometrical analysis of the static structure will have a "pseudo-dynamical" character. Similar to the use of FIRST-generated interactions as constraints in geometric simulation, it is hypothesised that to a good approximation the supported dynamics of the protein structure is *encoded* in the interactions identified in its native state. If correct, this would allow dynamic functions to be investigated using the static structure.

The parameters of the residue geometry network (RGN) are built by identifying the strongest non-bonded interactions using FIRST [235, 236], which, as discussed above, identifies hydrophobic tethers and quantifies the strengths of hydrogen bonds and salt bridges using a geometry-based scoring scheme applied to the static experimental structure. The information stored in the scoring function allows the (a) implementation of an energy cutoff such that only the strongest hydrophilic interactions are involved in the network and (b) construction of the weighted and unweighted networks. A similar approach to network construction has been employed previously using the computational tool BONGO [38],

which predicts the structural effect of single residue polymorphism. In their approach, nodes are removed iteratively to identify those that participate more strongly in building up the edges of a graph. By comparing wild-type and mutant proteins, BONGO is able to identify disease-associated mutations. Impressively, this algorithm is able to distinguish between disease-associated and non-disease-associated mutations with a positive predictive value and negative predictive value of 78.5% and 34.5%, respectively. Here, this approach is built upon using the FIRST algorithm to estimate the energy of hydrogen bonds and salt bridges, facilitating the removal of less influential hydrogen bonds from the graph and allowing weighted networks to be built for allosteric analysis.

In what follows, the RGN will be shown to recapitulate characteristics of protein structure and investigate dynamical functions. Evolutionary rate (*dN/dS*) is calculated as the ratio between non-synonymous mutations in protein coding genes (*dN*), which change the amino acid sequence and are a function of the selective pressures, and synonymous mutations (*dS*), which do not affect the amino acid sequence and therefore remain neutral with respect to selection pressure. Using a previously assembled comprehensive data set of 795 proteins [76], the RGN is used to investigate whether residue centrality is a major constraint on residue evolutionary rate. For the unweighted RGN (unRGN), where all network edges are assigned equal weights, a strong, negative correlation is identified between degree, betweenness, and closeness centrality measures and the evolutionary rate. Using the weighted RGN (wRGN), the same trend is found, as well as an increase in the weighted betweenness centrality correlation when compared to the same correlation measured using the unRGN. The importance of added chemical insight is demonstrated using more complex network analytics to study dynamical functions. For example, residues that form few local connections while maintaining high global centrality are, unexpectedly, found to be more highly conserved than hub residues. The subtle dynamical role played by these residues, whose corresponding nodes form hinges in the RGN, is investigated using several proteins from the data set. To develop the theme of deriving dynamical functions from static structure, the expected visiting time, which measures node signal traffic during random walks through the network, is employed to investigate the allosteric response of rhodopsin to light absorption. This method has previously been used in combination with molecular dynamics simulations to identify residues that regulate allostery [170]. Despite only using the crystal structure, residues that score high expected visiting time in the RGN are found to overlap strongly with well-known regulators of the allosteric response. These residues are often not identified when applying expected visiting time to the traditional distance-cutoff AANs. The code for RGN construction and analysis has been released with the aim of broadly empowering the

scientific community with a low-cost approach to understand, modify, and design protein structures.

Construction

The script for for building and analysing the RGN is written using the Python coding language [20]. Network analysis was achieved using the python package NetworkX [93]. For each protein, the corresponding network consisted of nodes, each representing one residue, and edges that corresponded to a particular interaction. In an unweighted network, all of the edges have the same weight while in a weighted network the edges are assigned weights according to the characteristics of the interaction they represent. When constructing the RGN, covalent interactions were represented by an edge between (a) adjacent residues and (b) cysteines connected via disulphide bridges. The geometric tool FIRST is then used to generate the non-covalent edge components of the network. Proteins analysed by the RGN are downloaded from the protein data bank (http://www.rcsb.org). For FIRST analysis, hydrogen atoms are added to proteins using the Reduce module within Molprobity [56]. Reduce adds hydrogens to ligands, but does not add explicit H atoms to water molecules. As a result, in the cases where H atoms of water molecules were not already present in the PDB file when downloaded, the water is removed. The data set contains both X-ray crystallographic and NMR data. In the case of NMR structural ensembles, the "best conformer", as identified in the PDB file, was used. If a specific conformer of the NMR ensemble was not specified in the PDB file, the first conformer was selected. The default settings of FIRST (syntax -non) were used during the construction of the RGN. The distance constraint (DC) AAN is constructed using the Bio3D suite [88], whereby edges are placed between a Cα atom and other Cα atoms that lie within an imposed DC (in Å). The closeness centrality properties of this unweighted network can then investigated in an identical manner to the RGN, as detailed below.

Network Analytics

Centrality attempts to identify nodes that are the most important, or influential, in a system. The centrality of a particular node can be measured using several different algorithms that highlight different aspects of the network. Degree centrality is simply the number of edges connected to a node. The assumption here is that a better connected node will be more important. However, this measure of centrality does not consider the position of the node relative to other nodes in the network. Betweenness centrality and closeness centrality are more complex measurements that account for the topology of the network. Betweenness

centrality measures the involvement of a node in the shortest paths between all other pairs of nodes in the network. This measure is, generally, useful for identifying nodes that play a role in the flow of information. This property is calculated according to the equation:

$$C_B(v) = \sum_{s,t \in Paths(V)} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \tag{3.2}$$

where V is the set of all nodes, $\sigma(s,t)$ is the number of such shortest paths between nodes s and t, and $\sigma(s,t \mid v)$ is the number of such paths that involve node v. Weighted and unweighted betweenness centrality are computed equivalently, with path lengths derived from weighted and unweighted networks, respectively. Weighted path length is computed according to:

$$d_{ij}^{w} = \sum_{a_{uv} \in g_{i \leftrightarrow j}^{w}} f(w_{uv}) \tag{3.3}$$

where f is a map from weight to length, a is the connection status of u and v, and $g_{i \leftrightarrow j}^{w}$ is the shortest weighted path between i and j. Closeness centrality is the inverse of the average shortest path length between residue i and all other residues in the network, according to the equation:

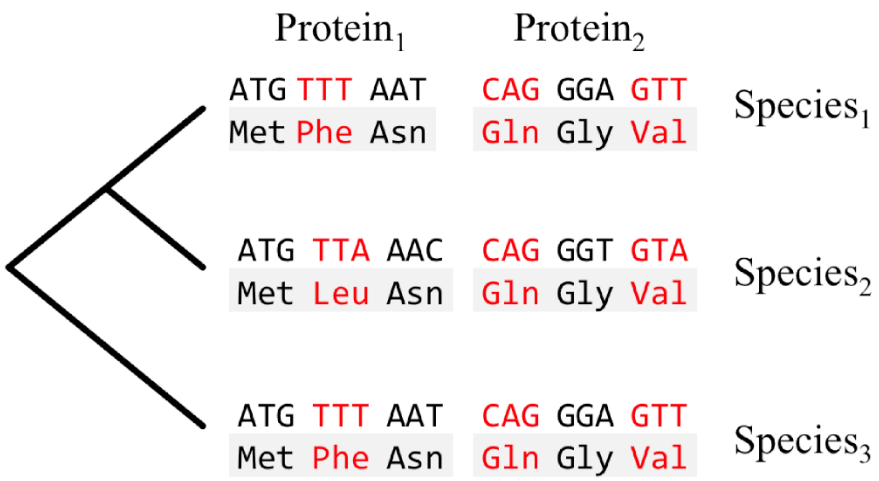
$$C_C(i) = \frac{n}{\sum_j d_{ij}}, \tag{3.4}$$

where n is the number of nodes in the system and d_{ij} is the shortest path length between nodes i and j.

3.3.1 Evolutionary Analysis

The PAML software package [248] has been employed to calculate the number of residue substitutions (*dN*) and the number of silent substitutions (*dS*), for a set of codons where the latter acts as a normalising factor. The ratio of these two factors, known as the evolutionary rate (*dN/dS*), provides insight into the rate of selection normalised by mutations at the DNA level [151]. The data set used by Franzosa and Xia [76] is employed to investigate the relationship between centrality and evolutionary rate, consists of 795 proteins derived from structural homology mapping of yeast (*Saccharomyces cerevisiae*). In particular, the multiple sequence alignments were calculated using ClustalW to generate an alignment between a translated open reading frame (ORF) from *Saccharomyces cerevisiae*, the mapped protein structure subunit sequence, and orthologous ORFs from *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus*. PAML is then used to calculate the

evolutionary rate. As the 4 species are closely related, a single value of the evolutionary rate was calculated for the entire tree (Figure 3.4). In particular, the module *codeml* within PAML is employed to calculate *dN/dS* using the tree [S. cerevisiae, S. paradoxus], S. mikatae, S. bayanus. *dN/dS* displays a greater validity for a larger data set [161], and *dN/dS* for individual residues does not provide realistic insight. To improve the accuracy of our measurements, residues were sorted into bins to increase the number of codons being analysed and provide a suitably large signal, as has been done in previous experiments [76].



$Species_1$: TTT CAG GTT

$Species_2$: TTA CAG GTA

$Species_3$: TTT CAG GTT

Fig. 3.4 Codon alignment for computing dN/dS for residues of a particular structural class (using tree topology given above), not by protein co-membership.

As FIRST uses an all-atom representation when identifying interactions, each node represents the interactions made by the atoms forming a particular residue in the polypeptide chain. The centrality of every node, and thereby each residue position in the dataset, was calculated, providing a dataset of 264,773 residue positions. Within *each* protein, the centrality of all the residues was calculated, allowing all residues to be ordered and partitioned into 20 bins according to their centrality value. Each of the bins represents 5% of the centrality values within a given protein; residues with the top 5% centrality values are assigned to bin 100, the highest 5-10 % are assigned to bin 95, and so on, resulting in 20 'rank' bins for each of the 795 proteins. The evolutionary rate within each bin (summed over the data

set) can then be calculated as outlined above to study relationships between centrality and evolutionary rates (Figure 3.5).

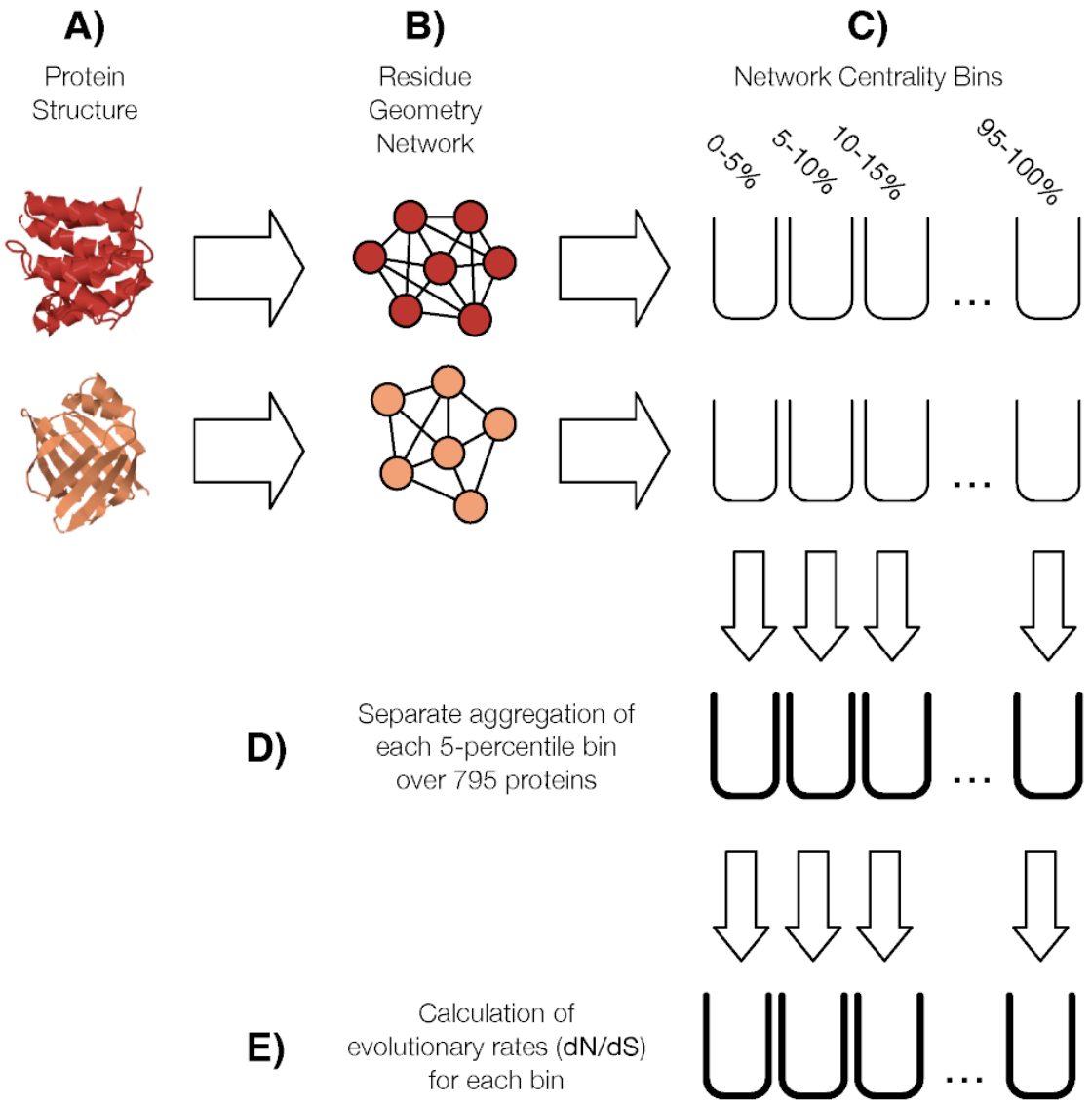


Fig. 3.5 Evolutionary analysis method. For each protein, the centrality for all residues is calculated and assigned to one of 20 bins depending on the centrality within each protein. Equal 5-percentile bins are then aggregated, allowing an accurate measure for the evolutionary rate to be calculated for each of the 20 summed bins. An identical procedure has been previously employed [76] using this data set, whereby a strong signal is attained by binning residues according to their relative solvent exposure, and the *dN/dS* is then calculated to look at "bin evolution".

The correlation coefficient can be calculated between the bin evolutionary rate and bin number (Figure 3.6). Note that if residues are randomly assigned to the ranked bins a correlation coefficient of 0.07 is found. To ensure validity of the results, the correlation

coefficient of this plot was calculated while manually varying certain parameters (*e.g.* H_{cut}). This allowed identification of the values that resulted in the best performance (strongest correlation) of the construction methods. To construct the wRGN for analysis using the weighted betweenness centrality metric, covalent and hydrophobic network weights were identified that optimised the correlation with the evolutionary data. Covalent energies have been varied between 0 and -40 kcal/mol and hydrophobic energies between 0 and -4 kcal/mol. Increasing the search range beyond this limits would not improve the agreement with evolutionary rate.

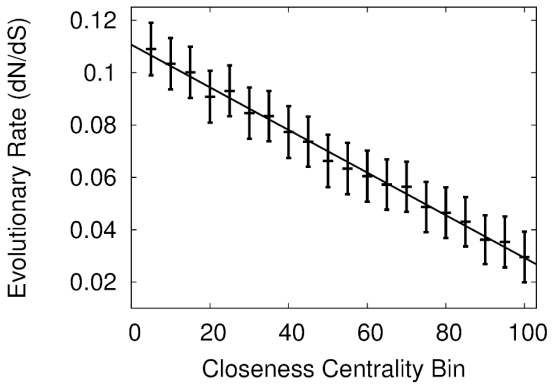


Fig. 3.6 The correlation coefficient between the evolutionary rate and the centrality bin is used to assess whether the different forms of centrality influence the evolutionary rate of the residues in the data set. The Pearson correlation coefficient is -0.997 using an unRGN network with H_{cut} = -3 kcal/mol between the bin evolutionary rate and the closeness centrality bin. The trend line for the data is shown in black, with standard error bars displayed for each calculation.

Correlation between Centrality and Evolutionary Rate

As mentioned, the RGN is constructed using FIRST-generated non-covalent interactions, therefore retaining the inexpensive computational resources of widely used DC techniques while providing additional insight into the geometry and chemistry of the residue environment. The relationship between high centrality residues, which have structural or functional significance [57], and evolutionary rate is investigated using the RGN and DC methods. In particular, the AAN for the protein data set is calculated using a DC of 4, 6, 8, and 10 Å. When constructing the AAN using a DC of 6 Å, an edge is placed between a residue and all residues that lie within 6 Å (measured between Cα atoms). This is a commonly used technique for the analysis of protein structure [246]. In a previous study [76], which employed the same data set, residue buriedness from the solvent was calculated and correlated with the

Table 3.1 Correlation coefficient for the best performing value of H_{cut} for the unRGNs (rows 1-3) and wRGN for betweenness centrality (row 4).

Centrality Measure	H_{cut} (kcal/mol)	Correlation Coefficient
Degree	-2.0	-0.994
Betweenness	-2.5	-0.995
Closeness	-3.0	-0.997
Weighted Betweenness	-2.5	-0.997

evolutionary rate. A correlation coefficient of -0.996 was identified between buriedness and evolutionary rate. In the analysis, the closeness centrality bins grouped using a DC of 6 Å also resulted in a correlation coefficient of -0.996 between the closeness centrality bins and evolutionary rate. Of course, these two physical observables are related, as residues in dense regions of the protein are likely to be shielded from the solvent. While residue density (as calculated by DC) and buriedness from the solvent, are strong selective pressures, they do not account explicitly for residue-residue interaction strengths. Therefore, such techniques cannot provide an understanding of the interaction network that stems from the chemistry of the environment. The RGN is therefore used to design a static AAN construction technique that reflects these important aspects of the system.

Certain parameters of the simulation are varied to investigate the effect on the accuracy of the RGN using the correlation coefficient between centrality bin and evolutionary rate as a meter, where a stronger correlation signified better performance of the RGN. In the unRGN, non-covalent and covalent interactions were all given the same weight and so the number of interactions is the only quantity that can be varied. The correlation between centrality and evolutionary rate was therefore optimised as a function of the H_{cut} (Figure 3.7A), which controls the minimum energy of a hydrogen bond or salt bridge that will be involved in the analysis. The unRGN is a coarse approximation; the interactions that occur between protein residues in nature vary strongly, which is not reflected in the analysis as all covalent and non-covalent interactions are assigned the same weight. Nonetheless, a strong, negative correlation can be found between centrality bin and evolutionary rate. The optimum value of the H_{cut} for the unRGNs (degree, betweenness, closeness) can be found in Table 3.1.

Residue Composition of Centrality Bins

The residues belonging to the unRGN closeness centrality bins were analysed further to study the nature of the non-covalent interactions in the data set, as well as examine the aspects of protein structure that are evident from the RGN. It is interesting to note that

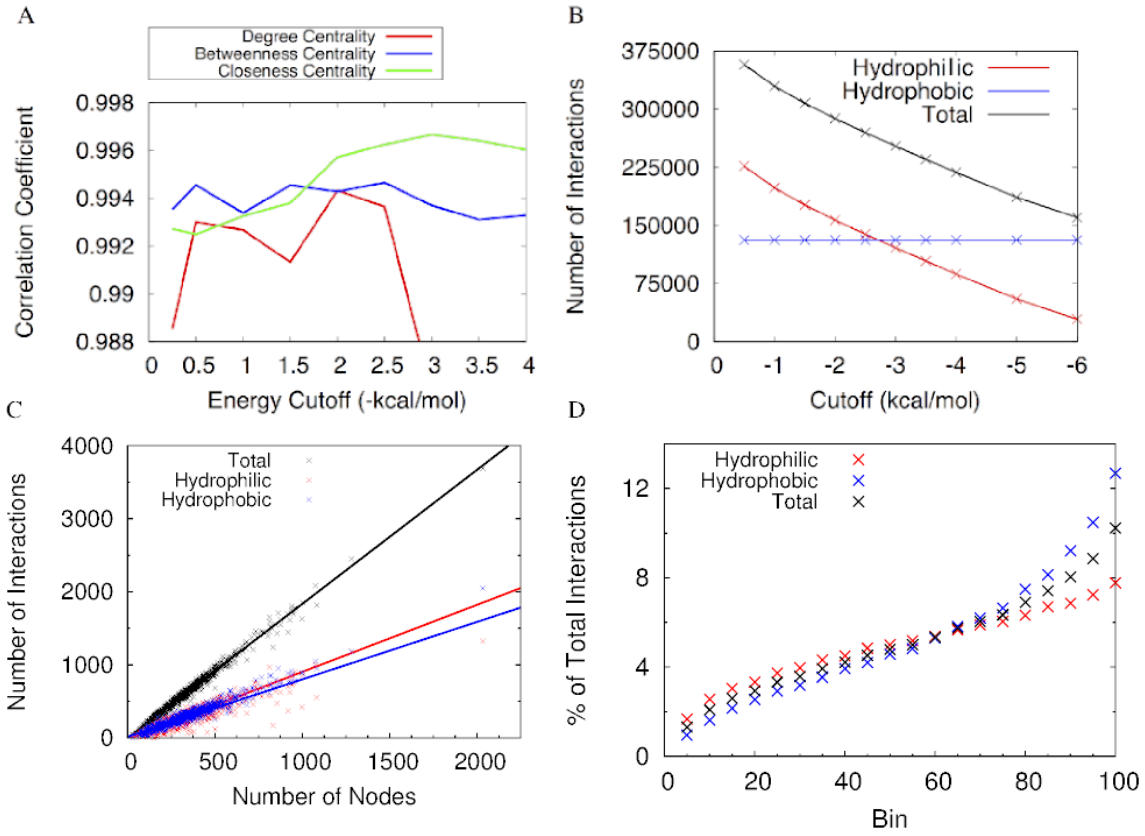


Fig. 3.7 (A) The effect of varying the H_{cut} parameter on correlation coefficient (as measured in Figure 3.6) for the unRGNs. (B) The number of hydrophilic, hydrophobic, and total interactions as a function of cutoff. Although the results of the analysis using a H_{cut} of -6.0 and -8.0 kcal/mol were less correlated with evolutionary rate than higher H_{cut} values, a correlation coefficient of > 0.99 was still observed. The robustness of the analysis to H_{cut} stems from the treatment of hydrophobic interactions. Namely, as the hydrophobic interaction energies are not explicitly calculated, they are not removed when lowering the value of H_{cut} and the total number remains constant. These interactions can therefore still identify high centrality residues at extremely low values of H_{cut} where very few hydrophilic interactions are involved, particularly in the centre of the protein where hydrophobic interactions are mostly concentrated. (C) For each protein the number of nodes has been plotted against the number of hydrophilic, hydrophobic, and total non-covalent interactions. (D) The percentage of the total number of hydrophilic, hydrophobic, and total interactions made by all residues in the data set are displayed for each bin. The closeness centrality bins for the unRGN were employed for B-D at a H_{cut} of -3.0 kcal/mol.

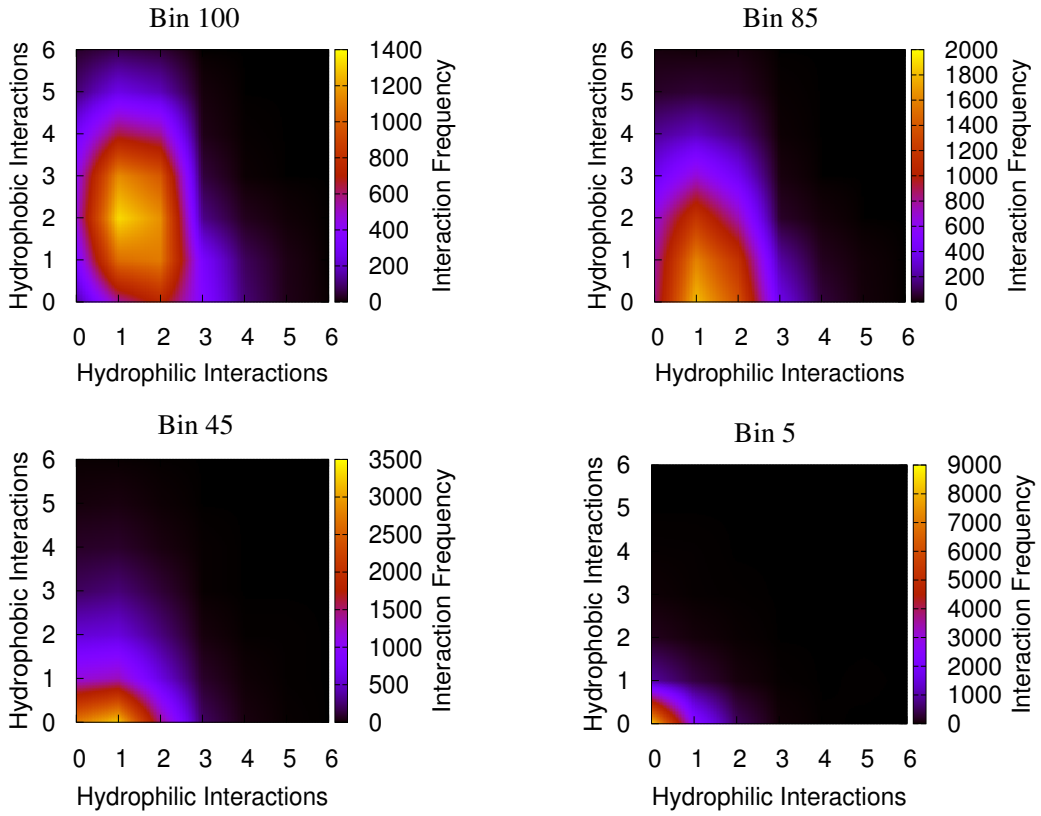


Fig. 3.8 The number of hydrophilic and hydrophobic interactions made by residues in each bin (grouped according to unRGN closeness centrality) were investigated, with the results of bins 5, 45, 85, 100 portrayed here to represent the trend. While the number of hydrophilic interactions formed by the residues does not vary greatly until very low centralities, the number of hydrophobic interactions can be seen to steadily decrease as centrality decreases.

the algorithm performs well when the number of hydrophilic interactions is similar to the number of hydrophobic ones (Figure 3.7B). This is expected to be a general property of the data set as the number of hydrophilic and hydrophobic interactions strongly correlates with the size of the protein (Figure 3.7C). Thus, balancing the influence of hydrophobic and hydrophilic interactions is achieved with a mid-range cutoff. The sigmoidal shape of the percentage of total interactions in each bin suggests that high centrality residues do exhibit a disproportionate number of interactions (Figure 3.7D). Similar trends were also observed for degree and betweenness centrality bins. The negative correlation between degree centrality and evolutionary rate, which suggests that residues forming few interactions have relaxed selection pressures, gives rise to this observation. Indeed, the importance of ‘hub’ residues, which has been noted in previous studies [246], is evident from the negative correlation between the degree centrality bins and evolutionary rate. evolutionary rate. This is due to

the fact that higher centrality residues are often embedded deep in the protein (as opposed to the periphery), which is also evidenced by the DC analysis. This is in line with the slower evolutionary rate of core residues relative to surface residues [76]. Furthermore, the residues displaying several hydrophobic and hydrophilic interactions are found to have the lowest evolutionary rate (Figure 3.8). This is likely to result from the smaller mutation space available for such residues that can form multiple hydrophilic and hydrophobic interactions.

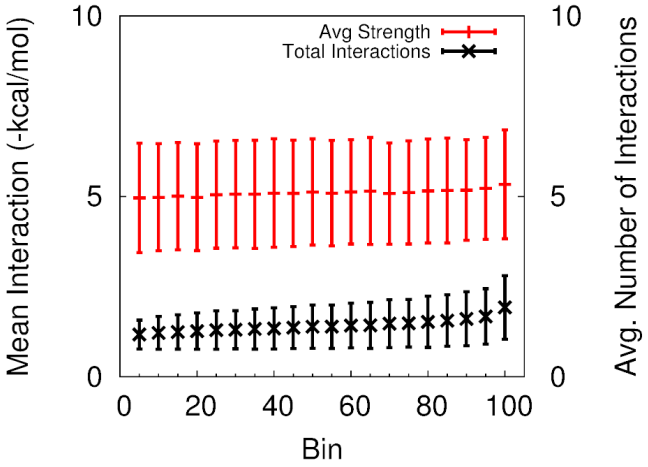


Fig. 3.9 The percentage total number of hydrophilic, hydrophobic, and the total interactions made by all residues in the data set are displayed for each bin. The weighted betweenness centrality bins were calculated using a H_{cut} = -2.5.

Analysis of the strength of hydrophilic reactions in each bin (Figure 3.9) revealed a steady increase in the average strength of the interactions despite not using weights in the analysis. However, residues with the highest closeness centrality bin have the highest standard deviation both for the average strength of an interaction, and for the number of interactions formed by each residue. This suggests not only that these residues generally employ stronger non-covalent interactions, but also that residues that have low degree centrality and/or form weak interactions are also found in the highest centrality bin.

The trends observed for the frequency of residue types in each bin (Figure 3.10) agree with the distribution of residues in protein structures. For the majority of residues, the frequency is found to rise or fall with decreasing centrality. For hydrophobic residues, including Val, Leu, Ise, Phe the frequency falls as centrality lowers while for Pro and Gly the frequency rises. For hydrophilic residues, it falls for Tyr, Cys, Met, Trp and rises for Gln, Asn, Ser, Thr. The frequency rises for all charged residues, namely Arg, Lys, Asp, Glu. In the case where the frequency falls with decreasing centrality it suggests that the residue, e.g. Leu, is more likely to be found in a high centrality region with a low evolutionary rate.

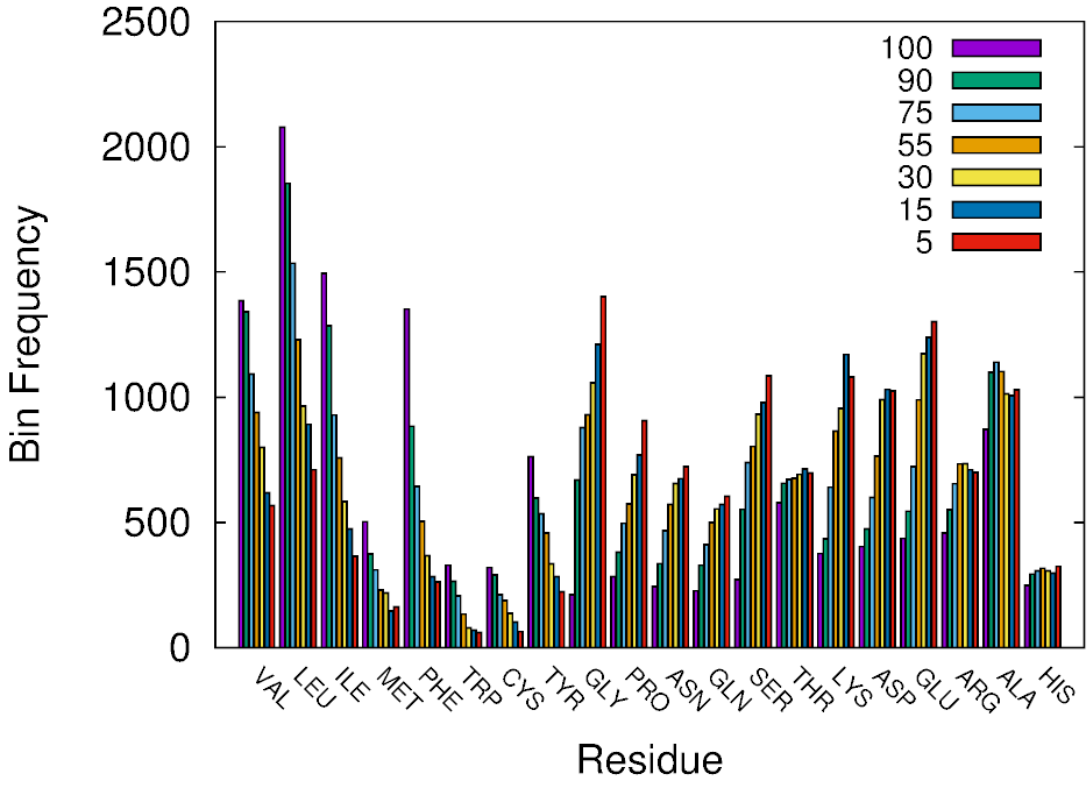


Fig. 3.10 The frequency of residues per bin has been displayed in the above histogram. Clear trends can be seen for the majority of residues, either rising or falling, across the range of bins. In general, the frequency falls for residues that are larger and hydrophobic, and rises for residues that are smaller and polar. For example, the average molecular weight found for residues where the frequency falls is about 150, while for residues where it rises it is roughly 130.

Indeed, for hydrophobic residues, the observed trend reflects their central position in the protein. Similarly, charged residues are more likely to be found in low centrality regions with high evolutionary rate, which is in line with these residues being found near the surface of the protein. This illustrates the ability of the RGN to identify trends in protein architectures.

The discussed trends could be useful for identifying idiosyncratic residues in the RGN, namely those whose behaviour does not conform to the expected trends. For example, Gly is generally found in lower centrality bins, and therefore has a higher evolutionary rate. This suggests that, in the cases where the residue is found to have a high centrality, it plays an important role. This concept will be explored later using residues that have both a low local connectivity (low degree) yet maintain a strong global connectivity (high closeness). Taken together, the RGN is able to identify key trends in protein structure.

Constructing the wRGN using weighted betweenness centrality

In addition to the unweighted networks, the wRGN was investigated for betweenness centrality. To do so, weights for the hydrophobic and covalent edges, which cannot be calculated within FIRST, were identified by optimising the correlation between weighted betweenness centrality and the evolutionary data. This required varying the hydrophobic interaction weight and the covalent bond weight parameters in addition to the H_{cut}. In particular, all hydrophobic interaction weights and covalent bond weights were assigned the same value in each calculation, and were investigated in the range of 0 to -4 kcal/mol and 0 to -40 kcal/mol, respectively. After varying these parameters, the highest correlation with evolutionary rate (r = -0.997) is found with the following variables; H_{cut} = -2.5 kcal/mol, hydrophobic interactions = -2.5 kcal/mol, and covalent interactions = -1 kcal/mol, which correspond to edge weights of 2.5, 2.5, and 1, respectively. This is an improvement on the highest correlation between unRGN betweenness centrality and evolutionary rate (from -0.995) (Table 3.1). Note that while the H_{cut} and the hydrophobic interaction strengths lie within the expected range, the covalent interaction energy is much lower than that found in nature. RGN edges are therefore not a straightforward representation of the enthalpic interactions that occur between the residues.

When constructing the network a question naturally emerges; what do the edges between residues represent? It is important to note that what is being modelled is not a static structure, but one that undergoes complex motions that have strong implications for function. While the backbone of covalent interactions helps determine the topology of the system, it is the non-covalent interactions that determine the unique ensemble of structures by restricting the conformational space accessible to the chain. Thus, the better performance of the diminished covalent bond weight (=1), compared to the order of magnitude higher strengths observed in nature, in the wRGN is predicted reflect of the low dynamic role such interactions have in comparison to non-covalent interactions with regards to a) protein structure, where covalent interactions are relatively constant compared to the continual breaking and reforming of non-covalent interactions and b) protein evolution, where mutations do not often change the covalent backbone.

The makeup of the centrality bins in the RGN illustrates the additional information that can be gained with knowledge of the specific interactions found in the protein. As discussed, standard DC-construction techniques do not account for residue types and their specific chemical interactions. Further attributes of the RGN will now be discussed, as evidenced by the evolutionary analysis and the wRGN network, that lie hidden when considering only DC AANs.

3.3.2 Hinge Residues

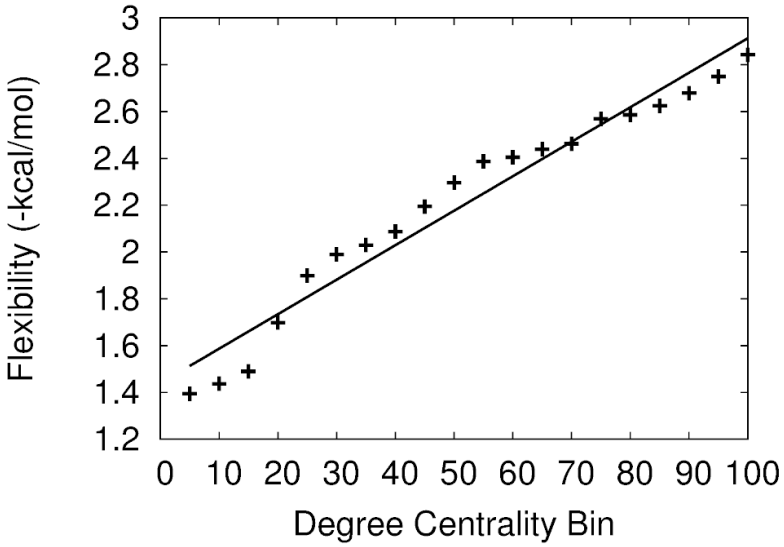


Fig. 3.11 Correlation between average bin rigidity index and degree centrality bin. The results show a strong, negative correlation ($r = -0.98$) between average rigidity index and degree centrality bin. Therefore, residues with a lower degree are more flexible.

Residues with both a high closeness centrality and low degree centrality are envisaged to behave as "hinges" in the network. Recall, CNA can be employed to investigate the flexibility of these residues. The low number of interactions made by RGN hinge residues suggests that, despite being well connected *globally* (as determined by high closeness centrality), these nodes support a greater amount of flexibility *locally* (low degree centrality). Indeed, a strong negative correlation between the degree centrality bin and rigidity index is found (Figure 3.11).

To further explore the network hinges residues were binned according to closeness and degree centrality, resulting in 400 degree-closeness centrality bins (Figure 3.12A). The correlation coefficient for the top (high closeness with increasing degree) and bottom (low closeness with increasing degree) row of Figure 3.12A has been measured to show how evolutionary rate changes with decreasing degree (Figure 3.12B). It can be seen that the hinge residues are more highly conserved than residues that display both high closeness centrality and high degree centrality. This trend opposes what would be expected for decreasing degree alone, given the negative correlation between degree centrality bins and evolutionary rate. Indeed, the analysis of this trend using low closeness centrality bins paints a different picture. Several residues have been used to exemplify this observation.

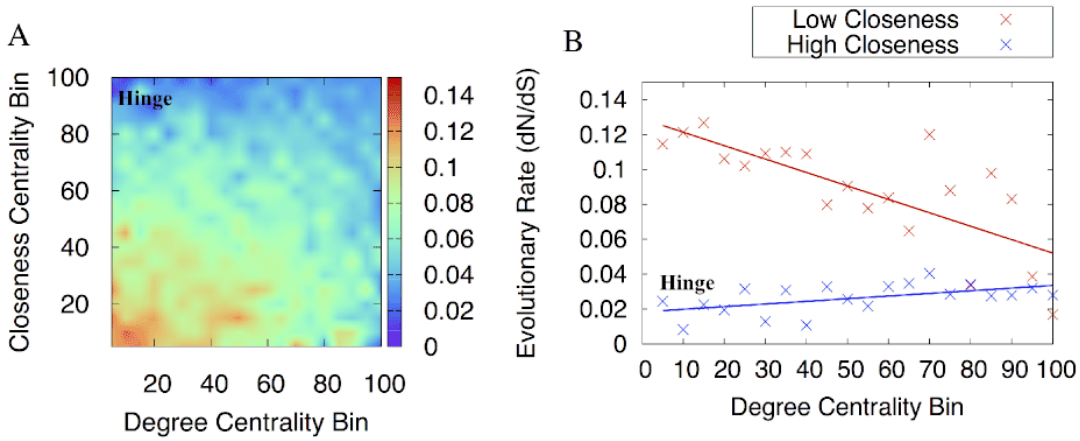


Fig. 3.12 (A) Heat map displaying the evolutionary rate for degree - closeness centrality bins. (B) the high closeness centrality and low closeness centrality rows have been displayed with trend lines. The trend lines show clearly that as degree decreases, the evolutionary rate of residues with high closeness centrality decreases (r=0.5, P value of 0.02) and the inverse trend is observed for residues with low closeness centrality (r=−0.7, P value of 0.0006). For the latter, when residues are randomly assigned to the ranked bins a correlation coefficient of -0.04 is found.

A well-conserved kinked α-helix is present in all geranyl-geranyl diphosphate synthase (GGPPS) protein structures [12]. This kink is found to occur just after residue G157 (Fig. 3.13) in a region of the protein that participates in ligand binding. G157 is found in the RGN analysis to form a hinge in the network: of 285 residues, it is ranked 7th highest for closeness centrality (0.182) and 270th for degree centrality (0.007). If the network is inspected more closely, two hub residues are found, namely F156 and L158, either side. F156 not only forms part of the ligand binding site [117], but also coordinates several residues of the ligand binding site. The latter is also true for L158. G157 also lies within 3 degrees of two residues that bind the phosphate moiety of the ligand. The low degree centrality in the RGN suggests that this residue has high flexibility. Indeed, a low CNA rigidity index (−1.3 kcal/mol) was identified for G157. Due to the close proximity to the active site and the binding pocket, it can be speculated that the removal of less influential hydrogen bonds residue is involved in dynamics associated with ligand binding. Indeed, glycine flexibility has been proposed as a mechanism to support induced-fit structural movements during ligand binding [163, 250, 223, 158]. This residue is not highlighted as a hinge residue using the DC technique (Figure 3.14), indicating the importance of incorporating chemical interactions into the AAN construction.

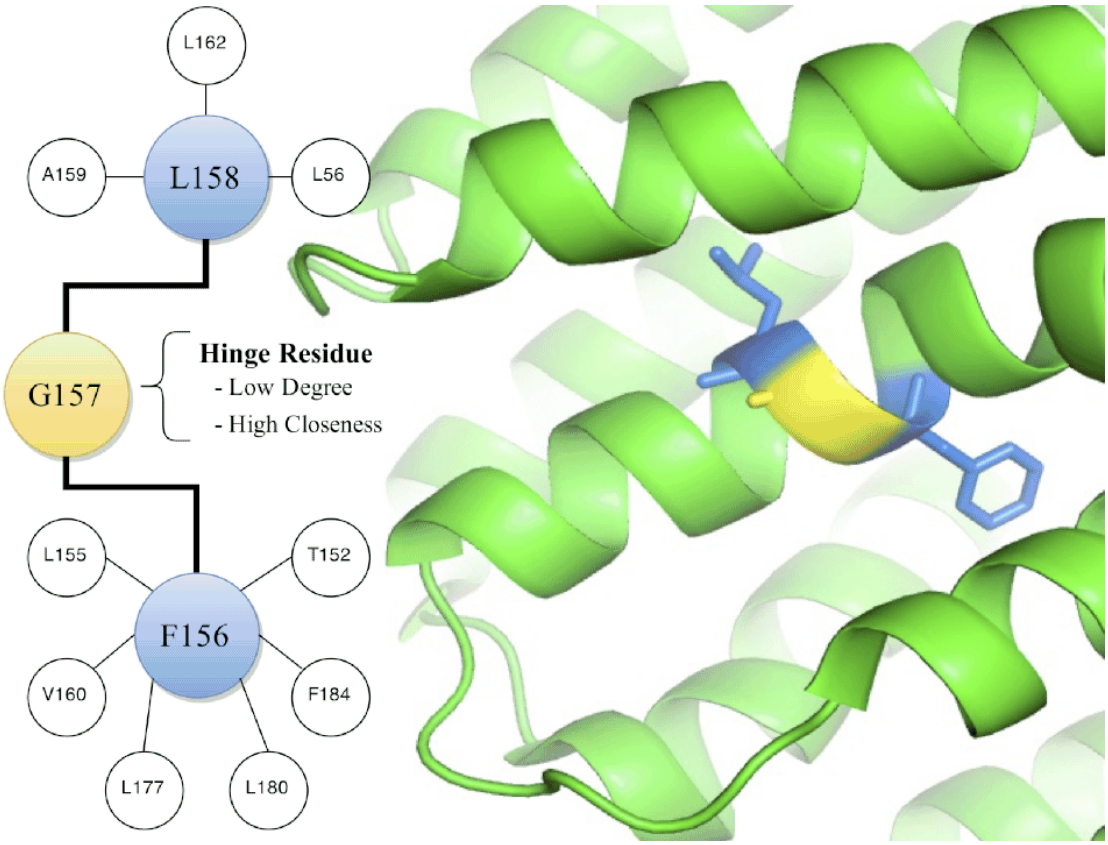


Fig. 3.13 A portion of the RGN of GGPPS is shown, displaying interactions made by coloured nodes. It can be seen that G157 interacts with only two residues (low degree). However, these residues form extensive interactions with the environment, and therefore give rise to the high closeness of G157. Critically, the blue nodes are known to play important roles in ligand binding.

H178 binds the phosphate moiety of G6P in the protein G6PD. Like G157, this residue also forms a hinge in the network: of 485 residues H178 is ranked 9th for closeness centrality (0.14) and 464th for degree centrality (0.004). By making few interactions, it is able to leave chemical groups free to bind the phosphate group via hydrophilic interactions. The low degree also results in its high identified flexibility (-1.1 kcal/mol) using CNA. Indeed, via site-directed mutagenesis H178 is found to contribute 1.4 kcal/mol net to the binding of G6P and is found conserved in all 27 G6PDs sequenced up to 1998 [49].

For the majority of residues, the strength of an interaction does not appear to strongly influence the centrality (Figure 3.9). In addition, the high closeness centrality bin also displays a greater spread, suggesting that residues that make fewer interactions can still have a high closeness centrality. The above hinge residues exemplify the importance of

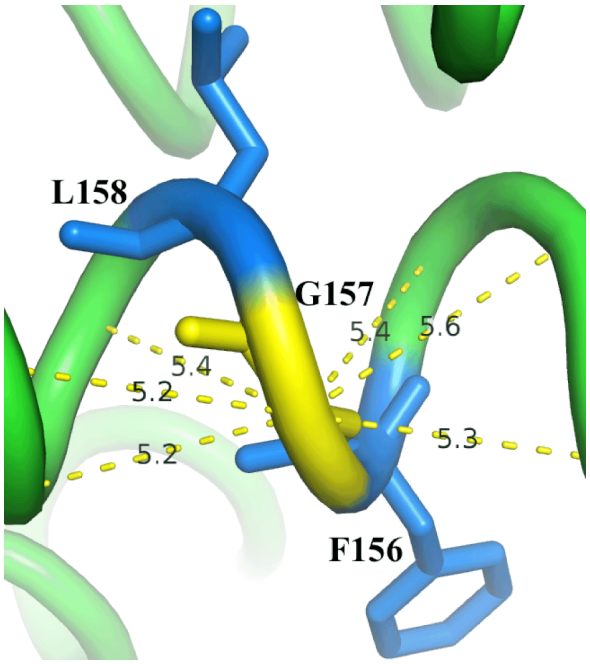


Fig. 3.14 The DC-construction technique is unable to identify G157 as a residue with dynamical function as connections are formed between this residue and all residues within 6 Å. The distance labels in the above diagram show that the DC-construction technique results in an additional 6 edges to those observed in the RGN.

considering more complex measures of centrality and how the combination of different centrality measures can help determine the role of a residue.

Hinge residues in proteins are found to behave as centres for global motion, displaying dynamic stability and strong conservation [145, 160]. The above results show that for proteins in the data set residues with lower degree generally have higher flexibility (Figure 3.11) and that among these residues those with high closeness are more strongly conserved than hub residues. The overlap between residues that display strong evolutionary dynamics and large structural dynamics suggests that conservation exists in the protein sequence to maintain motion [150]. The RGN hinge residues share several characteristics with protein hinge residues, which suggests that they can be used to identify centres for hinge motion and investigate evolutionary dynamics.

3.3.3 Expected Visited Time

The expected visiting time (EVT) measures the importance of each residue in the transfer of information through the network [170]. Signals are initiated at a particular residue and undergo random walks with the likelihood of propagation between two nodes determined

by the weight of the edge that connects them, before being absorbed at a second site. The EVT value for a residue is then calculated as the average visiting frequency of signals that pass through the corresponding node in the network for all absorbing sites. EVT analysis is therefore capable of identifying communication between distant sites and multiple pathways exploration, and has been previously used to study allosteric communication in proteins [170].

A Markov transition matrix, **T**, was derived from the wRGN and used to determine the signal transition probability to the nodes interacting with the signal node. The transition probability from node i to j (T_{ij}) is given a weight equal to the absolute value of the interaction between i and j (α_{ij}) divided by the degree:

$$T_{ij} = \frac{\alpha_{ij}}{d_i} \tag{3.5}$$

For hydrophilic interactions, the weight is equal to the interaction energy computed using FIRST. For hydrophobic and covalent interactions, weights that resulted in the strongest correlation of weighted betweenness centrality with evolutionary rate were employed (2.5 and 1.0 for hydrophobic and covalent weights, respectively).

The information flow through the wRGN is modelled using the absorbing Markov chain model [118], where the n x n "fundamental matrix" of the corresponding absorbing Markov chain is calculated according to:

$$\mathbf{F}^k = (\mathbf{I} - \mathbf{T}^k)^{-1} \sum_{l=0\rightarrow\infty} (\mathbf{T}^k)^l \tag{3.6}$$

In the above, $\mathbf{T}^k$ is the reduced transition matrix after the kth row and column were removed. The EVT for all nodes is calculated by averaging $\mathbf{F}^k$ over all absorbing nodes k:

$$\mathbf{M} = \frac{1}{n} \sum_k \mathbf{F}^k \tag{3.7}$$

EVT values have a bias towards residues that are nearby, whereas allostery occurs between distant sites in the protein. Here, the EVT values are scaled by multiplying the raw EVT value for each node by the shortest-path distance between the signal initiating site and the residue of interest [170]. The EVT values have been normalised to have a mean of 0 and standard deviation of 1.

Allosteric Analysis of GPCR signalling

The family of GPCRs includes proteins that are involved in signal transmission across the lipid membrane. GPCRs contain a transmembrane core of seven α-helices (H1-H7) that facilitate signal transduction. One such GPCR is rhodopsin, which is found in the rod cells of the retina where incident photons trigger a cascade of events that ultimately result in an electrical signal being passed from the eye to the brain. In order to absorb light, the protein binds 11-*cis*-retinal at the interface between H5, H6, and H7 via a covalent bond to Lys296 on H7. Isomerisation to *all-trans*-retinal occurs when an incident photon is absorbed by retinal. The *cis*-to-*trans* conversion undergone by this small compound is amplified by nearby residues, causing structural changes at distant, allosteric residues that stabilise the active conformation of the protein [119].

EVT is a random walk based measure of node importance during signal propagation. This holistic measure shares important features with allostery, including multiple pathway exploration and communication with distant sites. EVT analysis of rhodopsin has been previously carried out using constraints identified from molecular dynamics simulations [170], where correlations in atomic motion were used to build the edges of the network. By initiating a signal at the retinal molecule, a series of allosteric regulatory sites were identified using this dynamically-constructed network. In order to investigate whether the same insight into allosteric communication in rhodopsin may be obtained using the computationally cheaper AAN developed here, the wRGN has been constructed for rhodopsin and the EVT analysis has been performed on the resulting network. Finally, the results of the DC approach, which which is also computationally inexpensive but may be too simplistic to capture the chemistry of the interaction network in rhodopsin, are compared.

In general, the scaled-EVT values of the dynamical and RGN networks displayed sharp inhomogeneity, which peak at residues that regulate allosteric responses to photon absorption. Often, such residues were not highlighted in the DC network (Table 3.2). For example, spin label studies [7] have previously revealed structural changes at Phe313 in response to photon absorption. Arg135 is a key allosteric residue, as it forms part of the most highly conserved motif in GPCRs, known as the DRY motif [194]. This residue forms the strongest hydrophilic interaction ($E = -9.93$ kcal/mol) that is observed in the network with neighbouring residue Glu134, which also exhibits a high scaled-EVT value. This pair of residues can be found in topologically identical locations in most GPCR receptors and is predicted to stimulate the release of GDP as a result of photo-activation [2]. The ionic lock, which describes the interaction between residues Arg135 and Glu247, stabilises the inactive conformation and is broken in response to photon absorption, allowing a shift to the active conformation [6]. Arg135 is found to display the highest scaled-EVT value in the 326 node wRGN, but is not

Table 3.2 Scaled-EVT values for the RGN, Dynamical AAN [170], and DC AAN. Standardised values have been displayed to allow accurate comparison. Residues with high-scaled EVT values regulate allosteric change in rhodopsin and are often not identified using the DC network. The dynamical column has been left blank in cases where the residues have not been discussed in their study.

Residue	**RGN**	**Dynamical**	**DC**
W126	2.94	2.23	1.24
Y178	1.32	2.71	0.41
F103	0.58	2.25	-0.17
D83	-0.09	1.35	0.95
N55	-0.47	1.32	0.64
Y306	0.27	1.39	-0.82
F313	1.62	1.32	-0.82
V139	1.32	0.44	0.40
R135	3.79	1.07	0.55
E122	2.82		1.91
E134	1.35		0.50

highlighted in the DC network (Fig. 3.16). Due to the low resolution of rhodopsin (3.4 Å), an abnormally short distance between the interacting group of R135 and E247 is found in the X-ray crystal structure. The energy functional used in FIRST thus penalizes the interaction, assigning a positive energy to a bond that should relax into a low-energy bond [236]. Under the the H_{cut} used to construct the RGN, this interaction is not included in the network. If a modest [236] salt bridge of -5.0 kcal/mol is added between the network nodes and run the EVT analysis, an average difference in the observed value of 0.13 is found due to an increase in the EVT of R134 from 3.79 to 4.71. The introduction of this interaction results in a high scaled EVT value for E247 (2.78) that is not identified in the DC approach (-0.89).

G-proteins initiate downstream signalling events using GTP and GDP as a molecular switches for the active and inactive conformations, respectively. Activation is initiated on binding the so-called guanine nucleotide exchange factors (GEFs), which acts to stimulate the replacement of GDP by GTP. The process of GDP release relies on allosteric communication, as the GEF does not interact with GDP directly. A common Ga numbering (CGN) system (http://www.mrc-lmb.cam.ac.uk/CGN) has been identified for the Gα protein family and was herein used to compare the EVT of homologous structures. To study the signalling pathways, the average scaled EVT values of the 10 inactive Gα structures has been taken and the EVT analysis run with signals initiated at the sites in H5 that undergo contact rewiring in the GEF-bound state. These positions G.H5.[12,15,16,19,20,25] form between 3 and 5 contacts during the activating GEF interaction. These positions are used as the initiation site and

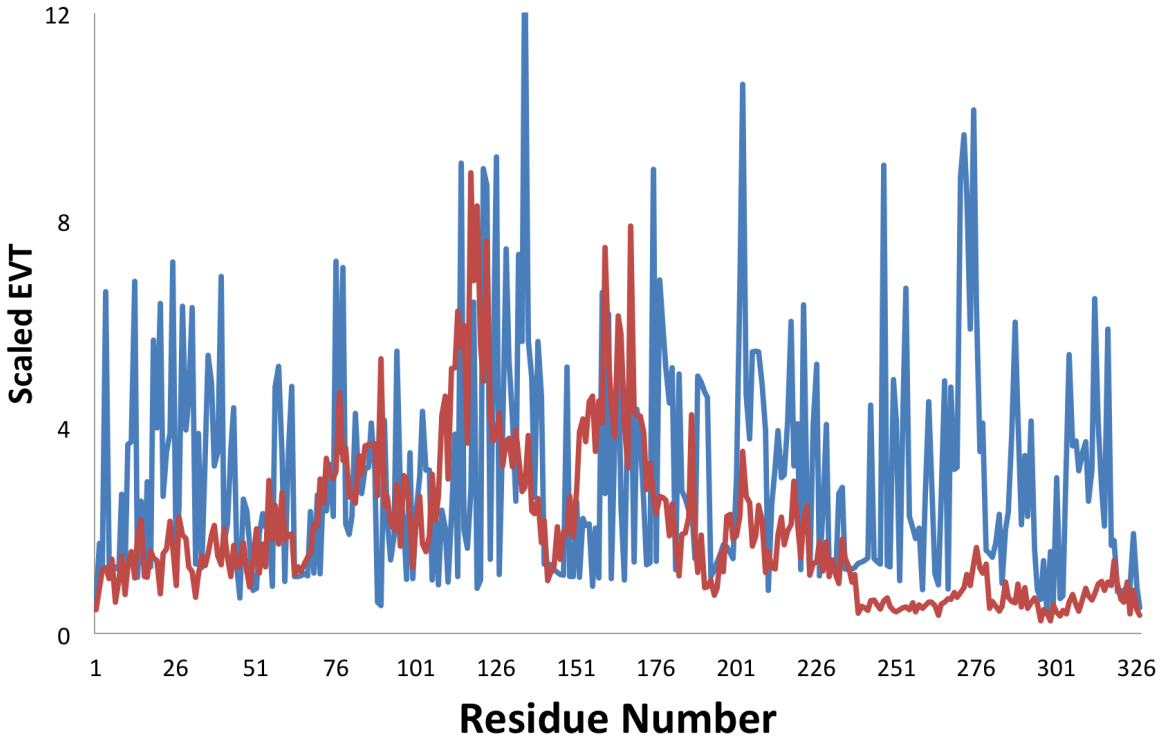


Fig. 3.15 Scaled-EVT values for RGN (blue) and DC (red) network (before the standard score is applied) allow the sharp inhomogeneities in the EVT values to be seen. These inhomogeneities are found to correlate with residues that play an allosteric role in Rhodopsin.

gone on to measure the scaled EVT at the 390 CGN positions. An inactive Ga EVT signal is calculated by taking the average scaled EVT across the structures. The CGN positions displaying the highest average EVT for the 6 initiation sites are found to overlap strongly with the conserved allosteric "wire" identified previously [70] (Figure 3.17).

The Gα:GEF interface directly impacts the C-terminal end of α- helix 5 (H5) and ultimately gives rise to helical domain opening that allows the dissociation of GDP. Note that average scaled EVT values refer to (1) average EVT values measured at CGN positions for the 10 inactive Ga structures and (2) the average EVT value relative to the 6 interface positions. The highest scaled EVT is found for positions G.H5.4 and G.H5.8 (Figure 3.18) . G.H5.4 (3.46) displays the highest average scaled EVT value and has been highlighted in previous experiments and contact networks as being important for Ga-GDP stability. Indeed, this site forms the most contacts in the inactive state, the majority of which are interrupted when binding to GEF. Mutation also gives rise to the greatest instability of the Ga-GDP state of any residues in H5. Ala and Cys mutations to G.H5.8 (3.35) accelerate GDP exchange, as it forms conserved contacts with H1. Indeed, several universally conserved contacts maintain the link between H5 and H1 in the absence of the GEF, including G.S2.6 (2.51) with G.S3.3

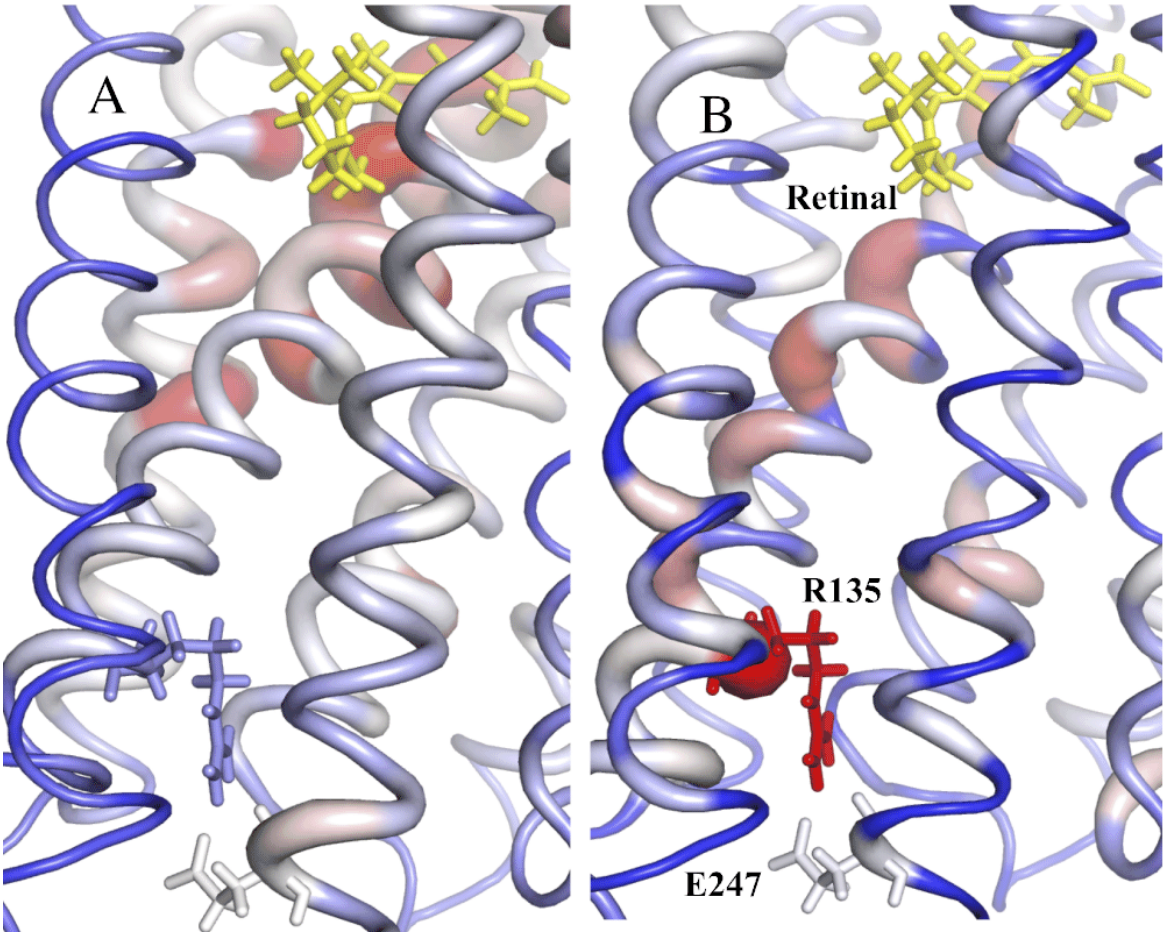


Fig. 3.16 The ionic lock, which stabilises the inactive conformation and is broken in response to photon absorption, was found to display below average scaled-EVT values in DC network (A) and significantly high scaled-EVT values in the wRGN (B). Residues with high scaled-EVT values are coloured red and have greater thickness.

(1.84) and G.H1.8 (2.56) with G.H5.8. In the above residues, receptor binding causes a reorganization of the contacts, ultimately disassociating H1 from H5 that facilitates GDP release. In addition to H1.8, a high scaled EVT value is found for G.H1.7 (2.41), whose mutation causes greatest decrease in stability in Ga-GPCR complex out of all mutations to H1, as well as a modest decrease in Ga-GDP stability [70].

3.3.4 Toward Protein Design

Predicting ΔΔG of Point Mutations

Scikit-learn [174] is a Python library that integrates machine learning algorithms. This software package has been employed to build a linear model [204] that predicts the ther-

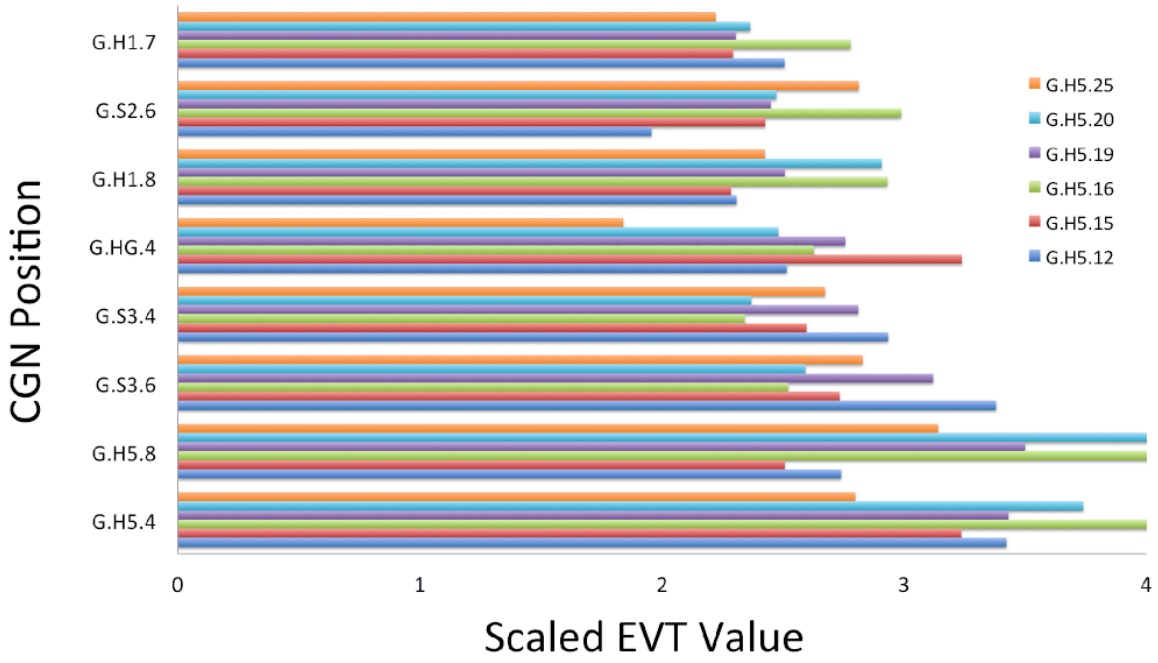


Fig. 3.17 The average scaled EVT at each CGN position has been measured for 10 inactive PDB structures (3AH8, 1ZCB, 1AS3, 3UMS, 1TAG, 3UMR,3FFB, 1GG2, 1GP2, 1GOT). Then, by considering signal absorption at positions that interact with the GEF, namely G.H.[12,15,16,19,20,25], we have taken the average scaled EVT measured with respect to these residues at the CGN positions to investigate where the signalling from these residues is most sensitive. Above are the residues with the highest average, with respect to the GEF interacting residues, scaled EVT values, within which several residues overlap with a conserved allosteric wire identified in Gα proteins.

modynamic properties of single nucleotide polymorphisms (SNPs). In particular, using as input amino acid type, position, and RGN centrality, a model is trained that predicts the experimentally observed $\Delta\Delta G$. The coefficients included: (1) For each residue, the WT and mutant hydropathy, volume, helix score, sheet score, turn score, acid dissociation constant, one hot encoding, (2) the direction (xyz) and magnitude of 23 closest residues and (3) the number of residues in the protein.

The Gibbs free energy (ΔG) is a thermodynamic measurement of protein stability [13]. The impact of a single nucleotide polymorphism (SNP) or point mutation to the native structure can therefore be measured using the change in the Gibbs free energy ($\Delta\Delta G$). Models that predict the $\Delta\Delta G$ of a SNP can be used to guide the decision process when scanning protein engineering design libraries [110]. Data from the "ProTherm" database (http://gibk26.bio.kyutech.ac.jp/jouhou/Protherm/protherm.html) has been acquired to test the usefulness of the RGN in aiding these models. Using residue characteristics, RGN

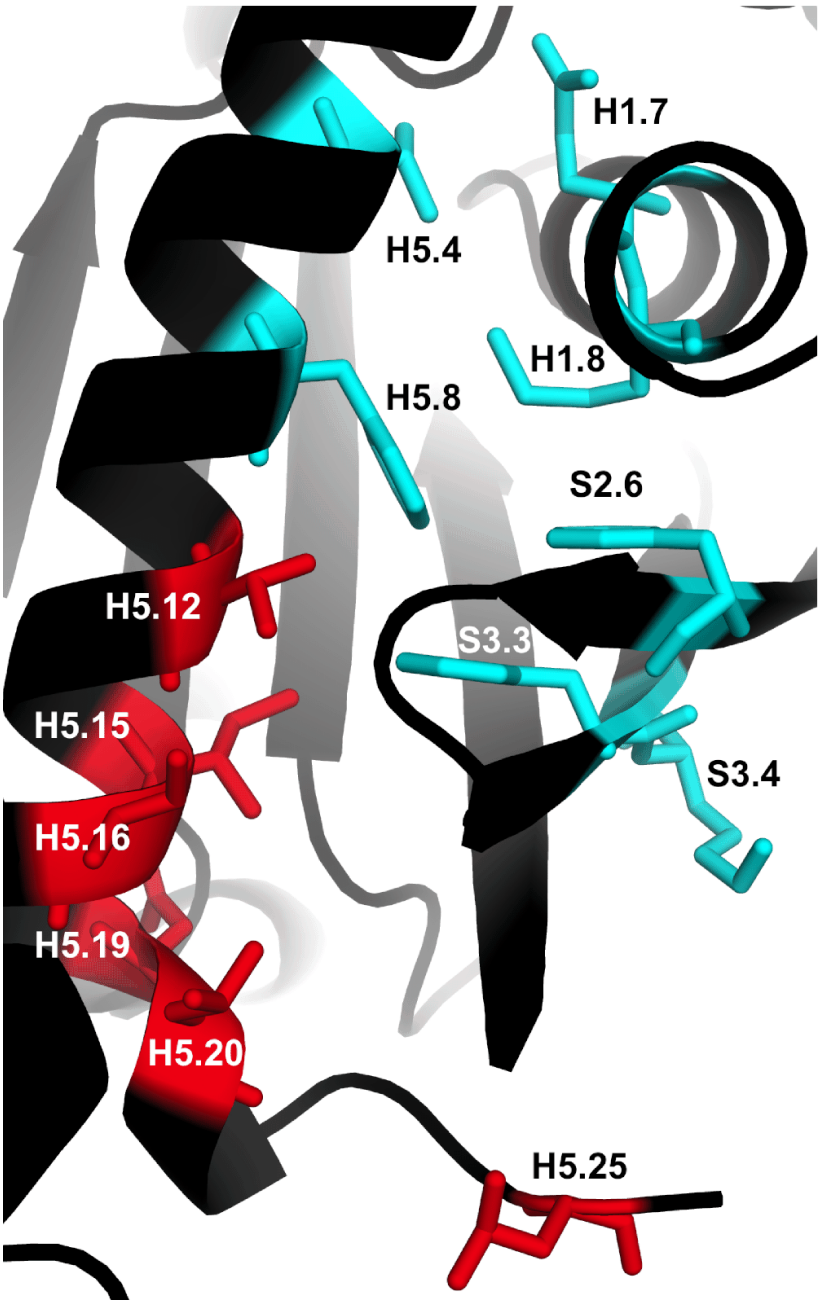


Fig. 3.18 CGN positions on the inactive Gα structure (pdb:1AS3). Residue positions highlighted using the average EVT analysis (cyan) form part of an allosteric wire involved in GDP release. Residues found in red form contacts with the GEF.

measured centrality, and Cα distances, a model with a coefficient of determination of 41% was achieved. Although this is not as strong as those identified by Jia et al. [110] (around 60%), the current methodology does not require energetic calculations using the Rosetta software [129].

Table 3.3 RGN differences between homologous mesophile and thermophile proteins. The average network measure for the 10 proteins in the mesophile and the thermophile datasets are displayed in column 2 and 3, respectively. Column 4 shows the p value of the paired t-test for the 10 mesophile-thermophile homologues. The properties were calculated for residues in the β-domain [233].

Measure	Meso	Thermo	p value
Interactions	561	591	0.003
Average Degree	1.8	1.9	0.03
Giant Cluster	0.74	0.84	0.08
Cliques	54	63	0.03
Avg. Clust. Coef.	0.08	0.09	0.05

Quality Control : Thermophillic, Mesophillic, and Amyloidogenic proteins

Thermophile proteins are more stable than their homologous mesophile forms. Brinda and Vishveshwara [32] have studied the thermophile and mesophile networks of 10 proteins to identify characteristics that confer stability. To do so, they used data from molecular dynamics (MD) simulations to place edges between residues that interacted strongly. They found that the number of interactions, the average number of residue interactions, and the giant cluster (size of the giant RC with all constraints in place) were higher in the thermophile, stable homologue. Similar conclusions can be found regarding the network characteristics of protein stability when using the RGN to compare these two groups of proteins. Note, the giant cluster is calculated as the number of nodes connected by non-covalent interactions at room temperature kT (H_{cut} = -0.593 kcal/mol [179]).

The more stable, thermophile form of the proteins shows distinct changes in their network properties (see Table 3.3). These results suggest that the RGN is able to identify hallmarks of protein stability. The average clustering coefficient has additionally been calculated for the two protein data sets, which are found to be lower for the mesophile proteins (0.08) than the thermophile proteins (0.09). This result was found to be significant at $p \leq 0.05$ using the paired t-test.

A Possible Application in Protein Design

De novo protein design would allow purpose-built function to be designed to tackle many of the current challenges in biomedicine and nanotechnology [106]. The Rosetta [134] suite is a widely used biophysical models for backbone flexibility. This model consists of a target structure, allowed backbone moves (e.g. dihedral changes), a predefined sequence space, a rotamer library, and an energy function. This model describes a characteristic space

that can be searched using Rosetta's iterative relaxation and design algorithm, consisting of a (1) "design step" whereby the amino acid identities and side chains are optimised and (2) "relaxation step" where the backbone and side chains are optimised using a hybrid stochastic/gradient descent optimization [48]. A scenario can be imagined whereby the RGN analysis could be incorporated into the design step. Here, EVT analysis could be used to design allosteric communication pathways or the identified hallmarks of protein stability could select conformations that are likely to have increased thermostability. Although speculative, the ease of RGN construction and analysis means that it's incorporation into more complex models would be trivial.

Summary

Geometric simulation identifies constraints that collectively describe the covalent and non-covalent interaction network of a protein. This network can be used in constrained dynamic simulations to generate the conformal ensemble of a static protein structure [235]. The RGN is a static network construction technique that uses only these fixed constraints to build the amino acid network. The centrality of residues measured using the unweighted RGN are found to correlate strongly with evolutionary rate, showing that residues that have a greater (network) importance evolve more slowly. This follows the observation that residues displaying high centralities are more likely to be found in the centre of the protein and in enzyme active sites. Residues that form hinges in the RGN were also found to play dynamical roles in protein function, hinting that dynamical functions can be investigated using the RGN. Residue characteristics, such as size and hydrophobicity, are clearly evident from their evolutionary behaviour, suggesting that the network model is able to recapitulate chemical aspects of the protein structures.

Investigating allostery using network methods has previously been accomplished using a dynamical network constructed using information derived from MD simulations. To test the ability of the RGN to investigate this purely dynamical function, weights must be assigned to the hydrophobic and covalent edges in the RGN. These values are estimated by optimising the correlation with evolutionary rate as they cannot be measured energetically. Random walk simulations of the weighted RGN are able to identify several allosteric residues as having high expected visiting times in the GPCR signalling pathway. This suggests that the computationally cheap, static AAN construction technique presented here can successfully identify allosteric residues in proteins, which had previously only been accomplished using dynamical techniques or with the aid of sequence alignments.

Chapter 4

Constrained Geometric Simulation of the Fenna-Matthews-Olson Complex

Outline

Chapter 3 exemplifies the powerful information that is encoded in the network of interactions identified using geometric simulation. This technique is a well-established approach for simulating the conformal ensemble associated with a single protein structure. The Lagrangian-based approach uses fixed constraints to store information related to bond lengths and angles, which can be used to efficiently explore protein dynamics. Indeed, the strong sampling efficiency allows the full conformal ensemble to be explored even for large proteins. This makes geometric simulation highly-applicable to the study of the 21 000 atom Fenna-Matthews-Olson complex (FMO), whose dynamic spectral density partly depends on slow conformal modes that extend beyond the time scales that can be investigated using Newtonian-based techniques.

The key role of the FMO protein structure in tuning the optical properties of the embedded pigment molecules, which act to transfer excitation energy through the complex, will be developed. The conformal ensemble gathered using constrained geometric simulation is shown to contain spatial correlations and a hierarchy of dynamical relationships that appear to play a role in stabilising the "working parts" of the energy transfer wire. The all-atom geometric simulation permits further analysis of the molecular details of these observations, allowing a number of general organisational principles that act to stabilise excitonic energy transfer in the presence of the slow protein motion to be identified, such as the low variation in excitonic couplings. The possible relevance of the correlated conformal motion toward the observed quantum coherence in 2DES experiments is highlighted throughout.

4.1 FRODA

FRODA (Framework Rigidity Optimised Dynamic Algorithm) [235] takes as input the decomposition of the protein into rigid and flexible clusters from FIRST to efficiently simulate all-atom motion. To generate a new conformation, the atom positions are perturbed in random directions by a magnitude of 0.1 Å. This is followed by an iterative cycle that allows the constraints to be efficiently and accurately re-introduced by fitting the atomic positions back on to so-called "ghost templates", which store the information relating to the constraints identified by FIRST (Figure 4.1). It should be noted that, unlike molecular dynamics (MD), FRODA is not a dynamical technique and hence does not supply any characteristic time scales for the formation of individual structures. However, the constraints based approach allows large systems and microsecond functions [237, 238] to be investigated, making its application to studying conformal disorder in the (20 000 atom) FMO complex highly relevant. In contrast to force field based molecular mechanics (MM) simulations, the constraints-based Lagrangian dynamics that is implemented in FRODA effectively flattens the potential energy surface and substantially reduces the number of degrees of freedom to be explored. It has been shown that these simplifications actually allow FRODA to outperform MD in the sampling of transient pockets at protein-protein interfaces [153]. Such efficiency does come at the cost of being limited to a fixed constraints topology, generating an athermal ensemble and neglecting long-ranged electrostatic interactions. Yet even within these approximations, FRODA-generated atomic fluctuations have been shown to agree with MD simulations and, critically, NMR experiments [80]. The constrained dynamics have been used to investigate the flexibility of the nicotinic acetylcholine receptor ion channels, which led to the identification of key residues that are predicted to facilitate rapid communication between the binding site and the transmembrane gate [22], in addition to cisplatin cross-linking in calmodulin [141], and myosin flexibility during the ATPase cycle [221]. The agreement of FRODA with experimental protein motions and its ability to provide insight into a wide range of functions [115, 130, 72] suggests that the dynamics of the native state emerge naturally from a simple network of contacts.

To fully explore the conformal landscape, 1 280 00 conformations are generated and every 250th structure is stored to gain a representation of the ensemble. Monomer one (as labelled in the PDB structure) of the FMO complex has been analysed in two simulations; a simulation in the absence of the other two monomers (the monomer simulation) and a simulation of the full trimeric structure (the trimer simulation). Although monomeric and trimeric simulations have been performed, the discussed analysis is with respect to the trimer simulation except where explicitly stated.

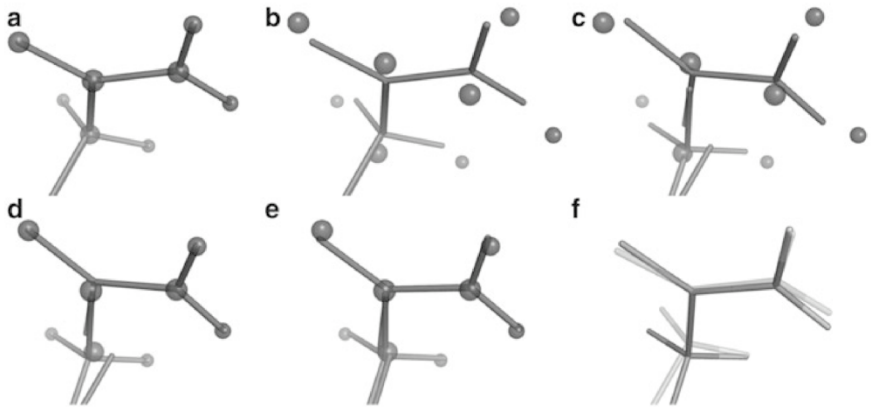


Fig. 4.1 Template-based geometric simulation. (a) A portion of the protein is shown with reduced atomic radii for clarity. Geometric (ghost) templates are defined by the bonding geometry in the rigid cluster and fitted over atom positions. In each example there are two (top and bottom) ghost templates. (b) Each step generates a new conformer and begins with the random displacement of atoms (effectively breaking the bonds). This is followed by an iterative cycle that realigns (c-e) the ghost templates with the atom positions using a least-squares fit to the new positions and the atom positions with the ghost templates. Note that in this latter step atoms that are shared between two ghost templates are placed equidistant from the associated templates. This iterative cycle is repeated, until the atomic positions and template vertices are within an acceptable tolerance. This figure has been taken from Ref. [235].

4.2 Constrained Geometric Simulation

FMO Structure Preparation

Geometric simulation calculations were performed on the holo form of the trimeric 1.3 Å X-ray crystal structure of *Prosthecochloris Aestuarii* (PDB accession code 3EOJ). Hydrogen atoms were added to the structure using the Molprobity software [56]. A manual investigation of the structure resulted in the transfer of a proton from the ε to the δ nitrogen of His 6 (and the equivalent residues in monomers 2 and 3). The hydrogen atom of Tyr 9 and Tyr 338 sidechains was also rotated to form hydrogen bonds with pigments 3 and 4, respectively. The AMBER11 software [35] was employed for structural relaxation. The description of the protein was accomplished using the FF99SB force field, the water molecules using the TIP3P model, and the pigments by the force field developed by Ceccarelli et al. [36]. Missing heavy atoms were added using the *leap* module of AMBER11. The structure was solvated and heated to 300 K over a period of 300 ps, followed by 1000 steps of conjugate gradients minimisation. Strong restraints (1000 kcal/mol/Å^2) were applied to the heavy atoms of the protein and pigments throughout the equilibration procedure. This relaxation process was conducted by Dr. Daniel Cole [44]. A number of water molecules were found to play important roles in the pigment-protein hydrogen bonding network. Therefore, at the end of the equilibration procedure, the closest 600 water molecules to the pigments were retained to model the effects of specific pigment-water and protein-water hydrogen bonds. Note that the central porphyrin ring of a pigment molecule forms a rigid cluster (RC) due to the structure of the covalent bonds. The central Mg atom of each pigment was rigidly connected to the protein *via* the ligands identified in the X-ray crystal structure.

4.2.1 Dynamics of the Fenna-Matthews-Olson Complex and the Impact on Excitonic Coupling

The FMO complex possesses a trimer quaternary structure in which each monomer assumes a clam-like architecture [168] that sequesters seven pigments [228]. The orientation of the FMO complex in the membrane positions an eighth pigment, which is located on the outside of each FMO monomer, near the baseplate, which is connected to the much larger light harvesting antenna complex known as the chlorosome [239].

Figure 4.2 shows the structure of the 20 000 atom FMO complex analysed using the RC decomposition at a H_{cut} of -4.6 kcal/mol. An unusual feature of the FMO complex is the persistence of a large RC at low values of H_{cut}. This RC encompasses the majority of

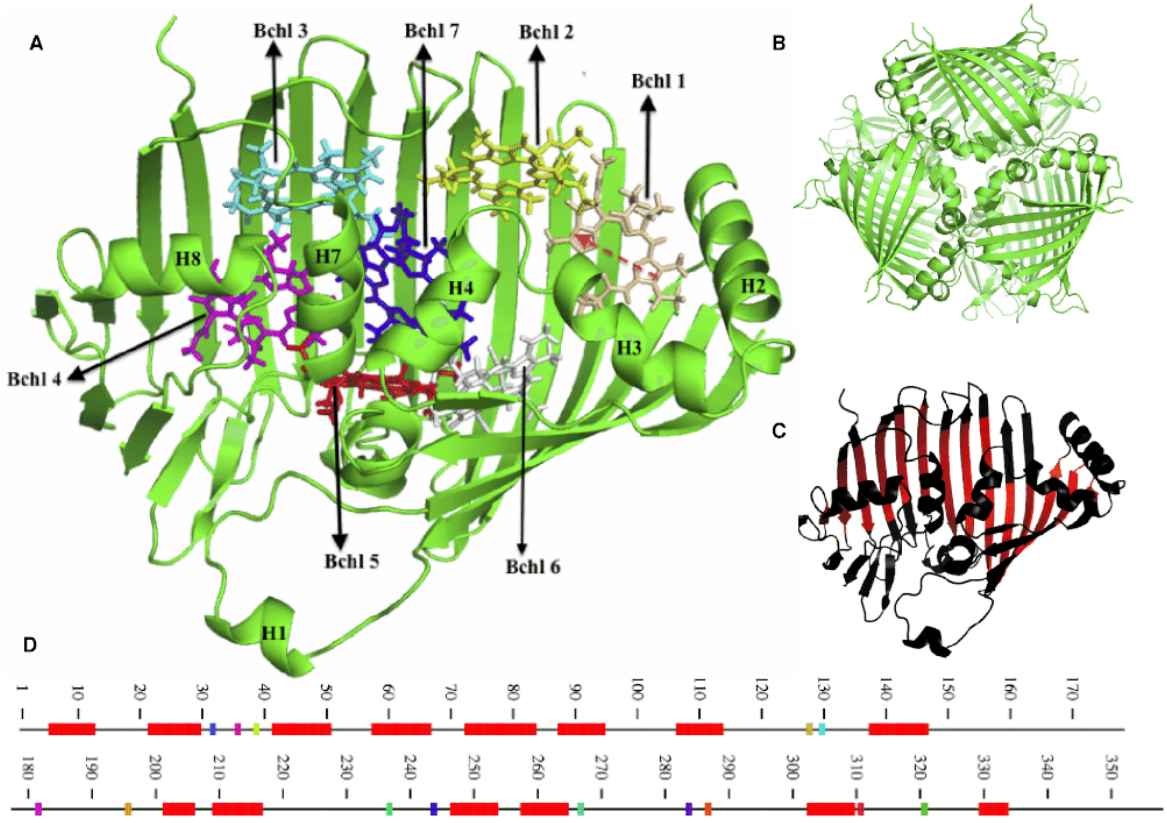


Fig. 4.2 (A) monomer one of the FMO complex with the pigment 1 to 7 and α-helices 1-5, 7, and 8 labeled. pigment 1 has been used to illustrate the Q_y axis, which lies along the line formed by connecting the nitrogen atoms of pyrrole I and pyrrole III (red arrow). (B) The orientation of each monomer in the full trimer structure. The RCs identified by FIRST using $H_{cut} = -4.6$ kcal/mol are shown in red and flexible regions in black. These regions have been visualised on the monomer structure (C) and along the primary sequence (D).

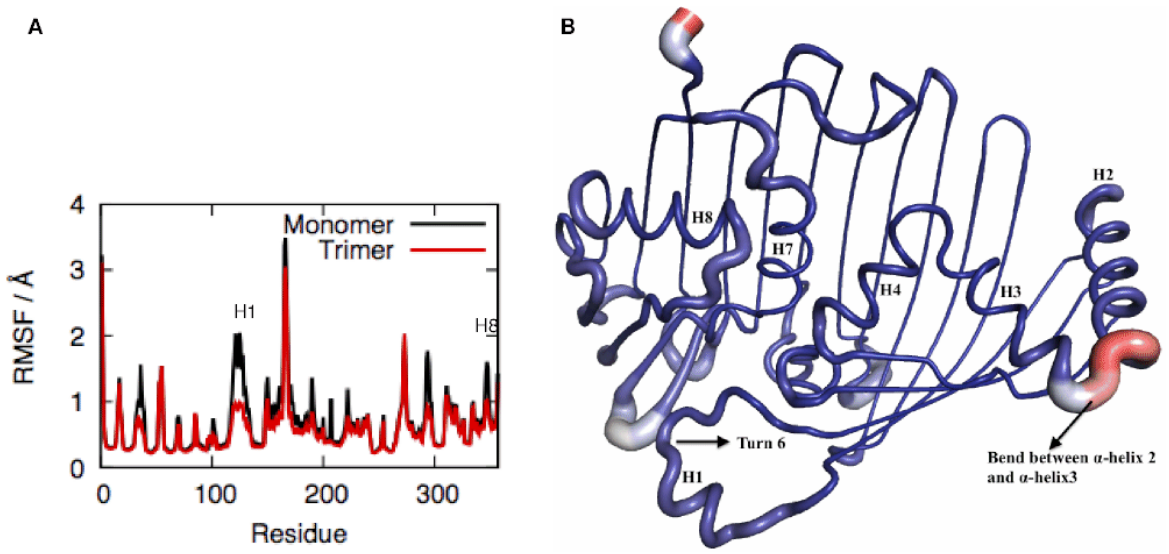


Fig. 4.3 (A) RMSF (Å) per residue in monomer one from the monomer and the full trimer FRODA simulations. (B) The RMSF values from the trimer simulation have been displayed by colour (increasing from blue to red) and tube thickness. The improved matrix facility of PTRAJ [186] has been employed to generate the RMSF of the C_α atoms of the stored conformations.

the clam shell (CS) architecture, which sequesters the optically-active pigments from the environment. The implications of the size and robustness of the CS RC will be discussed.

The internal structure of the FMO complex, which mostly comprises the pigments and the α-helices, is by comparison relatively flexible. Recall, it is widely accepted that the α-helices play a role in modulating the site energies of the pigments. The choice of $H_{cut} = -4.6$ kcal/mol, therefore, allows us to efficiently explore the large-amplitude motion of the pigments and their environment within the approximation of a rigid CS structure. The FIRST RC analysis gives valuable information regarding the expected flexibility in the FMO complex. FRODA takes, as an input, the FIRST rigidity analysis at $H_{cut} = -4.6$ kcal/mol, which includes 713 hydrogen bond and 236 hydrophobic non-bonded constraints.

The root mean squared fluctuations (RMSF) calculated for each residue, averaged over the ensemble of output configurations, has been displayed in Figure 4.3. In the trimer simulation, the analysis reveals few regions with high conformal mobility – most residue fluctuations are lower than 1 Å. One exception is the bend that connects α-helices 2 and 3 (residues close to 166). This is most likely due to the presence of three sequential glycine residues, which support the greatest backbone mobility. α-helix one forms direct contact with adjacent monomers in the full trimer structure and, as a result, has an RMSF in the monomer simulation that is more than twice that found in the trimer simulation. Less expectedly, in the

monomer simulation we find much larger motions of α-helix eight and turn 12 than in the trimer simulation. Recall that the α-helices, in particular α-helix eight, has been shown to strongly influence the optical site energies of the pigments [156]. It should be emphasised that the full trimeric structure should be used in dynamical studies of environmental fluctuations in the FMO complex. The average RMSF for each pigment may be related to the site energy disorder, and therefore the static inhomogeneous width and structure of the excitonic lineshapes in the linear absorption and linear and circular dichroism spectra [230, 5, 206, 44]. However, there is no significant difference in the RMSF of pigments 1–7, averaged over the atoms of the porphyrin rings, which are small and range from 0.36 (pigment 6) to 0.77 Å (pigment 5). The contrast between these weak fluctuations and the much larger differences in site energy distributions [206], which generates a particularly broad spectral function on site 7 (that has a low RMSF), suggests that the site energy dispersion depends more strongly on interactions with the protein environment.

The results clearly show that the pigments of the FMO complex, protected by a remarkably robust CS structure, display low conformal flexibility. This is particularly evident when comparison is made with the rigidity of structures that have previously been investigated with FRODA analysis. Wells et al. [234] measure rigidity by counting the number of C_α atoms that belong to the five strongest RCs at a $H_{cut} = -3.0$ kcal/mol in relation to the total number of C_α atoms (denoted f_5, Figure 7 in Ref. [234]). Of the 51 proteins analysed, only three proteins from a particular rigid protein family had a higher ratio than that revealed here for the FMO complex ($f_5 = 0.56$). It should be noted, however, that the FMO complex is constructed from more than twice the number of residues as those found in the rigid protein family.

Cross Correlation Analysis

By minimising the flexibility, and thus the thermal exploration of conformal states, PPCs can protect against the disorder of the Hamiltonian parameters that disrupt the basic structure of the EET pathways. The seemingly excessive rigidity of the FMO protein suggests that this super-structure may have been selected for purposes additional to simply holding the pigments in position. Indeed, this rigidity promotes and coordinates substantial spatial correlations in the more flexible secondary elements of the structure. Investigating these relationships requires measurement of their relative motions, which is accomplished using the cross-correlation coefficient.

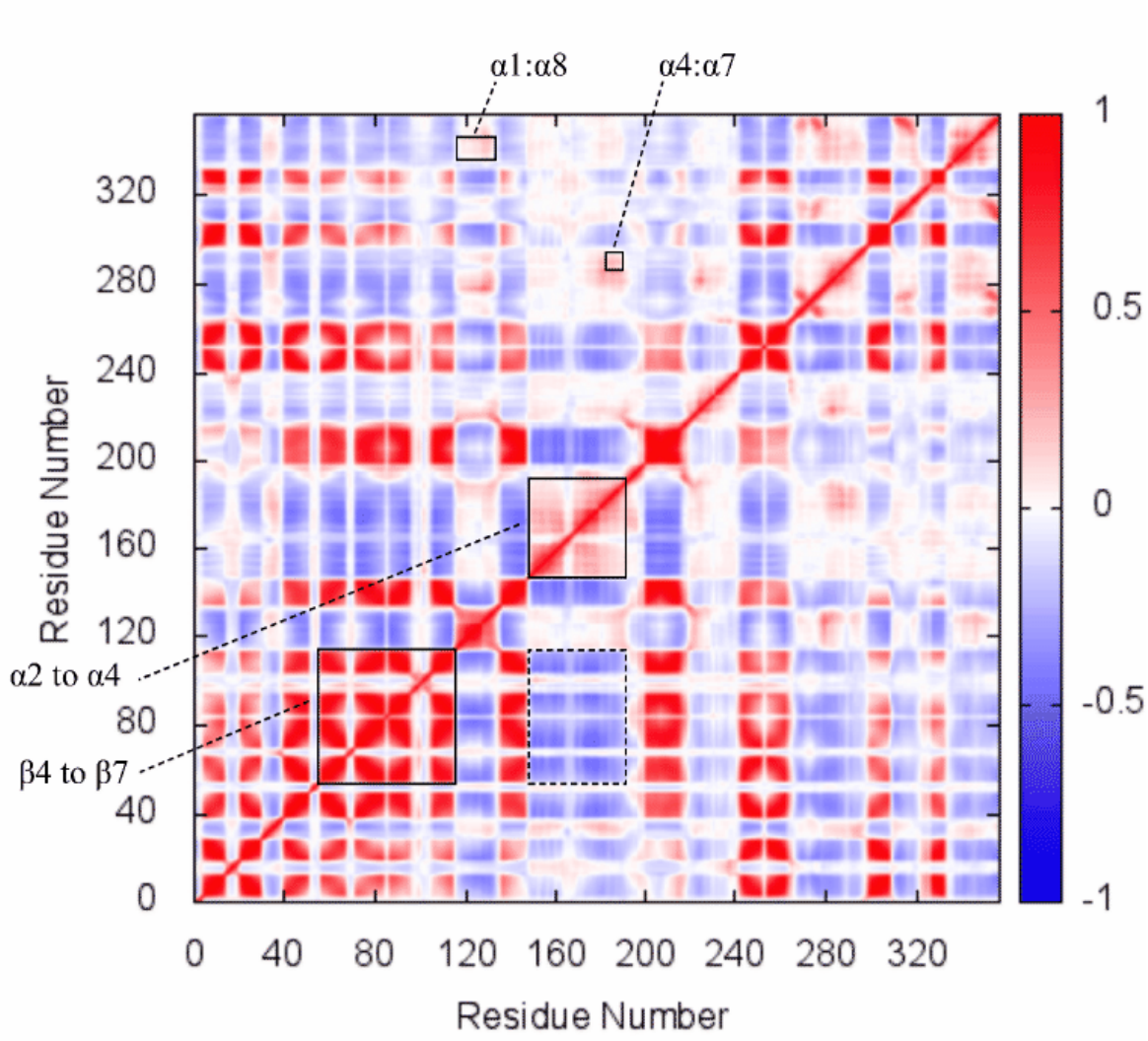


Fig. 4.4 The cross-correlation matrix for monomer one in the trimer simulation. Several noteworthy correlations between α-helices are highlighted. Also shown are β-sheets 4 through 7, which make up a large proportion of the CS structure – note also, that these structures are anti-correlated with the α-helices.

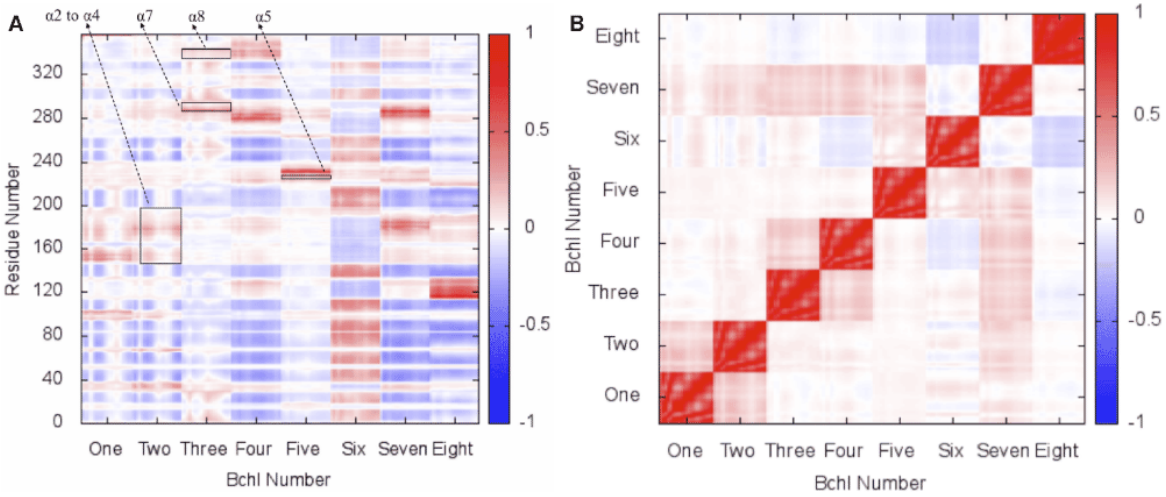


Fig. 4.5 (A) Correlations between the pigments and important secondary structure elements of the protein. (B) Correlations between the 8 pigment pigments of the FMO complex.

The cross-correlation coefficient between pairs of atoms is generated under the PTRAJ environment [186] as follows:

$$C_{ij} = \frac{\langle \triangle r_i \cdot \triangle r_j \rangle}{\sqrt{\langle \triangle r_i \cdot \triangle r_i \rangle \langle \triangle r_j \cdot \triangle r_j \rangle}} \tag{4.1}$$

Cross-correlation scores range from +1 (correlated) to -1 (anti-correlated), allowing long-range correlated motions to be identified in the ensemble.

The rigidity of the CS gives rise to extremely well-correlated motions within this sub-structure across the conformal ensemble (Figure 4.4). Strikingly, these residues are also, generally, anti-correlated with the functional α-helices in the protein. It was also found that α-helices 4 and 7, as well as α-helices 2–4, displayed correlated motion with each other. The correlation between α-helices 2, 3, and 4 is particularly noteworthy given that they span a distance of roughly 24 Å and implies that long-range communication is present in the protein environment.

The CS RC is predicted to be involved in the maintenance of globally correlated conformal movement. The cross-correlation analysis provides strong evidence for conformal correlations between pigments that are theoretically predicted to display strong excitonic coupling and form delocalised states [5, 4] namely; between pigment 1 and 2 and pigment 3 and 4. Pigment 7 also displays strong (near +1) correlations with pigments 1–5 (Figure 4.5B). Figure 4.5A shows the presence of correlations between the pigments and the protein environment. Firstly, with the exception of pigment 6, the pigments display anti-correlated motion with the CS RC. Moreover, important correlations have been identified between secondary

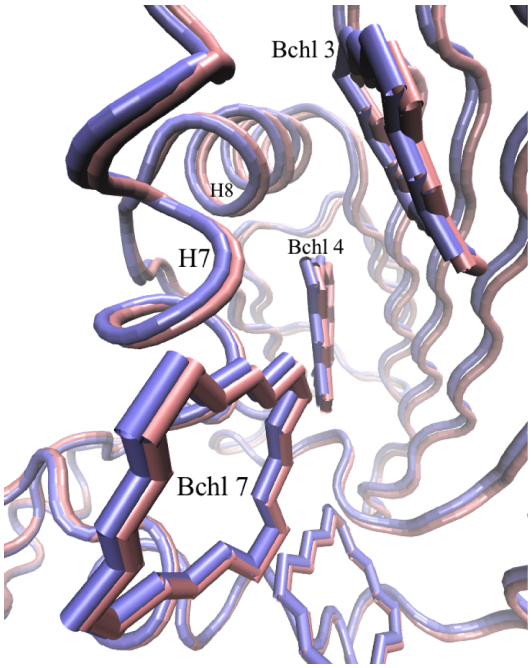


Fig. 4.6 The synchronised motion of α-helix 7 with pigments 3, 4, and 7, as is exemplified by PC1 (shown).

structure elements in the protein that, critically, modulate the pigment site energies. Pigments 1, 2, and 7 display correlated motion with α-helices 2 through 4 while pigments 3, 4, and 7 display positive correlations with α-helix 7.

Principal Component Analysis

Using the bio3d software package [88], principal component analysis (PCA) is employed to extract biologically relevant motion. Bio3d employs the R environment and has powerful statistical computing capabilities to generate the PCs. The mode that describes the largest variation represents large amplitude concerted motions and are, therefore, intimately related to protein function [75]. These motions are thought to occur on timescales much longer than those where energy transfer occurs, and are used in the current thesis to represent the static disorder identified in ensemble experiments. The specific advantages of PCA in identifying low-frequency motions will be discussed in greater detail in the Chapter 5.

The largest-variation principal component, namely PC1, provides additional evidence for the concerted movement of the functional α-helix seven with pigments 3, 4 and 7 (Figure 4.6). The PCA also identifies low frequency motions involving α-helices 1–4 and 8. It has been shown previously that the electrostatic dipoles of α-helices can significantly alter the local site energies of pigments [156, 4, 44]. In particular, α-helices 7 and 8 are thought to be

major contributors to the energy sink formed around pigments 3 and 4. The observations of correlated (Figure 4.5A) and concerted (Figure 4.6) motion between pigments 3 and 4, and α-helices 7 and 8, make it highly plausible that spatial site energy correlations also exist. As will be investigated in the next chapter, this general mechanism could preserve the relative site energies of the pigments in the configurational ensemble and appears to maintain pigment 3 as the ultimate sink for EET in each conformal realisation of the FMO complex. EET depends on the excitonic coupling between pigments, and the extent to which this coupling is impacted by the conformal motions can be investigated using purely structural techniques, as has been done in the past [5].

Variance in Excitonic Couplings

The excitonic coupling that occurs between pigments results from the Coulombic interaction between their transition densities [230]. When the static differences and environmentally induced fluctuations between the excitation energies of the individual pigments are comparable or smaller than these couplings, new excited states of the system appear that are coherently delocalised over the pigments [5], and which are spectrally shifted from the original (uncoupled, localised) transition energies. Under these conditions, transitions between states are driven by the (relatively) weak coupling to the environment (Redfield limit). In cases where the energy differences are very large, or the environmental interactions very strong, the relevant excitations are localised on the molecules and the (perturbative) dipole-dipole coupling is the source of incoherent energy transfer (Forster transfer) between them [230]. The strengths of these couplings are dependent on the mutual orientation and separation of the molecular transition dipoles. The degree of delocalisation of the excitonic states is a sensitive function of the strength of dipole-dipole couplings and the differences in site energies of the pigments involved, such that coupling strengths significantly greater than the energy differences leads to fully delocalised excitons. However, if these effects are important for function, the FMO structure must be robust against any major variations in the relative pattern of the pigments' positions and orientations.

In the FMO complex, the coupling interaction between pigments m and n (V_{mn}) has been shown to be well-reproduced by the point dipole approximation [182]. This approximation uses the strength and direction of the transition dipole moment to estimate the distribution of excitonic couplings that are present in the ensemble of FRODA-generated structures according to:

$$V_{mn} = f\frac{\mu_{vac}^2}{R_{mn}^3}[\vec{e}_m \cdot \vec{e}_n - 3(\vec{e}_m \cdot \vec{e}_{mn})(\vec{e}_n \cdot \vec{e}_{mn})] \tag{4.2}$$

Table 4.1 The inter-pigment excitonic couplings in trimer simulations of the FMO complex (cm^{-1}). The couplings identified by Adolphs and Renger [5] (column 2) are compared with the mean excitonic couplings identified in the current analysis (column 3). The standard deviation for these couplings over the ensemble of output structures is displayed in column 4, where the coupled pigments displaying large σ have been highlighted in bold. Dr. Daniel Cole compiled a script for this analysis.

Pigments	Adolphs and Renger [5]	This Work	σ
1-2	-98.2	-92.8	7.8
1-3	5.4	5.8	0.7
1-4	-5.9	-5.9	0.6
1-5	7.1	7.1	0.9
1-6	-15.2	-14.6	4.5
1-7	-13.5	-9.8	2.0
2-3	30.5	29.4	2.0
2-4	7.9	6.4	0.8
2-5	1.4	1.5	1.2
2-6	13.1	10.9	1.2
2-7	8.5	2.7	2.9
3-4	-55.7	-47.0	13.0
3-5	-1.8	-0.1	1.5
3-6	-9.5	-9.4	0.4
3-7	3.1	15.8	10.0
4-5	-65.7	-58.8	7.3
4-6	-18.2	-16.1	1.5
4-7	-58.2	-63.1	7.4
5-6	88.9	91.3	14.0
5-7	-3.4	-2.6	5.4
6-7	36.5	35.0	5.7

where $\vec{e}_{mn}$ is a unit vector joining the Mg ions of pigments m and n and $\vec{e}_m$ is a unit vector along the transition dipole moment of pigment m (defined by the line joining the N_B and N_D atoms of the pyrrole ring). $\mu_{vac} = 6.1$ D is the transition dipole strength of the Bchl pigment in vacuum and its value has been experimentally-determined [125]. The factor f describes the enhancement of the transition dipole moment in the protein environment, dielectric screening of the Coulombic interactions, and neglect of the dielectric cavities of the pigments. It has been shown, in comparisons with full solutions of the inhomogeneous Poisson equation, that $f = 0.8$ is a suitable scaling factor for the FMO complex [44].

Table 4.1 shows the excitonic couplings for the pigments in monomer one from the trimer simulations, and shows that they are extremely robust (with typical variances less than 5%) with respect to the allowed conformal variations supported by the FMO complex. These values are compared with a previous study that used an identical method but was based only on the static crystal structure [5]. The mean excitonic couplings of the ensemble of FRODA-generated structures, in general, agree with the calculations performed by Adolphs and Renger [5]. However, there are several noteworthy deviations between the previously calculated static couplings and the present data, which includes the effects of environmental fluctuations through the inclusion of structural flexibility. The largest deviation was found to be that between pigments 3 and 7. While the coupling calculated directly from the crystal structure was low (3.1 cm^{-1}), the current analysis has found that the averaged coupling over the conformal ensemble is higher, albeit with relatively large fluctuations (15.8 ± 10.0 cm^{-1}). A similar situation is identified for pigments 3 and 4, as well as pigments 5 and 6. The coupling between pigments 3 and 4 displays the second largest deviation from the static coupling and also second highest σ. This link is also a part of the major EET pathway proposed for excitations entering the FMO complex via site 6 [154, 177, 34]. The small σ identified between pigments 2 and 3 (2.7 cm^{-1}) shows that the excitonic coupling in one of the main EET pathways is well preserved despite thermal motion in the environment. The robust conservation of the coupling between these two pigments can be attributed to the conservation of their transition dipole orientation. This is predicted to be accomplished by a bend between β-strands two and three. This bend is (a) able to span the distance between the two pigments due to its long length and (b) has the highest concentration of prolines in the amino acid sequence. Proline residues can be employed to rigidify loops [19] due to their ability to reduce the conformal flexibility of the protein backbone. It may be that a conformally restricted bend allows the two pigments to remain connected while also allowing communication with other functional elements (α-helices, pigments, etc.).

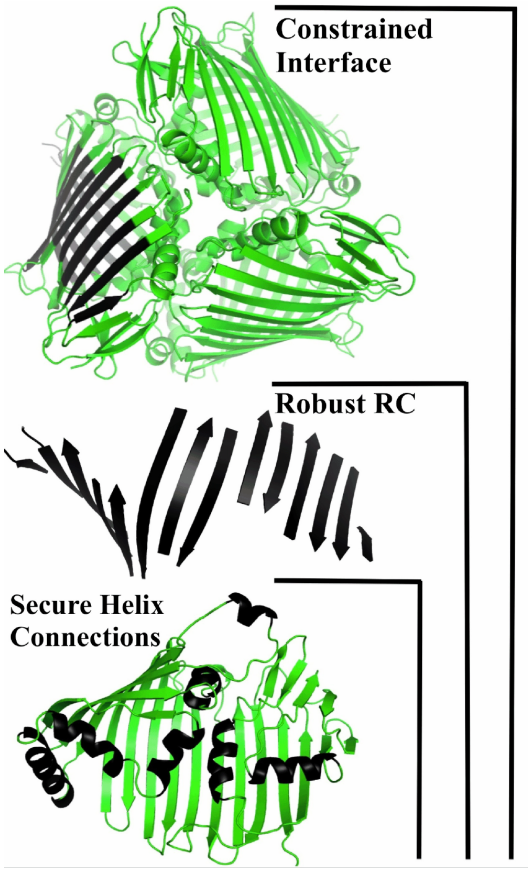


Fig. 4.7 Non-covalent interactions are employed to reduce the conformal fluctuations (uncertainty) in the motion of the FMO complex. As displayed above, these interactions can be grouped together into three main bands. Importantly, the observed correlations allow communication between the groups, an effect termed *trickle down structural organisation.*

4.3 Analysis of a Hierarchy of Protein Motions

The dynamical analysis highlights the use of three main levels of structural organisation to reduce mobility and constrain the conformations of the protein to ones that support coherent EET (Figure 4.7). In what follows, the mechanisms by which specific non-bonded interactions preserve correlated motion between functional elements of the PPC in each of these three layers of structural organisation, presumably to safeguard efficient EET, will be discussed.

Hydrophobic Interactions Define Inter-Pigment Correlations

EET is initiated from either pigment 1 or 6, after which excitons relax towards pigment 3 for entry to the reaction centre [33]. The nearly degenerate pigment 1 and 2 have been predicted

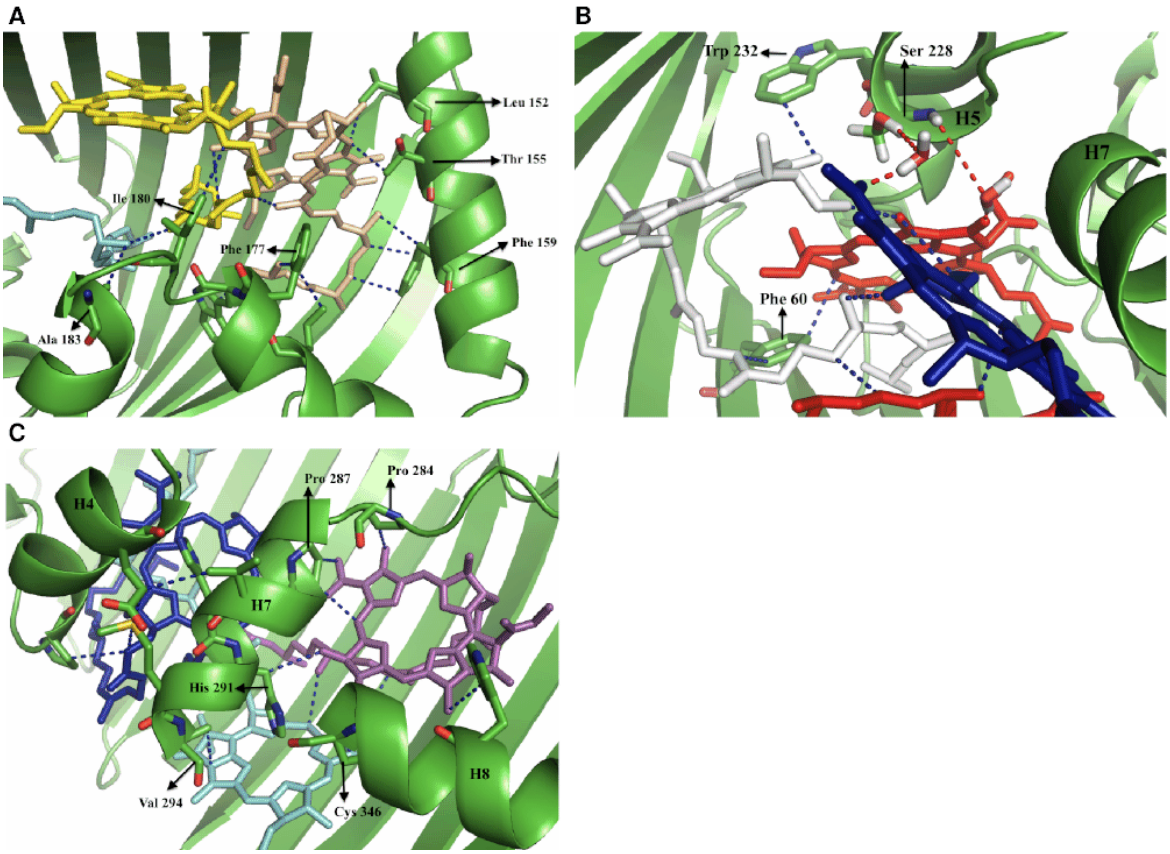


Fig. 4.8 (A) Extensive hydrophobic interactions between pigment 1 (wheat) and α–helix 2 can be seen. Several hydrophobic interactions between pigment 1 and pigment 2 (yellow) are shown. Hydrophobic interactions are also found with residues in (Phe 177) and adjacent to (Ile 180) α–helix 3. (B) A hydrophobic interaction was identified between pigment 6 (white) and turn nine residue Trp 232. The hydrophobic cluster that forms between pigment 5 (red), 6, and 7 (blue) has been shown. A hydrogen bonding cluster was also identified, which involved pigment 5, pigment 7, Ser 228, and two water molecules. (C) Hydrophobic interactions are found to occur between pigment 3 (cyan) and pigment 4 (magenta). pigment 3 and 4 are also found to interact with α–helix 7 (via Val 294 and His 291, respectively) and α–helix 8 (via Cys 346). Hydrophobic interactions were also identified between pigment 4 and residues near in sequence or in α–helix seven. Note, an interaction was also identified with with His 291, which coordinates pigment 3. A hydrophobic cluster involving pigment 7, α–helix 4, and α-helix 7 can also be seen.

to form a strongly delocalised pair of excitonic states, leading to a delocalisation-induced energy splitting and enhancement of EET in the pathway from sites 1 to 3 [40]. Indeed, it has also been proposed that the excitonic splitting might be tuned to exploit a specific frequency of the environmental noise to achieve this enhancement, as a result of the so-called 'phonon antenna' mechanism [185]. These quantum EET pathway dynamics require that the energy differences between sites 1 and 2 remain smaller than their excitonic coupling. The current analysis shows that these pigments display correlated motion, and that their excitonic coupling varies negligibly across different conformal realisations. Although pigment 1 and 2 interact directly with each other, their correlated movement is also maintained via interaction with nearby α-helices that display correlated motion with the pigments (namely, α-helices 2 through 4, see Figure 4.8A). Pigment 5 and 6 display correlated motion, and display extremely high excitonic coupling [33]. This is accomplished through non-covalent interactions between the two pigments (Figure 4.8B). Trp 232 is also well positioned to facilitate an indirect contact between pigment 6 and turn 9 (which itself coordinates pigment 5). The identified correlation between pigment 5 and 7 relies on interactions with overlapping biological elements. Pigment 5 displays correlated motion with α-helix 5, which precedes turn 9 and, in turn, interacts with α-helix 7 and thereby provides an indirect connection between pigment 5 and 7. These correlations are partially mediated by hydrogen bonds with water molecules.

It has been found that all the pigments (except pigment 6, which will be discussed in the next section) are correlated with pigment 7. From pigments 5 and 6 the EET relaxes to pigments 4 and 7. These two pigments are also found experimentally to participate in the same EET pathways [33]. Indeed, pigment 3 and 4 interact directly with pigment 7. However, correlated motion is presumably enhanced further by their mutual interactions with α-helices 7 and 8 (Figure 4.8C). Critically, these α-helices have previously been shown to affect the site energies of these pigments [4]. The sensitivity of pigment 7 to the motion of the other pigments suggests that it may play a unique role in the EET cascade. The importance of pigment 7 has already been noted by Skochdopole and Mazziotti [209] in connection with their research of quantum redundancy in the FMO complex. The quantum redundancy provides protection against the temporary or permanent loss of one or more chromophores by offering alternative pathways to the reaction centre. Therefore, although pigment 7 is only present in one pathway, it may be a vital component for conserving important correlations in both EET pathways. This idea is further supported by the striking demonstration of Olaya-Castro and Fassioli [164]. Namely, optimised energy transfer through the FMO complex, as a function of temperature, is achieved by *maximal* excitonic correlations between site 7 and the rest of the pigments of the FMO complex. The correlations suggest that the

pigment conformal motions remain in unison through *slow* correlated fluctuations managed by secondary structure elements.

The Robust CS RC

By using a low H_{cut}, rigidity properties are investigated that (a) are stable (due to the low cutoff) and (b) give rise to geometric structures in the protein that are robust against time. Connecting regions of a protein with particular roles assigned by Nature can reveal subtle design principles that are employed to improve a particular function. The uniqueness of the CS RC is a significant observation, and implies that it plays a critical role. In fact, such a large RC encompassing several secondary structure elements at low values of H_{cut} has not been seen in other systems [234]. Indeed, the CS has an important structural role in the protein by sequestering the pigments and aiding the organisation of the LHC by facilitating contacts with the baseplate, and perhaps promoting trimer aggregation in the periplasmic space between baseplates and reaction centres, which is another possible level of organisation that is not dealt with here. For example, the observation that 150-200 trimers contact the baseplate may demand tight packing that is facilitated by the CS RC [26]. However, the CS RC is also efficient at transmitting structural information over distances of more than 50 Å.

In a previous work, the α-helices were found to have an unanticipated ability to modulate the site energies of pigments 3 and 4 [156]. Natural selection facilitates the accumulation of residues that improve a protein's function (in this case, energy transfer). Therefore, it can be assumed that Nature has employed α-helices to manage the optical properties of the pigments and create an energetic funnel toward pigment 3. The remarkable anti-correlation between the α-helices and the CS suggest the existence of a higher level of communication, which further encourages correlation between the spatially separated functional units. Since the CS is well-correlated over large distances, it is an extremely well fitted candidate for mediating the long-range correlation of the functional α-helices. α-helix 7 is strongly anti-correlated with the CS RC and dictates the movement of pigment 7. Recall, this pigment is correlated with many of the pigments in the system. α-helix 7 has bends either side that connect to the bottom and top of the CS. The reaction to the CS motion is therefore particularly strong. As a result, the motion of pigment 6 (which is the only pigment that is correlated with the CS) seems to be anticorrelated with pigments 4 and 7. The flow of dynamic information from the CS to α-helix seven, and on to pigment 7 and the other pigments is a possible approach employed by the FMO complex to reduce uncertainty and maintain a robust electronic landscape for EET.

The Constrained Monomer Interface

Relative to the monomer simulation, the trimer displayed a decreased RMSF in several key regions of the FMO complex (Figure 4.3). One noteworthy example is the significantly decreased RMSF of α-helix 1 that occurs due to interactions with the adjacent monomer. It is important to note that α-helix 1 is (a) strongly anticorrelated with the motion of the CS in its own monomer and (b) strongly correlated with the motion of the α-helices and pigments in the adjacent monomer. Indeed, the motion of α-helix 1 of monomer 1 is found to be correlated with α-helices 2 through 4 ($C_{ij} \approx 0.5$), and pigments 1 and 2 ($C_{ij} \approx 0.6$), of monomer 2. As the FMO complex displays rotational symmetry, α-helix one may act to communicate between the functional units of the trimer and thereby help constrain the conformations to (coherent) EET-competent ones. With the above observations in mind, it is predicted that the trimeric structure adds a further level of organization that is exemplified by the interactions of α-helix one.

4.4 Constrained Geometric Simulation of Photosystem II

Photosystem II (PSII) is a multisubunit PPC found in the thylakoid membranes of all types of plants, algae, and cyanobacteria [10]. At the heart of PSII is the reaction centre, where charge separation is initiated and cofactors (chlorophyll, plastoquinone, pheophytin) are found arranged in two symmetrical branches. Of these two branches, only one is involved in the primary electron transfer from water to plastoquinone [207]. At the base of the active (left) and inactive (right) branch is a pair of chlorophyll a (Chl) molecules. Each branch contains one accessory Chl and one pheophytin. The left, active branch of PSII can be seen in Figure 4.9 and comprises pigments labelled 4-5, 6, and 8. The preferential energy transfer, which prerequisites electron transfer, dynamics along the left chain of the two symmetrical branches have been attributed to a set of residues that collectively lower the site energy of Chl 6. The protein environment is found to be the greatest impact on the site energies, with water and cofactors having a negligible effect [249]. This claim is also supported by a previous mutagenisis study [199] that showed residue D179 within α-helix 10 impacts the site energy of pigment 6 and therefore the optical spectra. The MD investigation concluded that the protein environment accounts for the differences in activity between the two chains and that particular residues give rise to the energy sink in the active chain.

The crystal structure of PSII (PDB : 3ARC, 1.9 Å) used in Zhang et al. [249] has been employed for geometric simulation. This protonation states were automatically determined by Gromacs. As is the case in the FMO complex, PSII has an extremely rigid structure. This may be a general characteristic of light harvesting complexes, which could employ

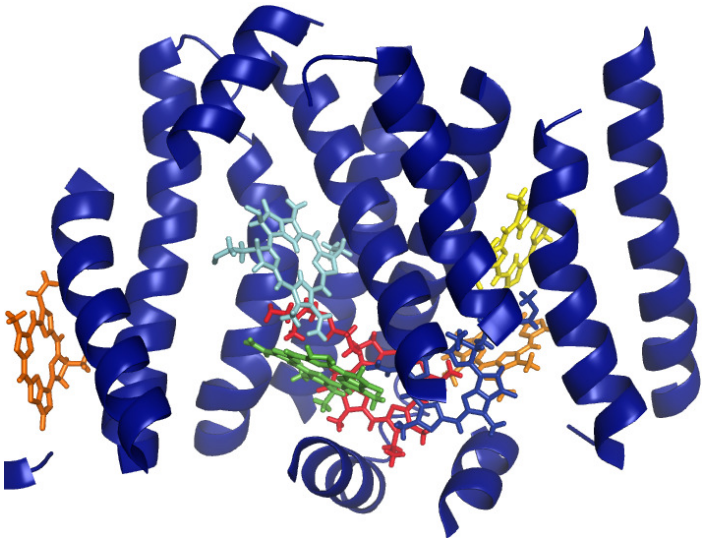


Fig. 4.9 Chains A,D, and E of PSII, which largely belong to the largest RC (RC1). The pigments found in this RC have also been displayed; namely 4 (red), 5 (blue), 6 (green), 8 (cyan), 7 (orange), 9 (yellow), 10 (dark orange).

rigid structures to minimise energetic disorder. Further analysis will be needed to test this hypothesis. The rigidity of PSII is evident by the large RC that is found at H_{cut} = -3.5 kcal/mol. This RC encompasses portions of 22 α-helices as well as the central 6 cofactors. The large RC is found to dissolve into several smaller ones, with a sharp transition at H_{cut} = -4.0 kcal/mol. Most interesting is the second largest RC (RC2) that encompasses 3 central α-helices and Chl 4, 5, and 6 (Figure 4.10). Chl 4 and 5 form the special pair and are the site where initial photoinduced charge separation occurs. Low frequency motion has been investigated by extracting the PCs from the constrained geometric simulation. PC1 describes positive spatial correlations between charged protein residues and pigment 6. α-helix 10, the short α-helix at the base of RC2, is also correlated with the active chain. Therefore, stronger interactions give rise to positive correlations in the active pathway (Figure 4.11), while cofactors of the inactive pathway form a distinct RC from the special pair.

At H_{cut} = -4.5 kcal/mol all α-helices are either flexible or form independent RCs. Of the cofactors, only the special pair belong to the same RC, while the porphyrin rings of the remaining cofactors form independent RCs. The correlation matrix shows the strong correlations between the special pair, and stronger correlations are still identified between the special pair and pigments in the active chain at a H_{cut} = -4.5 kcal/mol. As with H_{cut} = -4.0 kcal/mol, low frequency, positive correlations were found between the active chain and residues, including E329, E189, and D170, that have been shown to affect the active chain site energies [249].

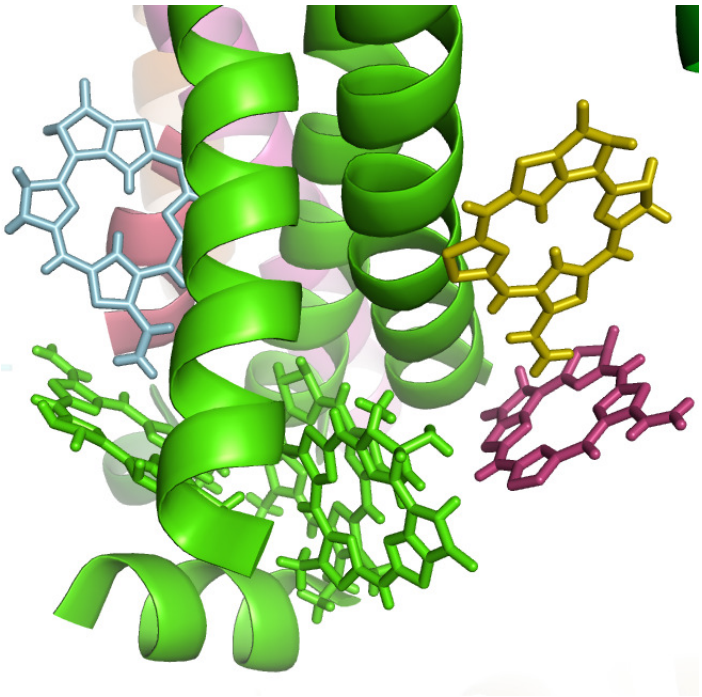


Fig. 4.10 A large RC (green) encompasses Chl 4,5, and 6, as well as α-helix 10 at the base (H_{cut} = -4.0 kcal/mol). This RC thereby introduces correlated motions between pigments involved in the active branch.

As discussed in Chapter 2, the protein plays a key role in tuning the optical properties of pigments and have been suggested to give rise to correlated inhomogeneous broadening through spatial correlations. This correlated disorder can act to maintain optimal pigment distance and transition dipole moment orientation. RC2 represents a collection of geometric constraints that act to enforce these aspects of energy transfer, protecting large variations in excitonic couplings from the warm, disordered photosynthetic environment. Furthermore, the spatial correlations between the active chain and structural elements that modify their site energy could give rise to correlated disorder in the reaction centre of PSII.

Summary

The trimeric FMO complex of GSB is a well-studied example of a photosynthetic pigment–protein complex, in which the electronic properties of the pigments are modified by the protein environment to promote efficient excitonic energy transfer from antenna complexes to the reaction centres. By a range of simulation methods, many of the electronic properties of the FMO complex can be extracted from knowledge of the static crystal structure. However, the recent observation of long-lasting quantum dynamics in the FMO complex

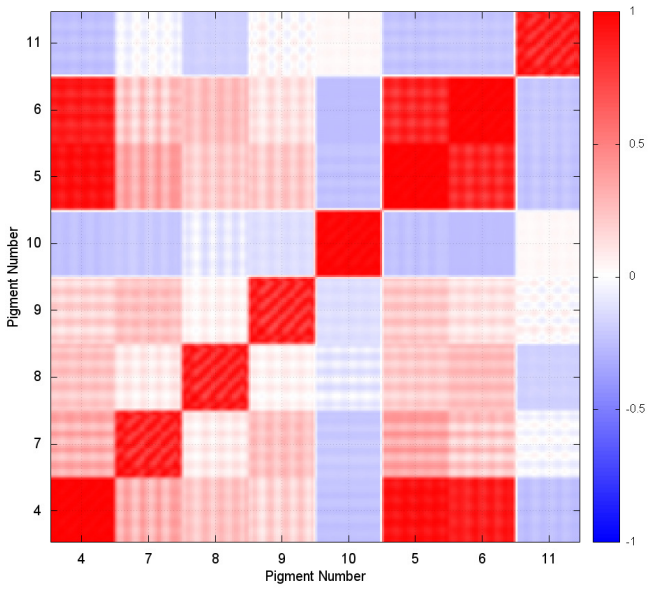


Fig. 4.11 Correlation matrix for cofactors in PSII (H_{cut} = -4.0 kcal/mol). Strong, positive correlations can be found between the special pair (pigments 4 and 5) and pigments 6 and 8 in the active branch.

point to conformal motions as a key factor in protecting and generating quantum coherence under laboratory conditions. While fast inter- and intra-molecular vibrations have been investigated extensively, the slow, conformational dynamics which effectively determine the optical inhomogeneous broadening of experimental ensembles has received less attention.

Geometric simulation is employed to generate a conformal ensemble of the full FMO complex. This technique has a high sampling efficiency, allowing the conformal landscape to be fully explored. The anharmonic dynamics are assumed to describe long-time scale, large-amplitude motions that retain the bonding and non-covalent interaction networks identified in the crystal structure. Statistical and principle component analyses of these dynamics reveal highly correlated low frequency motions between functionally relevant elements, including strong correlations between pigments that are excitonically coupled. We find similar behaviour in PSII, where rigid structures appear to be employed to organise pigment-protein dynamics. The analysis of the FMO complex reveals a hierarchy of structural interactions which enforce these correlated motions, from the level of monomer-monomer interfaces right down to the α-helices, β-sheets and pigments. In addition to inducing strong spatial correlations across the conformational ensemble, we find that the overall rigidity of the FMO complex is exceptionally high. These observations support a claim that the FMO

complex has evolved to reduce spatial disorder. In Chapter 5, quantum chemistry calculations will be used to investigate the optical properties of pigments 3 and 4 and identify whether the spatial correlations induce correlations in their energetic properties.

Chapter 5

Evidence for Correlated Static Disorder in the Fenna-Matthews-Olson Complex

Outline

In Chapter 4, geometric simulation is employed to simulate the anharmonic, long-time scale dynamics associated with the experimentally derived structure of the 21 000 atom Fenna-Matthews-Olson (FMO) complex. Within the conformal ensemble, strong spatial correlations between the pigments and protein structures, such as α-helices, that have been shown elsewhere to modify their optical properties were found. In particular, pigments 3 and 4 display strong correlations in the mode that describes the largest variation, as identified using principal component analysis (PCA), which is used to identify statistically significant large-amplitude motions explored within the conformal ensemble. Pigments 3 and 4 are the primary participants in the two lowest energy excitons that show long-lasting quantum beats in 2DES spectroscopy. The current analysis is therefore aimed to develop the hypothesis that positive inter-pigment correlations in the energetic disorder resulting from slow conformal motions are used in PPCs reduce inhomogeneous dephasing.

Calculating these properties is a daunting theoretical problem, as it requires (i) a method for calculating extended structural dynamics on timescales that are much longer than those accessible to traditional molecular dynamics simulations and (ii) accurate large-scale electronic structure calculations capable of describing inhomogeneous environments of several thousand atoms. Point (i) is addressed in Chapter 4, and allows us to bias a supplementary geometric simulation along the PCA eigenvectors that describe the largest-variation deformation in the ensemble. Of course, the spatial correlations between pigments 3 and 4 along the low-frequency mode do not guarantee correlations in their site energies, which

are sensitive to long-range electrostatics in the protein. In what follows, snapshots along the low-frequency mode are analysed using state-of-the-art electronic structure techniques to provide insight into the excited state energies. The results demonstrate for the first time that spatially correlated motions of the protein lead to *significant* positive correlations of the excited state energies on different pigments.

5.1 Computational Challenges in the Study of Static Disorder

The coupling of the pigment excited state energies to the protein vibrations can be characterised using the spectral density, which contains fluctuations in (i) the excitonic coupling that are responsible for energy transfer and delocalisation of the excited states and (ii) the local pigment site energies, which are tuned by the protein environment. The protein environment is able to dissipate the excess energy of the excitons and distribute it over many degrees of freedom as they relax downward to the excited state of the reaction centre. The proposition that correlated fluctuations in the inter-pigment site energy fluctuations could give rise to the observed quantum beats in 2D electronic spectroscopy [138, 169] prompted discussions as to the functional role of such correlations [181, 195, 1, 159, 66]. Structure-based simulations have been called upon to investigate the protein environment in light-harvesting systems.

Classical molecular dynamic (MD) simulations have been used to investigate spectral disorder in the FMO complex and uncovered weak correlations between the movements of pigments [165]. However, these spatial correlations did not result in significant correlations of local site energy fluctuations, which were calculated using semi-empirical quantum chemistry calculations. However, in addition to the limited timescales that can be investigated using MD simulations, this approach suffers from another factor. As discussed by Jing et al. [114], the mismatch between ground state molecular mechanical/quantum mechanical (MM/QM) potential energy surfaces is problematic when post-processing MD trajectories with QM excited state calculations – contributions to the site energies caused by high frequency intramolecular pigment vibrations are significantly over-estimated.

The complex interplay of electric fields and pigment-protein interactions that determines pigment site energies fluctuations does not necessarily follow from the observed mechanical rigidity. This reliability from controlled disorder would be a highly desirable and remarkable achievement in a self-assembling system of excitonic wires. Fidler et al. [68] have presented experimental evidence for such correlated disorder in the FMO complex. However, these experiments are unable to distinguish between the intra-pigment vibronic excited state

effects [245, 66, 218] and correlated disorder as the source of the electronic coherences. To truly simulate static disorder requires simulating all-atom anharmonic dynamics on long timescales, which has not been achieved previously. The computational efficiency of the FRODA technique allows one to explore large, long time scale motions that, as stated previously, go beyond the harmonic approximation that limits normal mode analysis approaches.

5.1.1 Using Principal Component Analysis to Generate Static Disorder

An extensive ensemble of structures permitted within the FIRST constrained dynamics have been generated in Chapter 4. The agreement of FIRST with experimental protein motions and its ability to provide insight into a wide range of functions suggests that the dynamics of the native state emerge naturally from a simple network of contacts [22, 115, 130, 221, 145]. Critically, this method has also been used to study long time scale gating mechanisms in ion channels [22, 130] and microsecond enzyme functions [238, 238], highlighting the ability of this technique to investigate functionally relevant motion on timescales relevant to static disorder. This Lagrangian-based constraints approach is actually beneficial for another reason outlined above by Renger et al. [183] for normal mode analysis, whereby the intrapigment contributions to the spectral density are removed. Normal mode analysis [147] has been used to calculate low-frequency, collective motions that, critically, fix the intra-pigment bonds and angles, thus allowing the high frequency component to be filtered out. Indeed, the neglected pigment modes have very small Huang-Rhys factors [114] and should not contribute to disorder, though they are clearly important for fast EET dynamics. The calculated spectral density, although containing some correlations in site energy fluctuations, did not impact the correlations of exciton population dynamics and dephasing [183]. However, the breakdown of the harmonic approximation at low frequencies in NMA has been highlighted and suggests that the impact of very soft, slow conformal motions can only be accounted for phenomenologically.

As discussed in the next subsection, although large-scale quantum mechanical excited state calculations are becoming more routine, QM simulations comprising thousands of atoms are still a costly endeavor. As a result, the number of structures of the ensemble to be analysed is reduced from more than 5000 stored to eight, whilst ensuring that as large a dynamical range as possible is captured by the ensemble. To do this, principal component analysis (PCA) is employed, also termed essential dynamics (so-called because the motion described by the low-dimensional subspace has been linked to protein function [8]). PCA of FIRST structural ensembles have been extensively investigated [54, 55]. The essential dynamics described by PCA have been shown to be consistent over a wide range of H_{cut}

values in FIRST simulations [55], and is therefore not too dependent on initial choices made by the user. A comparison between the sampling of FIRST and MD conformal ensembles has found that the MD projections are nearly completely contained within the FIRST runs, which supports the assertion that FIRST trajectories have a greater sampling efficiency than MD simulations [54, 55].

The first principal component (PC1) describes the largest percentage of the total motion. Indeed, large secondary structure motions will be shown to be the root cause of site energy fluctuations, suggesting that the effectively static regime of energetic disorder probed by optical experiments has been accessed. In a previous study, the overlap between the experimental dynamical modes of myoglobin with the motion described by the PC1 of the FIRST dynamics [55] suggests that the first principal component is able to identify biologically relevant motion. Here, PC1 captures approximately 10% of the ensemble variance in the FIRST-generated conformational ensemble. This motion is therefore the best estimate of the large-amplitude, long-time scale, delocalized and anharmonic motions that are associated with static disorder in the FMO complex.

Finally, to further verify the significance of the motion along PC1, the scalar product is employed to compare the PC1 vector to normal mode eigenvectors identified using coarse-grained elastic network modelling (ENM) [220]. For the first 10 non-trivial low-frequency modes (7-16), PC1 displays overlaps above 0.1 with several of the ENM modes and above 0.4 with two of them. Strong agreement between the motion identified using PC1 with the low-frequency modes of ENM is therefore found. Furthermore, strong spatial correlations (>0.9) are observed between pigments 3 and 4 for the first 10 PCs (in addition to PC1). The motions of α-helix 7 and α-helix 8 along PC1 give rise to the (correlated) energetic disorder. The similarity in the motion of these α-helices in the PC1 eigenvector to the equivalent motion in other, higher PC eigenvectors is therefore examined. The results indicate that similar motions of these helices are explored in multiple other PCs. This suggests that the motions explored in this study, and the behavior of the static disorder as a result, are significant features of the natural dynamics of the entire structure. Therefore, PC1 captures the largest fraction of the ensemble variance whilst representing the observed spatial correlations between pigments 3 and 4, and is hence a good choice for investigation in this study.

5.2 Biased Motion along Principal Component 1

To explore the limits of the motion described by the eigenvectors of PC1, the FRODA module is used to efficiently generate motion along the largest-variation mode. This approach has been shown to be a rapid and efficient method for investigating flexibility and internal motion

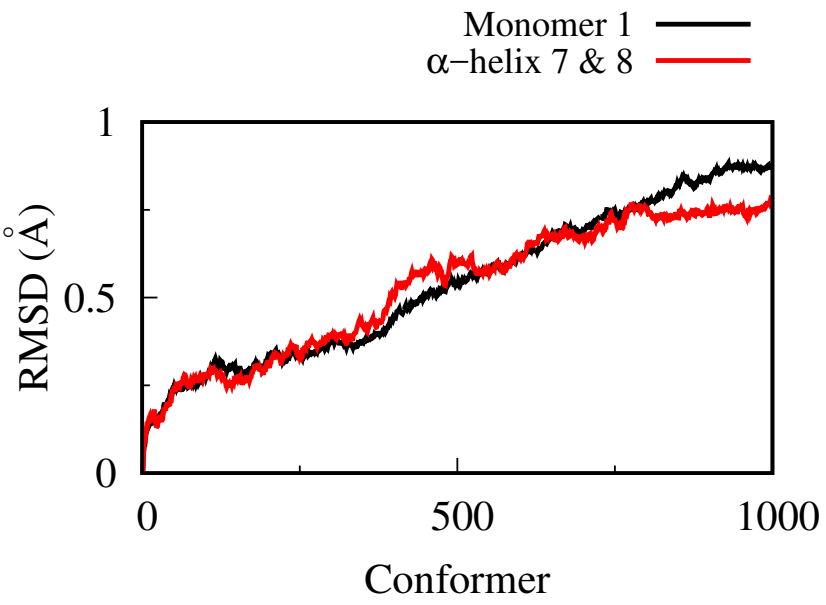


Fig. 5.1 RMSD of the atomic positions averaged over monomer one (black) and α-helix 7 and 8 (red). The motion described within the initial "linear phase" from 50 to roughly 800 is the most reproducible and biologically relevant motion. As the constraints hinder further exploration the RMSD begins to flatten (> 800).

in the protein structure [113]. In each step of the biased simulation, the algorithm perturbs all atoms by a step of the order of 0.01 Å along the PC1 eigenvector, along with a similarly sized random perturbation. The atom positions are then relaxed to satisfy steric exclusion, bonding geometry and hydrophobic-tether distance constraints. This information is encoded in the ghost templates, as is described in Figure 4.1. As the motion retains the interaction network and bonding geometry of the input structure, the generated conformations along PC1 are assumed to be biologically relevant.

Biased geometric simulations along low-frequency modes are characterised by an initial phase of "easy" motion, in which the root-mean-square deviation (RMSD) increases linearly and the motions are highly reproducible. Eventually, steric clashes and bonding constraints hinder further exploration along the bias direction and the RMSD stops rising linearly. The motions described in the initial "easy" phase hold the most biological significance [112] as these are motions which the protein can easily explore through random thermal motion. Figure 5.1 shows the RMSD of the atoms in monomer one and the atoms belonging to the residues that form α-helices 7 and 8. The expected linear increase is found over the first 750 conformations, with the subsequent motion (particularly for the α-helices) beginning to be impeded by the constraint network or steric contacts.

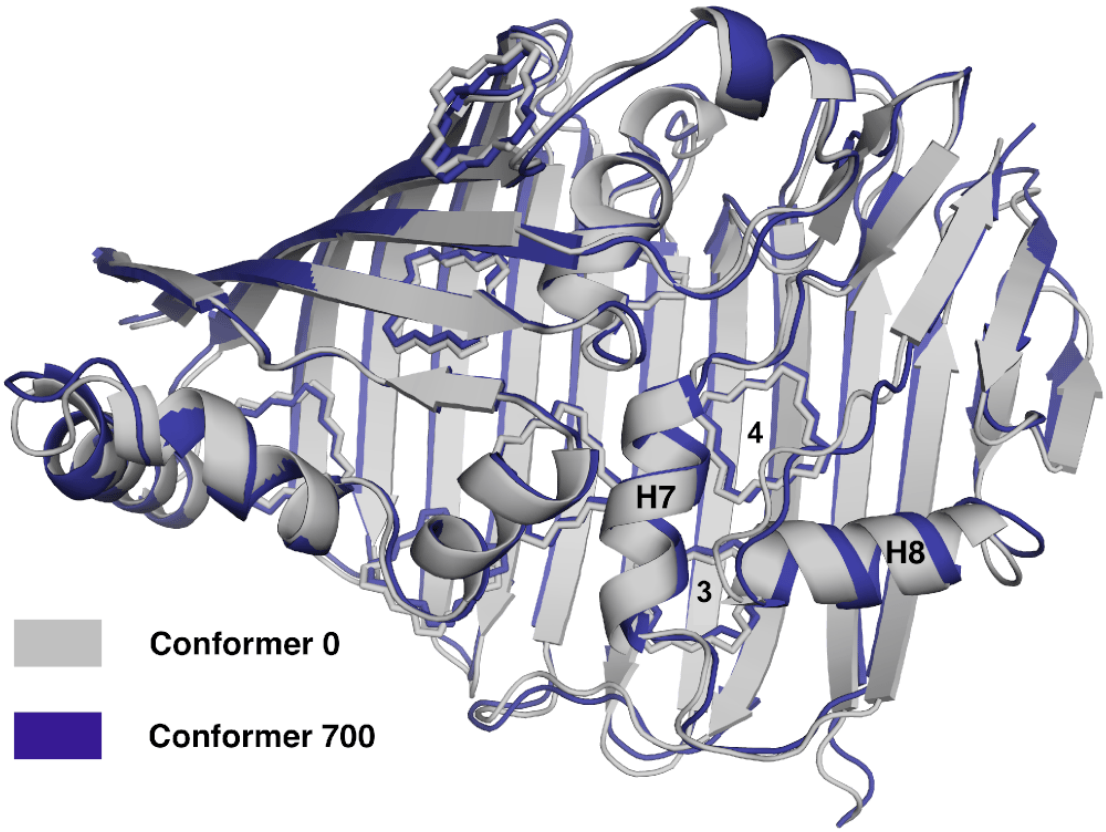


Fig. 5.2 The FMO complex crystal structure (conformer 0) is biased along PC1, from which eight snapshots are selected. These snapshots fall within the reproducible motion generated by the biased FRODA simulation (Section 5.2). The computational efforts have therefore been focused on conformations belonging to the linear regime, analysing every 100th conformation. The 50th conformer has additionally been analysed, as there is a particular large amount of motion in the very early stages of the simulation. Above, the limits of the PC1 motion are displayed, highlighting α-helices that play an important role in determining the energetic landscape for pigments 3 and 4.The RMSD of α-helix 8 (2) is greater than the average RMSD of monomer 1 (0.8).

The pattern of displacement described by PC1 is depicted in Fig. 5.2. The majority of the motion in PC1 is centred around α-helices 7 and 8 – two structural elements that have been shown to be important in creating the excitonic energy funnel leading to pigments 3 (the energy sink) and 4 [184, 44]. As pigments 3 and 4 are also the primary participants in the two lowest energy excitons that show long-lasting quantum beats in 2DES spectroscopy [68], the investigation focuses on the motion and site energies of pigments 3 and 4 and the protein environment. To do so, 2000-atom cluster centred on the pigment of interest and containing all protein residues, water molecules and pigments within 15 Å are extracted from eight conformations. The pigment-protein environements are then simulated in their entirety using quantum chemistry calculations.

5.3 Density Functional Theory

A quantum state can be described using $|\Psi(t)\rangle$ whose evolution is described using the Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$

where $\hat{H}$ is the Hamiltonian that describes the interactions with the degrees of freedom under study. The dimension of Hilbert space used to calculate $|\Psi(t)\rangle$ grows exponentially with the number of particles in the system, meaning that the Schrödinger equation can only be solved exactly when a few particles are present in the system. This means that the Schrödinger equation on its own cannot be used to answer many problems in many-body physics. In 1927, Llewellyn H. Thomas [225] and Enrico Fermi [67] independently attempted to solve the electronic Hamiltonian using the full electronic density as the fundamental variable of the many-body problem. This lead them to derive a differential equation for the density without employing the one-electron orbital. Pierre Hohenberg and Walter Kohn built upon this early work and presented two theorems showing that that the electron density uniquely determines the Hamiltonian operator and thus all the properties of the system. The Hohenberg-Kohn (HK) theorems describe the ground state of an interacting electron gas in an external potential $v(r)$;

Theorem 1: The electron density uniquely determines the Hamiltonian, and thereby all the properties (external energy, potential energy, total energy) of the system. If two systems have the same density in different potentials, their difference is equal to a constant.

Theorem 2: A universal functional $F[n]$ for the energy can be defined in terms of the density. The exact ground state is the global minimum value of this functional, and will therefore deliver the ground state energy if and only if the input density is the true ground state density.

The universal functional lies at the heart of density functional theory and depends on the kinetic energy $T[\rho]$ and the electron-electron interaction E_{ee} according to:

$$F[\rho] = T[\rho] + E_{ee}$$

The treatment of the kinetic and internal potential energies are the same for all structures. As a result, the universal functional applies to a variety of systems ranging from single atoms to large proteins. Knowing the universal functional would give the Schrödinger equation exactly. However, the explicit form of the $T[\rho]$ and E_{ee} functionals remains a major challenge. The HK theorems state that there exist many trial density where the density does not correspond to the ground state antisymmetric wave function. Haim Levy applied the constrained search method to obtain the ground state energy variationally, making use of the observation that the density that minimises the total energy is the exact ground state density. [140]. To do so, he rewrote the the universal functional as:

$$F[n] = \min_{\Psi \to n} \langle \Psi | \hat{T} + \hat{U}_{ee} | \Psi \rangle$$

where a constrained minimisation is conducted over the subspace of all antisymmetric ground state wavefunctions (Ψ) that give rise to the same density (n).

In 1965, Kohn and Lu Jeu Sham devised a practical scheme for determining the ground states [128]. Instead of using the kinetic operator, which is inherently non-local, Kohn and Sham calculated the kinetic energy of a system of non-interacting electrons that produced the same electron density of the interacting system of electrons. Importantly, by using a non-interacting reference system the Hamiltonian can be expressed in the form of Slater determinants. This allows the energy functional to be expressed in terms of orbitals that minimise the non-interacting electronic kinetic energy. The reference potential, which ensures that the density of the non-interacting reference system is the same as the true density of the interacting system, can then be determined by minimising the Kohn-Sham functional with respect to the density of N integrated particles. Applying the variational principle, namely, identifying what condition the orbitals must fulfil in order to minimise this energy expression, to the Kohn-Sham functional gives an equation for minimising the ground state density. This reference potential depends on the solution of the Kohn-Sham orbitals (which satisfy the one-electron Kohn-Sham equations) through the electronic density. The equation is therefore solved self-consistently, making sure the density used to construct the reference potential coincides with the obtained solutions of the density:

$$n(\mathbf{r}) = 2\sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2$$

Exchange and Correlation

If the exchange-correlation (XC) energy and potential were known, then the Kohn-Sham DFT approach would be exact. The exchange interaction stems from the indistinguishability of electrons in our system. Pauli's exclusion principle states that fermions with the same spins occupy orthogonal orbitals that are spatially distinct, and this decreases the energy. In an electron gas, where all the electrons interact, there is correlated movement that also results in the lowering of the total energy [171]. Homogeneous electron gas, while dissimilar from the inhomogeneous molecular systems under study, has well understood properties and represents the simplest system of correlated electrons. The two most common approaches for treating the XC problem are the local density approximation (LDA) and the generalised gradient approximation (GGA). In LDA, the homogeneous electronic system is considered as being locally homogeneous. GGA, which improves on the LDA, is performed using the using the Perdew, Burke and Ernzerhof (PBE) XC functional [175].

ONETEP

A scaling problem inhibits conventional DFT methods from tackling large systems, such as proteins. The computational load of DFT methods is of the order N^3, where N is the number of particles in a system. The formulation of linear scaling methods began in the late 1990's with the hope of investigating larger systems [86]. ONETEP (Order-N Electronic Total Energy Package) [210] is a DFT approach of order N that, when implemented in parallel, allows systems many with many thousands of atoms to be studied without a loss of accuracy [45].

A basis set, chosen to represent the operators and wave functions, can be defined as a collection of vectors found in a space where the problem can be solved. In the case of DFT, the basis set is a set of one-particle functions that are generally not orthogonal and can be used to build the molecular orbitals. These basis sets are complete as they span the space needed to represent any wave function. In the methodology for linear scaling a density matrix operator is used:

$$\hat{\rho} = \sum_i \omega_i |\psi_i\rangle \langle \psi_i|$$

The density matrix includes a variable ω which is related to the probability of a particular orthonormal ψ state i occurring. This allows the density to be projected onto a space governed by eigenvalues (one for occupied state; zero for unoccupied states). These orthonormal states can be expressed as a linear combination of basis states:

$$|\psi_i\rangle = \sum_j c_j^{(i)} |\phi_j\rangle$$

In ONETEP, the Kohn-Sham orbitals are constructed in terms of Wannier functions, which are constructed from a Fourier series of Bloch wavefunctions [213]. Importantly, as the Wannier functions decay exponentially, this property is also valid for the density matrix and can be exploited to obtain a linear scaling method. This separability of the density matrix, which stems from the decay rate, allows the density matrix to be written in quadratic form [100]. A basis set makes up non-orthogonal generalised Wannier functions (NGWFs) that represents occupied Kohn-Sham orbitals in a subspace of Hilbert space.

Linear-scaling methods make use of a property within electronic systems known as "nearsightedness" [178]. Kohn identified nearsightedness of electronic systems, which implies that local electronic properties, such as the density, depend significantly on the effective external potential only at nearby points. This is implemented in ONETEP by a mathematical truncation, whereby the NGWFs and density kernel only have non-zero elements within a spherical region. The exponential decay rate of the NGWFs and the density matrix is found to be proportional to the HOMO (highest occupied molecular orbital)-LUMO (lowest unoccupied molecular orbital) gap. As a result, there is no dependence on the size of the system and the essential information scales linearly with N. Implementing nearsightedness in quantum mechanical calculations significantly decreases the computational expense and the losses in accuracy can be recovered by converging physical characteristics (such as NGWFs radii) of the system for particular properties. ONETEP employs a technique whereby the NGWFs are expanded in an underlying basis set of primitive functions, chosen to be the periodic cardinal sine functions (psinc). Using the relation of these functions to plane waves, the kinetic energy can be calculated efficiently using fast fourier transforms at the centres (grid points) of the psinc functions that lie within the truncation region of the NGWFs [100].

The pseudopotential approximation allows the inert core electrons and the shielded ionic potential of the nucleus to be replaced by a weak potential. As the valence electrons do not need to be orthogonal to the (pseudo) core electrons, their representation is smoother and can be represented by a lower kinetic energy cutoff. The valence states can therefore be accurately represented by a much smaller plane-wave basis set.

The initiation phase involves building the NGWFs. A psuedoatomic solver [191] can be used to generate states that can be expected to form an "ideal" atomic orbital basis for a given calculation. This means that the pseudopotential of single isolated ions are used to generate the effective potential, which for a given XC functional, allows calculation of the Kohn-Sham states in a self-consistent procedure. The effective potential and electronic charge density must be solved, where the charge density is equal to the diagonal elements [n($\mathbf{r}$)=ρ($\mathbf{r}$,$\mathbf{r}$)] of the density matrix:

$$n(\mathbf{r}) = 2\sum_{\alpha\beta} \phi_\alpha\ (\mathbf{r}) K^{\alpha\beta} \phi^*_\beta\ (\mathbf{r}) = 2\rho(\mathbf{r},\mathbf{r})$$

The total energy is a functional of the density kernel and the NGWF. As the energy is a functional of both NGWFs and of the density kernel, it must be minimised with respect to both functions. Optimisation involves nested self-consistent loops that achieve minimisation as follows:

Outer loop - optimises the NGWFs with fixed density kernel.

Inner loop - minimises the density kernel with fixed NGWF.

These calculations, constrained by idempotency and normalisation, continue until the gradient for the calculated NGWFs falls below a threshold [212].

The plane wave basis set has clear benefits over the atomic orbital basis set. Atomic orbitals fill the simulation cell in a non-uniform way making their refinement problematic. They are also defined by a number of independent factors and are not orthogonal, making it difficult to improve their quality. The ONETEP psinc basis set is constructed from plane waves that maintain orthogonality while being localised. Critically, the quality of these plane waves can be improved using only the kinetic energy cutoff. The psinc basis is defined using plane waves $e^{-iq \bullet r}$ with the wave vectors belonging to a cube in reciprocal space [211]. The grid spacing for the psinc functions is related to the psinc kinetic energy cutoff and the ONETEP cube defined by the wave vectors [99].

Excited state energies can be investigated using ONETEP, allowing the absorption spectrum to be calculated for optically active molecules. The absorption spectrum measures the capability of a molecule to absorb radiation at particular wavelengths. In general, as a beam of light interacts with matter, it is modified as atoms convert a portion of the electromagnetic energy into internal energy. This internal energy could be electronic, translational, vibrational, or rotational. The band structure for an electronic system is given as a band index graph, where the allowed momentum and corresponding energy that an electron may have in each band is shown and allows identification of the HOMO-LUMO gap [230]. In ONETEP, following the method of Ratcliff et al. [166], the low-energy unoccupied conduction states are represented by a second independent set of NGWFs in addition to the NGWFs that are used for the valence states. A projection operator is employed to identify the contributions from the conduction states to the Hamiltonian. Optical spectra are calculated using Fermi's golden rule, which uses quantum mechanical concepts to calculate the transition rate between initial and final states, as an approximation. Successful application of this method has been used to identify transitions that are responsible for spectral peaks. While powerful, this method does have some limitations, including an inability to represent delocalised states and unbound states.

Implementation

Supercomputers allow researchers to employ thousands of processors in parallel to simulate computationally demanding systems [105]. They have access to a shared memory through a sophisticated interconnect system, allowing information to be efficiently shared between different nodes. The program calls on message passing interface (MPI) subroutines for communication between processors. The data parallelisation strategy employed by ONETEP splits the workload between (a) atomic data and (b) simulation cell data. Point (a) refers to data directly associated with the atom (such as the expansion of the NGWFs in the psinc basis set and pseudopotentials) that scales linearly with the number of atoms. To make the computational load for each processor similar, they are assigned an equal number of atomic quantities (that contain all the information for an atom) so the number of NGWFs on each processor is the same. The data that is distributed to processors is biased to atoms that are in close proximity. This minimises the data that needs to be communicated with different atoms, such as between those whose NGWFs support regions overlap. Point (b) refers to data that is specific to the cell size, such as charge density, Hartree potential, XC potentials. For these calculations, the simulation cell is divided into sections along the lattice vector and each section is allocated to a processor. The bulk of the computation involves computing the electronic energy, which has contributions from the band structure energy and the XC energy.

ONETEP Parameters

Spherical clusters of radius 15 Å were extracted for analysis (using a script compiled by Dr. Daniel Cole). The ground state electronic structure was computed using the Perdew–Burke–Ernzerhof (PBE) XC functional using an implicit solvent model with a relative dielectric of 80 [61, 139]. Note that the dieletric constant does not represent the environment, but it used in the optimisation of the density kernel. The single-particle density matrix is represented by *in situ* optimised orbitals, known as non-orthogonal generalised Wannier functions (NGWFs) [213], which are themselves expanded in a periodic cardinal sine (psinc) basis with an energy cut-off of 1020 eV. The NGWFs were localised in real space with cut-off radii of 10 Bohr. Two conduction states were optimised for every pigment present in the cluster, and optical spectra were computed using Fermi's golden rule in a joint valence and conduction basis [166]. Conduction calculations in the FMO complex were previously found to converged to within approximately 15 cm^{-1} with respect to changes in cut-off energy, NGWFs radius, system size and dielectric medium [210]. No QM structural optimisation was performed. To compensate for the systematic underestimation of

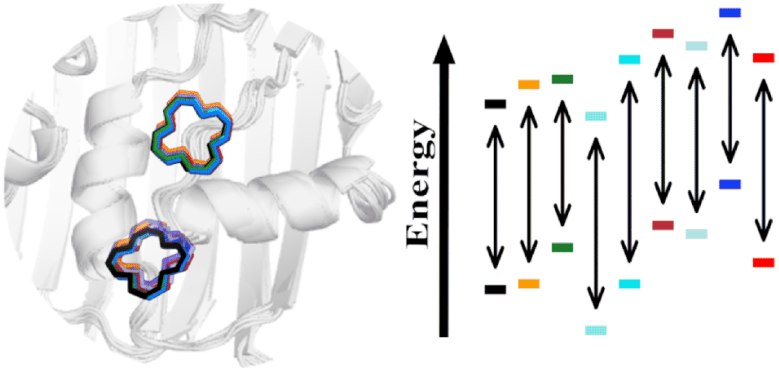


Fig. 5.3 For each pigment in each snapshot, the local pigment protein environment is extracted by taking all the atom that lie within 15 Å of the central Mg atom, which has been shown previously to converge the site energy [44]. ONETEP is then employed to calculate site energies, allowing us to investigate structurally induced site energy fluctuations.

the HOMO-LUMO gap [210] calculated site energies have been uniformly shifted by 3710 cm^{-1}.

By calculating the site energies using the ONETEP code, both valence and conduction state density matrices are represented using NGWFs. By optimizing the NGWFs functions *in situ*, ONETEP achieves near-complete basis set accuracy for the description of both valence [74] and conduction [166, 44] states for systems comprising many thousands of atoms, including entire proteins [46]. This is particularly important for the description of energy transport in PPCs, where pigment optical transition energies depend critically on electrostatic interactions with their environment. In this regard, the *ab initio* methods presented here are seemingly preferable to other structure-based methods for extraction of site energies, which approximate some or all of the PPC by classical point charges [116, 206, 111]. Indeed, it has been shown that these virtually parameter-free methods, when applied to the static crystal structure, give reasonable agreement with the experimental optical spectra of FMO [44].

5.4 Excited State Energy Calculations along the Largest Variation Trajectory

The variation in the site energies of the eight PC1 conformers is summarized in Fig. 5.4. The average site energies of pigments 3 and 4 along PC1 are 12,210 cm^{-1} and 12,513 cm^{-1}, respectively, and site 3 remains the energy sink in all conformations. These average values are in good overall agreement with previous calculations. For example, Adolphs et al. computed

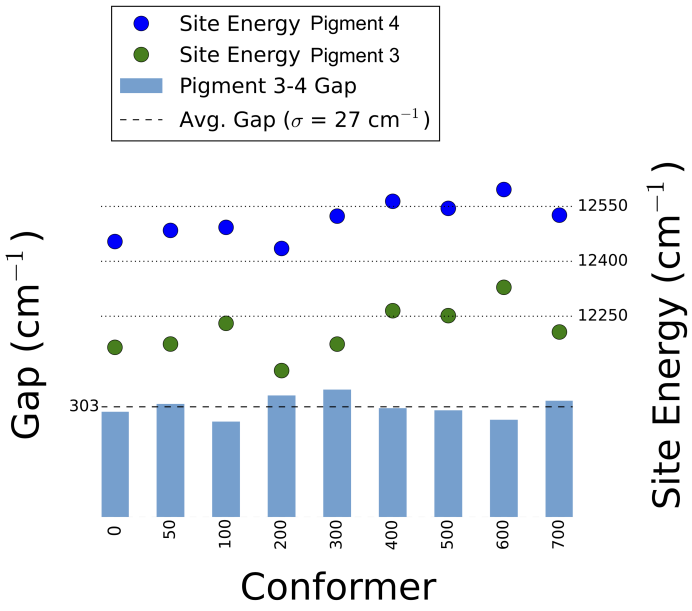


Fig. 5.4 Eight conformers were selected from the all-atom biased PC1 simulation and the site energies of pigments 3 and 4 were calculated using ONETEP. The difference in the site energies (gap) between pigments 3 and 4 is calculated for each conformer. The correlation coefficient between the site energies is R=0.91

values of 12,190 cm^{-1} and 12,380 cm^{-1} using a genetic fit to the experimental optical spectra and 12,195 cm^{-1} and 12,395 cm^{-1} using a classical, structure-based Poisson-Boltzmann approach built on phenomenological point charge distributions [4, 156].

Turning to energy variations over the entire PC trajectory, large standard deviations are found in the observed site energies of pigments 3 and 4 of 67 cm^{-1} and 52 cm^{-1}, respectively, corresponding to a full-width at half maximum (FWHM) of 158 cm^{-1} and 122 cm^{-1} for a Gaussian site-energy distribution. Encouragingly, these values are only slightly higher than the assumed or fitted theoretical literature values for the site energy disorder/inhomogeneous broadening of the lowest energy sites ($50-100$ cm^{-1}) [5, 64, 131, 183]. If these variations of the site energies were uncorrelated and Gaussian, the expected variations of the energy gap between sites 3 and 4 would be 85 cm^{-1}. Such energy gap variations would lead to extremely rapid inhomogeneous dephasing of inter-exciton coherences, faster in fact than the optical inhomogeneous dephasing times. However, as can be seen in Fig. 5.4, the environment-induced site energy fluctuations maintain a remarkably homogeneous energy gap ($303 \pm 27 cm^{-1}$) across different realisations of the structural ensemble, in spite of the much larger variances of the individual site energies. As is evident to the eye in Fig. 5.4, this arises from very strong correlation (R=0.91) of the site energies along complex, non-

monotonic trajectories which could not arise in a normal mode analysis, and suggests the occurrence of qualitative changes in the local environment of the pigments during the motion.

This is the key finding of this paper, which establishes the hitherto conjectured correlation of site energy disorder due to slow conformal motion and supports the experimental observable of correlated disorder [68]. The underlying PC1 motion will now be analysed to determine the physical mechanisms behind a) the relatively large fluctuations in individual site energies, and b) the very low static disorder in the site energy differences. It is very difficult to break down the influence of side chain, backbone and inter-pigment contributions to the quantum mechanical site energies and so, in what follows, a correlation analysis is performed to determine which structural elements drive and/or correlate the site energy shifts.

First, hydrogen bond donors have been shown to cause a substantial red-shift of the pigments' site energies and make an important contribution in maintaining the energy funnel and, critically, the energy sink at pigment 3 [4]. However, these hydrogen-bonds have relatively stable energies, measured to be around 3–8 kT, and they remain rigid in the constraints-based FRODA dynamics used here. As their influence is short range, this suggests negligible changes to the energy gaps between the pigments under slow conformal motion.

Next, the impact of larger secondary structure elements on the site energies of the two pigments are considered. In particular, the electric field of the dipoles associated with α-helices 7 and 8 have been shown to align with the difference in charge density between the excited state and the ground state of the pigments which cause red shifts in their site energies of around 200 cm^{-1} [4]. These shifts are the largest single contributions to the establishment of pigment 3 as the energy sink. Resulting from a dipolar interaction, the magnitude of the red-shift is likely to be sensitive to the distances between the pigments and these α-helices. To study the relationship between the pigments and their environment along the PC1 trajectory a cross-correlation analysis is employed. This analysis uncovers a high positive correlation in the motion of the pigments and the residues of α-helix 7 (Fig. 5.5, left). Similar to the behaviour of the pigment-protein hydrogen bonds, this strong positive correlation indicates that pigments 3, 4 and α-helix 7 move together as a single unit, which preserves the energy shifts on these sites from this α-helix. On the other hand, low correlations are found between the negative end of α-helix 8 (V349–K354) and the pigments, indicating a reasonably high variation in the distance between pigments 3 and 4 and the negative dipole of α-helix 8 (0.8 and 0.7 Å, respectively). However, these two distances are additionally found to be well correlated with each other ($R = 0.8$), suggesting that α-helix 8 motion can lead to substantial shifts in the pigment energies without changing their relative energy gap, which has been found in the quantum mechanical simulations.

The correlation coefficient has also been calculated between the pigment-residue distances and the pigment site energies (Fig. 5.5, right). The 16 strongest correlations involve residues that are located at the negative end of α-helix 8, as well as in nearby loops and the clam shell. There is a particularly strong correlation between the V349–pigment 4 separation and the site energies of the pigments 3 (E_m(B3)) and pigment 4 (E_m(B4)), reinforcing the idea that the pigment's proximity to α-helix 8 is a key contributor to the site energy variation observed in Fig. 5.4.

Summary

The ability to predict how transition energies change under very slow conformal motion opens the way for a detailed and realistic exploration of the potential functional impact of energetic disorder on the optoelectronics of PPCs. A number of theories point to the use of disorder in controlling average excitonic delocalisation lengths, with consequences for both optical spectra and energy transfer properties within and between antenna complexes [69, 39]. The critical links between slow conformal dynamics and photoexcitations in processes such as non-photochemical quenching [190], fluorescence blinking dynamics [82] and charge transfer in reaction centres has also been discussed [188]. The latter case should be especially interesting, as charges couple much more strongly to the protein environment and are consequently more susceptible to the influence of disorder. Under these conditions, it is thought the PSII reaction centre has evolved two distinct electron transfer pathways, each adapted for particular realisations of the unavoidable disorder [187]. In the FMO complex, by contrast, it seems that the disorder for excitonic processes can be suppressed below a level that might affect its function by high structural rigidity and correlated motion of major secondary elements.

With reference to the latter point, recent work in structural biology suggests that, in addition to protein structure, protein dynamics are also conserved in evolution due to their beneficial role in function [150]. In Chapter 4, the FMO complex is shown to score very highly against other protein structures using a comparative global measure of rigidity [234], and this rigidity hugely constrains the possible motions of its substructures, which has been shown to preserve excitonic structure. Calculating the excited state properties of 2000 atom clusters of the FMO complex are simulated by applying linear scaling electronic structure calculations to snapshots from constrained geometric simulations allows the energetic structure to be investigated. The resulting analysis shows that while slow, large amplitude conformal motion leads to large variations in the Q_y optical transition energies of pigments 3 and 4, correlations in the energy shifts greatly reduce variations in their energy gap across

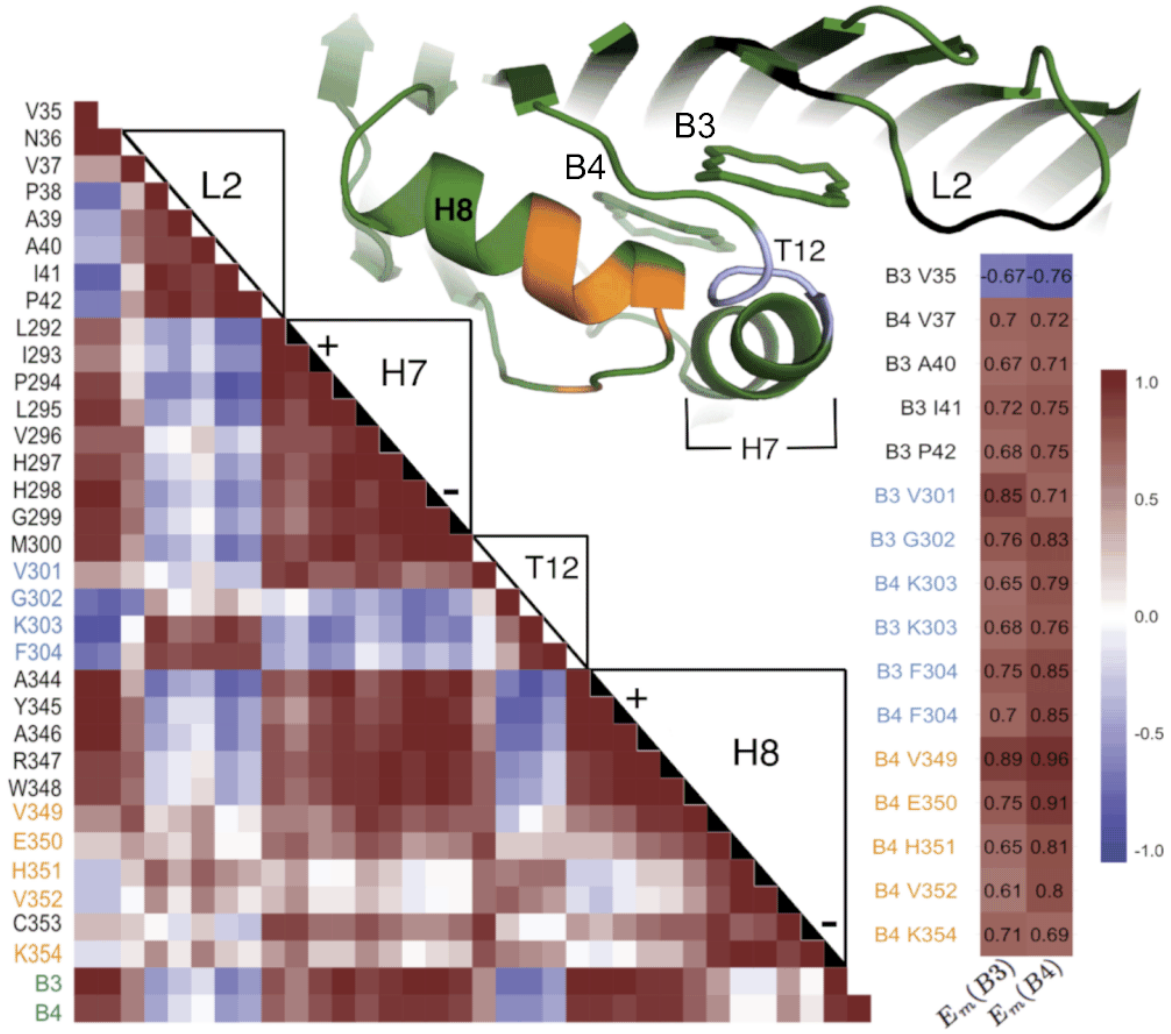


Fig. 5.5 Cross correlation analysis (left) between the motion of the pigments and nearby protein structures. The correlation coefficient (right) has been calculated between the pigment site energy fluctuations and the distance connecting the centre of mass of the pigment with the Cα atoms in the environment. The top correlations found when considering all the distances in the system is displayed, with coloured labels corresponding to coloured section of the protein model (above). When measuring the correlation between the pigment-Cα separation and the pigment site energies, only pigment-Cα pairs with distance variations above 0.1 were retained to focus the search on motions that are likely to have a strong impact on the site energies. The residues found to have the greatest influence on the site energy fluctuations are in α-helix 8, turn 12 and loop 2.

the ensemble. Although an admittedly speculative idea for the FMO complex, the concept of using mechanical structure to confine or rectify unavoidable thermal motion into neutral or beneficial trajectories could be a potentially powerful tool in artificial chromophore arrays and nanodevices [215], and the theoretical approach presented helps open the door to exploring this in biological, molecular and solid state examples.

Chapter 6

Concluding Remarks

In the present thesis, constrained geometric simulation has been employed to (a) construct the residue geometry network and (b) investigate the static disorder in the FMO complex. With reference to (a), the relationship between amino acid centrality and evolutionary rate (dN/dS) has been investigated. The identified strong, negative correlation between residue centrality and mutation rate suggests that this computationally cheap approach to AAN construction provides biologically relevant results. To illustrate the ability of this static construction method to identify residues that fulfil a dynamical function the RGN has been used to identify hinge residues and, in combination with an estimated visiting time analysis, allosteric regulatory sites in the G-protein coupled receptor signalling pathway. The low-cost of the RGN analysis allows the results to be used as an "intuition pump", highlighting interesting residues in the protein structure for more labour intensive theoretical or experimental investigations. Alternatively, a potential application lies alongside protein design technologies, where incorporation of the RGN analysis into existing models can select structures with particular functions.

In a separate investigation (point b), the constrained dynamics that emerge from the geometric analysis have been employed to simulate an all-atom survey of the conformal motions undergone by the FMO complex. This analysis uncovers multiscale, hierarchal structural relationships that are suggested to facilitate coherent EET and aid interpretation of the coherences in the experimental data. It is speculated that that the correlations of structural elements within *each* realisation might act to stabilise the static, average Hamiltonian structure of each configuration against inevitable deformations, helping the ensemble to perform efficient EET by ensuring that most of the members have functional electronic structure. To investigate the impact of conformal motions on the Hamiltonian requires the site energies to be calculated.

Previous structure-based methods for extraction of site energies have often been limited by the need to approximate some or all of the PPC by classical point charges, and it is known that the simulation of "static" or ensemble disorder using classical molecular mechanics is computationally infeasible. Using PCA, the largest amplitude motions from the exhaustive geometric simulation of the full trimeric FMO complex is calculated. On a set of representative snapshots of this motion, we employ fully quantum mechanical simulations of pigment site energies on clusters of nearly 2000 atoms with the ONETEP density functional theory code. These are among the largest fully quantum mechanical excited state calculations ever performed on a biological complex, and this is made possible by the advanced linear-scaling capabilities of the ONETEP software. We find the maintenance of a homogeneous gap (303 $\pm$ 27 cm^{-1}) between optical excitation energies in the functionally important pigments 3 and 4 across different realisations of the conformal ensemble. Critically, the energy gap is maintained against a background of larger, non-monotonic variances in the individual site energies which, if uncorrelated, would result in extremely rapid inhomogeneous dephasing of inter-exciton coherences, faster even than the optical inhomogeneous dephasing times. Cross-correlation analysis reveals the source of this correlated disorder, which is achieved through highly correlated motions of key secondary structure elements of the protein that emerge from the highly constrained and rigidly hierarchical protein structure of the FMO complex. This result offers strong evidence for correlated-disorder theories that have attempted to explain the much weaker than expected effect of the substantial inhomogeneity observed in the ground state absorption of FMO on the longevity of inter–excited state coherences [68].

While significant, these results would greatly benefit from electronic structure calculations of the full FMO complex, providing insight into the absorption spectrum along the slow, conformal motions. It is hoped that this work will inspire useful strategies when designing the energetic structures of nano-scale light harvesting devices, including tailor-made antenna complexes. While the structure of a protein is notoriously difficult to predict, specific DNA architectures are now routinely designed and modified to incorporate pigments at specific positions [102]. In addition to pigments, DNA can interact with proteins. Indeed, a class of proteins have evolved that bind specific sequences [123]. DNA-pigment-protein complexes could use principals from the above thesis when designing the energetic structure of the embedded pigments. One can imagine a situation where the DNA behaves as a passive organisational scaffold, while the protein is employed introduce rigidity, site energy shifts, and spatial correlations that give rise to interpigment correlated energetic disorder. This design principle would preserve the energetic structure of the antenna system, but could also give rise to delocalised exciton states. Combining engineered quantum coherent transport with the ability to program multichromophoric geometries could unleash antenna systems

that operate beyond the classical regime, providing biologically inspired solutions to the problem of efficient and directed light harvesting.

References

[1] Abramavicius, D. and Mukamel, S. (2011). Exciton dynamics in chromophore aggregates with correlated environment fluctuations. *J Chem Phys*, 134(17):174504.

[2] Acharya, S. and Karnik, S. S. (1996). Modulation of GDP release from transducin by the conserved Glu134-Arg135 sequence in rhodopsin. *J Bio Chem*, 271(41):25406–25411.

[3] Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem Rev*, 106(5):1589–1615.

[4] Adolphs, J., Müh, F., Madjet, A., and Renger, T. (2008). Calculation of pigment transition energies in the FMO protein: from simplicity to complexity and back. *Photosynth Res*, 95(2-3):197–209.

[5] Adolphs, J. and Renger, T. (2006). How proteins trigger excitation energy transfer in the FMO complex of green sulfur bacteria. *Biophys J*, 91(8):2778–97.

[6] Ahuja, S., Hornak, V., Yan, E., Syrett, N., Goncalves, J., Hirshfeld, A., Ziliox, M., Sakmar, T., Sheves, M., Reeves, P., Smith, S., and Eilers, M. (2009). Helix movement is coupled to displacement of the second extracellular loop in rhodopsin activation. *Nat Struct Mol Biol*, 16(5):168–75.

[7] Altenbach, C., Cai, K., Khorana, H. G., and Hubbell, W. L. (1999). Structural features and light-dependent changes in the sequence 306-322 extending from helix VII to the palmitoylation sites in rhodopsin. *Biochemistry*, 38(25):7931–7937.

[8] Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. (1993a). *Proteins: Struct, Funct, Bioinf*, 17(4):412–425.

[9] Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. (1993b). Essential dynamics of proteins. *Prot: Struct, Funct, Bioinform*, 17(4):412–425.

[10] Amerongen, H. and Croce, R. (2013). Light harvesting in photosystem II. *Photosynth Res*, 116.

[11] Anna, J., Scholes, G., and van Grondelle, R. (2014). A little coherence in photosynthetic light harvesting. *J Biol Sci*, 64(1):14–25.

[12] Artz, J., Wernimont, A., Dunford, J., Schapira, M., Dong, A., Zhao, Y., Lew, J., Russell, R. G. G., Ebetino, F. H., Oppermann, U., and Hui, R. (2011). Molecular characterization of a novel geranylgeranyl pyrophosphate synthase from plasmodium parasites. *J Biol Chem*, 286(5):3315–3322.

[13] Austin, R., Frauenfelder, H., Chan, S., Schulz, C., Chan, W., Nienhaus, G., and Young, R. (2010). *The Physics of Proteins: An Introduction to Biological Physics and Molecular Biophysics*. Springer New York.

[14] 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.

[15] Bahar, I. and Rader, A. (2005). Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Bio*, 15(5):586 – 592.

[16] Barabási, A. (2016). *Network Science*. Cambridge University Press.

[17] Barrett, G. and Elmore, D. (1998). *Amino Acids and Peptides*. Cambridge University Press.

[18] Bartlett, A. and Radford, S. (2009). An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat Struct Mol Bio*, 16(6):582–588.

[19] Barzega, A., Moosavi-Movahedi, A., Pedersen, J., and Miroliaei, M. (2009). Comparative thermostability of mesophilic and thermophilic alcohol dehydrogenases: Stability-determining roles of proline residues and loop conformations. *Enzym Microb Technol*, 45(2):73 – 79.

[20] Beazley, D. (2009). *Python Essential Reference*. Addison-Wesley.

[21] Begall, S., Červený, J., Neef, J., Vojtěch, O., and Burda, H. (2008). Magnetic alignment in grazing and resting cattle and deer. *Proc Nat Acad Sci*, 105(36):13451–13455.

[22] Belfield, W. J., Cole, D. J., Martin, I. L., Payne, M. C., and Chau, P. L. (2014). Constrained geometric simulation of the nicotinic acetylcholine receptor. *J Mol Graph*, 52:1–10.

[23] Bhattacharyya, M., Bhat, C. R., and Vishveshwara, S. (2013). An automated approach to network features of protein structure ensembles. *Protein Sci*, 22(10):1399–1416.

[24] Bhattacharyya, M., Ghosh, A., Hansia, P., and Vishveshwara, S. (2010). Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. *Protein: Struct, Funct, and Bioinform*, 78(3):506–517.

[25] Blankenship, R. (2008). *The Basic Principles of Photosynthetic Energy Storage*, pages 1–10. Blackwell Science Ltd.

[26] Blankenship, R. (2014). *Molecular Mechanisms of Photosynthesis*. Wiley.

[27] Boehr, D. D., Nussinov, R., and Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol*, 5(11):789–796.

[28] Bohr, N. (2016). On the constitution of atoms and molecules. In *Niels Bohr, 1913-2013*, pages 13–33. Springer International Publishing.

[29] Bolte, P., Bleibaum, F., Einwich, A., Günther, A., Liedvogel, M., Heyers, D., Depping, A., Wühlbrand, L., Rabus, R., Janssen-Bienhold, U., and Mouritsen, H. (2016). Localisation of the putative magnetoreceptive protein cryptochrome 1b in the retinae of migratory birds and homing pigeons. *PLoS ONE*, 11(3):1–17.

[30] Boyer, R. W. and Hensley, P. (2015). A further review of'orch or'theory: The universe in consciousness. *NeuroQuantology*, 13(2).

[31] Branden, C. (1999). *Introduction to Protein Structure*. Taylor & Francis Group.

[32] Brinda, K. and Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophy J*, 89(6):4159 – 4170.

[33] Brixner, T., Stenger, J., Vaswani, H. M., Cho, M., Blankenship, R. E., and Fleming, G. R. (2005). Two-dimensional spectroscopy of electronic couplings in photosynthesis. *Nature*, 434(7033):625–628.

[34] Caruso, F., Chin, A. W., Datta, A., Huelga, S. F., and Plenio, M. B. (2009). Highly efficient energy excitation transfer in light-harvesting complexes: The fundamental role of noise-assisted transport. *J Chem Phys*, 131(10):105106.

[35] Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, R., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., and Kollman, P. A. (2009). *AMBER 11*. University of California, San Francisco.

[36] Ceccarelli, M., Procacci, P., and Marchi, M. (2003). An ab initio force field for the cofactors of bacterial photosynthesis. *J Comput Chem*, 24(2):129–142.

[37] Chachisvilis, M., Kühn, O., Pullerits, T., , and Sundström, V. (1997). Excitons in photosynthetic purple bacteria : wavelike motion or incoherent hopping? *J Phys Chem B*, 101(37):7275–7283.

[38] Chang, S., Jiao, X., Li, C., Gong, X., Chen, W., and Wang, C. (2008). Amino acid network and its scoring application in protein-protein docking. *Biophys Chem*, 134(3):111 – 118.

[39] Cheng, Y. and Silbey, R. (2006). Coherence in the B800 ring of purple bacteria LH2. *Phys Rev Lett*, 96(2):028103.

[40] Chin, A. W., Datta, A., Caruso, F., Huelga, S. F., and Plenio, M. B. (2010). Noise-assisted energy transfer in quantum networks and light-harvesting complexes. *New J Phys*, 12(6):065002.

[41] Choi, J., Laurent, A., Hilser, V., and Ostermeier, M. (2015). Design of protein switches based on an ensemble model of allostery. *Nat Commun*, 6.

[42] Christakis, N. and Fowler, J. (2011). *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives – How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do*. Brown.

[43] Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci*, 104(49):19428–19433.

[44] Cole, D. J., Chin, A. W., Hine, N. D. M., Haynes, P. D., and Payne, M. C. (2013). Toward ab initio optical spectroscopy of the Fenna–Matthews–Olson complex. *J Phys Chem Lett*, 4(24):4206–4212.

[45] Cole, D. J. and Hine, N. D. M. (2016). Applications of large-scale density functional theory in biology. *J Phys Condens Matter*, 28(39):393001.

[46] Cole, D. J. and Hine, N. D. M. (2016, in press). Applications of large-scale density functional theory in biology. *J Phys Condens Matter*.

[47] Collini, E., Wong, C., Wilk, K., Curmi, P., Brumer, P., and Scholes, G. (2010). Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature. *Nature*, 463(7281):644–647.

[48] Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., and Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci*, 23(1):47–55.

[49] Cosgrove, M. S., Naylor, C., Paludan, S., Adams, M. J., and Levy, H. R. (1998). On the mechanism of the reaction catalyzed by glucose 6-phosphate dehydrogenase. *Biochem*, 37(9):2759–2767.

[50] Crick, F. (1996). *Of Molecules and Men*. Great minds series. Prometheus Books.

[51] Croce, R. and Van Amerongen, H. (2014). Natural strategies for photosynthetic light harvesting. *Nat Chem Biol*, 10(7):492–501.

[52] Cushing, J. (1994). *Quantum Mechanics: Historical Contingency and the Copenhagen Hegemony*. University of Chicago Press.

[53] Dahiyat, B. I., Benjamin Gordon, D., and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci*, 6(6):1333–1337.

[54] David, C. and Jacobs, D. (2014). Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol Biol*, 1084:193–226.

[55] David, C. C. and Jacobs, D. J. (2011). Characterizing protein motions from structure. *J Mol Graph.*, 31:41–56.

[56] Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S., and Richardson, D. C. (2007). Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*, 35(suppl 2):W375–W383.

[57] del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Sci*, 15(9):2120–2128.

[58] Dimitrov, S. D. and Durrant, J. R. (2013). Materials design considerations for charge generation in organic solar cells. *Chem Mater*, 26(1):616–630.

[59] Dostál, J., Pšenčík, J., and Zigmantas, D. (2016). In situ mapping of the energy flow through the entire photosynthetic apparatus. *Nat Chem*, 8(7):1755–4330.

[60] Duck, I. and Sudarshan, E. (1998). *Pauli and the Spin-Statistics Theorem*. World Scientific.

[61] Dziedzic, J., Helal, H. H., Skylaris, C.-K., Mostofi, A. A., and Payne, M. C. (2011). Minimal parameter implicit solvent model for ab initio electronic-structure calculations. *EPL*, 95(4):43001.

[62] Echave, J. (2012). Why are the low-energy protein normal modes evolutionarily conserved? *Pure Appl Chem*, 84(9):1931–1937.

[63] Einati, H., Mishra, D., Friedman, N., Sheves, M., and Naaman, R. (2015). Light-controlled spin filtering in bacteriorhodopsin. *Nano Lett*, 15(2):1052–1056.

[64] Engel, G., Calhoun, T., Read, E., Ahn, T., Mancal, T., Cheng, Y., Blankenship, R., and Fleming, G. (2007). Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature*, 446(7137):782–786.

[65] Erskine, P., Fokas, A., Muriithi, C., Rehman, H., Yates, L., Bowyer, A., Findlow, I., Hagan, R., Miles, J., Wallace, B., Wells, A., Wood, P., and Cooper, J. (2015). X-ray, spectroscopic and normal-mode dynamics of calexcitin: structure-function studies of a neuronal calcium-signalling protein. *Acta Crystallogr Sect D*, 71(3):615–631.

[66] Fassioli, F., Nazir, A., and Olaya-Castro, A. (2010). Quantum state tuning of energy transfer in a correlated environment. *J Phys Chem Lett*, 1(14):2139–2143.

[67] Fermi, E. (1927). Un metodo statistico per la determinazione di alcune priorieta dell'atome. *Rend Accad Naz Lincei*, 6(602-607):32.

[68] Fidler, A. F., Harel, E., Long, P. D., and Engel, G. S. (2011). Two-dimensional spectroscopy can distinguish between decoherence and dephasing of zero-quantum coherences. *J Phys Chem A*, 116(1):282–289.

[69] Fidler, A. F., Singh, V. P., Long, P. D., Dahlberg, P. D., and Engel, G. S. (2014). Dynamic localization of electronic excitation in photosynthetic complexes revealed with chiral two-dimensional spectroscopy. *Nat Commun*, 5(3286).

[70] Flock, T., Ravarani, C., Sun, D., Venkatakrishan, A., Kayikci, M., Tate, C., Veprinstev, D., and Babu, M. (2015). Universal allosteric mechanism for Ga activation by GPCRs. *Nature*, 524(7564):173–179.

[71] Fogel, G., Corne, D., and Pan, Y. (2007). *Computational Intelligence in Bioinformatics*. IEEE Press Series on Computational Intelligence. Wiley.

[72] Fokas, A. S., Cole, D. J., and Chin, A. W. (2014). Constrained geometric dynamics of the Fenna–Matthews–Olson complex: the role of correlated motion in reducing uncertainty in excitation energy transfer. *Photosynth Res*, 122(3):275–292.

[73] Fowler, J. H. and Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 337:a2338.

[74] Fox, S. J., Pittock, C., Fox, T., Tautermann, C. S., Malcolm, N., and Skylaris, C.-K. (2011). Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules. *J Chem Phys*, 135(22):224107.

[75] Frank, J., editor (2012). *Molecular Machines in Biology*. Cambridge University Press.

[76] Franzosa, E. and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol and Evol*, 26(10):2387–2395.

[77] Frieden, B. and Gatenby, R. (2011). Information dynamics in living systems: Prokaryotes, eukaryotes, and cancer. *PLoS ONE*, 6(7):1–6.

[78] Friedrich, P. (2014). *Supramolecular Enzyme Organization: Quaternary Structure and Beyond*. Elsevier Science.

[79] Frith, C. D. and Frith, U. (2007). Social cognition in humans. *Current Biology*, 17(16):R724–R732.

[80] Fulle, S., Christ, N. A., Kestner, E., and Gohlke, H. (2010). HIV-1 TAR RNA spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J Chem Info Mod*, 50(8):1489–1501.

[81] Gaillard, T., Dejae, A., and Stote, R. H. (2009). Dynamics of $\beta 3$ integrin I-like and hybrid domains: Insight from simulations on the mechanism of transition between open and closed forms. *Prot: Struct, Funct, Bioinform*, 76(4):977–994.

[82] Gall, A., Ilioaia, C., Krüger, T. P., Novoderezhkin, V. I., Robert, B., and Van Grondelle, R. (2015). Conformational switching in a light-harvesting protein as followed by single-molecule spectroscopy. *Biophys J*, 108(11):2713–2720.

[83] Gaspar, M. E. and Csermely, P. (2012). Rigidity and flexibility of biological networks. *CoRR*, 1204.6389.

[84] Gélinas, S., Rao, A., Kumar, A., Smith, S. L., Chin, A. W., Clark, J., van der Poll, T. S., Bazan, G. C., and Friend, R. H. (2014). Ultrafast long-range charge separation in organic semiconductor photovoltaic diodes. *Science*, 343(6170):512–516.

[85] Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci*, 80(12):3696–3700.

[86] Goedecker, S. (1999). Linear scaling electronic structure methods. *Rev Mod Phys*, 71:1085–1123.

[87] Goodey, N. and Benkovic, S. (2008). Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol*, 4:474–482.

[88] Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696.

[89] Gray, H. B. and Winkler, J. R. (2004). Electron tunneling through proteins. *Q Rev Biophys*, 36(3):341–372.

[90] Greene, L. H. and Higman, V. A. (2003). Uncovering network systems within protein structures. *J Mol Biol*, 334(4):781 – 791.

[91] Grémiaux, A., Yokawa, K., Mancuso, S., and Baluška, F. (2014). Plant anesthesia supports similarities between animals and plants: Claude bernard's forgotten studies. *Plant Signal Behav*, 9:e27886.

[92] Guyoneaud, R., Borrego, C. M., Martínez-Planells, A., Buitenhuis, E. T., and Garcia-Gil, J. L. (2001). Light responses in the green sulfur bacterium *Prosthecochloris aestuarii*: changes in prosthecae length, ultrastructure, and antenna pigment composition. *Arch Microbiol*, 176(4):278–284.

[93] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA.

[94] Hameroff, S., Kaszniak, A., and Scott, A. (1994). Toward a scientific basis for consciousness.

[95] Hameroff, S. and Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Math. Comput. Simulation*, 40(3):453 – 480.

[96] Hameroff, S. and Penrose, R. (2014). Consciousness in the universe: A review of the 'orch or' theory. *Phys Life Rev*, 11(1):39 – 78.

[97] Hamming, R. (1980). The unreasonable effectiveness of mathematics. *Am Math Monthly*, 87(2):81–90.

[98] Harel, E. and Engel, G. S. (2012). Quantum coherence spectroscopy reveals complex dynamics in bacterial light-harvesting complex 2 (LH2). *Proc Natl Acad Sci*, 109(3):706–711.

[99] Haynes, P. and Payne, M. (1999). Corrected penalty-functional method for linear-scaling calculations within density-functional theory. *Physical Review B*, 59(19):12173.

[100] Haynes, P. D., Skylaris, C.-K., Mostofi, A. A., and Payne, M. C. (2006). Onetep: linear-scaling density-functional theory with local orbitals and plane waves. *physica status solidi (b)*, 243(11):2489–2499.

[101] Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*, 96(1):86–103.

[102] Hemmig, E. A., Creatore, C., Wünsch, B., Hecker, L., Mair, P., Parker, M. A., Emmott, S., Tinnefeld, P., Keyser, U. F., and Chin, A. W. (2016). Programming light-harvesting efficiency using dna origami. *Nano Letters*, 16(4):2369–2374.

[103] Hidalgo, C., Blumm, N., Barabási, A., and Christakis, N. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*, 5(4):1–11.

[104] Hildner, R., Brinks, D., Nieder, J., Cogdell, R., and Hulst, N. (2013). Quantum coherent energy transfer over varying pathways in single light-harvesting complexes. *Science*, 340(6139):1448–1451.

[105] Hine, N. D., Haynes, P. D., Mostofi, A. A., Skylaris, C.-K., and Payne, M. C. (2009). Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with onetep. *Comp Phys Communicat*, 180(7):1041–1053.

[106] Huang, P.-S., Boyken, S. E., and Baker, D. (2016). The coming of age of de novo protein design. *Nature*, 537(7620):320–327.

[107] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., et al. (2003). The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

[108] Huelga, S. and Plenio, M. (2013). Vibrations, quanta and biology. *Contemp Phys*, 54(4):181–207.

[109] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

[110] Jia, L., Yarlagadda, R., and Reed, C. C. (2015a). Structure based thermostability prediction models for protein single point mutations with machine learning tools. *PloS one*, 10(9):e0138022.

[111] Jia, X., Mei, Y., Zhang, J., and Mo, Y. (2015b). Hybrid QM/MM study of FMO complex with polarized protein-specific charge. *Sci Rep*, 5(17096):1–10.

[112] Jimenez-Roldan, J., Freedman, R., Römer, R., and Wells, S. (2012). Rapid simulation of protein motion: merging flexibility, rigidity and normal mode analyses. *Phys Biol*, 9(1):016008.

[113] Jimenez-Roldan, J. E., Wells, S. A., Freedman, R. B., and Roemer, R. A. (2011). Integration of FIRST, FRODA and NMM in a coarse grained method to study protein disulphide isomerase conformational change. *JPCS*, 286(1):012002.

[114] Jing, Y., Zheng, R., Li, H.-X., and Shi, Q. (2012). Theoretical study of the electronic–vibrational coupling in the qy states of the photosynthetic reaction center in purple bacteria. *J Phys Chem B*, 116(3):1164–1171.

[115] Jolley, C. C., Wells, S. A., Hespenheide, B. M., Thorpe, M. F., and Fromme, P. (2006). Docking of photosystem I subunit C using a constrained geometric simulation. *J Am Chem Soc*, 128:8803–8812.

[116] Jurinovich, S., Curutchet, C., and Mennucci, B. (2014). The Fenna-Matthews-Olson protein revisited: A fully polarizable (TD)DFT/MM description. *ChemPhysChem*, 15(15):3194–3204.

[117] Kavanagh, K. L., Dunford, J. E., Bunkoczi, G., Russell, R. G. G., and Oppermann, U. (2006). The crystal structure of human geranylgeranyl pyrophosphate synthase reveals a novel hexameric arrangement and inhibitory product binding. *J Biol Chem*, 281(31):22004–22012.

[118] Kemeny, J. and Snell, J. (1983). *Finite Markov Chains*. Undergraduate Texts in Mathematics. Springer New York.

[119] Kessel, A. and Ben-Tal, N. (2010). *Introduction to Proteins: Structure, Function, and Motion*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press.

[120] Kim, T.-Y., Schlieter, T., Haase, S., and Alexiev, U. (2012). Activation and molecular recognition of the GPCR rhodopsin – insights from time-resolved fluorescence depolarisation and single molecule experiments. *Eur J Cell Biol*, 91(4):300 – 310. Structure and Function of Membrane Receptors.

[121] Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

[122] Klein, D., Radestock, S., and Gohlke, H. (2011). *Analyzing Protein Rigidity for Understanding and Improving Thermal Adaptation*, pages 47–66. CRC Press.

[123] Klug, A. and Rhodes, D. (1987). Zinc fingers a novel protein motif for nucleic acid recognition. *Trends in Biochemical Sciences*, 12:464 – 469.

[124] Knoll, A., Canfield, P., and Konhauser, K. (2012). *Fundamentals of Geobiology*. Wiley.

[125] Knox, R. and Spring, B. Q. (2003). Dipole strengths in the chlorophylls. *Photochem Photobiol*, 77(5):497–501.

[126] Kohn, W. (1996). Density functional and density matrix method scaling linearly with the number of atoms. *Phys Rev Lett*, 76:3168–3171.

[127] Kohn, W. and Sham, L. J. (1965a). Self-consistent equations including exchange and correlation effects. *Phys Rev*, 140:A1133–A1138.

[128] Kohn, W. and Sham, L. J. (1965b). Self-consistent equations including exchange and correlation effects. *Phys Rev*, 140:A1133–A1138.

[129] Kortemme, T. and Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proc Natl Acad Sci*, 99(22):14116–14121.

[130] Kozuska, J. L., Paulsen, I. M., Belfild, W. J., Martin, I. L., Cole, D. J., Holt, A., and Dunn, S. M. J. (2014). Impact of intracellular domain flexibility upon properties of activated human 5-HT3 receptors. *Br J Pharmacol*, 171(7):1617–1628.

[131] Kreisbeck, C. and Kramer, T. (2012). Long-lived electronic coherence in dissipative exciton dynamics of light-harvesting complexes. *J Phys Chem Lett*, 3(19):2828–2833.

[132] Krischer, M. K. and Sevecke, K. (2008). Early traumatization and psychopathy in female and male juvenile offenders. *Int. J. Law Psychiatry*, 31(3):253–262.

[133] Krüger, D. M., Rathi, P. C., Pfleger, C., and Gohlke, H. (2013). Cna web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucleic Acids Res*, 41(W1):W340–W348.

[134] Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368.

[135] Laman, G. (1970). On graphs and rigidity of plane skeletal structures. *J Eng Math*, 4(4):331–340.

[136] Lambert, N., Chen, Y., Cheng, Y., Li, C., Chen, G., and Nori, F. (2012). Quantum biology. *Nat Phys*, 9(1):10–18.

[137] Lambert, N., Chen, Y.-N., Cheng, Y.-C., Li, C.-M., Chen, G.-Y., and Nori, F. (2013). Quantum biology. *Nat Phys*, 9(1):10–18.

[138] Lee, H., Cheng, Y.-C., and Fleming, G. R. (2007). Coherence dynamics in photosynthesis: protein protection of excitonic coherence. *Science*, 316(5830):1462–1465.

[139] Lever, G., Cole, D., Hine, N., Haynes, P., and Payne, M. (2013). Electrostatic considerations affecting the calculated HOMO–LUMO gap in protein molecules. *J Phys Condens Matter*, 25(15):152101.

[140] LEVY, H. (1982). Stochastic dominance rules for truncated normal distributions: A note. *J Finance*, 37(5):1299–1303.

[141] Li, H., Wells, S. A., Jimenez-Roldan, J. E., Romer, R. A., Zhao, Y., Sadler, P. J., and O'Connor, P. B. (2012). Protein flexibility is key to cisplatin crosslinking in calmodulin. *Protein Sci*, 21:1269–1279.

[142] Li, Z., Yang, Y., Zhan, J., Dai, L., and Zhou, Y. (2013). Energy functions in de novo protein design: Current challenges and future prospects. *Ann Rev Biophys*, 42:101146.

[143] Lichtenthaler, F. (1995). 100 Years "Schlüssel-Schloss-Prinzip": What Made Emil Fischer Use this Analogy? *Angew Chemie*, 33(23-24):2364–2374.

[144] Liu, Y. and Bahar, I. (2012a). Sequence evolution correlates with structural dynamics. *Molecular Biology and Evolution*, 29(9):2253–2263.

[145] Liu, Y. and Bahar, I. (2012b). Sequence evolution correlates with structural dynamics. *Mol Biol Evol*, 29(9):2253–2263.

[146] Löwdin, P.-O. (1966). Quantum genetics and the aperiodic solid: Some aspects on the biological problems of heredity, mutations, aging, and tumors in view of the quantum theory of the DNA molecule. volume 2 of *Advances in Quantum Chemistry*, pages 213 – 360. Academic Press.

[147] Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373 – 380.

[148] Madjet, M., Abdurahman, A., and Renger, T. (2006). Intermolecular coulomb couplings from ab initio electrostatic potentials application to optical transitions of strongly coupled pigments in photosynthetic antennae and reaction centers. *J Phys Chem B*, 110(34):17268–17281.

[149] Marsh, J. and Teichmann, S. (2014a). Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, 36(2):209–218.

[150] Marsh, J. and Teichmann, S. (2014b). Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, 36(2):209–218.

[151] McInerney, J. O. (2006). The causes of protein evolutionary rate variation. *TREE*, 21(5):230–232.

[152] McKaughan, D. (2005). The influence of niels bohr on max delbrück: Revisiting the hopes inspired by "light and life". *Isis*, 96(4):507–529.

[153] Metz, A., Pfleger, C., Kopitz, H., Pfeiffer-Marek, S., Barringhaus, K. H., and Gohlke, H. (2011). Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. *J Chem Inf Model*, 52:120–133.

[154] Mohseni, M., Rebentrost, P., Lloyd, S., and Aspuru-Guzik, A. (2008). Environment-assisted quantum walks in photosynthetic energy transfer. *J Chem Phys*, 129(17):174106.

[155] Monod, J., Changeux, J., and Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology*, 6(4):306 – 329.

[156] Müh, F., Madjet, M., Adolphs, J., Abdurahman, A., Rabenstein, B., Ishikita, H., Knapp, E., and Renger, T. (2007). α–helices direct excitation energy flow in the Fenna–Matthews–Olson protein. *Proc Natl Acad Sci*, 104(43):16862–16867.

[157] Nagel, Z. D., , and Klinman, J. P. (2006). Tunneling and dynamics in enzymatic hydride transfer. *Chem Rev*, 106(8):3095–3118.

[158] Narayana, N., Cox, S., Xuong, N., Ten Eyck, L., and Taylor, S. (1997). A binary complex of the catalytic subunit of cAMP-dependent protein kinase and adenosine further defines conformational flexibility. *Structure*, 5(7):921 – 935.

[159] Nazir, A. (2009). Correlation-dependent coherent to incoherent transitions in resonant energy transfer dynamics. *Phys Rev Lett*, 103:146404.

[160] Nevin Gerek, Z., Kumar, S., and Banu Ozkan, S. (2013). Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol Appl*, 6(3):423–433.

[161] Nielsen, R. (2006). *Statistical Methods in Molecular Evolution*. Statistics for Biology and Health. Springer New York.

[162] Oberai, A., Ihm, Y., Kim, S., and Bowie, J. U. (2006). A limited universe of membrane protein families and folds. *Protein Sci*, 15(7):1723–1734.

[163] Olah, G. A., Mitchell, R. D., Sosnick, T. R., Walsh, D. A., and Trewhella, J. (1993). Solution structure of the cAMP-dependent protein kinase catalytic subunit and its contraction upon binding the protein kinase inhibitor peptide. *Biochemistry*, 32(14):3649–3657.

[164] Olaya-Castro, A. and Fassioli, F. (2011). Characterizing quantum-sharing of electronic excitation in molecular aggregates. *Procedia Chem*, 3(1):176 – 184.

[165] Olbrich, C., Strümpfer, J., Schulten, K., and Kleinekathöfer, U. (2011a). Quest for spatially correlated fluctuations in the FMO light-harvesting complex. *J Phys Chem B*, 115(4):758–764.

[166] Olbrich, C., Strümpfer, J., Schulten, K., and Kleinekathöfer, U. (2011b). Theory and simulation of the environmental effects on FMO electronic transitions. *J Phys Chem Lett*, 2(14):1771–1776.

[167] Olbrich, C., Strümpfer, J., Schulten, K., and Kleinekathöfer, U. (2011c). Theory and simulation of the environmental effects on fmo electronic transitions. *J Phys Chem Lett*, 2(14):1771–1776.

[168] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). Cath – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093 – 1109.

[169] Panitchayangkoon, G., Hayes, D., Fransted, K., Caram, J., Harel, E., Wen, J., Blankenship, R., and Engel, G. (2010). Long-lived quantum coherence in photosynthetic complexes at physiological temperature. *Proc Natl Acad Sci*, 107(29):12766–12770.

[170] Park, K. and Kim, D. (2011). Modeling allosteric signal propagation using protein structure networks. *BMC Bioinformatics*, 12(S-1):S23.

[171] Parr, R. and Weitao, Y. (1994). *Density-Functional Theory of Atoms and Molecules*. International Series of Monographs on Chemistry. Oxford University Press.

[172] Patthy, L. (2009). *Protein Evolution*. Wiley.

[173] Pedersen, M. Ø., Linnanto, J., Frigaard, N.-U., Nielsen, N. C., and Miller, M. (2010). A model of the protein–pigment baseplate complex in chlorosomes of photosynthetic green bacteria. *Photosynth Res*, 104(2):233–243.

[174] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *J Machine Learn Res*, 12:2825–2830.

[175] Perdew, J. P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868.

[176] Pinker, S. (2011). *The Better Angels of Our Nature: The Decline of Violence In History And Its Causes*. Penguin Books Limited.

[177] Plenio, M. B. and Huelga, S. F. (2008). Dephasing-assisted transport: quantum networks and biomolecules. *New J Phys*, 10(11):113019.

[178] Prodan, E. and Kohn, W. (2005). Nearsightedness of electronic matter. *Proc Natl Acad Sci*, 102(33):11635–11638.

[179] Radestock, S. and Gohlke, H. (2008). Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Eng Life Sci*, 8(5):507–522.

[180] Rankin, A M; Philip, P. J. (1963). An epidemic of laughing in the bukoba district of tanganyika. *Cent Afr J Med*, 9(5):167–170.

[181] Rebentrost, P., Mohseni, M., and Aspuru-Guzik, A. (2009). Role of quantum coherence and environmental fluctuations in chromophoric energy transport. *J Phys Chem B*, 113(29):9942–9947.

[182] Renger, G. (2008). *Primary Processes of Photosynthesis: Principles and Apparatus*. Number 1. RSC Publishing.

[183] Renger, T., Klinger, A., Steinecker, F., Schmidt am Busch, M., Numata, J., and Müh, F. (2012). Normal mode analysis of the spectral density of the Fenna–Matthews–Olson light-harvesting protein: How the protein dissipates the excess energy of excitons. *J Phys Chem B*, 116(50):14565–14580.

[184] Renger, T. and Müh, F. (2013). Understanding photosynthetic light-harvesting: A bottom up the oretical approach. *Phys Chem Chem Phys*, 15(10):3348–3371.

[185] Rey, M., Chin, A. W., Huelga, S. F., and Plenio, M. B. (2013). Exploiting structured environments for efficient energy transfer: The phonon antenna mechanism. *J Phys Chem Lett*, 4(6):903–907.

[186] Roe, D. R. and Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theor Comput*, 9(7):3084–3095.

[187] Romero, E., Augulis, R., Novoderezhkin, V. I., Ferretti, M., Thieme, J., Zigmantas, D., and Van Grondelle, R. (2014). Quantum coherence in photosynthesis for efficient solar-energy conversion. *Nat Phys*, 10(9):676–682.

[188] Romero, E., van Stokkum, I. H., Novoderezhkin, V. I., Dekker, J. P., and van Grondelle, R. (2010). Two different charge separation pathways in photosystem II. *Biochem*, 49(20):4300–4307.

[189] Rosenberg, A. (2008). *Darwinian Reductionism: Or, How to Stop Worrying and Love Molecular Biology*. University of Chicago Press.

[190] Ruban, A. V., Berera, R., Ilioaia, C., Van Stokkum, I. H., Kennis, J. T., Pascal, A. A., Van Amerongen, H., Robert, B., Horton, P., and Van Grondelle, R. (2007). Identification of a mechanism of photoprotective energy dissipation in higher plants. *Nature*, 450(7169):575–578.

[191] Ruiz-Serrano, Á., Hine, N. D., and Skylaris, C.-K. (2012). Pulay forces from localized orbitals optimized in situ using a psinc basis set. *J Chem Phys*, 136(23):234101.

[192] Safo, M. K. and Abraham, D. J. (2003). *X-ray Crystallography of Hemoglobins*, pages 1–19. Humana Press, Totowa, NJ.

[193] Sahu, S., Ghosh, S., Fujita, D., and Bandyopadhyay, A. (2014). Live visualizations of single isolated tubulin protein self-assembly via tunneling current: effect of electromagnetic pumping during spontaneous growth of microtubule. *Scientific Rep*, 4:7303.

[194] Sakmar, T., Menon, W., Marin, E., and Awad, E. (2002). Rhodopsin: Insights from recent structural studies. *Annu Rev Biophys Biomol Struct*, 31(1):443–484.

[195] Sarovar, M., Cheng, Y.-C., and Whaley, K. B. (2011). Environmental correlation effects on excitation energy transfer in photosynthetic light harvesting. *Phys. Rev E*, 83:011906.

[196] Sathyapriya, R., Vijayabaskar, M. S., and Vishveshwara, S. (2008). Insights into protein–dna interactions through structure network analysis. *PLoS Comput Biol*, 4(9):1–15.

[197] Savikhin, S., Buck, D. R., and Struve, W. S. (1997). Oscillating anisotropies in a bacteriochlorophyll protein Evidence for quantum beating between exciton levels. *Chem Phys*, 223(2):303 – 312.

[198] Schaper, S. and Louis, A. (2014). The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. *PLoS One*, 9(2).

[199] Schlodder, E., Renger, T., Raszewski, G., Coleman, W. J., Nixon, P. J., Cohen, R. O., and Diner, B. A. (2008). Site-directed mutations at d1-thr179 of photosystem ii in synechocystis sp pcc 6803 modify the spectroscopic properties of the accessory chlorophyll in the d1-branch of the reaction center†. *Biochemistry*, 47(10):3143–3154.

[200] Schmidt am Busch, M., Müh, F., El-Amine Madjet, M., and Renger, T. (2010). The eighth bacteriochlorophyll completes the excitation energy funnel in the fmo protein. *J Phys Chem Lett*, 2(2):93–98.

[201] Scholes, G. D., Fleming, G. R., Olaya-Castro, A., and van Grondelle, R. (2011). Lessons from nature about solar light harvesting. *Nat Chem*, 3(10):763–774.

[202] Schrödinger, E. (1926). An undulatory theory of the mechanics of atoms and molecules. *Phys Rev*, 28:1049–1070.

[203] Schrödinger, E. (1944). *What is Life ?: The Physical Aspect of the Living Cell & Mind and Matter*. Cambridge University Press.

[204] Seber, G. and Lee, A. (2012). *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley.

[205] Shi, L., Liapakis, G., Xu, R., Guarnieri, F., Ballesteros, J. A., and Javitch, J. A. (2002). $\beta 2$ adrenergic receptor activation: Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. *J Biol Chem*, 277(43):40989–40996.

[206] Shim, S., Rebentrost, P., Valleau, S., and Aspuru-Guzik, A. (2012). Atomistic study of the long-lived quantum coherences in the Fenna-Matthews-Olson complex. *Biophys J*, 102(3):649–660.

[207] Shkuropatov, A. Y., Khatypov, R. A., Volshchukova, T. S., Shkuropatova, V. A., Owens, T. G., and Shuvalov, V. A. (1997). Spectral and photochemical properties of borohydride-treated d1-d2-cytochrome b-559 complex of photosystem ii. *FEBS letters*, 420(2-3):171–174.

[208] Sistla, R. K., Brinda, K. V., and Vishveshwara, S. (2005). Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins*, 59(3):616–626.

[209] Skochdopole, N. and Mazziotti, D. A. (2011). Functional subsystems and quantum redundancy in photosynthetic light harvesting. *J Phys Chem Lett*, 2(23):2989–2993.

[210] Skylaris, C.-K., Haynes, P., Mostofi, A., and Payne, M. (2005a). Introducing ONETEP: linear-scaling density-functional simulations on parallel computers. *J Chem Phys*, (122):084119.

[211] Skylaris, C.-K., Haynes, P. D., Mostofi, A. A., and Payne, M. C. (2005b). Using onetep for accurate and efficient on density functional calculations. *J Phys Condens Matter*, 17(37):5757.

[212] Skylaris, C.-K., Haynes, P. D., Mostofi, A. A., and Payne, M. C. (2006). Implementation of linear-scaling plane wave density functional theory on parallel computers. *physica status solidi (b)*, 243(5):973–988.

[213] Skylaris, C.-K., Mostofi, A. A., Haynes, P. D., Diéguez, O., and Payne, M. C. (2002). Nonorthogonal generalized wannier function pseudopotential plane-wave method. *Phys Rev B*, 66:035119.

[214] Sljoka, A. (2006). *Counting for Rigidity, Flexibility and Extensions Via the Pebble Game Algorithm*. Canadian theses. York University (Canada).

[215] Stein, I. H., Steinhauer, C., and Tinnefeld, P. (2011). Single-molecule four-color FRET visualizes energy-transfer paths on DNA origami. *J Am Chem Soc*, 133(12):4193–4195.

[216] Stent, G. (1989). Light and life: Niels bohr's legacy to contemporary biology. *Genome*, 31(1):11–15.

[217] Still, S., Sivak, D., Bell, A., and Crooks, G. (2012). Thermodynamics of prediction. *Phys Rev Lett*, 109:120604.

[218] Strümpfer, J. and Schulten, K. (2011). The effect of correlated bath fluctuations on exciton transfer. *J Chem Phys*, 9(134):095102.

[219] Stumpf, M., Thorne, T., Silva, E., Stewart, R., Jun, H., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Nat Acad Sci*, 105(19):6959–6964.

[220] Suhre, K. and Sanejouand, Y.-H. (2004). Elnemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res*, 32(2):W610–W614.

[221] Sun, M., Rose, M. B., Ananthanarayanan, S. K., Jacobs, D. J., and Yengo, C. M. (2008). Characterization of the pre-force-generation state in the actomyosin cross-bridge cycle. *Proc Natl Acad Sci*, 105:8631–8636.

[222] Tay, T.-S. (1984). Rigidity of multi-graphs i linking rigid bodies in n-space. *J Combinat Thery B*, 36(1):95–112.

[223] Teplyakov, A., Sebastiao, P., Obmolova, G., Perrakis, A., Brush, G., Bessman, M., and Wilson, K. (1996). Crystal structure of bacteriophage T4 deoxynucleotide kinase with its substrates dGMP and ATP. *EMBO JD*, 15(14):3487–3497.

[224] The C elegans Sequencing Consortium (1998). Genome sequence of the nematode C elegans: A platform for investigating biology. *Science*, 282(5396):2012–2018.

[225] Thomas, L. H. (2008). The calculation of atomic fields. *Math Proc Cambridge Phil Soc*, 23(5):542–548.

[226] Thyrhaug, E., Žídek, K., Dostál, J., Bína, D., and Zigmantas, D. (2016). Exciton structure and energy transfer in the fenna–matthews–olson complex. *J Phys Chem Lett*, 7(9):1653–1660.

[227] Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett*, 77:1905–1908.

[228] Tronrud, D. E., Schmid, M. F., and Matthews, B. W. (1986). Structure and x-ray amino acid sequence of a bacteriochlorophyll a protein from prosthecochloris aestuarii refined at 19 å resolution. *J Mol Bio*, 188(3):443 – 454.

[229] Turin, L., Skoulakis, E. M. C., and Horsfield, A. P. (2014). Electron spin changes during general anesthesia in drosophila. *Proc Nat Acad Sci*, 111(34):E3524–E3533.

[230] van Amerongen, H., Valkūnas, L., and van Grondelle, R. (2000). *Photosynthetic Excitons*. World Scientific.

[231] Velazquez-Muriel, J., Rueda, M., Cuesta, I., Pascual-Montano, A., Orozco, M., and Carazo, J.-M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Structural Biology*, 9(1):6.

[232] Venter, J. and et al (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[233] Verma, D., Jacobs, D., and Livesay, D. (2012). Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged. *PLoS Comp Biol*, 8(3).

[234] Wells, S., Jimenez-Roldan, J., and Römer, R. (2009). Comparative analysis of rigidity across protein families. *Phys Biol*, 6(4):046005.

[235] Wells, S., Menor, S., Hespenheide, B., and Thorpe, M. (2005). Constrained geometric simulation of diffusive motion in proteins. *Phys Bio*, 2(4).

[236] Wells, S. A. (2013). Geometric simulation of flexible motion in proteins. In Livesay, D. R., editor, *Protein Dynamics Vol II*, volume 1084 of *Methods in Molecular Biology*, pages 173–192. Humana Press, New York.

[237] Wells, S. A., Crennell, S. J., and Danson, M. J. (2014). Structures of mesophilic and extremophilic citrate synthases reveal rigidity and flexibility for function. *Proteins: Struct, Funct, Bioinf*, 82(10):2657–2670.

[238] Wells, S. A., van der Kamp, M. W., McGeagh, J. D., and Mulholland, A. J. (2015). Structure and function in homodimeric enzymes: Simulations of cooperative and independent functional motions. *PLOS ONE*, 10(8):1–21.

[239] Wen, J., Zhang, H., Gross, M., and Blankenship, R. (2009). Membrane orientation of the FMO antenna protein from chlorobaculum tepidum as determined by mass spectrometry-based footprinting. *Proc Natl Acad Sci*, 106(15):6134–9.

[240] West, D. (2001). *Introduction to Graph Theory*. Featured Titles for Graph Theory Series. Prentice Hall.

[241] Wijma, H. J., Floor, R., and Janssen, D. B. (2013). Structure-and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr Op Struct Biol*, 23(4):588–594.

[242] Williams, R. (2016). *Spectroscopy: New Uses and Implications*. Apple Academic Press.

[243] Wiltschko, R., Stapput, K., Thalau, P., and Wiltschko, W. (2009). Directional orientation of birds by the magnetic field under different light conditions. *J R Soc Interface*.

[244] Wlodarczyk, L., Snellenburg, J., Ihalainen, J., vanGrondelle, R., vanStokkum, I., and Dekker, J. (2015). Functional rearrangement of the light-harvesting antenna upon state transitions in a green alga. *Biophys J*, 108(2):261 – 271.

[245] Wu, J., Liu, F., Shen, Y., Cao, J., and Silbey, R. (2010). Efficient energy transfer in light-harvesting systems, I: optimal temperature, reorganization energy and spatial–temporal correlations. *New J Phys*, 12(10):105012.

[246] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., and Shen, B. (2014a). The construction of an amino acid network for understanding protein structure and function. *Amino acids*, 46:1419–1439.

[247] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., and Shen, B. (2014b). The construction of an amino acid network for understanding protein structure and function. *Amino acids*, 46(6):1419–1439.

[248] Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–6.

[249] Zhang, L., Silva, D., Zhang, H., Yue, A., Yan, Y., and Huang, X. (2014). Dynamic protein conformations preferentially drive energy transfer along the active chain of the photosystem II reaction centre. *Naure Commun*, 5(3).

[250] Zheng, J., Knighton, D., Xuong, NH adn Taylor, S., Sowadski, J., and Teneyck, L. (1993). Crystal structures of the myristylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations. *Protein Sci*, 2(10):1559–1573.

[251] Zhou, J., Yan, W., Hu, G., and Shen, B. (2014). Amino acid network for the discrimination of native protein structures from decoys. *Curr Protein Pept Sci*, 15(6):522–528.

[252] Zhou, J., Yan, W., Hu, G., and Shen, B. (2016). Amino acid network for prediction of catalytic residues in enzymes: a comparison survey. *Current Protein and Peptide Science*, 17(1):41–51.

[253] Zimmer, K. and Delbrück, M. (1935). Über die natur der genmutation und der genstruktur. *Nachr d Ges d Wiss, Göttingen, Neue Folge, Bd*, 1.