Investigation of Protein-Protein Interactions: Multibody Docking, Association/Dissociation Kinetics and Macromolecular Crowding

Xiaofan F. Li

September 2010

Biomolecular Modelling Laboratory,
Cancer Research UK London Research Institute
and
Department of Biochemistry and Molecular Biology,
University College London

In partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Computational Biophysics
at University College London.

I, Xiaofan F. Li, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Protein-protein interactions are central to understanding how cells carry out their wide array of functions and metabolic procedures. Conventional studies on specific protein interactions focus either on details of one-to-one binding interfaces, or on large networks that require *a priori* knowledge of binding strengths. Moreover, specific protein interactions, occurring within a crowded macromolecular environment, which is precisely the case for interactions in a real cell, are often under-investigated.

A macromolecular simulation package, called BioSimz, has been developed to perform Langevin dynamics simulations on multiple protein-protein interactions at atomic resolution, aimed at bridging the gaps between structural, kinetic and crowding studies on protein-protein interactions. Simulations on twenty-seven experimentally determined protein-protein interactions, indicated that the use of contact frequency information of proteins forming specific encounters can guide docking algorithms towards the most likely binding regions. Further evidence from eleven benchmarked protein interactions showed that the association rate constant of a complex, k_{on}, can be estimated, with good agreement to experimental values, based on the retention time of its specific encounter. Performing these simulations with ten types of environmental protein crowders, it suggests, from the change of k_{on}, that macromolecular crowding improves the association kinetics of slower-binding proteins, while it damps the association kinetics of fast, electrostatics-driven protein-protein interactions.

It is hypothesised, based on evidence from docking, kinetics and crowding, that the dynamics of specific protein-protein encounters is vitally important in determining their association affinity. There are multiple factors by which encounter dynamics, and subsequently the k_{on}, can be influenced, such as anchor residues, long-range forces, and environmental steering *via* crowders' electrostatics and/or volume exclusion. The capacity of emulating these conditions on a common platform not only provides a holistic view of interacting dynamics, but also offers the possibility of evaluating and engineering protein-protein interactions from aspects that have never been opened before.

Acknowledgements

This thesis will not be possible without the generous support and help from my colleagues, family and friends. I shall firstly extend my greatest gratitude to my supervisor, Dr Paul Bates, for his enormous support during the entire four years of my PhD. It is rare for a PhD supervisor to fully engage in coaching graduate students as devoted as Paul has done; the constant, sometimes heated, exchange of ideas between Paul and me has been going on for almost every working day out of the 1461 days during which I have been his student. I have pleasantly enjoyed these communications, and furthermore, gracefully appreciated his openness and actual support of my persistent attempt of working on a large problem. Throughout the PhD I have enjoyed as much academic freedom as one possibly can; this is exactly where I shall also devote my thankfulness to my supervisor for, which again has become increasingly rare, nowadays, in an often very competitive academic environment.

The second person I feel extremely grateful to have worked with is Mr Iain Moal, a fellow PhD student and colleague of mine. It was not until I collaborated with Iain on macromolecular docking that the potential of my software, BioSimz, was fully revealed and further extended. I truly believe that this collaboration has fructified a multiplication, rather than a mere summation, of our individual capacities. I wish him the best of luck in completing his PhD in 2011.

I need to thank all other members of the Biomolecular Modelling Lab, whom I have been spending part or the whole of my four years with: Dr Tammy Cheng who always gives insightful comments on my manuscripts, Dr Raphael Chaleil who managed the computing farm, and PhD student Melda Tozluoglu, as well as Dr Yanlan Mao and Dr Alexander Tournier. I should also extend my thanks to former members Dr Marc Offman and Dr

Marcin Krol, as they once helped me a lot on better understanding protein structures and interactions.

I owe a great deal of gratitude to Cancer Research UK and its London Research Institute, for their generous funding for me and excellent support for my project during the last four years. Frankly speaking, it is the largest help from an organisation I have ever received, and it has changed the path of my life. Therefore, I should give my sincere gratefulness to all people who donated to and supported this great charity, Cancer Research UK – together we can beat cancer.

Finally, time to pay a tribute to my family. I cannot express more in thankfulness and love for my wife, Qiujun. In the past four years, she has been an intelligent friend, a productive co-worker, a cheerful girlfriend and now, a loving wife of me. Together we worked through all the difficulties from life to science, forming a true companionship that I have not hesitated for a single moment. I shall also give my sincere thanks to all my other relatives for their continuous support; however, the most respect and tribute should be addressed to my mother, the woman who raised me up, guided me on what to do and whom to be – none of my achievements today would have been possible without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Life

1.1.1 What is life?

Life, a distinctive feature on planet Earth, may have existed for 3.7 billion years (Maher and Stevenson, 1988). To date, some 1.75 million species of life have been recorded (IUCN, 2010) . Yet the true bio-diversity may extend beyond recognition: the number of insect species alone has been estimated to be between 4 and 6 million (Novotny et al., 2002). Each of these "living creatures" adopts vastly different body shapes, self mobility and living habits; however, the fundamental principles that make them "alive" remain largely the same. In particular, the concept of life defines a process, rather than a factual substance, that strikes the balance between the mutually exclusive property pairs listed below:

Diffusion and organisation: any form of life, no matter large or small, must contain enclosed space separating the "inside" from "outside". In many cases, the diffusivity inside an organism describes the permittance of the internal exchange or transfer of its living content. On the other hand, living organisms have developed highly organised internal structures at molecular, sub-cellular, tissue, organ and system levels, influencing the stochastic flow of substance and information within their enclosure.

Positive and negative feedback loops: feedback controls, i.e., using the outcome of some process to influence the process itself, are widely used by organisms to regulate their biological activities. Positive feedback loops act as amplifiers to an otherwise slow process; negative feedback loops prevent wasting of energy and self-intoxication by throttling the progression of the process by using the product(s).

Excitability and adaptability: both account for the ability to respond to changes in an organism's surrounding environment. Excitation is the rapid physical or chemical changes made by an organism through the use of the stored energy in the same individual. Adaptability is an organism's answer to a persistent, long-term environmental pressure; the response modes, often in forms of bodily and physiological alteration, are gradually expressed through generations of offspring.

Catabolism and anabolism: together they constitute metabolism, which assigns the "dynamic" properties to life. Catabolism is the process of "consuming" stored energies and, as a result, releasing them back to the environment, while anabolism is the reverse course, absorbing energy from the environment to accomplish activities as small as chemical synthesis and as large as body growth.

Programmed growth and death: in higher organisms, cell growth and death are both under strict regulation as a whole body. Uncontrolled growth leads to tumourisation of cells, while with improper, or the lack of, programmed cell deaths can lead to failures in tissue and organ development.

Reproduction and mutation: the ability of an organism to replicate itself at some stage is one of the key differences between life and non-life. Moreover, the careful balance on the extent of replication accuracy plays a vital role in supplying the necessary stochasticity that powers the evolutionary selection process.

From a thermodynamic point of view (Schrodinger, 1944), life is about progressing into and maintaining an ordered, low entropy molecular system through the intake of free energy available from the environment, such as light, chemical compounds or other organisms. While the origin of the

apparent spontaneity of the entropy-decreasing process is still a debated subject, it is generally known that, in living organisms, many biological functions and procedures have been shown to be energetically favourable to proceed, such as those in catabolism. In these cases, the Gibbs free energies, ΔG, representing the amount of work obtainable from the functioning thermodynamic system, are negative. This can be suitably described as a downhill process where the potential energy is transformed into kinetic energies. For processes that go "uphill" on the energetic landscape, life has adapted various forms of supplying the required energy from energy-rich small molecules such as adenosine triphosphates (ATPs) and guanine triphosphates (GTPs), through to a flow of spontaneous, energetically favourable processes involving enzymes, usually kinases.

1.1.2 Biochemistry

In the vast majority of its possible forms, life is organised into units called cells. A cell is a membrane-enclosed dynamic system that is self-contained for most, if not all, biological processes. Unless differentiated for a special purpose, a cell usually adopts a spherical or ellipsoidal shape. Two major cell types are commonly seen in different forms of life: those of eukaryotes and prokaryotes. A eukaryotic cell encloses various types of highly-organised sub-cellular structures called organelles, such as the endoplasmic reticulum (ER) system and the Golgi apparatus. Prokaryotic cells differ from eukaryotes in that the former do not contain a nucleus to enclose their genetic material; moreover, most prokaryotes cells are unicellular, with no ability to aggregate and differentiate to form higher level living entities. A comparative schematic illustration of the internal cellular structures from two eukaryotic cell (animal and plant) and a prokaryotic cell (bacterium) is shown in Figure 1.1.

Unlike many other forms of matter, a cell is mainly an aqueous solution of large macromolecules bounded by a membrane. The most abundant molecules, in a typical cell, are proteins, deoxy- and ribonucleic acids (DNAs and RNAs), polysaccharides and phospholipids.

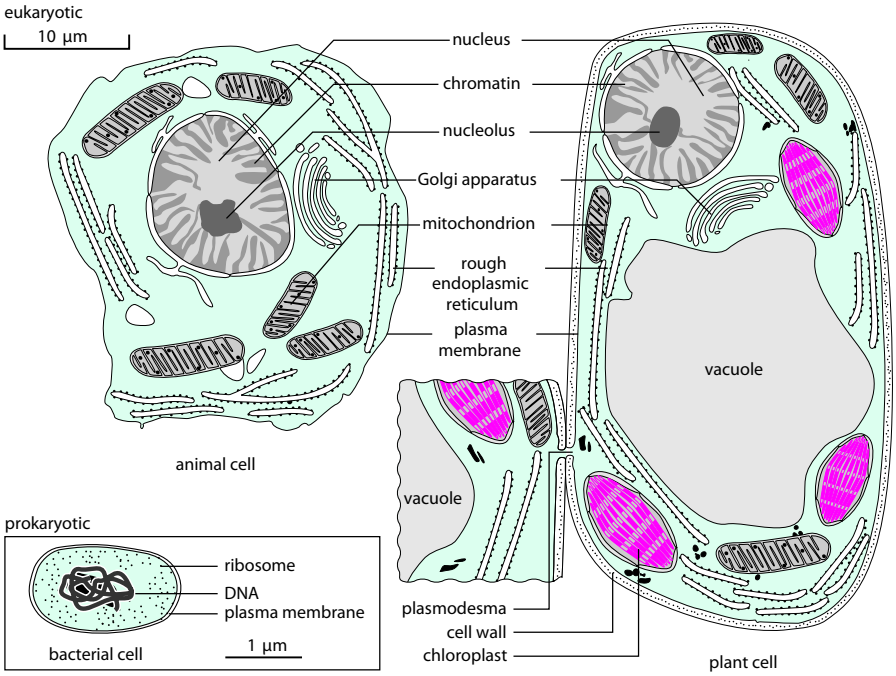

Figure 1.1: Structures of typical cells. A eukaryotic cell is an enclosure of various subcellular organelles connected by endoplasmic reticula (ERs), while prokaryotic cells are more simple in structure, containing only essential ribonucleic acids and proteins surrounded by a phospholipid-made plasma membrane and a cell wall made of peptidoglycans. Illustration adapted from Figure 1.7, Bolsover et al. (2003).

1.1.2.1 Phospholipids

Phospholipid molecules are the constructing units of cellular membranes, including the outer (plasma) membrane and various inner membranes such as ERs, Golgi apparatus, lysosomes and ultimately, the nuclear envelope. The unique hydrophilic-head, hydrophobic-tail character of phospholipids ensures integrity of the membrane in aqueous conditions, while also retaining a measured membrane liquidity required for cellular functions. A typical membrane chunk consists of two layers of phospholipid molecules, whose tails are buried in the middle. As a result, the vertical polarity of cellular membranes becomes amphipathic: the polar-apolar-polar layout thus forms an ideal anchor field for some amphipathic macromolecules, such as transmembrane proteins.

1.1.2.2 Polysaccharides

Polysaccharides are by far the most common energy source used by living organisms. Polymers are formed by the chaining together of monomer saccharide molecules, *via* glycosidic bonds, and usually contain multiple branching points. Depending on monomer types, the length and branching frequency of the polymer chain, polysaccharides can have distinctive appearances and physical/chemical properties, such as those displayed by starches, glycogen, cellulose, chitin as well as the coating polysaccharides in bacterial capsules. The consumption of polysaccharides as a source of energy constitutes a series of chemical reactions. The process begins with the hydrolysis of glycosidic bonds between saccharide monomers, and is followed by the glycolytic pathway that splits a 6-carbon saccharide into two 3-carbon pyruvate molecules. In eukaryote cells, pyruvate is completely oxidated to 3 units of carbon dioxide (CO_2) through the tricarboxylic acid cycle (TCA); in bacteria and other prokaryotes, the lack of a TCA cycle forces the organism to reduce pyruvate into other forms with relatively small amount of energy released. Metabolic enzymes are critically important in catalysing these often reversible reactions; for example, amalyse catalyses the hydrolysis of glycosidic bonds, while each of the 10 steps of the glycolytic pathway, shown in Figure 1.2, require a unique enzyme to overcome the activation energy barrier for each particular reaction. These 10 enzymes will be reintroduced in later chapters as environmental crowding agents for investigating specific interactions between other macromolecules.

1.1.2.3 DNA and RNA

Since the Nobel prize-winning work of Watson and Crick on modelling the molecular structure of DNA (Watson and Crick, 1953), its three-dimensional structure has been an overwhelming image featured in countless professional and popular publications. Essentially, DNA is a double-strand long-chain polymer of nucleotides. The two strands of the nucleic acids are tightly coupled by the binding between complementary base pairs through two or three parallel hydrogen bonds. On the other hand, the structure of RNA, due to lack of complementary strands, is much more flexible; many specific RNA topologies are still unknown and therefore are of current sci-

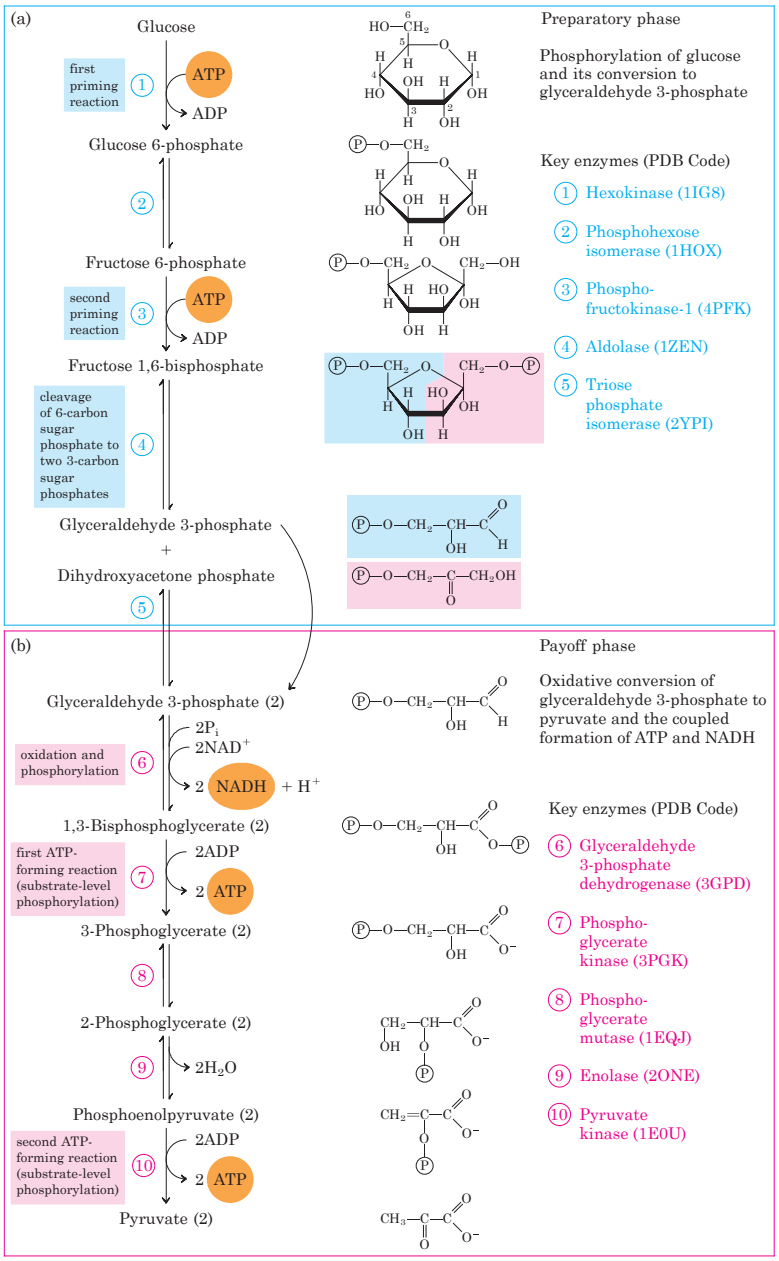

Figure 1.2: Two phases of the glycolysis process. The preparatory phase is an energy consuming process that attaches phosphate groups to either ends of the glycoside; the payoff phase releases more energy, in the forms of adenosine-5-triphosphate (ATP) and nicotinamide adenine dinucleotide with hydrogen ($NADH^+$), from the hydrolysis of 3-carbon intermediates containing high-energy phosphate groups. Each of the 10 steps is catalysed by a specialist enzyme, here with its protein data bank (PDB) code for the bacterial variant of the enzyme used in this work. Illustration adapted from Nelson and Cox (2004).

entific interest; which includes connecting the structure of RNA molecules, in addition to their sequences, to their specific biological roles in the cell.

In the majority of living organisms, DNA is the primary medium for storing genetic information through encoding the sequence of other macromolecules, such as proteins and RNAs. This role is matched with its highly-conserved structure, as well as the ultra-conservative rules for purine-pyrimidine base pairing.

1.1.3 Protein Structure and Interaction

1.1.3.1 Structure

Like DNA and RNA, proteins are also a class of polymer molecules. Also named polypeptide, a protein is composed of amino acid (AA) monomers chained together *via* the peptide bond. The peptide bond is formed between the carboxyl and amine groups of neighbouring amino acids; due to resonance caused by nearby charged groups, this bond exhibits partial single, partial double-bond properties under soluble conditions (Berg et al., 2002). This leads to two unique properties: firstly, peptide bonds are not rotatable, leading to the formation of the peptide plane that incorporates the neighbouring four atoms, N – C_α – C – O; secondly, peptide bonds display metastability, being less stable than pure single bonds but more stable than most double bonds. Therefore, polypeptides in water will eventually be hydrolysed; however, this process is extremely slow and in living organisms, it is accelerated by enzymes, specifically named proteases.

As opposed to only four common nucleotide types (A, G, C and T) for DNA and RNA molecules, amino acids that make up a protein show much more diversity in both number, 20 types, and in their diverse physical and chemical characteristics (see Figure 1.3). Therefore, proteins have a much greater variability in sequence, size, structure and biological activities compared to nucleotide based polymers; in evolution this may have helped them overtake RNA as the primary carrier of most biological functions (Poole et al., 1998; Dworkin et al., 2003).

The structure of a protein, due to its complexity, is commonly understood at four levels. The primary structure is a protein's linear amino acid sequence, counting from the amine end (termed the N-terminal) to the car-

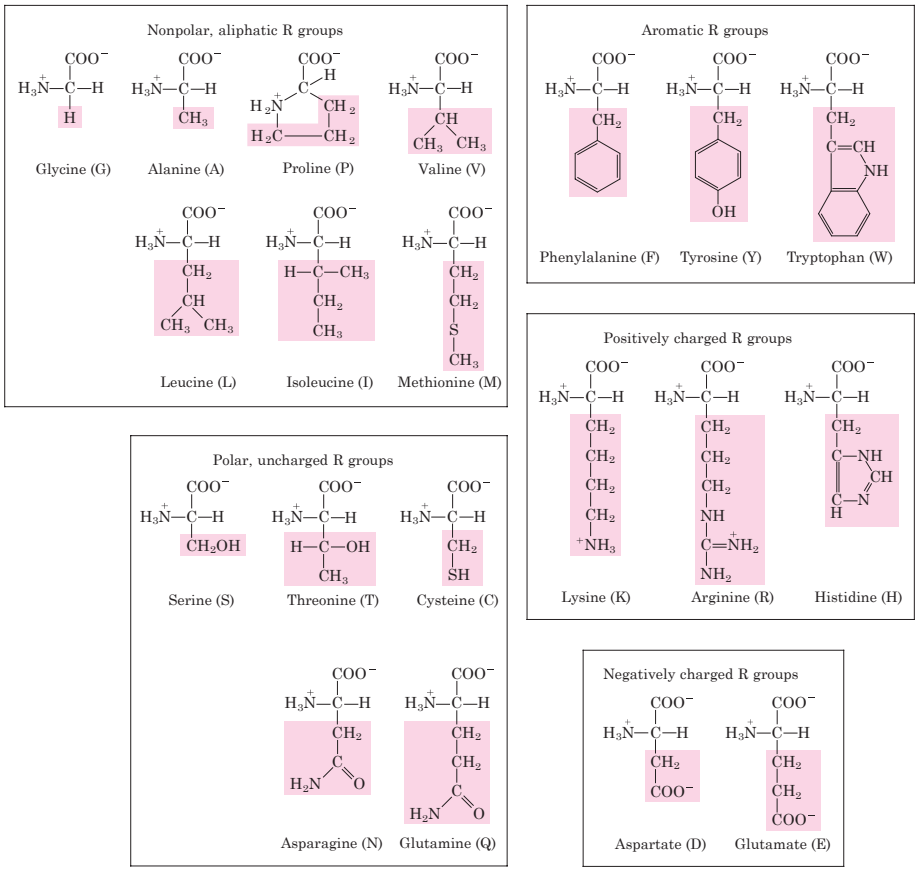

Figure 1.3: The 20 common amino acids of proteins. The chemical structure of each side chain group (R) is shown, with their respective name and one-letter code in brackets. In summary, seven of the AA types have an aliphatic side chain, three have aromatic rings in their side chains, five have polar -OH, -NH_2 or -SH groups attached to the end of their side chains, two have their R groups positively charged, two negatively charged, and one (Histidine) can be either positively or negatively charged depending on ionic conditions. Illustration adapted from Nelson and Cox (2004).

boxyl end (the C-terminal). Evolutionarily, a protein's primary structure (sequence) is the least conserved among the four levels, as a result of point mutations occurring to the host's genome. However, a swap in one letter of the protein's amino acid sequence, namely one residue, does not always lead to a marked change at a higher-level of protein structure, or its function, since some residues have similar shape and/or electrostatic properties. Therefore, single residue replacements, termed mutations, often survive and are retained over the course of evolution. Multiple mutations, perhaps accumulating over millions of years, can result in the diversity of protein sequences, eventually leading to some diversity in protein function for essentially the same protein fold shared among many different organisms. Exceptionally but not surprisingly, there are small chunks of "core" sequences that are highly conserved amongst species, which often make up the functioning site of a particular protein type. During its lifetime, the primary structure of a protein is very stable under natural conditions; however, post-translational covalent decorations do occur to proteins of certain types and functions, such as the palmitoylation and farnesylation of H-Ras (Dudler and Gelb, 1996; Rubio et al., 1999), a membrane-mounted small GTPase (see Figure 1.4) that plays a central role in the *ras*, and a number of other, signalling pathways.

The term "secondary structure" describes the essential constructing elements of a protein's overall three-dimensional structure. Common secondary structure elements (SSEs) include $\alpha-$helices, $\beta-$sheets, $\beta-$turns as well as the random coils that cannot be recognised with an identifiable structural pattern (see Figure 1.5). Mainly formed by inter-residue non-covalent interactions between backbone atoms, SSEs are generally constructed without overriding preference for specific residue types (with the exception of Proline, which usually terminates an $\alpha-$helix). However, there are patterns of consecutive residue types that predispose the formation of one secondary type compared to another; the recognition of such patterns forms the basis of computer algorithms to predict protein secondary structure (Jones, 1999). Under natural conditions, SSEs are normally very stable and are often formed spontaneously after the peptide chain is initially synthesised; the formation of multiple parallel H-bonds further reinstate the helical or sheet structure from minor wiggles and twisting forces caused by nearby

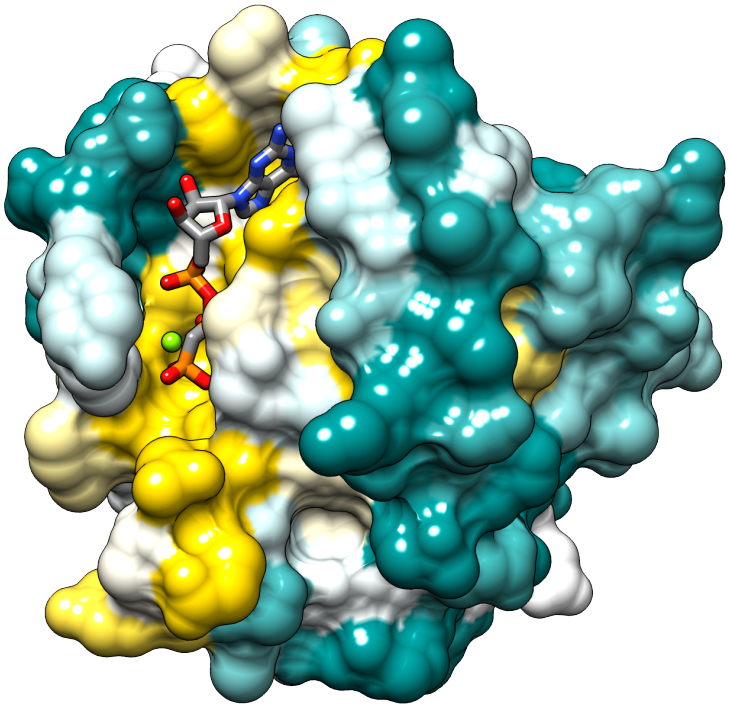

Figure 1.4: Structure of H-Ras (PDB:121P). The molecular surface of H-RAS is shown, coloured by degree of conservation. Bright yellow indicated the most conserved region; dark cyan represents the least conserved regions. An adenosine diphosphate (ADP) molecule is located in the ATP binding site of H-Ras, shown in balls and sticks. The ATP binding region is visibly the most highly conserved region on H-Ras.

spatial constraints. However, in highly salted conditions, excessive number of charged ions could systematically disrupt these proton bridges, resulting in the deformation of H-bonds and consequently, the SSEs.

The majority of proteins consist of a single polymer chain; the folded shape of a protein is called its tertiary structure, which describes the spatial arrangement of SSEs of a protein. Tertiary structure is mainly established by non-bonded inter-molecular interactions, or geometric complementarity, between the side-chain atoms; many of these interactions are residue-specific and are therefore prone to mutations. However, in many cases this will only slightly affect the orientations of the SSEs involved; the overall structure, despite the relatively small deviation from the original, does not change significantly. Sometimes, common structural patterns assembled by several SSEs are seen in a number of structural or functionally related proteins; these patterns are sometimes named super-secondary structures, structural motifs or more recently, domains. Over the years, structural domains have become increasingly important in protein classification and evolutionary studies; databases such as CATH (Greene et al., 2007), SCOP (Murzin et al., 1995) and Pfam (Finn et al., 2010) are primarily based on analysis of structural domains or motifs.

Importantly, a protein's tertiary structure is linked to its physiological functions that depend on interaction dynamics of the protein with other molecules. Moreover, the message conveyed by tertiary structure, i.e., atom positions in three dimensions, is not limited to spatial information; physical and chemical properties are also implied as long as residue and atom types are known. A "tertiary structure" of these properties, such as the spatial map of the protein's electrostatic potential or its hydrophobic affinity, can therefore be constructed uniquely for that protein, just like its unique layout of atoms in three-dimensional space.

Similar to other macromolecules, the tertiary structure of a protein is not a rigid-body. Both the temperature-induced random fluctuation between side-chain conformations, and the backbone bending and twisting movements that lead to rearrangement of the SSEs, constitute a protein's structural flexibility. The resulting structure is therefore not some static spatial occupancy, but a dynamic body that continuously changes its shape, although in most cases such alterations have minimal impact on the stability

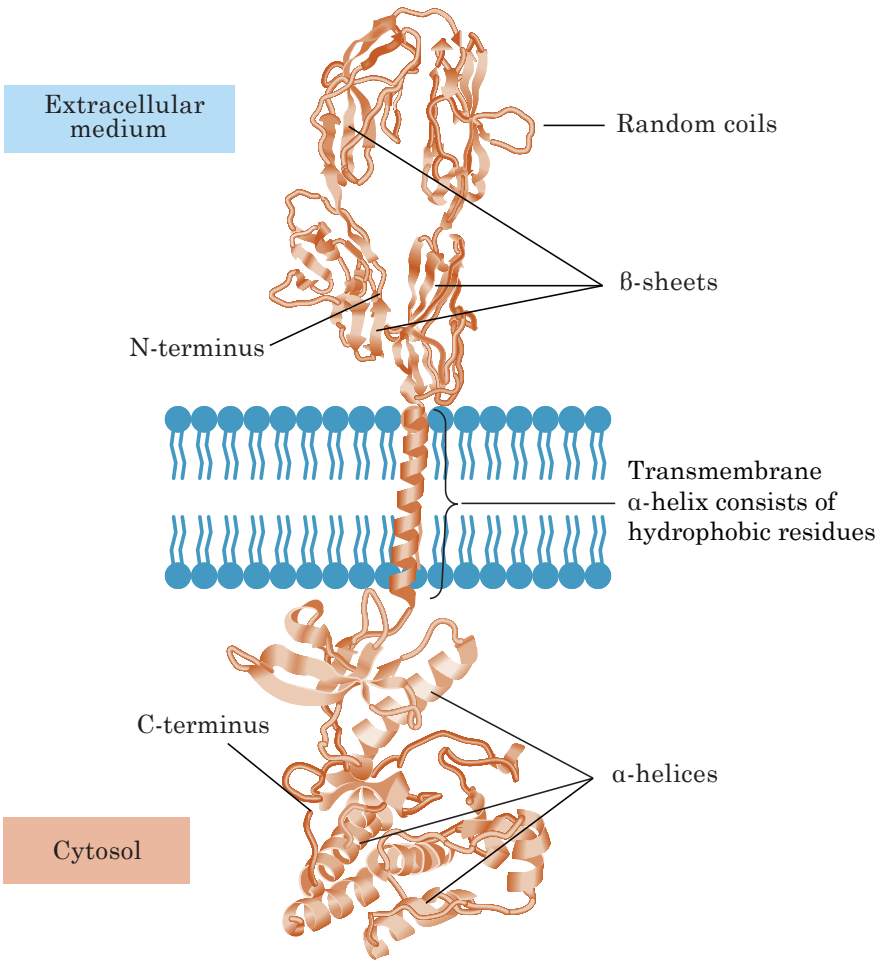

Figure 1.5: SSEs of a transmembrane growth factor receptor. Various secondary structure elements, including random coils linking them, are displayed as ribbons. The blue layer is a schematic representation of a lipid membrane, with the hydrophic $\alpha-$helix of the protein buried inside. Redrawn and altered based on the molecular model in Figure 9.15, Bolsover et al. (2003).

of the overall structure. It is also worth noting that some proteins, such as $\alpha-$endosulphine, a phosphoprotein involved in oocyte meiotic maturation, lack a stable tertiary structure at all (Boettcher et al., 2008). These extremely flexible proteins, often called disordered proteins, may be presented as flexible chains folded to SSE level, but subject to significant conformational changes within the cytosol due to the lack of a discoverable lowest-energy state (Papoian, 2008). Increasingly, such as for the case of endosulphine (Heron et al., 1998), disordered proteins have been found to have a regulatory role for multiple receptor-ligand interactions, probably because the relaxed structure leads to a higher level of domain or SSE-specific promiscuity.

The top level of protein structure, the quaternary structure, describes the spatial arrangements of multiple peptide chains. Depending on the number of symmetrically packed monomers, a number of different topologies can be formed, the most prevalent being dimeric, tetrameric and hexameric forms. The quaternary structure of a protein can be equivalently seen as a homogeneous protein-protein interaction that has very high affinity between monomers. From an evolutionary perspective, it is interesting to note, that the homo-polymerisation of structural subunit occurs more often to enzymes that need to process/carry more than one substrate at once to increase efficiency, such as haemoglobin (tetramer, Perutz et al., 1960) and pyruvate kinase (dimer and tetramer, Weernink et al., 1992). The inter-subunit interaction potentials that favourably keep the subunits together are the same as those conducting heterogeneous protein-protein interactions.

1.1.3.2 Interaction

In vivo and *in vitro*, proteins naturally associate and dissociate with each other, specifically and non-specifically, by non-bonded interacting forces. Protein interactions are of high biological importance and, along with protein folding and assembly, our lack of ability to routinely predict and understand such events was listed in the 100 unresolved scientific questions raise by the journal Science (2005) – "how do proteins find their partners?". Indeed, cellular processes, such as signal transduction, rely on cascaded specific protein-protein interactions to pass molecular messages that regulate activities within the cell, including cell growth and apoptosis, differentia-

tion and motility, molecular intake and secretion, as well as meiosis and mitosis. In essence, protein interactions are involved in almost all biological functions; it is therefore important to understand them in terms of both their dynamic and energetic characteristics.

Unlike the interactions between small molecules, protein interactions take place on a much more variable time scale, of which specific interactions may last between a few nanoseconds and, in extreme cases of specific binding, a matter of days, such as the case for the ribonuclease barnase and its inhibitor barstar (Hartley, 2001), as well as for the pancreatic trypsin and its inhibitor BPTI (Favre et al., 2000). The strength of such bindings is predominantly determined by complementarity between energy surfaces of the interacting proteins; due to the uniqueness of such a surface, a protein may only have one or a few specific binding partners at a particular interface, to which it shows higher affinity than those from random associations. The wide range of durations for retention of interaction also makes biological sense: some proteins are designed to inhibit others, thus a tight and lengthy binding is expected; for others, such as those in signalling pathways, a ligand binds to its receptor to physically or chemically modify it, hence only a small amount of time spent together would satisfy this need.

The mechanism of one protein binding to another is not intrinsically different from that of a protein folding itself up into its tertiary structure, or that of a multi-chain protein assembling its subunits. Non-bonded interaction potentials between atoms play a vital role in deciding the movement of an atom, or groups of atoms, in folding, assembly or binding processes. Among these potentials, the most common one arises from van der Waal's (VdW) forces. The VdW force describes the non-bonded, non-electrostatic interactions between nearby atoms due to the attractive London dispersion force (London, 1930) and the repulsive Pauli exclusion principle. In general, the London dispersion force is so weak that, unless there is a strong complementarity between surfaces of the proteins in question, protein-protein binding cannot be stabilised by London forces alone. A second source of attraction and repulsion is the Coulombic forces between partial charges on protein surfaces. These electrostatic interactions play a stronger role than VdW forces in forming specific protein binding (Nicholls et al., 1991; Sheinerman, 2000), further evidence of which will be shown in later chapters. As

a special case of dipole-dipole interaction, the electrostatic interaction between two strong electronegative atoms sandwiched by a hydrogen atom has the strength and stability comparable to a weak chemical bond, and is thus named the hydrogen bond (H-bond). H-bonds are widely seen in many biological macromolecules, such as base pairing in DNA, formation of protein SSEs and here in holding together protein binding partners. In addition to the above pairwise interaction forces, there are other factors that contribute, sometimes significantly, to established protein complexes. The most notable among these is the hydrophobic effect, which results from the entropy loss due to loss of fluctuating hydrogen bonds that would otherwise have been formed between water molecules filling the space occupied by the hydrophobic solutes. However, protein interactions dependent mainly on hydrophobic "forces" are generally less stable than those based predominantly on attractive electrostatic interactions, since it is more of an effect driven by the distribution of water molecules and not a force between the interacting partners. Whether molecular hydrophobicity can be transformed into complex stability further depends on the shape complementarity between binding interfaces, i.e. the "water tightness" of the buried hydrophobic interface. A reduced non-polar surface area exposed to water molecules would have the least cost in term of entropy for an aqueous environment. Further factors that affect the stability of protein-protein interactions include hydrodynamic effects (convection) and Brownian motion (diffusion). By its nature, the flow of solvents or solute molecules in a convection current will exert continuous forces on the binding partners in a particular direction; as long as the direction is not exactly orthogonal to the interface "plane" between the binding partners, it would have a negative impact on complex stability. Molecules undergoing Brownian motion, however, show diverse behaviour: in one way the random movement of solvents and solutes may cause instability in the established complex, alternatively such random exploration forces may assist the interacting proteins to correctly locate their specific binding interface. Analysis of these two contrasting effects is one of the main objectives of this thesis.

1.2 Computing

This section briefly reviews recent progress made in computing and computational approaches that are directed towards, or related to, the problems of modelling proteins and their interactions at the atomic level.

1.2.1 Hardware and Software

1.2.1.1 Computing Platforms

The average size of bio-macromolecules poses a significant problem for their computational modelling and simulation given even the current computing capability. However, attempts to represent a protein in its component atomic positions through the use of a computer started as early as 1970s using mainframes, which could provide not only relatively fast arithmetic calculations but also the required precision (Hermans and Vacatello, 1980). The 1980s, represented a "boom time" for all variants of molecular/protein modelling and simulating protocols; however, many have since disappeared for being incapable of adapting to better hardware infrastructure introduced through the later years. Given the computing characteristics of macromolecular modelling and simulation, a competent hardware platform would need to consider the following points:

Data precision for calculation and data storage. Historically, there are two main ways of representing real numbers, which are double precision (DP) and single precision (SP). The word "double" indicates the data width for representing these numbers are twice as long as the normal data width, which implies 32-bit (single-precision, SP). It is widely known that 32 bits are not sufficient in exactly representing many finite real numbers, such as 0.1; DP solves this problem by using 52 bits rather than 23 bits in SP for representing the fraction. Moreover, errors accumulated from SP calculations could lead to significant biases in matrix operations; in rigid-body dynamics this creates shape distortion in the moving object. The DP vs. SP argument would have been long gone after the introduction of 64-bit architecture which provides native (full-speed) support for computing double-precision numbers;

however, it is still worth mentioning today because of the rapid development of graphic processing unit (GPU) computing. SP numbers are intrinsically used for in GPUs for higher speed and bandwidth. The latest GPUs specially designed for scientific computing have added some native support to DP numbers, however, this significantly constricts their performance relative to SP calculations.

Parallelisation mechanism. Molecular systems are large in size; parallel computing platforms provide a quicker way to process the state variables of hundreds of thousands of atoms. Two types of architecture are commonly seen: shared-memory (SM) and distributed-memory (DM). DM parallel computing was the main form of parallelism until the end of 20th century, but had since been gradually overtaken by SM architecture with the introduction of multiple computing cores on a chip in addition to the bulk memory space. Most commodity PCs have about 4GB of random-access memory (RAM), which could theoretically store positional and other information for up to *circa.* 85 million atoms with the assumption of 50 bytes per atom (24 bytes for Cartesian coordinates, 8 bytes for atomic charge and 18 bytes for various type indices). Therefore, most modelling and simulation work can be accommodated well within a single process in one system. The clear benefit of doing this, rather than distributing the molecular system into multiple processes running on multiple computers, is the elimination of often very complex and time-consuming algorithm for neighbour exchanges, i.e. detecting and moving a particular set of atom from the memory of computer A to that of computer B. Even when a similar grid is needed to be implemented for a shared-memory architecture, to partition the simulated space/model into smaller blocks/chunks, it is still far more efficient than having to do distance evaluation across a network of computers. On the processor side, modern server farms are normally equipped with 4- or 8-core CPUs per blade server which, from the author's point of view, is adequate for molecular systems under 1 million atoms. For moderately-sized molecular systems, the benefits of doing multi-threaded calculations on a SM machine will offset the incapability of extending the task to more than one computers. The real potential of SM architecture, in the author's view, will

only be fully explored by a recent trend in shifting scientific computation from CPUs to GPUs. With the massively parallel but relatively simple "stream" processors, the typical GPU architecture is deemed more suitable for handling large number of pairwise distance evaluating calculations which are a dominating factor in the total performance of molecular modelling/simulation software. Notably, there have also been rare attempts to design specialist parallel hardware purely for the purpose of speeding up the molecular modelling/simulations, such as FASTRUN Fine et al. (1991), MD Engine (Uehara et al., 2002), MDGRAPE (Elmegreen et al., 2004; Kikugawa et al., 2009) and more recently, the Anton platform designed and developed by D. E. Shaw Research (Shaw et al., 2007). Under the hood, Anton is more similar to a hardware-implemented partition-grid on a SM architecture, than to the traditional cluster of computers communicating between each other (DM). Each computing node in Anton resembles a stream processor in a GPU but partially specially-wired for pairwise forcefield evaluations. However, the cost of a purpose-built chip for conducting molecular modelling/simulation has prohibited most academic/commercial organisations from doing so, while the flexibility of the Anton platform for more adaptive uses in the future, remains a question.

Task distribution and deployment. Software parallelism is all about completing a task using multiple processor resources, usually through the running of one application (process), on a specific computer or computer cluster and within a finite time period. A greater level of parallelism can be envisaged at the tasking level, which is to divide a large problem into a large number of smaller ones and to complete them individually and asynchronously. For example, the search for an energetically favourable structural conformation for a certain protein sequence, or a certain posture of two proteins interacting with each other, can both be investigated using a large number of commodity computing resources asynchronously using this divide-and-conquer strategy. With the presently popular concept of cloud computing, such jobs can be distributed across a cloud of computing nodes and performed when the resources are available. However, not be-

ing able to communicate much between the distributed tasks mean that they couldn't be split at the grid level, but at the run level. This consequently requires a good design in balancing the computing requirement for a single task, and the total number of tasks to be distributed across the cloud. A promising example is the Folding@home project (Pande et al., 2003), which currently employs circa 350,000 active CPUs performing various projects related to protein folding. In their recently published study (Kasson et al., 2010), this computing resource was used to investigate the transitional state in vesicle fusion process at atomic scale. Similar cloud computing project could also be set up for investigating protein-protein interactions; however due to the larger problem size (more atoms involved than in a typical folding simulation) and intrinsically less perfect forcefield, this may require more development time. An early attempt, Docking@home by Taufer et al. (2009), is still under active development and benchmarking. In some way, cloud computing seems to have offered the ultimate hope of simulating, atomically speaking, extremely large chunks of space within a living cell. However, such enterprises are again limited by the classical neighbour exchange problem faced by the DM architecture – the inter-communication between the cloud nodes is unavoidable if the total simulation space is partitioned. In this case, home PCs loosely connected to the internet may no longer be suitable for being the cloud nodes; professional services, such as the Elastic Compute Cloud (EC2) provided by Amazon, may play a significant role in macromolecular modelling and simulation in the future.

1.2.1.2 Development

With the development of hardware architecture there also came the creation and deprecation of developing tools and languages over the passing decades. At first glimpse, development tools and programming languages seem too technical to be related to the scientific problem of molecular modelling. However, the evolution of programming languages has increasingly enabled the modelling and simulation software to accommodate more atoms, adopt mixed modelling resolutions and as always, to automate more processes. Early packages, such as CHARMM (Brooks et al., 1983), AMBER

(Pearlman et al., 1995), UHBD (Madura et al., 1995), HYDROPRO (Garcia de la Torre et al., 2000), were often written in Fortran and procedural C, which guaranteed an acceptable performance given the then-limited computing resources. The limited capability in memory management and data encapsulation of early FORTRAN and C resulted in these packages containing a large number of individually compiled binary executables. Consequently, data exchanges between these programs were made via disk files, leading to inefficiency and sometimes confusion. In addition, maintenance of numerous binaries can be difficult, while adding new functionality as new binaries into the package would only worsen the situation. Apart from these operational problems, there has been a much larger hurdle right in front of the macromolecular modelling community. It is clearly known to the community that, yet more computing power will be needed for modelling a reasonably sized biological system at the atomic or sub-atomic scale. Therefore, compromises have to be made, for example, on how to retain high resolution at necessary places while keeping everything else in coarse-grain. Such a flexible scheme should be applied to both spatial and temporal measures. However, without the advanced software engineering approaches in recent years, it would have been extremely clumsy to realise this flexible modelling design – hence there is a general lack of multi-scale modelling or simulation packages currently available in the field.

The philosophy of software design and coding style evolves in line with computing power and the computing requirements of various research communities at different times. At the very beginning (1960s), there was ultimate freedom in designing and writing software code; this model quickly encountered increased occurrences of code clumsiness when the lines of code (LOC) of a project grew exponentially to accommodate more needs. Soon the adoption of procedures and functions made code re-usable on the "block" scale. This was sufficient at the time when most CPUs executed instructions at a speed around 1 MIPS (million instruction per second); in 2010 many quad-core CPUs are rated more than 60,000 MIPS.

In more software-oriented fields of study, structural design of code has been elevated from procedural and functional level to object level, with the introduction of objected-oriented (OO) languages such as C++ and Java in the early 1990s. In molecular modelling, however, this replacement did

not take place, as most packages had already grown too large to afford a re-write. Meanwhile, the initial concept of the object-oriented software model did not look appealing to the computational field as the overhead cost of having "objects" in modelling software will cause a significant drop in run-time performance. After the year 2000, new thoughts on designing software gradually emerged to account for many drawbacks of canonical, "everything-is-an-object", philosophy of OO programming. Consequently, objects have since then become more lightweight and function-oriented rather than structure oriented. This change was picked up by some computational groups whose function-oriented modelling packages were proven quite successful, such as the Poisson-Boltzmann solver program, APBS written in Objective C (Baker et al., 2001).

The latest major addition to programming design, especially in C++, has been the introduction of class and function templates, as well as the concept for generic and functional programming. There are two key differences between templates and base classes in OO: first, it is much looser and more lightweight than class inheritance; secondly, all conceptual checks are done at compile-time, not run-time as with the dynamic binding mechanism used in traditional OO. With templates, a specific function, for example, the calculation of moment of inertia, does not have to bind to a certain class of objects or, like in OO programming, to types from a common base class. The function can bind any class that obeys the concepts you assumed it to have, such as the position and mass properties of an object. This feature, along with many other performance-related improvement associated with template programming, can greatly help easing much of the complexity incurred in modelling and simulating molecules in a multi-resolution, multi-timestep and even multi-dynamics fashion.

1.2.2 Macromolecular Modelling

In the context of most scientific thought, macromolecular modelling specifically means modelling very large molecules at the atomic scale. However, taking its literal meaning, any approach that models the structure, function or behaviour of one or more molecules shall be legitimately reviewed under the concept of molecular modelling. And indeed there has been a whole spectrum of modelling studies aiming at tackling the macromolecu-

lar "structure, function or behaviour" problems on different size and time scales. Below a brief, non-exhaustive, review of such computational models is given.

1.2.2.1 Modelling at different resolutions

The primitive modelling of molecules starts with kinetics. Chemical reactions, as well as the non-bonded binding between proteins and/or ligands, occur at a certain rate,

$$r = \frac{d[A]}{dt} = -k[A]^n, \tag{1.1}$$

for a certain reactant A at concentration $[A]$. The number k is the rate constant of the process, independent of the reactant concentration, while n is the order of the reaction with typical values $n = 0, 1, 2$. For a system of reactants, if the rate constants and reaction order between each pair of them are known, we would then be able to simulate the reactive behaviour of these reactants. This approach, taught in most chemistry courses, is adopted and used in simulating protein-protein interaction networks. Essentially, the "reactivity" of a system of proteins can be represented by a graph, in which each connecting arc stands for a binding link between the two nodes (proteins); each arc is mathematically written as some variant form of Eq. 1.1. Given the initial values, the dynamic state of the network, i.e. the concentration of protein species at each timestep, could be solved by numerically integrating the system of differential equations over time.

This approach has been widely used in probing enzyme-inhibitor dynamics and regulation of biological processes such as protein synthesis and signalling pathways. In many cases, such as the investigation of protein phosphorylation dynamics in the TGF-β pathway (Schmierer et al., 2008), experimental verifications were found to be in line with computational prediction, proving the genuine usefulness of the ordinary differential equation (ODE) based models. One of the main appeals of such models is the simulation speed – even for large systems consisting of many tens of differential equations, it would require no more than a few seconds to perform a full run. Parameterisation, however, is key to success of these models. There might exist more than one set of parameters that would produce

the desired behaviour; or there might not be any parameters satisfying the boundary conditions. This quickly worsens as the dimension of parameter space increases – even advanced optimisation techniques, such as support vector machines (SVMs) or probabilistic models may not respond well to high-dimensional continuous-parameter problems. The major drawback of the ODE-based approach is the neglect of particle trajectories. As a result, interaction rates are completely dependent on external input, often empirical and inaccurate in terms of how the system should respond in a particular environment. Moreover, in biological systems, reacting agents, such as proteins, usually have very limited copy numbers. It is therefore very inaccurate to assign an otherwise, statistical, rate equation to a particular interaction process.

The Gillespie algorithm (Doob, 1945; Gillespie, 1976) provides a swift answer to the above problems. It works on the assumption that molecular interactions are events that occur with certain probabilities, which are estimated based on the fraction of molecular collisions with the proper energy and orientation. Therefore, two random variables are needed to determine the occurrence and time of the next reaction. When simulated, the Gillespie algorithm gives results that satisfy the distribution of the chemical master equation,

$$\frac{dP_k}{dt} = \sum_{\ell} T_{k\ell} P_{\ell}, \tag{1.2}$$

where P_k is the system in state k and T_{kl} is the transition rate constant from state l to state k. The algorithm has been used in a number of biochemical network studies (Adalsteinsson et al., 2004; Meng et al., 2004; Bhattacharya, 2010) and have since been adapted to fulfill special conditions of interaction, such as delays in reaction and non-Markovian properties in gene regulatory networks (Bratsun et al., 2005). Gillespie-based algorithms are computationally more costly than ODE-based approaches; a number of studies have focused on improving their performance on large networks (Slepoy et al., 2008; Cao and Samuels, 2009).

In general, most Gillespie-based algorithms treat interacting agents explicitly but do not consider their spatial distribution. More advanced approaches do so by incorporating two or three translational degrees of free-

dom – the positional coordinates in 2D or 3D. Smoluchowski (1917) outlined the velocity-independent distribution of particles undergoing a diffusional regime,

$$\frac{\partial S(x,t)}{\partial t} = \frac{1}{\zeta}\left[-\frac{\partial}{\partial x}F(x,t) + kT\frac{\partial^2}{\partial x^2}\right]S(x,t), \tag{1.3}$$

where $S(x,t)$ is the time-dependent distribution, $F(x,t)$ is the external force-field, ζ is the drag coefficient due to solvent viscosity, k is the Boltzmann constant and T is the temperature. Andrews and Bray (2004) modified and improved the above scheme, which was subsequently verified experimentally (Tournier et al., 2006). Further improvements based on Green's function were implemented, such as an event-driven, variable timestep particle simulation regime that allows spatial and temporal "jumps" to reduce the "idle" timesteps wasted on simulating diffusion (van Zon and ten Wolde, 2005). A similar scheme was also considered and implemented in this work, although it is based on entirely different mechanisms (see Section 2.1.3).

One problem persists: particle descriptions may be suitable for describing small molecules which do not differ markedly in size; for biological macromolecules the above particle descriptors are largely inadequate. There are models that represent proteins as spheres of different radii. However, these models are less frequently used directly for studying diffusion/interactions, since spherical protein models would still be too coarse. On the other hand, spherical models are more often used in theoretical modelling studies, such as the derivation of association rate constants using directional constraints for diffusion by Schlosshauer and Baker (2004). Due to its simplicity and relevance to protein sizes, spheres have been popular models for crowding agents. For example, in a recent protein folding study performed to take account of the crowded macromolecular environment (Jefferys et al., 2010), spheres were used as volumetric representation of environmental proteins whose surface specificity is not of primary interests.

All structural information for a protein needs to be utilised if specificity of interaction between proteins is to be rigorously studied. A quick way of doing this is to use lattice-based models to represent the backbone conformation of a protein. A huge benefit of discretising the configuration space into lattice points is the restraining of searching space; thus the method has

been adopted by many models for protein structure prediction, such as the TOUCHSTONE method (Kihara et al., 2001).

Desires to model more complex protein structures, coupled with the development in computing power, has driven macromolecular modelling technique into continuous space. In such models, protein backbones are mostly modelled at the atomic resolution, while in some packages, side-chain atoms are consolidated into a sphere, or other simple geometric bodies that resemble the shape of that residue. The above treatment of side-chain topology is also called the "united-atom" approach and was early explored in structure prediction (Liwo et al., 1993) and, combined with Langevin dynamics, in protein folding (Liwo et al., 2005). Popular molecular dynamics (MD) simulation packages AMBER (Case et al., 2005) and CHARMM (Brooks et al., 2009) support a finer united-atom representation of residues. Named "atom-groups", they are a conglomerate of nearby atoms in a common functional group, often bearing integer charges (-1, 0 or 1).

Eventually, the so-called full-atom approaches have increased in popularity, particularly over the last 10 years. However, the word "full-atom" can still be confusing with regard to the full extent to which macromolecules are represented at atomic detail. In essence, a full-atom model requires the dominating majority of the molecular structure to be explicitly modelled by spherical atoms with different radii; in reality, the following extensions are worth considering:

- Hydrogen atoms. A number of modelled topologies ignore H atoms for simplicity, such as the CHARMM 22 united atom parameterisation.
- Partial atomic charges. An atomic model would become "fuller" if the charges are distributed to each atom in a functional group, reflecting the polarity of the charged body.
- Internal flexibility. Theoretically all full-atom models should have bond-level internal flexibility as implemented in the classical MD packages AMBER and CHARMM. However, in adaptation to various needs, many have discarded backbone flexibility, or side-chain flexibility, or both. Technically speaking, all such models may be termed atomic models, as long as the energy function employed in the simulation still evaluates the interaction potential per atom pair.

- VdW approximation. Some models employ a simple hard-wall potential to bounce colliding particles off each other, while some others have adopted the full Lennard-Jones potential form. The benefit of taking the latter form is the reproduction of VdW attractive term (the London dispersion, described in Section 1.1.3), while the main drawback is the computational cost for doing many floating exponential calculations.

- Solvent modelling. MD packages tend to model water molecules explicitly to create a fully solvated environment. However, due to the huge computational cost of computing uninteresting trajectories of individual water molecules, many adopted various implicit solvation models, such as generalised Born (GB), Poisson-Boltzmann solvent accessibility (PBSA) and simpler, damped dielectric-constant models.

Full-atom models are typically driven by molecular mechanics (MM) which uses analogues of classical mechanics for modelling atoms, bonds and non-bonded interactions. In most cases this is already accurate, but for cases involving chemical reactions, i.e. the breaking or creation of bonds, quantum mechanics (QM) methods may be employed to derive the expected behaviour. QM methods rely on knowledge of sub-atomic entities such as electron configuration and atomic energy states, which would mount huge computational burden for molecules as large as a protein. Therefore, QM-based approaches are rarely used for investigating protein structures and protein-protein interactions except for a few examples (Lameira et al., 2008; Liu et al., 2001).

1.2.2.2 Modelling on different time scales

The appropriate time scale for a molecular simulation package is the length of biological time that can be simulated by that package using reasonable computation time and importantly, before the accumulation of errors diminishes the signal-to-noise ratio. The latter point is vital in that, without proper justification, no meaningful result can be achieved just by "brutally" increasing the computational performance thereby, running longer simulations.

Fully flexible atomic models are normally time-stepped on the femtosecond (fs) scale and simulations run up to nanoseconds (ns), which effectively

is an integration over 1 million to 10 million steps. Protein dynamics simulations beyond these timesteps, although thermodynamically stable, cannot accurately reproduce what happened in a real environment, for example in a crowded environment with other proteins, and therefore cannot be fully trusted biologically. It is already difficult to simulate the translational and diffusional behaviour of macromolecules based solely on MM calculations with explicit solvents – the reason why specially designed computational models need to be constructed for modelling particular macromolecular events, taking place over particular time periods.

Rigid-body atomic models are commonly seen in packages primarily dealing with macromolecular interactions. Diffusional behaviours are therefore an important aspect engraved into the dynamics, and there has been an increasing number of algorithms available for probing protein diffusion and interactions (Schreiber et al., 2009), indeed, this is a subject of the work presented in this study. However, even with unlimited computing resources, these simulations cannot be meaningfully stretched over a few tens of microseconds (μs) due to the accumulation of integration errors. For example, the assembly of larger biological complexes is expected to take place on the ms scale, but most of the current models are unable to be reproduce this behaviour.

One way to resolve this is to use coarse-grained models embedded with specific knowledge on protein dynamics and/or its specific binding affinity. A notable example of this is the series of studies on the assembly of a nuclear-pore complex (NPC) (Devos et al., 2006; Alber et al., 2007; Lezon et al., 2009). Typically, *a priori* knowledge can be acquired for overall charge distributions, surface shape and complementarity, as well as modes of molecular motion. Molecular assemblies can be stabilised using virtual bonds, as employed in elastic network models (Lezon et al., 2009).

To date, longer time scales are generally not suitable for most full-atom and coarse-grained molecular models, as both the energy function and the integration scheme employed will almost certainly not produce a biologically sensible outcome. On the other hand, empirical knowledge on binding, or rate constants derived from higher-resolution models or experiments, has to be used as direct input for coarser models, i.e. those using spheres and particles. Given most biological events take place on a millisec-

ond (ms) to second (s) time range and the importance of maintaining numerical accuracy by carefully constructed dynamic integrators, a timestep on the order of tens of nanoseconds may have to be used to reproduce the macroscopic behaviour of these long-term events seen in a real environment.

1.2.2.3 Modelling with different motion schemes

Each molecular model resorts to a certain motion regime, moving the modelled molecules from one state to another, based on the potentials calculated from pairwise interactions. Generally speaking, motion schemes can be summarised as follows:

- Physical schemes. The most commonly used is Newtonian dynamics, employed in MD packages, based on the Newton's second law of motion. Diffusion dynamics, including translational and rotational diffusion, is modelled based on the Langevin equation or its non-inertial equivalent, Brownian dynamics (BD), both of which are discussed in detail in Chapter 2.

- Quasi-physical schemes. Most elastic network models belong to this type, as well as schemes generating virtual forces to power the Newtonian dynamics integrator. A notable example is the use of swarm intelligence to predict protein-protein interactions (Moal and Bates, 2010).

- Non-physical schemes. These includes a wider spectrum of algorithms compared with their more physical counterparts. Various Monte-Carlo conformation generation schemes are a main part, and are accompanied by probabilistic models and normal mode based conformational selection schemes. Finally, evolutionary pressure can be used as driving forces for model selection, such as the genetic algorithm (GA) recently employed within our own laboratory to predict protein structure (Offman et al., 2008).

To summarise the above, motion schemes, on a variety of time scales and resolutions, constitute a three-dimensional parameter space that to varying degrees of success are covered by a number of macromolecular modelling and simulation packages. Interestingly, the correlation between any two of

these three principal parameters – motion, time scale and resolution – is high: if the parameter space was a cube with unit edges, then the packages tend to approach the diagonal line between (0,0,0) and (1,1,1).

1.3 Thesis Overview

Following on from the brief introductions of both the biology and computing sides, it is clear that the understanding, analysis and prediction of protein-protein interactions has been one of the main problems in both fields, attracting major interest from scientific communities across these fields. These are also the central themes of the study presented in this PhD thesis but with a particular emphasis on the time course of protein-protein interactions. Clearly, it is not possible for a single researcher to produce an all-encompassing protein molecular dynamics package, that is constructed on the latest parallel computing architecture and simulates numerous protein-protein interactions all on a variety of time scales. However, certain sub-problems in the macromolecular modelling field of protein-protein interactions can be tackled that for a number of reasons those developing the large MD packages have not addressed. These sub-topics are termed "the gaps" in our collective knowledge and are outlined below. A number of "hypotheses" are then presented as to how these gaps may be addressed. Research supporting the arguments in this discussion form the base for each of the subsequent chapters.

1.3.1 The Gaps

- There remains a gap in our understanding of how collective, compared to pairwise, protein-protein interactions work, i.e. the discontinuity from docking to binding kinetics. Although the well known Gibbs free energy takes the form, $\Delta G^0 = -RT \ln K_{eq}$, which links the equilibrium rate constant K_{eq} and the binding free energy ΔG^0, it is often difficult to compute the binding free energy with respect to an individual molecular environment in which the interaction takes place.

- The gap between *in vitro* and its *in vivo* protein kinetics. At least from experimental assays, much is known for specific protein binding; how-

ever, very little is known about how proteins actually find and and bind with their potential partners in a heterogeneous macromolecular environment; of particular interest is how proteins locate partners in crowded environments, i.e. environments with very high macromolecular concentration, typically found within cells.

- The discontinuity between time scales and resolutions of existing modelling packages. Those doing docking analysis, those investigating binding kinetics and those simulating crowded macromolecular environments don't normally overlap. It is often difficult to interpret the disparity between the results from different methodologies, let alone building a consistent, multi-aspect image of the molecular interactions of interest.

The methodological objective of this PhD thesis is to fill in these gaps and unite the three areas of docking, kinetics and macromolecular crowding under one roof.

1.3.2 Hypotheses

The proposed "under-one-roof" methodology is aimed at providing qualitative and quantitative insight on the following hypotheses, which constitute the core chapters of this thesis:

1. Binding mechanisms.

   (a) Inter-molecular movements (Chapter 3). For many protein-protein interactions, a high binding affinity between the receptor and the ligand is achieved through the so-called molecular steering mechanism: firstly, proteins make initial contacts and form a transient encounter with their binding partner(s); this is is followed by subsequent rolling/spinning movements of both molecules while in contact with each other. The extent to which these motions affect specific protein-protein binding is evaluated.

   (b) Intra-molecular movements (Chapter 4). For a long time it has been argued whether proteins bind through an induced-fit or

a simpler, conformational shifting (or its predecessor, lock-and-key) mechanism. Molecular dynamics has revealed the possibility of induced-fit upon protein binding; however, a recent study (Tsai et al., 2008) has reinstated the importance of conformational shifting. In this study, the debate is expanded from the perspective of encounter complex formation and is examined from an aspect that takes account of kinetics.

2. Association rate constants (Chapter 4). The Gibbs free energy equation, $\Delta G^0 = -RT \ln K_{eq}$, provides a link between docking and the kinetics of protein-protein interactions. This link should be directly reproducible through conducting atomic-level simulations on a multi-molecular interaction system. If this is the case, association rate constants (k_{on}) can therefore be predicted for a set molecular environment.

3. The consequences of macromolecular crowding (Chapter 5). It has been observed that environmental proteins, especially presented in a crowded concentration, can influence the binding/unbinding dynamics of many protein-protein interactions. The following questions are probed:

   (a) Whether the crowded association rate constants are predictable through directly simulating such an environment.

   (b) What are the driving mechanisms behind these changes. One factor, called the volume exclusion effects (VEEs), is shown to increase the activity of target proteins by reducing the effective volume of the solution (Minton, 1981). In addition to verifying this relationship, evidence is presented to show that environmental electrostatics may also account for these changes.

A general conclusion of the methodology developed and research work carried out in this thesis is provided in Chapter 6, and is complementary to the individual discussion sections for each of the above three core chapters. Chapter 6 also describes a number of avenues that can be taken for future development of the new methodologies presented, along with potentially interesting biological applications.

Chapter 2

Methodologies

A major part of the doctoral studies is devoted to the development of a macromolecular simulation framework. As reviewed in Section 1.2, there has been much progress in both the theoretical and application developments of the macromolecular simulation field. Given the increasingly readily available resources for molecular modelling and simulation, the need for developing yet another major simulation scheme has to be carefully justified. Based on the review in Section 1.2, the following gaps are identified from, or significantly under-represented by, currently available packages.

- There are few packages that work at the atomic scale whose main aim is to deal with multiple macromolecular interactions in a simulation box. However, it is envisaged that expanding the studies of protein-protein interactions onto a multi-molecule simulation platform may open up new territories on which to probe and manipulate such interactions; in particular, association kinetics and competitive binding.

- Specific protein-protein interactions are usually under-investigated in a diffusive simulation system. Current packages dealing with specific interactions usually rely on a large number of one-to-one, fixed-receptor trajectories to emulate "diffusion"; on the other hand, packages simulating multiple proteins tend to ignore specific interactions due to complexity and probably, a lack of significant signal due to the "noise" generated by a higher number of nonspecific interactions.

- Although there are some indications that traditional MD simulations are beginning to expand into the μs territory, it is not anticipated that

such systems will be capable of modelling a diffusive molecular system for quite some time.

- Most existing modelling packages are developed primarily for one or a very limited number of scientific applications. Many of them lack properly-designed infrastructure, and are therefore inflexible and unreliable to be expanded for further use involving modelling a large variety of molecular types and/or environments.

BioSimz, short for "*Bio*logical *Sim*ulation*s*", is developed to address as many of the aspects associated with the above four problems as possible and with proper consideration to the computing power currently available in most academic environments.

2.1 Theories

2.1.1 Dynamics

2.1.1.1 Diffusion and Markov chains

In general, the phenomenon of diffusion describes the fluctuating motion of diffusing particles in liquid and gas phases. The diffusional displacement of a particle has its origin in the enthalpy of the particle, while its random trajectory is the result of the molecule changing courses after frequent collisions with other molecules undergoing the same thermal motion. In a solvated environment, these trajectories of solute molecules or larger particles, such as those of the pollen grains observed by British botanist Robert Brown in 1827, are mainly determined by their collisions with solvent molecules, such as water.

Solvent collisions occur so often that given any short period of time, there are almost always more than one solvent molecule hitting a diffusing particle. Under ideal diffusion conditions, where no two particles interact with each other in the system, the displacement of any diffusing particle can be expressed in terms of a first-order Markov chain if the duration time of the diffusion is sliced into n discrete steps named "timesteps",

$$p(\mathbf{x}_n, t_n|\mathbf{x}_{n-1}, t_{n-1}; \mathbf{x}_{n-2}, t_{n-2}; \cdots; \mathbf{x}_0, t_0) = p(\mathbf{x}_n, t_n|\mathbf{x}_{n-1}, t_{n-1}), \tag{2.1}$$

where $\mathbf{x}$ is the displacement vector of the diffusing particle, t is the time line, and $p(\cdot|\cdot)$ is the conditional probability. The elimination of terms $\mathbf{x}_{n-2}, t_{n-2}; \cdots; \mathbf{x}_0, t_0$ marks the Markovian property, with which the displacement vector is memory-less, only depending on the state of $\mathbf{x}$ at the last timestep. To satisfy the Markovian property, individual collisions between solvent molecules should have only minimal effects on the momentum of the diffusing particle; for example, the collisions between solvent molecules are not Markovian as their trajectories can be (almost) completely deterministic once the initial conditions are given.

2.1.1.2 Fokker-Planck and the Einstein diffusion equations

Let $\delta t = t_n - t_{n-1}$ in Eqn. 2.1, we can then write the conditional probability of displacement vector $\mathbf{x}$ for two timesteps, resulting in the Chapman-Kolmogorov equation,

$$p(\mathbf{x}_n, t + \delta t|\mathbf{x}_{n-2}, t - \delta t) = \int p(\mathbf{x}_n, t + \delta s|\mathbf{x}_{n-1}, t) p(\mathbf{x}_{n-1}, t|\mathbf{x}_{n-2}, t - \delta t) d\mathbf{x}. \tag{2.2}$$

This is effectively an integral equation for the time evolution of the displacement probability. Solving the integral equation gives the Kolmogorov forward equation,

$$\frac{\partial p(\mathbf{x}_n, t|\mathbf{x}_0, t_0)}{\partial t} = -\frac{\partial}{\partial \mathbf{x}}(A(x)p) + \frac{1}{2} \cdot \frac{\partial^2}{\partial \mathbf{x}^2}(B(x)p), \tag{2.3}$$

which is also called the Fokker-Planck equation (Fokker, 1914; Planck, 1917). The LHS is the rate of change in displacement probability with respect to time, and the RHS shows that two factors contribute to the rate changes. The first term containing $A(x)$ is a first order derivative of the probability, representing the *drift*, or convection, of the moving particles due to systematic forces. The second term, a second-order derivative of some position-specific function $B(x)$ with p, represents the *diffusion*, or fluctuation, of the moving particles, corresponding to the residual average effect of the ran-

dom displacements in all directions.

Let us consider a diffusion-only process, then Eqn. 2.3 can be simplified to

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = \mathbf{D}\nabla^2 p(\mathbf{x}, t), \tag{2.4}$$

where ∇ is the three-dimensional gradient operator and $\mathbf{D}$ is the diffusion tensor. For simplicity, assume the diffusing particles are spherical. The tensor is therefore constant in all orientations such that a diffusion coefficient, D, can be used. Eqn. 2.4 thus becomes the Einstein diffusion equation. Solving this equation by multiplying x^2 on both sides and applying Green's theorem, we have

$$\langle x^2 \rangle = 6Dt, \tag{2.5}$$

for the mean-squared displacement $\langle x^2 \rangle$ in translational diffusion.

2.1.1.3 Diffusion and friction

In an aqueous environment, the diffusion coefficient D is found to be proportional to the solvent friction factor (Einstein, 1905; von Smoluchowski, 1906), as given by

$$D = \frac{k_B T}{\zeta}, \tag{2.6}$$

for all diffusing particles irrespective of their shapes and hydrodynamic properties, where ζ is the friction coefficient. Using Stokes' law, we have the Stokes-Einstein relationship

$$D = \frac{k_B T}{6\pi\eta r}, \tag{2.7}$$

where η is the solvent viscosity and r is the diffusing particle's radius. For rotational diffusion of spherical objects, a similar relationship stands,

$$D_{\text{rot}} = \frac{k_B T}{\zeta_{\text{rot}}} = \frac{k_B T}{8\pi\eta r^3}. \tag{2.8}$$

2.1.1.4 The Langevin equation

The diffusion process was also modelled by French physicist Langevin using a combination of deterministic and stochastic dynamics terms, which effectively linked the microscopic dynamics of Einstein's work and the macroscopic statistical theory initiated by Smoluchowski. In Langevin's theory, diffusion dynamics of individual particles is expressed as a stochastic differential equation (SDE) based on Newton's second law of motion. Taking the example of the translational diffusion of a sphere, the Langevin equation states

$$m\frac{d\mathbf{v}}{dt} = -\zeta\mathbf{v} + B(t), \tag{2.9}$$

where $\mathbf{v}$ is the instant velocity of the particle at time t, $\zeta\mathbf{v}$ represents the solution drag at velocity $\mathbf{v}$, and $B(t)$ is a random fluctuation force that originates from the collision impacts with solvent molecules. Solving the SDE using the equipartition theorem gives

$$\langle \mathbf{x}^2 \rangle = \frac{6k_BT}{\zeta} \cdot t + \frac{6mk_BT}{\zeta^2}\left(e^{-\zeta t/m} - 1\right). \tag{2.10}$$

For large $t \gg \delta t$, the second term diminishes, leaving

$$\langle \mathbf{x}^2 \rangle = 6 \cdot \frac{k_BT}{\zeta} \cdot t = 6Dt, \tag{2.11}$$

which agrees with the solution of the Einstein diffusion equation (2.5). Using the fluctuation-dissipation theorem (Kubo, 1966), this gives

$$\langle B(t) \rangle = 0, \tag{2.12}$$

$$\langle B(t_1)B(t_2) \rangle = 2k_BT\zeta\delta(t_1 - t_2), \tag{2.13}$$

which define the stochastic force term of the Langevin equation (2.9).

2.1.1.5 Distribution of velocities

For simplicity, assume the diffusion process is in one-dimension at velocity v, the solution to Eqn. 2.9 has the form

$$v = v_0 e^{-\gamma t} + \frac{1}{m}\int_{t_0}^{T} B(t) e^{-\gamma(T-t)} dt, \tag{2.14}$$

where $\gamma = \zeta/m$ is usually called the damping constant. It is apparent from this solution that γ is the decay factor for the particle's velocity, showing that the solution friction is essentially the result of massive collisions between a diffusing particle with solvent molecules. For the same reason, γ is sometimes referred to as the collision frequency. Consequently, the inverse of γ, usually denoted as $\beta = 1/\gamma$, is often interpreted as the averaged particle "relaxation time", the period of time between a particle colliding with a solvent molecule to re-approaching its velocity equilibrium.

Solvent collisions with a diffusing particle may occur in any direction, or in the case of one-dimensional diffusion, in either a positive or a negative direction. Therefore, we have

$$\langle B(t) \rangle = 0. \tag{2.15}$$

Hence for the particle's average velocity, we have

$$\langle v(t) \rangle = v(0) e^{-\gamma t}, \tag{2.16}$$

and for its mean-squared velocity

$$\langle v^2(t) \rangle = v^2(0) e^{-2\gamma t} + e^{-2\gamma t} \int_{t_0}^{T} \int_{t_0}^{T} e^{(t_1+t_2)\gamma} B(t_1) B(t_2) dt_1 dt_2, \tag{2.17}$$

where the cross term is eliminated since $\langle v(0)B(t) \rangle = 0$. Combining Eqns. 2.16, 2.17 and using the equipartition theorem, the variance of v is given by

$$\sigma_v^2(t) = \langle \langle v^2(t) \rangle - \langle v(t) \rangle^2 \rangle = \frac{k_B T}{M}(1 - e^{-2\gamma t}). \tag{2.18}$$

Independently we know that the particle velocity over a long period, should the diffusing particles be molecules, conforms to the Maxwell-Boltzmann distribution, of which we give the one-dimensional form here

$$p(v_e) = \sqrt{\frac{m}{2\pi k_B T}} \exp\left(-\frac{1}{2} \cdot \frac{m v_e^2}{k_B T}\right). \tag{2.19}$$

where v_e is the molecular velocity once the equilibrium of the energy distribution is reached. Consider the following limits that hold at an equilibrium state,

$$\lim_{t\to\infty} v_0 e^{-\gamma t} = 0, \tag{2.20}$$

$$\lim_{t\to\infty} \left(1 - e^{-2\gamma t}\right) = 1, \tag{2.21}$$

and insert them into the above Maxwell-Boltzmann distribution function (Eqn. 2.19), gives the conditional form

$$p(v, t|v_0, 0) = \sqrt{\frac{m}{2\pi k_B T(1 - e^{-2\gamma t})}} \exp\left(-\frac{m(v - v_0 e^{-\gamma t})}{2k_B T(1 - e^{-2\gamma t})}\right). \tag{2.22}$$

This is the probability density function of a normal distribution of the following form,

$$N\left(v_0 e^{-\gamma t}, \frac{k_B T}{m}(1 - e^{-2\gamma t)})\right). \tag{2.23}$$

The variance σ_v^2 of the above distribution agrees with the result of Eqn. 2.18.

2.1.2 Forcefield

In a real macromolecular environment, pairwise and multi-body interactions, can significantly influence the diffusive behaviour of individual molecules. Therefore, the classical Fokker-Planck equation, that represents the position of a particle as a probability density function, becomes unable to represent microscopic events; the Langevin equation, however, can be easily extended to include all interaction forces and efficiently account for the translational diffusion of a particle:

$$m\frac{dv}{dt} = F(t) - \zeta v + B(t), \tag{2.24}$$

where $F(t)$ is the total inter-molecular force exerted on the molecule. In essence, the interaction force can be written as

$$F(t) = F_{\text{hydro}} + F_{\text{VdW}} + F_{\text{elecs}} + F_{\text{desol}}, \tag{2.25}$$

where F_{hydro} accounts for the hydrodynamic effect on the molecule, F_{VdW} is the van der Waal's force, F_{elecs} is the electrostatic attractive or repulsive force and F_{desol} accounts for the desolvation effect, as discussed in Section 1.1.3.2. In this study, for practical reasons, a modified force term is used,

$$\boldsymbol{F}_{\text{tot}} = \boldsymbol{F}_{\text{L-J}} + \boldsymbol{F}_{\text{Coul}} + \boldsymbol{F}_{\text{H-bond}} + \boldsymbol{F}_{\text{ACE}}, \tag{2.26}$$

which includes the Lennard-Jones (LJ) VdW term, the Coulomb electrostatic term, a hydrogen-bonding term and a desolvation term.

The $\boldsymbol{F}_{\text{VdW}}$ term is expressed by taking the negative gradient of the Lennard-Jones potential of the following form,

$$\boldsymbol{F}_{\text{L-J}}(r) = -\nabla V_{\text{L-J}}(r) = -4\epsilon \left(\frac{12\sigma^{12}}{r^{13}} - \frac{6\sigma^{6}}{r^{7}} \right) \hat{\boldsymbol{r}}, \tag{2.27}$$

where $\hat{\boldsymbol{r}}$ is the normalised unit vector, while values of ϵ and σ are both taken from the CHARMM27 (MacKerell et al., 1998b) topology profile. The LJ parameters have been calibrated to reproduce energies in agreement with CHARMM27 VdW energies. Due to the use of dynamic timesteps, atom mass dependent ceiling functions are used to cap excessively large vdW potentials.

The electrostatic term, $\boldsymbol{F}_{\text{Coul}}$, uses the Coulombic potential function with variable dielectric constant ϵ, which is a linear function with respect to the distance d between the two charges in the range, 4.0Å $< d <$ 9.0Å. Correspondingly, $\epsilon_{\text{min}} = 6$ and $\epsilon_{\text{max}} = 78$. A quartic function is applied at each end of this linear function, to smooth out discontinuities, and an empirically parameterised partial charge is assigned to each atom according to CHARMM27. For PDB structures that have no hydrogen atoms stored, H-atoms are generated using dihedral angle information from the CHARMM27 topology profile.

The hydrogen bonding term, $\boldsymbol{F}_{\text{H-bond}}$, is a 6-4 potential that takes account of the additional affinity between H-bonding atoms not fully modelled by other forcefield components. To reflect the sensitive coupling between H-bond strength and angle, a quartic cosine function is attached to the 6-4 potential. In addition to the minimum and maximum distance cut-offs for fil-

tering out unphysical or negligible potentials, a bond-angle cut-off is chosen such that $\theta > \arccos(-0.9)$. For qualifying atoms, the hydrogen-bonding force term is calculated by

$$\boldsymbol{F}_{\text{H-bond}}(r) = -\nabla V_{\text{6-4}}(r) = -\epsilon_{\text{H}} \left(\frac{\sigma^6}{r^7} - \frac{\sigma^4}{r^5} \right) \hat{\boldsymbol{r}} \cos^4 \phi_{\text{bond}}, \tag{2.28}$$

where ϵ_{H} and σ are scaling constants calibrated to produce energies in line with experimental data, while ϕ_{bond} is the bond angle. The H-bonding term is calibrated such that, combined with existing LJ and Coulomb terms, it reproduces the potential values in agreement with those of experimentally measured hydrogen bonds.

The desolvation term, $\boldsymbol{F}_{\text{ACE}}$, is incorporated here to take account of the solvation free energy incurred when atom-water contacts are replaced by atom-atom contacts. The atomic energy (ACE) values, tabulated by Zhang et al. (1997), classify protein atoms into 18 different types that generate 162 pairwise solvation potentials. The method was calibrated on 9 protein complexes that have experimentally measured binding free energies; a correlation coefficient of 0.7 was achieved between experimental and calculated ACE-based, binding free energies. The ACE energies used in this study are subject to a 3rd-order decay w.r.t the distance between probing atoms. By considering ACE energies, VdW and electrostatic potentials could be considered to be double counted between atoms within contact distance. As a counter balance, adjustments have been made to the Coulomb potential functions so that they decay when probing distances fall below the atomic contact threshold, at the same rate as the ACE energies are weighed in.

For rotational rigid-body dynamics, the total torque is calculated as follows

$$\boldsymbol{\tau} = \sum_{i=1}^{n} \boldsymbol{\tau}_i = \sum_{i=1}^{n} \boldsymbol{F}_i \times \boldsymbol{r}_i \tag{2.29}$$

where $\boldsymbol{F}_i$ is the total force exerted on atom i of the molecule under consideration, and $\boldsymbol{r}_i$ is the distance from the atom to the molecule's rotating centre, i.e. centre of mass. The total torque is then plugged into the rotational Langevin equation (see Eq. 4.5).

2.1.3 Simulation

2.1.3.1 Rigid-body dynamics

Unlike conventional MD simulations, rigid-body dynamics is employed to handle translational and rotational movements of proteins, which are, from a rigid-body point of view, objects of irregular shapes and are composed by point masses (atoms). The collective motion of these point masses can therefore be investigated with respect to a representative point which, by convention and convenience, is usually the centre of mass of the object,

$$\mathbf{x}_{\text{cm}} = \frac{\sum_i m_i \mathbf{x}_i}{\sum_i m_i}, \tag{2.30}$$

where x_i and m_i are the position and mass of the i-th atom, respectively. The total mass and linear momentum of the molecule (protein) are therefore

$$m_{\text{total}} = \sigma_i m_i, \tag{2.31}$$

$$\mathbf{p}_{\text{cm}} = m_{\text{total}} \times \mathbf{v}_{\text{cm}}, \tag{2.32}$$

where v is the linear velocity, i.e. speed of displacement of a molecule in space with respect to its centre of mass. The change of linear momentum with respect to time is

$$d\mathbf{p}_{\text{cm}} = m_{\text{cm}} \cdot d\mathbf{v}_{\text{cm}} = m_{\text{cm}}\mathbf{a} \cdot dt = \mathbf{f}_{\text{cm}} dt, \tag{2.33}$$

where the total force with respect to the centre of mass, $\mathbf{f}_{\text{cm}}$, is simply the summation of all forces exerted on each of the atoms

$$\mathbf{f}_{\text{cm}} = \sum_i \mathbf{f}_i, \tag{2.34}$$

which accounts for the molecule's linear movement. The differential expression for updating linear velocity for rigid molecules is

$$\frac{d\mathbf{v}}{dt} = \frac{\mathbf{f}_{\text{cm}}}{m_{\text{total}}}, \tag{2.35}$$

following Newton's second law of motion. The angular movement, on the other hand, arises from the total torque applied to the molecule. The total

torque with respect to a molecule's centre of mass is the sum of all component torques felt by the atoms,

$$\boldsymbol{\tau}_{\text{cm}} = \sum_i \mathbf{r}_i \times \mathbf{f}_i, \tag{2.36}$$

where $\mathbf{r}_i = \mathbf{x}_i - \mathbf{x}_{\text{cm}}$. Analogous to the definition of linear momentum (see Eqn. 2.33), a similar relationship exists between angular momentum of a rigid molecule, $\mathbf{L}$, and its rotational velocity, $\boldsymbol{\omega}$,

$$d\mathbf{L} = d\left(\mathbf{I}_{\text{cm}} \cdot \boldsymbol{\omega}\right) = \boldsymbol{\tau} dt, \tag{2.37}$$

where $\mathbf{I}$, the 3×3 inertia tensor, changes with respect to the axis of rotation and is therefore also a function of t. Differentiating $\mathbf{L}$ with respect to t, we have

$$\begin{aligned} \boldsymbol{\tau} &= \frac{d\mathbf{L}}{dt} && (2.38)\\ &= \frac{d\mathbf{I}}{dt}\boldsymbol{\omega} + \mathbf{I}\frac{d\boldsymbol{\omega}}{dt} && (2.39)\\ &= \boldsymbol{\omega} \times \mathbf{I} \cdot \boldsymbol{\omega} + \mathbf{I}\frac{d\boldsymbol{\omega}}{dt}. && (2.40) \end{aligned}$$

Therefore the angular acceleration, analogous to its linear counterpart $\mathbf{a} = \mathbf{f}/m$, is written as

$$\frac{d\boldsymbol{\omega}}{dt} = \mathbf{I}^{-1} \cdot \left(\boldsymbol{\tau} - \boldsymbol{\omega} \times \mathbf{I} \cdot \boldsymbol{\omega}\right), \tag{2.41}$$

where the inertia tensor $\mathbf{I}$ is calculated by

$$\mathbf{I} = \mathbf{A} \cdot \mathbf{I}_0 \cdot \mathbf{A}^T. \tag{2.42}$$

Unlike $\mathbf{I}$, $\mathbf{I}_0$ is a fixed property, like molecular weight, for each rigid molecule. Tensor $\mathbf{I}_0$ is often chosen such that it has the form of a diagonalised matrix, i.e. containing three principal axes pointing from the molecule's centre of mass towards I_{00}, I_{11}, I_{22} respectively. Matrix $\mathbf{A}$, in the above equation, is the 3×3 rotation matrix that records the current rotation with respect to the initial posture aligned to the principal axes. The equivalent property of $\mathbf{A}$ in translational motion is the position vector $\mathbf{x}_{\text{cm}}$.

Combining Eqns. 2.35 and 2.41, we have the full rigid-body dynamics (also named the Newton-Euler dynamics) for the molecules in analytic forms.

2.1.3.2 Integration of paths

The full analytical form of a Langevin-based diffusion and interaction scheme is presented above; this section describes discretisation and subsequent numerical calculations to implement the scheme. The translational Langevin equation (Eqn. 2.24 is a first-order SDE w.r.t. the translational velocity v and takes the form

$$v' = f(v, t), \tag{2.43}$$

with initial value

$$v(t_0) = v_0. \tag{2.44}$$

A number of numerical integration schemes are available on different accuracy and complexity levels (Butcher, 2003), such as the Euler method (error in first-order, unstable), Verlet-based methods (error in second-order, stable, Verlet, 1967) as well as the class of Runge-Kutta schemes (locally fourth-order, stable). In this study, the midpoint method is chosen for its simplicity and relatively good approximation to exponential decay problems. Midpoint is one of the second-order Runge-Kutta integrating methods and has been shown to be superior to the classical 4th-order Runge-Kutta method for Langevin SDE problems (Vercauteren, 2005). It takes the form

$$v(t + \Delta t) = v(t) + \Delta t f\left(v\left(t + \frac{\Delta t}{2} f(v(t), t)\right), t + \frac{\Delta t}{2}\right), \tag{2.45}$$

where Δt is the timestep duration. Numerical results, from simulations based on this scheme, are shown and discussed in Section 2.3.2.

Due to discretisation error, individual interaction events will occasionally incur excessively large VdW and electrostatic potentials, which would transiently assign unrealistically huge amounts of momenta to the interacting molecules, leading the simulation to "explode". To deal with this, a

collision inspector has been constructed to check the translational and rotational velocities for each interaction event; should there be a velocity v such that $p(v) < 0.01$ from the corresponding Maxwell-Boltzmann p.d.f., a new velocity sampled from the p.d.f. will replace the erratic speed with a probability of $1 - p(v)$.

For crowded molecular environments with a large occupancy value (the volume of proteins occupy more than 20% of the volume of the solution), initial displacements of macromolecules in the simulation box, in a random manner, is likely to incur considerable overlap between adjacent molecules. To resolve this issue, a short, high-temperature, VdW-only simulation is used as a "distribution" procedure. Once all the molecules are distributed, in an unbiased fashion, throughout the simulation box, they are assigned linear and angular velocities, generated from the corresponding Maxwell-Boltzmann distribution at the designated simulation temperature. Before the production run, to balance the kinetic energy of the system, a second short equilibrium run is performed, with all components of the full force-field employed. For each simulation experiment, results were accumulated from each of three independent runs - starting from the same equilibrium state - and averaged.

Typical values for the key parameters used in the simulations are summarised in Table 2.1.

2.1.4 Readout

2.1.4.1 Structural assessment

To assess the quality of the modelled specific protein-protein interactions from a structural perspective, BioSimz uses translational, orientational and rotational thresholds, as well as all-atom root-mean-squared deviation (RMSD) between the structure of the encounter complex in question and the known structure of the crystal complex. The translational threshold is the maximum linear distance permitted between the centres of mass (CM) of the target and reference ligands when receptors are superimposed. The orientational threshold is the maximum in square-root deviation between the target and reference orientation vectors, which originate from the CM of the receptor to that of its ligand. The rotational threshold is the maximum

Table 2.1: Simulation setup & typical parameters, see Model section for detailed description.

Index	Property	Setting
1	Simulation box	$240 \times 240 \times 240 Å^3$ with periodic boundaries
2	No. target molecules	2∼29 for each receptor and ligand)
3	Environmental crowders	10∼20, 10 different types of enzymes
4	Timestep	1 picosecond
5	Temperature	310K
6	Typical simulation length	$0.5\mu s \times$ 10∼20 runs $\times$ 3 repeats
7	Forcefield	LJ & Coulomb, H-bond, desolvation & Brownian terms
8	Solvent	implicit water (variable dielectric constant $\epsilon = 6 \sim 78$)
9	Ionisation	pH=7.0 at 10mM, 50mM & 150mM, ions explicitly modelled
10	Solvent viscosity	0.71 mPa·s
11	Collision handling	conservation of momenta (all) & energy (elastic only)

angle allowed for the ligand to rotate away from its reference rotation matrix about an axis passing through the CM of the ligand. In simulation, the above three values are set to 14Å, 0.9 and 2.3rads respectively. It should be noted that the linear distance and self-rotation terms are relatively relaxed, whereas the orientation requirement is stricter; this is to reflect the fact that correct orientation is more important for the formation of encounter complexes. Protein-protein interactions that satisfy the above criteria are subject to a further RMSD check, to ensure they are structurally close to the experimentally defined complex state, before being identified as an 'on' event. It is widely adopted, originally by the committee of the Critical Assessment of PRedicted Interactions (CAPRI) (Janin, 2002), that a complex bearing an RMSD equal or smaller than 10Å from the crystal complex should be treated as having an "acceptable" quality; BioSimz employs this RMSD criterion to define whether a molecular contact is specific or not.

2.1.4.2 Binding scores

For most protein-protein interactions, the half-life of an established protein-protein complex is far longer than can be possibly simulated effectively at the atomic level, even if computer capacity was unlimited; from millisecond to a matter of days (barnase and barstar) or weeks (pancreatic trypsin and BPTI). Therefore, while BioSimz is capable of modelling the dynamics of both encounter complex formation and dissociation, it has to do so through separate simulations. This study focuses primarily on k_{on}, as the exact molecular mechanism for unbinding remains largely unknown. In simulation, an interaction is defined as the state and period for which two molecules have one or more atoms in contact (< 5Å). The retention time of an interaction event is therefore defined as the period from which the first pair of contacts is established until the last pair of atomic contacts breaks. A specific interaction event begins from the first moment (timestep) that the molecule pair/group satisfies the docking criteria (described in the previous subsection) with a known reference complex and ends when the criteria are no longer met. An interaction is defined nonspecific if the above criteria are not met throughout its lifetime.

A scoring scheme has been developed, here referred to as the binding score, to evaluate the quality of specific interactions. The scoring scheme

awards higher marks to interaction events that have longer retention times at specific binding sites. A simple four-section linear function is used, in which retention times between 1ps and 760ps score linearly from 1 to 40, those between 760ps and 2.08ns score linearly from 40 to 200, and those between 2.08ns and 4ns score 200 linearly from 200 to 400. To achieve a high score under this scheme, proteins need either a large number of median retention-time events, or a small number of long (stable) events. Mapping the strategies to mechanisms, essentially means that molecules either form more encounter candidates to compensate for the relatively low successful rate of them acquiring the final specific complex conformational state, or they have a smaller number of high-quality encounters that bear a larger chance of survival with penultimate formation of the native complex conformation. Interestingly, these binding strategies are analogous to the reproduction strategies of fish and mammals, either replicating by quantity or by quality (fitness) respectively. It is thereby reasonable to assume protein interactions acquire these binding strategies through selection pressure in the evolutionary process.

A detailed discussion of the binding scores is presented in Section 4.4.1 of this thesis.

2.2 Implementation

The models described in Section 2.1 are implemented in a standalone, fully-functional multi-molecular simulation package, i.e. BioSimz. As stated at the beginning of this chapter, until the work reported here, no simulation software performing atomic multi-molecular simulations in a box while investigating specific protein-protein interactions has been developed. The BioSimz project was initiated to fill this vacancy; at the time of writing this thesis, the author is still not aware of any other package capable of simulating multiple specific protein interactions simultaneously at atomic resolution.

A number of design and implementing principles were maintained throughout the development process, which ensured that BioSimz will provide unique molecular simulating capabilities that are currently beyond other modelling and simulating software. These principles are as follows:

- High-performance. The modelling design must be able to accommodate molecular systems containing not fewer than 1 million atoms. The simulation design should be parallelised to make use of multi-core CPUs and should deliver simulation results within a matter of days when given a CPU-intensive task, such as the simulation of a crowded macromolecular environment close to physiological concentration.

- Accuracy. Computation should be made as accurate as possible. This includes the choice of forcefield resolution (united atom or full atom, unit charge or partial charges), the regime of numerical integration, and the handling of rigid-body rotational dynamics.

- Ease of use. The number and types of simulating molecules, as well as all adjustable parameters should be made available through configuration profiles, and the constructed of such profiles should be further automated for batch processing.

- Scalability. The package should cope equally well with both dilute and crowded molecular simulations and should not waste CPU clock cycles or memory storage in regions of low molecular density.

- Reliability. The simulations shall not produce thermodynamic artifacts under both normal and crowded molecular concentrations. Molecules of irregular shapes should still be correctly handled. The package should be able to accept non-standard PDB data entries. The simulation software should not contain technical deficiency that may impede the reliability of running long simulations in large numbers.

- Extensibility. The software library should make the necessary reservation for future expansion. The overall design should take into consideration of the need for further algorithmic development, particularly algorithms to enable internal flexibility of the macromolecules.

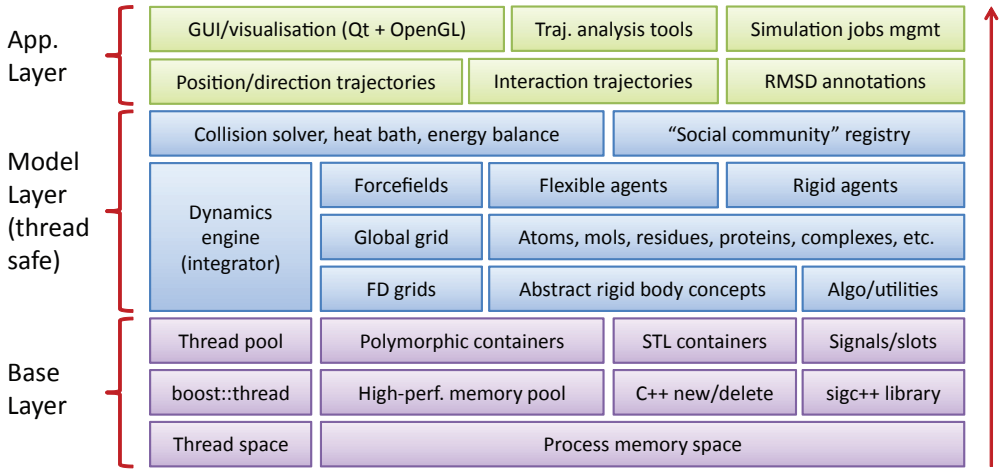

Figure 2.1: Overview of three-layer structure of the Biosimz library.

2.2.1 Framework Design

2.2.1.1 Code structure

BioSimz is built as a multi-threaded C++ library. The library is broadly organised in three layers: the base, model and application layers. Figure 2.1 illustrates the main modules within each of the three layers.

The base layer was constructed as a language extension on top of existing C++ facilities. The current "C++98/03" standard (Iso14882, 1998) has greatly expanded and standardised C++ from C, with support of classes, dynamic typing, templates and the introduction of the Standard Template Library (STL). However, it cannot yet deliver optimal performance for managing and operating small objects in their millions, which are required for large-scale macromolecular simulations. Therefore, polymorphic container classes, coupled with flexible memory pools, have been designed and implemented to supplement the lack of high-performance large-scale containers in C++ and its peripheral libraries (see Section 2.2.2). A number of design patterns were also implemented, such as singleton and factory (Gamma et al., 1994), to provide a universal interface for accessing objects of different types. The C++ Boost library (Abrahams and Gurtovoy, 2004) is also extensively used in the base layer, providing support for vector/matrix arithmetics, threading, file reading/writing and utility templates. Two external

libraries, `threadpool` and `sigc++`, were used to deliver thread pooling and event handling support. Both libraries are implemented as class templates, which usually have minimum running overheads as they are interpreted at compile-time.

The model layer was built on top of the base layer with no external dependencies. All atomic and molecular models, forcefield models, numerical integrators and grid models are located on this layer. The layer is designed such that applications could be written based on models from this layer without manipulating base-layer classes and objects, relieving application writers from having to know the low-level details.

The application layer is composed of a few standalone applications and some convenience library classes, such as trajectory management and statistical analysis. Currently, three applications are developed on top of the library: two command-line applications for running simulations and performing trajectory analysis, and a Qt-based graphical user-interface (GUI) to manage and display, in three-dimensions along with some annotation, molecular trajectories. As the application layer is well separated from the model and the base layers, making changes to these two layers will not require the applications to be modified accordingly.

The library and application framework is further supported by standalone applications and shell scripts for the creation, submission and management of individual simulation runs. A brief summary of BioSimz code statistics is shown in Table 2.2, which offers a glimpse of the size scale of the package from a technical aspect. Also it is worth noting that many of the BioSimz classes were implemented as class templates; all efforts have been made to re-use existing code where available.

2.2.1.2 Functional structure

The three core components of the BioSimz library responsible for conducting macromolecular simulations are: molecular systems, the object grid and forcefields. During each simulation timestep, information on the state variables of the simulated objects (proteins) is passed from systems to the grid, from the grid to forcefields, and then from forcefields to systems, completing the circle. Various data recording modules can be attached to this circular information flow for trajectory recording, analysis and visualisation.

Code Type	BioSimz	BioSimzApp	BioSimzStats	BioSimzLab	Total
C++ Sources	21,356	132	1,853	33,984	102,564
C++ Headers	42,947	3	83	2,206	
Makefiles	17,737	20	20	978	18,755
Comments	35,857	102	631	3,695	40,285
Total	117,897	257	2,587	40,863	161,604
No. of Files	665	3	23	97	788

Table 2.2: BioSimz code statistics. Numbers quoted here are lines of code (LOC), except for the last row. All blank lines in source files were removed from being counted in. The statistics was harvested by code metrics tool `cloc`. BioSimzApp, BioSimzStats and BioSimzLab are executable applications built on top of the BioSimz library for running simulations, performing statistical analysis and 3D visualisation of trajectories and annotations, respectively.

Module	Components	Module	Components
Algorithms	17	Mesh Operations	16
Concepts	11	Memory Management	10
Configurations	4	Molecular Models	31
Data Handling	50	Motion Schemes	6
Extended Datatypes	15	Software Patterns	42
Dynamics	11	Class Policy	1
Exception Handling	12	Simulations	13
Forcefields	13	Communities	6
Geometry	6	Solvent Models	1
Grid Operations	28	Class Traits	19
Interactions	9	Statistics	21
Class Interfaces	13	Visualisation	14
Total Modules	24	Total Components	369

Table 2.3: A list of BioSimz modules. "Components" are the number of C++ `struct` and `class` structures in each of the 24 modules.

The main working loop inside the BioSimz simulating engine is shown in Figure 2.2.

2.2.2 Module Designs

BioSimz has been constructed as a large-scale modelling and simulation library; therefore, the functional units of the library are organised in modules (namespaces in C++), each of which contains a number of classes that are functionally related to each other. A total number of 24 modules were constructed in the simulation package, as is listed in Table 2.3. This section discusses a few key module and class designs that have been vital to the library's structure and performance, and that have not previously been invented or applied in the field of molecular modelling and simulation.

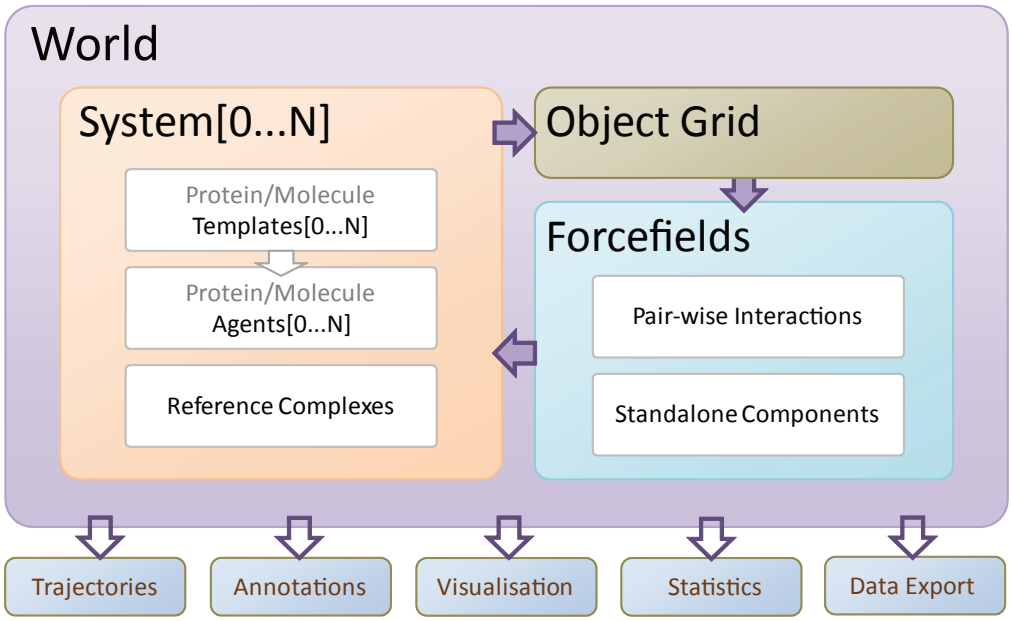

Figure 2.2: Overview of data and function flows in BioSimz.

2.2.2.1 Polymorphic containers

The container/iterator pattern (Gamma et al., 1994), widely used in C++ standard and Boost libraries, appears a natural choice for manipulating molecules, i.e. container of atoms. However, container classes in the standard library (`std::vector<T>`, `std::list<T>`) do not suffice the representation of molecular models for BioSimz:

- Both vectors and lists use the operators, `new,delete` to dynamically allocate and recycle memory resources as elements are added or removed from the container. For a small container computational costs for these operators is negligible, but its performance drops quickly when containers expand to, for example, tens of thousands of atoms. Vanilla C/C++ arrays are quick, but they are usually fixed in size and prone to boundary errors.

- Standard containers do not support polymorphic copying. For example, `Molecule` and `Protein` are two classes and the latter is derived from the former. Now, let us assume a Molecular `System` class has been created, that contains a molecule list, in which there may be both

proteins and molecules. Had `std::vector<Molecule>` been used, all instances of `Protein` would have been sliced to the size of `Molecule` when the base type copy constructor is called. This is the expected behaviour of normal copy constructors, but it would not be the expected behaviour of the modelling package.

- A common practice in C++ for handling polymorphism in containers is to define a vector of pointers (`std::vector<T*>`). In this author's view, this has somehow defeated the purpose of a container class, since it no longer encapsulates what it is supposed to own. Due to the frequent uses of large containers in molecular modelling software, memory leaks and corruption can be a major source of maintenance problems as cross-reference of objects increases.

- Sometimes it can be confusing whether a container owns the elements, or it just owns the references to the elements. A typical case for protein modelling is as follows: physically, the class `Protein` owns a list of objects of the Atom class; therefore, if a `Residue` object is to link to these `Atom` objects, they have to store the pointers. The calling convention for using an `Atom` from its `Residue` and `Protein` is then different depending on whether it is a pointer or an object; this behaviour is unwanted because it should not be relevant to the library developer.

- Standard containers lack some database-like operations such as indexing, searching, conditional filtering/manipulation, as well as multiple filtering using logical operators.

Therefore, polymorphic container classes, along with self adjustable memory pools, have been constructed, alleviating all of the above problems. Technically, polymorphic copying was realised through a placement `new` operator in the base class. By doing this, the correct size of the memory to copy is directly passed to the `new` operator even in the case of derived (expanded) classes, while no run-time type information (RTTI) overhead is incurred since it does not rely on virtual functions to infer the concrete type of the class.

As was benchmarked with the g++ compiler (version 4.3), the in-house containers achieved a 10 to 100-fold speed up compared to `std::vector<T>`

when pushing back a large number of elements; the larger the size of class `T`, the more pronounced the performance is for the in-house containers. Speed-up is primarily due to memory pooling, which also guarantees continuous memory access for objects in the container. There would also be a significant benefit in efficiency if the container content were to be cloned to elsewhere, for example, video RAM on a GPU card for further processing.

These in-house developed containers are partitioned into two types: one storing the actual content and one storing the pointers. Member functions of both content-storing and reference-storing classes have universal return types so that a library user does not need to be concerned whether the stored element is concrete or a reference. The following code snippet shows the benefit of this design:

```
1
2    Atom ca, cb;
3    Container< Atom > prot_container;
4    RefContainer< Atom > res_container;
5
6    prot_container.push_back( ca );
7    prot_container.push_back( cb );
8    res_container.push_back( ca );
9    res_container.push_back( cb );
10
11   std::cout <<prot_container[0].getPosition() <<std::endl;
12   std::cout <<res_container[0].getPosition() <<std::endl;
```

As is shown here, only the developer who sets up the container needs to think about whether to store the actual objects or their references.

2.2.2.2 Atom and molecular models

In BioSimz modelling and simulation, the number of atoms may reach over 1 million. Therefore, the `Atom` class should be made as small in size as possible to save storage space. The BioSimz `Atom` class is a composition of a three-dimensional vector with associated indices. Where possible, the indices are joined and manipulated through bitwise operators (>> and <<) so as not to waste the memory assigned to the full bit structure of an `int` type. The actual memory layout of the `Atom` class, plotted in Figure 2.3, has a size of 48 bytes per instance.

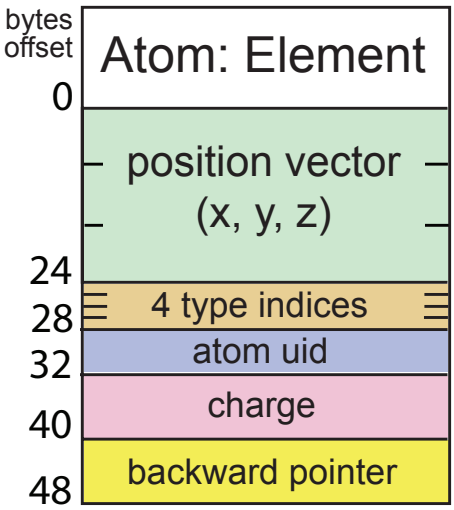

Figure 2.3: Memory layout of the `Atom` class. The four aggregated indices occupying the width of one `int` are chemical element index, atom name index, atom type index and ACP index, respectively.

In reality, all molecules are entities of bonded atoms. In modelling, however, there is no universal method to parameterise all types of molecules from small molecules to proteins and DNA, at least on the scale of molecular mechanics. Therefore, multiple "types" of molecules have to be modelled as derived classes from the `Molecule` parent. Hence, a further concept was introduced, that of a `Measure` concept, which includes a container of atoms plus the collective rigid-body properties of these atoms. The `Molecule` class derives from `Measure`, as well as many sub-molecular atom groups such as residues (see below). Figure 2.4 illustrates the hierarchical structure of molecular classes derived from `Measure` and `Molecule`. Hence, if we define

```
1  Molecule mol ("mol.pdb");
2  Protein prot ("prot.pdb");
```

we can conveniently make a container of molecules to include both objects

```
1  Container< Molecule > mols;
2  mols.push_back( mol );
3  mols.push_back( prot );
```

Subsequent manipulations can then be performed, container-wise, regardless of the molecule type

```
1  mols.act( force_eval( my_ff ) ); // evaluate the force for each
       mol
2  mols.act( movement_eval() );      // evaluate the movement
```

It is therefore possible in BioSimz that a molecular system may consist of many different types of molecular models, as was shown in Figure 2.4. These seemingly complex relationships ensure the operating interface for rigid-body objects is kept consistent between different types of molecules, for example, between a prosthetic group and a multi-chain protein complex. At the same time, these code modifications yield the maximum level of code reuse and thereby improves reliability. Using a similar pattern, the composition of different types of forcefield, or even the inclusion of coarse-grained forcefields, becomes manageable. The above code snippets are for illustrative purposes only; the actual code appears more complex as it deals with multiple issues, some of which have not been discussed here, for example, the encoded facility for treating some molecules as semi-flexible while others remain as rigid-body, nevertheless, the computer coding principles described above remain the same.

2.2.2.3 Communities

Another distinctive feature in BioSimz is the introduction of a social community system for housekeeping of molecular interactions occurred during the simulations. As is discussed in Section 2.1.4.2, time course information of retention of molecular interactions is used to evaluate the strength of specific protein binding. Therefore, it is necessary that the package should have this information recorded during the simulation, so that the time course analysis can be performed and relevant interaction trajectories can be reviewed specifically.

One needs to bear in mind that, unlike the investigation of pairwise interactions, simulation of multiple proteins interacting may encounter situations where multiple complexes can form simultaneously. Traditional ways of detecting complex formation cannot cope well with this; an intuitive resolution may be to use a three-dimensional matrix $N \times N \times T$ for N molecules simulated in T timesteps to describe the pairwise interaction courses. How-

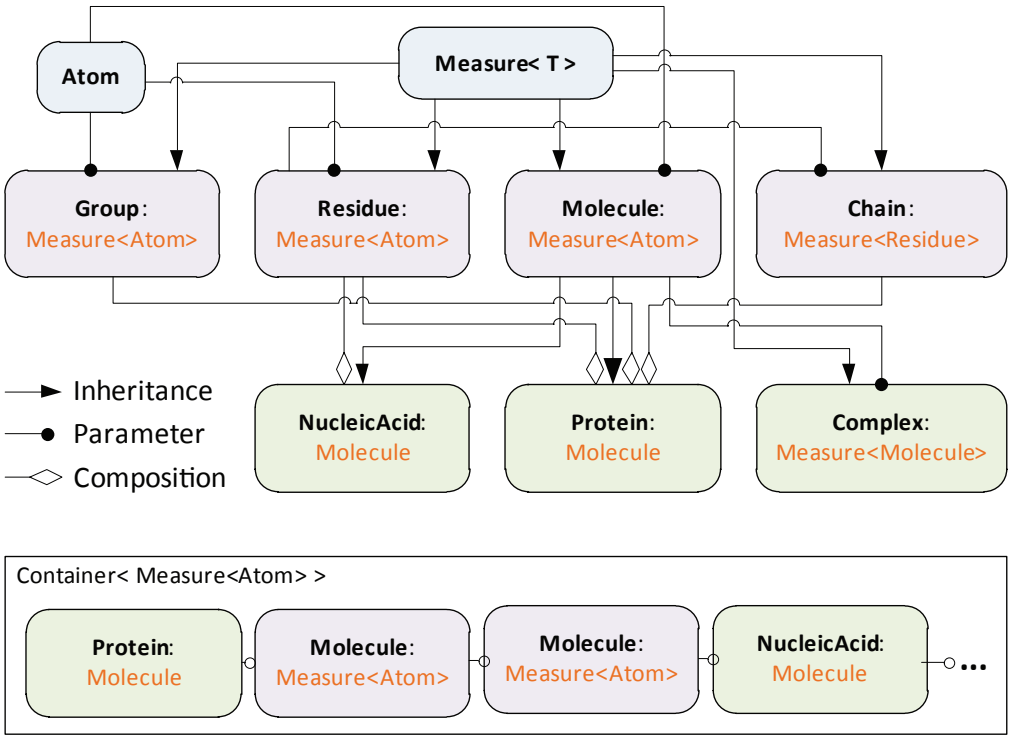

Figure 2.4: Relationship between `Measure` and its derived classes. There are three different types of relationships described in the key to this diagram: inheritance means the arrow-marked class is derived from the connecting class, parameter means the connecting class is a template argument of the solid circle-marked class, composition means the diamond-marked class contains the connecting class as its data member(s). The relationship needs not to be unique; for example, the `Protein` class has its great parent class `Measure<Atom>`, while it also owns several data members who themselves are children classes of `Measure<Atom>`. The bottom box shows a schematic view of a polymorphic container storing three different types of objects, all derived from `Measure<Atom>`.

ever, for large N and T the matrix is beyond normal size limit; yet most of the elements are zeros.

In BioSimz, the problem is solved by the introduction of a social community system with an interaction registry. The World concept, shown in Figure 2.2, corresponds to the `Borough` class which owns and manages a number of `Community` objects. A `Community` is the logical projection of a `System` of molecular models, which keeps a registry of all molecules that belong to the system. Interactions occur within `Members` of the same `Community`, at which time `Neighborhoods` are formed and immediately registered to the `Community`. Once an interaction no long exists, the `Community` will be notified about the departure of the neighbours and subsequently, de-register the `Neighborhood`. While this is a dynamic process during a simulation run, the history of formation and departure of neighbourhoods is always recorded like an event book of a chronicle order.

To efficiently model the above community actions, the C++ template library `sigc++` was used to create an event dispatching/handling mechanism within the simulation library. The novel use of template static variable further eliminated the need for defining a specific signal for every class type, i.e. no penetration into signal-emitting classes. This can be extremely suitable for some cases in BioSimz; for example, a collision solver need not be concerned about which molecular types it currently deals with, and only needs to process all events that are deemed too close.

The social community system is complementary to the Newtonian dynamics simulator: the latter produces a time-series of molecular trajectories in three translational and three rotational degrees of freedom, while the former, as described above, is an event-based registry of association and dissociation of interaction partners. BioSimz is therefore able to take the benefits from both time and event-based modelling of molecular interactions.

2.2.3 Peripherals

As is briefly mentioned in Table 2.2, three standalone applications were developed, on top of the library, to perform simulations, analyse trajectories and visualise results. The types of statistical analysis that BioSimz is capable of performing on simulated trajectories are summarised in Table 2.4. A snapshot of the Qt-based GUI application, `biosimzlab`, is shown in Figure

2.2.3.

2.2.4 Performance

The library is written with multi-threading support. Its computational performance is greatly dependent on the frequency of pairwise interactions occurred in a simulated molecular system. For an *in vitro* style simulation at less than 80g/L, typically $3 - 30\mu s$ trajectories can be generated per day on a quad-core processor. For crowded simulations, such as a 3×10^6 Da system at 330g/L, the performance decreases to approximately $0.5\mu s$ per day. To massively reduce computation time, typically 20 to 30 dual quad-core processors are employed to run multiple simulations in parallel.

Additional attempts have been made to further reduce the simulating time whenever possible. It is observed that under dilute conditions, many simulation circles are wasted in simulating the simple diffusion of target molecules which are still far from even the closest partner. Two improvements are therefore made to consolidate the "wasteful" timesteps, making the simulation scheme effectively operating under a variable-timestep mode:

Prolonging the interval before updating the neighbour list. The molecules are no longer subject to an update of all neighbouring molecules in adjacent grid blocks after *every* timestep. Instead a heuristic algorithm is used to gradually increase the time interval for the update to a molecule's intermolecular contacts until a ceiling time limit is reached. The resulting reduced updating frequency is reset to its original value when a molecule appears in the neighbour list of the target molecule.

Directly increasing the timestep size. This physically reduces the total number of timesteps simulated, albeit with the total amount of biologically time set for each simulation remaining the same. Since this approach affects *every* molecules in the simulation, it can only be triggered when none of the simulated molecules are interacting. Should any two of the molecules fall within a distance threshold, the timestep size immediately reverts to the original setting, in order to prevent collisions that result in molecules overlapping with each other.

Name of analysis	Note
`binding`	Search the trajectory and interaction registry for specific binding events
`rates`	Statistical analysis of the binding events with respect to time course
`diffusion`	Calculate mean-squared displacements for D_{tr} and D_{rot}
`rmsd`	Calculate the interface RMSD changes and dissociation scores
`energy`	Calculate the interaction potentials for specified proteins at specified timesteps
`angveloc`	Calculate linear and angular velocity profiles for simulated molecules
`trace`	Produce a cloud of ligand positions with receptor molecules superimposed
`frequency`	Produce a profile of ligand orientations with the receptors superimposed
`contact`	Produce a contact frequency profile for the receptor, w.r.t. its residues

Table 2.4: Types of analysis performed by BioSimzStats.

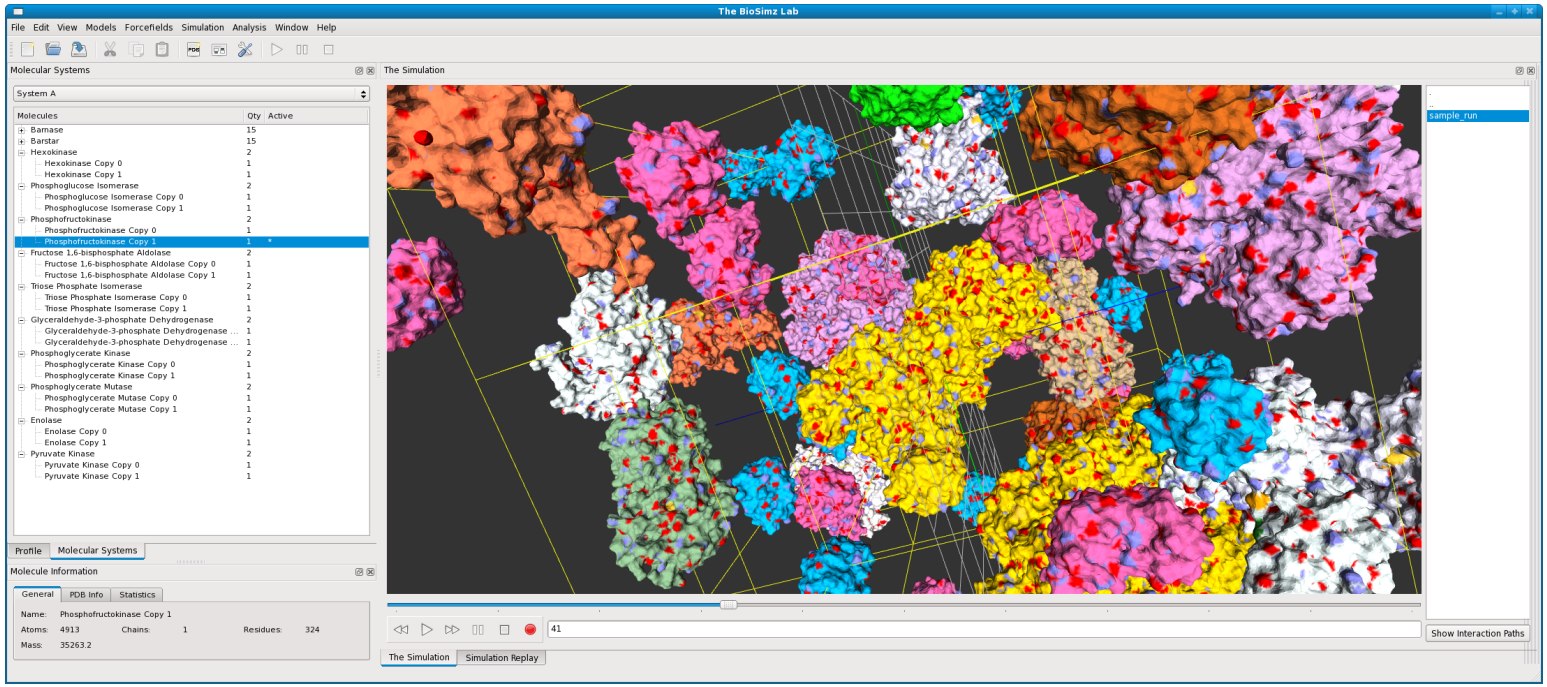

Figure 2.5: Screenshot of the GUI application, BioSimzLab. Molecules on display are from a snapshot of a crowded simulation trajectory, which includes barnase, a ribonuclease and its inhibitor barstar, plus 10 different types of bacterial glycolytic enzymes as macromolecular crowders. Each molecule type is coloured differently. Total concentration of this system is approximately 280g/L.

2.3 Discussion

2.3.1 Justification

Most simulation schemes based on the Langevin equation have to ensure that the systematic force $F(t)$ remains constant within each timestep Δt. Key questions arise as to how long the timestep should be set at, and subsequently, which analytical scheme for integrating the Langevin equation should be chosen. For the first question, most biomolecular simulations schemes for diffusion dynamics have a timestep size ranged from ps (for proteins) to ns (DNAs/nucleosomes), to coincide with the orders of magnitude of the momentum (or velocity) relaxation time, β, which is the inverse of the damping constant γ (see Section 2.1.1.4). The second question is in fact asking for a justification of the choice between using the full Langevin dynamics (LD) or its simplified, displacement-oriented form, the Brownian dynamics (BD).

Ermak (1975) provided the first BD algorithm for large molecules, with a precondition that $\Delta t \gg \beta$. This effectively means that the momentum relaxes to equilibrium much more rapidly than the average timestep, hence the inertia effects of the solutes are negligible. Weiner and Forman (1974) derived another solution suitable for $\Delta t \ll \beta$, while a more generic Monte-Carlo based algorithm for solving the Langevin equation regardless of the values of β was later introduced (Ermak and Buckholz, 1980). Currently, most BD simulations for biomolecules (Elcock, 2003; McGuffee and Elcock, 2006; Gabdoulline and Wade, 1997; Cerutti et al., 2003) are adapted from the Ermak-McCammon algorithm (Ermak and McCammon, 1978), which also bears the precondition $\Delta t \gg \beta$. The reason for the latter to be widely adopted, compared to the Monte-Carlo solution (Ermak and Buckholz, 1980), is the simplicity of the random displacement term,

$$\langle R \rangle = 0, \tag{2.46}$$

$$\langle R^2 \rangle = 2D\Delta t, \tag{2.47}$$

where its auto-correlation function is solely a function of the corresponding diffusion coefficient. In reality, this is largely true for the diffusive be-

haviour of macromolecules in a dilute solution. However, when molecules approach each other for interactions, the BD precondition $\langle m\frac{dv}{dt} = 0\rangle$ no longer holds; moreover, the extent of inertia effect can be large such that the Ermak-McCammon precondition, $\Delta t \gg \beta$ may not hold, either. Nevertheless, this is not be a problem for BD simulations which aim at "arriving and terminating at" the binding interface of two diffusing proteins, as the deterministic dynamics does not kick in, perhaps, until the very end of the these simulations.

In this study, protein-protein binding *dynamics* is treated as being equally important as protein-protein binding kinetics. Therefore, the full LD described in Section 2.1.1 is implemented, with which the force term, i.e. acceleration dv/dt, is explicitly calculated at each time step. This treatment effectively fuses the stochastic LD with rigid-body molecular mechanics, so that the molecular steering dynamics (Blundell and Fernandez-Recio, 2006) is more likely to be observed.

A crowded molecular environment, such as observed under *in vivo* condition, resembles a low Reynolds-number fluid, in which the inertia effect is often negligible. From first appearance it looks like as though a BD model suffices the need for crowded protein dynamics; however, the high viscosity comes from the collisions and interactions among the explicitly modelled macromolecules rather than the implicitly modelled solvent collisions, to which the damping constant in simulation applies. The low Reynolds-number is therefore the *outcome*, rather than an *input* in both simulation and reality. Therefore, the full LD can be justifiably used under crowded simulations of protein-protein interactions, where the viscosity of water $\eta = 6.95 \times 10^{-4}\text{Pa} \cdot \text{s}$ at $T = 310K$; the solvent viscosity (not the apparent viscosity of cytosol). Results show that the reproduced diffusion dynamics agrees with experimental observations for both dilute and crowded solutions (see Section 2.3.2.2).

2.3.2 Model Validation

The canonical (NVT) ensemble is implemented: the number of particles (N), along with the volume (V), of each system in the ensemble are the same, and the ensemble has a well defined temperature (T), given by the temperature

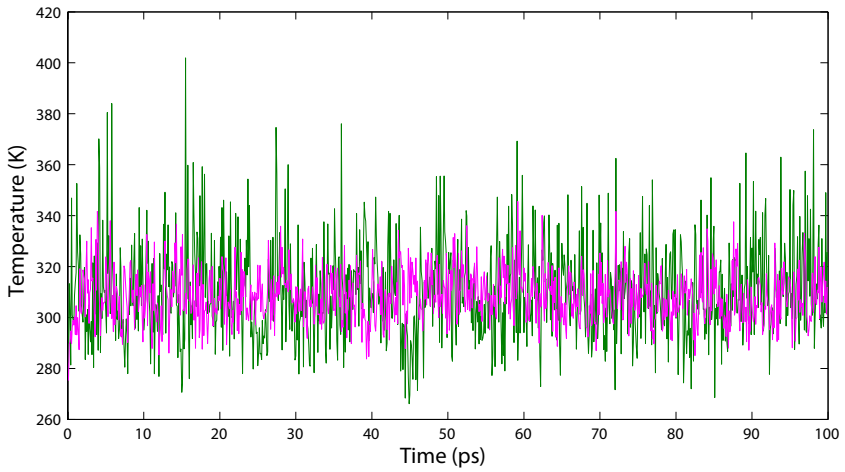

Figure 2.6: Thermal profiles of two simulation runs. The green trace shows a dilute barnase and barstar trajectory at 19.05g/L, whereas the pink trace shows the same proteins at 76.20g/L. In both runs, the heat-bath temperature was maintained at 310K. On-the-spot temperatures are calculated from the total kinetic energy of all molecules in the system.

of the heat bath with which it would be in equilibrium (see Figure 2.6 for a typical kinetic energy profile during a simulation). Importantly, over all the numerous test and production runs, and for all the various molecular systems simulated, energy profiles show stability over the complete time course of each simulation. However, fluctuations in kinetic energy do occur over time with respect to the protein concentration: a more crowded environment has less perturbation, which suggests that the movements of the molecules are subject to more restraints from neighbours. To ensure that molecules are physically well behaved for rigid-body interactions, both energy and kinetic momentum conservation is checked during each simulation: the total binding energy of an interacting pair of molecules should be no more than the loss of their kinetic energies upon interacting.

The energetic aspect of the specificity-sensitive LD scheme is examined through comparison between ligand-receptor RMSDs and the corresponding binding energies. A negative binding energy state means the partners are attracted towards each other, while a positive term means they are moving away. Similarly, a rising wave of binding energies in successive simulation timesteps means the interaction is becoming destabilised, and *vice versa*.

Figure 2.7 shows four examples of RMSD-binding energy comparisons for different protein interactions. The example trajectory of complex 1TMQ (alpha-amylase and ragi inhibitor) shows a very stable contact with which the RMSD keeps within 0.2-0.4Å of the native binding site throughout the 30,000 timesteps (30ns) of simulation. Correspondingly, the binding energy remains almost constant during the interaction time. In the second example, binding partners of complex 1E6E (Adrenodoxin-Reductase) displays a high correlation between RMSD and binding energy changes. Moreover, the peaks and valleys on the energy curve usually occur ahead of the highs and lows of the RMSD, showing the change in complex conformation is often led by changes in potential energy over a small period of time. The same RMSD-binding energy correlation can also be found in the example trajectory of complex 1WQ1 (Ras-GAP), where more dramatic change in RMSD (from approx. 14Å to 4Å) is also coupled tightly with variation in binding energies. The 1KTZ (TGFβ3-TβRII) example, however, displays weak binding energy terms throughout the trajectory; hence a gradual and stable drift from the binding site is observed. The above four examples suggest that the potential energy functions employed by BioSimz provide a reasonable approximation of the energetics associated with specific protein-protein interactions; therefore, the developed forcefield appears to be justified for conducting macromolecular simulations in normal and crowded conditions.

To further examine the effects on solvent electrostatics to macromolecular interactions, salt ions were explicitly added to the simulations and the differences in variations to the ionic strengths examined. In this test, Na^+ and Cl^- ions are explicitly modelled at three concentrations, i.e. 10mM, 50mM and 150mM. The result shows the ionic effect on binding scores is negligible if ionic strength is weak (10mM). At 50mM and 150mM, most interaction pairs display a slight decrease ($10\sim$ 20%) in binding scores. However, the relative affinities among a test set of eleven complexes (see Section 4.2.2) remain unchanged, except for highly electrostatic interactions such as displayed between barnase-barstar (PDB:1B27) and adrenodoxin-adrenodoxin reductase (PDB:1E6E), which appear to be damped more than others. These simulated results qualitatively agree with experimental observation, which suggests protein complex stability is negatively correlated to the solution's ionic strength.

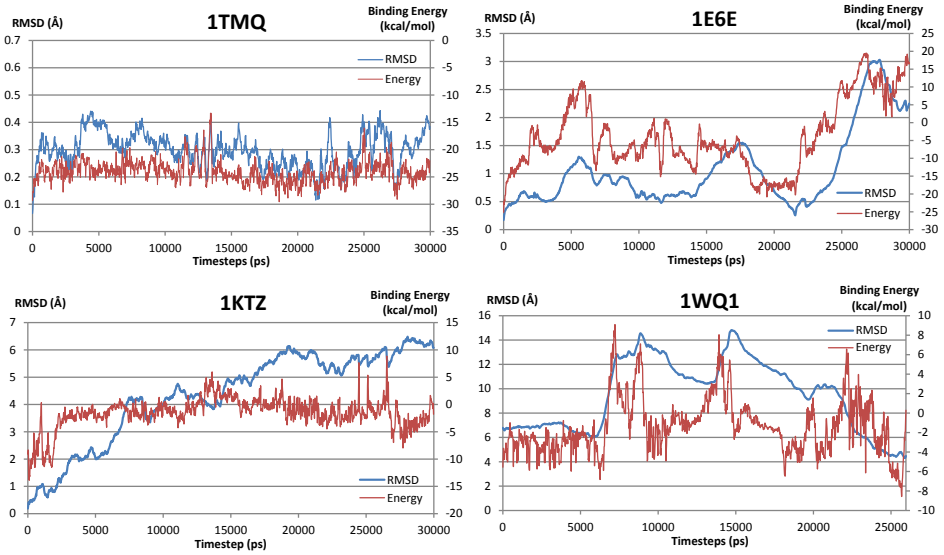

Figure 2.7: RMSD and energy profiles of protein-protein interactions at or near their respective binding sites. Blue curves show the trajectory of RMSD deviation of complex conformation from the native (crystallographic) state; red curves are the corresponding binding energies, calculated as the summation of all potential terms of the forcefield.

2.3.3 Approximations and Stability

Developers of molecular modelling software often face the following dilemma: given the size of the system, to what level of resolution can the system be modelled if the simulation is to cover time periods relevant to the biology of the system. Therefore, while a high resolution model would naturally include molecular flexibility this option has not as yet been implemented fully in BioSimz, rather, the focus has been more on simulating macromolecular interactions for as long a time period as possible, hence, the compromise made is that all molecules are treated as rigid-bodies. Moreover, for achieving reasonable translational and rotational dynamics, the Langevin timestep is set at the picosecond level. This effectively eliminates the possibility of doing side-chain level molecular mechanics, as 1ps would be too coarse for the side chain atoms to be sampled with a continuous trajectory. In this case, a Monte-Carlo conformational sampling method may be needed should internal dynamics be incorporated into the LD scheme. Despite the rigid-body treatment not being an appropriate method for highly-flexible macromolecules (DNA, RNA and disordered proteins), it is shown to be a sufficient description of molecules for which

relatively small conformational changes occur during diffusion and interaction. This is the case for the test examples described here.

The model employs a relatively simple treatment of electrostatics, although special rules (describe in the Forcefields section) have been applied to the dielectric constant. Alternative electrostatic modelling techniques include generalized Born (GB) models (Bashford and Case, 2000) and solutions to a Poisson-Boltzmann (PB) equation of the molecule (Baker et al., 2001). While they both provide a better approximation to simple Coulomb equations, the GB approach needs very careful parameterisation of the Born radii in highly electrostatic conditions (Bashford and Case, 2000), which are often the case in a crowded environment.

On the other hand, solving either the linearised or non-linear PB at each timestep under the influence of a changing number of neighbouring molecules is currently infeasible given the available computing power. Molecular simulation studies using PB have to pre-compute and store the electrostatic potentials in discrete meshes, which leads to significant errors on neglecting the low-dielectric region near the interaction interface. As was written in one review (Schreiber et al., 2009), these approximations "are worst when proteins are in close proximity, precisely where electrostatic interactions are expected to have the strongest influence" on the association rate constant, k_{on}. The comparison between electrostatic energies produced by PB and Coulombic methods shows that the major difference only kicks in when distance between molecular surfaces is less than 3Å(Camacho et al., 1999). The Coulomb-based treatment for electrostatics implemented in this work addresses this problem by re-parameterisation of the dielectric constant near the molecular surfaces (see Section 2.1.2), which is sufficiently accurate for the investigation of encounter complex formation.

Chapter 3

Macromolecular Docking

As was described in Section 1.1.3, macromolecular interactions, especially protein-protein interactions, are the cornerstone of biological processes and functions. The first step to understanding these interactions is usually to locate the binding interface region on each binding partner, and and then ascertain why these particular interfaces have been utilised.A second, perhaps alternative approach to the above, is to work out how interactions occur naturally, i.e. the binding mechanism from a dynamic point of view. This chapter applies the BioSimz simulation method to refine and improve molecular docking approaches, and attempts to reveal the macromolecular binding mechanisms through the analyses of simulation trajectories.

The majority of the content in this Chapter has been published as a prediction report for the Critical Assessment of PRotein Interactions (CAPRI) in *Proteins: Structure, Function, Bioinformatics* (Li et al., 2010). The work mentioned in this study was performed in collaboration with a fellow graduate student and colleague of mine, Mr Iain Moal.

3.1 Introduction

3.1.1 Theories

Theories of docking are split into two main categories, those of the lock-and-key model and those from the induced-fit hypothesis. The former theory postulates that protein shapes and surfaces are highly specific to their corresponding partners, emulating a key in a lock, such that they naturally

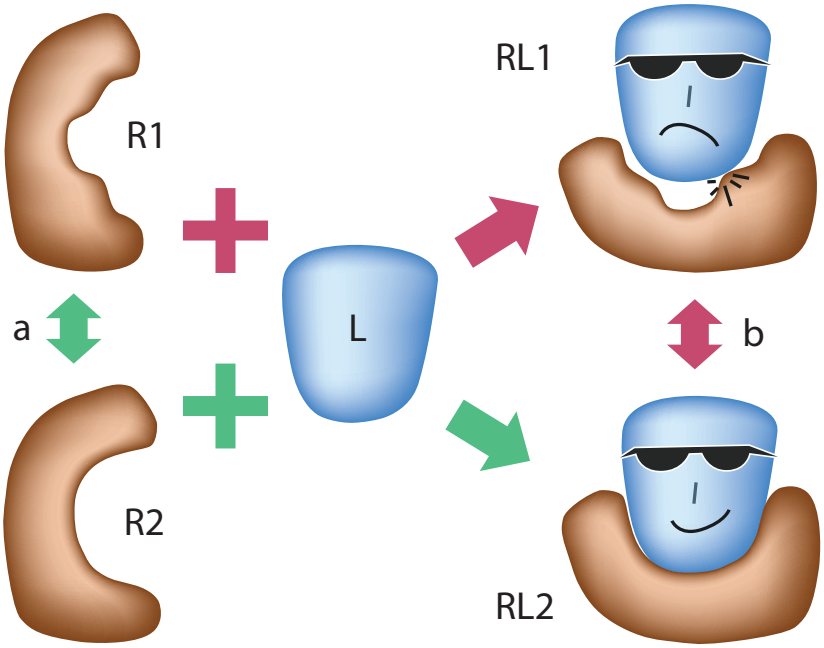

Figure 3.1: A schematic view of the two binding mechanisms. The induced-fit hypothesis follows the *red* path: $R1 + L \Rightarrow RL1 \Leftrightarrow RL2$. The conformational shift hypothesis follows the *green* path, i.e. $R1 \Leftrightarrow R2 + L \Rightarrow RL2$. Reversible process *a* is termed the pre-equilibrium process, while the corresponding reversible process *b* represents the conformational changes of the receptor upon association and dissociation.

bind upon meeting each other; the latter is characterised by initial binding in an open state, followed by structural rearrangements due to changes in potential energy caused by the binding partner. Over time, the lock-and-key model has evolved from proteins being rigid bodies to that of each protein binding partner being able to sample an ensemble of conformations, which are in constant flux due to the thermal energy of the system; therefore, complexes form when binding partners collide, with each partner being in the appropriate conformation for stable binding to occur. Perhaps to coincide with the phenomenon of state switching, the model has also been assigned with a number of names, such as "population shifting", "conformational selection" and "conformational shift"; the latter phrase is adopted in this study. To date, most protein binding studies have followed one of these two alternative binding hypotheses, as are illustrated in Figure 3.1.

The lock-and-key model is a century-old concept, first coined by Fischer

(1894), well before that highly specific three-dimensional structures could be experimentally elucidated. The model well explained that most enzymes are specifically "designed" to accelerate biochemical reactions of a single type. Indeed, from a modern perspective, many of the enzymes facilitating reactions between small molecules have identifiable "binding pockets" to accommodate substrates of certain shapes. In many cases, such as for serine proteases and aspartic proteases, the catalysing function is delivered by one or two key residues in the active site of the enzyme, acting exactly like the locking/unlocking mechanisms of a key and lock. For protein-protein interactions, many enzyme-inhibitor bindings are established by strong electrostatic interactions between key charged residues, such as in the case of bacterial ribonuclease barnase and its high-affinity inhibitor, barstar (Hartley, 2001).

The first mention of the induced-fit theory was made by Koshland (1958). Koshland stated that proteins change conformation favourably towards accepting their binding partner, within the final stage of complex formation, i.e. the fit occurs "*only after* the changes induced by the substrate itself". This theory became popular very quickly as it seemed to have explained all observations that did not fit the static lock-and-key model; perhaps more often than it ought to be, "conformational changes upon binding" was used as a get out clause when computational packages fail to dock proteins known to form complexes in a crystallised form. Nevertheless, evidence of induced-fit has been observed for many protein-protein interactions, such as within flexible SH2 and SH3 domains upon binding their peptide substrates (Hofmann et al., 2005).

Just as the induced-fit model seems a "one-size-fits-all" type of solution, there have been increasing voices suggesting otherwise (Bosshard, 2001). Additional binding mechanisms were suggested (Boehr and Wright, 2008), based on the conformational shift observed from studies of nuclear magnetic resonance (NMR) dynamics of ubiquitin; a highly-conserved small protein involved in many eukaryotic regulatory processes. Similar cases of conformational shifts influencing binding activities were also observed for enzymes and antibodies (James et al., 2003; Tang et al., 2007; Eisenmesser et al., 2005). Conformational shift, argued at least by one group (Tsai et al., 1999; Ma et al., 2002), may be a key factor for specific binding between pro-

teins, and even for those with highly flexible interfaces,or indeed disordered interfaces (Tsai et al., 2001).

With NMR providing the eye-opener for in-solution protein flexibility, the natural docking between proteins and their binding partners now appears to be less magical in terms of the induced fit model. Indeed, short high-frequency fluctuations in protein structure conformation only seems to have a small effect on binding dynamics; but longer term, low-frequency, dynamic fluctuations in conformation, as easily traced by NMR, may suggest proteins vary their shapes significantly and relatively frequently; therefore, the chance of binding partners meeting each other at the right time with the right pose, is not that slim.

The same group that argued for conformational shift also suggested that, the long-held view of protein binding at the energy minimum, represented by *one* bound/docked complex, may not be true either (Tsai et al., 2008). This view is shared and extended here. This implies bound, crystallographic complexes deposited in the PDB may represent only one of possibly a few closely-related conformations at which two proteins may bind and function.

Thoughts on multiple bound conformations has led to greater interest in investigating the less specific, more transient interactions between proteins and their binding partners. Encounter complexes are thereby loosely defined as proteins that make surface contacts near to the "correct" binding location. The study of encounter complexes falls between protein docking and protein diffusion studies – and traditionally more frequently investigated with the latter (Gabdoulline and Wade, 1997). One purpose of the work described in this thesis is to use computer simulations of encounter complex formation as a bridge between between docking and kinetic binding studies (see Section 3.3, 4.3 and 5.3).

3.1.2 Practices

To computationally solve the molecular docking problem, one has to balance the limitations of computer size, speed and accuracy. Therefore, a universal algorithm has not been developed and optimal solutions to any one particular macromolecular docking problem can be reached by a large number of algorithms; physical or non-physical, deterministic or random, domain specific or statistically modelled, supervised and unsupervised, stan-

dalone and collateral. Ways to approach a docked solution for two (occasionally, three or more) proteins are so numerous and diverse that even an incomplete examination of them will result in this section being excessively bloated – for a more detailed description of docking algorithms than can be given here, see for example, the comprehensive review of Halperin et al. (2002).

An even more concise review of the field would be to examine approaches that are currently under development in the molecular docking field, especially by those participating in the Critical Assessment of PRedictions of Interactions (CAPRI), an ongoing blind trial of macromolecular binding site predictions. In each round, one or a few protein complexes whose structures have recently been solved are put to test with registered entrants, which can be either human or non-intervention server predictors. Currently there are more than 30 participating groups, of which many have their named approaches. A tabularised brief review of the popular methods from CAPRI participants, as well as some classical methods, can be found in Table 3.1.

As was rightly noted in the latest report for CAPRI entries of 2009 (Lensink and Wodak, 2010), over the years there has been noticeable improvement in the correct scoring of known complexes, as opposed to the actual ranking of potential docking solutions. The ranking problem can further be divided into two sub-problems: a) whether an acceptable docking pose has been contained in the ranking set, and b) whether the pose can be highly rated by the algorithm. Evidence shown in the latest CAPRI report (Lensink and Wodak, 2010, Figure 2) suggests that problem a) has been more consistently encountered by the protein docking community. If the correct or near-correct ensemble of conformations is contained within a set of predictions, it is likely to be picked up by a number of scoring schemes. Therefore, it seems the larger hurdle to solving the macromolecular docking problem, is that current approaches cannot effectively generate a set of conformations, which regularly contain a correct, or near correct, answer. The situation is exacerbated by the knowledge that due to the large number of degrees of freedom that may need to be considered to dock two proteins, consisting of rotational,translational and a multitude of conformational degrees of freedom, many millions of potential solutions may need to be gen-

Name	Search Algorithm	Scoring Functions	Re-scoring, Ranking, Filtering and Refinement	Reference
3D-Dock Suite	Global rigid search (GRS): fast Fourier transform (FFT)	Residue-based pair potentials (RPScore)	Re-scoring and clustering; refinement of interface side-chains	Jackson et al. (1998)
3D-Garden	GRS in ensemble	Shape complementarity (SC) and Lennard-Jones (L-J) potential	Side chain and backbone dihedral refinement	Lesk and Sternberg (2008)
BiGGER	GRS: Soft surface representation by bit-mapping method	SC and favorable residue's contacts	Re-scoring with multiple filters (electrostatic, hydrophobic, side-chain contacts)	Palma et al. (2000)
ClusPro	GRS: FFT correlation approach (PIPER)	SC and desolvation energy, and electrostatics.	Re-scoring (desolvation and electrostatic energies) and clustering	Comeau et al. (2007)
DOT	GRS: FFT	SC, electrostatics and VDW	Re-scoring with ACE term	Mandell et al. (2001)
GRAMM	GRS: FFT. Smooth protein surface representation for soft docking	SC and L-J potential	Clustering of conformations	Vakser (1997)
GRAMM-X	GRS: FFT. smooth protein surface representation for soft docking	SC and L-J potential	Minimization and re-scoring with multiple filters	Tovchigrechko and Vakser (2006)
HEX	GRS: Fourier correlation of spherical harmonics	SC	N/A	Ritchie et al. (2008)
HADDOCK	GRS	Electrostatic, VDW and desolvation energy terms	MD simulated annealing refinement; filtering based on external data; clustering and re-ranking	Dominguez et al. (2003)
ICM	GRS: Monte Carlo (MC)	Empirical scoring function	Clustering and selection of conformations; refinement of interface side-chains	Abagyan et al. (1997)
MolFit	GRS: FFT	SC	Clustering of good solutions; filtering using a priori information and small/local rigid rotations around selected conformations	Segal and Eisenstein (2005)
PatchDock	GRS	SC and ACE	Clustering of conformations	Schneidman-Duhovny et al. (2005)
PPISP	GRS with bias from interface prediction	Biochemical data	Refinement with explicit-solvent MD	Qin and Zhou (2007)
PyDock	GRS: FFT	SC	Re-scoring by binding electrostatics and desolvation energy	Cheng et al. (2007)
RosettaDock	Local rigid search (LRS): MC with low and high resolution structure representation levels	Different scoring parameters for different resolutions	N/A	Lyskov and Gray (2008)
ZDOCK	GRS: FFT	SC, desolvation energy, and electrostatics.	Energy minimization and re-scoring	Chen et al. (2003)

Table 3.1: Popular docking approaches developed and used by CAPRI participants.

erated - this is especially true if there are no experimental or evolutionary pointers to the potential binding sites on each protein. It is for this reason that the work reported in this thesis focuses on the efficient simulation of encounter complex formation for protein-protein interactions, thereby providing a much smaller, and more appropriate, set of potential docking solutions that can be subsequently refined.

3.2 Materials and Methods

3.2.1 BioSimz

Rigid-body dynamic simulations of proteins at atomic resolution were run to provide information on potential binding regions, based on the frequency heatmap of surface contact in different regions of the interacting molecules. See Section 2.1.4.1 for more detail on the structural assessment of association dynamics in BioSimz, and for a detailed description and evaluation of the BioSimz package see Section 2.2.

The simulation starts with distributing all macromolecules randomly in a cubed box, sized $240 \times 240 \times 240 Å^3$, where periodic conditions are applied. A 10ns high-temperature run is carried out before each production run, allowing the kinetic energy of individual molecules to equilibrate. Throughout the production runs, the temperature is maintained at 298K. All simulations are run for 200ns at 1ps per timestep, resulting in a mean-squared displacement of 600 $Å^2$ in dilute solution. To even out statistical variance between individual runs, 10 runs are performed for each configuration.

In each simulation run, eight receptor and eight ligand molecules are put into the box, making up a total concentration of 1.92mM, which corresponds to between $18.5 gL^{-1}$ and $102.9 gL^{-1}$ depending on protein sizes. Crowded molecular simulations are also carried out to emulate *in vivo* binding conditions. Sixteen proteins, from ten types of bacterial enzymes of the glycolytic pathway (see Section 1.1.2.2 and Section 5.2), are modelled as macromolecular crowders, also at full-atom resolution. These molecules are ubiquitous throughout almost all life forms and often present in high abundance (Ishihama et al., 2008), hence they are thought to be representative of a typical

crowded macromolecular environment. The crowded simulations have a total protein concentration of 3.96mM, corresponding to between 168gL^{-1} and 252gL^{-1}, of which 149gL^{-1} are crowders. This is comparable to the estimated *in vivo* macromolecular concentration 300gL^{-1} (Zimmerman and Trach, 1991).

Time-course data, i.e. the retention time of protein-protein encounters formed during the simulations, are collected to produce a contact frequency map for the protein interaction of interest. A contact heatmap map, for the receptor of a receptor-ligand pair, is constructed as follows:

For all interaction poses found in the simulations, the receptors are superimposed, carrying over ligand centres of mass positions for all receptor-ligand interactions during the complete simulation time course. This produces a "point cloud" over the representative receptor surface. Since each point represents the location of a ligand relative to its closest receptor at the time it was recorded, positions recorded each timestep (1 ps), both the number and the retention time (relative hovering) of ligands as they make contact with the receptor can be visualised, see Figure 3.2. From this cloud, and the recorded full atom positions, the total number of receptor-ligand contacts, across the complete surface of the receptor, can be easily represented as a heat map, see Figure 3.3. To construct the equivalent heatmap for the ligand, the above process is repeated, but with the ligands superimposed carrying over the receptor centres of mass.

For the unusual case of CAPRI Target 43, described in detail below, dissociation simulations were also carried out for its 21 designed complexes to investigate their retention under *in vitro* conditions; see Section 4.2.3 for a detailed description of the simulation protocol for complex dissociation.

3.2.2 SwarmDock

A new flexible docking method developed within our laboratory called, SwarmDock (developed by Mr Iain Moal), was used to perform docking studies, guided by the contact frequency heatmaps generated from BioSimz. The method is named after the particle swarm optimisation (PSO) algorithms (Kennedy and Eberhart, 2002), from which the docking method is derived. For any optimisation problem, PSO employs a "swarm" of agents, each of which discovers and shares the local optimal solution of the poten-

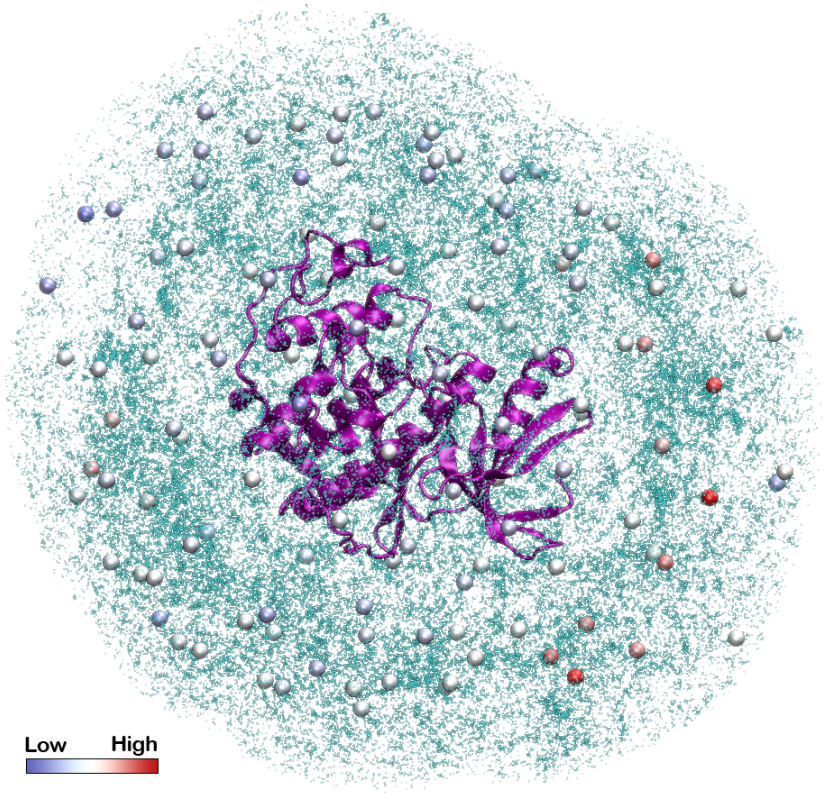

Figure 3.2: A ligand cloud density map. The centre of mass of the ligand at each timestep is plotted as a cyan dot; a non-homogeneous cloud is formed by the aggregation of large numbers of dots in certain regions. Spheres hovering above the receptor molecule are equally distributed around the receptor surface at a given distance. The spheres are coloured by the density of cyan cloud in its vicinity: the more *red* its colour is, the higher the density of the ligand cloud in that region; for *blue* spheres, it's the opposite. *White* spheres have a moderate density of ligand centre of mass cloud in its vicinity.

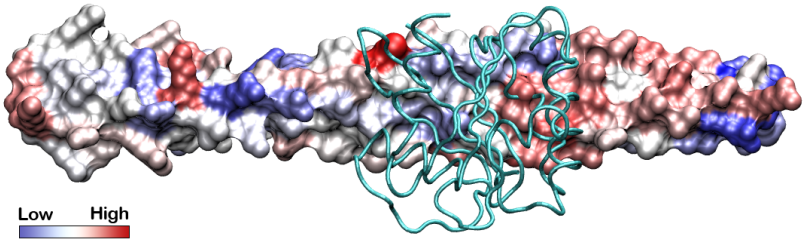

Figure 3.3: A heatmap for simulation surface contacts. CAPRI Target 37 is used as an example here: the long double-stranded α helices are selected as the receptor (displayed as a solid molecular surface), the ligand, a globulin protein, with its backbone represented by a cyan tube, is positioned at the experimentally determined binding site. The receptor's surface is coloured with the frequency of contacts made by the ligand in the simulations. A strong anchoring spot and a weaker contacting region are visible in the correct binding region.

tial surface it is travelling through. The movement of each agent through multi-parameter space, on each iteration of the algorithm, is determined by a combination of its own potential, as well as the potentials passed from other swarm members. Therefore, a team of local minima searching agents will ideally bear the power of searching for the global minimum, which represents the optimal solution to the given problem. SwarmDock uses a variant of the local best PSO, in which the search space consists of a Cartesian coordinate space, a quaternion space for orientations and a set of conformational space parameters for both interacting macromolecules. Currently, the conformation space parameters are constructed from the first five normal modes (lowest frequencies) for each unbound protein, calculated using the computer package ElNemo (Suhre and Sanejouand, 2004). A swarm of "docking particles" then navigate through the above described parameter space, searching for the global energy minimum between two interacting proteins. The potential energy function used to guide the search consists of VdW and electrostatic terms, with L-J parameters and partial charges taken from the CHARMM19 forcefield (MacKerell et al., 1998a). A complete description of the algorithm, along with extensive benchmarking, has been published elsewhere (Moal and Bates, 2010).

3.2.3 Filtering with Contact Frequency Maps

The ligand contact cloud maps (see Figure 3.2) from BioSimz simulations indicate the probabilities of the likely binding interface over the complete surfaces of each interacting protein. This information can then be used to directly dock the two proteins. However, it is often the case that one protein surface shows a more definitive heatmap than the other. To test whether the information from just one heatmap can help enhance our docking algorithm (SwarmDock), one contact frequency cloud for a protein of a complex pair, usually the larger termed the "receptor", is used to guide the docking of the other protein, the "ligand". A point in the cloud around the receptor is one position of a ligand molecule at one timestep; as described above, the density of the point cloud in any one particular region of space, is a product of both the number of receptor-ligand interactions and their retention times. For specificity and convenience, only points within certain cut-off distances from the receptor surface are included. The minimum cut off is set to 3Å

and the maximum to the sum of the longest axis of the ligand plus 7Å.

Equally distributed points around the receptor, that are used as the starting locations for the ensemble of ligands in each SwarmDock run, are then superimposed onto the contact frequency point cloud. The equally distributed Swarmdock points are then scored to reflect the density of the surrounding BioSimz trajectory points. A number of scoring schemes, similar but varying on how the neighbouring points are weighed, were evaluated and described in full in a recent publication (Li et al., 2010, Supporting Information). The scoring scheme used here is that for each trajectory point in the ligand cloud (a blue dot in Figure 3.2), the scores of the nearest 5 SwarmDock starting positions (coloured spheres in Figure 3.2) are incremented by 1. For the BioSimz guided docking runs, only the top half of all starting poses (half of the equally distributed SwarmDock starting points), according to the respective scoring scheme, are selected for PSO and subsequent docking steps; the remaining half are discarded as being unlikely to initiate the formation of the specific complex. The effectiveness of the above filtering approach is discussed below.

As an additional quality control for the effectiveness of filtering, the ability of BioSimz to locate the experimentally determined binding region, and to ascertain whether more frequent and prolonged ligand interactions occur within the known binding region compared to non-binding regions. The scores of the ten SwarmDock starting positions near the centre of mass of the experimentally determined bound ligand pose were tested against the null hypothesis that they are drawn from the same distribution as the other starting positions, with an alternative hypothesis that the scores of the binding region points are greater than the scores of the non-binding regions. Similar tests were also used to test whether the scores of the binding region are significantly lower than those of the non-binding region, an indicator that the true interface is disfavoured during simulation.

3.2.4 Test Cases

For benchmark studies, a total of 26 X-ray crystal complexes, along with the structures of their unbound components, were taken from the Protein-protein Docking Benchmark 2.0 (Mintseris et al., 2005). These include enzyme-inhibitor (1AVX, 1AY7, 1PPE, 7CEI, 1TMQ, 1EAW and 1HIA),

enzyme-substrate (1EWY and 1E6E), antibody-antigen (1QFW, 1JPS, 1NCA, 1VFB, 1AHW, 1NSN, 1I9R and 1FSK), and signal-effector/receptor (1KTZ, 1GCQ, 1GRN, 1FQJ, 1BUH, 1KAC, 1ML0, 1QA9, and 1HE8) complexes. For each complex, the larger of the binding partners is referred to as the receptor and the smaller as the ligand. Only the unbound structures are used in BioSimz for simulations and SwarmDock for docking.

3.3 Results

3.3.1 Signals at Binding Sites

Rigid-body LD simulations were run on all 26 test complexes. These trajectories were used to score the SwarmDock starting positions. To test whether the experimentally determined binding site region formed more frequent and tenacious interactions than the other regions surrounding the receptor during the time course of each simulation, one-tailed Wilcoxon rank-sum tests were performed at the 5% significance level. For 12 of the complexes (1KAC, 1KTZ, 1FQJ, 1E6E, 1EWY, 1AHW, 1GRN, 1BUH, 1VFB, 1FSK, 7CEI and 1AY7), the 10 starting points nearest to the ligand centre of mass had a significantly higher score than the remaining starting points further away from the binding site. Further Wilcoxon rank-sum tests showed that, for three of the complexes, the starting points near the binding site scored significantly lower (P-values: 1I9R 0.002, 1EAW 0.003, and 1TMQ 0.036). Interestingly, two of these, 1EAW and 1TMQ, involve a protrusion of one binding partner into a deep groove in the other, the formation of which cannot be predicted by the current version of BioSimz since only rigid-body dynamics can be performed; subsequent versions will include side-chain and limited backbone flexibility. See also a general introduction of Wilcoxon rank-sum test in the Appendix.

3.3.2 Filtered Docking

Two sets of SwarmDock runs were set up, one global, where the algorithm was run twice from each starting position (a set of equally distributed points around the receptor), and one filtered, also run twice, from half the starting positions where the lowest scoring half were discarded (points scored as

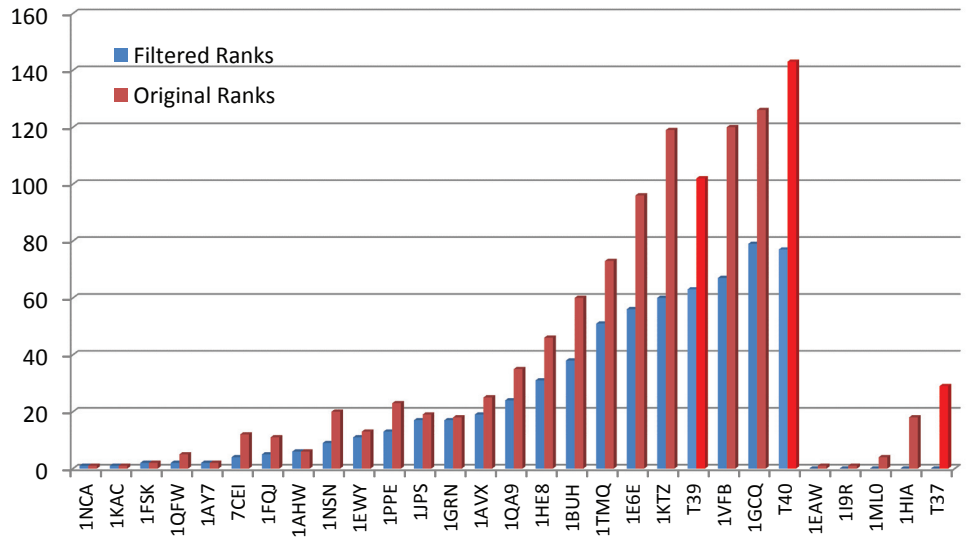

Figure 3.4: Rank improvement by filtering the docking (SwarmDock) starting points using the ligand hotspots. Among the 26 test complexes, ranks of 22 were improved, while 4 complexes failed to dock after filtering the starting points.

described above in section 3.2.3). The resulting structures were clustered and ranked. The difference in rank between the global and filtered runs is shown in Figure 3.4. The unfiltered runs successfully found the binding site for all complexes. Upon filtering, the rank was improved for 17 of the structures and remained the same for five structures (four of which were already in the top two), whereas four of the structures (PDB:1EAW, 1I9R, 1ML0, and 1HIA) did not find a successful docking hit after filtering. Of these, three did not form specific encounter complexes (PDB:1EAW, 1I9R and 1HIA), whereas the other was not detected by SwarmDock (PDB:1ML0).

To further test the validity of pre-filtering SwarmDock calculations, two additional SwarmDock runs were set up, one global and the other filtered. For the global docking, the algorithm was run twice from each starting point. For the filtered docking, the algorithm was run four times from the upper half of the points. The difference in number of successful docking hits between the global and filtered runs, plotted against the Wilcoxon rank-sum test P-values, is shown in Figure 3.5. The 12 complexes which have a significantly more populated binding region all found the binding site more

frequently by restricting the search space based on the BioSimz simulations. Of the 11 complexes for which the binding site scored neither significantly higher nor lower than the nonbinding region, five correctly docked more frequently, five less frequently, and one found the binding site an equal number of times. Of the complexes for which the binding site scores significantly lower, 1TMQ docked less frequently after restricting the search space, whereas the other two complexes (1I9R and 1EAW) failed to dock. Therefore, filtering out half of the less encountered regions for the receptor has yielded an overall positive outcome as long as a significant signal is found in the initial BioSimz simulation. Even if a significant signal is not present, our filtering process does not have a negative influence on the overall outcome by trimming off half of the initial docking starting positions.

3.3.3 CAPRI Targets

3.3.3.1 Targets 32 and 38

During the blind trial prediction period for these two targets, crowded simulations were performed on their unbound components, in addition to the dilute setting. For both settings, the correct binding site was found, However, without the overcrowding proteins, three false positive sites were also found, one of which was more prominent than the true binding region. Simulation with external crowding molecules, however, removed one false binding site altogether and significantly diminishes the strength of the two other false sites, leaving the true binding region. Correspondingly, the p-values for the non-crowded and crowded simulations are 0.354 and 0.025, respectively, demonstrating a significant improvement in the enhancement of the binding site by the inclusion of external crowding proteins. Similar results were found for Target 38, the signal-effector complex of centaurin-α1 and KIF13B (PDB:3FM8), for which a homology model was built for KIF13B FHA domain using the POPULUS server (Offman et al., 2008). In this target, the p-values for the crowded simulation was 0.057, again significantly improved over $p = 0.388$ for the non-crowded simulation. Analysis and discussion of the macromolecular crowding effect in general, are detailed in Chapter 5.

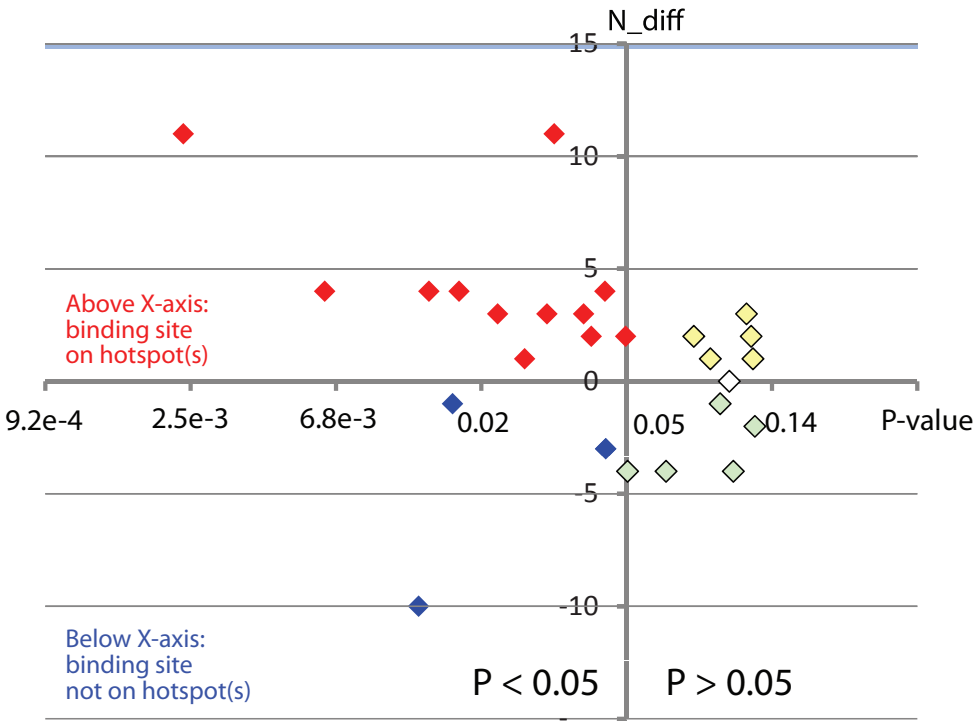

Figure 3.5: Significance test of the rank improvement. The x axis shows increasing Wilcoxon p-values, indicating a decreasing trend of significance of data, while the y axis records the difference in number of docked hits after filtering, as an indicative measure of the contribution from filtering out the non-hotspot regions. All 12 complexes (*red* marks) with significantly higher scoring binding regions correctly docked more frequently after imposing a restricted search space. The three complexes (*blue* marks) with a significantly lower scoring binding region performed better without restricting search space. For the remaining insignificant ($p > 0.05$) cases, five found the binding site more frequently, five less frequently and one equally frequently.

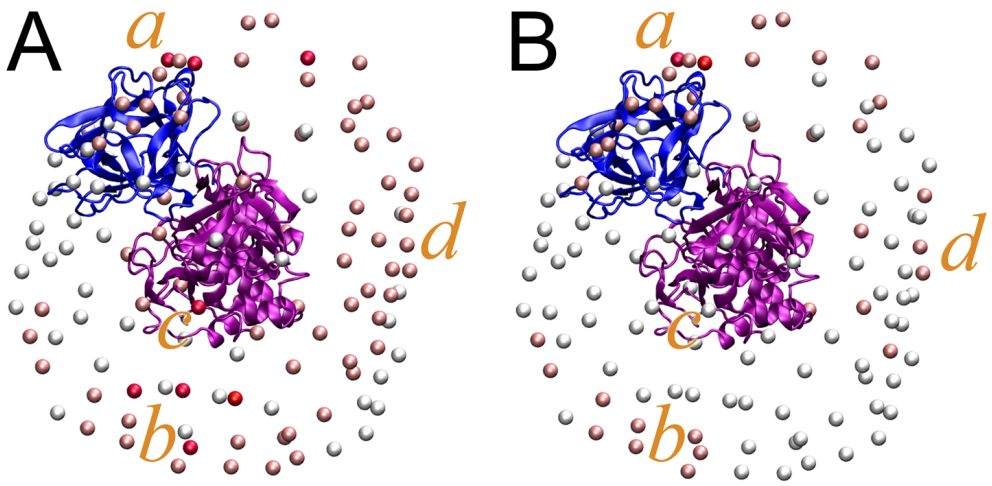

Figure 3.6: Elimination of false positive sites upon macromolecular crowding. SwarmDock starting positions are plotted in spheres surrounding the receptor, coloured with BioSimz contact cloud density. The *red* spheres corresponds to high contact density regions. (A) shows the density map from simulation without environmental crowders, where three false positive binding sites, b, c, d, are prominently visible. (B) is the same density map from simulation with environmental crowders, where the false positive binding site c was completely removed, while strengths of b, d were greatly reduced. Meanwhile, the strength of the true positive site, a, is retained.

3.3.3.2 Targets 39 and 40

CAPRI Target 40 (PDB:3E8L) is a complex of the double-headed arrowhead protease inhibitor and trysins. The protease inhibitor has two trypsin-bound binding sites, one of which contains a characteristic cysteine-lysine-isoleucine (CLI) protease inhibitor motif. This binding site is found in both uncrowded and crowded simulations, with p-values of 0.016 and 0.003 respectively. The other binding site is also found (with p-values of 0.159 and 0.169 for the uncrowded and crowded simulations), as shown in Figure 3.7, however, the high scoring region is not directly above the binding region, but to the side of it. A similar effect is seen for Target 39, the same binding partners as Target 38, but with the KIF13B FHA domain in the bound conformation. The unbound binding partners, superimposed upon the bound structure, are shown in Figure 3.8. Residues are coloured by the log of the number of contact events collected during the simulations. Although the interface residues do not make a significant number of contacts, the residues, which make the most contacts, appear near to the interface, opposing each other on both the receptor and the ligand. It would seem likely that the

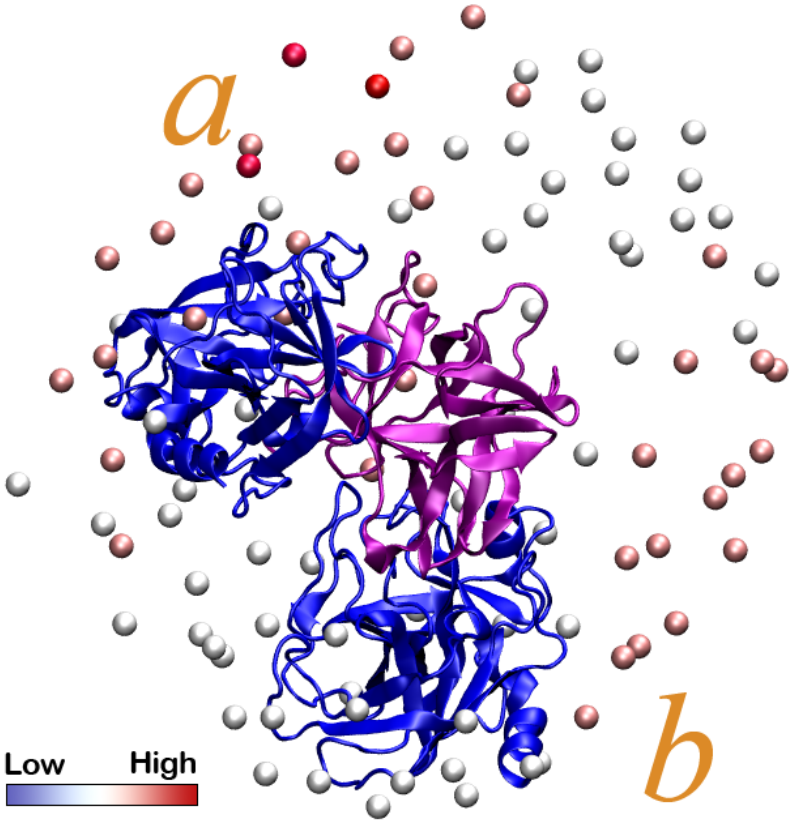

Figure 3.7: Dual binding sites found with BioSimz ligand binding cloud. The primary binding site of CAPRI Target 40, *a*, was found with the highest density of ligand presence, calculated from a number of BioSimz simulations, shown here mapped to the SwarmDock starting positions. The secondary binding site, *b*, also exhibited a moderate increase in cloud density near the binding sites. Bound conformations for the two experimentally determined binding sites are represented in ligand molecules shown in blue cartoon display.

initial encounter complex is formed here, followed by rolling of the proteins into the biological interface. This pattern, consistent with off-centre encounter complex formation, followed by a short 2D diffusional search, is also seen in the benchmark simulations for complexes 1BUH and 1QA9.

3.3.3.3 Targets 43 and 44

T43 and T44 are two rather unusual trials compared to most, if not all, previous CAPRI rounds. T43 contains 21 complex structures in which only one is the real structure solved from crystallographic studies; all remaining com-

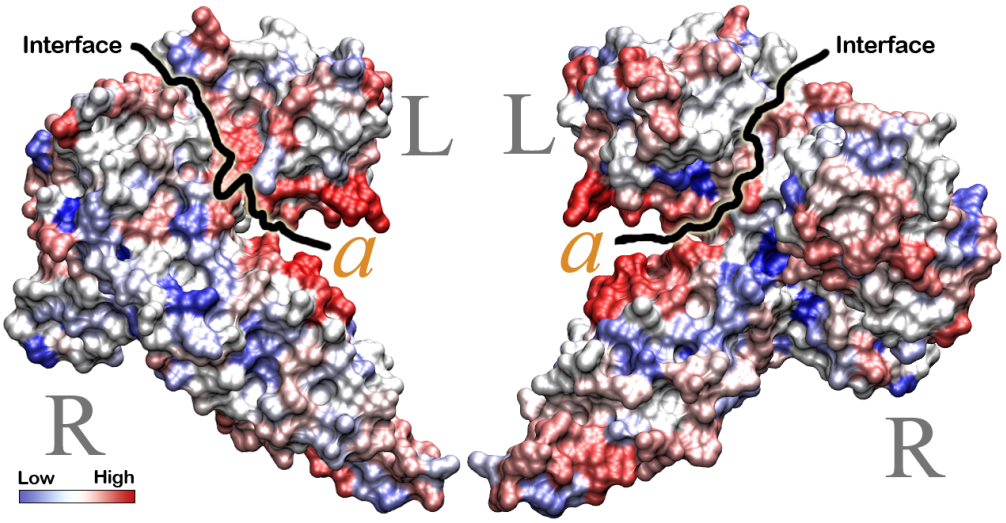

Figure 3.8: The anchoring residues. Surface contact heatmaps, viewed from front and back of the experimentally determined interface, for CAPRI Target 39, are displayed. The true binding interface is represented by the black line drawn between the ligand and the receptor. Region *a* is visibly the hottest spot for both ligand and receptor molecules, which is proximal to the specific binding interface.

plexes are designed. For T44, all of the 21 complex structures given were designed. The participating groups were asked to discriminate the crystal (native) structure from the designed complexes in T43 and to rank the designed targets in T44. All designed complexes were made available through the Baker group from University of Washington, St Louis. The trials might be part of the group's *de novo* interface design efforts for proteins that do not naturally bind.

By definition, T43 and T44 are two scoring-only targets, as the participants were not asked to alter the given structures. However, the difficulty came with the diversity of the modelled structures – each target contains several entirely different complex "families" rather than a set of different poses or conformations of the same receptor and ligand. By visual classification, T43 includes 5 receptors families and 21 different ligands, while T44 includes 4 receptor families with various ligands. In order to make the right judgement not only within a complex family but also in between different families, potential energy based scoring functions have to be used.

BioSimz simulations and SwarmDock docking/scoring methods were performed on all of the given structures. Table 3.2 and 3.3 present the scor-

Rank	Model	SD_{rank}	L_{rmsd}	C_{size}	B_{clus}	Ar_{rank}	Al_{rank}	D_{rank}
1)	10	1	1.046	58	Yes	3	1	17.0888
2)	4	1	1.083	19	No	0	0	18.0174
3)	6	1	0.354	74	Yes	3	1	15.632
4)	8	3	1.061	93	Yes	0	0	17.2262
5)	9	2	1.267	45	Yes	0	0	17.6508
6)	5	20	0.854	37	Yes	3	1	17.9287
7)	1	5	0.410	15	Yes	0	1	16.9812
8)	3	8	0.504	17	Yes	0	1	16.0287
9)	11	5	1.185	31	Yes	0	2	17.4597
10)	2	5	0.560	16	No	0	0	16.9510
11)	19	79	0.308	17	No	3	3	17.1189
12)	14	27	0.708	14	No	0	0	17.7264
13)	15	112	0.787	12	No	3	0	18.9913
14)	7	21	0.614	20	No	0	2	15.4912
15)	21	84	1.894	20	No	3	3	17.1113
16)	16	N/A	N/A	N/A	N/A	3	0	19.5187
17)	17	204	1.130	9	No	3	0	18.2994
18)	12	12	6.938	16	No	0	0	18.2452
19)	18	107	2.695	12	No	0	0	19.7383
20)	13	171	2.532	3	No	0	1	17.7149
21)	20	N/A	N/A	N/A	N/A	0	1	16.5006

Table 3.2: Target 43 scoring sheet. Two in-house algorithms were used to rank these models: (1) the SwarmDock flexible docking program using normal modes with particle swarm optimisation; (2) the BioSimz package, using rigid-body Langevin dynamics for simulating crowded macromolecular environments. The criteria shown are for SwarmDock rank (SD_{rank}), ligand RMSD in Å(L_{rmsd}), cluster size (C_{size}, with B_{clus} indicating if the cluster containing the model is the largest cluster). Ar_{rank} and Al_{rank} are ranks for chains A and B respectively, determined using BioSimz. D_{rank} is a score based on model dissociation studies using the BioSimz algorithm.

ing details for the two targets.

A full description of the approach used to rank structures for Target 43 is given here as it further exemplifies the utility of combining the BioSimz and SwarmDock algorithms. Firstly, all the models were separated and re-docked globally using SwarmDock. SwarmDock was run multiple times, searching overlapping patches on the surface of the receptor, from a set of equally distributed points around the receptor, which collectively gave an unbiased coverage of potential solutions. The structures found in these runs were minimised, clustered and ranked. All but five of the complexes (No. 4, 6, 8, 9 and 10) were discarded on the basis that they either did not dock

Rank	Model	SD_{rank}	L_{rmsd}	C_{size}	B_{clus}	Ar_{rank}	Al_{rank}	D_{rank}
1)	1	1	1.167	69	Yes	0	0	16.3744
2)	6	32	1.104	36	Yes	3	2	19.7654
3)	8	15	0.394	15	Yes	0	0	19.1811
4)	4	25	0.797	22	No	0	0	16.9905
5)	5	23	1.230	22	Yes	0	0	18.5996
6)	7	86	1.335	14	No	0	0	16.7699
7)	3	150	0.655	20	No	0	2	17.1649
8)	18	121	0.475	12	No	2	1	17.6775
9)	2	115	4.856	12	No	3	0	18.5311
10)	13	55	0.974	6	No	1	0	17.2667
11)	15	107	1.663	7	No	1	3	18.3724
12)	20	165	0.762	9	Yes	1	0	17.382
13)	11	152	0.666	2	No	3	0	18.4147
14)	21	146	1.936	7	No	0	0	18.2067
15)	16	239	1.277	5	No	0	0	16.7891
16)	19	184	1.063	7	No	0	2	19.0746
17)	9	220	9.044	2	No	2	2	18.829
18)	14	107	5.653	4	No	0	0	19.3307
19)	10	90	7.630	1	No	0	0	18.2486
20)	12	N/A	N/A	N/A	N/A	1	0	19.2216
21)	17	N/A	N/A	N/A	N/A	0	0	17.5926

Table 3.3: Target 44 scoring sheet. Two in-house algorithms were used to rank these models: (1) the SwarmDock flexible docking program using normal modes with particle swarm optimisation; (2) the BioSimz package, using rigid-body Langevin dynamics for simulating crowded macromolecular environments. The criteria shown are for SwarmDock rank (SD_{rank}), ligand RMSD in Å(L_{rmsd}), cluster size (C_{size}, with B_{clus} indicating if the cluster containing the model is the largest cluster). Ar_{rank} and Al_{rank} are ranks for chains A and B respectively, determined using BioSimz. D_{rank} is a score based on model dissociation studies using the BioSimz algorithm.

or had an anomalous ligand RMSD (RMSDs calculated between potential ligand pose solutions and the the experimentally determined ligand docking pose), rank or cluster size. At this stage, models 4, 6 and 10 remained the most promising candidates, since as for over 70% of the docking benchmark 2.0, they docked with a rank of 1. A single large cluster was obtained for each of models 6,8,9 and 10, and this too, from extensive SwarmDock benchmarks, is a good indicator of a native bound complex conformation.

To distinguish between these models, association and dissociation dynamics were studied using the BioSimz package. The 21 receptor and ligand proteins were separated from their respective complexes and put into BioSimz simulations performing rigid-body LD for association. Frequencies at which encounter complexes formed were then monitored against the given complex conformation. Of the five promising models, two of them (No. 10 and 6) formed frequent encounter complexes in keeping with the bound structure; as the benchmark studies showed, this is usually indicative of a genuine, stable, protein-protein interaction.

All model complexes were then put into BioSimz, once again performing LD, but this time starting from a complex state. Multiple runs were performed to investigate their stability in retaining the conformation throughout a certain time period. In this test, Model 10 of T43 performed reasonably – there were a number of trajectories where the complex remained together with a low RMSD to the initial starting complex conformation, and for a relatively long period of time (see Figure 3.9). Model 6, however, dissociated rapidly compared to the other models and known bound structures, see Figure 3.9, trajectory coloured *red*. Meanwhile, the average RMSD path of Structure No. 6 displays larger deviation in step sizes, as the curve is more "wiggly" than that of Structure No. 10. This demonstrates that the latter complex may have a "rougher" binding free energy landscape, making it less of a natural binding pair. On the contrary, the curve of No. 10 appears relatively flat, implying a much smoother process of dissociation (and therefore, association).

Hence, T43 Model 10 was selected as the top choice since, out of the 21 models, it was the only one for which the rank, ligand RMSD, cluster size, association dynamics and dissociation dynamics, were typical of a low energy, and in terms of association/dissociation kinetics, robust, protein-

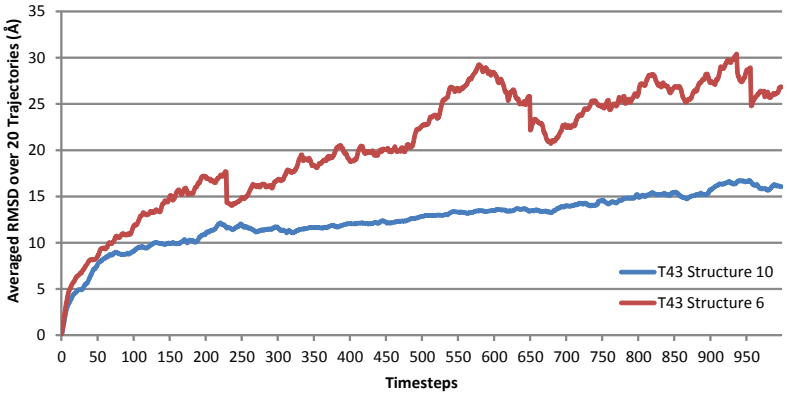

Figure 3.9: Average RMSD plotted for two CAPRI T43 models, over the time course of a BioSimz dissociation simulation. Two structures, No. 10 and 6 are plotted here. Structure No. 10 (*blue* line) has a greater tendency to retain its native binding conformation through the simulation course, compared to Structure No. 6 (*red line*).

protein complex. Later, the selection of Model 10 was verified by the Baker group to be indeed the crystal complex. For T44, however, many of the designed complexes (1, 6, 7, 9-11, 13) failed to express in the bacterial system; the only one that showed weak binding signs was Complex No. 2, which according to BioSimz scores (Ar_{rank} and D_{rank} in Table 3.3) also displayed signs of a moderate binding affinity. Overall, the BioSimz+SwarmDock approach has demonstrated that throughout the blind trials of CAPRI T43 and T44 that, at least for crystal complexes, a signal could be found between genuine and false-positive (or currently, the designed) interfaces between proteins binding partners.

3.4 Discussion

Macromolecular docking is an important problem that has perhaps been perceived, by molecular biologists, to be simpler than it actual is, especially considering the geometric and physicochemical complexity of molecular surfaces, as well as the complexity associated with the crowded and heterogeneous environments the interactions under study are embedded in. Traditional docking methods have focused on finding the exact pose two binding partners exhibit in crystal complexes; under crystal packing conditions,

protein interactions more often than not have one universal conformation, however, this does not mean that all the bound interacting pairs reached the optimal lowest-energy conformation before becoming super-saturated and upon crystallisation. Therefore, as indicated in a number of recent studies, such as the one conducted by Nussinov group (Tsai et al., 2008), an energetically optimal binding plane may exist for natural interactions rather than a single lowest energy binding conformation. This way, the imperfections of protein-protein binding is introduced; therefore, when considering the energetics of protein-protein interactions, it will be important to investigate how protein interactions settle onto this plane, and which paths through conformational space they took to achieve this. The solution of these problems is related to, but far more meaningful than, a one-off pursuit of working out the crystal complex conformation.

In this study, both BioSimz and SwarmDock packages have resorted to dynamics-based motion schemes mimicking the binding process. While a SwarmDock ligand agent tries to benefit the interface searching process from collective intelligence gathered from its neighbours, the total flock of them, when arriving at the receptor surface, resemble such a binding plane, on which each ligand agent occupies a locally energetically optimised pose. Although for the sake of classical docking experiments, the PSO algorithm still needs to settle at one lowest-energy point; it is this author's view, that more information could be retrieved, with respect to the possible binding plane, through the PSO simulations.

Currently, BioSimz emulates this optimal energetic binding plane through simulation of macromolecular interactions until the stage of encounter complex formation. Figure 3.5 implies that a preferential binding region can be observed, with statistical significance, in many of the interaction pairs in our benchmark set; this may be further improved by optimising the scoring scheme. As the diffusion and encountering dynamics for the described test set emulate real *in vitro* environments, preferential regions ("hotspots") may resemble the real contact regions for the binding partners. For many protein-protein interactions, it is intuitive that these regions will fully enclose the crystal-complex binding interface; again this is in agreement with Figure 3.5, in which 20 out of 27 complexes displayed some level of hot-spot/binding site correlation. For CAPRI T40, we even have found

both of the reported binding sites on the receptor in the right affinity order. However, sometimes BioSimz does not provide a positive correlation between hotspot generation and the native binding site. Nevertheless, a careful inspection of hotspots can potentially identify proximal association interfaces to the native, interfaces that initially form before conformational rearrangements occur to form the final complex. For example, the result for CAPRI T39 (Section 3.3.3.2 and Figure 3.8) from a set of BioSimz simulations indicated a neighbouring region, as opposed to the area directly overlapping the binding interface, as the hottest spot for encounter complex formation. Indeed, visual inspection of residues in the hottest regions of both the receptor and ligand, revealed complementary charged residue pairs, that may act as the first contact points. Interestingly, this observation supports a previously proposed binding mechanism (Blundell and Fernandez-Recio, 2006), see Figure 3.10, where the interacting proteins spin and roll over each other's surfaces before locating the native binding site. In the study reporting the data from which this mechanism was proposed (Tang et al., 2006), the authors suggested that these brief encounters, not necessarily bound near the final, established interface, might control not only the kinetics of the assembly process, but also the way the complex is put together. In our view, CAPRI T39 may be a sound example of such a binding process.

3.5 Conclusion

A combined simulation/docking approach to understanding and predicting protein-protein interactions has been employed. For approximately half of the complexes tested, rigid-body Langevin dynamics is sufficient to demonstrate significantly enhanced encounter complex formation at or near the biological interface. We have also successfully located the correct binding region for a number of CAPRI targets and particularly, we have correctly identified the native complex from a cluster of false positives and won the challenge. The protein-protein interaction simulations may also give mechanistic insights into the docking process. Moreover, this information can be used to restrict search space for computational docking and enhance docking results. For the other half, most complexes did not display a significantly diminished docking performance. Based on our predictions

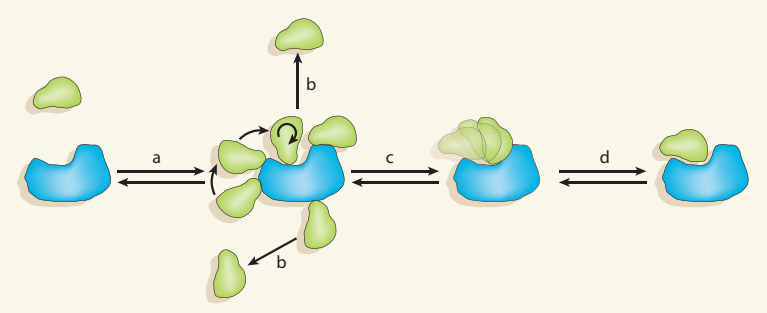

Figure 3.10: The possible binding path through a transient complex. Equilibrium *a* is the formation of transient encounter complexes by nonspecific collisions, guided mostly by electrostatic interactions; *b* indicates that many encounter complexes separate rapidly; in *c*, some productive encounter complexes reorientate and come closer to the final, specific orientation, guided mostly by desolvation, as water molecules move away from the protein surfaces. Equilibrium *d* leads to the formation of the specific complex, with final fitting of interacting surfaces. Illustration taken from Blundell and Fernandez-Recio (2006, Figure 1).

for several blind trial CAPRI targets, we have been able to observe a mode of complex formation that previously could only be hypothesised: initial encounter complex conformations may form that are proximal to the native binding region, then, through a process of two-dimensional searching, the binding partners roll and spin across each other's surfaces to locate the true, lower energy, native binding conformation.

The protein-protein docking studies described here point to the importance of detailed investigations into encounter complex formation, and their associated kinetics, if we are to fully understand macromolecular association/dissociation processes. The research tool developed for this purpose, BioSimz, after extensive benchmarking, appears to perform well. The next chapter explores further the dynamics and kinetics of protein-protein interactions, made possible by the use of this tool.

Chapter 4

Interaction Dynamics and Kinetics

In the context of macromolecular docking, protein-protein interactions are often described as simple ”binary switches”, either forming stable complexes or nonspecific contacts. This rather static and “frozen” (Schreiber et al., 2009) view of macromolecular complexes probably has its origins in the more deterministic studies of macromolecular structure, such as X-ray crystallography, rather than from the wider molecular biology community: illustrations of signalling transduction pathways tend to connect cascading proteins with either a “facilitating” or an “inhibiting” arch, while in crystallography the protein complex in question is either successfully obtained, or failed to crystallise.

The studies of interaction dynamics and kinetics, particularly those of macromolecular association and dissociation processes, added new dimensions in dealing with the problems of protein recognition and interaction.

4.1 Introduction

Protein interactions have a large diversity with respect to the biological functions they facilitate; therefore, it is naturally expected that these interactions bear the same diversity in their association and dissociation processes. For example, enzyme inhibitors usually bind with their receptors with very high affinity and stability, such as BPTI (bovine pancreatic trypsin inhibitor) and barstar (inhibits ribonuclease, barnase). The fast and stable binding of specific inhibitors to their enzyme receptors ensures these enzymes will not function at unnecessary places and thereby cause unwanted damage,

such as degradation of essential proteins or RNAs. In birds and mammals, antibodies also bear high binding affinities with their respective antigens; the loop regions on the variable (V) chains of an antibody are accounted for the specificity in binding. However, although antibody-antigen (Ab-Ag) binding is highly specific, a particularly fast binding response is not required. Therefore, the association speed of an Ab-Ag interaction is usually several orders of magnitude less than that of enzyme-inhibitor associations, while its dissociation rate is usually slow, comparable to that of enzyme-inhibitors. There is a further class of protein-protein interactions, namely signalling-pathway interactions, that have quite the opposite binding/unbinding characteristics to Ab-Ag binding, for example, the interactions between Ras and its effectors. Ras is a small GTPase that controls a number of intracellular signalling processes, and is a central binding partner of upstream and downstream regulatory proteins such as Grb2, Raf and G-proteins. The binding of Ras to PI3K, a kinase associated with cell growth, has a relatively high association rate constant as well as a high dissociation rate constant. This means that both members of the protein pair recognise each other rapidly, however, compared to enzyme-inhibitor and Ab-Ag complexes, the resulting complex also dissociates rapidly. It partially explains the rather promiscuous binding characteristics of signalling pathway proteins; their need to frequently bind and unbind with different partners to regulate cellular processes.

4.1.1 Diffusion

4.1.1.1 Translational diffusion

In a dilute solution, proteins undergo both translational and rotational diffusion before reaching at each other's binding interface and forming a complex. Translational diffusion coefficients for typical proteins span a spectrum of one order of magnitude, from $11.8 \times 10^{-11} \mathrm{m}^2\mathrm{s}^{-1}$ for a 14.4 kDa lysozyme, to $1.20 \times 10^{-11} \mathrm{m}^2\mathrm{s}^{-1}$ for pyruvate dehydrogenase (3.78 MDa). Compared to macromolecular association, *in vitro* diffusion of proteins and other macromolecules attracts relatively little attention, as it is not yet perceived to be of major biological interests. However, the capability to reproduce diffusion *in vitro* is a precondition for any macromolecular simulation

software designed to study binding kinetics. It is therefore important to be able to reproduce the quoted spectrum of diffusion coefficients for proteins, taking account of their wide variety in both shape and weight.

The fundamental mechanism of diffusion, or Brownian motion, was established a century ago (Einstein, 1905; Smoluchowski, 1917); the translational diffusion coefficient D_{tr}, can be described using the Stokes – Einstein relationship

$$D_{\text{tr}} = \frac{k_B T}{6\pi\eta R_S}, \tag{4.1}$$

where R_S is the Stokes radius of the diffusing particle (solute) and η is the solvent viscosity, determined by the type of solvent molecules and temperature of the system. However, Eqn. 4.1 only produces realistic results when the diffusing particle is a perfectly smooth sphere; for rod-shaped objects or those with rough surfaces, this relatively simplistic treatment may have a significant bias.

Cytoplasmic proteins are mostly spherical molecules; those that are not spherical tend to be at least ellipsoidal. Therefore, the radius of gyration is often used as an approximation of a protein's Stokes radius. For a protein consisting of N atoms, its radius of gyration is expressed as

$$R_{\text{g}}^2 \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{r}_i - \mathbf{r}_{\text{cm}}\right)^2, \tag{4.2}$$

where $\mathbf{r}_{\text{cm}}$ is the position vector of the centre of mass of the protein, and $\mathbf{r}_i$ is the position vector of atom i. Tyn and Gusek (1990) calibrated D_{tr} predictions on a dataset of 198 proteins (from 86 different types) with experimentally measured diffusion coefficients, and proposed the following empirical equation,

$$D_{\text{tr}} = \frac{5.78 \times 10^{-15} T}{\eta R_{\text{g}}}, \tag{4.3}$$

which yielded a accuracy of 87.4% at $\pm$20% accuracy. However, some rod-like proteins need to have their calculated D_{tr} adjusted in order to achieve the claimed correction rate. The work of He and Niemeyer (2003) eliminated this requirement by introducing the correlation with molecular weight, resulting in

$$D_{\mathrm{tr}} = \frac{6.85 \times 10^{-15} T}{\eta \sqrt{M^{1/3} R_{\mathrm{g}}}}, \tag{4.4}$$

where the square-root term accounts for the compensation of drift from the Stokes radii for molecules that have an irregular shape, such as ellipsoidal or cylindrical. The authors fitted and benchmarked Eqn. 4.4 with 203 experimentally measured diffusion coefficients, and achieved a rate of 86.7% in correct predictions within a $\pm$15% deviation error limit, and without the need of special adjustment for irregular-shaped proteins.

4.1.1.2 Rotational diffusion

Compared to translational diffusion, rotational diffusion has been reported less in studies of protein dynamics. Molecular dynamics packages need not deal with rotational diffusion, as the objects in motion are spherical atoms; diffusion models that treat proteins as spheres require no rotational diffusion either. However, the accurate reproduction of rotational diffusion dynamics is a key pre-condition to the investigation of the orientational adjustments made by two proteins upon forming an encounter complex. Analogous to translational Brownian motion (see Section 2.1.1), the rotational diffusion coefficient can be expressed in the following Langevin equation

$$\mathbf{I}\frac{d\omega}{dt} = -\zeta_{\mathrm{rot}}\omega(t) + \sum_{i=1}^{N} \mathbf{P_i}(t) + \mathbf{B}_{\mathrm{rot}}(t), \tag{4.5}$$

where I is the moment of inertia with respect to the rotating centre, w is the angular velocity, $\mathbf{P_i}$ is the torque exerted on the i-th particle with respect to the rotating centre, and $\mathbf{B}_{\mathrm{rot}}(t)$ is a three-dimensional Wiener process, the one-dimensional form of which is written as

$$< B_{\mathrm{rot}}(t) > \quad = 0, \tag{4.6}$$

$$< B_{\mathrm{rot}}(t) B_{\mathrm{rot}}(t_0) > \quad = 2k_B T \zeta_{\mathrm{rot}} \delta(t - t_0). \tag{4.7}$$

Note that the rotational Langevin equation has the exact same form as the translational, only with changes of notation to their angular equivalent. The average thermal angular velocity is

$$< |\omega(t)| >= \sqrt{\frac{3k_BT}{I}}. \tag{4.8}$$

However, the rotational diffusion coefficient, D_{rot}, is less used in evaluation of the random angular deviation of a molecule compared to D_{tr}. This is because D_{rot}, measured in unit of $\text{rad}^2\text{s}^{-1}$, has periodical boundaries at $\pm 2\pi$. It is apparently difficult to measure D_{rot} by angular displacements ($-2\pi < \theta < 2\pi$) over unit time periods, since such displacements are periodical. A rotational relaxation time τ, defined as the average time required for the molecule to be displaced from its original orientation by a mean angle θ_0, is more frequently used for benchmarking rotational diffusion. For typical *fluorescence polarisation* measurements,

$$\theta_0 = \arccos\left(e^{-1}\right). \tag{4.9}$$

Without loss of generality, assume the rotating object (molecule) represented in Eqn. 4.5 contains a dipole, on which the external torques $\mathbf{P}$ are exerted, from a single field that has strength $\mathbf{E}$, we have

$$\mathbf{P} = \mu\mathbf{n} \times \mathbf{E}, \tag{4.10}$$

$$\frac{d\mathbf{n}}{dt} = \omega \times \mathbf{n}, \tag{4.11}$$

where $\mathbf{n}$ is a unitless directional vector and μ is the dipole moment. The rotational friction coefficient ζ_{rot} for a sphere is

$$\zeta_{\text{rot}} = 8\pi\eta r^3, \tag{4.12}$$

which suggests the rotational friction is generally much larger than its translational equivalent for molecules of large radii, such as proteins. Therefore, it can be assumed that the rotational Brownian motion is non-inertial and Eqn. 4.5 can then be re-written as,

$$\frac{d\mathbf{n}}{dt} = \frac{\mathbf{B}_{\text{rot}}(t) \times \mathbf{n}}{\zeta_{\text{rot}}} + \mu\mathbf{E} \cdot (\mathbf{1} - \mathbf{n} \times \mathbf{n}). \tag{4.13}$$

The general Smoluchowski diffusion equation is expressed as

$$\frac{\partial \mathbf{p}}{\partial t} + \nabla \cdot (\mathbf{j}_d + \mathbf{j}_f) = 0, \tag{4.14}$$

where $\mathbf{p}$ is the distribution function in rotational coordinate space, $\mathbf{j}_d$ is the current density due to diffusion and $\mathbf{j}_f$ is the current density due to the field $\mathbf{E}$. Combining this equation with the re-organised Eqn. 4.13 gives

$$\mathbf{j}_d = -D_{\text{rot}} \nabla \mathbf{f}, \tag{4.15}$$

$$\mathbf{j}_f = f(d\mathbf{n}/dt) = \tfrac{\mu}{\zeta_{\text{rot}}} \mathbf{E} \cdot (\mathbf{1} - \mathbf{n} \times \mathbf{n}) f. \tag{4.16}$$

The solution of Eqn. 4.16 (Margenau and Murphy 1943, Section 5.2-5.4, Mazo 2009, Section 15.1) introduces a function $g(t)$ of the form

$$\frac{dg}{dt} = -\frac{2k_B T}{\zeta_{\text{rot}}} g, \tag{4.17}$$

which in turn has the solution

$$g = e^{-t/\tau_D}, \tag{4.18}$$

where

$$\tau_D = \frac{\zeta_{\text{rot}}}{2k_B T} = \frac{1}{2D_{\text{rot}}}. \tag{4.19}$$

is the Debye relaxation time. The second equality is derived from the Stokes-Einstein relationship.

Other forms of relaxation times are also quoted, such as the orientational relaxation time (τ_2) measured through fluorescence anisotropy and nuclear magnetic resonance (NMR) experiments. It has also been shown and verified that

$$\tau_2 \simeq \frac{\tau_D}{3} = \frac{1}{6D_{\text{rot}}} \tag{4.20}$$

in which the equalisation applies to spheres ($I_{xx} = I_{yy} = I_{zz}$), and the approximation applies to asymmetric tops ($I_{xx} \neq I_{yy}$ and/or $I_{yy} \neq I_{zz}$) (Sack, 1957; Ford et al., 1979; Coffey et al., 2002).

Actual measurements of D_{rot} and τ_2 for molecules of various sizes are regularly reported from experimental studies using fluorescence polarisa-

tion and NMR, although for proteins, the relevant measurement data are not frequently reported. Typical values of τ_2 vary greatly depending on the size and shape of the diffusing molecules. For example, τ_D of liquid chloroform ($CHCl_3$, 119Da) is 6.4ps ($\tau_2 = 2.13$ps) (Coffey et al., 2002), while a mid-sized protein, concanavalin A (51.3kDa), has τ_2 measured at 58ns (Inbar et al., 1973). A larger protein, aspartate aminotransferase (89.5kDa, sedimentation coefficient $s = 5.5$S) was determined to have a τ_2 of 130ns (Churchich, 1967), which is approximately 20,000 times longer than that of chloroform molecules. The same study also suggested that for a 5kDa protein segment, $\tau_2 \approx 9.0$ns. More interestingly, the study of Schuldiner et al. (1975) showed the significant changes in τ_2 that can occur for a small-molecule substrate before and after binding to its specific enzyme. Free (unbound) 2-(N-dansyl)aminoethyl β-D-thiogalactoside (dansylgalactoside, DG_2, 533Da) molecules were measured with $\tau_2 = 660$ps, while the same relaxation times for nonspecifically and specifically bound molecules were measured at 21.6ns and 150.0ns, respectively. The receptor enzyme in the above study, lactose permease (LacY), is a globular protein sized of 93.0kDa. It becomes apparent from the authors' data that rotational diffusion of DG2 was partially restricted, allowing for a loose, nonspecific, encounter with LacY. The final complex of the enzyme and substrate DG_2, however, does not allow relative rotational movement between them, as the relaxation time for bound DG_2 is comparable to a protein of similar size, such as the τ_2 of the separately studied aspartate aminotransferase. This implies that the final binding is as tight as a single covalently bonded structure; a much later structural study verified and revealed that lactose binding occurs within a deep cavity of LacY, resulting in the substrate molecule being highly restricted it its ability to translate or rotate in its bound position (Abramson et al., 2003).

A later study to those discussed above, based on the hydrodynamic properties of proteins, reviewed and computationally predicted D_{tr} and D_{rot} for 13 proteins from 6kDa to 230kDa, using their own computer program, HYDROPRO, on atomic-resolution models (Garcia de la Torre et al., 2000); this was the successor to their earlier attempt to calculate rotational diffusion coefficients using a bead model (Garcia de la Torre et al., 1987).

4.1.2 Association

The association and dissociation of biomacromolecules are dynamic processes resembling the chemical combination and decomposition reactions, albeit without the formation and deformation of chemical bonds. Therefore, representation and theories for macromolecular complex formation can also be borrowed from chemical reactions that describe small molecules. One of the more accepted representation of macromolecular interaction is as follows

$$A \; + \; B \; \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} \; A:B \; \underset{k_{-2}}{\overset{k_2}{\longleftrightarrow}} \; AB, \tag{4.21}$$

where $A : B$ denotes the "intermediate state" of the binding partners and AB is the final complex. This equation represents a two-stage process, in which the first stage involves diffusion and collision of molecules A and B, and the second stage covers the transition of the encounter complex $A : B$ into the established complex AB by overcoming a high-energy transition state (Eyring, 1935). The observable rate constants are accordingly calculated as

$$\begin{aligned} k_{\mathrm{d}} &= \frac{k_{\mathrm{off}}}{k_{\mathrm{on}}}, \\ k_{\mathrm{on}} &= \frac{k_1 k_2}{k_{-1} + k_2} \approx k_1, \\ k_{\mathrm{off}} &= \frac{k_{-2}}{k_1}, \end{aligned}$$

where k_{d} is the equilibrium binding constant, k_{on} is the association rate constant, and k_{off} is the dissociation rate constant. The association rate constants (k_{on}) for protein-protein interactions normally vary between 10^2 and $10^9 \mathrm{M}^{-1}\mathrm{s}^{-1}$, while the dissociation rate constants (k_{off}) vary between 10^{-6} and $10^2 \mathrm{s}^{-1}$. As was explained in Section 2.1.1, the physical nature of diffusion can be described by the Langevin equation. A basal collision rate constant,

$$k_{\mathrm{on}}^0 = 10^5 \sim 10^6 \mathrm{M}^{-1}\mathrm{s}^{-1}, \tag{4.22}$$

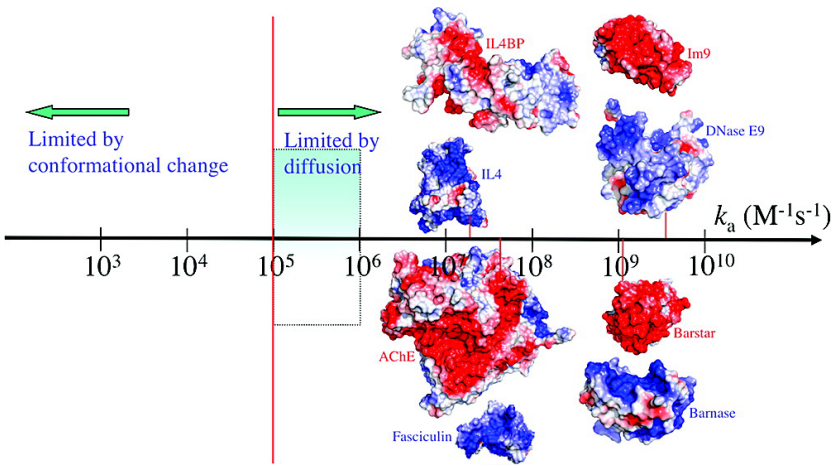

Figure 4.1: Wide spectrum of association rate constants. The red vertical line marks the start of the diffusion-controlled regime. The shaded range marks the absence of long-range forces. Illustration taken from Schreiber et al. (2009, Figure 1).

can therefore be computed based on Langevin diffusion for spheres at the size of a normal protein ($r = 20 \sim 30$Å) (Zhou, 1997; Schlosshauer and Baker, 2004). For proteins that have $k_{on} < k_{on}^0$, the reaction rates are said to be limited by conformational changes; for proteins that have $k_{on} > k_{on}^0$, the rates are diffusion controlled. An illustration of the wide spectrum of k_{on} values is shown in Figure 4.1.

A noticeable approximation in Eqn 4.22 needs to be taken into account when calculating k_{on}; the assumption that $k_2 \gg k_{-1}$ under normal conditions, implying the probability of the state transition from $A : B$ to AB is far larger than the probability that the encounter complex will revert back to the standalone binding partners. Clearly, this condition is only met when k_{on} is diffusion-limited. Accordingly, all of the proteins that have been referred to or investigated in this study do have an experimental k_{on} above the basal rate k_{on}^0. Proteins that have slower binding processes ($k_{on} < k_{on}^0$) are most often limited by their need to undergo large internal conformational changes, and are thereby out of scope for the kinetic studies carried out for this chapter.

The first attempt into deriving the theoretical association rate constant was to follow Smoluchowski's theory on reaction kinetics of two spheres (Smoluchowski, 1917),

$$k_{\text{on}}^{0*} = 4\pi Dr, \tag{4.23}$$

where r is the Stoke's radius of the solute and D is the (translational) diffusion coefficient. When considering the interaction potential between the two spheres, the above equation leads to an increased rate constant, as was initially described by Debye (1942),

$$k_{\text{on}}^{0*}(a) = 4\pi D \left(\int_a^{\infty} \exp\left(\frac{U(r)}{k_B T} \right) r^{-2} dr \right)^{-1}, \tag{4.24}$$

where $U(r)$ is the interaction potential energy, k_B is the Boltzmann constant and T is the temperature. Evaluation of this integral with $a = 20\text{Å}$ and $D = 2 \times 10^{-7}\text{Å}^2\text{ns}^{-1}$ gives a basal rate $k_{\text{on}}^{0*} = 6 \times 10^9\text{M}^{-1}\text{s}^{-1}$, which is approximately three to four orders of magnitudes larger than the previously cited basal rate k_{on}^{0} (Schreiber et al., 2009). It is apparent that the difference is due to the specific directional requirement for proteins to form a complex. However, by counting only protein-protein interactions in the correct binding direction, which approximately covers an area spanning 5° on the arc of each protein's sphere representation, one obtains a 10^6-fold reduction of the theoretical basal rate. The disparity between the corrected k_{on}^{0*} and the suggested k_{on}^{0} attracted an early investigation using Brownian dynamics simulations, in which the authors suggested there might be multiple hits per collision with slightly twisted (rotated) orientations each time, which leads to quick resampling of the collision directions thereby increases k_{on}^{0} (Northrup and Erickson, 1992). However, these estimations, based on Smoluchowski's and Debye's approximated diffusion and reaction models, neglected all detail concerning molecular surfaces; the subsequent explanation of compensation by multiple hits per collision was also based on a spherical approximation for each interacting protein. In this simplified case, the roughness of molecular surfaces, the short-range interaction forces, as well as desolvation effects, are neglected all together. Hence, the speculation on rate compensation by rotational movements was still subject to a thorough investigation employing more comprehensive simulation methodologies.

To date, there have been two classes of algorithms proposed for directly estimating the diffusion-controlled association constants between two molecules, taking into account external potential functions. The first is a

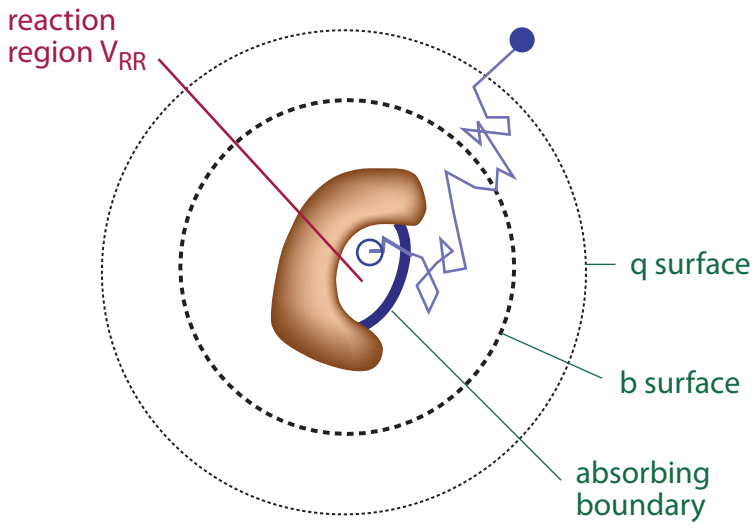

Figure 4.2: Boundary conditions of the Northrup and Zhou algorithms for calculation of k_{on}. The boundary conditions used in the algorithm of Northrup et al. (1984) was illustrated in *green* text, while the boundary condition used by Zhou (1990) is presented in *red* text. See main text for detail description of both algorithms.

distance-based scheme, proposed by McCammon and coworkers (Northrup et al., 1984), which expresses

$$k_{\text{on}} = 4\pi D b S_{\infty}, \tag{4.25}$$

given

$$S_{\infty} = \frac{S}{1 - (1 - S)k_{\text{on}}(b)/k_{\text{on}}(q)} \tag{4.26}$$

where $k_{\text{on}}(b)$ and $k_{\text{on}}(q)$ are calculated by Eqn 4.24. The factor S is the fraction of encounters that satisfy the boundary conditions for formation of a native complex. In order to correctly parameterise the equation, initiation and termination boundaries b, q $(q > b)$, should be set as far from the receptor centre as possible, thereby eliminating the influence from long-range potentials from being counted into the basal rate (see Figure 4.2).

The second method, based on reaction region volume V_{RR} (see Figure 4.2), was proposed by Zhou (Zhou, 1990). In his treatment, the association

rate constant for diffusion-controlled binding is given by,

$$k_{\text{on}}^{\text{diff}} = \gamma V_{\text{RR}} \exp\left(\frac{-\langle U\rangle^*}{k_B T}\right) \frac{S}{1-S}, \tag{4.27}$$

and for transition-state controlled binding by

$$k_{\text{on}}^{\text{react}} = \gamma V_{\text{RR}} \exp\left(\frac{-\langle U\rangle^*}{k_B T}\right). \tag{4.28}$$

In both expressions, $< U >^*$ is the average interaction energy throughout the encounter complex stage. Given the assumption that all encounters within S will settle into final complexes according to the second equilibrium reaction of Eqn. 4.21, $k_{\text{on}}^{\text{diff}}$ can be written as a function of the basal rate

$$k_{\text{on}}^{\text{diff}} = k_{\text{on}}^{0} \exp\left(\frac{-\langle U\rangle^*}{k_B T}\right), \tag{4.29}$$

where

$$k_{\text{on}}^{0} = \frac{\gamma V_{\text{RR}} S}{1-S}. \tag{4.30}$$

4.2 Methods

4.2.1 Diffusion

BioSimz simulation protocol for protein diffusion follows the Langevin equations for translational and rotational Brownian motion (Eqns. 4.5 and 2.24). For each of the simulated proteins, their translational friction coefficient ζ_{tr} is calculated from the protein's diffusion coefficient, which is in turn predicted based on the protein's radius of gyration and molecular weight (He and Niemeyer, 2003). Therefore,

$$\zeta_{\text{tr}} = \frac{k_B T}{D_{\text{pred}}}, \tag{4.31}$$

given by the Einstein relationship. The rotational friction coefficient, ζ_{rot}, is computed using the following procedures. Given the Stokes' law for rotational motion of a sphere (Brilliantov and Krapivsky, 1991),

$$\zeta_{\text{rot}} = 8\pi\eta r^3, \tag{4.32}$$

and the Einstein relationship (exists also for rotational Brownian motion) for spheres,

$$D_{\text{rot}} = \frac{k_B T}{\zeta_{\text{rot}}}, \tag{4.33}$$

we have

$$D_{\text{rot}} = \frac{k_B T}{8\pi\eta r^3} = \frac{k_B T}{6\eta V} = \frac{k_B \rho}{6} \cdot \frac{T}{M\eta}, \tag{4.34}$$

where V is the volume of the diffusing body (spherical) and ρ, M are the density and mass of the diffusing body respectively. Clearly, the result of Eqn. 4.34 reduces to a constant value for objects of a similar density, such as proteins. Therefore, it is possible to obtain a linear regression of D_{rot} with respect to $T/M\eta$ using known D_{rot} and τ_2 from molecules of varying sizes. Eqn. 4.34 was fitted to the experimental D_{rot} reported in (Garcia de la Torre et al., 2000), yielding the linear relationship,

$$D_{\text{rot,pred}} = 0.0856 \cdot \frac{T}{M\eta}. \tag{4.35}$$

Predicted rotational diffusion coefficients can then be used to work out friction coefficients, ζ_{rot}, using the Einstein relationship.

Simulations are thereby performed with both ζ_{tr} and ζ_{rot}, predicted from regression data of experimentally verified diffusion coefficients, for each particular protein under study. Trajectory data are then collected to computationally measure the actual translational and rotational D. The D_{tr} is measured by taking the mean-squared displacements over a set time period, while rotational diffusion is evaluated by measuring τ_2, i.e. the average time taken for protein molecules to rotate $\arccos(1/e) = 68.1^\circ$ about an arbitrary axis. Eqn. 4.20 is used to convert τ_2 back to D_{rot} for comparing with experimental D_{rot} data for each of the proteins studied.

A test set of 10 proteins was used to benchmark the diffusion simulations (Garcia de la Torre et al., 2000). All calculations were based on displacements averaged over 20 trajectories for each simulated protein. Diffusion simulations were run over 500ns (biological time). For translational diffusion coefficients, experimental D_{tr} were benchmarked; for proteins without experimentally-measured D_{rot}, predicted values are used from the literature

(Garcia de la Torre et al., 2000).

4.2.2 Association

Association of diffusing proteins is assessed by monitoring the structural similarity between their interactions and the reference, experimentally determined, complex. In addition, the frequency of protein-protein interactions near the reference structure's conformation, and the length of time these interacting pairs spent in the proximity of the reference conformation, are recorded. Quantitative measures used for investigating these aspects are full-atom RMSDs and BioSimz binding scores, introduced in Section 2.1.4.

The benchmarking protein set for studying association is different to the benchmarking set used for demonstrating and reproducing *in vitro* diffusion parameters. This is due to the fact that the protein/protein pairs used in experimental studies for measuring diffusion and association did not overlap. For association, eleven known complexes were used. This included enzyme-inhibitor complexes barnase-barstar (PDB:1B27), DNase E7-Im7 (PDB:7CEI); enzyme-substrate complexes adrenodoxin-adrenodoxin reductase (PDB:1E6E), ferredoxin-ferredoxin reductase (PDB:1EWY); signal-effector complexes Ras-GAP (PDB:1WQ1), Ras-PI3K (PDB:1HE8), TGFβ3-TβRII (PDB:1KTZ), Cdc42-Cdc42GAP (PDB:1GRN), CDK2-CksHs1 (PDB:1BUH); immuno-protein complexes CD2-CD58 (PDB:1QA9) and Adenovirus Knob 35-CAR (PDB:1KAC).

Simulations were run in a box sized at $240 \times 240 \times 240$Å. Eight receptors and eight ligands were put into the box for each run, making up a total protein concentration of 1.92mM. Each simulation run was performed for a biological time of 0.5μs, and for each interaction pair, such simulations were repeated 20 times – to provide better sampling introduced by the initial placement of proteins within the simulation box. The total simulation time for each interacting pair therefore sums to 10μs; a significant time period for the simulation of biological events. The simulation temperature was maintained at 310K (37°C), and the viscosity constant was set to 6.915×10^{-4}Pa·s, which is the viscosity of fluid water at 310K (Wangemann and Liu, 1996). The simulations were first performed using bound receptors and ligands separated from the complex structure; a second batch of simulations

were performed with unbound structures taken from their individual PDB entries. To inspect their possible influence on protein binding dynamics, artificially increased translational and rotational diffusion coefficients were used in some of the benchmark simulations.

4.2.3 Dissociation

To further study the macromolecular interaction process, simulations of unbinding were also performed. In each of these simulations, one receptor and one ligand are put into the same simulation box used in association simulations, with the same parameters and components of the forcefield, except for the generation of random forces and torques (see next paragraph). The receptor and ligand are set to the bound conformation initially, and start to gradually dissociate as the perturbations of Brownian motion accumulate and consequently disrupt the established interface contacts. For a detailed discussion of the validity on simulating complex dissociation, and the implication of the outcome, please refer to the relevant paragraphs in Section 4.4.

For the purpose of enhancing dynamic movements during the dissociation process, necessary to restrict the time course of the simulations, friction coefficients are artificially increased. This can be easily justified, as under normal diffusive conditions, the dissociation rate constants, k_{off}, are usually quite small – even for the fastest-dissociating complexes, it is true that $k_{\text{off}} \ll 1 \times 10^6 \text{s}^{-1}$. The average lifetime of these complexes is therefore much longer than a few μs, exceeding the maximum magnitude of length of a single simulation. At any rate, given the current lack of precision in the model with respect to the accuracy of modelling H-bonds, desolvation and direction-preferential Brownian motion (no significant solvent collisions between interfaces), to obtain real-term dissociation rate constants isn't feasible. Moreover, even a short average lifetime for a complex can last hundreds of thousands of timesteps, making the measurement of dissociation difficult, especially when different complexes have k_{off} values that differ by several orders of magnitude. Given the above, a naive multiplier of 10 is therefore applied to the translational friction coefficient ζ_{tr} and a multiplier of $\sqrt[3]{10}$ is applied to the rotational friction coefficient. There have been few studies on simulating complex dissociation (Nesatyy, 2002; Swegat et al.,

2003), and even for these no comparisons between dissociations of different protein complexes were made. Consequently, it has not been possible to compare parameters used in this dissociation study with any other; the two multipliers are thereby deemed to be a necessary compromise for the simulation of dissociation events explored here.

The benchmark complex set used for the dissociation studies is the same 26-complex data set used for the docking studies (see Chapter 3 and particularly, Section 3.2.4), which is also a super set of the 11-complex data set used for the association studies. Each dissociation simulation was run for 30,000 timesteps; since the solution drag coefficient, ζ, was artificially increased to speed up the dissociation simulations, this timestep cannot be directly compared to real time.

A scoring scheme is developed to evaluate the stability of target complexes during dissociation. The mathematical form of the raw scoring function is expressed as

$$S^*_{\text{tot}} = \int_0^T w(t) \cdot c_1 s_1(t) \cdot c_2 s_2(t) \cdot c_3 s_3(t) \cdot (d_4 - \overline{d_{\text{rmsd}}(t)}) dt, \tag{4.36}$$

where

$$c_i s_i = b(d_i - d_{\text{rmsd},n}) c_i \sum_{n=1}^{N} H(d_i - d_{\text{rmsd},n}), i = 1, 2, 3 \tag{4.37}$$

in which $H(\cdot)$ is the Heaviside step function used as a binary signal delivery function. The three RMSD thresholds, $d_1 = 6.0$Å, $d_2 = 12.0$Å, $d_3 = 25.0$Å are set as the borderlines of bound, transient and distant encounter complexes, and are thereby given different weights, c. In this study, $c_1 = 50.0, c_2 = 5.0$ and $c_3 = 1.0$. Parameter $\overline{d_{\text{rmsd}}(t)}$ is the averaged RMSD over all trajectories at the designated timestep t. The maximum RMSD allowed to contribute to the score, d_4, is set to 26.0Å as is used in Eqn. 4.36. Snapshot conformations bearing higher RMSDs than d_4 are dropped, since they are no longer thought to represent interacting proteins, or protein interactions located in the right hemisphere, the centre for which is located at the centre of mass of the experimentally determined binding interface. Since in Eqn. 4.36 the use of multiplication naturally amplifies the differences in sub-scores, an inverted sigmoid function is assigned as the overall

weighting and smoothing function, $w(t)$, which has the form

$$w(t) = \begin{cases} \dfrac{1}{1+\exp\left(-\dfrac{20t}{T}\right)} - 0.5, t < T/2, \\ \dfrac{1}{1+\exp\left(\dfrac{20(1-t)}{T}\right)} + 0.5, t \geq T/2. \end{cases} \tag{4.38}$$

To compensate for the non-linearity resulting from the above multiplication of parameters, the log of the total score S^* is taken, the maximum value for which is scaled to 100,

$$S = \frac{\log S^*}{\log S_{\max} - \log S_{\min}} \cdot 100. \tag{4.39}$$

The final score, S, is the dissociation score used in this study for complex dissociation processes.

4.3 Results

4.3.1 Diffusion

For both translational and rotational benchmark simulations a good correlation was ultimately achieved between experimental and theoretical diffusion coefficients. The calibrated translational diffusion coefficients, measured by taking the mean-squared displacement over a set time period, have a correlation coefficient $\sigma = 0.985$ against experimentally verified translational diffusion coefficients (see Figure 4.3). Deviation between simulations and experiments tend to occur for smaller proteins, such as trypsin (6.52kDa, PDB: 4PTI) and hen egg-white lysozyme (HEWL, 14.3kDa, PDB: 6LYZ). This may reasonably be expected, since the method employed to theoretically generate D_{tr}, although corrected, still biases towards irregular shaped objects. Smaller proteins have relatively larger surface irregularity; the roughness of the surface, relative to the protein's radius, thereby facilitating rapidly altering frictional forces between interacting surfaces. In comparison, larger proteins, such as fructose 1,6-biphosphate aldolase (157kDa, PDB:1ADO), normally have very stable and predictable diffusion behaviour.

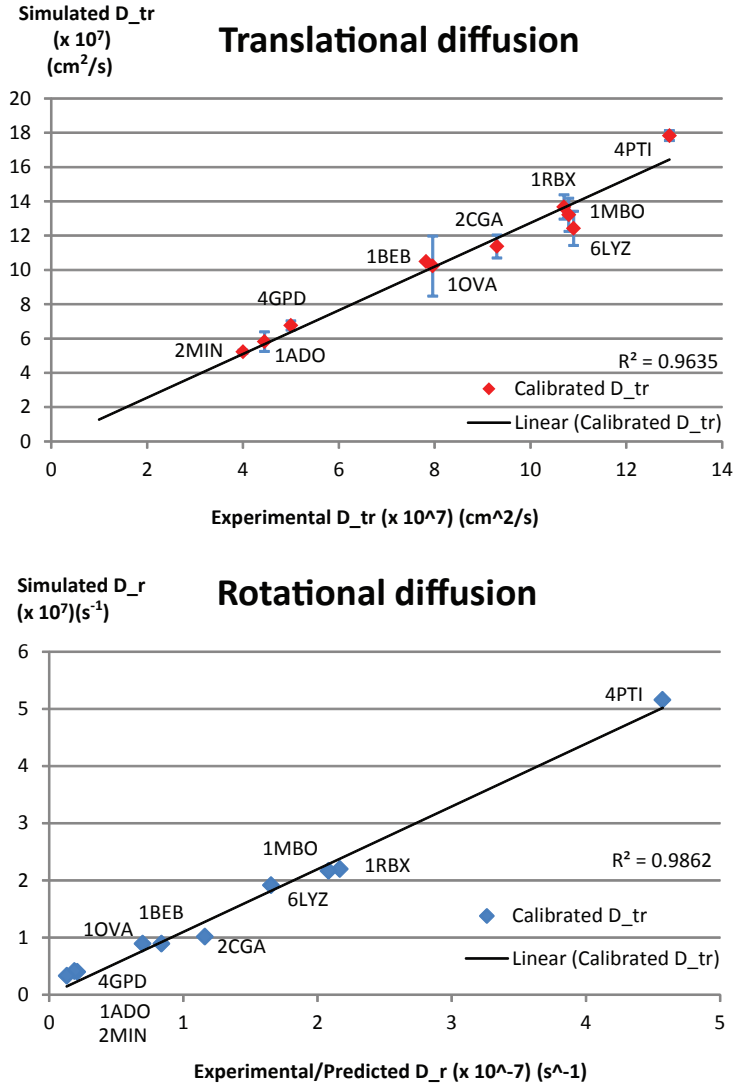

Figure 4.3: Simulated vs experimental/theoretical diffusion coefficients. For translational diffusion, the simulated D_{tr} are plotted against experimentally verified values. For rotational diffusion, the simulated D_{rot} are plotted against experimentally verified and theoretically predicted values. All data for verification purposes were taken from Garcia de la Torre et al. (2000).

Rotational diffusion is also replicated well in the simulations. There is an overall high correlation, $\sigma = 0.993$, between the theoretical D_{rot} values calculated in this study and the theoretical values determined by Garcia de la Torre et al. (2000). Moreover, there is also a high correlation, $\sigma = 0.963$, with experimentally determined D_{rot} vales, see Figure 4.3. It is interesting that rotational diffusion has exhibited greater conformity to experiments and theoretical calculations, compared to translational diffusion. This could be partly due to the fact that the rotational friction coefficient, ζ_{rot}, changes on the same order as the scale of molecular weight (see Eqns. 4.33 and 4.35), rather than the molecule's radius, which is one third of an order of the molecular weight as in the case of translational diffusion. This means that rotational Brownian motion is more sensitive to the change of molecular size than its translational counterpart, which implies the "hills and valleys" on the molecular surface may not have a significant biasing effect on D_{rot}.

Of course, one has to bear in mind that D_{rot} would be significantly directionally biased if the overall shape of the molecule is not spherical. Although the solvent collisions are essentially uniformly distributed across the macromolecular surface, the moment of inertia can be significantly different with respect to the axis about which the molecule rotates. A more precise way of representing D_{rot} for non-globular macromolecules would be to use a 3×3 rotational diffusion tensor, corresponding to the inertia tensor of the underlying macromolecule. However, as the proteins used for diffusional and association studies here are mostly highly globular, full three-dimensional diffusion tensors are deemed an unnecessary addition to the model at present. Similarly, when hydrodynamic forces are to be considered, a 3×3 translational diffusion tensor is required instead of a simple D_{tr}; for the same reason (globular proteins) translational diffusion tensors and any possible hydrodynamic effects except for rotational Brownian motion, are also omitted from this study.

4.3.2 Association

4.3.2.1 Correlation with k_{on}

The same diffusional environment, as used in the previous section, is applied to investigate the association dynamics of specific macromolecular in-

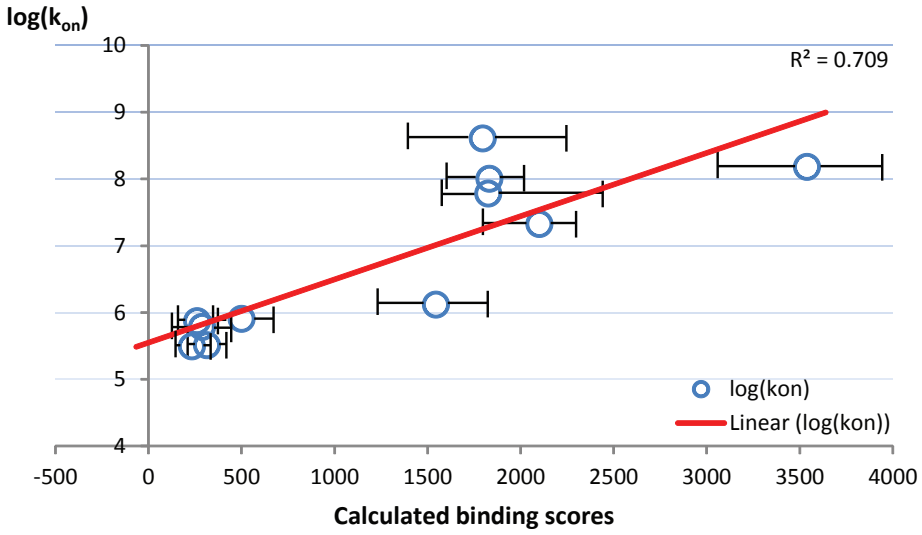

Figure 4.4: Binding scores vs. k_{on} in log space. Experimental measurements of k_{on}, plotted on the y-axis, are sourced from different studies (see Table 5.2). The line in *red* is a linear regression of the $\log(k_{on})$ and binding score value pairs; the corresponding correlation coefficient is 0.84.

teractions and, in particular, encounter complex formation. Binding scores (Sections 2.1.4.2) are calculated from the simulations of eleven complexes and are compared to their experimentally measured association rate constants. Results show that the binding scores correlate reasonably well with experimental $\log k_{on}$, of which the correlation coefficient ρ is 0.84 (see Figure 4.4).

4.3.2.2 Hotspots and binding dynamics

Trajectories of interaction events were analysed; the encounter complexes were superimposed with respect to the receptor (arbitrarily chosen between the two binding partners). Regions of the receptor surface that are more frequently accessed by ligands during the simulations are designated "hotspot" areas (see Figure 4.5). It is intuitive to infer that hotspots should overlap or be close to a protein's binding region; after examination of test cases we found this to be true for all of the eleven test set cases except for Cdc42/Cdc42-GAP (PDB:1GRN). Figure 4.5 shows the contact frequencies during the simulation for association of CDK2-CksHs1 (PDB:1BUH). It is

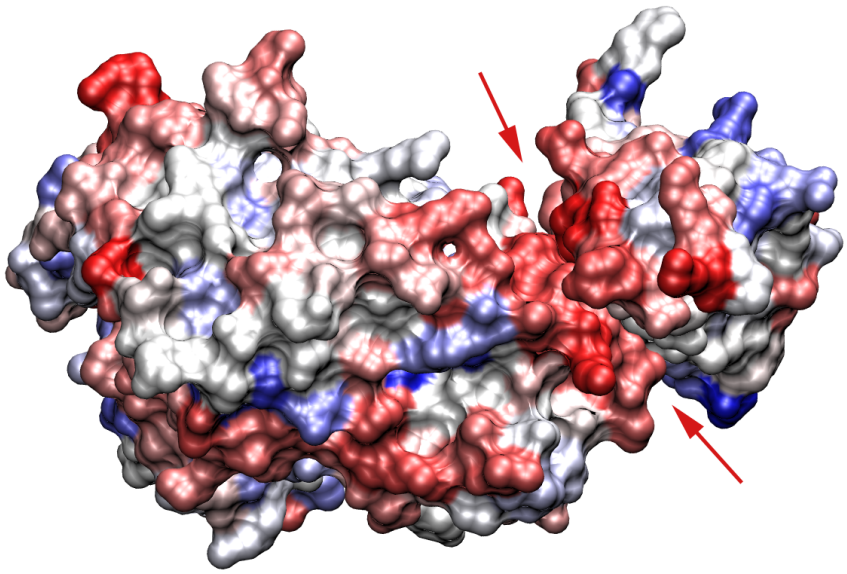

Figure 4.5: Hotspots (high-frequency contacting regions on complex CDK2-CksHs1 (PDB:1BUH). The molecular surface is coloured as a blue-white-red gradient heatmap showing the different contact frequencies the binding partners encountered during a total simulation period of 1μs. Blue means the number of contacts in this region is lower than average; white means the number of contacts is about average. Red means there are more contacts in this region than the average frequency. The binding interface is formed between the two arrows.

clear from the illustration that the hottest regions on both the ligand and receptor are the correct binding interfaces. In a second example (PDB:1KAC), the coxsackie and adenovirus receptor has the hottest spot directly over its binding interface; the ligand, Adenovirus Knob 35, has the most frequently visited region adjacent to the correct binding interface. This, and many other cases from our simulation test set, again, support the much discussed the spinning-rolling mechanism for protein-protein association (Blundell and Fernandez-Recio, 2006; Tang et al., 2006), which has been discussed in detail in Section 3.4 and in an accompanying paper recent published (Li et al., 2010).

With relevance to the kinetics, this mechanism may have accounted for the higher k_{on} (10^7 to $10^8 M^{-1}s^{-1}$) of many specific protein-protein interactions. As was described in Section 4.1.2, geometric encounter rates, under position and orientation constraints, are approximately $10^5 \approx 10^6 M^{-1}s^{-1}$ (Schlosshauer and Baker, 2004). Any k_{on} higher than this basal rate would have to rely on some forms of "encounter dynamics" after the initial colli-

sion occurs.

4.3.2.3 Native and designed interfaces

Further predictive tests were performed on two relatively large data sets. The first set is the Docking Benchmark V2.0 (Mintseris et al., 2005), in which all of the 120 member complexes have their structures determined by protein X-ray crystallography; the second is a 80-strong list of computationally designed complexes offered as CAPRI Target 35 by David Baker (personal communication via the CAPRI committee). The objective here was to distinguish the designed complexes from native, and classify the designed complexes into potentially binding or non-binding (work performed in conjunction with Mr Iain Moal); the Baker group had a list of designed complexes that actually bound, although we were informed (CAPRI Committee) that most, if not all, designed complexes didn't form native complexes. Results from BioSimz simulations of complex association indicated a clear division in binding score distributions for the two sets (see Figure 4.6); the average score for native complexes is approximately 1.5 times larger than for the designed set. Approximately 75% of the false positives (designed complexes) during BioSimz simulation runs displayed a normalised binding score less than 36, while the fraction of the native complexes that scored below 36 was approximately 38%. When used as the sole input dimension of a support vector machine (SVM) classifier, these binding scores achieved a total accuracy of 71% (142 correct and 58 incorrect). The accuracy at this level was thought to be reasonable, because a) binding scores are linked to association rate constants, and many genuine complexes do have very low k_{on} values, and b) the false positive set is made up of designed interfaces that are supposed to reproduce as many stabilising properties as a real interface possesses, and it is rational to think that some of the designed complexes have achieved the objectives set and, therefore, scored reasonably well for rates of association during simulation.

4.3.3 Dissociation

Motivated by the need to further distinguish native bound complexes from designed complexes that are unlikely to form native complexes, attempts to

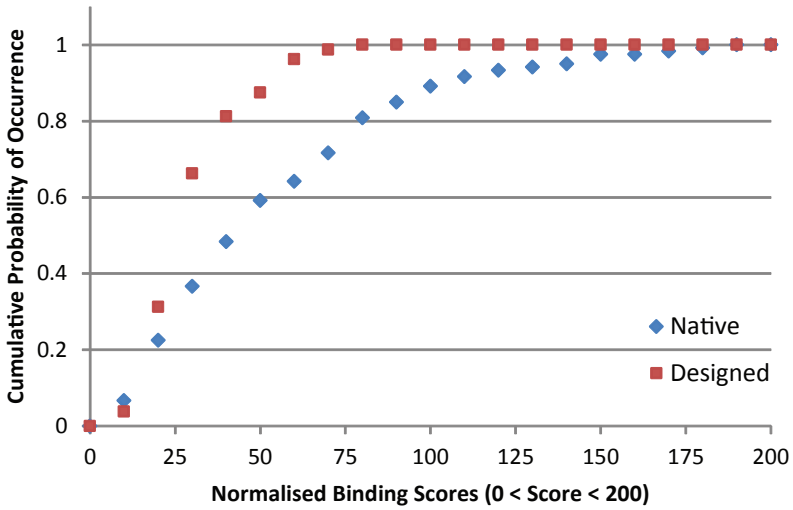

Figure 4.6: Difference in binding score distributions of native and designed complexes. Data points shown are the cumulative probabilities, i.e. $P(s) = \sum_{x=0}^{s} p(x)$. Data from the native complex set are plotted in *blue* diamond markers, whereas those from the designed complex set are plotted in *red* square markers.

fit the dissociation scores with k_{off} were also made. From the derivation of the association score(see Section 4.2.3), it can be easily seen that the scores ideally should correlate with k_{off} and on a negative basis, i.e. the more stable the complex is, the higher dissociation scores but k_{off} for the complex will be lower. Our results did show this negative correlation with k_{off}; however, the signal, $\rho = -0.24$, was too weak to be considered consistent.

This suggests that our scoring scheme, or indeed the dissociation simulation, must be missing some factors that are important (see Section 4.4.3 for discussion). One of these more important factors may be the internal flexibility of the binding partners. By not allowing internal movements within the molecules, impacts from anywhere on their molecular surface will pass through their rigid-bodies, disrupt the fine and tight binding between the interaction partners. Therefore, unlike the positive signal for the loose encounter complexes, the bound complexes were shown to be too prone to external impacts during the simulations of dissociation, thereby significant correlation to real k_{off} was lost. Nevertheless, it was noticed that the direction of correlation was still correct, which could potentially be used as a binary classifier at the very least.

Therefore, we performed further tests for the purpose of creating a native vs. false positive complex classifier based on dissociation scores. The 27 test complexes used for the docking analysis (see Chapter 3) were borrowed here to derive a distribution of dissociation scores for native complexes, and two data sets from CAPRI Target 43 and 44 were used as false-positive benchmarks since all of these complexes, except for one in T43, have their binding interface computationally designed. The results demonstrated, to some extent, the recognition capability of this dissociation classifier: it is possible to distinguish the distributions of scores between true and false positive sets, such that a cut-off score of 60 retains 82% of the native complexes above the threshold but only 45% of the false positives. The dissociation scores we developed here also helped the selection of the crystal complex from false positives for CAPRI Target 33. While the correct complex (ranked No. 2 otherwise) did not have a particularly high dissociation score (50.0), the candidate scoring higher in all other criteria (ranked No. 1) had a particularly low dissociation score (24.3). The dissociation score measurement thus helped us in removing this excessively unstable complex and

the winner was successfully identified.

When tested with the CAPRI T35 data set discussed in Section 4.3.2.3, the addition of dissociation scores to the SVM parameter space slightly increased the native/non-native classifying accuracy from 71% to 73%, meaning 146 correct and 54 incorrect predictions.

4.4 Discussion

4.4.1 Time Course

There have been numerous attempts to computationally predict k_{on} from interface energetics or macromolecular simulations, as discussed comprehensively in a recent review (Schreiber et al., 2009). To date, most researchers in the field have used either direct computation or simplified molecular simulation to achieve their goals. In direct computation, much effort has been directed into accurately calculating the binding free energy, ΔG, or its related values, such as binding energy of the molecular interface or free energy difference, $\Delta\Delta G$, upon mutation of key interface residues. On the other hand, simulation methods, mostly based on Brownian dynamics (Ermak and McCammon, 1978), derive k_{on} from the number of trajectories falling within a structural/directional threshold of the binding pose. Although the validity of both approaches has been verified theoretically or through simulations (Section 4.1.2), there is still a crucial link that is missing between the energetics of macromolecular interaction and rates of association: the time course of interactions.

Observing time courses for macromolecular interactions is a major component of this work. The author attempts to reproduce this "real" link between energy and kinetics for protein interactions in simulation, even if the relevant theory, characterised by $\Delta G^{\circ} = -RT(\ln K_{eq})$, was established centuries ago. By performing Langevin simulation of molecular diffusion/interaction and *not* terminating the trajectories in which the proteins are found to be at a binding pose, time courses of such interactions can be harvested and statistically analysed.

A valid concern over our seemingly simplistic binding scores (see Section 2.1.4.2) is where it stands from a theoretical perspective, and why the

scores are, as described in the Results, correlated with the logarithm of k_{on}: is this correlation simply fortuitous given the data set we parameterised on? Recall that Eqn. 4.29, proposed for calculating k_{on}, is based upon interaction potential and interface volume V_{RR} (Zhou, 1990). The variable V_{RR} and S can hereby be assumed to be constant, provided the ensemble of protein interface areas under study are broadly similar to each other. We will come to the issues of V_{RR} and S later on. With these simplifications, the Zhou equation can be re-written as

$$k_{on}^{*} = k_0 \exp\left(\frac{-\langle U\rangle^{*}}{k_B T}\right), \tag{4.40}$$

where k_0 is effectively a scaling constant. The interaction potential $\langle U\rangle^{*}$ is therefore

$$\langle U\rangle^{*} = -k_B T \ln k_{on}. \tag{4.41}$$

Taking the negative gradient on both sides, this transforms to

$$\langle F\rangle^{*} = \nabla \cdot (k_B T \ln k_{on}). \tag{4.42}$$

Here, the average force term, $\langle F\rangle^{*}$, is an indicative term suggesting how large on average a force needs to be in order to "pull away" the two binding partners in an established complex. Based on the Langevin equation, the total force exerted on a molecule is

$$F^{*} = F_{friction} + F_{U} + F_{random}, \tag{4.43}$$

where F_U is the force due to the interaction potential, and where the other two forces are labelled after their respective causes. In a binding pose, both binding partners are in a static state; therefore we have

$$F_{friction} = -\zeta v = 0, \tag{4.44}$$

and importantly,

$$F_{U,total} = -F_{random}. \tag{4.45}$$

The above equality must hold in order for the complex to be in a stable state.

Given the fact that the VdW component in $F_{\mathrm{U,total}}$ can provide as much repelling force as required, this equality is necessarily reduced to the following requirement,

$$\langle |F_{\mathrm{random}}| \rangle < |F_{\mathrm{U,\,attract}}| \tag{4.46}$$

where $F_{\mathrm{U,\,attract}}$ is the total attractive force that holds the complex together. Experimental evidence shows that deterministic forces are on the scale of 10^{-2} to 10 piconewton (pN), while stochastic forces can amount to 10^{-3} to 10^{-2}pN (Finer et al., 1994; Ishijima et al., 1996; Evans et al., 1995). It is already known that F_{random} conforms to a normal distribution of mean 0 and variance $2k_BT\zeta$, and the probability of F_{random} being larger than some certain $F_{\mathrm{U,\,attract}}$ is increasingly small when the negative potential $-U$ is large. Therefore, the average time needed for a random force large enough to appear and successfully counter the attractive forces is increasingly large. Hence, there is general stability, with occasional instability, when the strengths of two forces overlap. For molecular displacements, the same principle also holds: the larger $-U$, the more time on average it would take for a random displacement, or a combination of random displacements, to be large enough to jump out of the energy funnel. The source of random force and torques do not just come from solvent collisions; internal (heat) movements of the interacting molecules contribute even more to instability. In summary, we have the following relation

$$\langle t_{\mathrm{retention}} \rangle \sim \langle F_{\mathrm{random}} \rangle \propto -U \propto \ln k_{\mathrm{on}} \tag{4.47}$$

established for molecular interactions, where $t_{\mathrm{retention}}$ specifically means the retention time of the binding partners staying within the encounter zone. The functional form of the first monotonic relationships shown in Eqn. 4.47 was not analytically determined in this study due to the model lacking internal molecular movements. However, a section-based scoring function (Section 2.1.4.2) was put in place to reflect the link between $t_{\mathrm{retention}}$ and interaction potential U. As our results have shown in the previous section, this works rather well with a high correlation, $\rho = 0.84$, between calculated and experimentally measured k_{on}.

To test the validity of our theory, and its compliance with Eqn. 4.29, we

further tested the inclusion of interaction volume V_{RR} and survival fraction S to examine whether it improves our binding scores. The V_{RR} was estimated by exponentiating the interface area by the exponent factor, $3/2$, which effectively rescaled the values from area to volumetric measurement; the interface areas were estimated using the InterProSurf server (Negi et al., 2007). Survival fraction S was calculated from our simulations as the ratio of long-term specific encounters among all encounters that had once approached within the threshold of being a specific encounter complex.

We found that both V_{RR} and S terms further improved the correlation of computed binding scores with experimental k_{on}. The inclusion of S raised the correlation from 0.84 to 0.88; while the inclusion of V_{RR} further pushed the correlation up to 0.90. These results clearly demonstrated that the time course measurement of retention for specific encounter complex formation fits well to the existing theories and models of macromolecular association kinetics, and indeed extended them.

Nevertheless, with this relatively high correlation, $\rho = 0.90$, the accuracy of k_{on} calculated from simulations has ample room for improvement. More accurate k_{on} calculations have indeed been published, such as from the binding-site energetics based approaches (Song et al., 2004). Other attempts include fitting association rate constants, based on interface-areas (Schlosshauer and Baker, 2004) and RMSD thresholds for specific/nonspecific encounters (Gabdoulline and Wade, 1997). However, we stress that none of these studies has achieved accurate calculations for k_{on} using universal and consistent parameters for a cluster of randomly selected proteins, whose association rate constants span across several orders of magnitude. Moreover, the k_{on} derived from this study were obtained from dynamic simulations; this implies that different k_{on} values could be predicted from running the simulation with different parameters and environmental settings. It was this degree of freedom that inspired and enabled our further investigation of protein-protein interactions in a crowded macromolecular environment (Chapter 5).

4.4.2 Kinetics and Binding

From a docking point of view, some proteins will only bind with their partner after undergoing significant conformational change, while for others,

this may be achieved by passive internal movement within the limit of thermal energy. From a kinetic point of view, all specific protein interactions maintain a certain, often characteristic, rate of association under the same physical and chemical conditions. The association rate constants of protein interactions may therefore have links to their underlying binding mechanisms; as was pointed out in Figure 4.1, $k_{on} = 1 \times 10^5 M^{-1}s^{-1}$ is the boundary dividing interaction speed controls between "limited by conformational changes" and "limited by diffusion".

However, there could still be questions over this simplistic relationship. Recall the theories of induced-fit and conformational shifting for binding mechanisms, and the arguments on which one is the dominating party (see Section 3.1.1). The same hypotheses might well be reflected by the interaction kinetics and thus could be tested by simulation, with proteins in different conformational states. According to the induced-fit theory, under free diffusion, proteins remain in the unbound conformation (or one of their unbound conformations) until after making contact with their binding partner, preferably at a specific orientation favouring the formation of the final complex. This is equivalent to using unbound proteins in our simulations. On the other hand, the conformational shift theory states that proteins constantly switch into different states that can either be bound or unbound while undergoing free diffusion, and only interact with their binding partner when they are in or near to the correct (bound) conformation. This is equivalent to simulating protein interactions using all bound conformations, since the time spent in a certain state is proportional to the entire period of simulation, therefore, the overall kinetics will not change as long as our measurement, the binding scores, remains a quantity to be fitted rather than based on the actual k_{on}. Now in the context of our simulation environment, the problem of induced-fit vs. conformational shift boils down to comparing the binding scores of "unbound" and "bound" simulations, at least for the diffusion-governed protein interactions. It is worth restating that the binding scores, as proven in the last subsection, are measurements for the encounter complex formation and do *not* cover the period for the formation of the final complex conformation from an encounter complex ensemble. It is exactly at this encounter stage where conformational shifting and induced-fit are differentiated.

The simulated results were revealing: the "bound" simulations showed a correlation of 0.84 (0.90 with interface volume adjustment) with the experimentally measured k_{on}, while the "unbound" simulations returned a correlation of only 0.28 – still positive, but very weak. One may think that this would have been obvious, since with unbound conformation there would be less affinity – reduced true binding interface complementarity. The argument is apparently valid, but the individual difference in binding affinity, between unbound and bound, does not warrant the loss of correlation with experimental k_{on} across a wide diversity of protein interaction pairs. In other words, the scores from the "unbound" simulations were expected to be smaller than those of "bound" simulations, but as most of the proteins in the test set do not undergo major conformational change upon binding, it was expected that the calculated k_{on} values for the "unbound" simulations would be almost as similar to the experimentally determined k_{on} values as for the "bound" simulations (correlation coefficient, $\rho = 0.84$). However, this is not observed in our simulations, as $\rho = 0.28$ is hardly a noticeable correlation. Therefore, for proteins that have k_{on} limited by diffusion, it appears that the conformational shift mechanism dominates; it is more likely that when specific protein partners meet each other, they are already in a conformation at or close to the bound state. On reflection, it is perhaps unlikely that proteins can adjust their conformational state notably during the initial encounter stage, as proteins would still be too far apart for any strong forces to be generated. Further more, large adjustments upon binding (typical case of induced-fit) often slows down the rate of association, and are thereby more likely to be attached to interactions having particularly low k_{on} – a subject not studied here.

However, it should be noted that the "unbound" simulations did not lose the signal for specific binding; they just did not correlate with k_{on}. This was shown in Chapter 3 where all binding predictions were made using unbound-unbound conformations, and binding "hot spots" were still found. The unbound binding scores do have a moderate correlation with molecular weights M_w (0.46) and particularly, the cubic root of M_w (0.56), which implies the protein radii. The corresponding correlation coefficient for binding scores in "bound" simulations with molecular weights M_w is 0.16 – almost no correlation. This shows, even the specific collisions be-

tween unbound binding partners are influenced strongly by their geometric, thereby diffusional, properties (Section 4.3.1), rather than from individual interaction potential surfaces. In terms of the binding mechanism, perhaps the unbound conformations will first give a preferential coverage for the location of the binding site; the binding partners may hover over the specific binding area with a higher frequency compared to elsewhere (demonstrated in Chapter 3). Meanwhile, the bound state of the protein, shifted from the ensemble of its conformations during the hovering process, enables the correct binding energy funnel to be created, so that specific encounters can form with the right association kinetics, in line with the experimental k_{on} values. These conformational changes (shifts), although occurring near or on the binding site, need not to be "induced".

To summarise, for protein interactions that have k_{on} below the diffusion threshold ($1 \times 10^5 M^{-1}s^{-1}$), unbound protein conformations can offer anchoring points for the subsequent conformational changes to occur before binding (Section 3.3.3.2); for the diffusion-limited cases, as discussed above, the required conformations for specific binding may be directly supplied through conformation shifting, resulting in the equilibrium between associated and unassociated partners that corresponds well to the k_{on} value.

We recognise that the rigid-body nature of our simulations can limit further quantitative investigation into binding-mechanism preferences between induced-fit and conformational shift. Nevertheless, the simulation approach taken here offered a different aspect from which to study the dynamics of protein association with binding kinetics in mind, and the dominance of the conformational shift mechanism was indeed observed.

4.5 Conclusion

With the careful measurement of retention times for specific encounter complexes, the link between specific protein docking and interaction kinetics has been quantitatively verified. We found that the duration and, thereby, the stability of complex formation, correlates linearly with the logarithm of the association rate constant, k_{on}. This correlation was tested true for 11 protein-protein interactions whose k_{on} values span several orders of magnitude, and further verified on 120 native and 80 designed complexes, which

display significant difference in binding affinity between native and non-native complexes. A dissociation scoring scheme was also implemented and tested positive for distinguishing native/non-native interfaces, although the scores are not correlated with dissociation rate constants, k_{off}, mainly due to the fact that the internal motion of each protein is not currently modelled. Based on the comparative analysis of results from the simulations of complex association using bound and unbound binding partners, we examined alternative binding mechanisms. We postulated that for protein interactions whose k_{on} is limited by diffusion ($> 10^5 \text{M}^{-1}\text{s}^{-1}$), the dominating binding mechanism is conformational shift, rather than induced-fit, which should be more commonly observed for slow-associating proteins.

Chapter 5

Macromolecular Crowding

5.1 Introduction

In a living cell, nearly all functional behaviours have their roots in various type of macromolecular interactions, such as the formation of a well structured cytoskeleton framework, the cascaded cellular signals for growth control and the immune response to antigens. The intracellular environment is therefore crowded with structural and functional macromolecules at 300 $\sim$ 400g/L (Zimmerman and Trach, 1991), a concentration at which the condensed matter takes up 30% $\sim$ 40% of the cellular volume (Fulton, 1982). This far exceeds the normal soluble conditions of *in vitro* experiments and, from a chemical point of view, can be considered as too condensed to conduct efficient reactions. Worse still, the copy number of some molecule species may be so low that the collision-reaction theory becomes incapable of explaining why interactions between them would still be possible, often at a rate comparable to, if not better than, its *in vitro* equivalent (Ellis, 2001b,a).

The distinctive characteristics of volume-exclusion effects from aggregation of macromolecules was observed as early as in the 60s (Laurent and Ogston, 1963). Its influence in cellular interaction dynamics was first reviewed by Fulton (Fulton, 1982) and was later assigned the name "molecular crowding effects", a broad term that summarizes all unexplained phenomenon that may be related to the abundant molecular presence in the cell. Individual reports of how MCEs affect a particular macromolecular process are regularly published with observations of improved protein binding/-

folding dynamics, most of which have been recently reviewed (Zhou et al., 2008). However, many of these experimental approaches suffer from the drawback that the commonly used crowding agents, usually polyethylene glycol (PEG) or ficoll (Roque et al., 2007) of varying sizes. Disappointingly, are neither charged nor globular, and hence very different from the local environment in which proteins interact in a cell. Nevertheless these studies have inspired the theoretical work by Minton and coworkers, who made the first important breakthrough in understanding MCEs (Minton, 1981). In this and their subsequent work (Minton, 2000, 2001; Hall, 2003), the term "excluded volume effect" was widely used to describe the small chunks of inaccessible space voided by surrounding macromolecules; usually, they are too small for a macromolecule to diffuse into and are rendered useless. The result is that the effective diffusional volume of the solution is reduced, raising the reactant's effective concentration, i.e. chemical activity. Zhou *et al* (Zhou et al., 2008) further formulated a number of types of pairwise macromolecular interactions based on the excluded volume framework. However, due to the limitations of their model, which represents molecules as hard spheres, little could be speculated beyond the phenomena of volume exclusion and, consequently, the nonspecific interactions between the spheres. Later, a number of ambitious atomic-based Brownian dynamics (BD) simulations within a crowded environment were published by Elcock and coworkers (Elcock, 2002, 2003; McGuffee and Elcock, 2006). They found that the 2nd virial constant B_2, an indication of the level of nonspecific pairwise interactions, increased as the concentration gradient increases, in line with the experimental data. Due to lack of accurate forcefield modelling, the model was unable to capture the dynamics of specific interactions under crowded conditions. There have also been a few simulation studies targeted at molecular crowding using coarse-grained models (Sieber et al., 2007; Homouz et al., 2008), most of which used spheres with various radii to represent proteins of different sizes. By eliminating the need for calculating atomic pairwise potentials these models are capable of running with a much larger timestep than is possible using atomic models; however, lack of molecular details made it impossible to distinguish between specific and nonspecific interactions. Recently, Elcock et al. (McGuffee and Elcock, 2010) performed an atomic simulation on a proportion of the E. coli cytosol to re-

produce the *in vivo* translational diffusion coefficients for molecules such as the green fluorescent protein (GFP). However, once again, specific molecular binding processes were not directly simulated.

Our understanding of the macromolecular crowding effects (MCEs) is that, in a real cell, crowding has far more implications than just, simply, the volume exclusion effects. Many macromolecular interactions display a long, relatively unstable encounter complex state before a bound conformation is formed (Ubbink, 2009; Blundell and Fernandez-Recio, 2006; Tang et al., 2006). Intuitively, the spatial constraints imposed by crowded molecules could help the formation of an encounter complex and prolong its existence; whether such constraints do indeed help molecular binding, i.e. forming specifically oriented encounter complexes, remains to be investigated. Moreover, MCEs are likely to play an important role in sophisticated cellular processes, such as macromolecular assembly, transport and signalling, as these phenomena are almost impossible to be replicated *in vitro*. Effective study of all the above requires the specific interactions between target molecules to be probed, while taking into account the characteristics of the crowded molecular environment in which they are immersed. However, there is a lack of both theory and tools with which to understand MCEs in the context of their contributions to complex cellular functions.

Here we investigate, by the implementation of a novel computational model, the synergy between molecules in an overcrowded environment to illuminate the MCEs that have been previously unobserved, as well as their underlying molecular mechanisms. The model is implemented in our in-house simulation software package, BioSimz, from which a crowded molecular simulation is demonstrated in Figure 5.1.

5.2 Methods

A wide range of known protein-protein interactions were investigated in a molecular crowding context. This included enzyme-inhibitor complexes barnase-barstar (PDB:1B27), DNase E7-Im7 (PDB:7CEI); enzyme-substrate complexes adrenodoxin-adrenodoxin reductase (PDB:1E6E), ferredoxin-ferredoxin reductase (PDB:1EWY); signal-effector complexes Ras-GAP (PDB:1WQ1), Ras-PI3K (PDB:1HE8), TGFβ3-TβR-II (PDB:1KTZ),

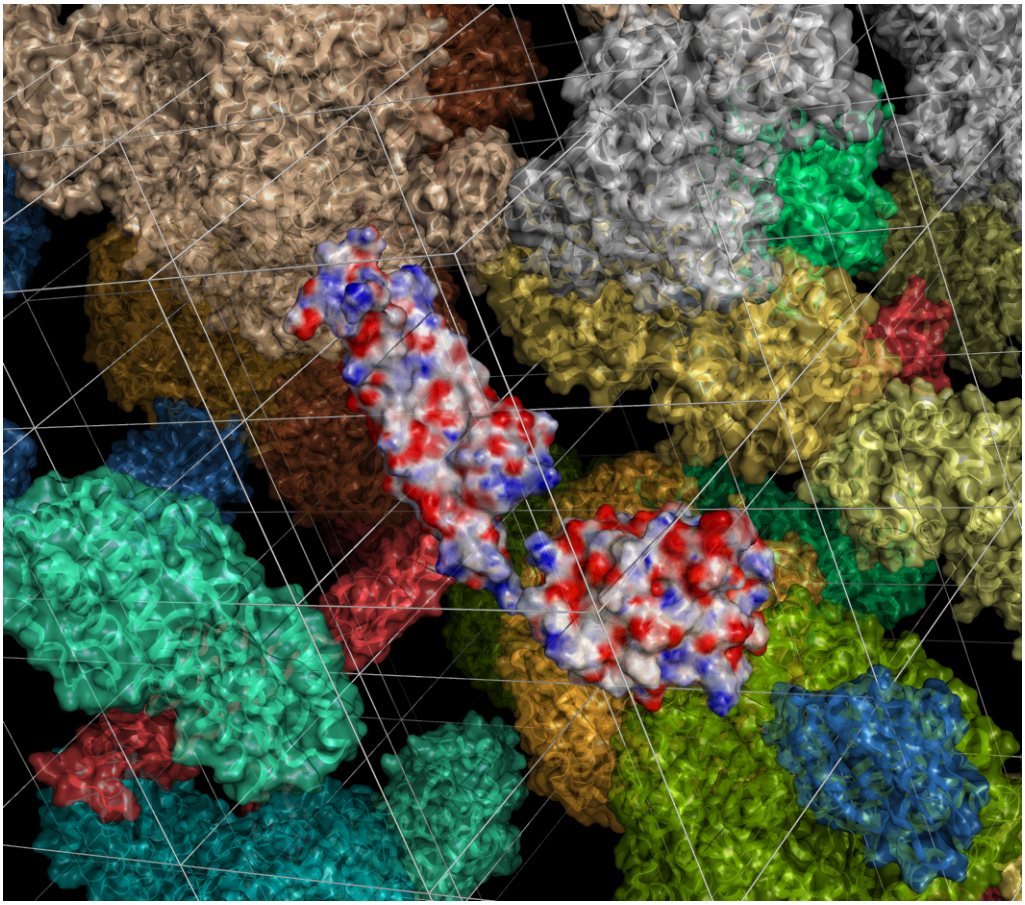

Figure 5.1: Snapshot of a simulation of crowded proteins, including TGF-β3, its receptor TβR-II and 10 different types of bacterial glycolytic enzymes as macromolecular crowders. Each molecule type is coloured differently, while the 2 target molecules approaching each other's binding site are shown in foreground with surfaces coloured by their surface electrostatic potentials (*red*: negatively charged; *blue*: positively charged). Total concentration of this system is approximately 240g/L.

Table 5.1: List of metabolic enzyme crowders

Index	Name	Weight (Da)	PDB Code
1	Hexokinase	52,209.9	1IG8
2	Phosphoglucose Isomerase	125,225.0	1HOX
3	Phosphofructokinase	35,263.2	4PFK
4	Fructose 1,6-bisphosphate Adolase	37,046.1	1ZEN
5	Triose Phosphate Isomerase	53,326.5	2YPI
6	Glyceraldehyde-3-phosphate Dehydrogenase	73,379.5	3GPD
7	Phosphoglycerate Kinase	45,315.5	3PGK
8	Phosphoglycerate Mutase	56,864.5	1EQJ
9	Endolase	93,739.4	2ONE
10	Pyruvate Kinase	198,428.0	1E0U

Cdc42-Cdc42GAP (PDB:1GRN), CDK2-CksHs1 (PDB:1BUH); immunoprotein complexes CD2-CD58 (PDB:1QA9) and Adenovirus Knob 35-CAR (PDB:1KAC). Various binding mechanisms may be present during the formation of the above complexes in a cellular environment (Ben-Shimon and Eisenstein, 2005). It is, therefore, of much interest and importance that the consistencies and differences of their behaviour, under varying molecular crowding conditions, are thoroughly examined and compared.

Ten types of bacterial enzymes involved in the glycolytic pathway were used as crowder molecules (see Table 5.1, hereafter termed as "environmental crowders"). These enzymes usually have an abundant presence in cells, as they are vital in metabolic pathways (Zimmerman and Trach, 1991); it is also natural to assume that in a local environment, the enzymes may aggregate and hover close by each other because they work in a cascade. These enzymes catalyse the glycolysis of a glucose/fructose molecule into two pyruvate molecules; therefore, it is safe to assume that they do not form specific complexes with the target proteins selected above.

All simulations were run in a cubic box sized at $240\times240\times240$Å, with periodic boundary conditions applied on all sides. This size of the box is large enough to eliminate boundary effects for long-range interactions that are normally below 40Å, and are no more than 100Å in the extreme case of protein-RNA binding (Tworowski and Safro, 2003). Typically, 2 to 29 copies of each target protein were distributed in the box at a receptor/ligand ratio of 1:1, plus 10-20 environmental crowders. Therefore, the simulated molecular systems had a target concentration between 0.12mM

and 3.36mM, which is 18.5g/L to 102.9g/L depending on molecular types; depending on the simulation runs, the environmental crowders make up from 92.6g/L to 185.2g/L of solutes in the box. Total protein concentrations thereby varied between 18.5g/L and 288.1g/L, the latter of which approaches the approximated cytosolic protein concentration of 300g/L (Zimmerman and Trach, 1991). The total protein occupancies of the molecules with respect to the simulation box varied between 6.89% and 38.71%. In all simulations the copy ratio among the environmental crowder types were kept at or around 1:1 – this may not be the physiological ratio in regard to a whole cell, but in a small catalytic compartment where the presence of all enzymes should be guaranteed, i.e. at least 1 copy, a 1:1 ratio is a reasonable assumption. The actual number of crowder copies placed in a simulation box was 1 or 2 per protein type, thus making the total number of crowders in any simulation to be between 10 and 20. The total occupancies of the crowders varied from 6.57% to 13.14%.

Both structural and kinetic investigations were carried out on molecular trajectories obtained from simulations, using root-mean-squared distances (RMSD) and a deviation threshold to identify whether each interaction event occurred at the correct, specific, binding interface. The deviation threshold $(\delta \boldsymbol{x}_t, \delta \boldsymbol{R}_t)$ was defined by the relative position and orientation of a target ligand molecule with respect to its receptor at time t (see Section 2.1.4.1 for details). A retention time threshold (40ps) was set up to distinguish the events that stayed within all structural specificity measures, with such events labelled, "specific interaction events". This by no means implies the two binding partners are guaranteed to go on to form a bound complex; it is merely a threshold indicating a specific encounter complex is formed, given the average retention time of encounters being roughly 10 times this length (Bui and McCammon, 2006). The stability of this specific encounter complex was evaluated in the form of a binding score, described in detail in Section 2.1.4.2, and as justified by the previous theoretical discussion, see Section 4.4.1.

5.3 Results

5.3.1 Association Rate Constants

In Chapter 4, a significant correlation between the predicted binding scores and experimental k_{on} values was reported for the eleven test set protein-protein interactions, studied under dilute conditions. To ascertain if a similar correlation holds for target proteins in a crowded environment, the same target data set was simulated with molecular crowders. As pointed out by a recent review on molecular crowding (Elcock, 2010), there has been a lack of consistent and rigorously built test sets for computational modellers to benchmark protein-protein interaction models against experimentally validated data involving crowded environments, be they simple polymers such as Ficoll, PEG, Dextran or more sophisticated colloid molecules, such as proteins. This is more problematic for the likes of this study, as protein crowders are rarely used, let alone validated and benchmarked, in experimental crowding studies as they tend to hinder the experimental read-out. Indeed, among all eleven test set complexes, quantitatively assessed crowding kinetics can only be found for barnase and barstar (Phillip et al., 2009); even in this case, barnase and barstar were crowded with PEG molecules rather than protein crowders.

Nevertheless, our predicted, crowded k_{on} values, in the form of binding scores, match well with the experimentally verified k_{on} changes for the barnase/barstar interaction with and without environmental crowders (correlation coefficient=0.97). Moreover, it is worth noting that we successfully predicted the immediate damping effect on barnase-barstar k_{on} with median-level of crowding, as well as the partial recovery of the association rate constant when the crowding level increases (see Figure 5.2).

Having achieved the above correlation for crowded barnase/barstar and the previous correlation on eleven non-crowded protein-protein interactions (Section 4.3.2), we were reasonably confident in predicting the k_{on} changes for the remainder of the complexes in the test set. We found that seven out of the eleven complexes (PDB:1WQ1; 1KTZ; 1GRN; 1BUH; 1EWY; 1KAC; 7CEI) showed an increased k_{on} with environmental crowding, five of which increased by more than two-fold; the remaining four complexes (PDB:1B27; 1HE8; 1E6E; 1QA9) have their association kinetics slowed down

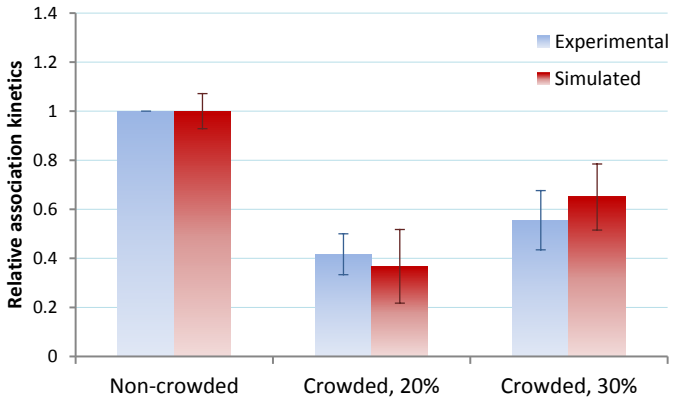

Figure 5.2: Comparison of experimentally measured k_{on} and simulated binding scores in various crowding conditions for barnase-barstar. Two crowding setups at approximately 10% and 30% mass concentration ratios are shown. Mass concentration is expressed as weight-percent of the viscogenic agent from the total weight of the solution. The non-crowded association rate constants from experiments and non-crowded binding scores from simulations are scaled to 1, as only relative k_{on} are available from experimental data (Phillip et al., 2009).

by crowding (see Figure 5.3 and Table 5.2). An immediately observable pattern from the test set results shows that environmental crowding has greater positive effects on protein interactions with a relatively low k_{on}, whereas for highly affinity protein binding cases, crowding is likely to act as a damping factor for their association dynamics. This is in line with the intuition that *in vivo* molecular crowding acts as a selective force, which helps the slower members to bind faster while penalising fast-binding proteins (see the Section 5.4 for more discussion). It is worth noting, however, that k_{on} is not the *only* measure of quality assessment for molecular interactions, especially *in vivo*. A reduced association for particular types of proteins is often compensated by a decreased dissociation rate, which keeps the reaction dynamics (K_d) little affected in a crowded environment (Phillip et al., 2009).

The characteristic dynamics of specific interaction events, under the influence of environmental crowders, stimulated the subsequent investigation of the underlying principles that govern them. A particular interest is to understand why low association-rate interactions are likely to be boosted by crowding.

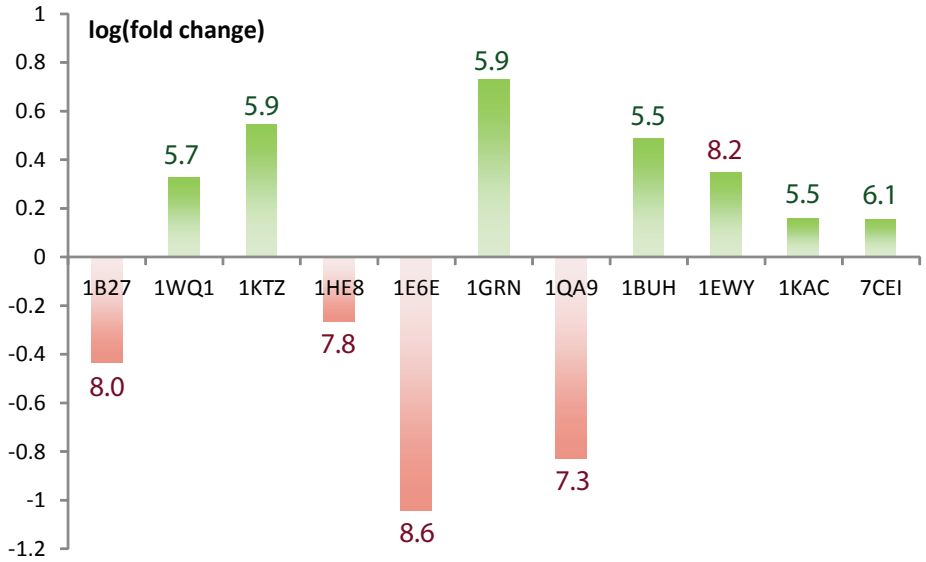

Figure 5.3: Magnitudes of changes in k_{on} values compared to those from non-crowded simulations. Bar plot shows the order of magnitude changes, in log space, for each k_{on} of the test complexes. Green bars indicate a positive fold change (increase), and red bars a negative change (decrease). The numbers labelled on each bar are the original, experimental non-crowded k_{on} values for these protein-protein interactions.

Table 5.2: Changes of k_{on} and percentage increases of molecular steering due to crowding effects

Index	Complex	PDB Code	Exp. Non-Cr. k_{on}	Pred. Cr. k_{on}	k_{on} Fold change	Cr. steering increase
1	Barnase-Barstar	1B27	1×10^8 (Wang et al., 2004)	3.7×10^7	0.37	4.08%
2	Ras-RasGAP	1WQ1	6×10^5 (Antonny et al., 1991)	1.3×10^6	**2.13**	**16.7%**
3	TGFβ3-TβRII	1KTZ	7.4×10^5 (Baardsnes et al., 2009)	2.7×10^6	**3.62**	**32.0%**
4	Ras-PI3K	1HE8	6×10^7 (Pacold et al., 2000)	3.3×10^7	0.54	-17.7%
5	Adrenodoxin-Reductase	1E6E	4×10^8 (Lambeth et al., 1980)	3.6×10^7	0.09	3.15%
6	Cdc42-Cdc42GAP	1GRN	8×10^5 (Leonard et al., 1998)	4.3×10^6	**5.37**	**51.2%**
7	CD2-CD58	1QA9	2.1×10^7 (Davis et al., 1998)	3.1×10^6	0.15	9.21%
8	CDK2-CksHs1	1BUH	3.2×10^5 (Bourne, 1996)	9.8×10^5	**3.06**	**77.2%**
9	Ferredoxin-FNR	1EWY	1.5×10^8 (Bhattachryya et al., 1986)	3.4×10^8	**2.23**	5.68%
10	Adenovirus Knob-CAR	1KAC	3.1×10^5 (Lortat-Jacob et al., 2001)	4.5×10^5	**1.44**	**27.7%**
11	DNase E7-Im7	7CEI	1.3×10^6 (Hosse et al., 2009)	1.9×10^6	**1.43**	**29.0%**

5.3.2 Interaction Dynamics

The above demonstrates that molecular steering may have an effect on association kinetics at typical *in vivo* concentrations. How, in a crowded environment, molecular steering may enhance or reduce the association activity of the target proteins, is an important question to address in a biological context; do there exist generic nonspecific or specific steering effects for all protein-protein interaction events within the cell?

Average linear and angular velocities of the target molecules were calculated from the (environmentally) non-crowded and crowded simulations, respectively. The ratios of change in angular velocities (rotational speeds) between crowded and non-crowded conditions, as well as those between specific and nonspecific interactions are plotted in Figure 5.4.

At all times, angular velocities for specific interactions are higher than those of nonspecific interactions (Figure 5.4(b); this implies that proteins undergo more rotational movement when they approach their binding sites, further adding to the proof of the "spinning-and-rolling" binding mechanism for specific binding that has been demonstrated in other sections of this study (Section 3.4 and 4.3.2.2). With the crowders are added ($\sim$20% occupancy), a very interesting disproportionation of ratios is observed: the angular velocities of proteins, during specific interactions, have increased significantly (24.5% averaged over test set complexes, Figure 5.4(a)), whereas their angular velocities during nonspecific interactions have undergone a mild decrease (-6.93%). As a result, with molecular crowding, the averaged angular velocities during specific interactions have increased by 50.3% (Figure 5.4(b)), compared to velocities during the nonspecific interactions. The same velocity comparison for a non-crowded environment results in only an increase of 11.3%. The linear (translational) velocities are also increased for specific encounters; however, no change was found for linear velocities for nonspecific encounters between crowded and non-crowded conditions.

This distinctive differences between crowded and non-crowded conditions, particularly when crossing over the differences between specific to nonspecific interactions, have revealed that macromolecular crowding seems to provide extra momenta, particularly angular momentum, for proteins that are near or at their binding sites. For at least a majority (nine out of eleven) of the test complexes, molecular crowding appears to have "gone

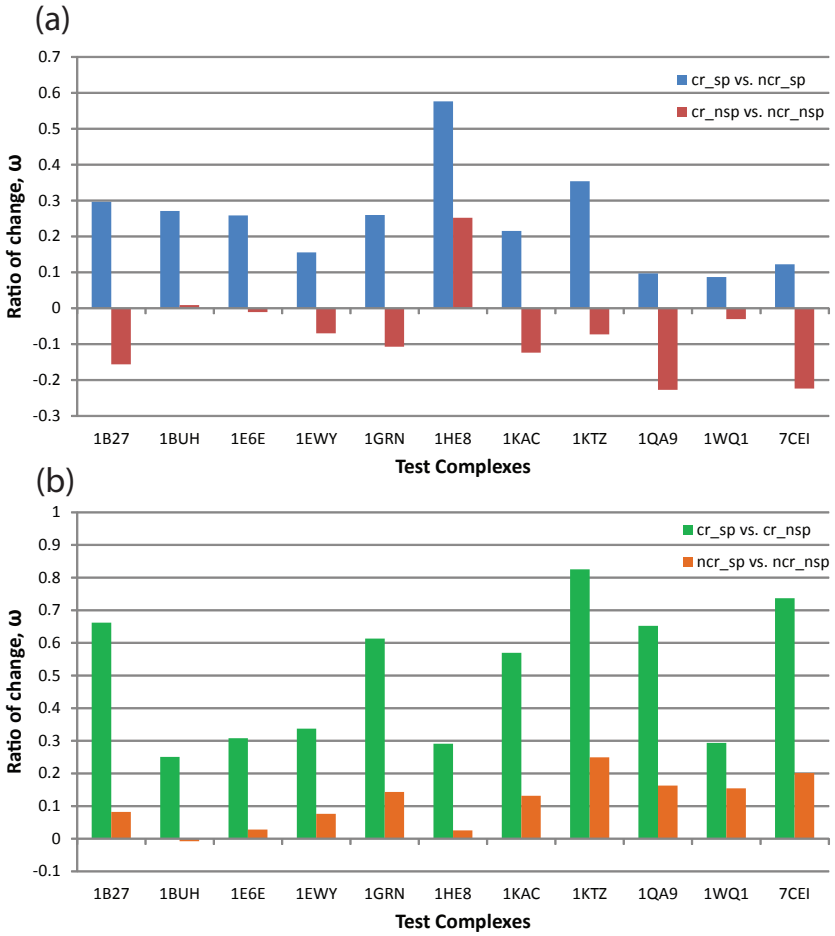

Figure 5.4: Ratios of change in angular velocities for protein-protein interactions under different crowding conditions. **(a)** This plot shows the disproportionate magnitude between the angular velocities of protein pairs undergoing specific and nonspecific interactions when the reaction environment becomes crowded. The *blue* bars are the ratios of change between *crowded specific* and *non-crowded specific* interactions. The red bars are the same ratios between *crowded nonspecific* and *non-crowded nonspecific* interactions. **(b)** This plot shows the elevated specific/nonspecific ratio under crowded conditions. The *green* bars are the ratios of change between *crowded specific* and *crowded nonspecific* interactions, and the *orange* bars are the same ratios between *non-crowded specific* and *non-crowded nonspecific* interactions.

for the extra mile" to further reduce momenta of nonspecific encounters. Hence, the simulation results presented seem to suggest that, with the confinement of crowders, the dynamics of specific encounters is accelerated, while encounters that are not at an energetically favourable conformation tend to be stuck there (with lowered velocities); they will eventually dissociate, even under relatively mild random forces induced by their environment (Section 4.3.1).

5.3.3 Crowded steering effects

Immediately one wonders about the source of the extra momenta discussed above, preferentially assigned to specific interaction events thereby facilitating molecular steering. We suspected the driving force behind this is the environmental potential between target molecules and crowders, where long-range, ambient electrostatics may have a significant contribution. Comparing the interaction dynamics data from three crowding conditions tests this hypothesis: non-crowded, crowded with crowder electrostatics turned off, and crowded with normal, charged environmental crowders. Atoms on target molecules remained normally charged in all three cases. Results, in general, showed that k_{on}-enhancing protein-protein interactions support the above hypothesis: for example, there is a 5.4-fold increase in k_{on} if normal crowders are used for complex Cdc42-Cdc42GAP (PDB:1GRN); with uncharged crowders, there is a moderate 2.1-fold increase from the corresponding dilute k_{on} value. For electrostatics-dominated interactions, which have displayed reduced k_{on} values in crowded conditions, eliminating crowder charges leads to partial recovery of the rate. Hence, environmental electrostatic interactions may act as an amplifier to the existing upward trends of interaction dynamics of the target protein pairs. In fact, for all test set complexes, the existence of charged crowders leads to an increased proportion of interaction events converting from nonspecific to specific. This enhancement is notably pronounced for encounters that are neither too short (> 0.5ns) nor too long (< 1.5ns), which is in line with average life time ($\sim$ 1ns) for encounter complexes (Bui and McCammon, 2006) (see Figure 5.5). After removing all charges from crowders, the level of specific to nonspecific enhancement is approximately halved. The extra number of mid-retention interaction events, boosted by environmental elec-

trostatics (as well as other crowding factors), may come from two sources. One is from the pool of nonspecific interactions converting to specific interactions and this accounts for the additional increase of association rate for some protein pairs. The other source is from longer-retention events, affecting the stability of established complexes. Because environmental electrostatics accounts for a large portion of the overall influences from crowders, the simulations have shown that crowding may have more negative impact on the more electrostatics-driven interactions, such as the case for barnase-barstar. This also explains why our simulation shows a deeper dip for the crowded k_{on} for barnase-barstar compared to the dip found for experimental data (Phillip et al., 2009): the interplay of electrostatics between barnase/barstar molecules and the charged protein crowders, used in this study, may have interfered with the highly-electrostatic specific binding between barnase and barstar molecules. In the corresponding experimental study, the crowding agents, non-charged PEG molecules, would not have had this electrostatic interference to barnase/barstar interactions.

We then ask whether certain interactions are preferentially biased by molecular steering aided by environmental crowders. To systematically investigate this, a measure of successful steering ratio is introduced, which is the ratio of the specific interaction events that have longer retention time than the threshold (40ps), indicating a successful steering, among all interaction events that spend any time within the specific binding region. Although this ratio varies among different complexes, all but one (adrenodoxin-reductase, PDB:1E6E) receive an increase in their respective ratio of successful steering in the presence of environmental crowders (see Table 5.2). Some complexes, particularly those of a low k_{on}, have a notably enhanced increase in this ratio; even those with high k_{on} interactions mostly show a gentle upward trend with the help of crowder molecules. The correlation coefficient between k_{on} and the *increase* in successful steering rate is -0.67, a significantly negative correlation; the lower the k_{on} value, the more likely the interaction will get a boost in successful steering rate from macromolecular crowding. For example, the Cdc42-Cdc42GAP complex (PDB:1GRN) has a k_{on} of $8 \times 10^5 M^{-1}s^{-1}$. This complex reports a 51.2% increase in successful steering ratio in a crowded environment as opposed to non-crowded. Correspondingly, complex 1GRN also has a 5.4-fold increase

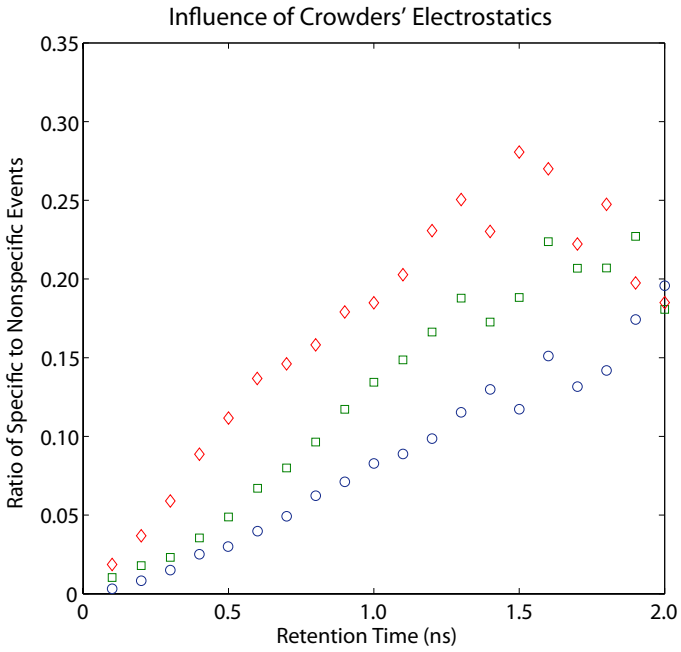

Figure 5.5: Influence of crowder electrostatics on the ratio of specific and nonspecific encounters over different retention times. The blue data points (○) show the ratios changes without environmental crowders. The red data points (◊) show the same ratios with the presence of normal electrostatically active crowders. The green data points (□) show the ratios with the presence of the same environmental crowders as for the red profile, however with their charges disabled.

in predicted k_{on}, the largest among the test set complexes.

5.4 Discussion

Using BioSimz, we have constructed a computational model capable of conducting rigid-body protein-protein interaction simulations in a crowded macromolecular environment. It is for the first time that the dynamics and kinetics of specific protein-protein interactions can be directly investigated in a multi-macromolecular environment where all simulated bodies and forcefield parameterisation are at the atomic scale. Calibrated on the limited available, experimentally determined k_{on} values in crowded (Section 5.3.1) and non-crowded (Section 4.3.2) conditions, we revealed the changes of binding dynamics with the existence of environmental crowders, and predicted the possible mechanisms that generated the changes, of which ambient electrostatic interactions were shown to play an important part.

Conventionally, the molecular crowding effects have been considered to be almost synonymous to volume exclusion effects; extensive theoretical, simulational and experimental studies have focused on how volumetric difference plays a part in limiting the free diffusion of the target macromolecules (Minton, 1981; Zhou et al., 2008). Until very recently, most existing studies have tried to make a generic, binary assertion that crowding either helps or hinders the macromolecular binding process. By the work reported in this thesis, it is shown that crowding, either by the reactants themselves or by environmental protein crowders, has a diverse spectrum of effects on protein-protein association dynamics under a central scheme: crowding always tends to increase the chances of specific encounter complex formation through facilitating predominately angular movements of target proteins. Whether the molecular steering effect converts more nonspecific encounter to specific ones, or destabilises (thus shortens the lifetimes of) existing specific encounters, remains dependent on the nature of the molecular interaction itself. One major contributor of this extra momentum is through ambient, nonspecific electrostatic interactions with environmental crowders.

An important factor of molecular crowding is the volume exclusion effect placed onto the system of interacting molecules. As pointed by Minton

(2001), diffusion tends to be less limited for smaller molecules; it would be more difficult for the large target molecules to diffuse (thus interact) through the crowded solution. By examining the data from the simulations, this statement holds for nonspecific interactions, as the correlation coefficient between molecular weights and the numbers of nonspecific interactions is -0.62. This is a stark contrast to the same correlation coefficient calculated for specific interactions, which yielded $\rho = -0.11$. This, in addition to the results shown in Section 5.3.1, further verifies that the formation of specific encounters, unlike that of nonspecific encounters, is not affected by diffusion conditions, be it non-crowded or heavily crowded.

Molecular simulations performed here point to enhanced molecular steering, leading to changes in k_{on} for target molecules. Indeed, the correlation is strong and positive (0.72) between *changes* in molecular steering, i.e. rotational movements, and *changes* to k_{on} when the interaction environment becomes crowded. For example, Ras-GAP (PDB:1WQ1) shows a 2.1-fold k_{on} increase under crowded conditions, and this projects to a 16.7% increase in the successful steering ratio. However, counter-examples do exist, as for ferredoxin and its reductase (PDB:1EWY) – here a marginal 5.68% increase in the successful steering ratio is recorded, but a 2.2-fold increase in k_{on} is observed in crowded simulations. This indeed may be a case where the volume exclusion effect kicks in, enforcing longer retention of the already formed encounter complexes.

Finally, it is interesting to note, that many of the low k_{on} protein interactions are involved in signalling pathways (five in the test set); therefore, the rate enhancement for their association under crowded conditions (four out of five cases) may have particularly meaningful implications. It is extremely difficult to replicate the biological behaviour of a signalling transduction pathway through *in vitro* experiments; hence, the intracellular content that crowds these proteins, albeit different from the example crowders used in this study, may play an important role in governing the correct binding cascade being formed. Through atomic detailed simulation algorithms, such as the model described here, investigation into specific protein-protein binding and/or competition in a specific crowding context, i.e. crowders that may exert specific influence towards certain targets, has perhaps for the first time become within reach.

5.5 Conclusion

In conclusion, we have demonstrated that MCEs influence the outcome of macromolecular interactions in multiple ways across frequential, spatial and temporal measures. We showed that when molecules approach each other, they are more likely to spin and roll towards their specific binding sites under the influence of neighbouring crowders. We also revealed that, in addition to the commonly perceived volume exclusion effect, the electrostatic steering effect by crowder molecules contributes to the rate enhancement and persistence of specific binding between target molecules. At a stage where no previous experimental or computational efforts have been able to investigate the physical nature of *in vivo* macromolecular crowding, our approach offers the first glimpse into the molecular mechanisms of MCEs in atomic detail, responding to the "quantitative challenges" (Hall, 2003) on the subject of "specific molecular interactions" (Elcock, 2010) for the macromolecular crowding problem.

Chapter 6

Concluding Remarks

6.1 General Conclusion

Cellular structures and functions rely on a complex network of regulated protein interactions, which are further based on complex molecular binding mechanisms. In this study, further resolution to three key, progressive, questions (Figure 6.1) was attempted: a) how does a protein find its partner, b) at what rate do they interact and c) how are interactions affected by their crowded, *in vivo*, environments. While these questions are usually studied in isolation, we believe that unique observations have been made by carrying out holistic investigation of the three problems under one roof, bringing together molecular docking, interaction kinetics and macromolecular crowding.

The construction of the macromolecular simulation package, BioSimz, enables us to create a simulation environment that accommodates multiple macromolecules conducting pairwise and multiple interactions (Chapter 2). Performing simulations under the general Langevin dynamics scheme, enabled the mapping of potential, low energy, protein-protein interaction sites to trajectory density profiles; the resulting heatmaps were demonstrated to have guided our in-house docking protocol, SwarmDock, to better focus on the potential binding sites (Chapter 3). Under dilute conditions, the distributions of duration of the specific encounter complexes, formed during simulations, were found to strongly correlate with their association rate constants, k_{on} (Chapter 4). After adding different types of proteins, such as the ten glycolytic enzymes, as environmental crowders into the simula-

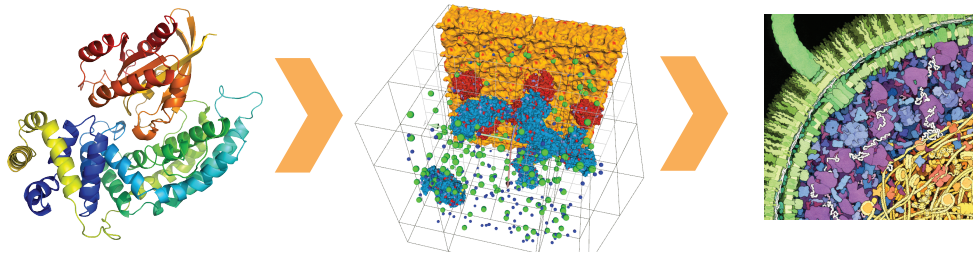

Figure 6.1: Three steps to tackle the molecular interaction problem. Step one, the pairwise binding problem. Step two, association and dissociation kinetics of a system of macromolecules . Step three, partial emulation of a dynamic cellular environment. The simulation box with proteins, ions and a membrane on one of its faces is modelled and rendered with BioSimz. The artistic illustration of an *E. coli* cell (right most) was copyrighted by David Goodsell.

tion box, target protein-protein interactions with highly electrostatic binding surfaces, and generally with a high k_{on}, tend to have their interaction kinetics damped, while the slower interactions seem to have their k_{on} elevated (Chapter 5).

Throughout this whole study, the quest for a more thorough understanding of the underlying mechanisms for macromolecular interactions has been maintained. On the intermolecular side, the binding mechanism of molecules spinning and rolling, that describes how molecules search each other's surfaces to locate stable contacting areas, has been demonstrated, with direct evidence from neighbouring contacts (Section 3.4) and angular velocities (Section 5.3.2). With respect to intramolecular movements, the debate on induced-fit versus conformational shift, as the main source of bound-form conformations, has been investigated, with the conclusion that, due to the better agreement between binding scores from simulations of bound targets and the experimental k_{on}, conformational shift may be the predominant factor for fast, diffusion-controlled molecular interactions (Section 4.4.2). Finally, the spinning and rolling mechanism has been shown to apply to not only protein binding *in vitro*, but also in a crowded macromolecular environment, in which the steering force/torque can be further enhanced by ambient electrostatics of the surrounding crowder molecules (Section 5.3.3).

This thesis has now come to conclusion with the above theories and observations, which have formed a self-contained set where each of its element

explains each other.

6.2 Future Directions

A brief overview of how the current work may be continued is summarised below.

6.2.1 Models

The BioSimz library is currently equipped with only one motion scheme, which is rigid-body Langevin dynamics. Improvements can be made through adding internal flexibility to the system. This requires work on the following two fronts:

Backbone flexibility. A suitable movement model is required to represent and store the motion of protein backbones. It seems apparent that, given the picosecond time frame, a full-atom MD movement model is inadequate to model the large, deterministic movement incurred during molecular collisions. The fixed-ends movement (FEM) method, allowing extensive backbone flexibility, providing considerable promises in being able to deliver a high-performance backbone movement whilst in agreement with polymer physics. The FEM algorithm has already been implemented and tested in the BioSimz library; the next step is to design and benchmark a set of force propagation rules so that the movement engine can be attached to the full-atom forcefield without incurring internal atom conflicts.

Sidechain flexibility. Conformational changes of sidechains are too "delicate" at the picosecond scale, such that a Newtonian motion scheme is deemed not applicable. Therefore, Monte-Carlo sampling from sidechain conformers should be used to generate acceptable sidechain conformations.

The introduction of molecular internal flexibility, no matter how well it would be implemented, will inevitably hinder the current computational performance of BioSimz simulations. This negative influence can be minimised by implementing the new algorithms, especially the Monte-Carlo

conformational sampling, on a massively parallel GPU computing platform, such as CUDA (Section 1.2.1.1).

6.2.2 Applications

Much of the effort in this thesis project involved the development of the simulation package itself. Upon completion and further improvements, many more questions of scientific interest to the molecular and systems biology communities can be explored using, or with the help of, the BioSimz package. Here I list two problems, which are deemed to be mostly interesting from the author's point of view:

The macromolecular aggregation problem. It is commonly known that certain proteins, such as DNA polymerases, can aggregate to a confined space, e.g. around DNA, when carrying out functions that require them to be present in large numbers. Lysozymes are reported to have a similar tendency and a very transient homogeneous cluster can be formed when protein concentration is high (Porcar et al., 2010). This problem is indeed related to the macromolecular crowding problem, only differentiated by the specificity of the interactions: in this case, proteins in a transient cluster seem to have a very weak, but definitely specific binding with each other. These bindings may have extremely high k_{off} values such that they dissociate quickly after being associated; nevertheless, aggregations may act like a "glue", keeping the homogeneous proteins within a small vicinity of each other, as a consequence of rapid cluster member association/dissociation dynamics. This problem could be an ideal target for BioSimz simulations to investigate: by fixing a number of lysozymes at certain places, the diffusion distribution of other lysozymes can be examined throughout a number of simulations. If they are found to have a higher occurrence near the restricted lysozymes, while other protein types don't, then the specific aggregation/clustering effects would indeed be demonstrated through simulation.

The ligand competition problem. Competition for the same receptor binding site, by different ligand types, or even simply different mutants of the same ligand type, is of great interest in both academic and drug

discovery fields. Conventional MD is far too slow for screening and evaluating the affinity of different ligand molecules towards a medically interested binding site; on the other hand, most docking packages only perform relatively static geometric and energetic fitting trials between ligands and the receptor. This package not only dynamically explores the curvature of binding energy funnels, but also simulates multiple ligand/receptor types at the same time in one box; this naturally leads to the consideration of target ligand filtering, perhaps on an ensemble of ligand mutants, to obtain the receptor-ligand pair displaying the required kinetics.

Indeed, with the opening up of new dimensions in molecular quantities and interaction time-courses for macromolecular simulations performed at atomic resolution, the problem and solution space in which researchers explore and gain understanding in the field of molecular interaction and recognition, has never become so wide.

Appendix A

The Wilcoxon Rank-Sum Test

The Wilcoxon rank sum test is a non-parametric method of testing statistical significance that two populations have the same distribution. Like the Student's t-test, the Wilcoxon rank sum is usually used for testing whether two populations are different from each other. However, a rank sum test differs from the t-test in that the former is solely based on the order in which the observations fall, and that it does not assume the prior distribution of the two samples to be tested.

Generally, suppose there are n_A and n_B sampled observations from populations A and B respectively. We wish to test the hypothesis that the distribution of samples in A is the same as that in B, which is written as

$$H_0 : A = B. \tag{A.1}$$

A departure from H_0, to be revealed in H_1 by the Wilcoxon rank sum test, is termed a location shift. Possible scenarios for location shifts can take the form

$$H_1 \ : \ A > B, \tag{A.2}$$
$$H_1 \ : \ A < B, \text{or, more relaxed,} \tag{A.3}$$
$$H_1 \ : \ A \neq B. \tag{A.4}$$

Where the first two of the above conditions are mutually exclusive, the test of either condition is termed a one-side rank sum test. If the direction (up or down) of the shift in ranks is not of interest, or the two conditions are not

mutually exclusive, the rank sum test is then conducted in a two-sided way, i.e., satisfying the third condition.

The actual ranks are generated by mixing and sorting samples from population A and B, using a certain scoring function. The Wilcoxon rank sum test statistics is the sum of the ranks for observations from either one of the two sample populations. If population A and B are from the same distribution, we expect their sums of ranks to be of the form

$$E(T) = \frac{n_A(n_A + n_B + 1)}{2}. \tag{A.5}$$

The distribution of sum of ranks has its variance given by

$$\sigma_T^2 = \frac{n_A n_B}{12}(n_A + n_B + 1). \tag{A.6}$$

Provided n_A and n_B represent large enough samples (> 10), T can be approximated as normally distributed. P-values for one-sided tests can then be determined using the standard normal P-value table. For two sided tests, both tails need to be considered, resulting a doubled P-value.

Bibliography

Abagyan, R., Batalov, S., Cardozo, T., Totrov, M., Webber, J., and Zhou, Y. (1997). Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. *Proteins*, Suppl 1:29–37.

Abrahams, D. and Gurtovoy, A. (2004). *C++ Template Metaprogramming: Concepts, Tools, and Techniques from Boost and Beyond*. Addison-Wesley Professional.

Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., and Iwata, S. (2003). Structure and Mechanism of the Lactose Permease of Escherichia coli. *Science*, 301(5633):610–615.

Adalsteinsson, D., McMillen, D., and Elston, T. (2004). Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks. *BMC Bioinformatics*, 5(1):24+.

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A., and Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature*, 450(7170):695–701.

Andrews, S. S. and Bray, D. (2004). Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical Biology*, 1(3):137–151.

Antonny, B., Chardin, P., Roux, M., and Chabre, M. (1991). GTP hydrolysis mechanisms in ras p21 and in the ras-GAP complex studied by fluorescence measurements on tryptophan mutants. *Biochemistry*, 30(34):8287–8295.

Baardsnes, J., Hinck, C. S., Hinck, A. P., and O'Connor-McCourt, M. D. (2009). TβR-II discriminates the high- and low-affinity TGF-β isoforms via two hydrogen-bonded ion pairs. *Biochemistry*, 48(10):2146–2155.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10037–10041.

Bashford, D. and Case, D. A. (2000). Generalized Born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51(1):129–152.

Ben-Shimon, A. and Eisenstein, M. (2005). Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid can be used for Detection of Active Sites and Enzyme-Ligand Interfaces. *Journal of Molecular Biology*, 351(2):309–326.

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry, Fifth Edition: International Version (hardcover)*. W. H. Freeman, fifth edition.

Bhattacharya, S. (2010). Towards a matrix mechanics framework for dynamic protein network. *Systems and Synthetic Biology*, 4(2):139–144–144.

Bhattachryya, A. K., Meyer, T. E., and Tollin, G. (1986). Reduction kinetics of the ferredoxin-ferredoxin-NADP$^+$ reductase complex: a laser flash photolysis study. *Biochemistry*, 25(16):4655–4661.

Blundell, T. L. and Fernandez-Recio, J. (2006). Cell biology: Brief encounters bolster contacts. *Nature*, 444(7117):279–280.

Boehr, D. D. and Wright, P. E. (2008). Biochemistry. How do proteins interact? *Science (New York, N.Y.)*, 320(5882):1429–1430.

Boettcher, J. M., Hartman, K. L., Ladror, D. T., Qi, Z., Woods, W. S., George, J. M., and Rienstra, C. M. (2008). Membrane-induced folding of the cAMP-regulated phosphoprotein endosulfine-α. *Biochemistry*, 47(47):12357–12364.

Bolsover, S. R., Hyams, J. S., Shephard, E. A., White, H. A., and Wiedemann, C. G. (2003). *Cell Biology: A Short Course*. Wiley-Liss, 2 edition.

Bosshard, H. R. (2001). Molecular recognition by induced fit: how fit is the concept? *News in physiological sciences : an international journal of physiology produced jointly by the International Union of Physiological Sciences and the American Physiological Society*, 16:171–173.

Bourne, Y. (1996). Crystal structure and mutational analysis of the human CDK2 kinase complex with cell cycle regulatory protein CksHs1. *Cell*, 84(6):863–874.

Bratsun, D., Volfson, D., Tsimring, L. S., and Hasty, J. (2005). Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14593–14598.

Brilliantov, N. V. and Krapivsky, P. L. (1991). Stokes laws for ions in solutions with ion-induced inhomogeneity. *The Journal of Physical Chemistry*, 95(16):6055–6057.

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614.

Brooks, B. R., Bruccoleri, R. E., Olafson, D. J., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217.

Bui, J. M. and McCammon, J. A. (2006). Protein complex formation by acetylcholinesterase and the neurotoxin fasciculin-2 appears to involve an induced-fit mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42):15451–15456.

Butcher, J. C. (2003). *Numerical Methods for Ordinary Differential Equations*. Wiley, 2nd edition.

Camacho, C. J., Weng, Z., Vajda, S., and DeLisi, C. (1999). Free energy landscapes of encounter complexes in protein-protein association. *Biophysical journal*, 76(3):1166–1178.

Cao, Y. and Samuels, D. C. (2009). *Discrete Stochastic Simulation Methods for Chemically Reacting Systems*, volume 454 of *Methods in Enzymology*, pages 115–140. Elsevier.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688.

Cerutti, D. S., Wong, C. F., and Mccammon, J. A. (2003). Brownian dynamics simulations of ion atmospheres around polyalanine and B-DNA: Effects of biomolecular dielectric. *Biopolymers*, 70(3):391–402.

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, 52(1):80–87.

Cheng, Blundell, T. L., and Fernandez-Recio, J. (2007). pyDock: Electrostatics and desolvation for effective scoring of rigid-body proteinprotein docking. *Proteins*, 68(2):503–515.

Churchich, J. (1967). The rotational relaxation time of aspartate aminotransferase. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 147(3):511–517.

Coffey, W. T., Kalmykov, Y. P., and Titov, S. V. (2002). Langevin equation method for the rotational Brownian motion and orientational relaxation in liquids. *Journal of Physics A: Mathematical and General*, 35(32):6789+.

Comeau, S. R., Kozakov, D., Brenke, R., Shen, Y., Beglov, D., and Vajda, S. (2007). ClusPro: performance in CAPRI rounds 6-11 and the new server. *Proteins*, 69(4):781–785.

Davis, S. J., Ikemizu, S., Wild, M. K., and Merwe, P. A. (1998). CD2 and the nature of protein interactions mediating cell-cell recognition. *Immunological Reviews*, 163(1):217–236.

Debye, P. (1942). Reaction Rates in Ionic Solutions. *Transactions of the Electrochemical Society*, 82:265+.

Devos, D., Dokudovskaya, S., Williams, R., Alber, F., Eswar, N., Chait, B. T., Rout, M. P., and Sali, A. (2006). Simple fold composition and modular architecture of the nuclear pore complex. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2172–2177.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737.

Doob, J. L. (1945). Markoff Chains–Denumerable Case. *Transactions of the American Mathematical Society*, 58(3):455–473.

Dudler, T. and Gelb, M. H. (1996). Palmitoylation of Ha-Ras facilitates membrane binding, activation of downstream effectors, and meiotic maturation in xenopus oocytes. *Journal of Biological Chemistry*, 271(19):11541–11547.

Dworkin, J., Lazcano, A., and Miller, S. L. (2003). The roads to and from the RNA world. *Journal of Theoretical Biology*, 222(1):127–134.

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560.

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121.

Elcock, A. H. (2002). Atomistic simulations of competition between substrates binding to an enzyme. *Biophysical journal*, 82(5):2326–2332.

Elcock, A. H. (2003). Atomic-level observation of macromolecular crowding effects: Escape of a protein from the GroEL cage. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2340–2344.

Elcock, A. H. (2010). Models of macromolecular crowding effects and the need for quantitative comparisons with experiment. *Current Opinion in Structural Biology*, 20(2):196–206.

Ellis, J. R. (2001a). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Current Opinion in Structural Biology*, 11(1):114–119.

Ellis, R. (2001b). Macromolecular crowding: obvious but underappreciated. *Trends in Biochemical Sciences*, 26(10):597–604.

Elmegreen, B. G., Koch, R. H., Schabes, M. E., Crawford, T., Ebisuzaki, T., Furusawa, H., Narumi, T., Susukita, R., and Yasuoka, K. (2004). Simulations of magnetic materials with MDGRAPE-2. *IBM J. Res. Dev.*, 48(2):199–207.

Ermak, D. and Buckholz, H. (1980). Numerical integration of the Langevin equation: Monte Carlo simulation. *Journal of Computational Physics*, 35(2):169–182.

Ermak, D. L. (1975). A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *J. Chem. Phys.*, 62:4189–4196.

Ermak, D. L. and McCammon, J. A. (1978). Brownian dynamics with hydrodynamic interactions. *The Journal of Chemical Physics*, 69(4):1352–1360.

Evans, E., Ritchie, K., and Merkel, R. (1995). Sensitive force technique to probe molecular adhesion and structural linkages at biological interfaces. *Biophysical journal*, 68(6):2580–2587.

Eyring, H. (1935). The Activated Complex and the Absolute Rate of Chemical Reactions. *Chemical Reviews*, 17(1):65–77.

Favre, I., Moss, G. W. J., Goldenberg, D. P., Otlewski, J., and Moczydlowski, E. (2000). Structureactivity relationships for the interaction of bovine pancreatic trypsin inhibitor with an intracellular site on a large conductance Ca2+-activated K+ channel. *Biochemistry*, 39(8):2001–2012.

Fine, R., Dimmler, G., and Levinthal, C. (1991). FASTRUN: A special purpose, hardwired computer for molecular simulation. *Proteins*, 11(4):242–253.

Finer, J. T., Simmons, R. M., and Spudich, J. A. (1994). Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature*, 368(6467):113–119.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, 38(suppl 1):D211–D222.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges*, 27:2984–2993.

Fokker, A. D. (1914). Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Ann. Phys.*, 348(5):810–820.

Ford, G. W., Lewis, J. T., and McConnell, J. (1979). Rotational Brownian motion of an asymmetric top. *Physical Review A*, 19(2):907–919.

Fulton, A. B. (1982). How Crowded Is the Cytoplasm? *Cell*, 30:345–347.

Gabdoulline, R. R. and Wade, R. C. (1997). Simulation of the diffusional association of barnase and barstar. *Biophys J*, 72(5):1917–1929.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. M. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional, 1 edition.

Garcia de la Torre, J., Huertas, M. L., and Carrasco, B. (2000). Calculation of Hydrodynamic Properties of Globular Proteins from Their Atomic-Level Structure. *Biophysical Journal*, 78(2):719–730.

Garcia de la Torre, J., Lopez Martinez, M. C., and Garcia Molina, J. J. (1987). Approximate methods for calculating rotational diffusion constants of rigid macromolecules. *Macromolecules*, 20(3):661–666.

Gillespie, D. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434.

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucl. Acids Res.*, 35(suppl_1):D291–297.

Hall, D. (2003). Macromolecular crowding: qualitative and semiquantitative successes, quantitative challenges. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, 1649(2):127–139.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443.

Hartley, R. W. (2001). Barnase-barstar interaction. *Methods in Enzymology*, 341:599–611.

He, L. and Niemeyer, B. (2003). A Novel Correlation for Protein Diffusion Coefficients Based on Molecular Weight and Radius of Gyration. *Biotechnology Progress*, 19(2):544–548.

Hermans, J. and Vacatello, M. (1980). Modeling water-protein interactions in a protein crystal. *Biophysical journal*, 32(1):87–88.

Heron, L., Virsolvy, A., Peyrollier, K., Gribble, F. M., Le Cam, A., Ashcroft, F. M., and Bataille, D. (1998). Human -endosulfine, a possible regulator of sulfonylurea-sensitive KATP channel: Molecular cloning, expression and biological properties. *Proceedings of the National Academy of Sciences of the United States of America*, 95(14):8387–8391.

Hofmann, G., Schweimer, K., Kiessling, A., Hofinger, E., Bauer, F., Hoffmann, S., Rösch, P., Campbell, I. D., Werner, J. M., and Sticht, H. (2005). Binding, domain orientation, and dynamics of the Lck SH3-SH2 domain pair and comparison with other Src-family kinases. *Biochemistry*, 44(39):13043–13050.

Homouz, D., Perham, M., Samiotakis, A., Cheung, M. S., and Wittung-Stafshede, P. (2008). Crowded, cell-like environment induces shape changes in aspherical protein. *Proceedings of the National Academy of Sciences*, 105(33):11754–11759.

Hosse, R., Tay, L., Hattarki, M., Pontesbraz, L., Pearce, L., Nuttall, S., and Dolezal, O. (2009). Kinetic screening of antibody-Im7 conjugates by capture on a colicin E7 DNase domain using optical biosensors. *Analytical Biochemistry*, 385(2):346–357.

Inbar, M., Shinitzky, M., and Sachs, L. (1973). Rotational relaxation time of concanavalin A bound to the surface membrane of normal and malignant transformed cells. *Journal of Molecular Biology*, 81(2):245–253.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics*, 9(1):102+.

Ishijima, A., Kojima, H., Higuchi, H., Harada, Y., Funatsu, T., and Yanagida, T. (1996). Multiple- and single-molecule analysis of the actomyosin motor by nanometer-piconewton manipulation with a microneedle: unitary steps and forces. *Biophysical journal*, 70(1):383–400.

Iso14882 (1998). ISO/IEC 14882:1998: Programming languages – C++. Technical report, International Organization for Standardization.

IUCN (2010). IUCN Red List of Threatened Species. version 2010.1. Technical report, IUCN 2010.

Jackson, R. M., Gabb, H. A., and Sternberg, M. J. E. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem1. *Journal of Molecular Biology*, 276(1):265–285.

James, L. C., Roversi, P., and Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. *Science (New York, N.Y.)*, 299(5611):1362–1367.

Janin, J. (2002). Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Genetics*, 47(3):257.

Jefferys, B. R., Kelley, L. A., and Sternberg, M. J. E. (2010). Protein Folding Requires Crowd Control in a Simulated Cell. *Journal of Molecular Biology*, 397(5):1329–1338.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202.

Kasson, P. M., Lindahl, E., and Pande, V. S. (2010). Atomic-resolution simulations predict a transition state for vesicle fusion defined by contact of a few lipid tails. *PLoS computational biology*, 6(6):e1000829+.

Kennedy, J. and Eberhart, R. (2002). Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948.

Kihara, D., Lu, H., Kolinski, A., and Skolnick, J. (2001). TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10125–10130.

Kikugawa, G., Apostolov, R., Kamiya, N., Taiji, M., Himeno, R., Nakamura, H., and Yonezawa, Y. (2009). Application of MDGRAPE-3, a special purpose board for molecular dynamics simulations, to periodic biomolecular systems. *J. Comput. Chem.*, 30(1):110–118.

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104.

Kubo, R. (1966). The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255–284.

Lambeth, J. D., Seybert, D. W., and Kamin, H. (1980). Adrenodoxin reductase/adrenodoxin complex. Rapid formation and breakdown of the complex and a slow conformational change in the flavoprotein. *Journal of Biological Chemistry*, 255(10):4667–4672.

Lameira, J., Alves, C. N., Moliner, V., Marti, S., Kanaan, N., and Tunon, I. (2008). A Quantum Mechanics/Molecular Mechanics Study of the ProteinLigand Interaction of Two Potent Inhibitors of Human O-GlcNAcase: PUGNAc and NAG-Thiazoline. *The Journal of Physical Chemistry B*, 112(45):14260–14266.

Laurent, T. C. and Ogston, A. G. (1963). The interaction between polysaccharides and other macromolecules. *The Biochemical Journal*, 89:249–253.

Lensink, M. F. and Wodak, S. J. (2010). Docking and scoring protein interactions: CAPRI 2009 (in press). *Proteins: Structure, Function, and Bioinformatics*.

Leonard, D. A., Lin, R., Cerione, R. A., and Manor, D. (1998). Biochemical studies of the mechanism of action of the Cdc42-GTPase-activating protein. *Journal of Biological Chemistry*, 273(26):16210–16215.

Lesk, V. I. and Sternberg, M. J. E. (2008). 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. *Bioinformatics*, 24(9):1137–1144.

Lezon, T. R., Sali, A., and Bahar, I. (2009). Global motions of the nuclear pore complex: insights from elastic network models. *PLoS computational biology*, 5(9):e1000496+.

Li, X. F., Moal, I. H., and Bates, P. A. (2010). Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding (in press). *Proteins: Structure, Function, and Bioinformatics*.

Liu, H., Elstner, M., Kaxiras, E., Frauenheim, T., Hermans, J., and Yang, W. (2001). Quantum mechanics simulation of protein dynamics on long timescale. *Proteins*, 44(4):484–489.

Liwo, A., Khalili, M., and Scheraga, H. A. (2005). Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2362–2367.

Liwo, A., Pincus, M. R., Wawak, R. J., Rackovsky, S., and Scheraga, H. A. (1993). Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide. *Protein Science*, 2(10):1715–1731.

London, F. (1930). Zur Theorie und Systematik der Molekularkräfte. *Zeitschrift fur Physik A Hadrons and Nuclei*, 63(3):245–279.

Lortat-Jacob, H., Chouin, E., Cusack, S., and van Raaij, M. J. (2001). Kinetic Analysis of Adenovirus Fiber Binding to Its Receptor Reveals an Avidity Mechanism for Trimeric Receptor-Ligand Interactions. *Journal of Biological Chemistry*, 276(12):9009–9015.

Lyskov, S. and Gray, J. J. (2008). The RosettaDock server for local protein-protein docking. *Nucl. Acids Res.*, 36(suppl_2):W233–238.

Ma, B., Shatsky, M., Wolfson, H. J., and Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein science : a publication of the Protein Society*, 11(2):184–197.

MacKerell, A. D., Bashford, D., Bellott, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998a). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *The Journal of Physical Chemistry B*, 102(18):3586–3616.

MacKerell, A. D., Brooks, C. L., Nilsson, L., Roux, B., Won, Y., and Karplus, M. (1998b). *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*, volume 1 of *The Encyclopedia of Computational Chemistry*, pages 271–277. John Wiley & Sons: Chichester.

Madura, J. D., Briggs, J. M., Wade, R. C., Davis, M. E., Luty, B. A., Llin, A., Antosiewicz, J., Gilson, M. K., Bagheri, B., Scott, L. R., and McCammon, J. A. (1995). Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Computer Physics Communications*, 91(1-3):57–95.

Maher, K. A. and Stevenson, D. J. (1988). Impact frustration of the origin of life. *Nature*, 331(6157):612–614.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng.*, 14(2):105–113.

Margenau, H. and Murphy, G. M. (1943). *The Mathematics Of Physics And Chemistry*. Van Nostrand.

Mazo, R. M. (2009). *Brownian Motion: Fluctuations, Dynamics, and Applications (The International Series of Monographs on Physics)*. Oxford University Press, USA.

McGuffee, S. R. and Elcock, A. H. (2006). Atomically detailed simulations of concentrated protein solutions: The effects of salt, pH, point mutations, and protein concentration in simulations of 1000-molecule systems. *J. Am. Chem. Soc.*, 128(37):12098–12110.

McGuffee, S. R. and Elcock, A. H. (2010). Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Computational Biology*, 6(3):e1000694+.

Meng, T. C. C., Somani, S., and Dhar, P. (2004). Modeling and simulation of biological systems with stochasticity. *In silico biology*, 4(3):293–309.

Minton, A. (2000). Implications of macromolecular crowding for protein assembly. *Current Opinion in Structural Biology*, 10(1):34–39.

Minton, A. P. (1981). Excluded volume as a determinant of macromolecular structure and reactivity. *Biopolymers*, 20(10):2093–2120.

Minton, A. P. (2001). The Influence of Macromolecular Crowding and Macromolecular Confinement on Biochemical Reactions in Physiological Media. *Journal of Biological Chemistry*, 276(14):10577–10580.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005). Protein-Protein Docking Benchmark 2.0: an update. *Proteins*, 60(2):214–216.

Moal, I. H. and Bates, P. A. (2010). SwarmDock and the use of normal modes in flexible protein-protein docking (in press). *International Journal of Molecuar Science*.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540.

Negi, S. S., Schein, C. H., Oezguen, N., Power, T. D., and Braun, W. (2007). InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics*, 23(24):3397–3399.

Nelson, D. L. and Cox, M. M. (2004). *Lehninger Principles of Biochemistry, Fourth Edition*. W. H. Freeman, fourth edition edition.

Nesatyy, V. (2002). Dissociation of noncovalent protein complexes by triple quadrupole tandem mass spectrometry: comparison of Monte Carlo simulation and experiment. *International Journal of Mass Spectrometry*, 221(3):245–262.

Nicholls, A., Sharp, K. A., and Honig, B. (1991). Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, 11(4):281–296.

Northrup, S. H., Allison, S. A., and Mccammon, A. J. (1984). Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *The Journal of Chemical Physics*, 80(4):1517–1524.

Northrup, S. H. and Erickson, H. P. (1992). Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proceedings of the National Academy of Sciences of the United States of America*, 89(8):3338–3342.

Novotny, V., Basset, Y., Miller, S. E., Weiblen, G. D., Bremer, B., Cizek, L., and Drozd, P. (2002). Low host specificity of herbivorous insects in a tropical forest. *Nature*, 416(6883):841–844.

Offman, M., Tournier, A., and Bates, P. (2008). Alternating evolutionary pressure in a genetic algorithm facilitates protein model selection. *BMC Structural Biology*, 8(1):34+.

Pacold, M. E., Suire, S., Perisic, O., Lara-Gonzalez, S., Davis, C. T., Walker, E. H., Hawkins, P. T., Stephens, L., Eccleston, J. F., and Williams, R. L.

(2000). Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase γ. *Cell*, 103(6):931–944.

Palma, P. N., Krippahl, L., Wampler, J. E., and Moura, J. J. (2000). BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39(4):372–384.

Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J., and Zagrovic, B. (2003). Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1):91–109.

Papoian, G. A. (2008). Proteins with weakly funneled energy landscapes challenge the classical structurefunction paradigm. *Proceedings of the National Academy of Sciences*, 105(38):14237–14238.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. T. (1960). Structure of haemoglobin: A three-dimensional fourier synthesis at 5.5 Angstroms resolution, obtained by X-ray analysis. *Nature*, 185(4711):416–422.

Phillip, Y., Sherman, E., Haran, G., and Schreiber, G. (2009). Common Crowding Agents Have Only a Small Effect on Protein-Protein Interactions. *Biophysical Journal*, 97(3):875–885.

Planck, M. (1917). Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. *Sitzungsber. Preuss. Akad. Wiss.*, 24:324–341.

Poole, A. M., Jeffares, D. C., and Penny, D. (1998). The path from the RNA world. *Journal of molecular evolution*, 46(1):1–17.

Porcar, L., Falus, P., Chen, W.-R., Faraone, A., Fratini, E., Hong, K., Baglioni, P., and Liu, Y. (2010). Formation of the Dynamic Clusters in Concentrated Lysozyme Protein Solutions. *The Journal of Physical Chemistry Letters*, 1(1):126–129.

Qin, S. and Zhou, H.-X. X. (2007). A holistic approach to protein docking. *Proteins*, 69(4):743–749.

Ritchie, D. W., Kozakov, D., and Vajda, S. (2008). Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*, 24(17):1865–1873.

Roque, A., Ponte, I., and Suau, P. (2007). Macromolecular crowding induces a molten globule state in the C-terminal domain of Histone H1. *Biophysical Journal*, 93(6):2170–2177.

Rubio, I., Wittig, U., Meyer, C., Heinze, R., Kadereit, D., Waldmann, H., Downward, J., and Wetzker, R. (1999). Farnesylation of Ras is important for the interaction with phosphoinositide 3-kinase gamma. *European journal of biochemistry / FEBS*, 266(1):70–82.

Sack, R. A. (1957). Relaxation Processes and Inertial Effects II: Free Rotation in Space. *Proceedings of the Physical Society. Section B*, 70(4):414+.

Schlosshauer, M. and Baker, D. (2004). Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Science*, 13(6):1660–1669.

Schmierer, B., Tournier, A. L., Bates, P. A., and Hill, C. S. (2008). Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proceedings of the National Academy of Sciences*, 105(18):6608–6613.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research*, 33(Web Server issue).

Schreiber, G., Haran, G., and Zhou, H. X. (2009). Fundamental aspects of protein-protein association kinetics. *Chemical reviews*, 109(3):839–860.

Schrodinger, E. (1944). *What Is Life?* Cambridge University Press.

Schuldiner, S., Spencer, R. D., Weber, G., Weil, R., and Kaback, H. R. (1975). Lifetime and rotational relaxation time of dansylgalactoside bound to the lac carrier protein. *Journal of Biological Chemistry*, 250(23):8893–8896.

Science (2005). So Much More to Know. *Science*, 309(5731):78b–102.

Segal, D. and Eisenstein, M. (2005). The effect of resolution-dependent global shape modifications on rigid-body protein-protein docking. *Proteins*, 59(3):580–591.

Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J., Chao, J. C., Eastwood, M. P., Gagliardo, J., Grossman, J. P., Ho, C. R., Ierardi, D. J., Kolossváry, I., Klepeis, J. L., Layman, T., McLeavey, C., Moraes, M. A., Mueller, R., Priest, E. C., Shan, Y., Spengler, J., Theobald, M., Towles, B., and Wang, S. C. (2007). Anton, a special-purpose machine for molecular dynamics simulation. In *ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture*, pages 1–12, New York, NY, USA. ACM.

Sheinerman, F. (2000). Electrostatic aspects of proteinprotein interactions. *Current Opinion in Structural Biology*, 10(2):153–159.

Sieber, J. J., Willig, K. I., Kutzner, C., Gerding-Reimers, C., Harke, B., Donnert, G., Rammner, B., Eggeling, C., Hell, S. W., Grubmuller, H., and Lang, T. (2007). Anatomy and Dynamics of a Supramolecular Membrane Protein Cluster. *Science*, 317(5841):1072–1076.

Slepoy, A., Thompson, A. P., and Plimpton, S. J. (2008). A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks. *The Journal of Chemical Physics*, 128(20):205101+.

Smoluchowski, M. V. (1917). Versuch einer mathematischen theorie der koagulationskinetic kolloider lösugen. *Zeitschrift f physik chemie*, 92:129–168.

Song, Y., Zhang, Y., Shen, T., Bajaj, C. L., McCammon, J. A., and Baker, N. A. (2004). Finite element solution of the steady-state Smoluchowski equation for rate constant calculations. 86(4):2017–2029.

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

Swegat, W., Schlitter, J., Kruger, P., and Wollmer, A. (2003). MD simulation of protein-ligand interaction: Formation and dissociation of an insulin-phenol complex. *Biophysical Journal*, 84(3):1493–1506.

Tang, C., Iwahara, J., and Clore, G. M. (2006). Visualization of transient encounter complexes in protein-protein association. *Nature*, 444(7117):383–386.

Tang, C., Schwieters, C. D., and Clore, G. M. (2007). Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature*, 449(7165):1078–1082.

Taufer, M., Armen, R., Chen, J., Teller, P., and Brooks, C. (2009). Computational multiscale modeling in protein–ligand docking. *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society*, 28(2):58–69.

Tournier, A. L., Fitzjohn, P. W., and Bates, P. A. (2006). Probability-based model of protein-protein interactions on biological timescales. *Algorithms for molecular biology : AMB*, 1:25+.

Tovchigrechko, A. and Vakser, I. A. (2006). GRAMM-X public web server for protein-protein docking. *Nucl. Acids Res.*, 34(suppl_2):W310–314.

Tsai, C. J., Ma, B., and Nussinov, R. (1999). Folding and binding cascades: shifts in energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 96(18):9970–9972.

Tsai, C. J., Ma, B., Sham, Y. Y., Kumar, S., and Nussinov, R. (2001). Structured disorder and conformational selection. *Proteins*, 44(4):418–427.

Tsai, C.-J. J., del Sol, A., and Nussinov, R. (2008). Allostery: absence of a change in shape does not imply that allostery is not at play. *Journal of molecular biology*, 378(1):1–11.

Tworowski, D. and Safro, M. (2003). The long-range electrostatic interactions control tRNA-aminoacyl-tRNA synthetase complex formation. *Protein science : a publication of the Protein Society*, 12(6):1247–1251.

Tyn, M. T. and Gusek, T. W. (1990). Prediction of diffusion coefficients of proteins. *Biotechnology and Bioengineering*, 35(4):327–338.

Ubbink, M. (2009). The courtship of proteins: Understanding the encounter complex. *FEBS Letters*, 583(7):1060–1066.

Uehara, K., Tse, J., Yasuaki, H., Miyagawa, H., Kitamura, K., and Toyoda, S. (2002). Large Scale Molecular Dynamics Simulation Using MD-Engine. In Pollard, A., Mewhort, D. ., and Weaver, D. ., editors, *High Performance Computing Systems and Applications*, volume 541 of *The Kluwer International Series in Engineering and Computer Science*, chapter 16, pages 115–116–116. Springer US, Boston.

Vakser, I. A. (1997). Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins*, Suppl 1:226–230.

van Zon, J. S. and ten Wolde, P. R. R. (2005). Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space. *The Journal of chemical physics*, 123(23):234910+.

Vercauteren, N. (2005). Numerical investigation of solutions of Langevin equations. Master's thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.

Verlet, L. (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review Online Archive (Prola)*, 159(1):98+.

von Smoluchowski, M. (1906). Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Ann. Phys.*, 326(14):756–780.

Wang, T., Tomic, S., Gabdoulline, R. R., and Wade, R. C. (2004). How Optimal Are the Binding Energetics of Barnase and Barstar? *Biophysical Journal*, 87(3):1618–1630.

Wangemann, P. and Liu, J. (1996). Osmotic water permeability of capillaries from the isolated spiral ligament: new in-vitro techniques for the study of vascular permeability and diameter. *Hearing Research*, 95(1-2):49–56.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Weernink, P. A., Rijksen, G., Mascini, E. M., and Staal, G. E. (1992). Phosphorylation of pyruvate kinase type K is restricted to the dimeric form. *Biochimica et biophysica acta*, 1121(1-2):61–68.

Weiner, J. H. and Forman, R. E. (1974). Rate theory for solids. V. Quantum Brownian-motion model. *Phys. Rev. B*, 10:325–337.

Zhang, C., Vasmatzis, G., Cornette, J. L., and Delisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of Molecular Biology*, 267(3):707–726.

Zhou, H. X. (1990). On the calculation of diffusive reaction rates using Brownian dynamics simulations. *Journal of Chemical Physics*, 92:3092+.

Zhou, H. X. (1997). Enhancement of protein-protein association rate by interaction potential: accuracy of prediction based on local Boltzmann factor. 73(5):2441–2445.

Zhou, H. X., Rivas, G., and Minton, A. P. (2008). Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences. *Annual Review of Biophysics*, 37(1):375–397.

Zimmerman, S. and Trach, S. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *Journal of Molecular Biology*, 222(3):599–620.