UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering, Science and Mathematics

School of Chemistry

Application of simulation methods for the identification of allosteric binding site in human glucokinase

By Nadia Vahdati

A thesis submitted for the qualification of Doctor of Philosophy at the University of Southampton

April 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF CHEMISTRY

Doctor of Philosophy

Application of simulation methods for the identification of allosteric binding site in human glucokinase

by Nadia Vahdati

Allosteric binding site, in generic terms, refers to a binding site distinct from the active site, where binding of an effector can enhance or decrease enzyme's activity. From a drug design point of view targeting allosteric binding sites can offer advantages in terms of selectivity, saturability, and the opportunity for discovering new chemotypes. Current known allosteric drugs have often been identified by chance with high-throughput screening. Although experimental methods can be successful in identifying such binding sites, owing to the cost and time associated with experiments, it would be useful to be able to aid the prediction of the location of such alternative binding sites with computational methods.

In this thesis two computational approaches, molecular dynamics (MD) and normal mode analysis (NMA) have been applied to an allosteric enzyme human glucokinase (GLK) for which known allosteric activators have been discovered, as a test-case. First an X-ray structure, bound to glucose and an allosteric activator, has been examined as a benchmark. The apo form of the enzyme has also been studied for further understanding of the dynamics of this enzyme in an unbound form. We then turned our attention to an X-ray structure, bound only to glucose, where the allosteric binding site is not visible. This allowed for a real scenario where the allosteric binding site is not observed in the static structure in the absence of a suitable activator and for which we aimed to reveal the binding site.

In this system we have successfully revealed the allosteric binding site and would have been able to predict the location with reasonable confidence.

Acknowledgements

I would like to take this opportunity of thank Professor Jonathan Essex for all his support and guidance throughout this research. Thank you to my industrial supervisor Dr. Richard Ward for his involvement and encouragements. Many thanks to the entire Essex group for making my time here full of good memories. A special thank you to all the system administrators in the group during my time here, Julien Michel, Juan Camerona-Fernandez, Mishal Patel and Patrick Schöpf.

Thank you to my family and friends for their continuous love, support and patience throughout my studies, especially Robert. A special thank you to Patrick, George and Diederik for their support in the final stages of this journey.

Last but not least thank you to EPSRC and AstraZeneca for funding this project.

Contents

Chapter 1

Introduction

Allosteric binding sites, in generic terms, refer to a binding site distinct from the active site, where binding of an effector can enhance or decrease the protein's activity. In the classic view of allostery, in a multi-subunit protein, binding of an effector to one subunit would induce conformational changes in that subunit which then propagate to the remaining subunits, enhancing further binding of the effector. The new view of allostery suggests that instead of induction of new conformations in the equilibrium, allosteric regulation takes place via the redistribution of the existing protein conformational ensemble (*1*). In addition, the recent view posits that allostery does not necessarily involve a shape change (backbone conformational change), leading to a thermodynamic definition of allostery, which may involve enthalpic, enthalpic and entropic, or solely entropic effects.

From a drug design point of view targeting allosteric binding sites can offer advantages in terms of target selectivity, saturability of their effect and the opportunity for discovering new chemotypes for enzymes for which finding drug-like active site substrates has been challenging. Current known allosteric drugs have been identified by high-throughput screening and resolved by crystallography (e.g. HIV-1 non-nucleoside reverse transcriptase inhibitors) (*2*). Other methods such as phage display and tethering

has also led to the identification of allosteric binding sites. Although experimental methods can be successful in identifying these sites, the cost and time associated with such experiments limits their use in routine studies of protein targets. As such, there is a great interest in computational approaches to identify novel allosteric binding sites, discussed in chapter 2.

To establish a suitable protocol for the identification of potential binding sites, a test-case is required for which known allosteric modulators have been identified. Human glucokinase (GLK) is an allosteric enzyme that catalyses the ATP-dependent phosphorylation of glucose, which plays a critical role in glucose homeostasis, as evidenced by naturally occurring mutations. Inactivating mutations in the GLK gene have been linked to maturity-onset diabetes of the young, type 2 (MODY2) (*3*) and the activating mutations to persistent hyperinsulinemic hypoglycemia of infancy (PHHI) (*4*). The discovery of allosteric activators for this enzyme in the recent years has offered a promising new therapeutic approach for the treatment of diabetes.

In this thesis three X-ray structures of this enzyme have been studies, two of which are publicly available (*5*) and one that has been provided by AstraZeneca (unpublished data). One of the publicly available X-ray structures is in complex with glucose and an allosteric activator at distinct sites (PDB ID: 1v4s). In the super-open inactive conformation, both the active and allosteric binding sites are absent, since the drastic conformational change leading to this state results in the collapse of both binding sites (PDB ID: 1v4t). The X-ray structure provided by AstraZeneca (unpublished data) is very similar to the publicly available closed state X-ray structure (PDB ID: 1v4s), but the enzyme is only in complex with glucose. In the absence of the allosteric ligand in this structure, the allosteric binding site cannot be observed as a vacant cavity. The research in this thesis has focused on the identification of the allosteric binding site in both the super-open conformation and the X-ray structure provided by AstraZeneca.

At present there is a gap in rational drug design in handling allosteric binding sites and there is a need for the advancements of methods, in particular computational methods which play an important role in rational drug design. The key issues from a structure-based design point of view are: first, to be able to correctly model the protein flexibility using computational techniques; second, to be able to identify any pockets (that may be

allosteric) that appear as a result of conformational change; and third, to design ligands that effectively bind to such allosteric binding sites.

Several conventional computational methods to simulate protein dynamics include normal mode analysis (NMA), Monte Carlo (MC) and molecular dynamics (MD). However, it is frequently impractical to carry out MD simulations of even moderately large macromolecules for times long enough to observe phenomena of interest, like an allosteric transition. For this reason, NMA is gaining popularity. It has been shown that a few normal modes are sufficient to correlate with the direction of known conformational changes. Although much faster than MD, the limitation of this method is the harmonic approximation of the energy potential surface in an energy minimum, where small oscillations can be derived, which may not be sufficient to capture conformational changes associated with the dynamic nature of the protein structure that may be captured in an MD simulation. Here we assess whether the application of this method is suitable for capturing conformational changes associated with allosteric binding sites appearing, particularly in GLK.

Molecular dynamics and normal mode analysis have been utilised in this thesis to treat the protein dynamics. The possibility of reducing the binding site search to the principal motions has also been explored. This has been combined with binding pocket detection studies to determine if the binding site would be predicted without the bias of using the prior experimental knowledge in the study.

In the following chapters current approaches to capturing allosteric mechanisms through experimental and computational approaches will be discussed (chapter 2) and the computational methods utilised in this thesis will be outlined (chapter 3). The background to GLK, the target of interest in our studies will be presented in chapter 4, followed by a benchmarking of the application of the chosen computational methods on a structure for which the allosteric site is occupied. In chapters 6 and 7, results will be presented for an X-ray structure, for which we attempt to reveal the allosteric binding site.

Chapter 2

Identification of allosteric binding sites

In this chapter the general concept of allostery, the advantages of targeting allosteric binding sites from a drug design point of view, and current approaches for identifying such binding sites will be discussed. From a drug design point of view targeting allosteric drugs can offer advantages in terms of selectivity, saturability of their effects and the opportunity for discovering new chemotypes. Current known allosteric binding sites have been identified indirectly through experimental methods. Owing to the highly advantageous nature of such sites from a drug design point of view, attention has been steered towards such sites, both in the experimental and computational communities (*2, 6, 7*).

2.1 Introduction to allostery

The purely lock and key analogy of protein ligand binding originally postulated by Fischer over 100 years ago is now outdated (*8*). We now know that enzymes are flexible, and that the shape of the binding sites can be considerably modified by substrate binding (*9*). Ligand binding is a continuous process of combining a conformational selection stage and partially fitting structures, followed by minor structural rearrangements within the complex (*10*). This process of dynamic recognition is called induced-fit (*11*). All proteins either undergo some conformational change ranging from very small to large collective motions to achieve their functional role, and

especially in response to ligand binding. This property gives some proteins the ability to be allosterically modulated in order to shift substrate binding affinities, alter enzymatic activity, or regulate protein-protein interactions, which is a branch of the induced-fit phenomenon (*11*).

Allosteric regulation is the regulation of an enzyme or protein by binding an effector molecule at the protein's allosteric site, distinct from the active site (*9*). Effectors that enhance the protein's activity are referred to as allosteric activators, whereas those that decrease the protein's activity are called allosteric inhibitors.

Originally, only proteins with multiple subunits were thought to be able to achieve cooperative ligand binding, such as that observed in haemoglobin, where binding of oxygen to one subunit induces a conformational change in that subunit which then propagates to the remaining three subunits available for oxygen binding, thereby enhancing oxygen affinity. This definition of allostery has been expanded to any case in which an event at one site on a protein impacts function, dynamics, or distribution of conformations at another site, including in single domain proteins (*12*).

The new view of allostery is that instead of inducing new conformations in the equilibrium, allosteric regulation takes place via the redistribution of the existing protein conformational ensemble (*1*). Binding of an allosteric modulator changes the environment of the target protein, leading to a shift in the distributions of the conformational substates (*13*).

Early models of protein allostery, only applicable to oligomeric protein assemblies of identical protein molecules, assumed each subunit either in a relaxed state (R) or tense state (T), where the R state would be more receptive to ligand binding than the T state (*5*).

Two classical allosteric models, Monod-Wyman and Changeux (MWC) (*14*), also known as the concerted model, and Koshland-Nemethy-Filmer (KNF) (*15*), or the sequential-model, have been used to describe the allosteric mechanism in multi subunit assemblies. In the MWC model, a conformational change in one subunit causes an equivalent change in all others. The ratio of the R and T state is determined by thermal equilibrium. Binding of a ligand to either state increased the equilibrium in favour of the R state, therefore making the ensemble more receptive to further ligand binding (*16*). In

the KNF model, the subunits of each assembly do not need to be in the same state. Binding to one subunit by induced-fit changes the state of that subunit from T to R, affecting the other subunits, which leads to more receptive conformations of the subunits but not necessarily the R state.

The two classical view of allostery explicitly assume that the multiplicity of the quaternary structure (i.e. the arrangement of multiple folded protein molecules in a multi-subunit complex) does not change during the allosteric transition, where the dramatic changes in the structure and function of the individual subunits are recognised but not the quaternary multiplicity. Jaffe (*17*) expands the classical allosteric models for oligomeric systems to include a dynamic equilibrium of protein structures wherein a protein monomer can exist in more than one conformation and each monomer conformation can lead to a different quaternary structure of finite multiplicity and different functionality. A prototype system where this concept can be demonstrated is the allosteric regulation of porphobilinogen synthase (PBGS) by magnesium. PBGS is capable of assembling into at least two different oligomers; predominately a high-activity octamer; in humans it can form a low-activity hexamer, leading to different functionality.

Recent views of allostery are more complex than the previously presented two state models. It is becoming broadly accepted that proteins exist as complex classical ensembles of conformers and that proteins unfold and fold continuously in localised regions, which can result in a large number of conformations with subtle conformational differences (*18*). Binding of an allosteric modulator can lead to a change in this distribution of protein molecules that reside in the same or different conformational states.

Tsai et al. (*1*) discusses allostery in terms of thermodynamics, postulating that communication across proteins could be mediated not only by changes in the mean conformation but also by changes in the dynamics fluctuations about the mean conformation. Allosteric communication could not only involve enthalpic contribution in the case of conformational changes at the binding site, but it could also involve entropic contribution (*19-22*). In cases where no conformational change is observed in the backbone of the protein system (possible side-chain reorientations), allostery may involve only entropic contributions. Therefore, we cannot purely view allostery in

terms of conformational change. The authors divide allosteric proteins into three groups. Allostery controlled by entropy in the case of no or subtle changes in the backbone; by enthalpy and entropy where minor conformational changes are observed; and by enthalpy in the case of relatively large domain conformational changes (*1*). The authors categorise Type I, to a conformational change of less than 1 % of total number of residues, ~ 3 residues between the allosteric protein pairs, an example of which is hemoglobin (PDB ID: 2hhbA and 1hhoA). A conformational change between 3 % and 10 % (~10 residues), is classed as subtle conformational change, an example of which is FixJ (PDB ID: 1dbwA and 1d5wA). Between 10 % and 35 % of unmatched residues in the protein pairs, refers to minor changes, named type II, an example of which is arf6 (PDB codes: 1e0sA and 1hfvA). Finally, type III, with more than 35% or ~ 60 unmatched residues is classified as domain movements, for which purR (PDB codes: 1dbqA and 1wetA) is an example.

Recent updated views of allostery are useful in understanding allosteric mechanisms and thus the prediction of allosteric sites. The different levels of conformational change between allosteric protein pairs, once fully understood may require the application of various computational methods and protocols depending on the type of allosteric mechanism, which could lead to the prediction of allosteric binding sites and the design of allosteric drugs.

Allosteric modulators (effectors) offer a number of advantages over conventional agonist or antagonist drugs. Allosteric drugs frequently exhibit little or no intrinsic activity where their mode of action is to enhance or inhibit the action of the endogenous agonist (*23*). Targeting an allosteric binding site in a protein can also offer advantages in terms of selectivity. Although the allosteric site could be a binding site used by an unidentified endogenous ligand (an orphan allosteric site), this is not necessarily the case. Therefore the allosteric binding site would not have evolved to accommodate an endogenous substrate, but in the presence of a suitable synthetic ligand such sites may adapt to yield a suitable cavity for an allosteric ligand (*24*). Such binding sites will have much higher sequence diversity across target families compared to the endogenous site which would possibly bind similar natural substrates, i.e. ATP in kinase family. This offers the additional benefit of fewer side effects.

A key property of allosteric modulators is that their action is saturable, which enables possible dissociation of the intensity of effect from the duration of the effect (*25*); once the allosteric binding site is fully occupied, further increase in the allosteric modulator concentration will not lead to further target-based effects (i.e., once the allosteric sites are fully occupied, no further allosteric effect is observed). This can be used as a strategy for prolonging the duration of modulation without influencing the magnitude of the effect, by increasing the concentration of the allosteric modulator at the receptor compartment. This also reduces the risks of over-dose in comparison to conventional orthostatic drugs (the endogenous binding site on a receptor, i.e. active site in an enzyme). (*23*).

An additional advantage of targeting the allosteric binding site from a drug design point of view is the opportunity of discovering new chemotypes for enzymes for which finding drug-like active site substrates have been challenging, e.g. in G protein-coupled receptors (*26*).

A possible implication in the use of allsoteric drugs is that different allosteric modulators can affect the affinity of the same agonist/antagonist at the orthostatic site to varying degrees, indicative of the production of different allosteric conformational states, which can have physiological consequences (*25*). As an example, HIV can undergo resistance mutations (*27*) where the prolonged presence of allosteric modulators can lead to a mutated HIV virus that has a new conformation which will not be inhibited. In this case a combination of several orthostatic and allosteric drugs can reduce the risk.

Most allosteric binding sites have been identified through experimental methods high-throughput screening (HTS), and resolved by crystallography. As such, the allosteric binding site in human liver glycogen phosphorylase (GlyP) (*28*), kinesin spindle protein (KSP) (*29*), fructose-1,6-bisphophatase (F168BPase) (*30*), protein tyrosine phosphatase 1B (PTP1B) (*31, 32*), β-Lactamase (*33*), p38 kinase (*34*), glucokinase (*5*) and HIV-1 RT (*35*) have all been identified with this route. The allosteric binding site of factor VIIa (FVlla) (*36*) was discovered by phage display and resolved through crystallography. Tethering was used for the discovery the allosteric binding site in Caspase-3 and Caspase-7 (*37*).

Kinetic experiments can be used to understand the mechanism of allostery for a target (*18*) by measuring the rates and affinity of binding. Methods such as X-ray crystallography, nuclear magnetic resonance (NMR), fluorescence or cross-linking experiments can be utilised to determine the location of the binding site.

In the following sections the experimental and computational methods utilised in the literature to identify regions involved in allosteric regulation will be discussed.

2.2 Experimental methods

2.2.1 High-throughput screening

High throughput screening (HTS) is the central function in most pharmaceutical drug-discovery processes. Traditional HTS can be outlined in several steps (figure 2.1).

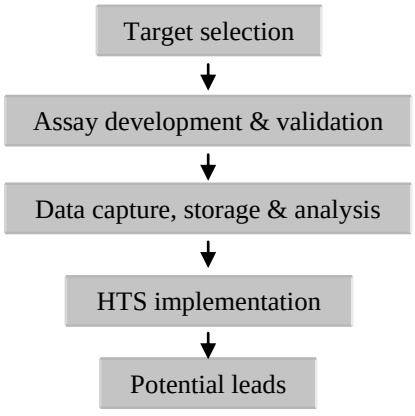


Figure 2.1: Steps involved in high throughput screening. Figure adapted from ref. (*38*).

The process begins with the target selection which is motivated by the disease area, but factors important for the HTS process should also be considered. These factors include technical issues such as the ability to design a robust assay for the target of interest, optimisation of the process for quality, throughput and cost. The next important feature

of a high quality assay is the biochemical data obtained from the assay and statistical performance.

Most of the allosteric binding site discoveries described above were serendipitous, where the HTS was designed for targeting the active site, but featured unusual outcomes which led to X-ray crystallography and the identification of the site of action, which was distinct from the active site.

As an example, PTP1B, a promising target for the treatment of obesity and type II diabetes, the discovery of the allosteric binding site in this system is of particular interest as the discovery of pharmaceutically acceptable inhibitors that bind to the active site remains substantially challenging (*32*). Owing to a high sequence identity (~72%) with T-cell protein tyrosine phosphate (TCPTP), active site inhibitors are equipotent on TCPTP. In addition to the selectivity issue, it has proven challenging to design phosphotyrosine (TyrP) mimetics with suitable pharmalogical properties such as cell permeability and bioavailability.

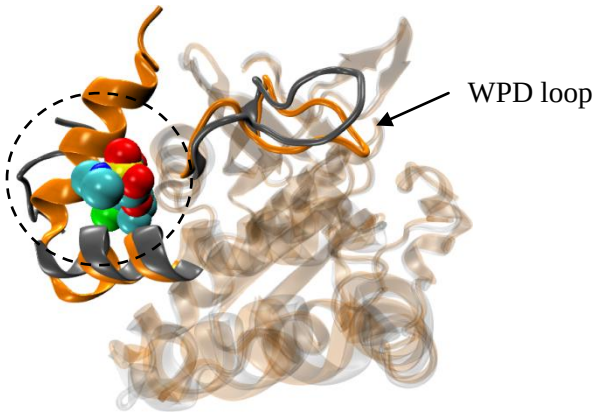


Figure 2.2: The allosteric binding site of PTP1B. The active state structure (PDB ID: 1pty) is depicted in orange for regions that undergo significant conformational change; the remaining residues are displayed in a transparent representation. The corresponding residues in the inactive state (PDB ID: 1t48) in gray, bound to the allosteric inhibitor, which influence the catalytic WPD loop (Trp179, Pro180, Asp181) conformation at the active site, and cause disorder in the terminal residues.

Binding of the allosteric inhibitor at a site ~20 Å away from the active site propagates a conformational change to the WPD loop at the active site, hindering the binding of a natural substrate (figure 2.2).

2.2.2 Peptide phage display

Phages are viruses that only infect bacterial cells and are not infectious to humans. Phages are made up of two components; the genetic material and a protein coat. The phage genome cause infected bacteria to make more phage and the protein coat protects the DNA when the phage goes from cell to cell. In phage display, new genetic material is inserted into a phage gene. The bacteria process the new gene so that a new protein or peptide is made. This protein or peptide is exposed on the phage surface. It was developed first as a technology to link the phenotype of a bacteriophage (phage) surface peptide with the genotype encoding that peptide, packaged within the same phage particle (virion) (*39*). This has led to a screening technique of large diverse peptide libraries for receptors and other protein or non-protein targets. When the phage enters a cell its DNA is expressed causing more phages to be made. The coat protects the genetic material from being damaged as the phages reproduce and go from cell to cell. Phages used in phage displays are normal phages that have been genetically modified to express one extra protein that will be on the exterior of the phage.

Generally a diverse set of genes are inserted into the phage's existing genome (*40*). Each phage receives only one gene; therefore, the modified gene expresses a single protein or peptide, creating a population of phage displaying related but diverse proteins or peptides. The related proteins keep most of their chemical and physical properties of their parent protein, providing a library of proteins that can be screened with a target of interest. The exposure of the library to a target receptor can lead to the binding of some members of the library to this receptor through an interaction with the displayed peptide on the phage surface. The proteins isolated by this method are possible drug candidates because of their tight and specific binding to disease target.

This method was used to identify the allosteric inhibitors of factor VIIa (FVIIa), a critical protease in the blood-clotting cascade. Two classes of peptides were found to bind to two distinct sites (*41*, *42*) (figure 2.3). One class changes the hydrogen bond

network at one end of the protein, which causes a hydrophobic collapse, leading to the disruption of the oxyanion hole. The other class does not influence the oxyanion hole, but subtly disrupts the substrate-binding site.

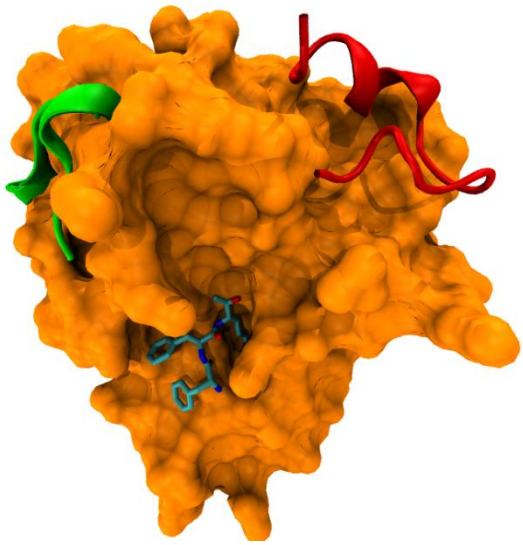


Figure 2.3: FVIIa represented in an orange surface with the relative locations of the two classes of peptide inhibitors, indicated in red and green at two distinct sites (not bound at the same time). To demonstrate the location of the active site, an active site substrate has been depicted in a cyan stick representation.

2.2.3 Tethering

Tethering is a useful technique in fragment based drug discovery. This method has been utilised in the discovery of allosteric binding sites in caspases (*37*). This method makes use of surface cysteine residues. The method relies on the reversible formation of a disulphide bond between native or an engineered cysteine residue in the protein and member of a library of thiol-containing fragments (*43*). This amplifies the affinity of the fragment for the target molecules, enabling detection at lower concentrations.

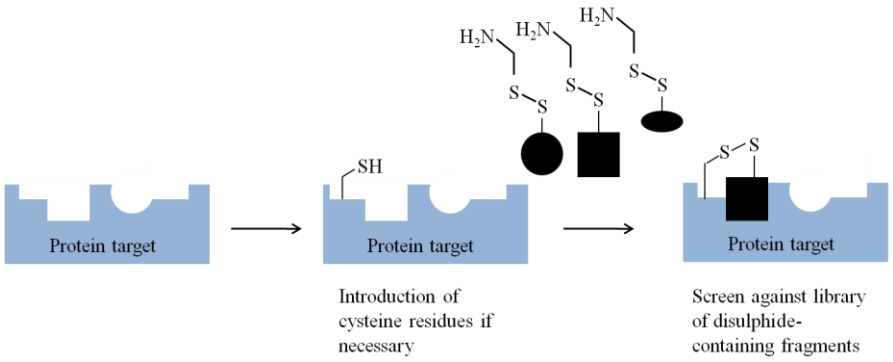


Figure 2.4: Tethering fragment-based drug discovery. (Figure adapted from reference (*43*))

The cysteine residue should be within 5 to 10 Å of the target site on the surface of the protein to facilitate the thiol-disulphide exchange (figure 2.4). The library of disulphide-containing fragments are then reacted with the target protein, under reducing conditions. Assuming there are no non-covalent interactions between the fragments and the protein and that the reactivity of the disulphide in each fragment is the same, the reaction pool at equilibrium should consist of the protein disulphide-bonded to each fragment in equal proportion. However, if one of the fragments has inherent affinity for the protein of interest, and if it binds to the protein near the desired or introduced cysteine, then the thiol-disulphide equilibrium will be shifted in favour of the disulphide for this fragments, and this protein-fragment complex will predominate (*43*).

A library of ~10,000 disulphide-containing fragments were screened with active caspases-3 and -7 containing five surface-exposed cysteines in addition to the catalytic cysteine. Mass-spectrometry analysis demonstrated the binding location of the fragments to a single cysteine residue, at a site distinct from the active site.

X-ray crystallography demonstrated very different binding modes of inhibitors although at the same site (*2, 4, 37*). Figure 2.5 depicts the location of the allosteric inhibitor at the dimer interface.

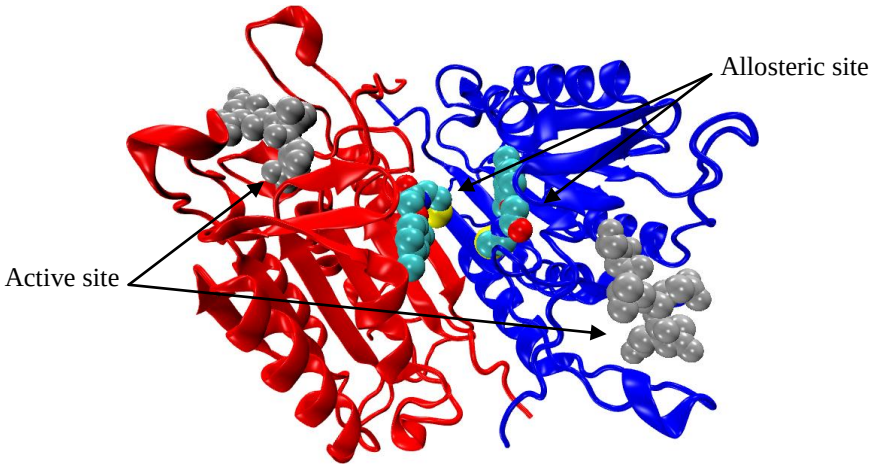


Figure 2.5: The allosteric site in caspase-7 at the dimer interface, represented in cyan spheres. The backbone structure refers to the active state structure bound to the peptide substrate at the active sites, in gray sphere representation.

The application of this method can be useful in particular when a region in the protein structure has been identified to have a key role in the regulation of the target, but for which a binding pocket is not necessarily observed.

2.2.4 NMR-based screening methods

NMR can provide information concerning the dynamics of proteins. This usually involves measuring relaxation times to determine order parameters, correlation times, and chemical exchange rates. The technique is able to measure movements over a broad range of timescales, ranging from 10 ps to ms, which capture the fast and slower motions. The overall rotational diffusion can also be measured (5-50 ns), as well as protein domain movements which take place on a very slow timescale (ms - days) (*44*). The application of nuclear magnetic resonance (NMR) spectroscopy for structure determination of proteins and nucleic acids with molecular mass exceeding 30 kDa is largely constrained by two factors, fast transverse relaxation of spins of interest and complexity of NMR spectra, both of which increase with increasing molecular size (*45*). NMR in combination with the transverse relaxation optimised spectroscopy (TROSY) detection method (*46*) can extend its application to larger systems such as a 91 kDa

allosteric proteins. TROSY uses spectroscopic means to reduce transverse relaxation based on the fact that cross-correlated relaxation caused by the interference of the dipole–dipole interaction and chemical shift anisotropy gives rise to much smaller transverse relaxation rates at high fields in a system of two coupled spins (1/2), *I* and *S*, such as the ^{15}N-^{1}H moiety in the protein backbone and the ^{13}C-^{1}H moiety in the aromatic group of amino acids, allowing NMR studies of much larger proteins and nucleic acids (*47*). NMR measurements of residual dipolar couplings in solution have also been applied to characterise dynamic processes (*48*). Therefore, allosteric regulation on these sorts of time-scales can be established where it may even lead to the identification of regions where binding could influence the regulation.

NMR-based screening is an extensively utilised approach in drug discovery efforts, where small molecule ligands or fragments are screened for a macromolecular target, monitoring the changes observed in NMR parameters that occur upon their binding. For small molecules, the parameters include the longitudinal, transverse, and double-quantum (DQ) relaxation; diffusion coefficients; and intermolecular and intramolecular magnetisation transfer.

In fragment-based drug design, NMR screens of small fragments can give an indication of hot spot locations on the protein surface, in addition to the affinity of the ligand or fragment. This method is useful as even ligands with very weak affinities can be detected (*49*). This method can be used to detect small cavities which may bind certain classes of fragments, which are in close proximity to each other, and could be joined up into one ligand. Presumably the same is applicable in the identification of potential allosteric binding sites. Jahnke et al. report a strategy for the identification and optimisation of allosteric protein kinase inhibitors (*50*). If possible, targeting the allosteric binding site in kinases offers advantages in terms of selectivity, in comparison with the active site that shares a high degree of sequence homology to other targets in the kinase family, binding ATP. The authors discuss an NMR detection method involving a spin-labelled adenine analogue (figure 2.6).

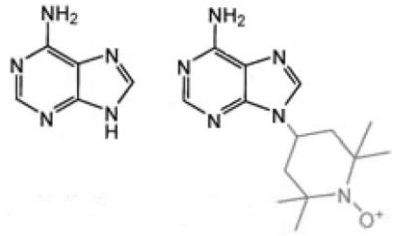


Figure 2.6: Adenine on the left and adenine spin label on the right. The spin-labelled adenine analogues can be utilised in the NMR screen of allosteric inhibitors of kinases.

The free electron on the spin label which is paramagnetic, exerts drastic and long-ranging paramagnetic relaxation enhancement effects on any nuclear spin within a distance of 15-20 Å (*50*). The use of spin labelling of the protein side chains as in the SLAPSTIC experiments (**s**pin **l**abels **a**ttached to **p**rotein **s**ide chain as a **t**ool to **i**dentify interacting **c**ompounds) (*51*) can be efficient in detecting a protein ligand interaction, while spin labelling of a given ligand allows the detection of a second ligand that binds simultaneously, within the vicinity of the first ligand. As such this technique can be useful in fragment-based drug design. This can be extended to the identification of allosteric ligands when the binding site of the known ligand is within the vicinity of the limits of this method. The application of this method in kinases is useful as if a spin-labelled adenine binds at the ATP binding site, no other ligand can bind at this site, and therefore any ligand binding within the 15-20 Å will be detected, causing a paramagnetic relaxation enhancement on the second ligand.

2.3 Computational approaches

Although experimental methods can be rather successful in the identification of allosteric binding sites, the cost and time associated with such experiments limits their use in routine studies of protein targets. As such, there is a great interest in computational approaches to identify novel allosteric binding sites.

2.3.1 Identification of regions involved in allosteric regulation

A predictive method named COREX (*52, 53*) based on monitoring stability in the protein system, considers a thermodynamic statistical ensemble in which a state is characterised by having some region(s) in a non-folded state, starting from a high resolution template structure. This region can be as short as a few residues or as long as the entire protein, but depending on the size of system, the computation is exhaustive or performed using a sampling technique (*54*). The algorithm generates a large number of different conformational states through the combinatorial unfolding of a set of predefined folding units. The Boltzmann weight of each state is determined from the calculated Gibbs energy, and the probability of each state is determined. In comparison with an evolutionary method such as MD, the goal of this method is to identify the most probable state (i.e. the minima in the landscape). This method can be used to monitor the effect of ligand binding on this ensemble probability. In application to Escherichia coli dihydrofolate, the linkage between two binding site were explored and the shift in the probability distribution of the conformational ensemble was demonstrated for the unbound and bound states (*55*). While the algorithm provides significant insights, it uses a reduced model for the degrees of conformational freedom available to a residue, as each residue exists in either a folded or unfolded state (*7*).

Another approach is the Statistical Coupling Analysis (SCA) (*56, 57*), where pairs of residues that tend to be mutated together in multiple sequence alignments suggest coupling between protein sites (*58*). This method makes use of functional information buried within the evolutionary record from the sequences of a family of proteins (*2*). All sequences within a family are examined and divided into allosterically and non-allosterically regulated family members. This method has been able to validate previous evidence on the networks of amino acids that transmit allosteric signals in the G-protein coupled receptor, haemoglobin and chymotrypsin serine protease families (*57*). A drawback of this approach is that a large collection of sequence alignments is required for the study.

Evolutionary trace analysis is yet another method in the identification of functionally important protein surfaces (*59*); in its most basic form it requires a multiple sequence

alignment of a protein family and an evolutionary tree, based on sequence identity, which can approximate the functional classification of the protein sequences. However the application of either method, the SCA or evolutionary trace analysis, has not led to the identification of novel or orphan allosteric sites.

Another dynamical approach is to focus on the protein vibrations around a static structure through atomistic or coarse-grained normal mode analysis (*60, 61*). This method has been applied in this thesis. Even at the atomic level a pitfall of this approach lies in the harmonic approximation of the energy surface, where perturbations that sample multiple wells on the energy surface would be out of reach, for instance when multiple stable conformations of the enzyme are observed. Despite the approximation in this approach, it has been shown to accurately predict conformational transition between two distinct state of protein system (i.e. opening/closing of two domains) (*62-65*). This method will be discussed in more detail in chapter three.

In most cases methods based on dynamics generate an ensemble of conformations which are then analysed with a range of techniques such as cross-correlations (*66*), contact correlations (*67*), principal components (*67*), or local unfolding correlations (*68*). Most of these correlation methods are based on the calculation of an average structure, which may not be suitable for large conformational changes. Other methods exist that aim to account for both quasi-harmonic and anharmonic correlations in Cartesian space (*69, 70*) or methods that introduce mechanical perturbations and monitor the subsequent motions of residues, in which they can detect substantial population shifts (*71, 72*).

A very recent method is an entropy-based approach for analysing protein conformational ensembles, generated by molecular dynamics or Monte Carlo simulations, called MutInf (*58*). The method is based on the concept of mutual information. When conformational flexibility of one residue affects the conformational flexibility of another residue, changes in entropy and conformational information are exchanged (*73*). Mutual information quantifies the amount of conformational dependence between protein residues. The method calculates the configurational entropies from conformational ensembles using only internal coordinates to focus on low frequency motions, including the anharmonic side chain rotamers.

In application to human interleukin-2 (IL-2Rα), the method correctly identified local correlations in sequence and in distance space, in addition to the long-range correlations. After clustering the matrix of mutual information between residues, a few clusters demonstrated a strong pattern of correlation, of which two refer to the functional sites in IL-2Rα. One is at the IL-2Rα-competitive protein interface/inhibitor binding sites and the other at a highly flexible loop that moves to reveal a cryptic binding pocket for the allosteric ligand.

Despite the promising observations, authors point to a few limitation of the method at this stage. As the method focuses on coupled residue conformations rather than on vibrations, the method may not efficiently capture the role of semi-rigid elements in mediating correlation between more flexible sites. In the application to IL-2Rα, motions faster than 1 ps were not included in the analysis, and considered as non critical for ligand binding cooperativity in this system. The faster-time scale motions can help mediate cooperativity between flexible sites and therefore there is a need to incorporate such fast motion into the analysis in the future.

After the improvements have been made, future applications of this method should determine the robustness of the approach in the identification of allosteric binding sites.

Using structural level computational approach for the identification of allosteric binding sites first requires the correct representation of protein dynamics, however small the allosteric changes may be, as the static structure is only a snapshot in the pre-existing ensemble of conformations. The use of methods such as MD and NMA have been mentioned earlier to handle the protein dynamics. Whether correlation studies are used to point to a particular region of the system or not, a correct binding site identification algorithm is key in characterising such sites. Once a site location has been identified the next challenge lies in the accurate representation of the binding site and finally to be able design a ligand to bind that site with the desired effect. Computationally there are many binding site detection methods which will be briefly reviewed in the following section.

2.3.2 Binding site identification methods

The ability to identify binding sites on a protein surface that could bind to small, drug-like compounds is a very important and useful concept in drug design (*74*). Of particular interest are the allosteric binding sites.

One of the major difficulties in predicting protein-ligand complexes is taking into account the flexibility of the target protein (*75*). In the case of enzymes, it is widely accepted that flexible loop regions have a critical functional role (*76-78*). Loop flexibility allows correct positioning of catalytic residues. Triosephosphate isomerise is a classical example of an enzyme in which a loop acting as a rigid lid closes the active site upon ligand binding (*76*). Regardless of the method of dynamics treatment that precedes this step, any pocket detection algorithm will process one static frame at a time; therefore it is important that the structure(s) fed into this step are as representative as possible of the conformational states of the system in hand. Conformational change in enzymes has been classified into two types, domain motion and loop motion. In the case of domain motion, domains joined by a flexible hinge move relative to each other. In the case of loop motion, the loops adopt different conformations (*79-81*).

Another challenge stems from the fact that all cavities on the surface of a protein are not suitable for ligand binding, and it is not well understood what distinguishes binding sites from other cavities (*82*).

Computational methods for identifying binding sites can be broadly put into three groups; geometric methods, energy-based methods and probe mapping/docking algorithms that use a variety of small molecules and functional groups as probes and hence provide direct input for fragment-based drug design. Of those, the geometric and energy-based methods will be discussed here, while the docking/mapping of fragments is not directly applicable to the preliminary identification of the location of the binding site in this thesis; once the location of the binding site is identified, one may apply fragment docking to the binding for the de Novo design of suitable ligands (*83-86*).

The geometric approaches cover the protein surface with a layer of spherical probes and then filter out those that clash with the protein or are not sufficiently buried (*87*). These methods generally identify pockets with no account for any measure of druggability

(*74*). Among geometric methods, SURFNET (*88*), LIGSITE (*89*), PASS (*90*), CAST (*91*), and PocketPicker (*92*) are all based on different detailed algorithms, either by fitting virtual spheres on the surface of the protein or by defining grid points which correspond to a binding site. Many more algorithms exist in the literature but these methods give no indication of favourable binding as they solely rely on geometric criteria to find cleft and surface depressions, capitalising on the studies that show that the actual ligand-binding site usually coincides with the largest pocket on protein surfaces (*93, 94*). This may prove fairly successful when looking for the active site of a protein where it is likely that the site would be large, but when looking for alternative binding sites, more subtle methods are required, as these sites are not necessarily very large.

Two known published algorithms of this type (energy-based) are PocketFinder (*87*) and Q-SiteFinder (*95*). Both calculate the van der Waals interaction energy of methyl probes with the protein on a grid, and retain probes with favourable interactions. PocketFinder determines the clusters of such probes and sorts the clusters by their volume. In Q-SiteFinder the clusters of probes with favourable interaction are ranked according to their total interaction energies. While no systematic comparison of the two methods have been carried out, both identify the known ligand binding sites among the highest ranked pockets. An et al. noted that in a large set of complexes studied, the top two largest predicted binding sites cover close to 93% of real binding sites. However, the results of these methods still do not necessarily provide sufficient structural information for fragment-based design (*74*).

Another approach is the Multiple Copy Simultaneous Search (MCSS) which processes numerous ligand copies simultaneously, each transparent to the others but subject to the full force of the receptor (such as ionic bonds, hydrogen bonding and van der Waals forces) (*96*). The method determines where specific functional (chemical) groups have local potential energy minima in the binding site. The minima are analysed and are then connected with $(CH_2)_n$ linkers to form candidate ligands. This method can lead to too many minima on the surface of the protein, and it is difficult to determine which of the minima are actually relevant (*74*).

The application of all of the above methods in an extensive way is dependent on the reliability and practicality of the tool/method. When studying one static structure, the processing of the information gathered from the binding site detection tool can be relatively simple. However when analysing large number structure in the MD conformational ensemble approach, it must be possible to at least partly automate the search. This relies on sufficient information from the detection tool, instead of a purely visual inspection.

Another major challenge in binding site detection is the actual drugability of the binding site, and the relevance of the location of the detected binding site to the proteins' regulation. In this case it may be necessary to rely on additional knowledge in a particular protein target family, such as naturally occurring mutations and/or experimental kinetic observations. Structural information in protein families such as flexible loop regions in proteins and hinge regions may focus the search.

It may prove very useful to combine residue correlation information from correlation analysis of MD simulations with pocket detection algorithms to eliminate less relevant sites. In the case of the active site, it is likely that the binding site will be the largest site but this will not necessarily be the case for an allosteric binding site.

2.4 Summary

The advantages of targeting allosteric binding sites in protein have been discussed. Such sites offer better selectivity, are saturable and offer the opportunity for the design of new chemotypes. It is now widely accepted that allostery does not necessarily need to involve large conformational change but it may be a shift in the pre-existing conformational equilibrium.

In targeting allosteric binding sites, it may be necessary to classify the allosteric regulation and understand some of conformational differences in the bound and unbound states before embarking on identifying allosteric sites in the system. Different approaches may be necessary for different types of allostery and target families. Different computational methods discussed in this section are still subject to limitations and have not directly led to the discovery of novel allosteric binding sites.

In this thesis we go back to standard routine simulation methods and combine these with binding site search algorithms to further understand the challenges involved in identifying allosteric binding sites, in particular for a monomeric allosteric system.

In the following section the computational methods used in this thesis will be discussed in detail. In addition, the two binding site detection methods Q-SiteFinder (*95*) and Pocket-Finder (an implementation of LIGSITE) (*89*) will be discussed and the reason for choosing the two methods will be outlined. Note that the method "Pocket-Finder" used in our study is a geometric approach, named by Laurie et al. (*95*) and is different to "PocketFinder", an energy based method developed by An et al. (*87*) mentioned earlier in this section.

Chapter 3

Computational methods

In the previous chapter computational methods typically used to account for the protein flexibility, in the context of allostery have been discussed. Here, normal mode analysis (NMA) and molecular dynamics (MD) utilised in this project will be discussed in detail. In addition, the analysis methods applied to the simulations and the binding site search methods will be described.

NMA and MD are both mainly carried out at the molecular mechanics level when applied to large systems, such as proteins with thousands of atoms. In comparison with quantum methods (QM) where the electrons are explicitly represented in the calculation, at the molecular mechanics level the electronic motions are ignored, and the energy of the system in calculated as a function of nuclear positions, described by a force field. Most force fields describe the system by four main terms, including bond stretching, angle bending, torsional rotation and a term for the non-bonded interactions, the van der Waals and electrostatic interactions. These will be discussed in the following section.

3.1 Force fields

Most systems studied in molecular modelling are too large to be considered by quantum mechanics. Force field methods ignore the electronic motions and calculate the energy of a system as a function of nuclear positions only (*97*), which limits the method to applications that do not involve drastic electronic redistribution such as bond making/breaking.

Thus the potential energy of a system is described in terms of the intra- and inter-molecular forces within the system (Equation 3.1) (*97*).

$$U(\mathbf{r}^N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \quad (3.1)$$

$U(\mathbf{r}^N)$ is the potential energy as a function of positions **r** of N particles. The first three terms deal with the specific internal degrees of freedom within molecules, describing the bonded interaction. The first term is a harmonic potential between bonded atoms that gives the contribution to the energy when the bond length l_i deviates from the equilibrium value $l_{i,0}$, where *k* is the stretching constant. The second term is a harmonic potential in the valence angles of the molecules, where k is the force constant, θ_i and $\theta_{i,0}$ are the angle and the corresponding angle at equilibrium. The third term is a torsional potential describing the periodic variation in energy due to bond rotations, where V_n is referred to as the barrier height of rotation, *n* is the multiplicity term (i.e., the number of minima in the function as the bond is rotated through 360°), ω the torsion angle and γ is the phase factor.

The last term represents non-bonded interactions, describing interactions between atoms in separate molecules or between atoms separated by three or more bonds in the same molecule. Van der Waals interactions (attraction/repulsion) are short-range and die off rapidly as the atoms move away from each other, and repulsion occurs when the distance

between the interacting atoms (*i* and *j*) becomes marginally less than the sum of their contact radii (σ), described by the Lennard-Jones potential, where ε_{ij} and σ_{ij} are the Lennard-Jones well-depth energy and collision diameter parameters and r_{ij} is the inter-atomic distance.

The Coulomb electrostatic potential is used to model the electrostatic interaction between non-bonded pairs of atoms, where q is the partial atomic charge and ε_0 the dielectric constant of free space (permittivity) and r the distance between two charges. The electrostatic term is harder to calculate as it does not drop off rapidly over distance, as do the van der Waals attractions. Additionally, long-range electrostatic interactions are often of interest in simulations. A basic method to deal with the calculation is to apply a cut-off, but this results in an abrupt discontinuity. An improved method is switching, where the electrostatics are smoothed between the inside and outside of the cut-off radii. More improved and complex methods also exist, e.g. Particle Mesh Ewald (PME) (*98*) where the calculation of electrostatic forces and energies are separated into short- and long-range interactions. The short-range interactions are calculated explicitly, whereas the long-range interactions are summed in Fourier space. The parameters of atoms (includes atomic mass, partial charges, equilibrium values for bond lengths and angles), together with the above potential energy terms constitute the force field. For the more commonly used force fields (e.g. CHARMM (*99*), AMBER (*100*), GROMOS (*101, 102*)), parameters are often calibrated to experimental results and quantum mechanical calculations of small organic model molecules (e.g. calculations of conformational energies and hydrogen bonding (*103*)), and their ability to reproduce physical properties measurable by experiment is tested.

More sophisticated force fields may contain further terms, to refine the system representation; however, the equation above is the basic skeleton for most force field methods. Alternative force fields have been developed, which are derived from other types of experimental data, e.g. enthalpy of sublimation and enthalpy of vaporisation (e.g. OPLS (*104*)). A limitation of most current empirical force fields is that they are unable to model bond formation or breakage, thus making these force fields inadequate for modelling chemical reactions. Additionally, since the atom parameters are considered as constants, typically, the effective partial fixed charges are assigned to the atoms independent of the environment, which are adjusted to account for the influence

of induced polarization in an average way (*105*). The functional form of the Coulomb interaction potentials thus created is not capable of adapting the charge distributions to changes of polarity in the environment. Such force fields, which are commonly termed ‘‘additive’’, are currently used for most biomolecular simulations. To address these limitations, mixed quantum mechanical - molecular mechanical force fields (e.g. reactive force fields (*106*)) are currently under development. In this study, the all-atom force field AMBER03 (*100*) has been used.

3.2 Normal mode analysis

Normal mode analysis (NMA) is one of the major modelling techniques used to probe the large-scale, shape changing motions in biological molecules (*107, 108*). It has been introduced to biomolecules in the early 1980s and made the modelling of large protein motions accessible (*109, 110*). The largest fluctuations (lowest frequency modes) are thought to be associated with functionally relevant motions (*111*).

Normal modes are coupled vibrations that are found by assuming that the molecular potential energy can be approximated as a quadratic or harmonic function of the energy surface in the vicinity of a well-defined energy minimum, and then solving a generalised eigenvalue problem to give an analytical description of the motion (*110, 112*). Thus the anharmonic potential energy surface of a molecule is approximated by a harmonic surface, around this conformation, which is also the fundamental restriction of this method (*113*). A stable conformation corresponds to a local minimum of a smooth, slowly varying potential. The potential well is not smooth, but full of local minima and energy barriers of smaller height, referred to as conformational substates (differing in terms of arrangements of side-chains) (*113*)..

A standard NMA involves three steps, including the minimisation of the system as a function of atomic Cartesian coordinates, the construction of a Hessian matrix, elements of which are the second derivatives of the potential energy with respect to the mass-weighted Cartesian coordinates and finally, the diagonalisation of the Hessian matrix, which is the bottleneck of this method, to produce the eigenvectors and eigenvalues associated with the normal modes (*111*). A protein system with N atoms consists of 3N normal modes, of which the first six have eigenvalues of zero, corresponding to rigid-body translational and rotational movements of the protein that occur at no energy cost

at all. The reason the diagonalisation step is regarded as the bottleneck of the process is due to the size of the Hessian matrix which increases as the square of the number of atoms. In terms of computational memory this corresponds to 1-2 GB of computer memory for ~4000 atom protein system (*114*).

Normal mode vectors describe in which direction each atom moves, and how far it moves relative to the other atoms (not an absolute amount of displacement). Additional information (e.g. temperature) is required for fixing the global amplitude of the atomic displacements (*113*). Each eigenvector is often referred to as a normal mode with certain vibrational frequency. The frequency is determined by the eigenvalue. The overall dynamics of the molecular system can be described by a superposition of a number of linearly independent normal modes.

The very first NMA calculation of a biological molecule was performed in the early 1980s for bovine pancreatic trypsin inhibitor, which is a small protein containing only 58 residues (*107*). The approach was based on the use of dihedral angle space, neglecting the other degrees of freedom, resulting in the reduction of the Hessian matrix size. Results showed that for this system most modes with frequencies above 50 cm^{-1} behave harmonically at room temperature. Some anharmonicity were observed in modes with frequencies lower than 50 cm^{-1}. 174 modes with frequencies below 120 cm^{-1} were generally nonlocal where the protein molecule behaves like a continuous elastic body. The method produced a frequency spectrum in agreement with experiments, but until more recently, where methods have advanced by dividing a protein into blocks (*63, 115*), or simplified by coarse-graining (*116-118*) and faster computers became available, relatively small systems could have been studied.

In most cases as the number of degrees of freedom in proteins is large, various methods have been used to reduce the complexity and size of NM calculations of large molecules. These methods are usually based on the reduction of the number of degrees of freedom of the system.

For large symmetric protein systems a great simplification of the NM calculation is possible without any loss of information by using methods based on group theory as applied to fully flexible all-atom icosahedral viruses (*119*), which often consist of more than 50,000 residues. This is done by representing the system in "symmetry

coordinates", thereby significantly reducing the size of the Hessian matrix. This method has been implemented in the software package CHARMM (*120*).

The DIMB (Diagonalisation In a Mixed Basis) method developed by Perahia and Mouawad (*121, 122*), performs an iterative diagonalisation based on mixing of subsets taken from two different basis sets: low-frequency NM and Cartesian coordinates. The method was applied to systems as large as the aspartate transcarbamylase (3N = 79,758) (*123*). The method is based on the principle that approximate low frequency modes may be refined if they are coupled to higher frequency degrees of freedom.

In recent years methods have been developed where each residue (or sets of residues) have been considered as a rigid block (figure 3.1) having six translation-rotation degrees of freedom (Rotational-translational block (RTB) or block normal mode (BNM) method, a coarse-grained procedure) (*124, 125*), which reduce the size of the Hessian matrix considerably and consequently the computational time.

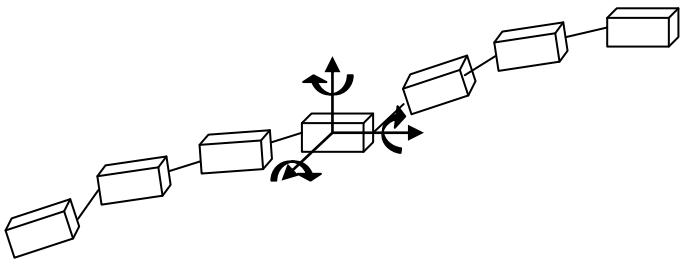


Figure 3.1: Placing one or sets of residues in each rigid block to reduce the degrees of freedom in the RTB method.

The quality of the modes depends on the number of residues taken in a block. The one residue per block yields low-frequency modes comparable to those obtained for all degrees of freedom (0.84 to 0.96 correlation), while increasing the number of residues in each block results in a rapid deterioration of the mode. The block size can be system dependent and influenced by computational resources available (*124*).

Since in RTB, all atoms in the system are included in each block, this approach accounts for the heterogeneity of amino acid types, but the results involve artificial rigidification of the motions (*119*). Another disadvantage is that the initial 3Nx3N full Hessian matrix will have to be stored for the blocks to be extracted.

Another approach is the elastic network model (ENM) (*116*), which is based on a simplified potential where the structure is only maintained by springs between neighbouring atoms, which is consistent with the solid-like nature of proteins. To avoid the minimisation stage, the potential is constructed such that the geometry corresponds to the minimum energy. For large system, the size of the matrix can be reduced by only considering the coordinates of the Cα atoms. Although this model yields modes in agreement with observed conformational changes, in application to the project at hand, this method cannot be used directly because of the level of generalisation. Such methods demonstrate little or no effect of point mutations or small bound ligands/drugs on the large-scale motions (e.ge. associated with channel gating) (*126*). This unified representation of all atoms may under-represent the otherwise subtle effects of electrostatics or van der Waals interactions between protein and ligands. The lack of ligand-induced changes within the ENM method and the absence of residue sidechains, limits the use of this methods in the application to the identification of drug binding sites, owing to the need for the presence of the side-chains in characterising a binding pocket. However such approximate methods could be a good initial step in the process of identifying the expected major global movements and then with that detail in mind one could move to more elaborate methods.

Methods such as ENM and RTB neglect the coupling between the retained degrees of freedom and those that were disregarded, which may be important for a detailed description of the conformational change mechanism. One way to take all degrees of freedom into account is to make use of iterative schemes to diagonalise the Hessian matrix, such as methods based on Rayleigh Quotient (*127*) or perturbation method (*115*) or the mixed basis method (*121*). The first two methods have not been applied to very large system, but the mixed basis method is the only one that has been implemented directly within the molecular simulation package CHARMM (*120*) and applied to systems as large as 2878 residues.

Another method, the Guassian Network Model (GNM) (*117*), inspired by the ENM, uses an NxN matrix instead of the 3Nx3N force constant matrix, where the residue-residue contacts are described. This method is isotropic, with no indication of direction of motion. However, the anisotropic network mode (ANM) (*128*), is an extension to this method, which adds the direction details for the residue fluctuations. ANM is basically the coarse-grained (CG) version of ENM except in assuming uniform mass for each amino acid (*129*).

In both methods, the structures are coarse-grained representing each residue with only one site (bead) and representing the interaction of these by a harmonic potential (springs) between sites that are sufficiently close to lie within a cut-off distance. The main difference lies in the dynamic quantities that can be derived from each method (*130*). GNM only characterises the size of fluctuations; ANM determines the 3N components, and therefore the directions of the fluctuation vector. Despite the more informative nature of ANM, GNM demonstrates more robust evaluation of the energy landscape and is the preferred over ANM for the evaluation of the mean-square fluctuation and the square displacements in low frequency modes (*128*).

Despite the advancement of simplified methods, for the application to the identification of alternative binding site (or allosteric sites), the above coarse-grained methods cannot be applied for the protein motion, as the atomic details are required by the following step; the ligand binding site search. In addition, it is believed that in GLK, the side-chain interactions at the allosteric site contribute to the stability of the closed state, confirmed by the activating mutations that can occur in this region. Another solution would be to model the sidechains back onto frames obtained by CG methods, but this can introduce artefacts that will affect the reliability of our results.

The united-atom force field, where only polar hydrogens are explicitly defined, can reduce the size of the system. As the non-polar hydrogens may not have a crucial functional role, it would be acceptable to reduce the system size in this way. In the Amber united-atom force used (ff03ua) (*131*), as well as the polar hydrogen atoms, the aliphatic hydrogens on all alpha carbon atoms are also represented explicitly to minimise the impact of the united-atom approximation on protein backbone conformations.

Aromatic hydrogens are also explicitly represented. In a system with 6952 atoms, by using this united-atom force field the size of the system is reduced to 4962 atoms.

Many studies, comparing NMA, experimental data and MD simulations, still confirm the usefulness of this method in application to the study of protein motion. Earlier studies suggested that most of the motion in the protein system (e.g. the closing/opening of a protein) could be described by a few normal modes (*62-65*). Recent studies have highlighted that a few normal modes may not be sufficient to capture a functional motion. Van Wynsberghe et al. (*132*) further illustrated this point through the block normal mode analysis study of the voltage gated ion channel, KvAP. Covariance matrix plots of individual modes and combination of various number of modes, demonstrated that the correlation between the flexible domains are not correctly represented by the first or second modes and in fact for this system ~ 50 modes are required to arrive at a converged covariance matrix. The authors highlight the point that individual modes are only a single piece of the overall motion of the molecule and despite the fact a few normal modes may be sufficient in visualising and representing structural flexibilities, motional correlation, particularly important to allosteric regulation cannot be represented by a few normal modes. As motional correlations can be both positive and negative, the correlation of one pair of residue from one low-frequency mode can be negated or even reversed by a group of modes that display an opposite correlation (*132*). Although a few normal modes can often overlap well with the large-scale conformational changes, direct motional coupling cannot be elucidated when considering a few normal modes.

One other issue in standard NMA is the lack of an explicit water representation. The presence of water would increase the friction and dampen the atomic fluctuations (*133, 134*).

Owing to the availability of resources, in this project full atomistic NMA has been carried out. This will allow us to be sure that the results will not be in any way influenced by potential inaccuracies in approximate CG methods. Considering the background understanding of NMA and its limitations, if the allosteric binding site cannot be observed along the atomistic NMs, this method will not be suitable for application to the novel binding site search study, in particular in the context of allostery.

3.2.1 Standard NMA theory

The theory is based on the idea that, at the energy minimum on an energy surface, the potential energy function V(**r**) has a quadratic form and can be expanded in a Taylor series in terms of mass-weighted Cartesian coordinates **r**, $r_i = \sqrt{m_i}\Delta x_i$, where Δx_i is the displacement of the i^{th} coordinate from the energy minimum and m_i is the mass of the corresponding atom, $(r_1 = \sqrt{m_1}x_1,\ r_2 = \sqrt{m_1}y_1,\ r_3 = \sqrt{m_1}z_1,\ r_4 = \sqrt{m_2}x_2, \ldots)$.

For small atomic displacements the potential energy function is expanded as a Taylor series:

$$V(\mathbf{r}) = V(\mathbf{r}^0) + \sum_{i=1}^{3N}\left(\frac{\partial V}{\partial r_i}\right)_0 (r_i - r_i^0) + \frac{1}{2}\sum_{i,j=1}^{3N}\left(\frac{\partial^2 V}{\partial r_i \partial r_j}\right)_0 (r_i - r_i^0)\left(r_j - r_j^0\right) + \ldots \quad (3.2)$$

The index 0 refers to a reference structure that should be a minimum structure. The first term is the minimum value of the potential which may be set to zero. The second term is the first derivative of the energy potential which is zero at a minimum; both $(\partial V/\partial r_i)_0 = 0$ and $V(\mathbf{r}^0) = 0$. Additionally for sufficiently small displacements, the terms beyond second order may be neglected, reducing $V(\mathbf{r})$ to the quadratic form:

$$V(\mathbf{r}) = \frac{1}{2}\sum_{i,j=1}^{3N} H_{i,j}\ (r_i - r_i^0)\left(r_j - r_j^0\right) \quad (3.3)$$

where $H_{i,j} = (\partial^2 V/\partial r_i \partial r_j)_0$ refers to the force constant relative to coordinates *i* and *j*.

In matrix format equation (3.3) can be expressed as follows:

$$V(\mathbf{r}) = \frac{1}{2}\,\mathbf{r}^{\mathrm{T}}\mathbf{H}\mathbf{r}$$

(3. 4)

where **r** is now a vector of 3N coordinates differences ($r_i - r_i^0$), $\mathbf{r}^{\mathbf{T}}$ the transpose of **r**, and **H** the 3Nx3N Hessian matrix, the elements of which are the second derivatives of the potential with respect to the components of **r**.

Because NMA is applied to the study of dynamics, it is necessary to account for kinetic energy as well as potential energy. The equation of motion can be written as:

$$\mathbf{M}\frac{\mathrm{d}^2\,\Delta\,\mathbf{r}}{\mathrm{d}t^2} + \mathbf{H}\,\Delta\,\mathbf{r} = 0$$

(3.5)

where **M** is a diagonal matrix contains the masses of the particles (*135*). Each mass is repeated three times, once for each of the particle's three Cartesian coordinates. A solution to this differential equation is the 3N-dimensional vector $\mathbf{u}_k(t) = \mathbf{a}_k \exp\{-i\omega_k t\}$ where $\mathbf{a}_k$ is a complex vector containing both amplitude and phase factor, and ω_k is the frequency of the mode of motion represented by $\mathbf{u}_k(t)$.

The amplitude, C_k of normal mode *k* at temperature *T*, is defined as:

$$C_k = \frac{\sqrt{2k_BT}}{\omega_k}$$

(3.6)

where k_B is the Boltzmann constant and ω_k the frequency (*136*).
Substituting $\mathbf{u}_k(t)$ into equation 3.5, the equation of motion can be described by:

$$\mathbf{H}\mathbf{u}_k = \omega_k^2\mathbf{M}\mathbf{u}_k$$

(3. 7)

which is a generalised eigenvalue equation. In the matrix form, the complete set of $\mathbf{u}_k(t)$, $1 \leq k \leq 3N$, and the corresponding squared frequencies ω_k^2 may be arranged into the columns of the matrix **U** and elements $\lambda_k = \omega_k^2$ of the diagonal matrix **Λ** to rewrite the set of 3N equations represented by equation 3.7 in compact notation:

$$\mathbf{HU} = \mathbf{MU\Lambda} \tag{3. 8}$$

It is normal to rewrite equation 3.8 by using mass-weighted Cartesian coordinates so that the dependence on the mass matrix, **M** is removed from the right-hand side, by introducing the inverse square root of the mass matrix, $\mathbf{M}^{-1/2}$, which is equal to the diagonal matrix of the inverse square roots of the atomic masses (*137*).

$$(\mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-\mathbf{1/2}})\left(\mathbf{M}^{\mathbf{1/2}}\mathbf{U}\right) = \mathbf{\Lambda}\,(\mathbf{M}^{1/2}\mathbf{U}) \tag{3. 9}$$

Equation 3.9, in short notation can be written as:

$$\widetilde{\mathbf{H}}\widetilde{\mathbf{U}} = \mathbf{\Lambda}\widetilde{\mathbf{U}} \tag{3.10}$$

The mass weighted second-derivative matrix, $\widetilde{\mathbf{H}}$, may be solved to obtain the eigenvalues, λ, and the eigenvectors, $\widetilde{\mathbf{u}}_k(t) = \mathbf{M}^{1/2}\mathbf{u}_k$. Because the matrix has 3N×3N dimensions, there will be 3N different solution, each of which represents an independent displacement that they system can make (*137*). These displacements are the normal modes, for which the associated frequency λ, is the square root of the modes' eigenvalue.

In this study the low frequency modes will be examined to establish if the allosteric binding site does appear along the global deformations, which are expected to be the opening and closing of the two domains in GLK.

3.3 Molecular dynamics

One of the principal tools in the theoretical study of biological molecules is the method of molecular dynamics simulations (MD). This computational method calculates the time dependent behaviour of a molecular system.

Molecular dynamics simulations have proven to be useful in identifying pathways between the opened and closed conformations of protein systems, which have not been possible to achieve through experimental techniques (*138*). However, to observed allosteric transitions, timescales of microsecond or milliseconds may be required (*48, 139-141*). Although the advancement of technology has largely extended the time scales and system sizes accessible for MD simulations, typical timescales performed are on the nanoseconds scale (*142*).

Methods have been developed to overcome the current time-scales limitations of MD; as such, steered molecular dynamics (SMD) (*143, 144*), targeted molecular dynamics (TMD) (*145, 146*), conformational flooding (*147*) and essential dynamics sampling (*148, 149*) extend the rate of sampling compared with standard MD simulation by inducing an external force or biasing the degrees of freedom of the simulated system.

Steered molecular dynamics (SMD) imposes restraints e.g. a harmonic potential on particular atoms, restricting the simulation to a particular degrees of freedom which steers the system along a prescribed path. This accelerates a process that would otherwise be too slow to be observed in a conventional MD simulation. The release of ADP from the cAMP-dependent-protein-kinase A (PKA) was observed to dissociate by the application of the SMD method (*150*).

TMD establishes a distance constraint between two structures, the initial and final, such that the sum over the distances of all atoms between the two structures has to be reduced in every step of the simulation. A successful application of the TMD method has been observed in the study of the conformational transition pathway of human glucokinase,

from a super-open inactive conformational state to the closed active conformational state (see figure 4.2) (*151*).

In conformation flooding, an artificial potential that destabilises the initial conformation and thereby lowers the free energy barriers of structural transitions is applied, allowing the system to explore new regions of phase space (*147*). Using a multi-variate Gaussian potential padded upon the original landscape, the molecule can be nudged out of an energy minimum, thus achieving a wider conformational search (*152*). In application to carbonmonoxy myoglobin (MbCO), the method was effective and practically robust in computing motions that correlate well to previously found ligand diffusion pathways thought the protein (*153*).

Similarly, in essential dynamics, the sampling is enhanced relative to conventional MD by geometric constraint along a selection of principal modes. The method yields an enhanced sampling of the configurational space by extrapolation of the essential subspace, as approximated from relatively short MD simulations (*149*). Original studies on the histidine containing phosphocarrier protein (HPr) from Escherichia coli demonstrated clusters of conformations of an order of magnitude larger than that found for with conventional MD simulation of comparable length (*149*).

In addition to the methods above, digitally filtered molecular dynamics (DFMD) (*154*) can be also be used to enhance or suppress the vibrational motions within the system, thereby allowing better observation of motion in regions of interest in the system, by favouring low-frequency motion. By performing a series of MD simulations while monitoring the spectral density and the adjustment of digital filters, it is possible to focus on motions in regions of interest at any desired frequency.

In our studies we hope to be able to predict protein conformational flexibility assuming no prior knowledge of the protein mechanism; thus, methods such as TMD would only prove useful in the case where structural knowledge is available. As we hope to be able to apply our protocol to any generalised system, methods such as DFMD are not directly applicable, as regions of interest must be pre-defined. Once the overall conformational flexibility of the system has been identified, one could use methods mentioned in this section to enhance the motions in regions of interest.

3.3.1 Molecular dynamics theory

In MD, the method is based on classical mechanics, where the rate and direction of motion in a system is governed by the forces that atoms of the system exert on each other as described by Newton's equation (Equation 3.11):

$$\mathbf{F}_i = m_i \mathbf{a}_i \tag{3. 11}$$

where m_i is the mass of atom *i*, and $\mathbf{F}_i$ is the force acting on atom *i*, and $\mathbf{a}_i$ is acceleration which is the first derivative of velocity with respect to time and the second derivative of position with respect to time.

The equation of motion can also be expressed with respect to the potential energy function *V* of the system. The force on atom *i* in the system can then be determined as the negative of the gradient of the potential energy (Equation 3.12).

$$\mathbf{F}_i = -\nabla V \tag{3.12}$$

The combination of equations 3.11 and 3.12 leads to equation 3.13, from which the positions of atoms r_i with respect to time *t*, can be related to the derivative of the potential energy *V*.

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{3.13}$$

In a more realistic many body system, the force on each particle will change whenever the particle changes its position, or whenever any of the other particles with which it interacts change position. The equations of motion must be calculated simultaneously, which rapidly becomes complicated and cannot be solved analytically for more than two particles. In a many body system the equations of motion are solved using a finite difference method, which assumes that positions and dynamic properties can be calculated using a Taylor series, with position, velocity and acceleration as shown in equations (3.14-3.16) respectively.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\boldsymbol{a}(t)\delta t^2 + \cdots \tag{3.14}$$

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t) + \boldsymbol{a}(t)\delta t + \frac{1}{2}\boldsymbol{b}(t)\delta t^2 + \cdots \tag{3.15}$$

$$\boldsymbol{a}(t+\delta t) = \boldsymbol{a}(t) + \boldsymbol{b}(t)\delta t + \cdots \tag{3. 16}$$

where $\boldsymbol{r}$ is the position, $\boldsymbol{v}$ is the velocity, $\boldsymbol{a}$ the acceleration and $\boldsymbol{b}$ is the first derivative of acceleration with respect to time.

The essential idea is that the integration is broken down into many small stages, each separated in time by a fixed time step δt. The total force on each particle in the configuration at a time t is calculated as the vector sum of its interactions with other particles. From the force one can determine the accelerations of the particles, which are then combined with the positions and velocities at time t to calculate the positions and velocities at time $t + \delta t$. The force is assumed to be constant during the time step. The important factors to consider when choosing the algorithm are first to conserve energy and momentum, second, that the method is computationally efficient, and third that the algorithm permits a long time step for integration.

The simplest finite difference method is the Verlet algorithm (*155*). The positions and particle acceleration at time *t*, and the positions from previous step, **r**(*t* – *δt)* are used to calculate new positions. However the method has a major drawback due to unsatisfactory treatment of velocities. Velocities do not appear explicitly in the Verlet scheme. The velocities have to be calculated or estimated which means that the method is not self-starting.

A modified Verlet scheme in which velocities appear explicitly is the velocity Verlet algorithm (*156*) in which the positions, velocities and acceleration are all stored at the same time:

$$\boldsymbol{r}(t+\ \delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\,\boldsymbol{a}(t)\delta t^2 + \cdots \tag{3.17}$$

$$\boldsymbol{v}(t+\ \delta t) = \boldsymbol{v}(t) + \frac{1}{2}\,\delta t[\boldsymbol{a}(t) + \boldsymbol{a}(t+\delta t)] \tag{3.18}$$

As seen in equation 3.18, the acceleration must be calculated at both *t* and *t* + *δt*. Therefore, first the positions at *t* + *δt* are calculated according to equation 3.17, using the velocities and accelerations at time *t*. The velocities at time $t + \frac{1}{2}\delta t$ are then determined by the following equation:

$$\boldsymbol{v}\left(t + \frac{1}{2}\delta t\right) = \boldsymbol{v}(t) + \frac{1}{2}\,\boldsymbol{a}(t)\delta t \tag{3. 19}$$

The forces can then be determined from the new positions, allowing the calculation of ***a***(t + *δt*). In the final step, the velocities at time *t* + *δt* are determined by:

$$\boldsymbol{v}(t+\ \delta t) = \boldsymbol{v}(t+\frac{1}{2}\delta t) + \frac{1}{2}\,\boldsymbol{a}(t+\delta t)\,\delta t$$

(3. 20)

In addition to Velocity verlet, other methods that improve upon the Verlet algorithm include the leap-frog (*157*), and Beeman(*158*) algorithms.

The important factors in choosing an integration algorithm include the stability (i.e. conservation of energy), accuracy, speed and computational efficiency. The stability of the integrator algorithm is dependent on the timestep used. The biologically relevant behaviour of systems takes place mostly on relatively large time scales, often in the range of seconds, but time steps of a few femtoseconds (fs) are usually used to conserve the energy of the system. The small timestep means that the trajectory covers an insufficient proportion of the phase space, whereas, the use of a large time step leads to forces changing too quickly, instabilities in the integration algorithm and degradation in energy conservation due to high energy overlaps between atoms. Therefore, there is a trade-off between accuracy and economy. A solution is to dampen the high-frequency vibrations by constraining the bonds involved to their equilibrium values, while still allowing the remaining degrees of freedom to vary under the intramolecular and intermolecular forces present. The most commonly used method is the SHAKE algorithm (*159*), where the equations of motion are solved, whilst simultaneously satisfying the imposed bond constraints.

For simulations to be realistic and comparable to experimental data, simulation conditions must create the correct environment for the system in question. Some of those conditions include the temperature, pressure, boundary set-up, solvent and so on, some aspects of which will be discussed in the following sections.

3.3.2 Thermodynamic conditions

The thermodynamic state of a system is usually defined by a number of parameters which include the temperature, volume, pressure and number of particles which make up

the system. These conditions characterise a particular thermodynamic state, called an ensemble. The microcanonical ensemble (NVE) is the traditional thermodynamic ensemble in MD, characterised by a fixed number of atoms, N, a fixed volume, V, and a fixed energy, E. The two most common alternatives are the canonical ensemble (NVT), and the isobaric-isothermal ensemble (NPT). In NVT, as its name implies, the number of atoms, the volume and temperature are fixed, whereas in NPT, the number of atoms, the pressure and temperature are the fixed conditions.

In simulating the canonical ensemble, there are different approaches to controlling the temperature which add and remove energy from the simulation in a realistic manner. The main temperature scaling methods include the Berendsen thermostat (*160*), Nosé-Hoover thermostat (*161*), Andersen thermostat (*162*) and Langevin dynamics (*163*). In the Berendsen method the system is coupled to an external heat bath that is fixed at a desired temperature. The heat is removed or supplied as necessary. The velocities are scaled at each step, so that the rate of change of temperature is proportional to the difference in temperature between the bath and the system. This method however does not generate rigorous canonical averages where the velocity scaling artificially prolongs any temperature differences among the components of the system (*164*). The Andersen thermostat introduces a stochastic element to the temperature by having random collisions of molecules with an imaginary heat bath at desired temperature. However, the presence of random collisions creates uncorrelated particle velocities. As a result true molecular kinetics are not preserved by the Anderson thermostat (*165, 166*).

The Langevin thermostat (*163*) follows the Langevin equations of motion (*167*) rather than Newton's equations of motion. At each time step all particles receive a random force and have their velocities lowered using a constant friction. The frictional force decreases the temperature because it is a fixed positive value. The random force is selected from a Gaussian distribution and adds kinetic energy into the particles of the system, with its variance being the function of the selected temperature and time step. This results in the random force being balanced with the frictional force and maintains the system temperature at the desired value.

Similar to the temperature control in molecular dynamics simulation, depending of the study, it may be desirable to maintain the system at constant pressure. The pressure is

controlled by adjusting the volume of the unit cell (by small amounts at each step) to make the computed pressure approach the target pressure. Many of these algorithms are counterparts to thermostats described in this section. Of those, a common pressure control method is the Berendsen barostat (*160*), where the system is weakly coupled to a pressure bath and the pressure is controlled by dynamically adjusting the size of the unit cell and rescaling the coordinates of non-fixed atoms during the simulation. In the extended pressure-coupling system methods, first introduced by Andersen (*162*), an extra degree of freedom, corresponding to the volume of the box, is added to the system coupled to a fictitious pressure bath. The volume can vary during the simulation, with the average volume being determined by the balance between the internal pressure of the system and the desired external pressure.

3.3.3 Spatial boundary conditions

The correct treatment of boundaries and boundary effects is crucial to simulation methods because it enables 'macroscopic' properties to be calculated from simulations using relatively small numbers of particles. Dependent on the number of particles in the study, one should consider suitable boundaries that would allow an appropriate derivation of the bulk properties.

Periodic boundary conditions (PBC) create the bulk condition by replicating the cubic box that contains the protein and solvent through space. Through this replication any unwanted surface effects are eliminated and should a molecule leave the box during simulation then it is replaced by an identical molecule that enters from the opposite side. Only the coordinates of the central box are stored and used, since all the boxes are identical.

3.3.4 Long-range interactions

The most time-consuming part of an MD simulation is the calculation of the non-bonded forces. The long-range interactions in the system include the van der Waals interaction and the electrostatics. The electrostatic term is harder to calculate as it does not drop of

rapidly over distance, unlike the van der Waals attractions. Additionally, long-range electrostatics are often of interest in simulations. A basic method to deal with the calculation is to apply a cut-off, but this method introduces significant nonphysical effects especially at and also beyond the cutoff distance (*168*), and simply increasing the cut-off distance can dramatically raise the computational cost.

An attractive alternative to the use of a cut-off distance are a variety of methods that include the Ewald summation (*169*), the reaction field method (*170, 171*) and cell multipole method (*172*).

In our studies the long-range interactions are handled using particle-mesh Ewald (PME) (*173, 174*) which is a variation of the standard Ewald sum method (*169*). In comparison to the Ewald sum method where the potential due to the partial charges of a system, together with all of their periodic images are incorporated by splitting the Coulomb interactions into a short-range term and a long-range smoothly varying term that are handled by a direct sum and reciprocal sum by Fourier method (*175*) respectively; in PME the reciprocal space Ewald sums are B-spline interpolated on a grid and the convolutions necessary to evaluate the sums are calculated via fast Fourier transformations, which accelerates the solution (*114*).

3.3.5 Solvent models

Solvent, usually water, has a fundamental influence on the structure, dynamics and thermodynamics of biological molecules, both locally and globally. One of the most important effects of the solvent is the screening of electrostatic interactions. Solvent effects can be incorporated in the system at varying levels.

Implicit and explicit solvent models are the two generalised models for incorporating solvent influence within the system. The main difference between the two models is that implicit models employ a homogenous medium to represent the solvent where as the explicit model uses atomistically represented molecules (*176*). The explicit model is more physically realistic, but does however have practical limitations for very large systems, in terms of computational cost, as most of the computational time is spent on the simulation of the solvent surrounding the system.

Explicit solvent models rely on using hundreds or thousands of discrete solvent molecules (*177*). They are the most widely used methods for carrying out simulations in solvent. Such calculations converge only slowly because of the large number of particles involved. They generally require orders of magnitude more CPU time than corresponding gas phase calculations on the same molecule. Because explicit models are so computationally demanding, there is a significant interest in developing more rapid implicit solvent models.

Implicit models treat the solvent as a continuous medium having the average properties of the real solvent, and surrounding the solute beginning at the van der Waals surface. A variety of continuum models have been described including the generalized Born (GB), Surface Area (SA) model (*178*), where the total solvation free energy is given as the sum of a solvent-solvent cavity term, a solute-solvent van der Waals term and a solute-solvent electrostatic polarisation term.

Amongst the large number of water models for biomolecular simulation, including the rigid, flexible and polarisable models, the most widely used models are the rigid type. TIP (*179*) (Transferable Interaction Potential) models and SPC (*180*) (Simple Point Charge) models.

MD simulations within this study have used the explicit solvent model for water, TIP3P (*179*).

3.4 Analysis methods

In the study of the protein systems presented in this thesis, a number of analysis tools have been used. The main methods of analysis include principal component analysis (PCA), Root Mean Squared Fluctuations (RMSF), Root Mean Squared Deviations (RMSD) and atom distance measurements.

3.4.1 Principal component analysis

Once a suitably long simulation has been run, principal component analysis (PCA) can be applied to the simulation, including the regions or atoms of interest in the analysis. The analysis starts by removing the centre of mass rotation and translations, as these are irrelevant to the internal motions. This is achieved by least-square fitting each frame of the trajectory to a reference or average structure. This results in a Cartesian molecular coordinate system for which the atomic motions can be obtained. The correlation between atomic motions can then be determined through a covariance matrix C_{ij} of the positional fluctuations, for coordinates i and j:

$$C_{ij} = \langle M_{ii}^{\frac{1}{2}}(x_i - \langle x_i \rangle) M_{jj}^{\frac{1}{2}}(x_j - \langle x_j \rangle) \rangle \tag{3. 21}$$

Where $\langle\,\rangle$ denotes the average over all instantaneous structure sampled over the simulation. M is a diagonal matrix containing the masses of the atoms (mass-weighted analysis) or the unit matrix (non-mass weighted analysis). C is a symmetric 3N x 3N matrix, where N refers to the number of atoms, and can be diagonalised with an orthonormal transformation matrix R, to obtain principal components (PCs) and corresponding eigenvalues:

$$\boldsymbol{R^T C R} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{3N}) \tag{3. 22}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{3N}$ are the eigenvalues. The columns of R are the eigenvectors, or PCs of the system, where R defines a transformation to a new coordinate system. Therefore, the trajectory can be projected onto the eigenvectors to give the PCs $q_i(t)$, i= 1, ..., 3N:

$$\mathbf{q}(\mathrm{t}) = \boldsymbol{R}^{\boldsymbol{T}} M^{\frac{1}{2}} (\mathbf{x}(\mathrm{t}) - \langle \mathbf{x} \rangle)$$

(3. 23)

As the first few PCs contain the largest mean square fluctuations, typically the first few PCs describe the collective, global motions of the system. To visualise motions sampled along one (or more) principal components, the trajectory can be filtered along those PCs (*181*). For a PC *i* this can be obtained by:

$$\mathbf{x}^{\mathbf{f}}(t) = \langle \mathbf{x} \rangle + M^{-\frac{1}{2}} R_{*i} q_i(t)$$

(3. 24)

where R_{*i} denotes the i^{th} eigenvector of the matrix C_{ij} (the i^{th} column of $\boldsymbol{R}$).

The root-mean-square fluctuation of each principal mode corresponds to the square root of the corresponding eigenvalue, λ_i . Large eigenvalues correspond to large fluctuations and low frequency correlated motions, thought to be important for enzymatic catalysis. The number of eigenvectors to study is an arbitrary decision depending on the motions of interest in the system, but some criteria do exist, one being the Kaiser criterion (*182*). This procedure only retains eigenvectors which have eigenvalues greater than 1. The rationale behind this criterion is that interpretation of proportions of variance (here, positional fluctuations), smaller than the variance contribution of a single variable, are of dubious value. An alternative method is the Cattell Scree test (*183*), which uses simple line plot representations of the eigenvalues displaying the relative importance of each component in fitting the data to major conformational reorganisation. Cattell suggests that the point at which the line smoothes out is selected as the cut-off for the eigenvector choice, disregarding principal components to the right of this point. Generally, it has been found that the former method retains too many factors and the latter, too few, but both are found to be acceptable for use in application to large global conformational

changes (*184*). However, these are only guidelines and in practice, the extent to which a solution is interpretable, giving sensible results, is additionally taken into account.

The similarity between the eigenvectors can be quantified by calculation of the inner-products (*185-188*). The averaged conformation of each of the two simulations for which the eigenvectors are to be compared are taken and brought into best-fit positions, reorientating the eigenvectors accordingly. The inner-product is then calculated as in equation 3.25, where *i* and *j* represent the eigenvectors to be compared of vector sets **w** and **u** respectively.

$$\mathrm{I}_{ij} = \mathbf{w}_i . \mathbf{u}_j \tag{3. 25}$$

If the two eigenvectors are identical, the inner-product must be unity due to normality. A value of 0 indicates no similarity between the eigenvectors. The sum of the squared inner-products of two sets of N vectors can be compared by calculation of the root mean squared inner-product (RMSIP), as shown in equation 3.26.

$$RMSIP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathbf{w}_i . \mathbf{u}_j)^2} \tag{3. 26}$$

where $\mathbf{w}_i$ represents the i^{th} eigenvector associated with data set **w** and $\mathbf{u}_j$ represents the j^{th} eigenvector associated with data set **u**. This value can be used to evaluate the overlap of the conformational space spanned by the different systems.

In this project, PCA has been performed using the "g_covar" and "g_anaeig" modules of GROMACS 3.2.1 (*181*).

3.4.2 Root-mean-squared displacement

Root-mean-squared displacement, commonly referred to as RMSD, is a measure of the average distance between selected atoms and a reference structure, where δ refers to the distance between N atoms (usually Cα, or other atoms of the backbone) (equation 3.27).

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \qquad (3.\ 27)$$

When fitting structures, the aim is to reduce the RMSD value to a minimum, by finding relative orientations of the molecules that achieve this (*97*).

3.4.3 Root-mean-squared fluctuation

Root-mean-squared fluctuation (RMSF) is a measure of the deviation of the position of the atoms from the average structure from the MD trajectory (equation 3.28).

$$\mathrm{RMSF} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_i - \bar{x})^2} \qquad (3.\ 28)$$

where x_i and $\bar{x}$ denote the instantaneous and time-averaged coordinates of residue *i* (typically Cα), respectively. *T* refers to the number of trajectory frames. RMSF gives a measure of the fluctuations about a time-averaged structure. It is useful in highlighting regions of high mobility in an MD simulation, in the case of high RMSF values, and restricted mobility in the case of a low values (*189*).

3.5 Protein-ligand binding site identification

The ability to identify binding sites on a protein surface that could bind to small, drug-like compounds is a very important and useful concept in drug design (*74*). Of particular interest are the allosteric binding sites which offer advantages in terms of selectivity, eradicate competition with natural substrate at the active site, and allow for the possibility of new chemo-types, moving away from drug ligands designed on the basis of natural substrate scaffolds.

Many binding site detection methods have been mentioned in the previous chapter; here the two methods Q-SiteFinder and Pocket-Finder used in this study are discussed in more detail.

The two methods were chosen for our studies as they proved to be reliable for predicting the known binding sites observed through crystallography in the X-ray structure of the closed-state GLK studied here, providing useful information such as the grid coordinates and the name and number of the interacting residues. In addition, the tools provided a relatively easy interface to post-process. However, the tool interface does not provide the possibility for multiple submissions. This is generally the case for most binding site search methods, where methods have not been optimised for batch mode analysis. Thus, for both methods used here, a server interface was used to analyse snapshots from the MD trajectories. This has therefore been a laborious task, where perl scripts had to be written to analyse the outputs of the large scale search in this study. As both methods use a similar interface, for ease of analysis, as well as using Q-SiteFinder, the geometric method (Pocket-Finder) was also utilised, mainly as a means of comparison and to ensure that the results were not based only on one method.

3.5.1 Pocket-Finder

Pocket-Finder is a geometric-based method, based, like others, on the observation that the binding site is commonly found in the largest pocket on the protein surface. This is likely to be the case for the active site, but for second sites, like the allosteric binding site in GLK, this would not necessarily be true. Nevertheless, alongside the energy-based method Q-SiteFinder, Pocket-Finder was utilised for comparison.

The algorithm (Pocket-Finder) as described by the authors (*95*), is an implementation of LIGSITE (*89*) which is based on the POCKET algorithm (*190*). In POCKET, a probe sphere of radius 3 Å is passed across the protein along the 3D Cartesian grid directions. A point on the grid is defined as an interaction between the protein and probe sphere if the centre of the protein atom is found to be within the probe sphere. A pocket is defined if an interaction occurs followed by a period of no interaction, followed by another interaction, referred to as the protein-site-protein (PSP) event. LIGSITE improves on POCKET by including four cubic diagonal search directions in addition to the existing three; thus seven directions are scanned (figure 3.2), which makes the method less dependent on the orientation of the protein in the 3D grid. Similar to LIGSITE, Pocket-Finder establishes the buriedness of each grid point in the protein. At the start all the grid points are set to zero. When a grid point is identified as a pocket in the PSP event, the grid point is incremented. Every grid point can eventually take values from zero, suggesting that the grid point is not part of any pocket, to a value of seven (as there are 7 search directions), which suggests a deeply buried pocket. Pockets are defined by cubes of retained grid points. The grid points themselves are only retained if they are within the defined threshold of number PSP events, which here is set to a minimum of 5 times. In Pocket-Finder a grid resolution of 0.9 Å and a probe radius of 1.6 Å is used. A pocket is a collection of probes in the site, and it is ranked depending on the number of probes identified in each site. Ten site are reported by the server, ranked in the order of probe numbers, with the largest on top.

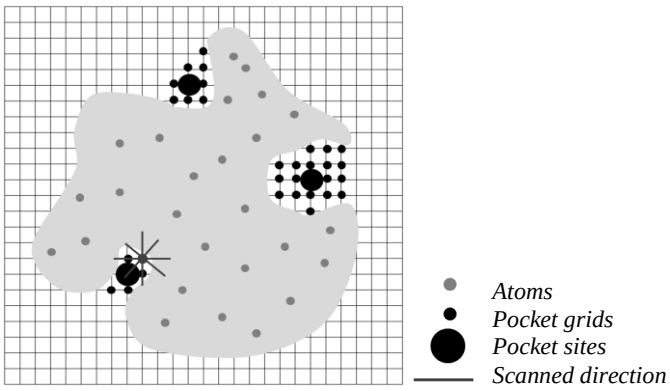


Figure 3.2: Pocket-Finder scans the grid for protein-solvent-protein events, in 7 directions. Figure adapted from reference (*191*).

3.5.2 Q-SiteFinder

In contrast to the geometric approach in Pocket-Finder, the pocket in Q-SiteFinder is defined only by energetic criteria. The method calculates the van der Waals interaction energies of a methyl probe with the protein. Initially, the coordinates are rotated about the geometric centre to minimise the volume of the box enclosing the protein. This reduces the number of grid points requiring analysis. The program Liggrid calculates the non-bonded interaction energy of the methyl probe with the protein at each grid position, using the GRID force-field parameters (*192*) to estimate the interaction energies. A grid resolution of 0.9 Å on a 3D grid enclosing the whole protein is implemented. The probes with the most favourable binding energy are retained based on an interaction energy threshold. The protein coordinate is then rotated back to the original orientation. Probes are clustered according to their spatial proximity, and the total interaction energies of probes within each cluster are calculated. Probe clustering uses a variable known as the connection range, which determines the maximum distance between two probes that can be connected as part of the same cluster. This value should be greater than the probe grid resolution used to generate the probe output file. A value of 1.0 Å for a grid resolution of 0.9 Å is used in this method. This connects all adjacent sites but not those on the diagonal of the cube. The probe clusters are ranked according to their total interaction energies, with the most favourable being identified as the first predicted site.

The clustering program also calculates sites volume, and can identify which protein atoms are within a defined range of cluster sites.

3.5.3 Comparison of the two methods

The same pre-processing of the protein is carried out by both methods in defining protein atoms, removal of solvent and addition of hydrogen. The protein in both methods is then rotated about the geometric centre to minimise the volume of the box enclosing the protein. This reduces the number of grid points requiring analysis.

In comparison with Pocket-Finder, Q-SiteFinder can predict binding sites with much higher precision, mapping ligands at binding sites with a tighter map. Pocket-Finder on the other hand predicts large sites that are much larger than a known ligand at the binding site. Although this precision in Q-SiteFinder may be useful for structure-based drug design, when in search of unknown binding site, this information may be too discriminating. Primarily a guide to the location of the binding site is useful, and then more accurate methods can be used to precisely map the binding site. Since binding sites can be flexible and subject to induced conformational changes upon ligand binding, a tight prediction that would map a ligand that is complexed with the protein in an x-ray structure, is not necessarily the information we are looking for when scanning simulation trajectories for potential binding sites. However, Q-SiteFinder does offer the advantage of an energy-based method which should be more informative in comparison with a method that is purely geometrical, where sites may be predicted but would not be suitable for binding at all.

3.5.4 The interface and analysis of the output

Both tools, Q-SiteFinder and Pocket-Finder, can only be used through a server interface, which makes the task of post-processing challenging. Recently the interface to Q-SiteFinder was updated, but the output format is similar. The interfaces for both tools are depicted in figure 3.3.

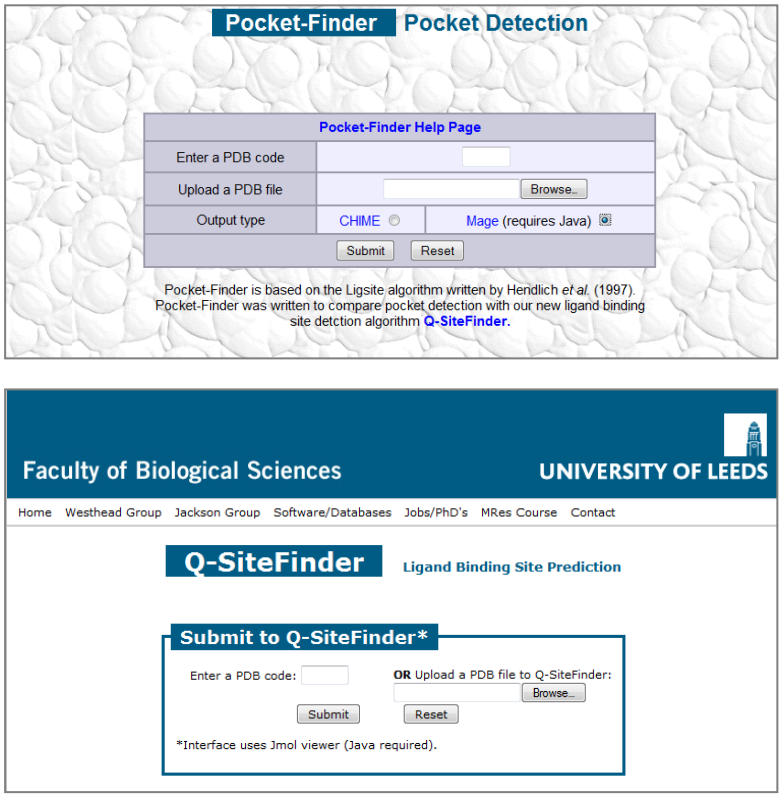


Figure 3.3: Pocket-Finder (*193*) and Q-SiteFinder (*194*) interfaces on top and bottom, respectively. In both cases, the input can be provided in the form of a PDB code or a file containing protein coordinates.

In both tools either a PDB identity code or a file containing protein (and ligand) coordinates can be provided. The output for both methods is presented as a HTML page providing a 3D visualisation window which allows full rotation of the structure (figure 3.4). The binding sites can be toggled on/off. In addition, information regarding interacting residues for each predicted site, and the volume of the predicted binding site, is provided.

A .pdb format file can be downloaded which contains the original protein coordinates, and the coordinates of the grid points at each binding site, in order of rank, separated by a 'TER' card. In addition, the coordinates on the interacting residues are also included. However, despite including atom names, the residue numbers and names have not been provided. In our study where the identification of the allosteric binding site is dependent

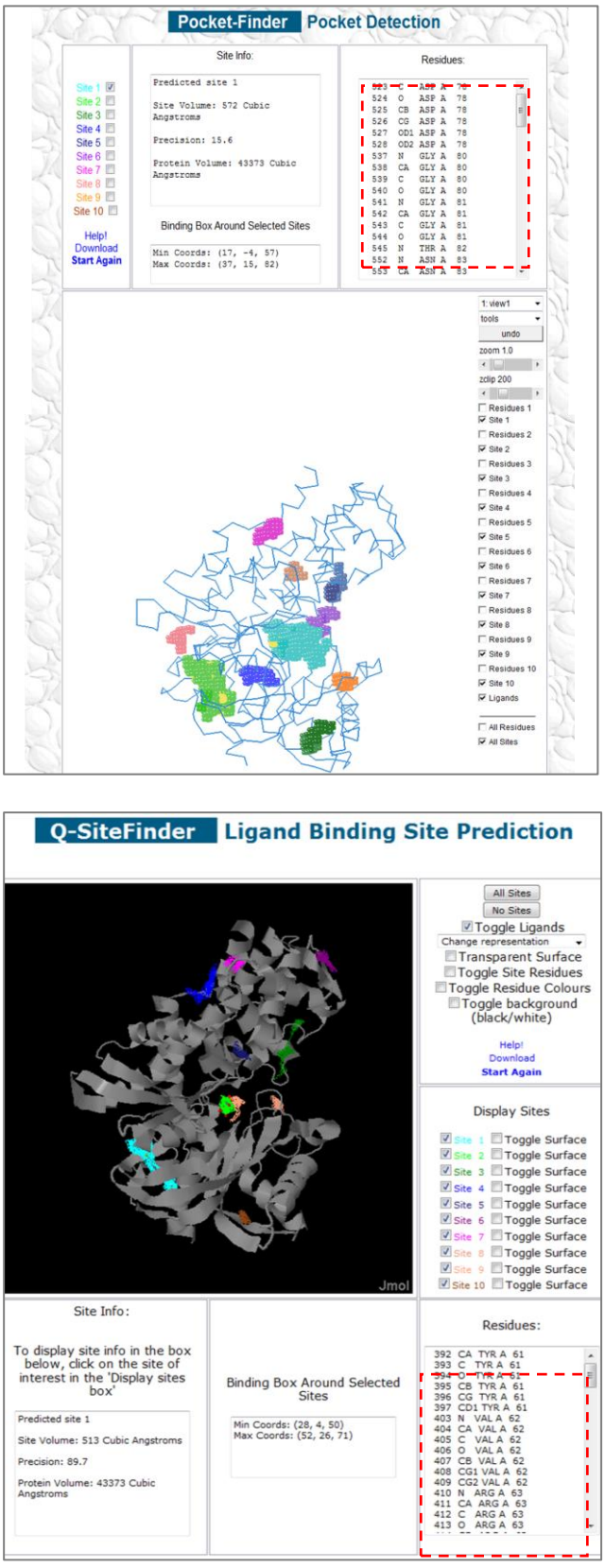


Figure 3.4: Pocket-Finder and Q-SiteFinder interfaces on top and bottom, respectively. Boxed in red are the windows that list the interacting residues for each site. This information was used in searching the HTML source pages, when analysing the outputs.

on knowledge of the residues in the vicinity of the binding pocket, this lead to more tedious analysis of the source file for the HTML page.

To handle the analysis of large data sets of all the source files obtained for frames from MD trajectories, I developed a perl script to search for relevant contact residues name/number as texts. This information was provided in a window on the output HTML page (figure 3.4), and was available in full in the source page. A counter was used in the script to establish how many of the known allosteric binding site residues were present in a particular predicted binding site. If more than 3 were recorded, this site was further analysed in plots that will be presented in the results chapter.

3.6 Summary

In this chapter the methods that have been utilised in this thesis have been outlined. Normal mode analysis and molecular dynamics will be used to capture protein dynamics and flexibility.

Additional analysis methods such as RMSD and RMSF give an indication of residue flexibilities. Principal component analysis will be applied to simulation trajectories to establish the likelihood of revealing the allosteric binding site along the principal modes.

Two binding site detection tool Q-SiteFinder and Pocket-Finder have been chosen for the binding site search. Q-SiteFinder is an energy-based method which offers more realistic prediction than a purely geometric method. For comparison Pocket-Finder is also utilised. Both tools are free software that have been implemented in form of an online server which can give useful information regarding the location of the binding sites, grid coordinates, interacting residue coordinates and site volumes. The methods have not be implemented for batch mode submission; therefore every simulation snapshot will have to be submitted individually and the results are post-processed with in-house perl scripts.

Chapter 4

Human glucokinase (GLK): Background

Human glucokinase (GLK), also known as hexokinase D or hexokinase IV, is a member of the hexokinase family (Hexokinase I-IV) that catalyse the ATP-dependent phosphorylation of glucose to glucose 6-phosphate, the rate limiting step in glycolysis (*195*).

GLK is highly expressed in the pancreatic β-cells, hepatocytes and brain. The mechanism of GLK in the brain is less well understood (*196*); however, in the pancreatic β-cells it plays a key role as a glucose sensor by integrating blood glucose levels and glucose metabolism with insulin secretion. In the liver GLK regulates glucose uptake and glycogen synthesis (*197*).

GLK plays a critical role in glucose homeostasis as evidenced by naturally occurring mutations that cause lasting glycemic disorders in humans (*195*). Inactivating mutations in the GLK gene have been linked to a form of diabetes called type 2 maturity onset diabetes of the young (MODY2) (*3, 198*). On the other hand, activating mutations, enhancing the enzymes catalytic activity by lowering the threshold for glucose stimulated insulin release cause persistent hyperinsulinemic hypoglycaemia of infancy (PHHI) (*4, 199*). These diseases demonstrate the extent to which GLK action has to be regulated and emphasise the potential therapeutic value of modulating the activity of this key metabolic enzyme.

In 2003 the first allosteric activator of GLK (GKA) was reported by Roche scientists (*200*). Since then, many more GKAs have been reported and patented by other pharmaceutical companies (Eli- Lilly, OSI, AstraZeneca, Pfizer and Banyu) (*201*). The discovery of the first activators have offered promising new therapeutic approach for the treatment of diabetes since the activators can augment both hepatic glucose metabolism and glucose-induced insulin secretion (*202*).

4.1 Glucokinase role in glucose homeostasis

Control of blood glucose levels is one of the body's critical homeostatic mechanisms (*203*). Circulating levels of glucose are controlled by two enzymes, insulin and glucagon, the main effectors in the homeostatic feedback loop. The liver and pancreas on the other hand play a central role in the control of blood glucose by detecting changes in the glucose concentration.

If the levels are too high, β-cells in the islets respond by releasing insulin. If the levels are too low, α-cells in the islets secrete glucagon (*203*). At high glucose levels, the liver takes glucose and replenishes depleted glycogen stores, and then synthesises fatty acids. During starvation and diabetes the liver releases glucose to blood from glycogenolysis and gluconeogenesis (*204*).

The role of GLK in the glucose homeostasis is depicted in figure 4.1. For both, the pancreatic β-cell and the liver hepatocyte, glucose is transported into the cell by a high capacity, low affinity glucose transporter, GLUT2. In these cells glucose is predominantly phosphorylated by GLK yielding glucose-6-phosphate at the expense of ATP (*205*). Hexokinase I-III are also expressed in hepatocytes, however at much lower levels and only account for a small fraction of glucose-phosphorylation activity, but are crucial as a back-up mechanism in the event of glucokinase activity malfunction.

In the liver, once transported into the cells by GLUT2, glucose phosphorylation is catalysed by the L-type GLK (GLK_L) isoform, and then stored as glycogen by insulin stimulation (figure 4.1). In the adipose and muscle cells, insulin stimulates glucose uptake and metabolism by triggering the translocation of the glucose transporter isoform GLUT4 from intracellular storage vesicles to the plasma membrane. Alteration in the

rate of liver glucose production and its implication on insulin resistance (*206*), as well as impaired pancreatic islet function in older animal models (*207*), which is a reflection of alterations in glucose transport, phosphorylation and utilisation, indicate the key role of liver in the control of whole-body glucose homeostasis.

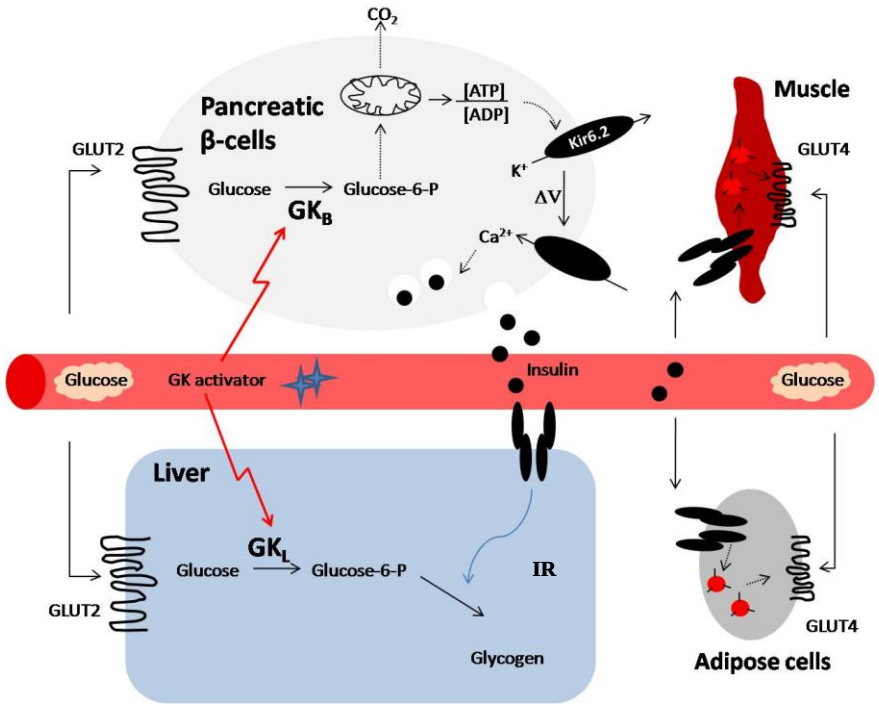


Figure 4.1: Central role of glucokinase in whole-body glucose hoemostatis. (Adapted from reference (*202*)). In both the pancreatic β-cell and the hepatocyte, glucose is transported into the cell by GLUT2. In the pancreatic β-cells, B-type GLK (GK_B) constitutes part of the glucose sensor. Glucose is phosphorylated by GLK yielding glucose-6-phosphate. Glycolysis and oxidative mechanism of glucose increases the ATP:ADP ratio, leading to inactivation of the Kir6.2 potassium channel and to subsequent depolarisation of the membrane, causing the release of the insulin hormone into the blood stream. In the liver, glucose phosphorylation is catalysed by L-type GLK (GK_L), then stored as glycogen. In the adipose and muscle cells, insulin stimulates glucose uptake and metabolism by triggering the translocation of the glucose transporter isoform GLUT4 from intracellular storage vesicles to the plasma membrane. The GLK activators augment both glucose-induced insulin secretion in β-cells and hepatic glucose metabolism, and as a result lead to improved clearance of glucose from the blood stream.

In comparison with the other hexokinase isoforms, the specific function of GLK is based on the particular kinetic properties of this enzyme, which include a low affinity for

glucose, a cooperativity with glucose and a lack of end-product inhibition at physiological concentrations of G6P (glucose-6-phosphate) (*208*). In addition, only in the liver GLK activity is regulated through protein-protein interaction by the GLK regulatory protein (GKRP), which acts as a competitive inhibitor with respect to glucose and also regulates the nucleocytoplasmic localisation of the enzyme (*209*).

When glucose concentration is low, GLK remains bound to GKRP in the nucleus, where it is unavailable to phosphorylate glucose. An increased glucose concentration after feeding leads to the dissociation of GLK from GKRP and translocation to the cytoplasm, where it phosphorylates glucose to G6P, starting it on its way toward glycogen synthesis (*210*).

In the liver, this association between GLK and GKRP is ligand-dependent. Ligands that bind GKRP are fructose-6-phosphate (F6P) and fructose-1-phosphate (F1P), binding to the same site (*211*). F6P favours GKRP-GLK complex whereas F1P weakens this interaction and leads to the release of GLK (*212*). Site directed mutagenesis experiments indicate that the binding interface for GKRP lies close to the binding site of allosteric GKAs (*213*).

4.2 Structure of human glucokinase

Structurally glucokinase is a 448 residue monomeric enzyme folded into two domains, comprising a large and a small domain connected by a hinge made up of three flexible loops. In the active form the domains are separated by a deep cleft which forms the active site, binding glucose and ATP for phosphorylation, and an allosteric site that is at the interface between the two domains, about 20 Å away from the active site, surrounded by a flexible connecting region.

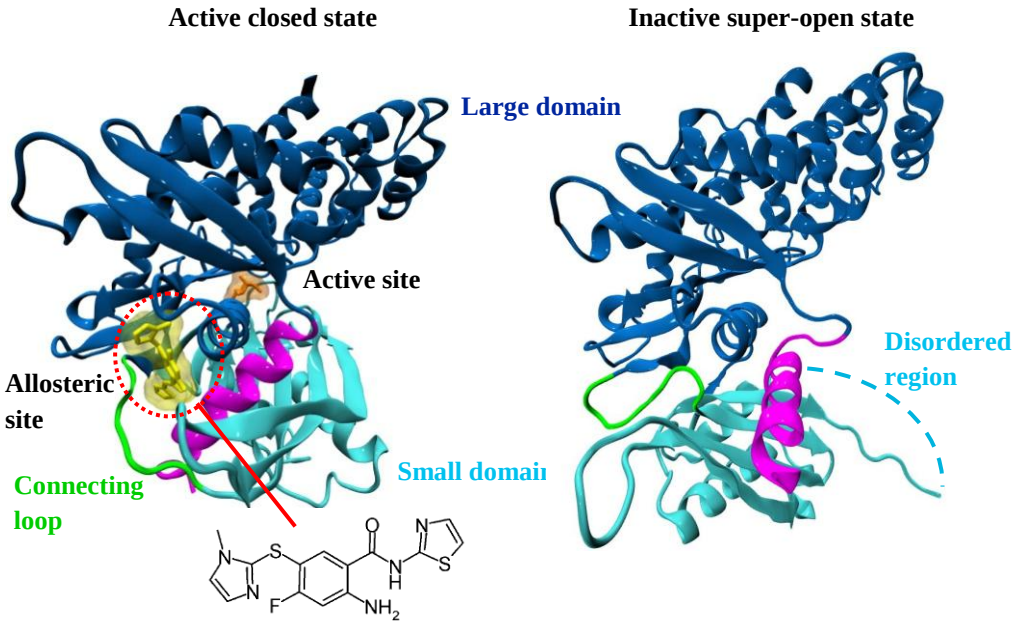


Figure 4.2: On the left human glucokinase active closed state (PDB ID: 1v4s), in complex with glucose (orange) and allosteric activator ligand (yellow); the structure of compound A, [2-Amino-4-Fluoro-5-[(1-Methyl-1h-Imidazol-2-yl)Sulfanyl]-N-(1,3-Thiazol-2-yl-Benzamide] has been illustrated below. On the right is the super-open inactive state (PDB ID: 1v4t). The crystal structure has missing residues in the disordered region (Resides 157-179). Large conformational changes occur in the small domain on transition from the closed state to the super-open, releasing the terminal α-helix, and rotating it by ~90°. The active and allosteric binding sites collapse and the structure becomes inactive.

Alternative splicing of the glucokinase gene (GCK) results in three tissue-specific forms of glucokinase. Isoform 1 is specifically expressed in pancreatic islet beta cells (B-type GLK) (*214*), which has a distinct N-terminus; the remainder of the protein is identical to isoforms 2 and 3. Isoform 2 and 3 are the major and minor isoforms expressed in liver (L-type GLK), respectively, with a distinct N-terminus; the remainder of the protein is identical to isoform 1.

Since at the time of the study only two X-ray structures of human glucokinase were publicly available, representing two distinct conformations of the enzyme (*5*) (Figure 4.2). Since, more structures have been added to the protein databank in complex with allosteric activators. Under different crystallisation conditions with short truncations at the NH2-terminal end (11 or 15 amino acids) of the hepatic isoform, two X-ray structures were obtained. The active form bound to glucose and the allosteric activator (compound A, here) is referred to as the closed state, with two clefts, one for the active site, binding glucose and the other for the allosteric activator, compound A. The flexible

loop connecting the two domains is the feature that structurally differentiates GLK from the other hexokinases in the family. The additional flexibility in this region offers the space required for the allosteric activator.

In the inactive form, the conformation is very different, referred to as the super-open form. In other hexokinases, the inactive state, of the enzyme adopts a conformation which is referred to as the open with respect to the distance between the two domains, but in GLK, the domain open further apart. In the super-open, the orientation and conformation of the small domain is significantly different to the closed state. The main part of the small domain rotates about 99° (Figure 4.2). In this state, due to the significant structural rearrangements and disorder of residues 157-179, the glucose binding site becomes exposed to solvent. This state is regarded as inactive as the necessary residues for the enzymatic activity are absent from the active site region, as the small domain undergoes large conformational change. The allosteric binding site also completely disappears in this state.

Intermediate conformation(s) have been speculated, suggesting an open conformation, similar to that of hexokinse I in the absence of glucose (*215*). Targeted molecular dynamic studies between the super-open and closed conformation, indicate several low energy intermediate states (*216*). Kinetic studies and tryptophan fluorescence (*217*) further support this theory by demonstrating the existence of four distinct conformations of the enzyme. It has been hypothesised that glucose and the allosteric activator could only bind to the active closed form of the enzyme as both binding sites appear absent from the super-open inactive X-ray structure. However, recent binding kinetics studies of glucose and the allosteric activator (GKA) to GLK, reconfirmed multiple conformational states and further demonstrated that GLK can bind GKA even in the absence of glucose. They suggest that in the absence of any ligand GLK may be able to sample several conformational states between the super-open and closed conformations (*218*). Recent kinetics finding will be discussed in the following section.

The focus of this thesis is the identification of novel allosteric binding sites. For GLK an apo structure would be required. In the super-open both binding sites are absent from the X-ray structure (PDB ID: 1v4t). Even though there is a body of kinetic evidence suggesting that the allosteric ligand could bind GLK even in the absence of glucose, this

is likely to be a rather rare event in the conformational equilibrium. In addition, this structure was resolved at a very low crystallographic resolution of 3.5 Å while missing residues 157-179, of which some form part of the active site residues in the closed state. For this reason, this structure could not be the sole apo form used in our studies.

In the absence of a crystal structure of GLK only bound to glucose in the public domain, an X-ray structure of GLK (hepatic isoform 3) bound to glucose was crystallised by colleagues at AstraZeneca (unpublished data, referred to as the GLK_AZ structure hereafter) (figure 4.3). The structure was resolved to 1.90 Å. A number of mutations were carried out to facilitate crystallisation, at locations where the impact on the secondary structure construct should be negligible. The mutant X-ray structure has a 97% sequence identity with the X-ray structure bound to glucose and the allosteric activator (PDB ID 1v4s (See appendix A). The following residues have been mutated in comparison with the 1v4s sequence; M11S, A12S, L13N, T14S, L15Q, E27A, E28A, E51A, E52A, E94A, E95A, E96A. At the N-terminus residues M11, A12, L13 have been resolved in addition to 1v4s sequence, but this structure is 5 residues shorter than 1v4s at the C-terminus.

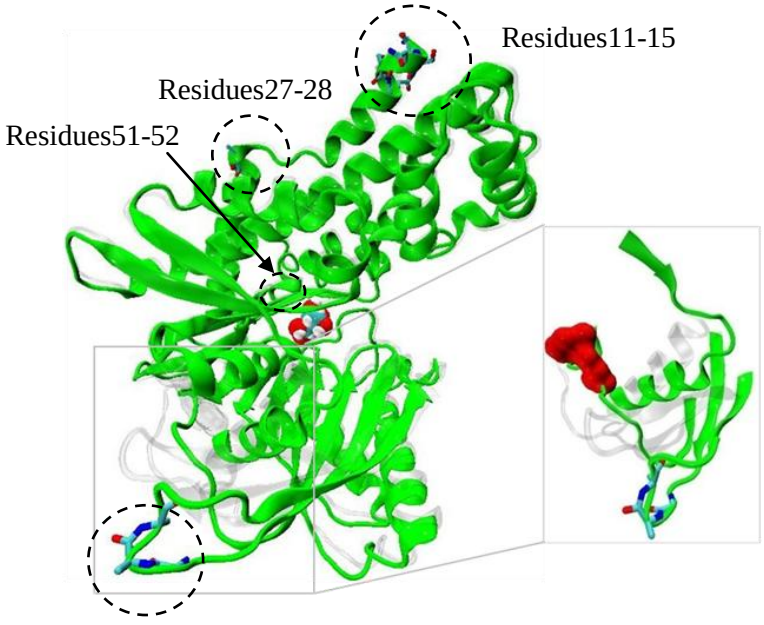


Figure 4.3: The X-ray structure of GLK bound to glucose only (GLK _AZ). In green the AstraZeneca structure, with glucose at the active site. Mutation residues are in stick representation, circled with dotted black line. In grey, the structure of 1v4s has been overlaid to highlight the differences. The main difference is in the allosteric region, zoomed in on the right. In this image, in red the location of the allosteric ligand has been highlighted in 1v4s. In GLK_AZ the connecting loop, crosses through the region occupied by the allosteric ligand in 1v4s.

Of those mutations, residues 94-96 in particular contribute to the stability of this X-ray structure in the absence of the allosteric activator in GLK_AZ. This β-turn is very flexible with high B-factor values. This flexibility allows allosteric ligands of various sizes to bind in the allosteric binding pocket. Below in figure 4.4, an overlay of various GLK structures that have been co-crystallised with allosteric activators by AstraZeneca has been depicted to demonstrate this flexibility. Owing to confidentiality reasons the specific ligands have been removed from this image. Original compound A, allosteric activator from the wild-type structure (PDB ID 1v4s) has been shown to highlight the allosteric site.

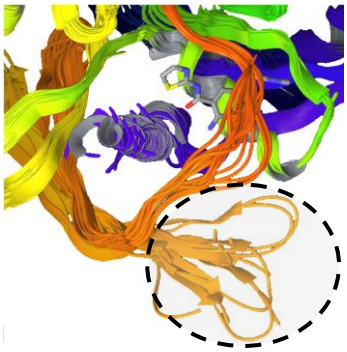


Figure 4.4: Various allosteric bound AstraZeneca, GLK structures (image obtained from Richard Ward, AstraZeneca) (*219*). The specific ligands have been removed due to confidentiality reasons. The ligand displayed is the allosteric ligand (compound A) from the original structure (PDB ID 1v4s), to indicate the vicinity of the allosteric binding site. The highly flexible β-turn has been highlighted, which moves to accommodate larger ligands in the allosteric pocket.

None of the above mutations have been linked to naturally occurring inactivating or activating mutations that could lead to MODY or PHHI respectively (*220*) (see appendix D for the location of MODY and PHHI mutations). This structure (GLK_AZ), has been used in this study to represent an active closed state conformation in which the allosteric site is not fully accessible. Binding pocket search in this structure does not show a binding site in the allosteric region, and therefore this would allow a suitable starting point to search for the allosteric binding site (discussed in chapter 6). To eliminate the possibility that results could be biased by any of the crystallography mutations, a homology model of the original structure was built, converting the sequence to that of 1v4s, using the secondary structure fold

in the GLK_ AZ structure. The possibility that mutations in the β-turn, residues 94-96 may be activating the enzyme therefore, aiding the crystallisation process by stabilising the active form cannot be ruled out, as crystallising GLK in only the presence of glucose has proved challenging. In this study as well as the original X-ray structure, bound to both glucose and the allosteric activator, the super-open state has been studied. Additionally the GLK_AZ structure has been studied, with both the mutant sequence and the sequence homologous to 1v4s.

4.3 Human glucokinase kinetics

GLK has a unique kinetic property relative to other hexokinases in the family, which enables this enzyme to play a key role as a glucose sensor. It has a low affinity for glucose (*195*) which is significantly lower than that of other hexokinases. GLK displays a sigmoidal reaction rate, rather than the Michaelis-Menten hyperbolic kinetics (*221*), so its activity increases rapidly with blood glucose concentration over normal physiological range (*222, 223*), which allows GLK to operate as an ultrasensitive physiological glucose sensor (*196*).

In addition, GLK is insensitive to inhibition by the physiological concentrations of the reaction product, glucose-6-phosphate (G6P). Owing to the unique kinetic properties of GLK the higher the blood glucose concentration, the faster the liver converts glucose to G6P. Hepatic cells, unlike most cells, contain large quantities of the insulin-independent glucose transporter GLUT2 and are therefore freely permeable to glucose. Thus at low blood glucose concentration, the liver does not compete with other tissues for the available glucose supply, whereas at high blood glucose concentration, when the glucose needs of these tissues are met, the liver converts the excess glucose to glycogen, first by converting glucose to G6P, with the aid of GLK and then to glycogen, a process stimulated by insulin (*210*).

GLK is a monomeric enzyme (*224*), for which the positive cooperativity for glucose cannot be explained by the classical mechanisms of allostery for multi-subunit protein where substrate binding to one enzymatic subunit, stimulates the rest of the

subunits to adopt active conformations, suitable for binding the substrate (*225*). As GLK only has one active site, positive cooperativity has been explained with the aid of kinetic models.

In the past, two kinetic models have been postulated for GLK to describe this positive cooperativity, both requiring a slow equilibrium between two conformations of GLK with their relative affinities for glucose. The "mnemonical" model was first suggested in the late 1970s by Ricard et al. (*226-229*). In this model, the enzyme exists in two distinct conformations, with a slow transition between the two. This model suggested that the observed cooperativity was due to different glucose binding affinity to the two forms of the free enzyme not in equilibrium under steady state conditions. In the general slow transition mechanism, only one form of the enzyme, the free enzyme, may assume a second conformation, i.e. there is only one active enzyme-substrate complex or catalytic cycle. However, this model does not satisfy all the main features of glucokinase kinetics. The second model, "slow-transition" (*225*), involves more steps, and gives better account of glucokinase kinetics, with both conformations capable of entering the catalytic cycle, but still with different kinetic parameters for glucose, with the steady state rate, a sum of the rates of the two cycles.

In both models the equilibrium distribution between the two conformations, the super-open and closed, is dependent upon the glucose concentration, and the rate of conformational change is slower than the rate of catalysis under steady-state conditions. The inability of the two conformations to come to equilibrium during steady-state catalysis yields the cooperative kinetics (*230*). Later, Neet et al. (*231*) suggested a slow change in the distribution of the active species of GLK dependent on glucose concentration, leading to at least two slowly inter-convertible, kinetically distinct forms of GLK, which was further supported by fluorescence spectroscopy (*232*). X-ray crystallography confirmed the two states providing support for mneumonical model; a super-open conformation in the absence of glucose and a closed state bound to glucose and an allosteric activator (*5*). The super-open conformation represents the low glucose affinity, catalytically inactive form of the

enzyme. The closed conformation corresponding to the high affinity form of the GLK with glucose and ATP bound at the active site during catalysis.

Very recently a new kinetics model of human GLK (pancreatic isoform) using fluorescence spectroscopy was in agreement with a pre-existing equilibrium of at least two slowly inter-convertible GLK conformations, both able to bind glucose. Three additional phases were observed as intermediates between the super-open and closed states. This study has shown that GLK does not obey the induced-fit mechanism proposed by Heredia et al. (*233*), but is in agreement with the cyclic pre-existing equilibrium model proposed by Kim et al. (*234*).

The authors provide three pieces of evidence to support the existence of the three additional conformational states (*218*). These three states have been observed not only for glucose but for the allosteric activator (compound A in PDB ID:1v4s) binding. Targeted molecular dynamic studies also predicted three intermediate states between the super-open and closed states (*151*).

Titration experiments showed that the allosteric activator (compound A) could bind to GLK in the absence of glucose, although with a modest K_D of ~50-100 μM, supporting the equilibrium binding data and clearly establishing that GLK binds GKA even in the absence of glucose. This structure is subject of study in this thesis and will be discussed in detail in chapter 5.

The current hypothesis is that GLK might be able to perform an extensive sampling of the accessible conformational space delimited by the super-open and the fully closed limiting conformations and driven by the likely existence of three stable intermediates.

4.4 Therapeutic potential & known mutations

Since the discovery of the mutations in the GLK gene, GLK has been subject of therapeutic interest. Disabling mutations in GLK impair catalysis and result in maturity-onset diabetes of the young (MODY) named MODY2 (*235, 236*) and severe forms of permanent neonatal diabetes mellitus (PNDM) (*237, 238*).

Activation mutations cause persistent hyperinsulinemic hypoglycaemia of infancy (PHHI) (*199, 237-239*).

In healthy human pancreatic β-cells, a normal GLK activity results in a threshold for glucose stimulated insulin release of about 5 mM in humans; however, in the case the of alterations in GLK activity, for example as a consequence of naturally occurring deactivating or activating mutations, blood glucose levels can range as low as 1 mM to as high as 50 mM, (*201, 240*).

Patients with MODY2, have impaired liver metabolism and pancreatic islet function. This leads to defective regulation of hepatic glucose metabolism, involving an increase in glucose production when blood glucose levels should be normal, and subnormal clearance of glucose by the liver after a meal because of delayed suppression of hepatic glucose production and impaired conversion of glucose to glycogen (*195*). In the recent years the discovery of allosteric activator molecules binding to a similar region as the activating mutations in PHHI, has created a strong interest in GLK as a therapeutic target for the treatment of diabetes (*241-243*).

Over 200 mutations have been discovered; of those the majority are deactivating and are associated with MODY which are spread throughout the GLK sequence (*244*). The activating mutations appear to be concentrated in the region occupied by the allosteric activator in the X-ray structure of the closed state in complex with glucose and the allosteric activator (PDB ID: 1v4s) (*5*), but are not exclusive to the allosteric binding site (*201, 245*).

The following mutations have been linked to the activation of the enzyme; S64(Y/P/F), T65I, G68V, S69P, D73E, V91L, W99(R/C/L), N180D, M197(I/V), I211F, Y214C, E216D, G446S, V452L, S453A, V455M, A456V. The locations of the activating mutations have been depicted in figure 4.6. Most mutations are within or in close proximity to the allosteric binding site but not exclusive to this area. Residues N180 and M197 are more than 15 Å away from the allosteric site.

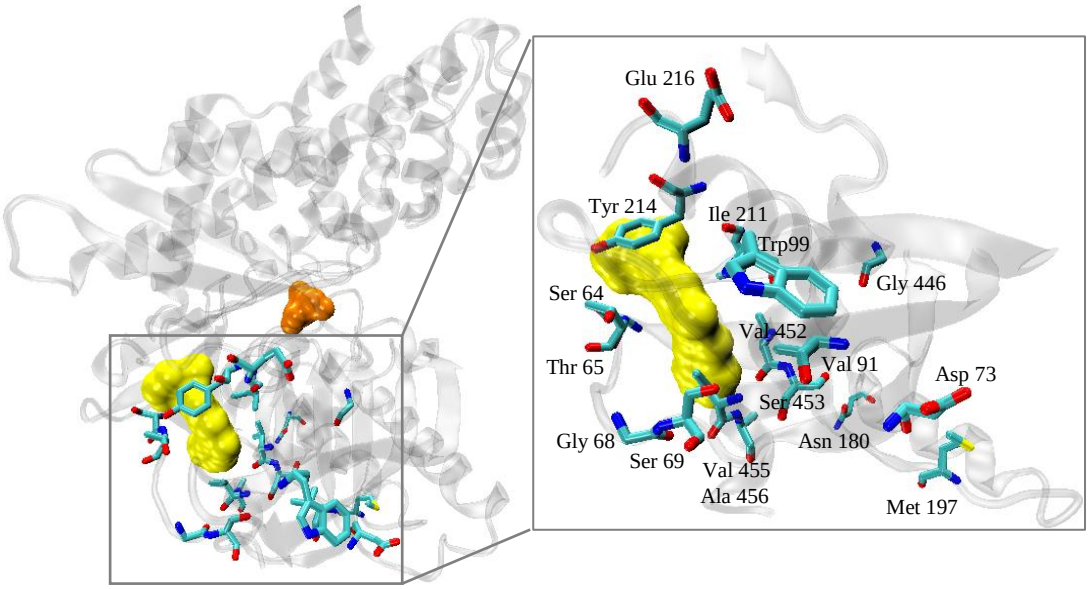


Figure 4.6: The location of activating mutations demostrated in the closed state GLK. (PDB ID: 1v4s). The protein backbone is in grey, the allosteric ligand and glucose represented by yellow and orange surfaces, respectively. On the right a zoomed snapshot of the region highlighting the location of the activating mutations in stick representation.

Of the mutations in figure 4.6, I211F has been shown to be the most active GLK mutation so far identified (*245*). Ile-211 directly interacts with the allosteric activator in the X-ray structure of the closed form in complex with glucose and compound A (PDB ID: 1v4s). The authors postulate that by replacing this residue with a bulkier phenylalanine side chain that would fill the pocket otherwise occupied by the activator, the glucose-bound conformation may be stabilised, in a similar mode of action as the allosteric activator. This mutation is fatal, which is why it has not been observed in patients with hypoglycaemia. T651I, Y214C, V455M, and A456V are predicted to stabilize the closed active conformation through reducing the hydrophobic surface area of the allosteric pocket exposed to solvent. The remaining activating mutations, which map outside of the pocket, are thought to stabilize the closed state by directly affecting domain–domain interactions (*246*).

GKAs bind to a similar region as the activating mutations, and enhance glucose-stimulated insulin release and glycogen synthesis, reduce gluconeogenesis in the liver and lower blood sugar in normal and diabetic animals. Additionally, preliminary human

trials have shown the effectiveness of GKAs in normal and patients with MODY (*200, 201*). However GLK activation can lead to PHHI or cause severe hypoglycaemia in healthy animals. Therefore, when designing activators the goal is to achieve optimal glycemic control while minimising or eliminating hypoglycemic risk (*200*).

4.5 Summary

GLK plays a critical role in glucose homeostasis by catalysing the ATP-dependent phosphorylation of glucose to G6P. In the pancreatic β-cells it plays a key role as a glucose sensor by integrating blood glucose levels and glucose metabolism with insulin secretion. In the liver GLK regulates glucose uptake and glycogen synthesis (*197*).

Inactivating mutations in the GLK gene have been linked to a form of diabetes called type 2 maturity onset diabetes of the young (MODY2) (*3, 198*). On the other hand, activating mutations, cause persistent hyperinsulinemic hypoglycaemia of infancy (PHHI) (*4, 199*). These diseases demonstrate the extent to which GLK action has to be regulated and emphasise the potential therapeutic value.

Structurally, there are two publicly available conformations of this enzyme. The active closed state is bound to glucose and an allosteric activator (PDB ID: 1v4s) and the inactive super-open state (PDB ID: 1v4t). In the absence of glucose and the allosteric activator the apo super-open state conformation differs significantly to that of the closed state, leading to the absence of either binding site from the super-open state.

An additional X-ray structure in a closed state provided by AstraZeneca, which has been co-crystallised with glucose, lacks the allosteric binding site, due to the conformation that the loop connecting the two domains adopts in the absence of an activator. In this structure, we aim to predict the allosteric binding site.

A better understanding of the nature of the allosteric binding site in GLK is the subject of interest in this thesis which could aid the development of a suitable protocol for the discovery of other allosteric binding sites (or novel unexplored cavities) in other targets of therapeutic interest.

The ability to predict such binding sites with theoretical methods would greatly aid and speed the discovery of such novel binding sites, leading to the design of therapeutic drugs with a better opportunity for selectivity and chemo-type diversity.

Chapter 5

Insight into the dynamics of human glucokinase

5.1 Aim

The aim of this chapter is to validate a protocol for the identification of allosteric binding sites in a system where one would expect to do so. Human glucokinase (GLK) has been studied as a test case due to the discovery of a synthetic allosteric activator and the importance of this target in the treatment of type II diabetes.

GLK is a highly dynamic protein system, with two distinct experimentally observed conformations; the active closed form and the inactive super-open form, as described in the previous chapter, that are both publicly available in the protein data bank (PDB). The allosteric binding site is present in the active form of GLK when bound to the allosteric activator. However, the pocket is not present in the super-open conformation. Different kinetic models have been discussed in the previous chapter. The latest kinetic study suggests that there is a pre-existing equilibrium of at least two slowly inter-convertible GLK conformations, both able to bind glucose (*218*). A previously suggested mechanism suggested an induced-fit model of glucose binding to the super-open form, inducing a global conformational change to the closed state. Simulations of both the closed and the super-open states, should give further insight into the mechanism of glucose binding, the nature of the allosteric binding site and the conformation transition mechanism. An X-ray structure provided by AstraZeneca, bound only to glucose, does not contain a cavity at the allosteric site. Simulations of the closed-state allosteric bound X-ray structure (PDB ID: 1v4s) can highlight if the binding site would collapse in the absence of the allosteric ligand from the binding site, and/or it would be

possible to still identify the binding site in the absence of the activator. In theory, in a pre-existing equilibrium kinetic model, even in the super-open state, it should be possible to identify the allosteric binding site, although it is possible that simulations of ns time-scale may not be sufficient to observe this rare event in the super-open state. Both molecular dynamics and normal mode analysis were applied to both conformations of GLK, discussed in the following sections.

The main purpose of this study is to determine the feasibility of identifying an allosteric pocket in an enzyme. If the methods adopted here cannot find the additional pocket, when the starting structure has the bound conformation (PDB ID: 1v4s), then they will likely fail when applied to an unbound conformations (GLK_AZ) discusses in chapter 6.

5.2 MD study of the active closed-state GLK

The holo crystal structure of human GLK in complex with glucose and the allosteric activator, compound A, (2-Amino-4-fluoro-5-[(1-methyl-1h-imidazol-2-yl)sulfanyl]-n-(1,3-thiazol-2-yl)benzamide) was obtained from the Protein Data Bank (PDB ID: 1v4s). The allosteric ligand will be referred to as "compound A", hereafter.

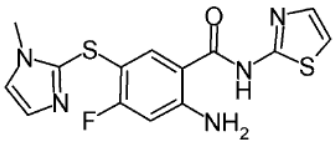


Figure 5.1: Allosteric activator, [2-Amino-4-Fluoro-5-[(1-Methyl-1h-Imidazol-2-yl) Sulfanyl]-N-(1,3-Thiazol-2-yl Benzamide] bound to the active closed form of GLK (PDB identity 1v4s)

For the closed active conformation crystal structure (PDB ID: 1v4s), four simulations were set-up: simulation A, with both ligands (glucose and compound A) present; simulation B, with glucose bound, and the allosteric activator removed; simulation C, with the allosteric activator present, and glucose removed from the active site; simulation D, in which both ligands were removed (see table 5.1).

Simulations	Closed state (PDB ID: 1v4s)
Simulation A	Glucose and allosteric activator bound
Simulation B	Glucose bound (activator removed)
Simulation C	Allosteric activator bound (glucose removed)
Simulation D	Both ligands removed

Table 5.1 Simulations set up for the closed state X-ray structure (PDB ID: 1v4s), depicted also in figure 5.2.

These four simulations should give detailed understanding of both the active- and allosteric binding sites. The binding site flexibility in the absence of either or both ligands and the interdependence of the presence of either ligand on the other binding site are monitored throughout each simulation.

5.2.1 System preparation and MD parameters for simulation of the active closed form of GLK

For each simulation, if ligand present, the ligand was protonated via the PRODRG (*247*) web server, and the protein combined with the coordinates of the protonated ligand, in WHATIF (*248*). All residues were monitored to ensure appropriate tautomeric and protonation states, in particular for residues that were flipped in the WHATIF protonation protocol to maximise hydrogen bonding, where no further adjustments were necessary.

The Amber package (*114*), tool XLEAP, was used to solvate the system with TIP3P water molecules, with a minimum distance of 12 Å from the protein (20804 water molecules). In addition, a total of 20 sodium counter-ions were added to the systems to neutralise the overall charge in each case.

All simulations, unless otherwise stated, have been carried out using the AMBER (*114*) molecular dynamics package using the AMBER03 force field (*100*) for the protein and gaff force field, AM1-BCC charge method (*249, 250*), for the ligand, with Particle-Mesh-Ewald (PME) boundary condition limited to a 10 Å cutoff.

The minimisation was carried out in stages, starting with the minimisation of solvent only while applying a restraint force constant of 500 kcal mol^{-1} $Å^{-2}$ on the rest of the particles, followed by the minimisation of protein with restraint on other particles, then the ligands, if applicable, and finally the entire system, removing all restraint; in each stage, reaching an energy value below the set root-mean squared deviation of 0.001 kcal-$mol^{-1}Å^{-1}$. In each case, the minimisation commenced with 100 steps of steepest descent, followed by conjugate gradient for the remainder.

For each, the minimised system was heated gradually to 300 K in the NVT ensemble, in six 50 K blocks, allowing 15000 MD steps for each block, using a Langevin thermostat with a 1 ps^{-1} collision frequency and a time-step of 2 femtosecond (fs).

The production MD simulations were run in the NPT ensemble, using Langevin thermostat with a 1 ps^{-1} damping parameter (*251, 252*) at 300 K and a time-step of 2 fs. The pressure was controlled by weak-coupling isotropic position scaling (*160*) with a relaxation time of 2 ps. All bonds containing hydrogen were constrained using the SHAKE algorithm (*159*).

For each simulation, 25 ns of MD production was collected after removing the initial ~1.5 ns, for equilibration at the beginning of the NPT production run, based on stabilisation of properties such as energy, volume, density and RMSD.

5.2.2 MD results for the active closed-state GLK

Four 25 ns simulations have been run for the active closed state GLK to show the influence of the presence and absence of both the active site ligand, glucose, and the allosteric activator at the allosteric site (Table 5.1 and figure 5.2). The analysis methods,

root-mean-squared fluctuations (RMSF), principal component analysis (PCA) and binding site search are discussed in this section.

The starting structures for the four simulations of the closed-state X-ray structure (PDB ID: 1v4s) discussed in this section, simulations A to D are depicted in figure 5.2.

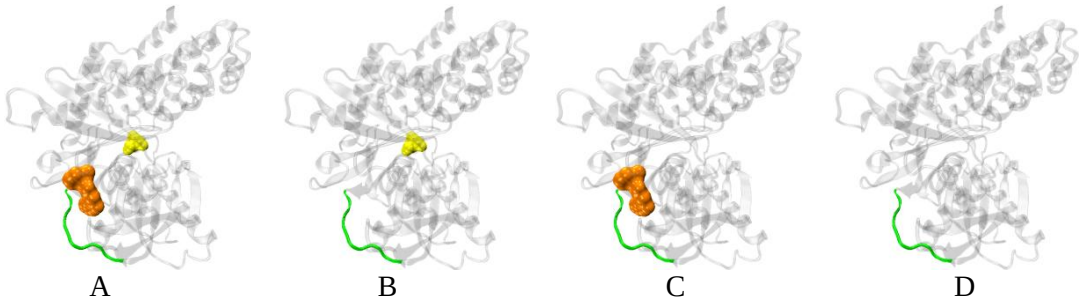
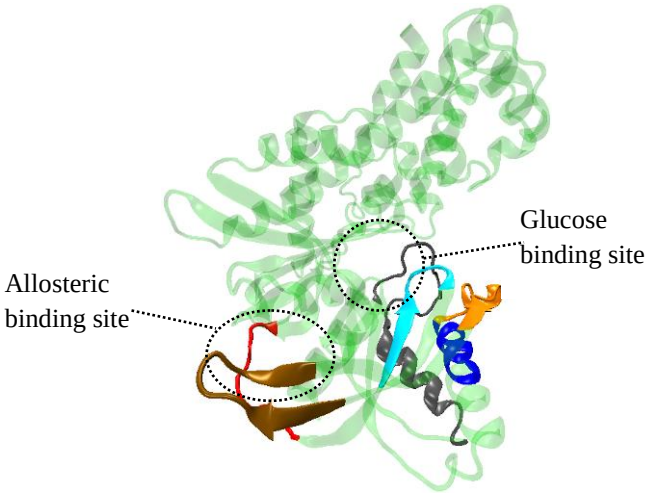


Figure 5.2 From left to right, simulations A to D set up for the closed state X-ray structure (PDB ID: 1v4s). The loop connecting the two domains is represented in green. Both ligands are shown in a VDW representation; glucose and the allosteric activator coloured in yellow and orange, respectively.

For clarity during the discussion, the small domain residues in the GLK closed form are colour coded in figure 5.3. A complete colour-code of the entire protein has been included in appendix B. The large domain adopts similar conformations in both the super-open and closed state; thus conformational changes in the small domain plays a key role in the large conformational change between the two states. A number of residues in the small domain form part of the active site, which is not present in the super-open state, and in addition the allosteric site resides in the small domain. Therefore the simulation analysis is focused mainly on the small domain.



Residues: (64~71 **Red**), (76~83 **Cyan**), (91~100 **Brown**), (108~119 **Orange**), (120~129 **Blue**), (165~194 **Gray**)

Figure 5.3: Structure of the active conformation (PDB ID: 1v4s). For clarity ligands have been removed. Residues in the lower domain, discussed below have been highlighted for ease of understanding.

An overlay of the crystallographic B-factors of the starting crystal structure (PDB ID: 1v4s) with the RMSF plot of the 25 ns 'simulation A', rescaled by Xmgrace (*253*), demonstrate a large degree of similarity, in particular from our point of interest, the connecting loop between the two domains near the allosteric site, residues 65-71 (figure 5.4). A strong peak at residues 94~100 on the RMSF plot, corresponds to a β-turn in the small domain (figure 5.3) close to the allosteric site. It has been observed that this β-sheet and in particular the β-turn resembles a highly mobile flap that can adapt multiple conformations in ligand bound GLK systems, figure 4.4 (unpublished data).

RMSF comparison plots have been used to illustrate the exact differences in residue motions in comparison with 'simulation A', which has been treated as the reference simulation, in complex with both glucose and the allosteric activator. The plots have been created by subtracting the individual residue RMSFs in each simulation from the reference simulation, 'simulation A' RMSF values (Figure 5.5). Thus, a negative RMSF indicates more flexibility.

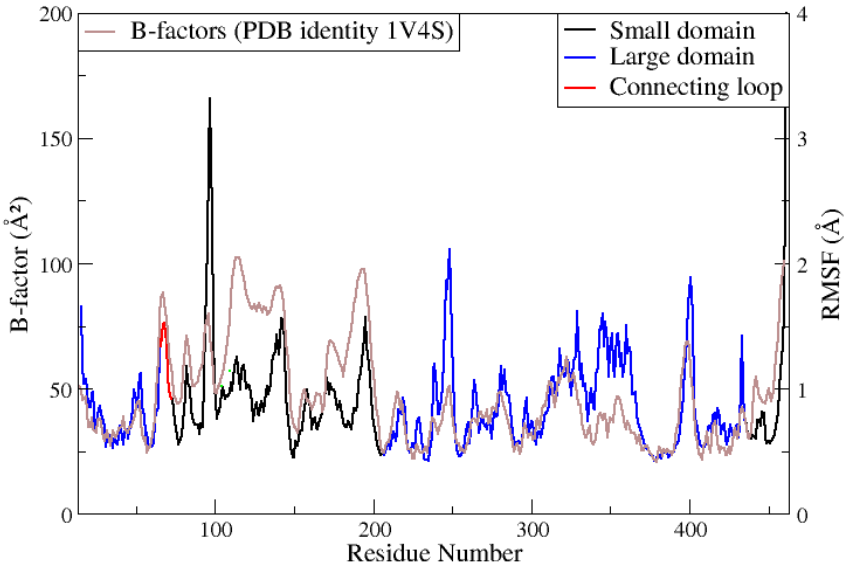


Figure 5.4: An overlay of the B-factors for the closed form GLK (PDB ID: 1v4s) in brown and the RMSF plot for 'simulation A'. The y-axis scale on the left and right of the plot refer to the B-factors and RMSF of 'simulation A', respectively.

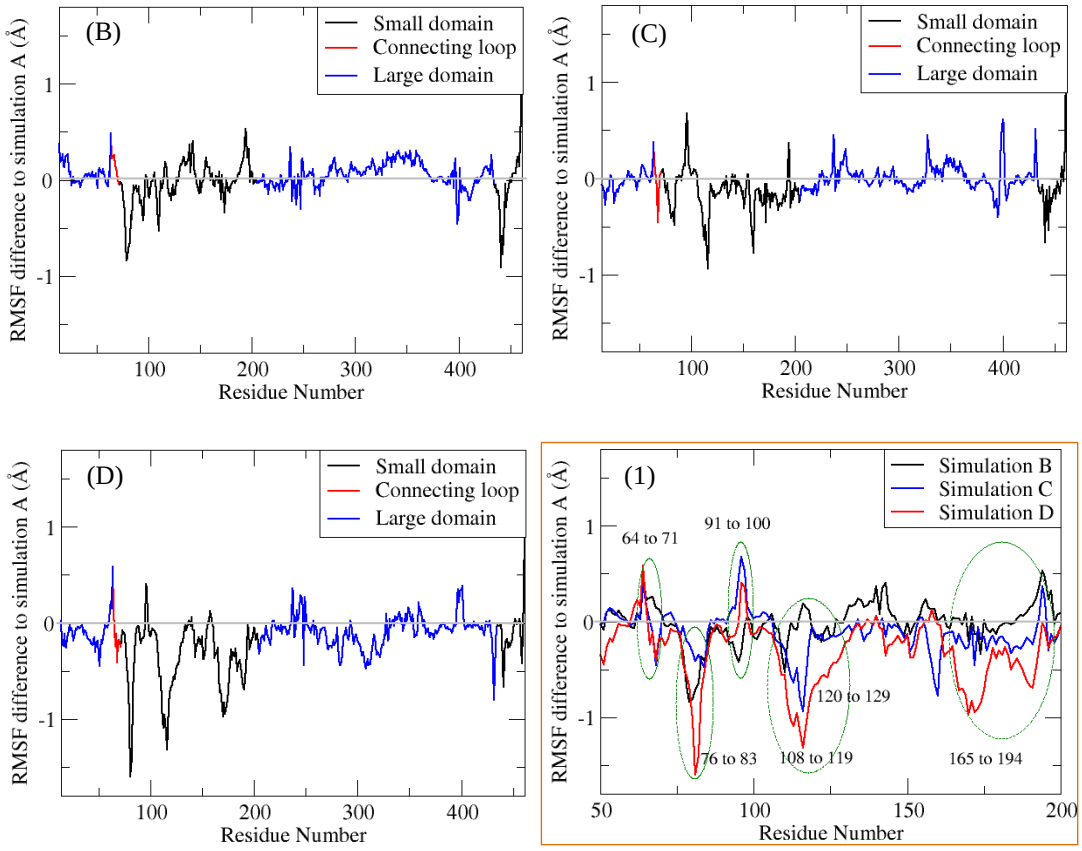


Figure 5.5: RMSF difference plots. Plots B-D are the relative motions of simulations B, C, and D to simulation A, respectively. A negative value is reprehensive of more flexibility in comparison with residues of simulation A. On the lower right, (1) boxed in an orange frame, is a close-up of the three difference plots (B-D), focusing on residues 50-200.

RMSF comparison for the four simulations shows that the motions in the large domain residues remain fairly similar to 'simulation A', regardless of the presence/absence of either ligand, although a slight increase in mobility is observed in the large domain in simulation 'D', in the absence of both ligands.

The small domain residue fluctuations in 'simulation B' are comparatively similar to simulation 'A'. In comparison, residues in the connecting loop are less mobile (Res 64-71). Residue V62~R63 leading to the connecting loop directly interact with the allosteric activator. In the absence of the allosteric activator in simulation B, these residues may stabilise by interacting with neighbouring residues. In addition, residue T65 has been shown (*254*) to undergo a natural activating mutation T65I, highlighting the importance of this residue in this region, in stabilising the close active state conformation.

Residues in the β-turn (residues 91-100) in close proximity to the allosteric binding site become slightly more mobile in simulation 'B' (black line in plot (1), figure 5.5) and in contrast noticeably more rigid in the absence of glucose, in simulations 'C' and 'D' (blue and red lines in plot (1), figure 5.5). This may indicate an important mobility that leads to the allosteric ligand access to the binding site. The other end of this β-sheet is in close proximity to the active site (figure 5.3). It appears that in the presence of glucose at the active site, where possibly local interactions keep this end of the β-sheet stable, the other end close to the allosteric site, compensates by becoming very mobile.

Residues of the small domain (residues 72-203) in simulations 'C' and 'D', demonstrate comparative trends in mobility, although much stronger fluctuations are observed in simulation 'D'. In the absence of glucose in simulations 'C' and 'D' residues at the beginning of the connecting loop (residues 64-71), similar to simulation 'B' become less mobile, however the second half of connecting loop is slightly more mobile in simulation 'C' and 'D' in the absence of glucose and both ligands, respectively. Although distant from the allosteric site, and close to the active site, residues 76 to 83, become more mobile in the absence of the allosteric ligand, in simulations 'B' and 'D', but only slightly more mobile in simulation 'C', in the presence of the allosteric activator. These residues are part of a β- sheet and turn (cyan in figure 5.3), close to the active site but are directly linked to the connecting loop (red in figure 5.3) in the

allosteric region, separated by 5 residues. These residues may take part in the allosteric regulation and communication.

Residues 108-119, very close to active site (figure 5.3), become significantly more mobile in the absence of glucose, in both simulations 'C' and 'D', however residues following these, (residues 120 to 129), an α-helix section only become more mobile in the absence of both ligands, in simulation 'D', suggesting an important role in stabilising the closed state in the presence of glucose and/or allosteric activator.

The peak at residue 157-161 observed only in the RMSF plot of simulation 'C' (figure 5.5) corresponds to a β-turn pointing into the allosteric site, from behind (relative to the orientation in figure 5.3). In the presence of glucose, this region is fairly stable. Interestingly, residues leading to this β-sheet are positioned at the active site (151 to 153) and residues 157-161 correspond to part of the disordered region in the super-open state (157-179).

The next major RMSF profile difference relative to simulation 'A', stems from residues 165 to 194, only in the absence of both ligands. Of those, residues 165 to 179 are part of a loop at the active site. Residues 168 and 169 directly interact with glucose at the active site, while residues 180 to 194 belong to an α-helix following this loop close to the active site (figure 5.5).

The root-mean-squared deviations (RMSD) of each simulation were plotted with reference to the first frame of the simulation and to the open form of Hexokinase I (PDB identity 2YHX) (figure 5.6). The structure of Hexokinase I has been used in this case, as there are no available X-ray structures of the intermediate open state for GLK (hexokinase IV). The super-open state of GLK (PDB ID: 1v4t) is structurally too different to be used for this purpose. Simulations 'A' to 'C', did not demonstrate a noticeable deviation from the starting structure (data not shown), however in 'simulation D', there is a clear deviation away from the starting frame, ~12.5 ns into the simulation (figure 5.6). In addition, an RMSD to the open form of Hexokinase I, shows that there may be a shift towards domain opening when both the glucose and compound A are removed.

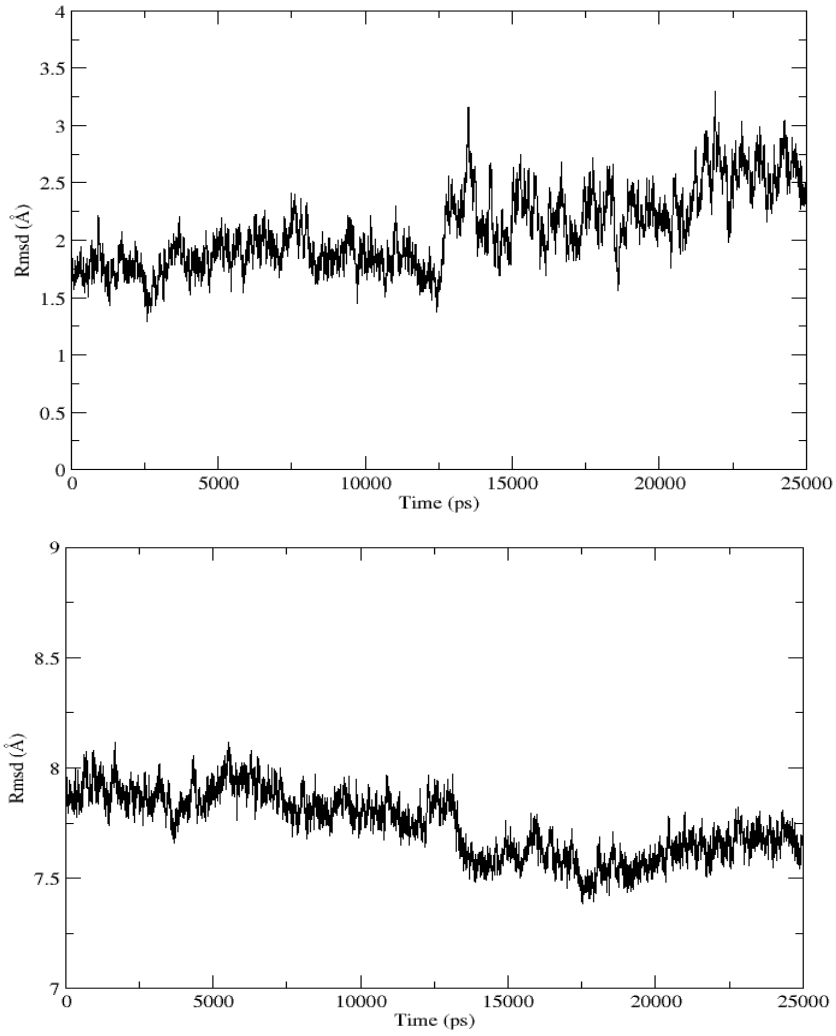


Figure 5.6: Cα RMSD plot of simulation 'D', with reference to the first frame of the simulation on top, and on the bottom with reference to an open conformation of Hexokinase I (PDB ID: 2YHX), in the absence of an open Hexokinase IV crystal structure.

Principal component analysis has further confirmed the opening and closing motion of the two domains in 'simulation D', see later in this chapter (figure 5.8), confirming that in the absence of both ligands the enzyme conformation shifts towards an open state. Owing to the relatively short simulation time (25ns), the complete conformational transition has not been observed here but an indication of a conformational shift towards and opening of the two domains has (Figure 5.6 and 5.8).

The mobility pattern that has emerged from the above simulations is in agreement with the observed mechanism of GLK where 'simulation B', in the presence of glucose, follows a similar RMSF pattern to simulation 'A'; however 'simulation C and D', in absence of glucose, or both ligands, respectively, share a similar pattern of freedom in mobility, although 'simulation D' demonstrates the highest flexibility. The above observations, demonstrating sensitivity of the simulations to the presence or absence of either ligand, provides confidence in the accuracy and suitability of the simulation set-up for this system. Thus, it would be appropriate to continue with similar simulation parameters with further studies of new starting structures of GLK, investigating the allosteric binding site.

It has been observed that in the presence of both ligands at the active and allosteric binding site, residues at the tip of both the large and small domains, are mobile, while residues that actually interact with the ligands and the neighbouring residues remain fairly stable (figure 5.5 and appendix B). While removing the allosteric ligand does not change the RMSF profile of residues in the large domains, in the small domain this leads to a small reduction of mobility of residues 64 and 65 of the connecting loop, while the others remain unchanged (figure 5.5, B). An increase in the mobility of following residues (residues 76-83), correspond to a β-strand and a loop close to the active site (figure 5.3). Residues of the β-turn (residues 91-100) remain or become slightly more mobile in the absence of the allosteric ligand.

When glucose is removed from the simulation in simulation 'C', creating a hypothetical system, in which the allosteric ligand would remain bound in the absence of glucose, residues of the small domain become more mobile, highlighting that the main driving force of the close-state conformation, is glucose, consistent with the allosteric transition from the super-open to the closed state in the presence of glucose. In the absence of both ligands, the system overall becomes more mobile, including residues of the large domain. Interestingly, the preceding residues of the connecting loop become increasingly stable, while those of the β-turn (residues 91-100) become highly mobile. Residues 108-129 of the small domain demonstrate greater mobility, as do residues 165-194 of an α-helix in the small domain (figure 5.3). An RMSD of this simulation ('D'), in the absence of both ligands, does demonstrate that the system begins to move away from the close state starting structure, at ~12-13 ns into the simulation, while an RMSD

comparison with an open state hexokinase I, further suggests that the simulation trajectory may have very slightly moved towards an open state.

Considering the simulation observations here, it is likely that the presence of glucose at the active site increases the mobility of residues in the allosteric region, promoting the binding of an allosteric activator with potentially more accommodating conformations of the region available.

In the following section, principal component analysis (PCA) was applied to each simulation to highlight the major motions sampled, and to establish if conformational changes in the allosteric binding site region, in particular the loop connecting the two domains, are among the major slow motions in the system.

5.2.3 Principal Component Analysis

An MD trajectory is a mixture of fast and slow motions. Although, on the 25 ns time-scale we would not expect to observe large conformational changes, principal component analysis (PCA) has been applied to extract the most significant motions sampled in the above simulations. It is also of interest to monitor residues in the connecting loop in the major principal components to establish whether the connecting loop motion is correlated with the large conformational changes, i.e. the opening and closing of the two domains, or an independent flexible loop that can adopt various conformations in the presence of the allosteric activator.

GROMACS (*181*) was used to carry out the PCA on frames collected from each trajectory. In each case, the covariance matrix and resulting eigenvectors and eigenvalues have been generated using the α-carbon atom trajectories.

For proteins in general, the first few eigenvectors describe the largest variance in the motion (*148, 255*). Scree plots depict the proportion of the motion described by the first few principal components (PCs), in the simulations above (Figure 5.7). The cut-off usually is chosen at the elbow of the slope.

The scree plots demonstrate that in the presence of the either or both ligands at the active site and the allosteric site, at the timescale of 25 ns, that overall, the backbone residues (just the Cα in PCA) do not undergo global motions comparable to simulation ‘D’, as

evidenced by a larger eigenvalue on the scree plot (figure 5.7). Nevertheless, all four scree plots show that a handful of modes are sufficient to capture the majority of the motion sampled.

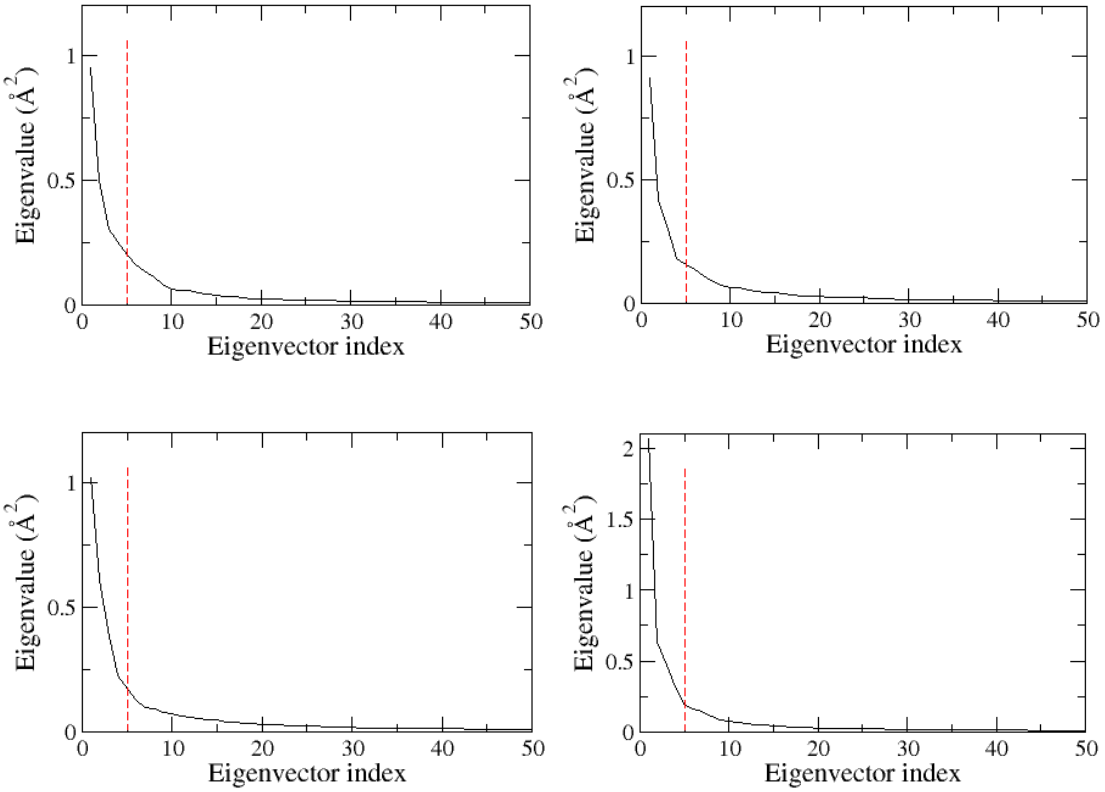


Figure 5.7: Scree plots of simulations A (top left) B (top right), C (bottom left), D (bottom right). In total 1344 eigenvectors calculated for each trajectory (448 residues, 3N). For clarity only the first 50 are shown here. The red line depicts the eigenvalues of the first five eigenvectors.

Below, the first three PCs for each simulation have been depicted in figure 5.8. The indication of increased global residue mobility suggested by the scree plot for simulation 'D' is further evidenced by the visual representation of the PCs (figure 5.8), demonstrated by the two extreme conformations sampled for each PC, with arbitrary frames placed in between. The first PC in simulations 'A' to 'C' in the presence of either or both ligands samples local residue fluctuations, owing to the conformational restriction incurred by the presence of either or both ligands, whereas in simulation 'D', PC 1 captures a global backbone mobility that corresponds to a subtle opening and closing of the two domains in the absence of both ligands.

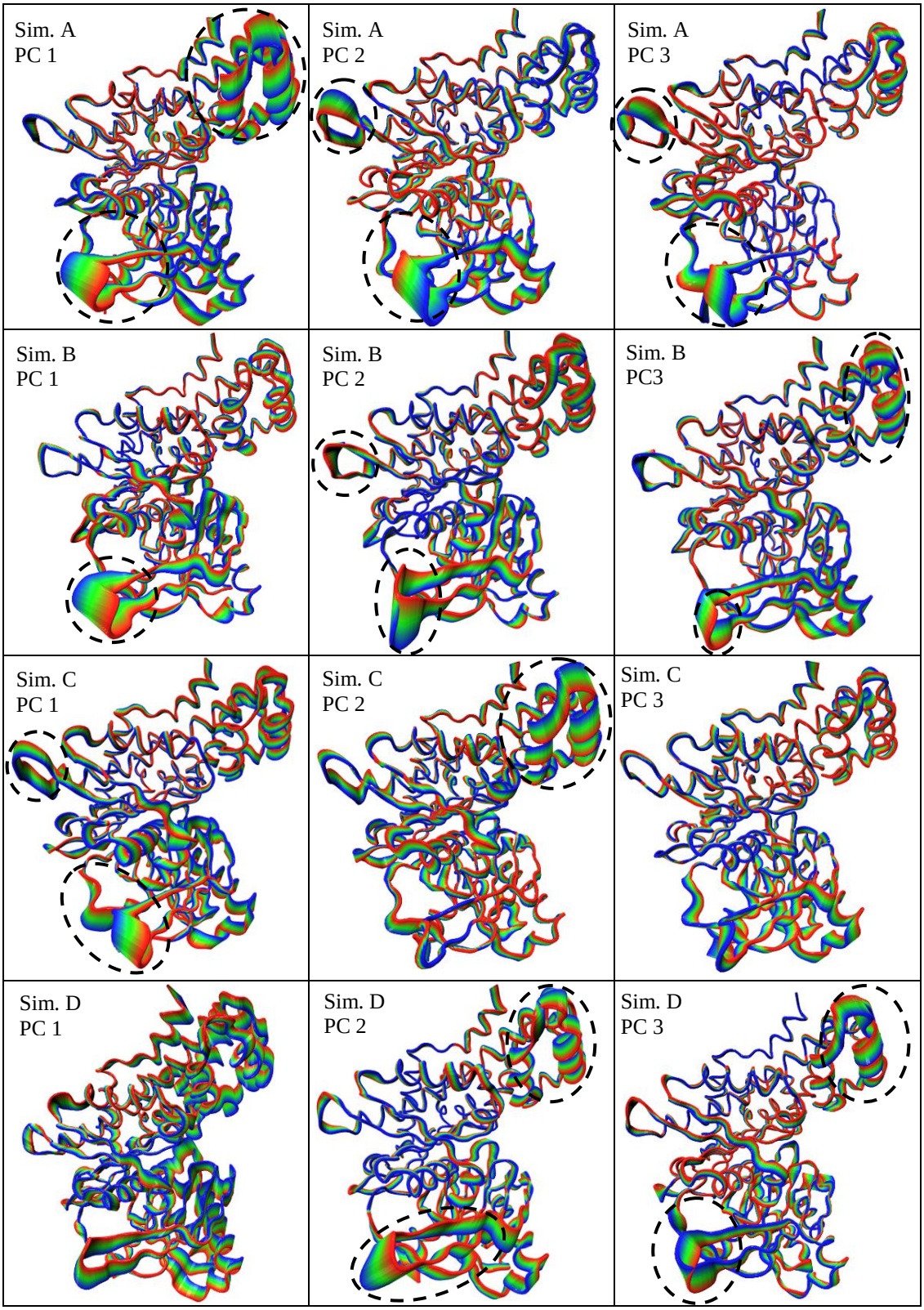


Figure 5.8: First three PCs in each simulation. From top, 'simulation A', 'simulation B', 'simulation C' and 'simulation D'. The timestep colour representation in VMD (*256*) has been used. Circled in dotted line are regions of significant mobility.

In figure 5.8, for all four simulations, the two extreme conformations of each PC sampled along the trajectory with respect to the average structure have been demonstrated in red and blue, with arbitrary frames interpolated in between the two extremes, degenerated with GROMACS (*181*). The regions of high mobility have been circled.

Focusing on the first PC in all four simulations, it is evident that residues in the β-turn mentioned previously (residues 93-99) are rather flexible, especially in the presence of glucose at the active site. The connecting loop also moves simultaneously with the β-sheet in the allosteric site region. However, in 'simulation D' where the entire protein is undergoing the opening and closing motion, the connecting loop and the β-sheet sample a mobility consistent with the rest of the residues. The flexibility of the this β-sheet is intensified in the presence of glucose at the active site in simulation 'B' and in the presence of both ligands in simulation 'A', and less intense in simulation 'C', in the absence of glucose and significantly reduced in the absence of both ligands in simulation 'D'. This mobility highlights significant correlation with the presence of either ligand, consistent with the RMSF plots, which suggests a region of significant importance in the allosteric binding site region.

The observation above further suggests that in the presence of glucose at the active site, the allosteric region samples a number of conformations, which should be able to accommodate the allosteric activator. This observation is promising in studying later active GLK crystal structures only bound to glucose, which should allow us to identify the allosteric site in the simulation, providing the simulation length is sufficiently long for sampling.

In the following section, the nature of the binding site flexibility during the four simulations will be discussed, to give a better understanding of the nature of the active and allosteric sites. Two binding site search tools Pocket-Finder and Q-SiteFinder (*95*) will be applied to frames from all simulation trajectories, to examine the performance of the two tools. This should allow us to assess the accuracy of the binding site search methods utilised here in frames from simulations where we would expect to identify the known binding pockets.

5.2.4 Binding site profile of the closed state X-ray structure

The two described tools (chapter 3), Q-SiteFinder and Pocket-Finder were utilised to investigate the dynamics of the active and allosteric binding sites. Shown below in figure 5.9, is the original crystal structure of the active, closed state GLK conformation, (PDB ID: 1v4s) with the top 10 binding sites predicted by both tools. It is important to validate that the tool can accurately predict the existing, known, binding sites and rank them well.

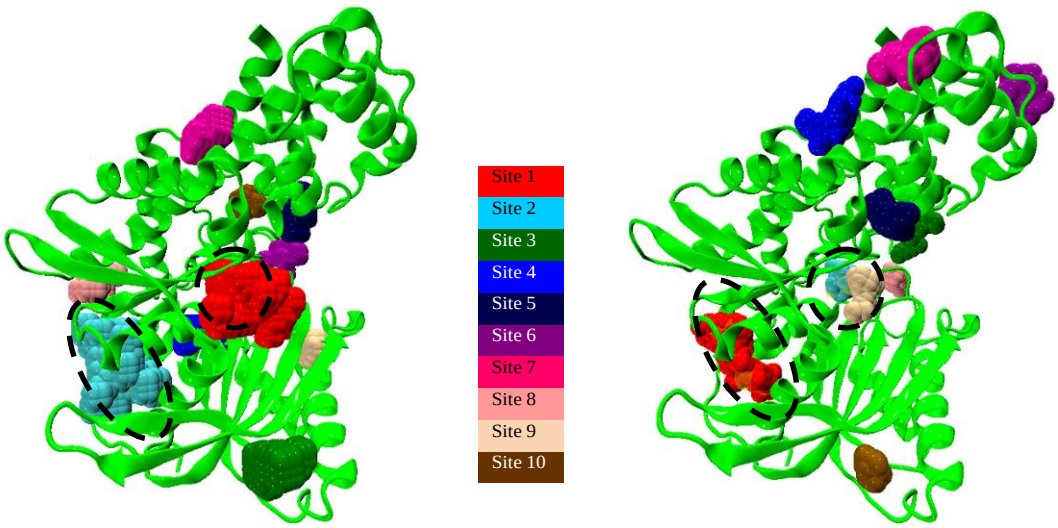


Figure 5.9: On the left, top 10 sites predicted for (PDB ID: 1v4s) by Pocket-Finder, and on the right the corresponding by Q-SiteFinder. In the centre is the colour code for the sites in order of rank. The dotted circle and oval indicate the glucose and allosteric ligand locations in the X-ray structure (PDB identity: 1V4S), respectively.

Both tools provide a precision value for the level of accuracy in mapping the existing ligand in the binding site. For the X-ray structure (PDB ID: 1v4s), Q-SiteFinder predicted a binding site ranked as the 1st site in the allosteric region, mapping the existing allosteric activator relatively well. Pocket-Finder on the other hand predicted a larger binding site, in this region, and ranked as the 2nd site. Images in figure 5.10 illustrate the degree of similarity between the actual ligand size and the predicted binding site using the two methods.

Similarly for the glucose binding site, Q-SiteFinder predicts a site ranked as the 2nd site, mapping glucose closely. Pocket-Finder again identifies a much larger site than the actual size of glucose, but ranks this site as the 1st. In addition to glucose at the active site, ATP also binds for the phosphorylation of glucose, but was not included in the crystal structure and the simulations. Therefore, it may be more accurate to imagine a larger site than a tight site mapped onto glucose.

As the two tools use different underlying search methods for the binding site search and rank, described earlier (chapter 3, section 3.5), the order of ranking is different. Q-SiteFinder tightly maps the ligand whereas Pocket-Finder predicts larger sites, ultimately combining some of the smaller sites predicted by Q-SiteFinder. The glucose binding site predicted by Pocket-Finder, ranked as the 1st, maps with three sites predicted by Q-SiteFinder, ranked as the 2nd, 8th and 9th.

The degree of overlap between the pockets predicted for both binding sites, and the respective ligands, has been demonstrated in figure 5.10.

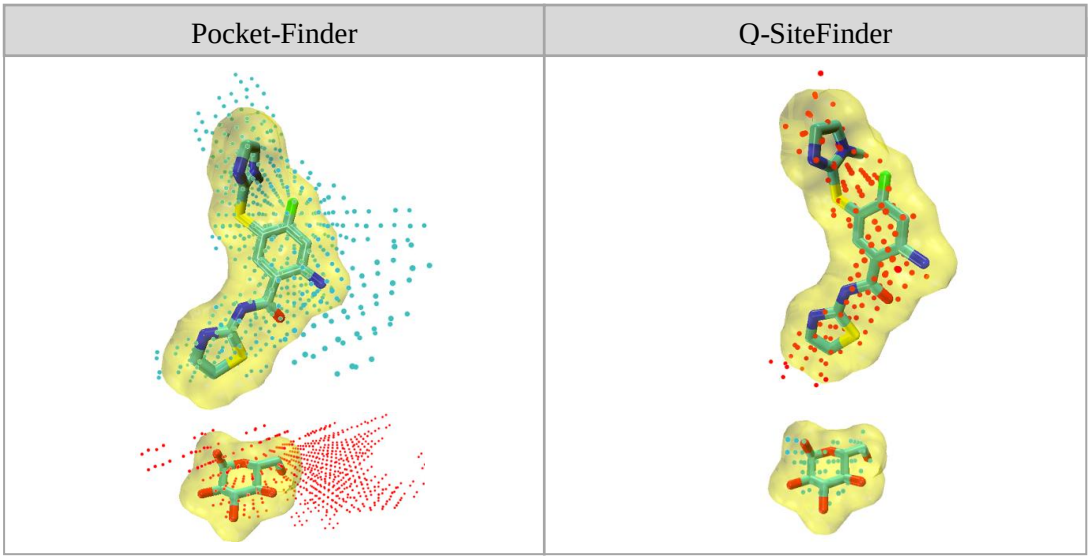


Figure 5.10: In point representation, the predicted grid points for the binding site. On the left top and bottom, the allosteric activator and glucose sites, respectively, predicted by Pocket-Finder. On the right, the corresponding, predicted by Q-SiteFinder. The point colours correspond to the colour code in figure 5.9. The van der Waals surface of each ligand is represented in a yellow transparent smooth surface.

In the closed form, as defined by X-ray crystallography (*5*), glucose binds to the bottom of the deep cleft between the large domain and the small domain and interacts with residues E256 and E290 in the large domain, T168 and K169 in the small domain, and forms hydrogen bonds with N204 and D205, (Figure 5.11).

The residue predictions above by both methods at the glucose binding site are all consistent to those predicted by the crystallographers in the X-ray structure, although several other interactions have also been suggested at each site here (Table 5.2).

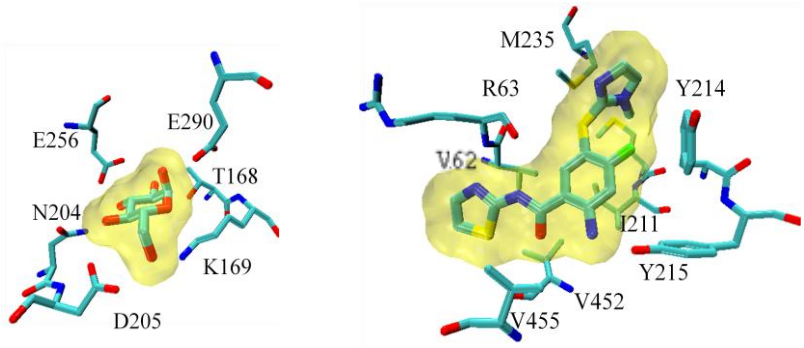


Figure 5.11: On the left glucose binding site in the X-ray structure (PDB ID: 1v4s) (*5*) and on the right the allosteric activator (compound A) at the allosteric site.

For the allosteric binding site, residues V452 and V455 have not been predicted using neither of the binding site search tools, but are believed to interact with the allosteric ligand (*5*).

The predicted interactions at either binding site by both Pocket-Finder and Q-SiteFinder have been summarised in table 5.2. Pocket-Finder identifies larger binding pockets at both sites, in comparison with Q-SiteFinder, and thus higher numbers of interaction points (residues) have been recorded for both sites. The interaction sites found are greater in numbers than the actual interactions by either ligand at each site. For the allosteric site, this may offer additional/alternative interaction points for which synthetic activators may be designed.

Q-SiteFinder (Active Site)					
Residue Name/Num.		**Residue Name/Num.**		**Residue Name/Num.**	
SER	151	THR	206	**GLU**	**256**
PHE	152	ILE	225	TRP	257
PRO	153	THR	228	GLY	258
THR	**168**	GLY	229	ALA	259
LYS	**169**	CYS	230	GLN	287
ASN	**204**	ASN	231	**GLU**	**290**
ASP	**205**	ASN	254		

Q-SiteFinder (Allosteric site)					
Residue Name/Num.		**Residue Name/Num.**		**Residue Name/Num.**	
TYR	61	ASP	158	CYS	220
VAL	**62**	ILE	159	GLU	221
ARG	**63**	**MET**	**210**	**MET**	**235**
SER	64	**ILE**	**211**	ARG	250
THR	65	**TYR**	**214**		
PRO	66	**TYR**	**215**		
GLN	98	HIS	218		

Pocket-Finder (Active Site)					
Residue Name/Num.		**Residue Name/Num.**		**Residue Name/Num.**	
ASP	78	PHE	152	GLY	229
GLY	80	**LYS**	**169**	CYS	230
GLY	81	**ASN**	**204**	ASN	231
THR	82	**ASP**	**205**	**GLU**	**256**
ASN	83	THR	209	GLY	258
PHE	84	ILE	225	ALA	259
ARG	85	GLY	227	GLN	287
SER	151	THR	228	**GLU**	**290**

Pocket-Finder (Allosteric site)					
Residue Name/Num.		**Residue Name/Num.**		**Residue Name/Num.**	
TYR	61	GLY	68	HIS	218
VAL	**62**	GLN	98	CYS	220
ARG	**63**	ILE	159	GLU	221
SER	64	**MET**	**210**	**MET**	**235**
THR	65	**ILE**	**211**	ARG	250
PRO	66	**TYR**	**214**		
GLU	67	**TYR**	**215**		

Tables 5.2: Predicted interactions at the active and the allosteric site, on the left and right, respectively for the X-ray structure (PDB identity 1v4s). On top, predicted by Q-SiteFinder and on the bottom, by PocketFinder. In bold are those reported by the crystallographers, to interact directly with glucose and the allosteric activator (*5*) (Figure 5.11).

It is evident that Q-SiteFinder can predict a known binding site with a higher level of accuracy; however the rigorous search may be a limitation in our studies where the binding site profile will continuously change through the simulation, and where the exact boundaries of the site may not be crucial. Once a plausible site has been identified, more accurate methods can be used for exact mapping of the binding site, which may then be used in *de novo* design of suitable fragments for the binding site.

For the aforementioned simulations, the binding site flexibility through the simulations was monitored at 1 ns interval, to gain insight into the nature of the binding sites, especially in the absence of either ligand from their respective binding sites. In addition, it is important to be able to correctly predict and highly rank the active and allosteric

binding sites in the simulations where we would expect the binding sites to remain intact, especially in the presence of the appropriate ligand.

5.2.5 Binding site profiles through simulations 'A' to 'D'

In snapshots for all four simulations, neither the active site nor the allosteric binding site collapsed. Mostly, the allosteric and the glucose binding sites were ranked very high in the 25 frames of each simulation, as the 1st and 2nd pockets. However, the size of the pocket changed. Binding site volume changes have been demonstrated with the aid of plots in figures 5.12-5.13.

Both tools have predominantly ranked the two binding sites, within the first 5 ranked sites. In particular, Pocket-Finder has ranked both sites within the top 2 sites. The ranking has therefore not been highlighted in the volume plots. In simulations, where either/both ligands were removed, the site has been identified by the calculation of the known contact residue count in the predicted sites, for the respective ligands (Table 5.2). The highest presence of contact residues has been used to identify the location of the active and allosteric sites, in the absence of either/both ligands, in 'simulations B to D'.

The active binding site volume profile (figure 5.12) is fairly different between the two methods for all 4 simulations; however in both, the pocket is highly ranked and in almost all frames, all 6 interactions with glucose mentioned in table 5.2 have been identified.

At no point in simulations 'C' and 'D', where glucose has been removed from the active site, did the binding site completely collapse, although it is noticeable that with Q-SiteFinder, greater fluctuations in pocket volume are observed along all four simulations, as the method is more sensitive to local variations.

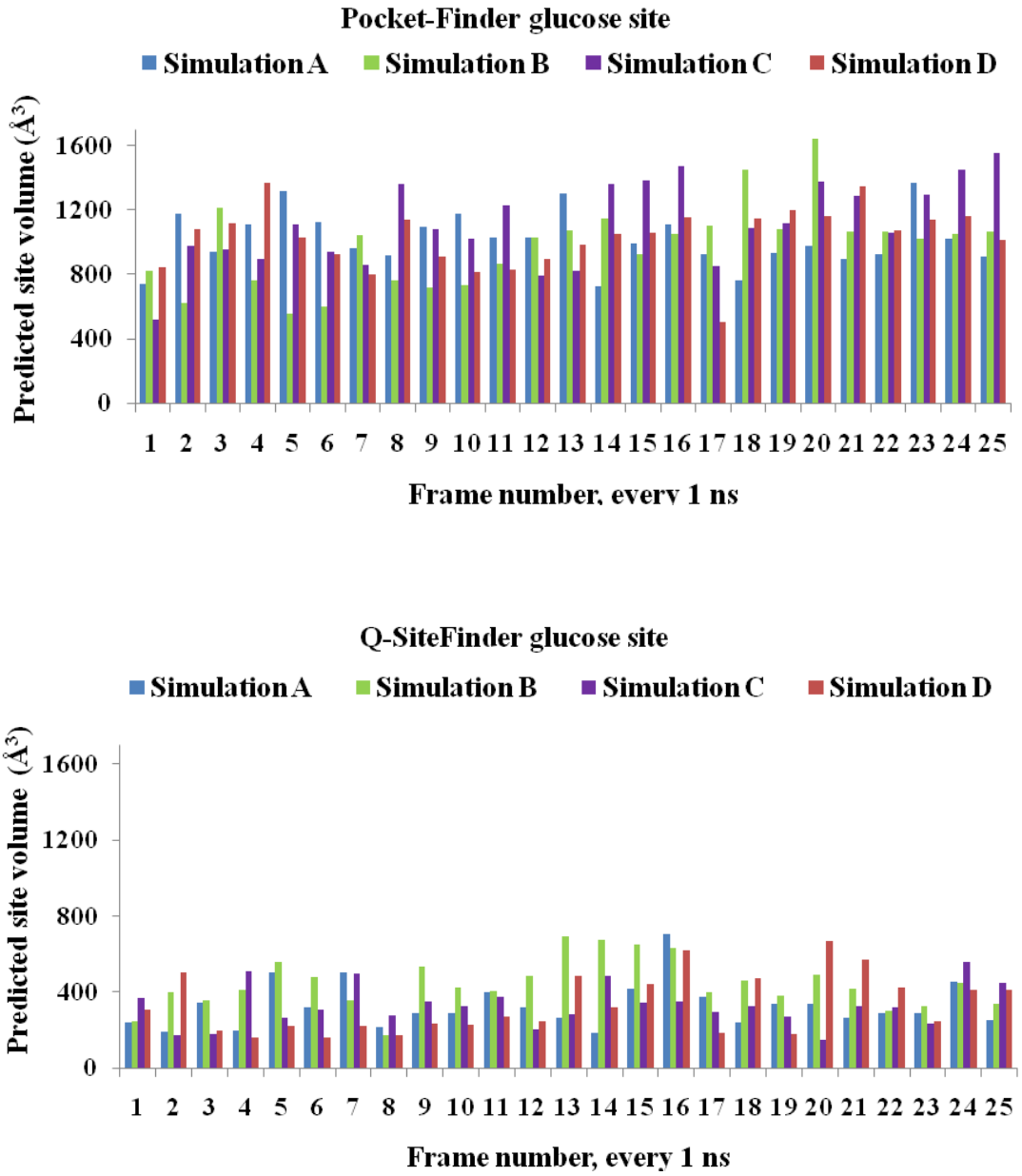


Figure 5.12: Active site volume profiles at every 1 ns interval for each of the 25 ns simulations, with Pocket-Finder (top) and Q-SiteFinder (bottom).

The allosteric site is ranked within the top two sites in 'simulation A' and 'simulation C', with almost all required interactions with the allosteric activator identified by both tools. The pocket volume change in the allosteric site is depicted in figure 5.13.

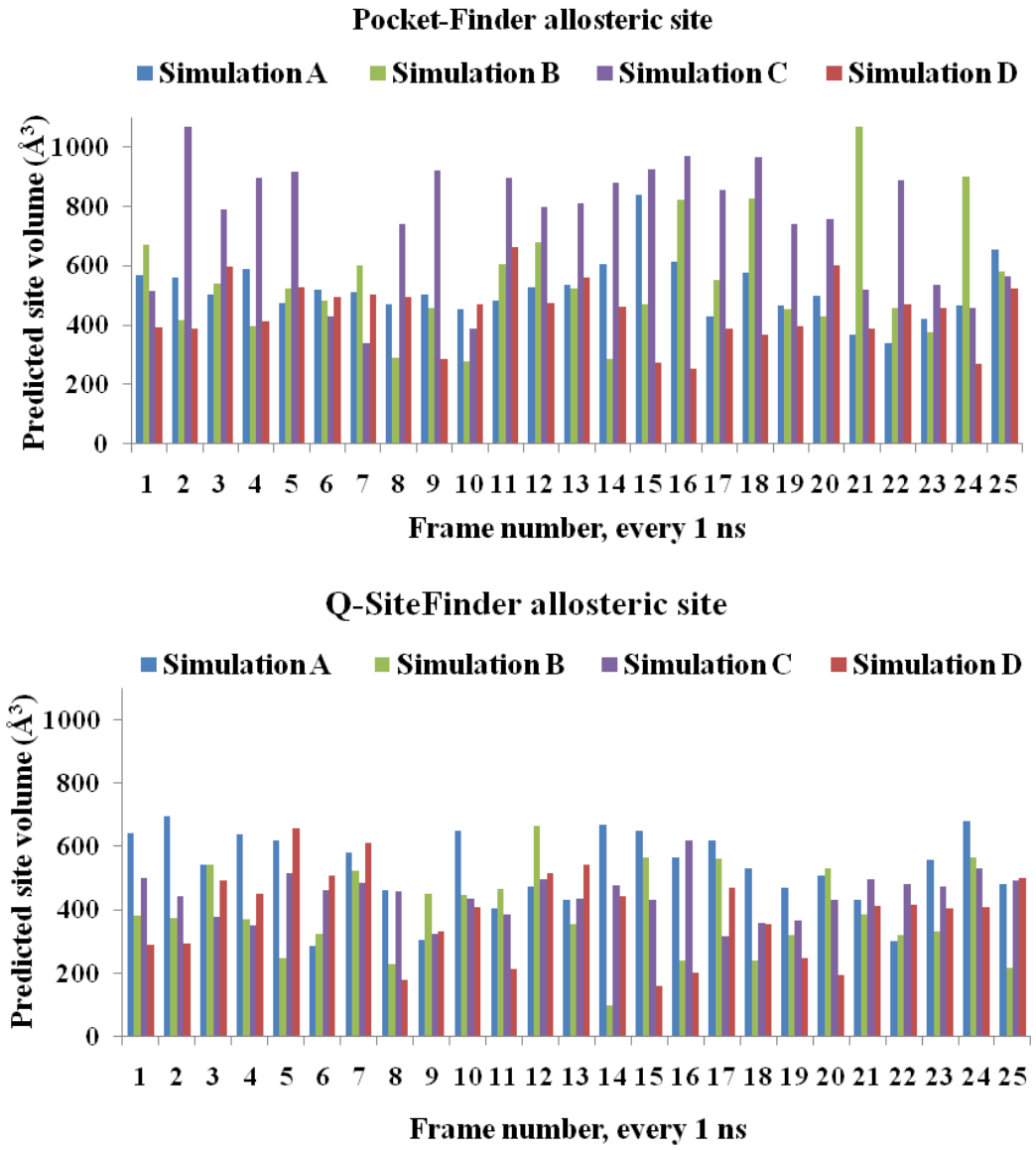


Figure 5.13: Allosteric site volume profiles at every 1 ns interval for each of the 25 ns simulations, with Pocket-Finder (top) and Q-SiteFinder (bottom).

Using Pocket-Finder, the allosteric site volume fluctuates more in the presence of the allosteric ligand in simulation 'C'. This volume profile is not observed in the Q-SiteFinder pocket search through the simulation frames. However, extra lower ranked sites involving a few of the allosteric activator interacting residues were observed nearby, in other sites predicted by Q-SiteFinder. In the absence of the glucose from the active site in simulation 'C', the allosteric binding site flexibility increases, despite the presence of the allosteric activator. The volume mainly extends in to the region close to

a β-turn (residues 157-161), that demonstrates increased flexibility in the absence of glucose at the active site, in simulation 'C'. As discussed in relation to the RMSF plots, this β-turn is part of a β-sheet that follows a loop at the active site (appendix B). With Q-SiteFinder, for the same frames of the trajectory, a separate binding site is predicted in the vicinity of this β-turn and in some frames, it is also highly ranked. We could view the separate site predicted by Q-SiteFinder as site fragmentation, or as a more precise binding site mapping method.

In the absence of both ligands from their respective binding sites, 'simulation D', searching with both tools, the allosteric site was still identified, although the sites were not always highly ranked with Q-SiteFinder and not all interactions were identified. However, once again nearby sites included some of the interacting residues missed from the major allosteric site. In Pocket-Finder, this is not the case, as the Q-SiteFinder split sites that are unified into one.

Owing to the level of sensitivity in Q-SiteFinder, the binding sites which may be grouped into one large site in Pocket-Finder are here split. This may have an impact on the ranking order, as the number of identified grid points in each site will be reduced. In addition, as Q-SiteFinder identifies several sites in one region, but is limited to returning results for only the top 10 sites, it will not show sites in as many regions of the system as Pocket-Finder, despite the same limitation on the number of sites reported.

The main objective of the binding site search in the above simulations was to establish the level of accuracy and plausibility of identifying highly ranked binding sites when expected. Both methods have identified the active and allosteric sites correctly, showing sensitivity to the changes in the protein conformation through the simulation. This allows us to utilise the two binding site search tools, with some degree of caution, in future studies of systems with less prior knowledge of the binding site locations. Some degree of system understanding may be required to eliminate sites which are highly unlikely to be important for binding.

Looking at particular protein families, some common regions may be targeted for binding. In kinases, one may be able to apply common knowledge of binding site regions in one enzyme to explore others in the family. Some analysis of the protein

dynamics, identifying the major conformational changes in the active/inactive states, and mutagenic studies may offer clues, when in search of alternative binding sites.

With GLK, if no prior knowledge of the allosteric binding site was available, one may have been able to use the information regarding the activating mutations in this region, in combination with a binding site search tool, to conclude the existence of such an allosteric binding site.

Considering that the allosteric binding site was observed in all chosen frames along the simulations, including simulations where the allosteric ligand was removed from the starting structure, 'B' and 'D', a binding site search along the principal components of the simulation will not be necessary.

In the following section we take a closer look at the correlated motions in the four simulations.

5.3 Normal mode analysis

All atom normal mode analysis was carried out to assess the major motions in the system. The full atomistic NMA would take full account of all motions, including those of side-chains that may be important to binding. In addition to validating the major motions obtained by MD and PCA (figure 5.8), if the allosteric binding site had appeared during the major slow normal modes in the system, then there would be no need for long time-consuming MD simulations. In this case, as the starting point is a state that already contains the allosteric site, the binding site should not disappear during NMA of a structure with the ligand present at the binding site. It is interesting to monitor the motion in this region for all four scenarios mentioned in table 5.3, to gain better understanding of the allosteric region and the performance of NMA and the sensitivity to the presence of the either ligand.

Nucleic Acid Builder (NAB) was utilised to run the normal modes calculation. (see appendix C, for run script). The routine is written in NAB language and converted to .c executable by the program. The original x-ray structure of the closed active form of

GLK was used (PDB ID: 1v4s). Four starting structures were set up, summarised in table 5.3.

Normal mode analysis starting structures	
NMA A	Glucokinase in complex with glucose and compound A
NMA B	Glucokinase only in complex with glucose
NMA C	Glucokinase only in complex with compound A
NMA D	Glucokinase (both ligands removed)

Table 5.3: Normal mode analysis starting structures for the closed-state x-ray structure (PDB ID: 1v4s).

Similar to the MD simulations, the AMBER 03 force field (*100*) was used for the protein and the gaff force field, AM1-BCC charge method was applied to the ligands (*249*). Each complex was initially minimised with the conjugate gradient method to reach an RMS energy gradient of 1.0 x 10^{-5} kcal $mol^{-1}Å^{-1}$, followed by 2nd derivative Newton-Rhapson minimisation method, terminated at a RMS energy gradient of 1.0 x 10^{-12} kcal $mol^{-1}Å^{-1}$, for the root-mean squared deviation convergence criterion, under the Generalised-born implicit solvent conditions (*257*). Owing to computer memory limitations, 50 normal modes (NMs) were collected for each structure, of which the first 6 have a value of zero, corresponding to the rotational and translational motions in the system.

5.3.1 Results

The frequency values for the first 44 non-zero modes range between ~2 to ~10 $cm^{-1,}$ with slight variations for each complex. Viewing the complete RMSF plots for each normal mode does indicate that the major motions of the system are occurring within the first few non-zero modes. However within the first 5 normal modes for instance, some of the large backbone fluctuations are missed in comparison with the RMSF when including

more modes (figures 5.14-5.17). Although, the major global motions would be described by a few modes, the subtle differences in the backbone motion by the inclusion of more modes, contributes to the mobility of the allosteric region.

The Amber package tool (*114*), PTRAJ was used to generate the root-mean-square fluctuation data (RMSF) for the normal modes, where the MSF in normal mode analysis is defined by:

$$< x_i^2 > = \frac{k_B T}{m_i} \sum_{j}^{modes} \left(\frac{\boldsymbol{W}_{ij}}{\omega_j} \right)^2 \tag{5.1}$$

Where m_i is the mass of the atom corresponding to the i^{th} degree of freedom, k_B is the Boltzmann's constant, T is the temperature, and $\boldsymbol{W}_{ij}$ is the i^{th} component in the eigenvector with frequency ω_j. The sum only goes over relevant modes (*258*).

The RMSF plots demonstrate subtle sensitivity to the presence of either or both ligands at the active and allosteric sites. In particular, in NMA (B), in the presence of only glucose, an increased mobility is observed in the loop connecting the two domains in comparison with other three NMAs, consistent with MD simulation 'B'. This increases the confidence in the application of NMA to a holo system, such as for example, GLK_AZ studied in the following chapter, where the protein structure is in complex with glucose, and we aim to predict the allosteric binding site from a static x-ray conformation in which the allosteric binding site is absent.

For clarity, the RMSF plots only include the Cα atoms, but the analysis has been carried on the full atomistic system. For the purpose of our study, primarily we need to establish if the loop connecting the two domains can be displaced by the normal modes.

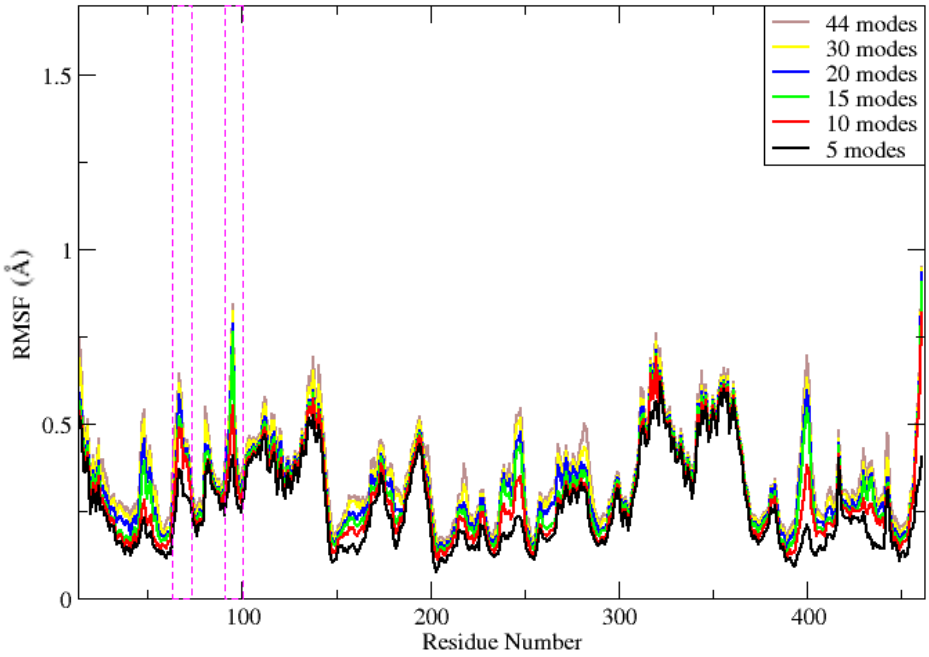


Figure 5.14: An overlay of RMSFs (Cα) for the normal modes for NMA (A) in the presence of both the glucose and the allosteric activator. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site.

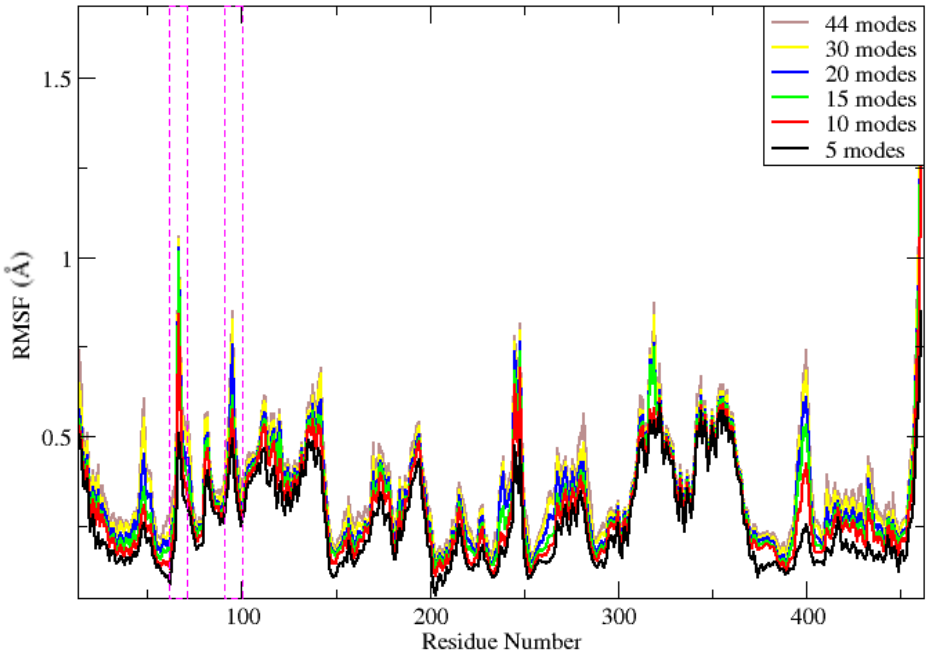


Figure 5.15: An overlay of RMSFs (Cα) for normal modes for NMA (B) in the presence of glucose. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site.

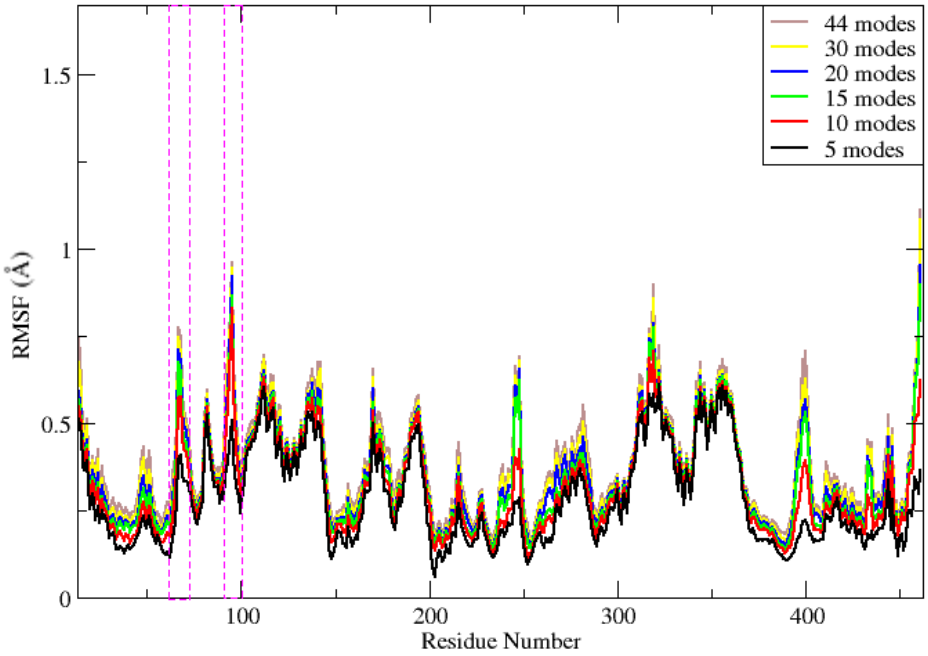


Figure 5.16: An overlay of RMSFs (Cα) for the normal modes for NMA (C) in the presence the allosteric activator. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site.

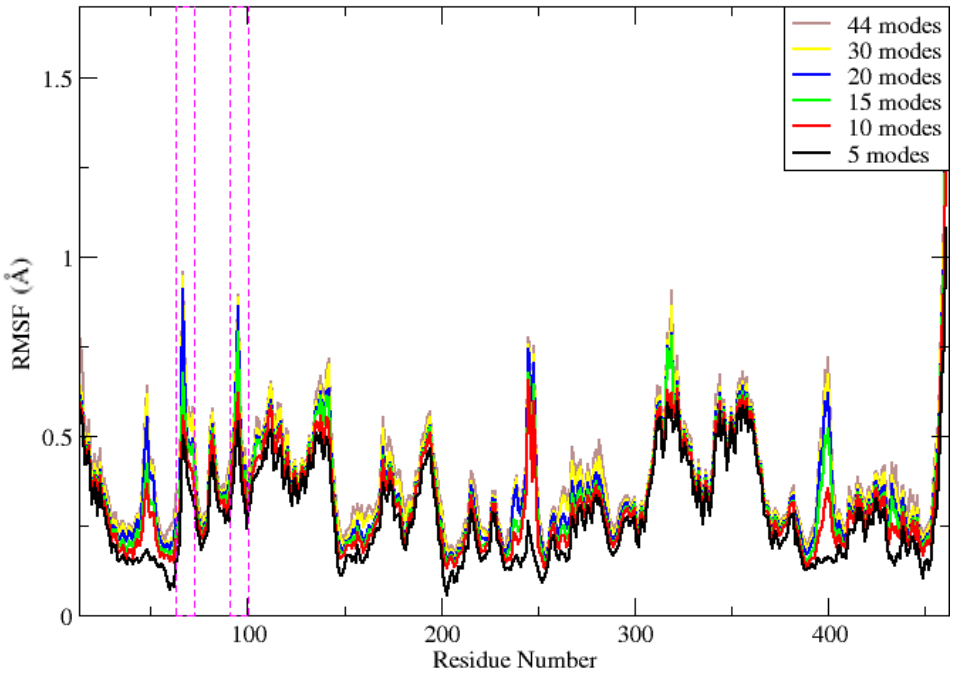


Figure 5.17: An overlay of RMSFs (Cα) for the normal modes for NMA (D) in the absence of both the glucose and the allosteric activator. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site.

The usual advantage of the use of normal modes is the notion that a few normal modes should be sufficient to capture the global motions of the system. In GLK, a few normal modes would correctly capture the opening and closing of the two domains, however in this case the conformational changes of the allosteric region do not occur within the first few normal modes. The MD simulations demonstrated that the connecting loop and β-turn, in close proximity to the allosteric site can be highly mobile in the presence of glucose at the active site. This has been supported by the normal modes, but requires the inclusion of more than a handful of modes, where ~ 20 modes appear to capture the majority of the backbone mobility.

The RMSF plots (figure 5.14-5.17) indicate that in this system, after the inclusion ~20 modes, the RMSF profile of Cα does not significantly vary by including more. Without prior knowledge, it would be challenging to establish which modes would be required to allow the investigation of additional binding sites, such as the allosteric binding site in GLK. The inclusion of at least all those modes that contribute to the backbone mobility gives the confidence that all such fluctuations are taken into account. Obviously in an ideal situation all degrees of freedom would be included but this is computationally too expensive at an atomic level of detail.

Figure 5.18 depicts the location and the degree of mobility captured by 1, 5 and 20 modes in NMA (B), generated by the normal modes visualisation tool NMDISPLAY (unpublished by Chris Moth), as a sister program to MDDISPLAY (*259, 260*). The frames are generated by converting the eigenvalues in the normal modes output file into displacements assuming 300 K temperature. The frames are arbitrarily generated by assigning a timestep of 1ps as the frame counter advances.

Although the opening/closing of the domains are observed, hardly any mobility is observed in the allosteric region in mode 1 (first non-zero). The inclusion of 5 modes captures some of the mobility in the loop connecting the two domains. The mobility of the allosteric region is intensified with 20 modes. We can use this information to look at the GLK_AZ structure, to establish if it is possible to predict the allosteric binding site along the normal modes.

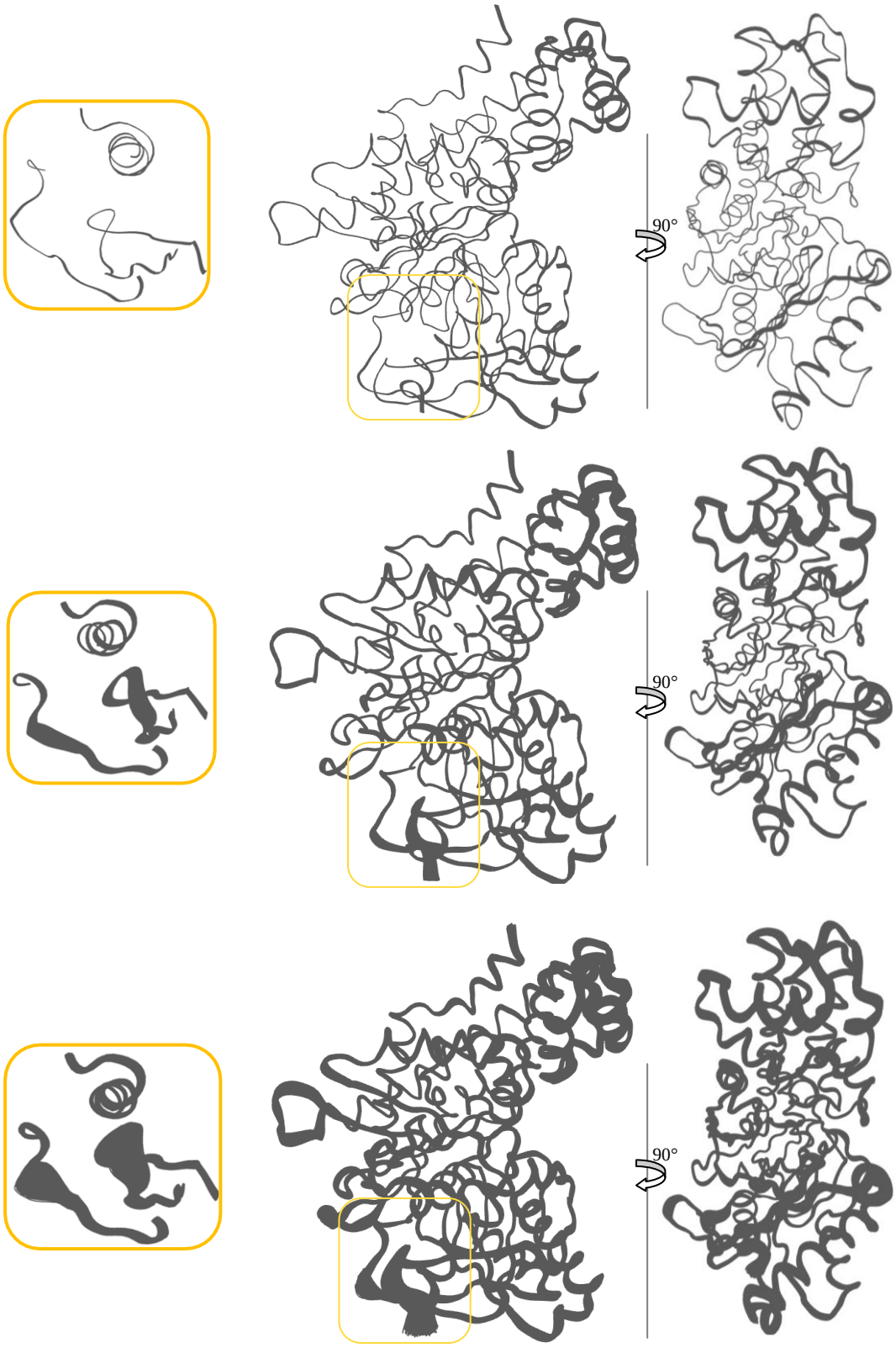


Figure 5.18: On top, the 1st non-zero mode for NMA (B), 1v4s in complex with glucose. On the left, the allosteric region, zoomed in. In the centre, the entire protein and on the right, the protein rotated by 90°. The 2nd row corresponds to 1st 5 modes combined and the 3rd row to the top 20 combined modes.

To better establish the relevance of each mode to motion near the allosteric binding site, the Cα RMSFs plots of the loop residues have been illustrated in figure 5.19. The mobility of the connecting loop is crucial to the allosteric binding sites opening. From hereafter, the mode numbering refers to non-zero modes, starting from 1 to 44. For clarity, only modes that demonstrated an RMSF of ≥ 0.1 Å has been displayed in figure 5.19.

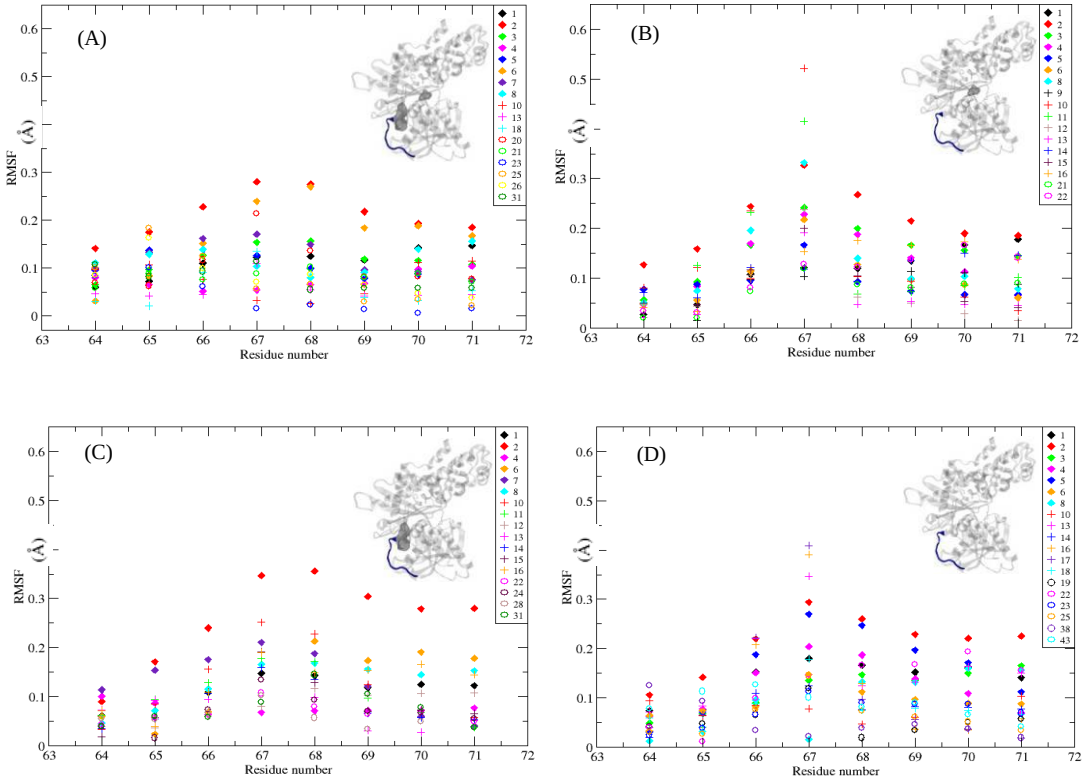


Figure 5.19: Cα RMSF for residues in the loop connecting the large and small domains in the four NMAs. Only normal modes where at least one residue has an RMSF ≥ 0.1 Å have been displayed for clarity.

As expected, the loop does not move significantly in the presence of the allosteric ligand (NMA A & C). Residues 67-71 which are not involved in direct interaction with the allosteric ligand move along mode 2 in all four NMAs referring to a twisting motion of

the two domains, however the loop just translates (data not shown) and the net effect of this move with other parts of system does not change the nature of the binding site. In the absence of the allosteric ligand from the binding site (NMA B & D), more fluctuation is observed in residue 67, along higher frequency modes. In contrast to residues T65 and G68 nearby, that can undergo activating mutations, E67 has not been shown to undergo naturally occurring mutations. The stability of residues at either side of E67 may stem from the role that these residues play in stabilising the active state closed form of the enzyme by undergoing activating mutations.

As, in all four NMAs, the starting structure was a structure which already encompassed the allosteric site and, a large change is therefore not expected in this region. Although it has been demonstrated that NMA is sensitive to the presence of the ligand and correctly highlights the increased motion of the loop in the absence of the allosteric ligand from the binding site, the above observation indicates that we cannot limit the allosteric binding site search to the first few modes in a closed-state starting structure. In addition, it is likely that the side chain movements will also contribute to the allosteric site opening.

In NMA (D), where both ligands have been removed from the active and allosteric binding site, a similar RMSF profile to NMA (B) is observed in residue of the connecting loop (figure 5.19). It is notable that in the absence of both ligands, the same level of fluctuation in the loop residues, in particular residue 67, has been shifted to slightly higher frequency modes, from 8, 10 and 11 in NMA (B) to 13, 16 and 17 in NMA (D). The presence of glucose at the active site in NMA (B) appears to contribute to promoting the motion of this residue to a lower frequency.

Within these 44 modes, residues in the loop show a dependence on the presence of the glucose or the ligand. By nature, NMA explores the motion around one stable conformation; therefore we cannot expect to observe motions that may be linked to the global conformational changes along the allosteric transition from the super-open to the closed state. In the absence of the allosteric ligand, it is expected that residues in this region adopt a compact conformation, to stabilise the closed state.

In the following section the nature of the allosteric binding site is explored by binding site search along the normal modes in NMA (B).

5.3.2 Binding site search along the normal mode eigenvectors

As the starting structure of all four NMAs is the X-ray structure of a closed state GLK, bound to glucose and the allosteric activator (PDB ID: 1v4s), considering the relatively small fluctuations along the first 44 normal modes, we would not expect a significant change in the allosteric region along the normal modes. However, in NMA (B) where the allosteric ligand was removed prior to minimisation and NMA, the minimisation process reduces the size of the allosteric binding site to adopt a more compact conformation, in particular by moving the loop connecting the two domains (figure 5.20). Binding site search on both the starting structure and the minimised structure reveals that the size of the allosteric site has decreased and the ranking of the site has been demoted to rank 5 (figure 5.21) using both binding detection tools with respect to the x-ray structure. Although Pocket-Finder also demonstrates a reduction in the size of the pocket, it still ranks the allosteric binding region, as the 2nd ranked binding pocket (figure 5.21). The binding site profile of the starting structure (x-ray structure) prior to minimisation is that demonstrated in figure (5.9).

Figure 5.20: an overlay of the allosteric region in the minimised NMA (B) structure in green and the corresponding original X-ray structure of the closed state glucose and allosteric activator bound structure (PDB ID: 1v4s) in grey. The allosteric ligand is in stick representation with a transparent van der Waals surface over it.

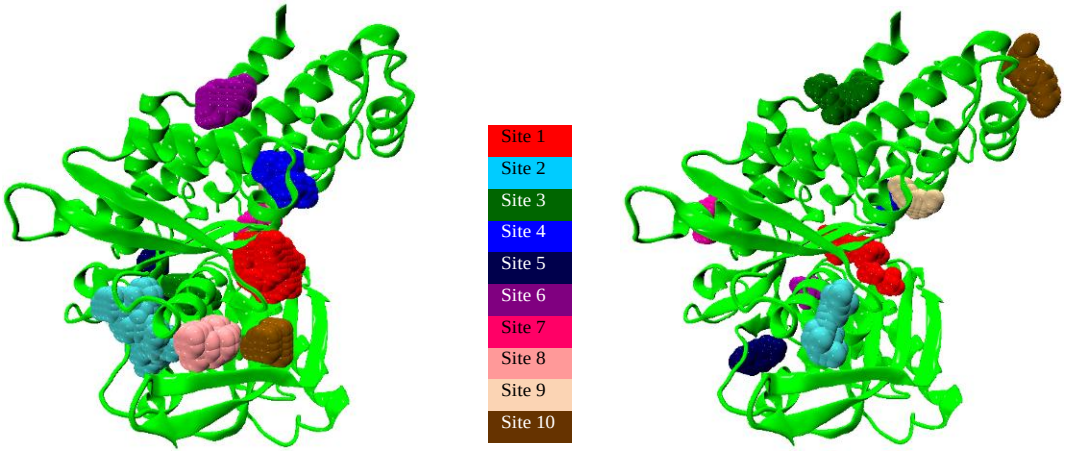


Figure 5.21: On the left, top 10 sites predicted for the minimised structure in NMA (B) by Pocket-finder, and on the right the corresponding by Q-SiteFinder. In the centre is the colour code for the sites in order of rank.

Binding site search on a few frames at the extreme of the conformational range away from the minimised structure from modes 1-20 of NMA (B), demonstrates that there is not a considerable conformational change in the allosteric region, to allow a significant improvement in the binding pocket prediction with respect to the minimised structure. Nevertheless, a frame at the extreme of the range of conformations found along the first 20 modes leads to a small increase in the size of the pocket predicted at the allosteric site by Pocket-Finder, retaining the 2nd ranked position. With Q-SiteFinder, the pocket volume increases in size and improves ranks from 5th to 3rd. The allosteric site pocket volume for the minimised structure is predicted as 250 Å^3 and 173 Å^3 by Pocket-Finder and Q-SiteFinder, respectively. The site volume is increased to 282 Å^3 with Pocket-Finder and to 215 Å^3 with Q-SiteFinder, in the extreme conformation from modes 1-20. This is in contrast to the X-ray starting structure (PDB ID: 1v4s) allosteric site prediction of 2nd ranked by Pocket-Finder with a volume of 461 Å^3, and ranked 1st by Q-SiteFinder with a volume of 513 Å^3.

For comparison, the active site, reduces in volume from the 572 Å^3, predicted by Pocket-Finder for the starting X-ray structure (PDB ID: 1v4s) to 396 Å^3 in the minimised structure and increased again along the 20 normal modes to 425 Å^3 in NMA (B). Despite the presence of glucose at active site in NMA (B), in the heavily minimised structure, residues near the active must contract around glucose.

In contrast in the MD simulation of the close state conformation (PDB ID: 1v4s) in complex with glucose only (simulation B), the allosteric binding site volume did fluctuate, demonstrating the flexibility of the allosteric region in the presence of glucose at the active site (figure 5.13).

The vibrations around a heavily minimised structure may not capture the full mobility of the allosteric region within the first 20 modes, however the application of NMA to the identification of the allosteric binding site cannot be ruled out, owing to the significantly shorter calculation time in comparison with MD.

We may be able to use this information in the study of the AstraZeneca structure, where we aim to reveal the allosteric binding site.

5.4 Summary of closed-state GLK

Treating the study of the closed state glucose and allosteric activator bound X-ray structure as a benchmark, the dynamics and flexibility of the system was studied with MD simulations and normal mode analysis.

MD simulation conditions were sufficiently accurate to be sensitive to the presence of either ligand at their respective binding sites. The system was most restricted in mobility in the presence of both ligands at the active and allosteric binding sites. In comparison to the reference simulation (‘simulation A’), ‘simulation B’ in the presence of glucose only, demonstrated small increase in mobility in the small domain. Residues in the connecting loop appear to have even become more restricted. Binding site search along this simulation shows that the allosteric site is slightly reduced in size after the initial minimisation and MD simulations, but does increase in size to the full size in some frames of the simulation, which is promising in being able to predict the allosteric site in the AstraZeneca structure.

In ‘Simulation C’ in the absence of glucose at the active site, the small domain is more mobile. In contrast residues in the β-turn of the β-sheet (residues 91 to 100), near the allosteric site becomes less mobile. There may be a link between the presence of glucose at the active site and the mobility of this region, contributing to the opening of

the allosteric binding site. The other end of this β-sheet is very close to the active site, which appears to be stable in the presence of glucose at the active site, and the other end closer to the allosteric site then seems to fluctuate more. In the absence of glucose (simulations C& D), the β-turn closer to the active site (residues 120-129) has the freedom to move, which appears to reduce the mobility of the β-turn (residues 91-100) closer to the allosteric site, in comparison with that in simulation B. This further strengthens the hypothesis that this β-sheet may play a key role in the allosteric binding site opening along with residues of the loop connecting the two domains.

Binding site search along the MD simulations appears reliable in predicting either binding site well and ranks them highly when expected, which can help in predicting unknown binding sites.

NMA studies have shown sensitivity to the presence of either ligand at their respective binding sites. It has been demonstrated that a few normal modes would not capture the total backbone flexibility. This is an issue, as the allosteric binding site opening is dependent on the movement of a flexible loop and not just side-chain rearrangements. By including ~20 modes in this system, most of backbone conformational fluctuations are captured.

Binding site search on the minimised NMA (B) structure, only in complex with glucose, further illustrates that in the absence of an allosteric activator the connecting loop adopts a slightly more compact conformation, reducing the size and access to the allosteric binding site. However, the less sensitive Pocket-Finder algorithm can still identify a 2nd ranked pocket at the allosteric site, although the size of the binding site is reduced. Q-SiteFinder identifies a 5th ranked site with significantly reduced size in comparison with that of the X-ray structure. Residues at the beginning of the connecting loop remain fairly rigid along the normal mode deformations, consistent with observation in the MD simulation. Of those, T65 is known to undergo activating mutations, stabilising the active close state GLK. Despite the rigidity of the residues at the beginning of the loop, later residues of the connecting loop demonstrate mobility along the normal modes. Binding pocket search on the most open conformation of the connecting loop of the 20 normal modes does improve the size and the rank of the allosteric binding site, but the

size and ranking remains lower than that of the X-ray structure. This information may be useful in the study of the GLK_AZ structure.

In the following section, the study of the super-open state GLK (PDB ID: 1v4t) will be discussed. According to the pre-equilibrium kinetic model of this system, it should be possible to observe conformations of the system that could bind glucose, even in the absence of the ligand itself. Simulation lengths of a few 100 ns are not sufficient to see an intermediate conformation between the super-open and closed state, but it may be possible to see the partially formed binding sites (active and allosteric) in some conformations of the protein in the simulation ensemble. NMA of this system would not be useful in the study of the allosteric binding site, as this structure is in a completely different conformational space and NMA would not capture any motion other than small fluctuations around that local minimum. GLK has shown several stable forms along the transition between the super-open inactive to the closed active state. By definition, NMA will only calculate the systems' fluctuations within one local minima and cannot overcome energy barriers associated with several stable states along the energy potential surface. Calculating the NMA too close to the fully closed state structure, does not offer the conformational variations that the system may experience.

5.5 Inactive super-open form GLK

The GLK super-open state is the more stable conformation in the absence of glucose, binding to glucose with low affinity in comparison with the high glucose affinity closed-state GLK. To date this is the only apo crystal structure of the human GLK available. Despite the low crystallographic resolution (~3.5 Å), and missing residues (157-179), this structure was studied in addition to the active form, to gain better insight of the allosteric region. Recent kinetic studies (*218*) have indicated that glucose binding may not be a prerequisite to allosteric activator binding, and therefore it may be possible to observe conformations along the simulation that could bind glucose or the allosteric activator. However, a simulation length in the hundreds of ns may not be sufficient to observe such slow conformational changes.

In previously suggested kinetic mechanisms (*233*), an induced-fit model was postulated upon glucose binding, shifting the equilibrium from the more stable super-open to the active closed state. It was believed that glucose would primarily bind to residues in the large domain (in particular D205) and induce the large conformational change. Although these models have now been contradicted with the recent kinetic studies (*218*), simulation of this system should give further insight in understanding the mechanism of binding in this system.

For the super-open inactive structure (PDB ID: 1v4t), two simulations were set-up. This was due to the modelling that was required to place a number of missing residues in the structure (residues 157-179). Two distinct conformations of the modelled loop were selected as starting structures.

Simulations	Super-open state (PDB ID: 1v4t)
Model 1	First conformation of the modelled missing residues
Model 2	Second conformation of the modelled missing residues

Table 5.4 Simulations set-up for the super-open state X-ray structure depicted in figure 4.2.

5.5.1 System preparation and MD set-up

The *apo* crystal structure of human GLK was obtained from the Protein data Bank (PDB ID: 1v4t). The missing residues (157-179) were modelled externally by Richard Ward (*219*) using the Schrödinger PRIME tool (*261*). Two conformations of the missing loops were generated, the first (model 1) based on the default setting of the loop sampling and the other (model 2) with the ultra extended and in each case, the lowest energy conformation was selected (Figure 5.21).

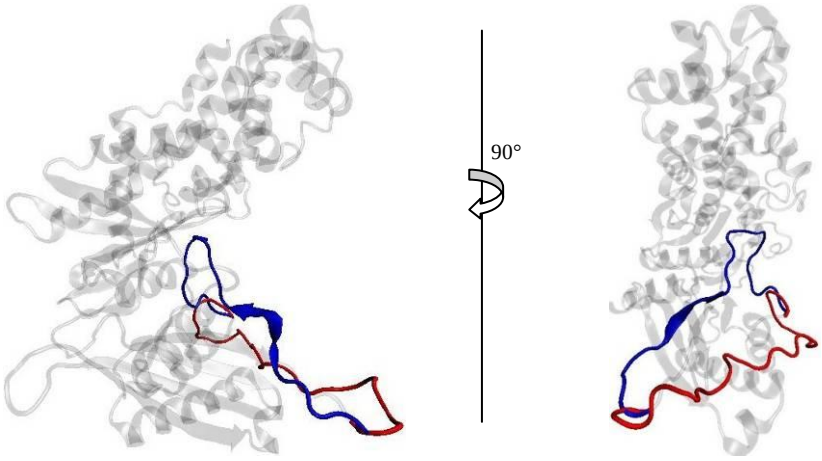


Figure 5.21: X-ray structure of the inactive super-open GLK. The missing residues (157-179) were modelled with PRIME, highlighted in red and blue, for the default conformation and an optimised conformation, respectively. On the right, a side view.

The modelled loop residues in "model 2", highlighted in blue (figure 5.21), adopt a very different conformation to "model 1". Of those, residues 168-169 are involved in glucose interactions in the active state, and the loop adopts a similar conformation to that in the closed active form (Figure 5.22).

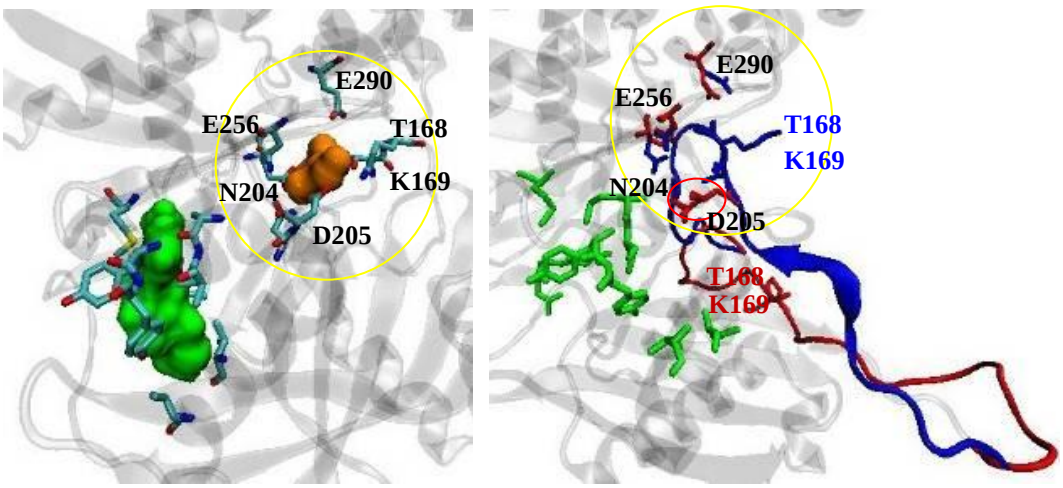


Figure 5.22: On the left, residues involved in interacting with glucose and the activator, ligands represented in orange and green surfaces, respectively, for the active closed state X-ray structure (PDB ID: 1v4s). On the right, the two modelled conformations of the super-open form. Residues involved in glucose interaction in the closed state are highlighted in red or blue stick representation circled in yellow. On the right allosteric activator interacting residues are coloured in green, and their conformations remain the same after optimisation of the missing loop. Circled in red is the location of D205 (as they are in the same space, D205 in model 2 has been covered with D205 from model 1). Note that T168 and K169 are in two different locations in the two models, where the residue labels have been accordingly coloured.

The residue positional difference between the closed (PDB ID: 1v4s) and the super-open structures (PDB ID: 1v4t) have been highlighted in figure 5.23. The residue displacements were calculated between the two states by treating the closed state structure (PDB ID: 1v4s) as a reference and calculating the residue displacements in the super-open state with respect to the reference state (closed state). The major differences are in residues of the small domain (65-200).

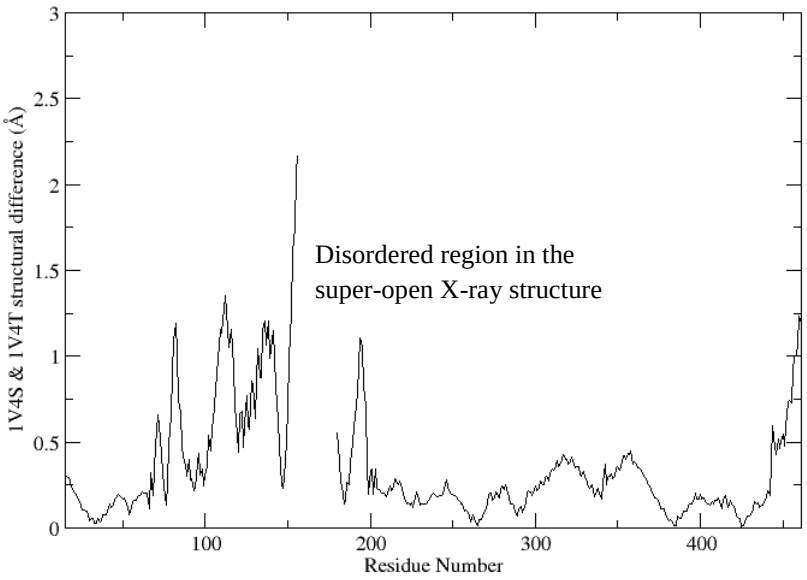


Figure 5.23: structural difference between the close-state X-ray structure (PDB ID: 1v4s) and the super-open (PDB ID: 1v4t) GLK structure.

Both structures were protonated using WHATIF (*248*). The Amber package tool (*114*), XLEAP was used to solvate the system with 20986 TIP3P water molecules, with a minimum distance of 12 Å from the protein atoms. In addition 21 sodium counter-ions were added to the system to neutralise the overall charge.

The same simulation set-up as the closed-state GLK described in section 5.1 was applied here. For both of the inactive super-open conformations, 100 ns of MD production was collected after removing the initial ~2 ns, assuming equilibration at the beginning of the NPT production run, based on stabilisation of properties such as energy, volume, density and RMSD.

5.5.2 Structural dynamic analysis

The crystallographic B-factor values for the Cα atoms in the structure of the super-open inactive form reveal that the enzyme is highly flexible. Residues 157-179 were

not resolved in the crystal structure (Figure 5.24). This can be an indication of a highly mobile region. Residues T168 and K169, involved in glucose binding in the active form are among the unresolved residues.

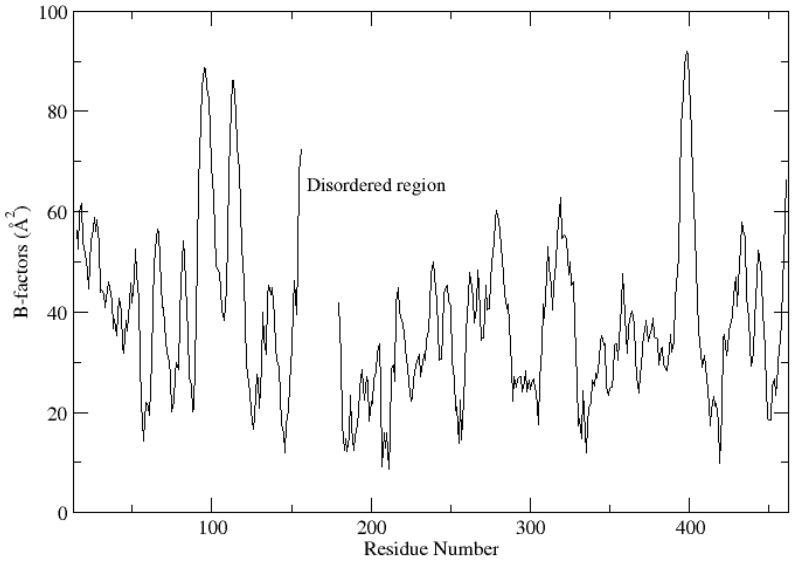


Figure 5.24: Crystallographic B-factors for the super-open inactive GLK. (PDB ID: 1v4t). The missing values refer to a disordered region, residue 157-179.

As expected the modelled missing residues show a large degree of mobility during the 100 ns simulations, more so in 'model 1', coloured in red in the RMSF plot (figure 5.25).

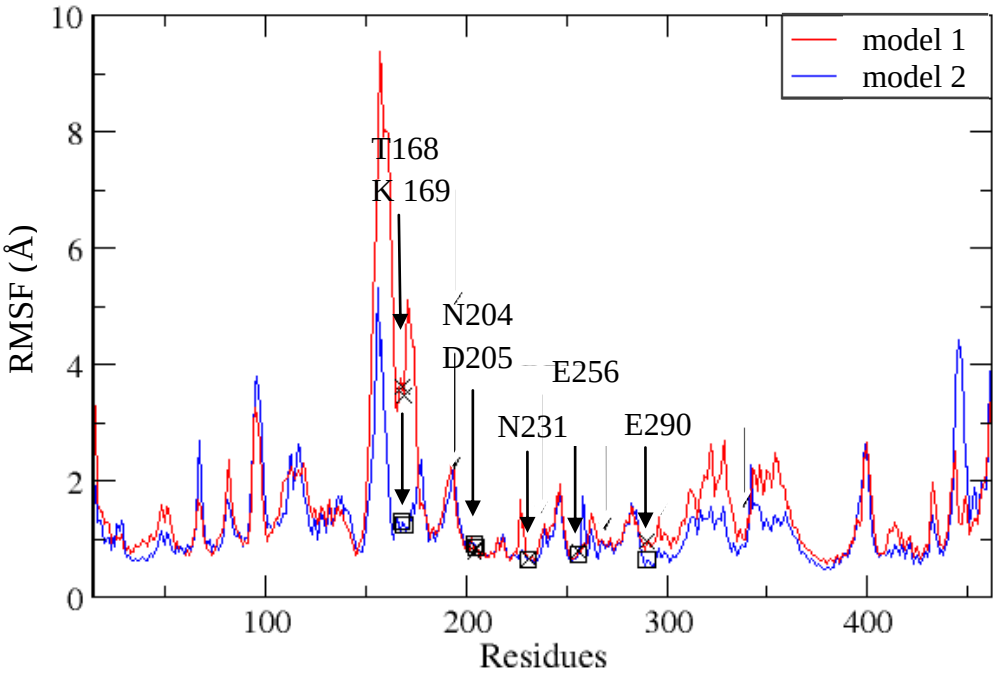


Figure 5.25: RMSF for the two 100 ns simulations. Colours red and blue refer to 'models 1 and 2'. The x and □ refer to the residues in 'model 1' and 'model 2' respectively. These residues interact with glucose in the closed state GLK.

The root-mean squared fluctuations in the two 100 ns simulations follow a similar pattern to the crystallographic B-factors. In model "1" the modelled missing residues (157-179) appear much more mobile than in that of model "2". This could be due to the fact that in "model 2", the loop adopted a conformation which led to some residues occupying a similar space to glucose in the active form, thereby stabilising the loop (figure 5.26). A recent study of the super-open state dynamics with GNM, highlighted the fact that residues involved in interacting with glucose in the closed state, N204, D205, N231 and E256 remain fairly stable in the super-open state, apart from T168 and K169 that are in the unresolved region (*262*). In 'model 1' residues of the disordered region begin to adopt a more compact conformation and move closer to the active site region. In 'model 2' these two residues are already at the active site, and therefore are very stable throughout the simulation. The conformation of residues in this region could be that of an intermediate conformation of the system along the equilibrium between the super-open and an open state.

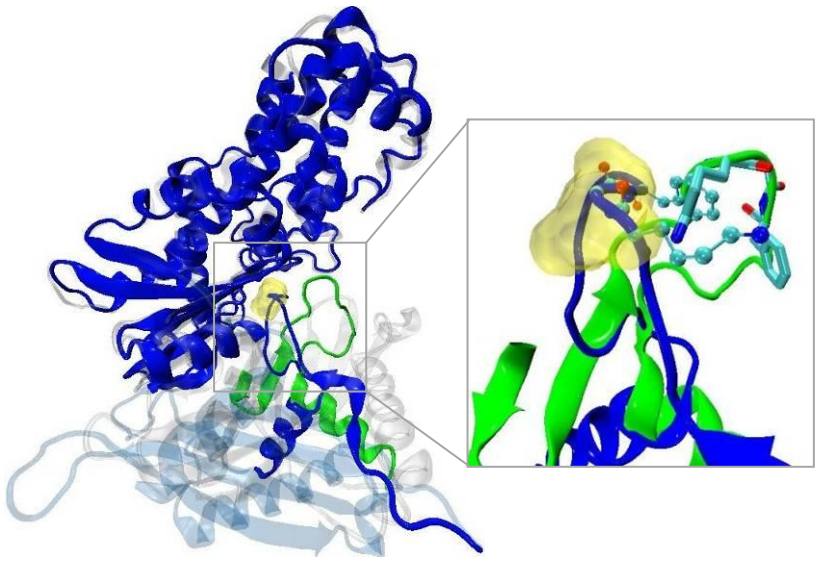


Figure 5.26: An overlay of the GLK super-open "model 2" starting structure (PDB ID: 1v4t) and the closed active form (PDB ID: 1v4s). Small domain residues are transparent in both (super-open: blue, closed: grey), except for residue that were modelled in the super-open structure (blue) and the corresponding in the closed state (green). The yellow surface represents the location of glucose in the closed form. In the close-up image, in addition, residues 169 to 171 have been displayed, represented in 'stick' and 'ball & stick' for the closed form and 'model 2', respectively.

It has been suggested (*262*) that the super-open inactive state is stabilised by salt-bridges D205-R447 and E216-K458, which are both broken upon glucose binding. The two suggested salt-bridges, D205-R447 and E216-K458, have been monitored in the two simulations (Figure 5.27). For the purpose of the plot, the distance between the carbon atom of the carboxyl side chain of D205 and the carbon of the guanidine group of R447 was measured. For E216, the carbon of the carboxyl side chain was used and for K458, the nitrogen of the amino group was used in the measurement. In "model 1", both salt bridges are fairly stable. In addition, at the same time a very slight distance change is also observed in the D205-R447 salt-bridges (Figure 5.27). However, in 'model 2', the D205-R447 salt-bridge has disappears at a very early stage of the simulation, which points to a structural conformation in the trajectory (model 2) that has moved away from the super-open conformation, although E216-K458 remains stable. As it has been suggested that glucose may first associate with D205 in the large domain before fully binding, the destabilisation of this salt-bridge does suggest that there may be

conformations present that resemble an intermediate state in the shift towards the closed conformational state.

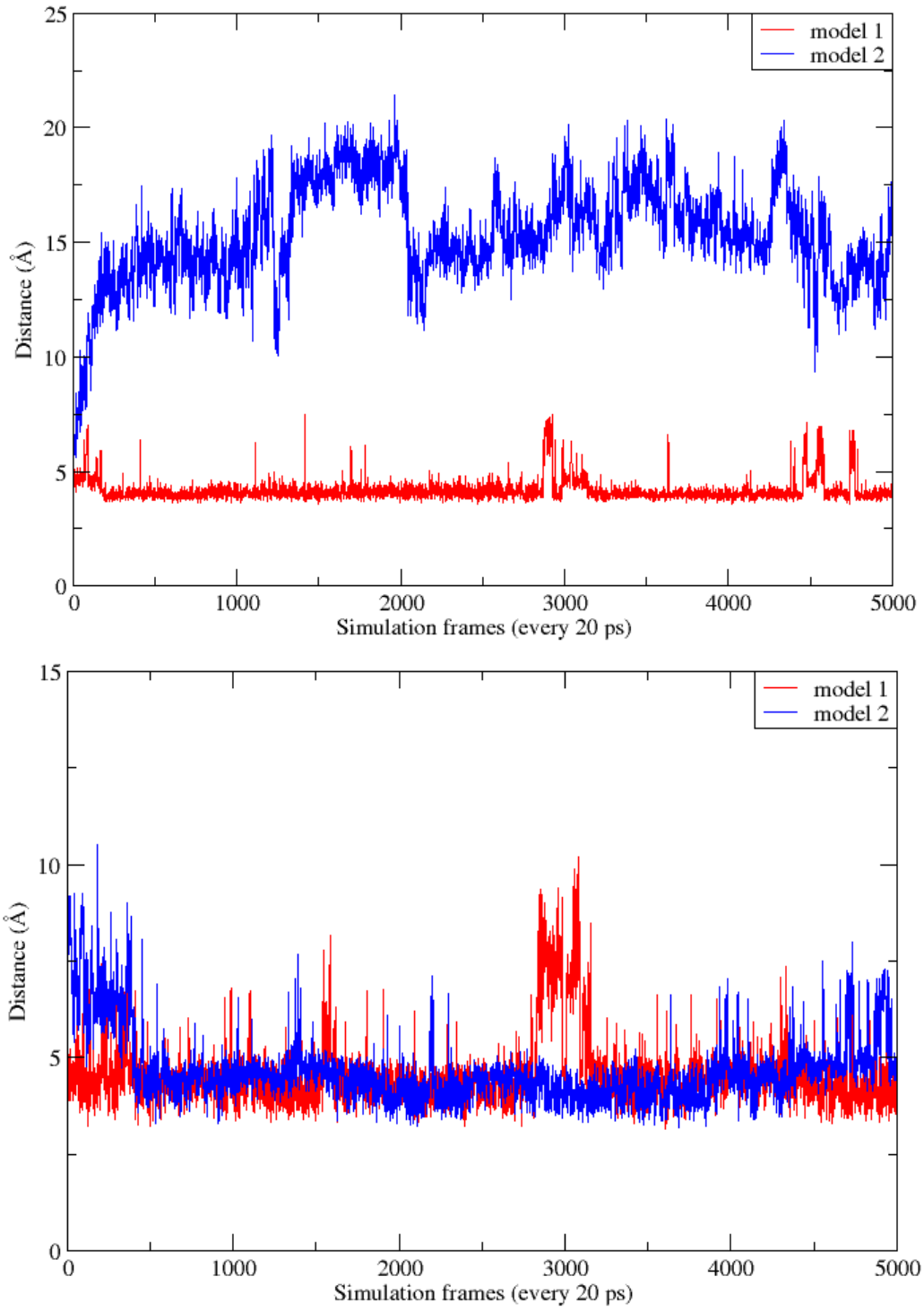


Figure 5.27: Distances in residues believed to be involved in stabilising salt-bridges of the super-open conformation, D205-R447 and E216-K458.

A binding site search was carried out on frames from both simulations to establish whether in those conformations that may resemble an intermediate state, either a full binding site or a partial site would be observed at the active or allosteric sites.

5.5.3 Binding site profiles through the simulations of the super-open state

In the closed form, glucose interacts with residues N204, D205, E256 and E290 in the large domain and residues T168 and K169 in the small domain. In the super-open state, residues T168 and K169 are among the unresolved loop residues which have been modelled. In the super-open state, the small domain rotates away from the large domain by ~ 99°, hence it is expected that residues of the small domain, involved in glucose interaction are too far away from the active site. Point-mutations have demonstrated that in the super-open conformation glucose binds to residues in the large domain (*262*).

The X-ray structure (PDB ID: 1v4t) and both the starting structures with the modelled loop were analysed using both pocket identification tools. There are some differences in the overall rank profile of the predicted sites before and after modelling the missing loop and between Q-SiteFinder and Pocket-Finder.

The original X-ray structure with missing residues shows different binding site profiles using the two methods (Figures 5.28-5.29). Although it would not be expected to pick up the glucose or the allosteric binding site in the super-open state, it is interesting to study the profile through the simulations. It is believed that glucose initially associates with residues in the large domain, in particular D205 (*262*).

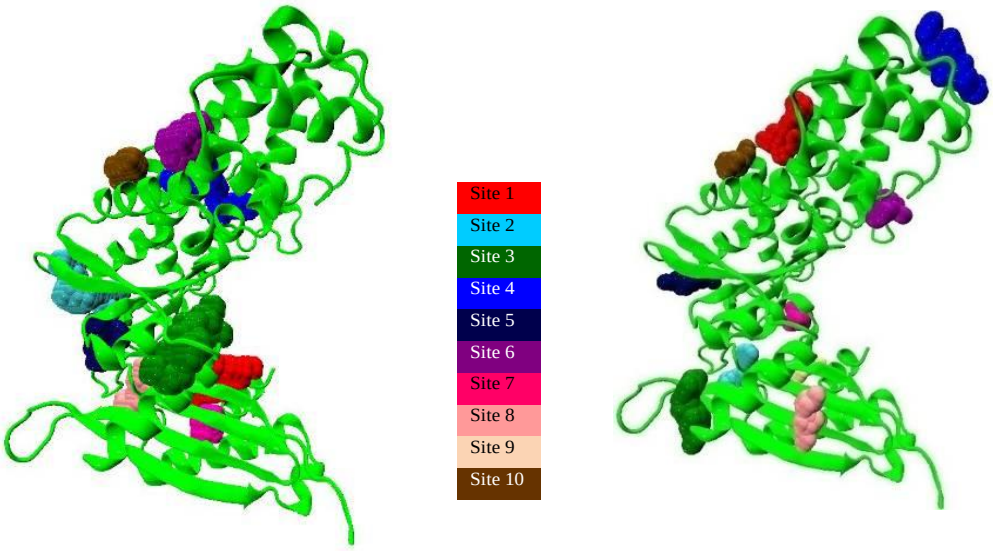


Figure 5.28: On the left, top 10 sites predicted for the super-open X-ray structure (PDB ID: 1v4t) in green, by Pocket-Finder and on the right the corresponding by Q-SiteFinder. In the centre, the colour code for the sites in order of rank. The missing residues have not been modelled in this case.

In the X-ray structure of the super-open state (PDB ID: 1v4t), a site involving D205 has been picked-up with Pocket-Finder, however not with Q-SiteFinder (demonstrated in figure 5.29). The number of glucose or allosteric activator interacting residues (in the active state), have been monitored in the super-open X-ray structure and the two super-open simulation starting structures where the missing residues have been modelled, 'model 1' and 'model 2'.

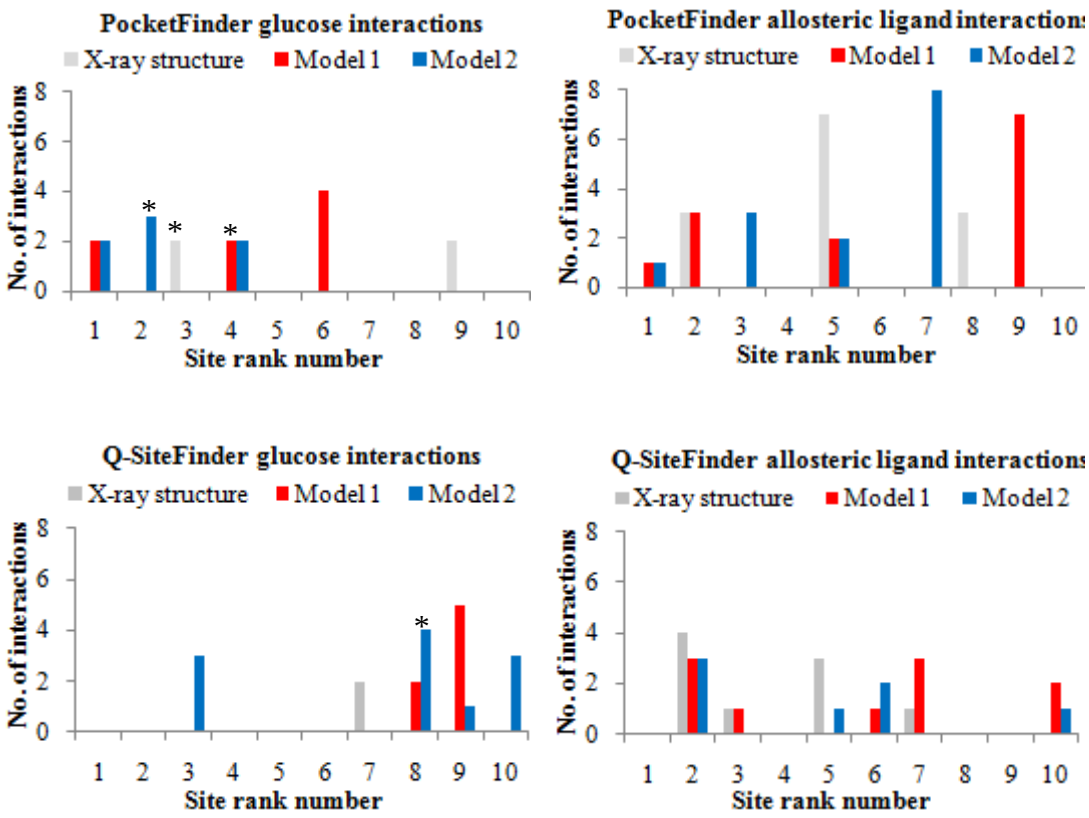


Figure 5.29: Glucose and compound A interaction residues identified in the corresponding GLK closed form, identified for the super-open form (PDB ID 1v4t), 'model 1' and 'model 2'. On top PocketFinder was used to search for sites in the original X-ray structure of the super-open GLK and the two simulation starting structure. On the bottom, the corresponding with Q-SiteFinder. The (*), indicates the presence of D205 in the identified binding site.

It is interesting that within the top 10 binding sites predicted by the two methods; at least a few of the active site residues are included (figure 5.29). Q-SiteFinder as expected ranks these sites lower. It is surprising that a number of residues pointing into the allosteric site are predicted by the two tools, although of those not all have been marked by the crystallographers to be directly involved in interaction with compound A in the closed active form (PDB ID: 1v4s), but at least there is an indication that a small cavity does exist involving the allosteric residues (figure 5.29).

The spread of the predicted sites for the two models are depicted below (Figure 5.30). The repair of the missing residues does, as expected, influence of the distribution of the

predicted sites, and the modelled residues are now involved in interactions in predicted sites.

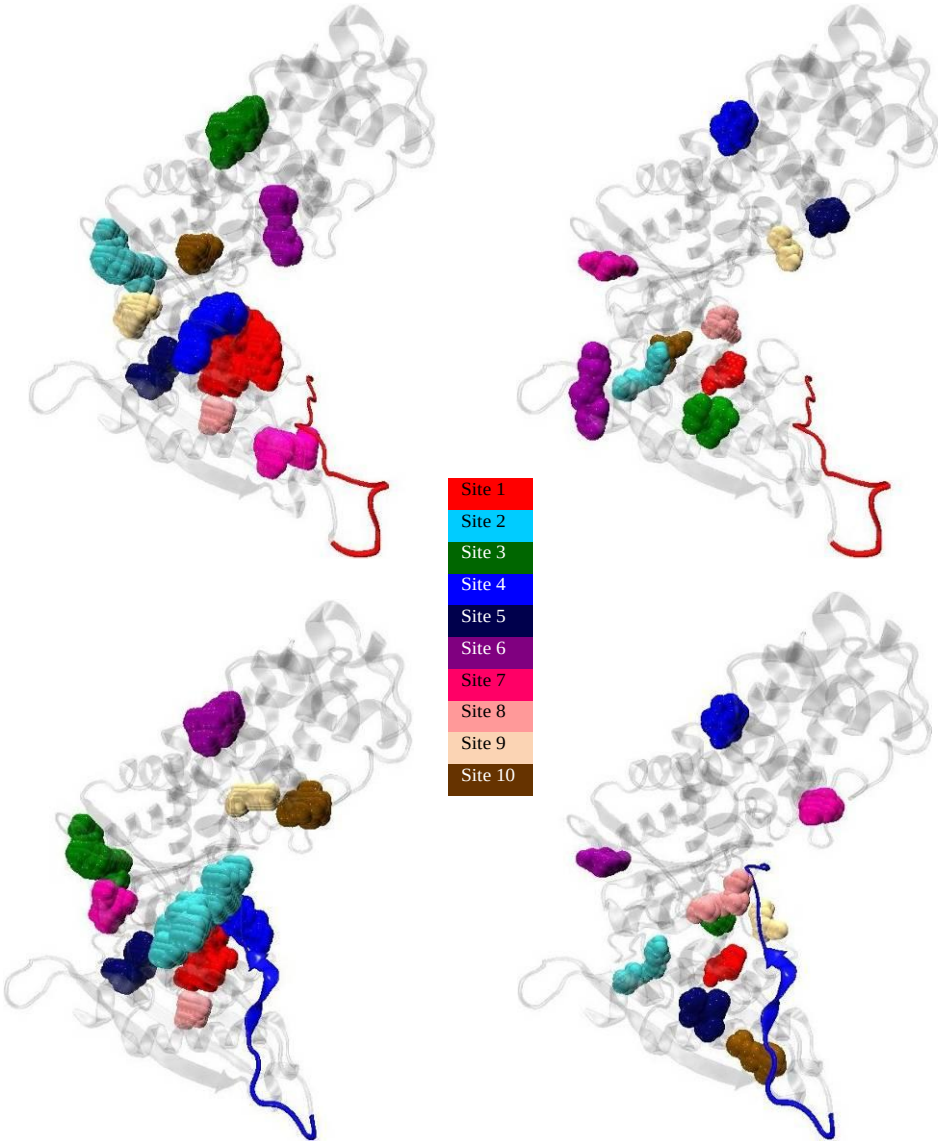


Figure 5.30: On top predicted binding sites for 'model 1', with Pocket-Finder on the left and Q-SiteFinder on the right. Below, the corresponding for 'model 2'. In grey, the protein where 'model 1' and 'model 2' are indicated by the red and blue modelled loops, respectively. The pockets are coloured in order of rank displayed in the table.

The two conformations of the disordered loop influence the ranking of the binding site in both binding site search methods. However, the sites are predicted in similar areas,

apart from the immediate vicinity of the active site, where the conformation of the modelled disordered loop does influence the region.

The predicted binding site profile was monitored in the two simulations of the super-open form, with ‘models 1 and 2’, as starting structures. Observations of destabilising salt-bridges, discussed earlier, indicated that conformations resembling an intermediate state could be present in the ensemble, where the glucose or the allosteric site might be detectable.

Each plot includes 25 equally-spaced frames along the corresponding 100 ns simulation on the horizontal axis while the number of relevant contact residues picked-up in each pocket is on the vertical axis. The pocket ranks are indicated by the colours. A total of 20 residues were classed as residues forming the active binding site, of which 6 do directly interact with glucose in the active form. The classification of residues are based on the combination of residues that Q-SiteFinder and Pocket-Finder, identified as pointing into either binding site (Table 5.2). The main purpose of using all residues is to not bias the data purely based on the particular activator bound to the crystal structure. Another ligand may interact differently, with additional or fewer interactions at the allosteric binding site.

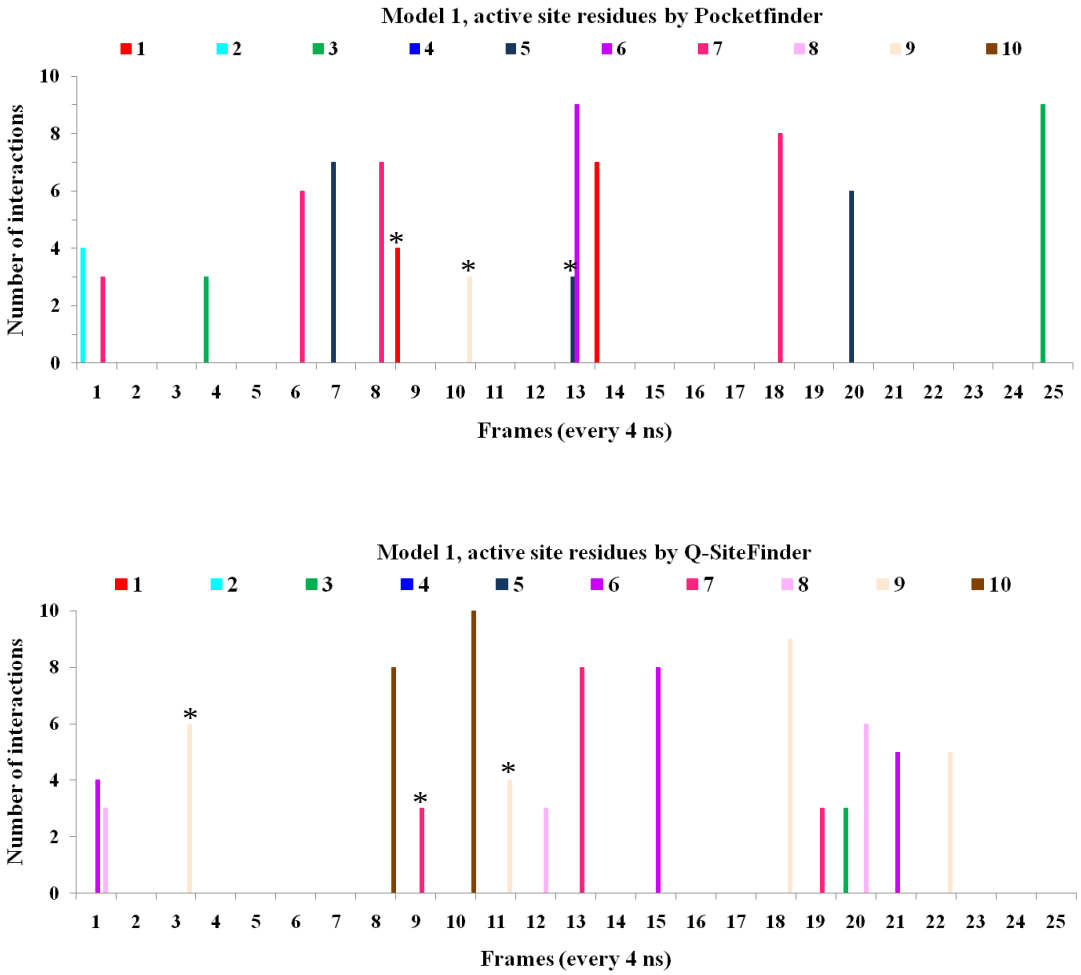


Figure 5.31: On top, predicted binding sites involving glucose interaction residues for 'model 1' with Pocket-Finder and below with Q-SiteFinder. On the horizontal axis, 25 frames from the 100 ns simulation. For each frame, 10 identified sites are coloured and numbered in order of rank. "*" indicates the presence of D205.

In the simulation of 'model 1' the glucose contact residues in the active form are detected in some frames, but are ranked fairly low. Q-SiteFinder does identify similar sites to PocketFinder, but at lower rank.

Residues forming the allosteric site in the active form have been monitored along the 'model 1' simulation. A total of 21 residues were classed as residues forming the allosteric binding site in the active form, of which 7 directly interact with the allosteric activator 'compound A'.

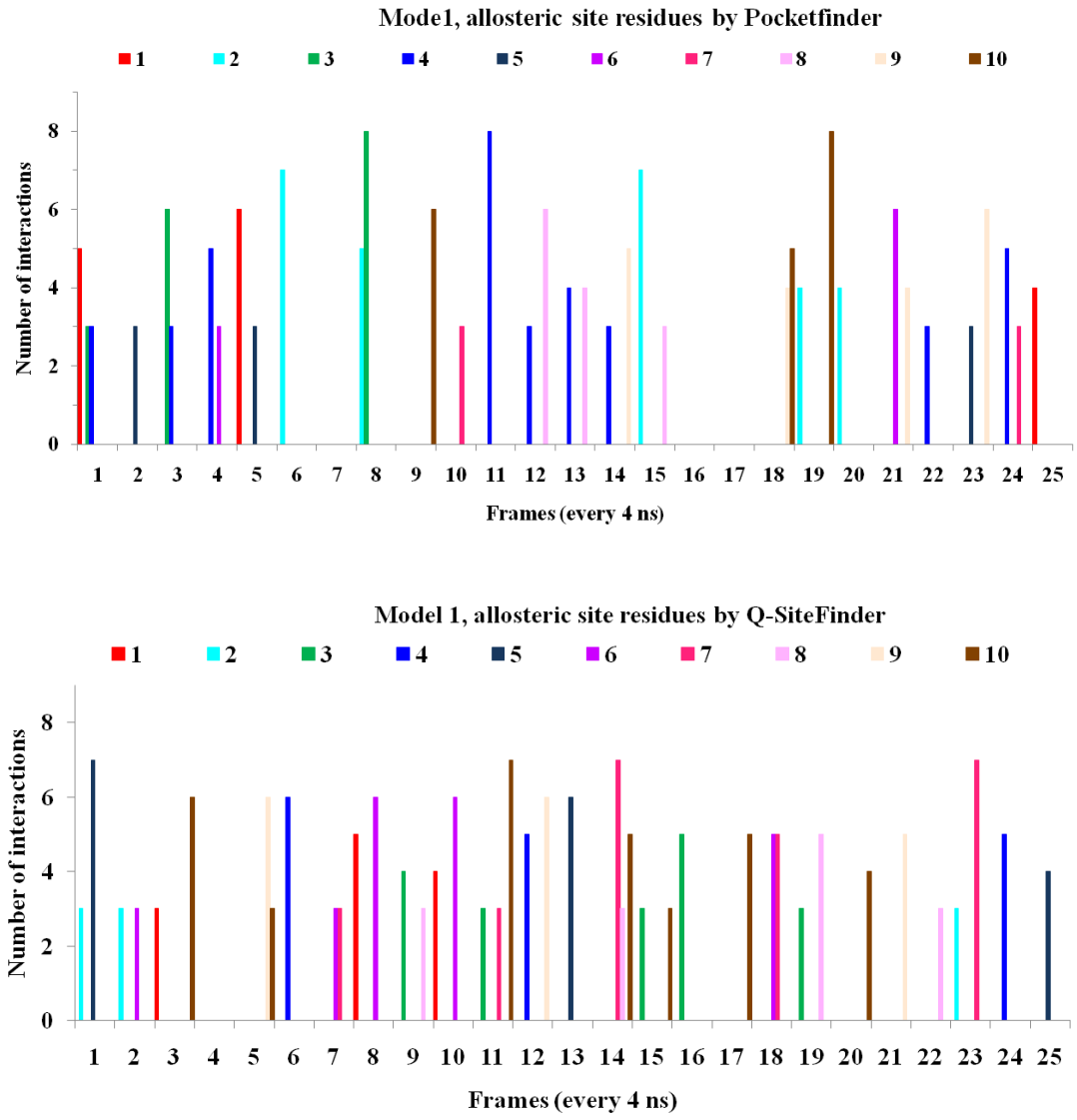


Figure 5.32: On top, predicted binding sites involving the allosteric activator interaction residues for 'model 1' with Pocket-Finder and below with Q-SiteFinder. On the horizontal axis, 25 frames from the 100 ns simulation. For each frame 10 identified sites are coloured and numbered in order of rank.

At most eight allosteric site residues are recorded in the predicted sites, although not always highly ranked. Owing to the large conformational change that occurs at the allosteric site region between the closed and super-open states, the presence of the full binding site would not be expected, although if a strong active site is formed, then we could imagine that the protein backbone could be adopting an intermediate

conformational state that may go on to form the full allosteric site. Similar to previous oberservations, it is clear that Q-SiteFinder splits the residues in both sites into more sites than does PocketFinder. The glucose binding site profile plots show a greater number of sites involving a small number of residues for each frame by Q-SiteFinder, in comparison with PocketFinder that has combined all the small sites into one strong large site which is then ranked highly for each frame. This is better observed in the following plots for 'model 2' (figure 5.33).

In 'model 2' the glucose active site is more apparent. Most frames include a top ranked site involving a large number of the residues that form the active site in the closed state. The active site residues are predomintely seen in the 1st ranked site which is in agreement with the possibility of intermediate conformations in the simulation of 'model 2'. In most primary predicted sites, D205 is included as well.

Although there is strong evidence for at least a partial glucose binding site in 'model 2', the allsoteric binding site does not appear fully (Figure 5.34). In comparison to 'model 1' where a similar number of interactions have been identified, the rank of the allosteric site has improved in 'model 2'. In some frames, more than one site includes allosteric site residues, which suggests the presence of several small site in this region, which could eventually combine to form a complete allosteric binding site.

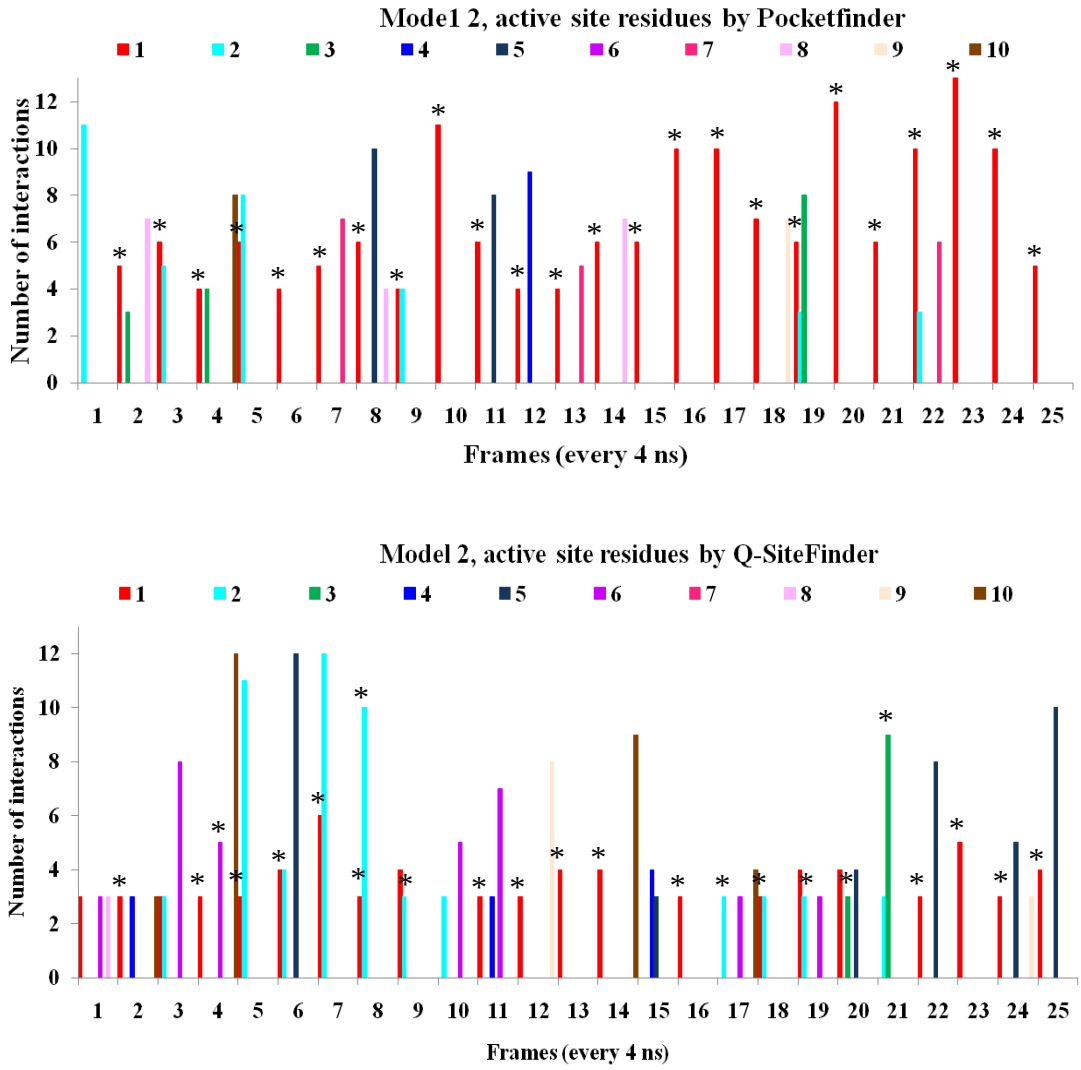


Figure 5.33: On top, predicted binding sites involving glucose interaction residues for Model 2 with Pocket-Finder and below with Q-SiteFinder. On the horizontal axis, 25 frames from the 100 ns simulation. For each frame 10 identified sites are coloured and numbered in order of rank. "*" indicates the presence of D205 as an interaction residue in the site.

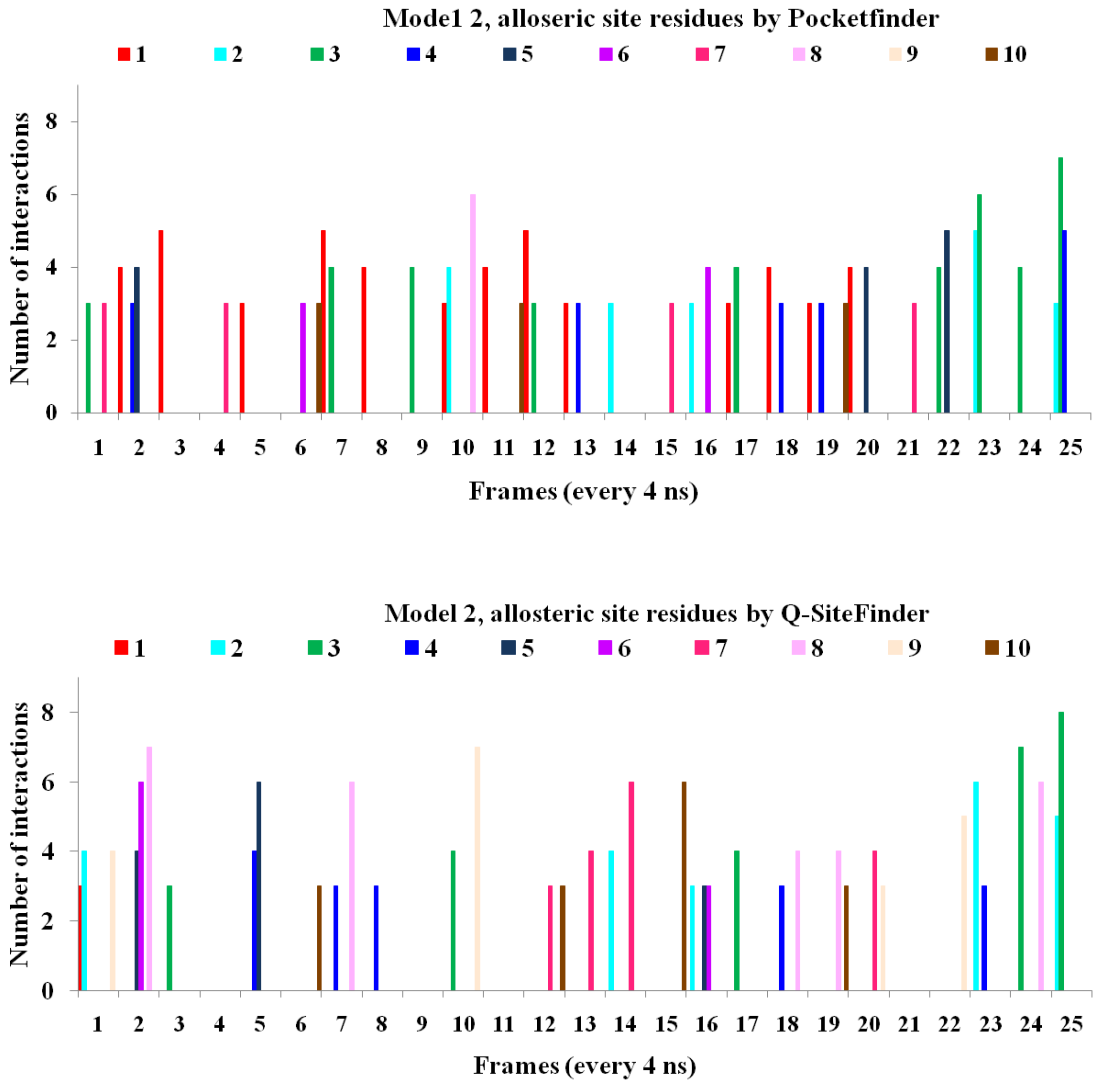


Figure 5.34: On top, predicted binding sites involving allosteric activator contact residues for Model 2 with Pocket-Finder and below with Q-SiteFinder. On the horizontal axis, 25 frames from the 100 ns simulation. For each frame, the 10 identified sites are coloured and numbered in order of rank.

It is promising that the allosteric site can be at least partially predicted for the super-open state, which is in agreement with a low affinity conformational state in the pre-existing equilibrium.

The location of the allosteric site residues identified by the two methods, shown in the plots above have been depicted in figures 5.35 to 5.38. Although sites involving the allosteric site residues (from the closed state) are fairly spread out, due to the spread of

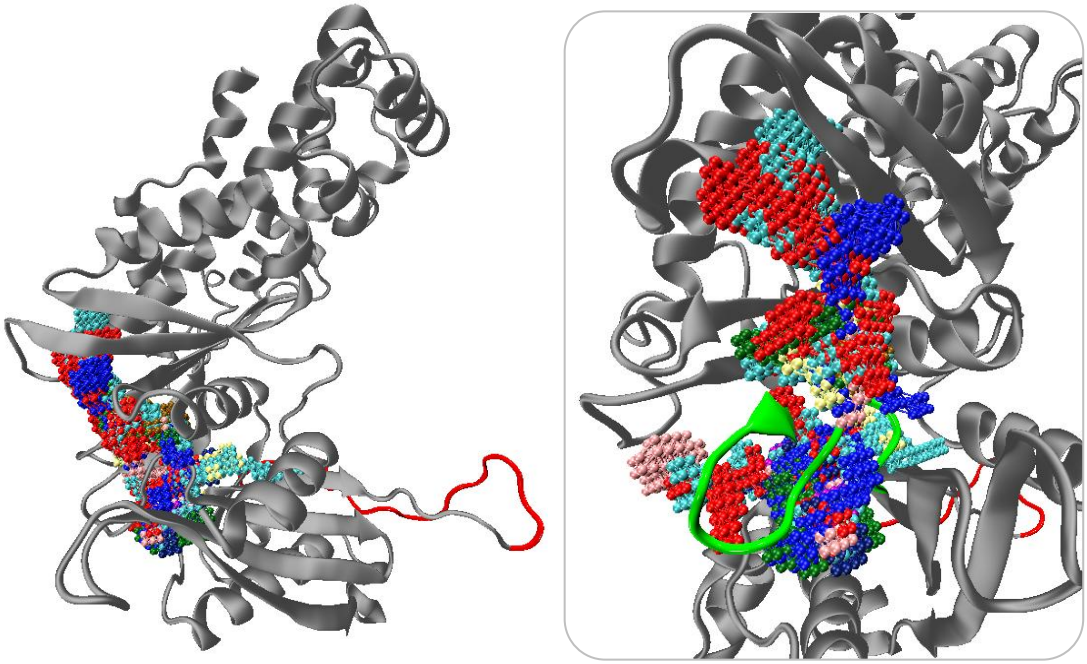


Figure 5.35: Model 1, binding sites that have been identified by Pocket-Finder to interact with at least 3 residues that point into the allosteric binding site of the closed-state conformation X-ray structure (1v4s). On the right, a zoomed image of the pockets, coloured in the order of rank, corresponding to plot 5.32. The loop connecting the two domains is in green.

the relevant residues in the super-open state, they are within the vicinity of the connecting loop between the two domains. In the figures (5.35-5.38) the rank of the binding sites have also been demostrated by the use of the relevant colours, corresponding to the binding site profile plots (figures 5.32 & 5.34). As Pocket-Finder on the whole identifies larger binding sites, it is more likely to observe sites that appear to be relatively far from the connecting loop, but where the allosteric interacting residues interact with one edge of a large/long predicted site.

The location of the binding sites may also suggest the potential direction in which the allosteric activators may gain access to the binding site. Consistently, in all four figures (5.35 to 5.38), to the left of the connecting loop (green) in the zoomed images, several pockets have been idenfied. Interestingly, in the closed state, a β-turn envelopes the allosteric binding site in a similar region to the extension to the left of the connecting loop in the super-open state. This β-sheet in the closed-state corresponds to the part of the disordered region of the structure in the super-open state.

In general Q-SiteFinder identifies smaller sites relative to Pocket-Finder, as highlighted before. Similarly, here smaller sites have been identified for models 1 & 2 by Q-SiteFinder (figures 5.36 & 5.38). In "model 1", the relative positions of the predicted sites involving the allosteric site residues are similar, identified by both algorithms (figures 5.35 & 5.36).

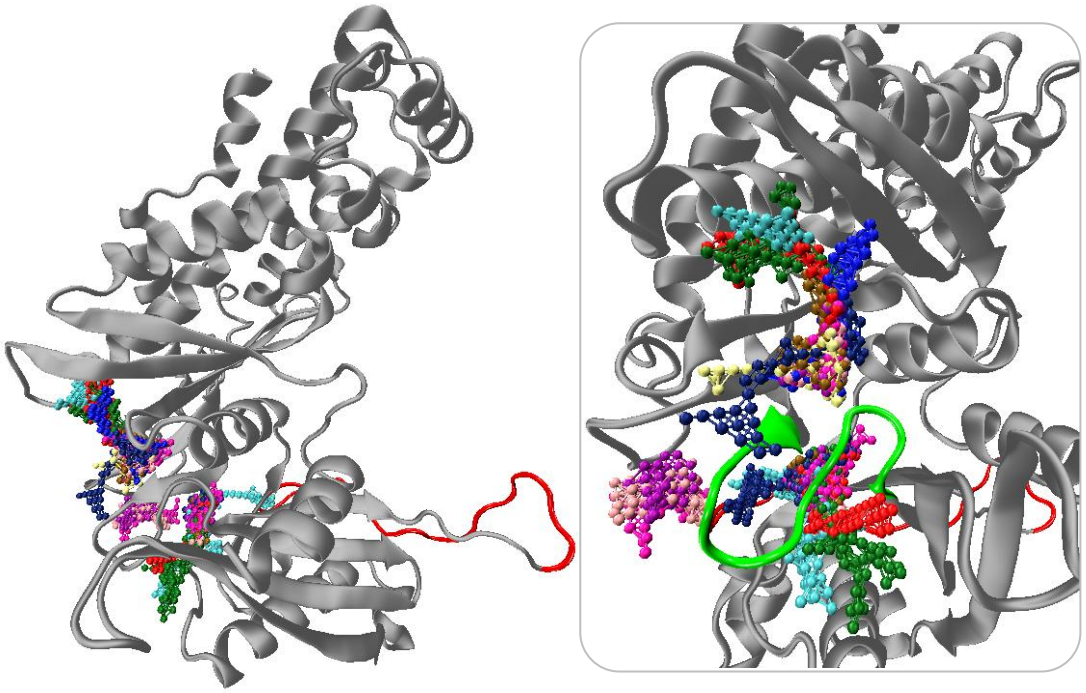


Figure 5.36: Model 1, binding sites that have been identified by Q-SiteFinder to interact with at least 3 residues that point into the allosteic binding site o the closed-state conformation X-ray structure (1v4s). On the right, a zoomed image of the pockets, coloured in the order of rank, corresponding to plot 5.32. The loop connecting the two domains is in green.

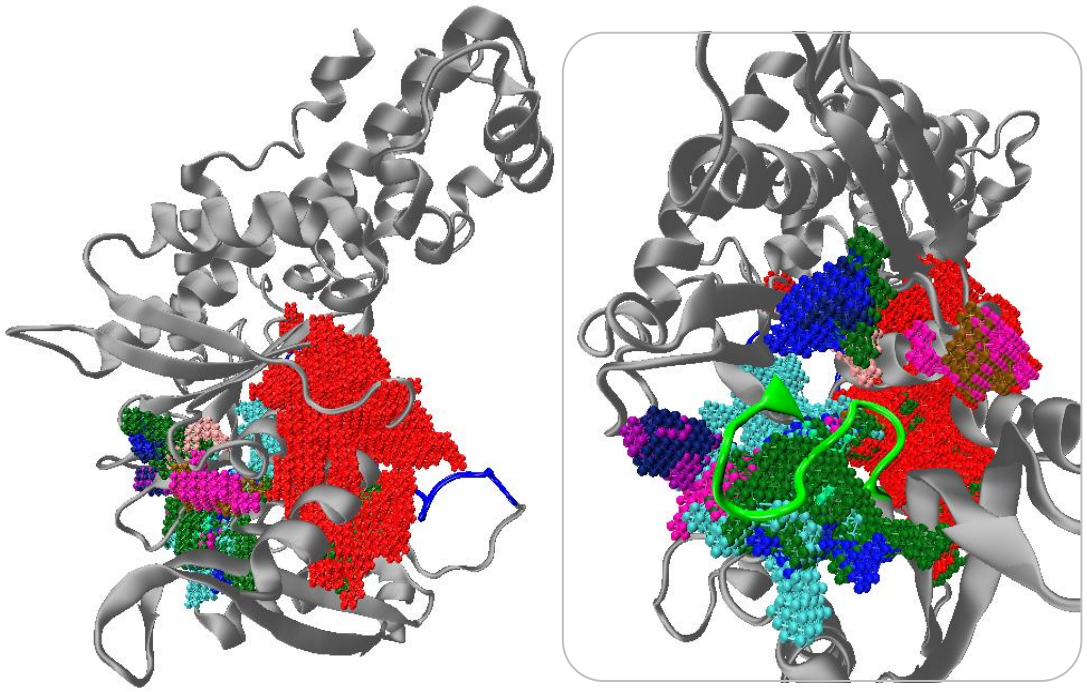


Figure 5.37: Model 2, binding sites that have been identified by Pocket-Finder to interact with at least 3 residues that point into the allosteric binding site of the closed-state conformation X-ray structure (1v4s). On the right, a zoomed image of the pockets, coloured in the order of rank, corresponding to plot 5.33. The loop connecting the two domains is in green.

In ‘model 2’, the spread of the sites involving allosteric site resiudes expands. As Pocket-Finder identifies larger sites, in ‘model 2’ this has led to the identifiction of both the allosteric and active site residues in the first predicted site of many frames with Pocket-Finder (figure 5.37). This may be due to the large conformational change in the small domain in the super-open conformation, that has allowed the active site and allosteric site residues to be closer in distance than in the closed state. Ignoring the first site in ‘model 2’ predicted by Pocket-Finder, leads to a similar spread pattern of predicted sites in this region to that of ‘model 1’.

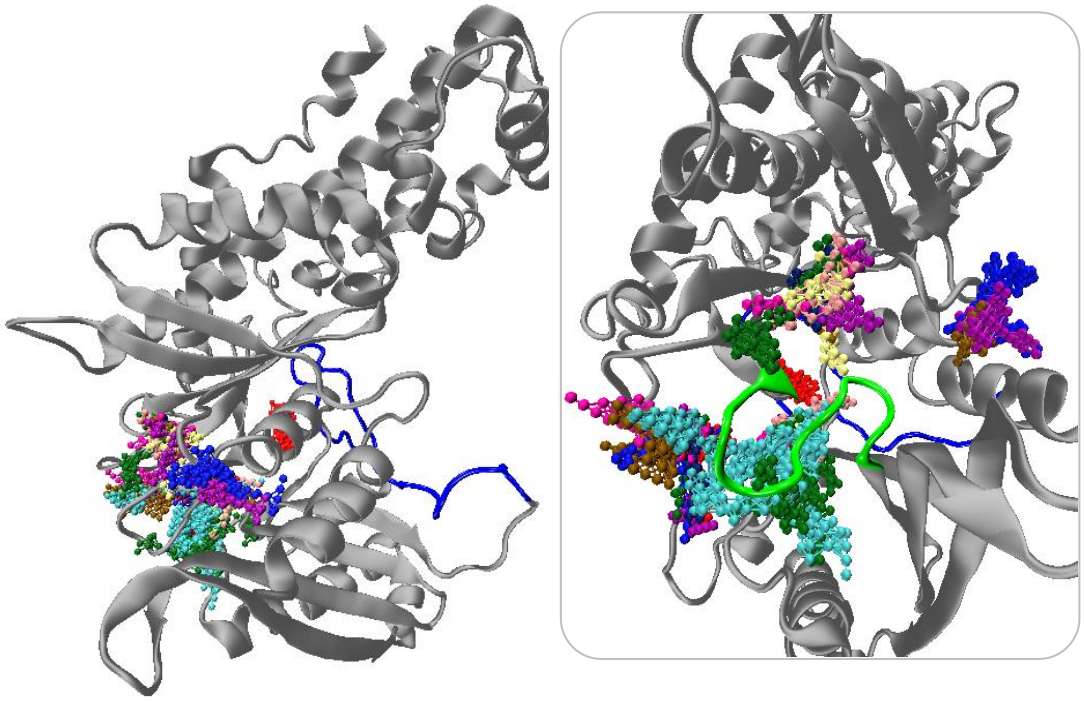


Figure 5.38: Model 2, binding sites that have been identified by Q-SiteFinder to interact with at least 3 residues that point into the allosteric binding site of the closed-state conformation X-ray structure (1v4s). On the right, a zoomed image of the pockets, coloured in the order of rank, corresponding to plot 5.33. The loop connecting the two domains is in green.

Sites prediced by Q-SiteFinder for ‘model 2’ are less spread in comparison with those predicted for ‘model 1’. Here (figure 5.38), the potential path of access to the allosteric binding site is highlighted with the higher populations of pockets close to the connecting loop (green), mainly to the left of the loop (relative position in the image).

5.6 Summary of the super-open state

Two conformations of the disordered region in the super-open state X-ray structure (PDB ID: 1v4t) were prepared as starting structures for the MD simulations. Owing to a stable conformation that the modelled loop adopted in 'model 2', being close to the active site, throughout the simulation a greater number of interactions similar to that of the active site in the closed state were observed. In addition to a better formed glucose binding site in 'model 2', D205, believed to be an important residue in glucose association to the enzyme, has been picked up in the predicted site more frequently than in 'model 1'.

The allosteric binding site profile did not improve with a better formed active site in 'model 2'. We observed a small pocket involving a few of the allosteric interacting residues, as a highly ranked site in the starting structure. During the course of the simulation, the site increases in size and a greater number of the allosteric residues are picked-up in the region, although the ranking of the sites are not always high. The binding site profile observed in the allosteric region is consistent with kinetic studies, indicating that the allosteric activator could bind the enzyme in the absence of glucose, although with fairly low affinity.

In this state, without prior knowledge and based on the binding site prediction, the allosteric binding site would have not necessarily been identified during the simulation, although the X-ray structure binding site search did reveal a small pocket in this location ranked as the 2nd site but neither the size, nor the location is optimum for binding in this conformation.

The conformation of the modelled loop in 'Model 2' is the more likely conformation but it is evident that the loop is highly mobile, as it has not been resolved in the X-ray structure.

In the following chapter, a mutant form of the closed state, a glucose bound GLK X-ray structure provided by AstraZeneca (not publicly available) is studied. In this structure the allosteric site is not identified with either pocket detection tool. This demonstrates that a static conformation is not sufficient to spot such allosteric sites. Similar methods

used in this chapter will be discussed in the following chapter, applied to the closed state AZ X-ray structure, where a full allosteric site is absent.

5.7 Summary of benchmark chapter

We gained a considerable insight into the conformational dynamics GLK, both in the closed active and super-open inactive states. The MD simulation protocol has been reliable and successful in capturing the sensitivity of the protein to the presence of either or both ligands at the active and allosteric sites. Both pocket detection methods are relatively reliable, although Q-SiteFinder may be a bit too sensitive, but it benefits from the advantage that it is an energy-based detection method.

NMA of the closed state demonstrated that a few normal modes are not sufficient to capture the entire flexibility of the protein backbone, but up to a minimum of ~20 modes are required for this system. If we use no prior knowledge it is important to capture the mobility of all residues in the protein system. This is further evidenced by the fact that the flexibility of the allosteric region in GLK, is in particular influenced by the inclusion of more modes. The loop connecting the two domains adopts a fairly compact conformation in the absence of an allosteric activator. As normal mode analysis is limited to vibration around one local minimum, it appears that this is not sufficient to fully capture the mobility of this loop. However, residues in the centre of the loop do move sufficiently along the 20 normal modes to improve the binding site prediction at the allosteric site in terms of rank and size.

It is interesting that even in the super-open state, evidence of the active and allosteric sites are observed, in particular in ‘model 2’ where the conformation of the modelled loop may be that of an intermediate conformation along the conformational transition to the closed-state.

Studies in this chapter indicate that MD would be the more reliable method in comparison with NMA, in predicting the allosteric binding site in the GLK_AZ structure (chapter 6). However, considering the improved allosteric binding site size and rank in comparison with the minimised structure in NMA (B), and the significantly shorter

calculation times in comparison with MD simulations, the application of NMA to GLK_AZ cannot be ruled out, although significantly more modes than a handful may be required.

In the following chapter the GLK_AZ structure will be discussed with the aim of predicting the allosteric binding site. The structure does not include a noticeable cavity at the allosteric site.

Chapter 6

Active closed state glucokinase: GLK_AZ

6.1 Aim

In the previous chapter the two publicly available conformations of human glucokinase were studied and discussed. Here an X-ray structure of GLK in the closed state, bound to glucose only will be discussed. This structure has been provided by the crystallography group at AstraZeneca (unpublished data), referred to as GLK_AZ, which is the liver isoform 3 (in contrast to 1v4s, isoform 2). Owing to low structural stability of GLK in the active state it has been challenging to crystallise the active closed state; for this reason several mutations were applied to the sequence that was co-crystallised with glucose (see appendix A for sequence comparison). In this structure, the connecting loop between the two domains adopts a conformation which passes through the region where the allosteric activator is bound in the publicly available closed-state X-ray structure (PDB ID: 1v4s), discussed in chapter 5. A binding site search on this structure revealed a small, low rank pocket close to the allosteric region, not including all the residues forming the allosteric site in the original closed state GLK structure bound to glucose and the allosteric activator (PDB ID: 1v4s). Yet, this mutant sequence has been utilised as a template in the co-crystallisation of allosteric activators at AstraZeneca, and hence the structure can accommodate a ligand at the allosteric region when the activator is in solution.

Without the prior knowledge of the allosteric binding site from the activator bound X-ray structure (PDB ID: 1v4s), the allosteric binding site in the GLK_AZ X-ray structure

would have been overlooked. The main aim of this chapter is to predict a full allosteric binding site that is highly ranked, which is not observed in the static X-ray structure. In addition, we can explore the possibility of there being additional or alternative binding sites that may be utilised.

Once again molecular dynamics and normal mode analysis were applied to study the dynamics of the GLK_AZ structure, in the presence and absence of the active site ligand, glucose. In addition, in the following chapter the sequence was also reverted back to the wild-type to ensure that simulation observations were not purely influenced by the crystallographic mutations. Both simulations were then followed by binding site searches.

6.2 System preparation and MD parameters for the simulation of GLK_AZ

The GLK_AZ X-ray structure has a 97% sequence identity with the original structure studied in this thesis (PDB ID: 1v4s). The following residue mutations, M11S, A12S, L13N, T14S, L15Q, E27A, E28A, E51A, E52A, E94A, E95A, E96A have been made in comparison with 1v4s (see figure 6.1 for locations). Residues M11, A12, L13 have been resolved in addition to those in the 1v4s structure. The mutant structure ends at residue 456 whereas the X-ray structure of 1v4s ends at residue 461. Sequence alignments are given in appendix A.

The crystallisation mutations are not directly involved in the active or allosteric sites. Simulation of the original closed state conformation (PDB ID: 1v4s), revealed a highly flexible region in the β-turn where three mutations have been inserted (residues 94 to 96). Additionally, residues 96 to 97 were not resolved in the X-ray structure (GLK_AZ), which has been externally modelled by Richard Ward (*219*) using PRIME, a Schrödinger tool (*261*). The GLK_AZ conformation is structurally highly similar to the 1v4s X-ray structure except for the allosteric region, which has been depicted in figure 6.1.

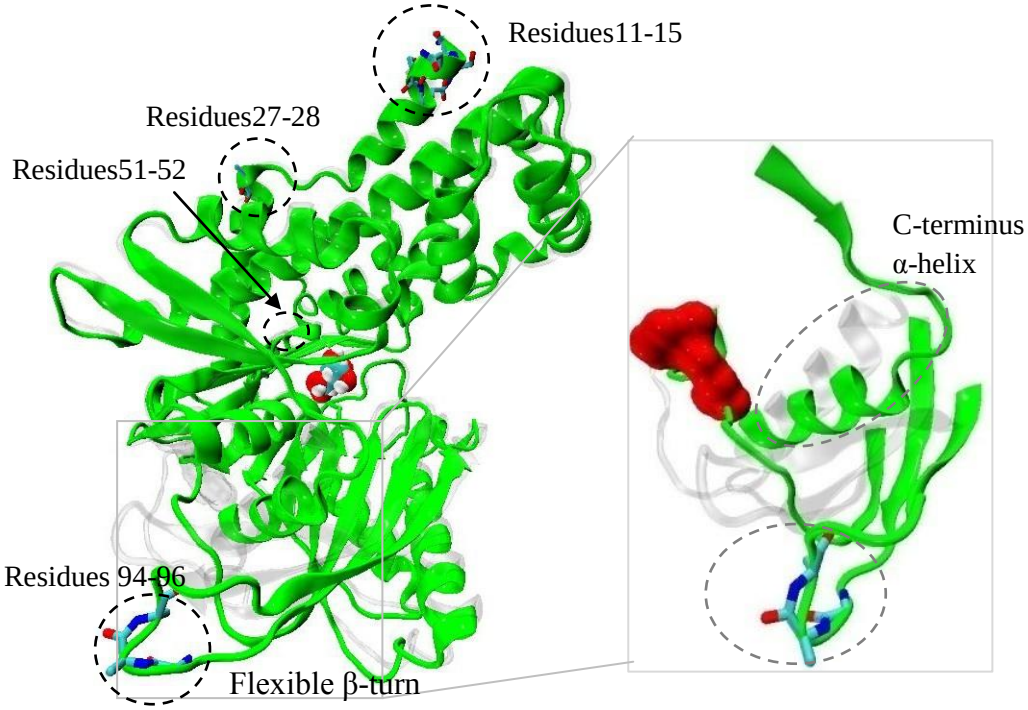


Figure 6.1: An overlay of GLK_AZ (green) and the original closed state GLK (grey) (PDB ID 1v4s). The glucose is in van der Waals ball representation and the mutation residues in stick representation, circled by dotted black lines. On the right, a close-up view of the allosteric region. The allosteric activator in 1v4s is represented in a red surface. The connecting loop from GLK_AZ passes through the allosteric activator region in 1v4s.

For the GLK_AZ mutant structure, two simulations were carried out; one in the presence of glucose in the active site (GLK_AZ-holo hereafter) and another, in the absence of glucose from the active site, as a pseudo apo (GLK_AZ-apo$_{pseudo}$ hereafter).

For each simulation, if ligand present, glucose was protonated via the PRODRG (*247*) web server, and the protein in WHATIF (*248*). Similar to the set-up in chapter 5, both simulations, have been carried out using the AMBER (*114*) molecular dynamics package using the AMBER03 force field for the protein and gaff force field, AM1-BCC charge method (*249, 250*), for the ligand, with Particle-Mesh-Ewald (PME) boundary conditions limited to a 10 Å cutoff.

The Amber package tool, XLEAP, was used to solvate the system with TIP3P water models, with a minimum distance of 12 Å from the protein (19204 water molecules). A total of 15 sodium ions were added to the systems to neutralise the overall charge in each case. Fewer ions were required to neutralise the system in comparison with 1v4s, owing to the mutations of negatively charged glutamic acid residues to alanines.

The minimisation was carried out in stages, starting with the minimisation of solvent only while applying a restraint force constant of 500 kcal mol^{-1} $Å^{-2}$ on the rest of the particles, followed by the minimisation of protein with restraint on other particles, then the ligands, if applicable, and finally the entire system, removing all restraint; in each stage, reaching an energy value below the set root-mean squared deviation of 0.001 kcal-$mol^{-1}Å^{-1}$. In each case, the minimisation commenced with 100 steps of steepest descent, followed by conjugate gradient for the remainder.

For each, the minimised system was heated gradually to 300 K in the NVT ensemble, in six 50 K blocks, allowing 15000 MD step for each block, using a Langevin thermostat with a 1 ps^{-1} collision frequency.

The production MD simulations were run in the NPT ensemble, using Langevin thermostat with a 1 ps^{-1} damping parameter at 300 K and a time-step of 2 fs. The pressure was controlled by isotropic position scaling with a relaxation time of 2 ps. All bonds containing hydrogen were constrained using the SHAKE algorithm. The non-bonded cut-off of 10 Å was employed.

For each simulation, 100 ns of MD production was collected after removing the initial ~1.5 ns assuming equilibration at the beginning of the NPT production run, based on stabilisation of properties such as energy, volume, density and RMSD.

6.3 MD analysis for GLK_AZ

The major differences between the active, closed state structure (PDB ID: 1v4s), bound to glucose and allosteric activator with GLK_AZ have been depicted in figure 6.1, where 1v4s was treated as the reference to establish the exact conformational difference

between the two structures. In figure 6.2, the relative displacements between residues of GLK_AZ and 1v4s have been illustrated with the aid of a plot, where the residues of 1v4s were treated as a reference and the displacement of residues of GLK_AZ relative to the reference, 1v4s, were calculated.

As demonstrated in figures 6.1 and 6.2, there is a significant conformational difference in the loop connecting the two domains (residues 64-71), followed by the β-turn (residues 91-100), where several mutations have been made. The final major difference is in the loop preceding the terminal α-helix. The overlay demonstrates the difference of this α-helix in the two structures (figure 6.1). In addition this α-helix is 5 residues shorter than that of 1v4s.

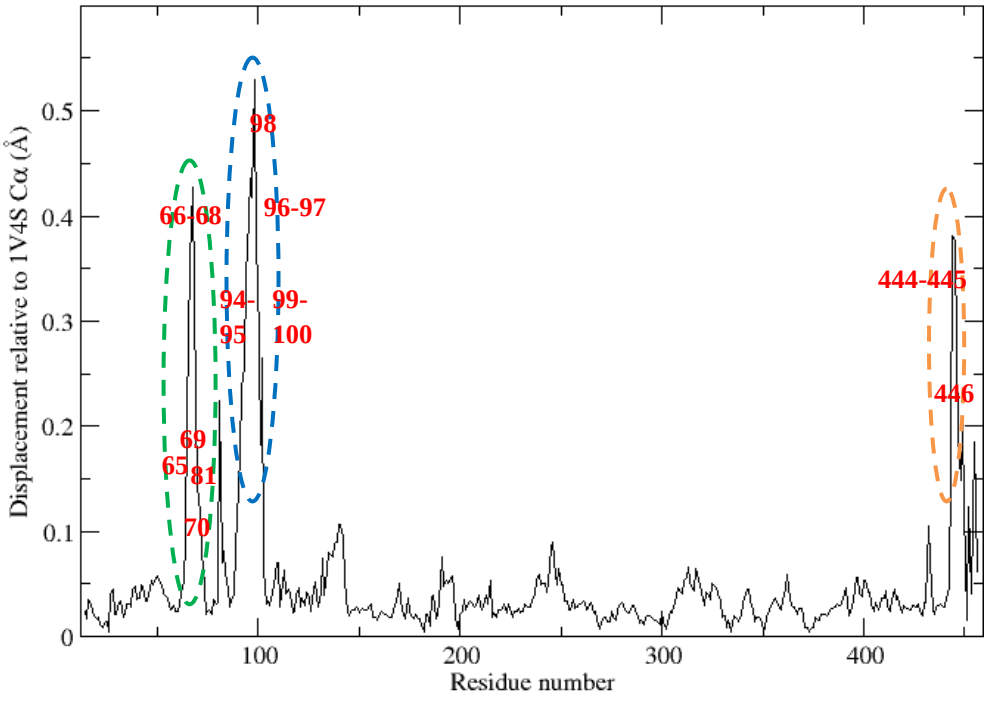


Figure 6.2: Displacement of Cα atoms of GLK_AZ relative to 1v4s. Significant displacements have been labelled. Circled in green are residues in the loop connecting the two domains, in blue the β-turn residues and in orange the loop leading to the C-terminus α-helix. Refer to figure 6.1 for an overlay of the two structures.

In what follows, results from both simulations, GLK_AZ_holo and GLK_AZ_apo, are presented. A comparison of the RMSF plots from the MD simulations (figure 6.3) demonstrates a very similar profile in most regions of the structure. However, a significant difference is observed in residues 92-99 in the β-turn, which further highlights the role of glucose in the flexibility of this region. In addition, residues 68-69 of the connecting loop display more flexibility in GLK_AZ_holo, in the presence of glucose at the active site.

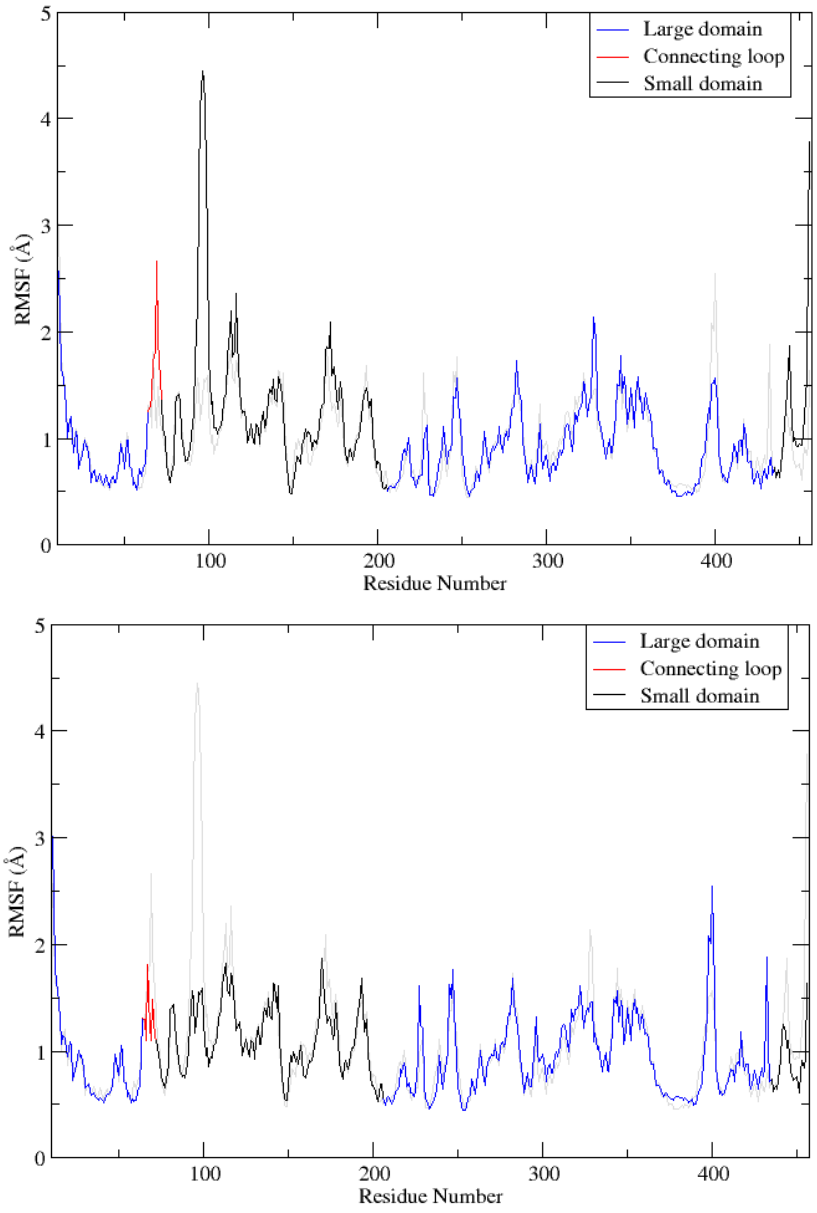


Figure 6.3: RMSF GLK_AZ_holo on top and GLK_AZ_apo$_{pseudo}$ on the bottom. In gray an overlay of the other plot for comparison.

The mobility reduction of the loop connecting the two domains and the β-turn in the GLK_AZ_apo$_{pseudo}$ simulation are consistent with simulation 'D' of 1v4s (figure 5.5), although the mobility reduction is not as prominent is 1v4s, simulation 'D'.

In the following section binding site detection through the simulations GLK_AZ_holo and GLK_AZ_apo$_{pseudo}$ will be presented.

6.3.1 Binding site profiles in the starting structure and through the simulations of GLK_AZ

In the X-ray structure the loop connecting the two domains goes through the region occupied by the allosteric ligand in 1v4s.

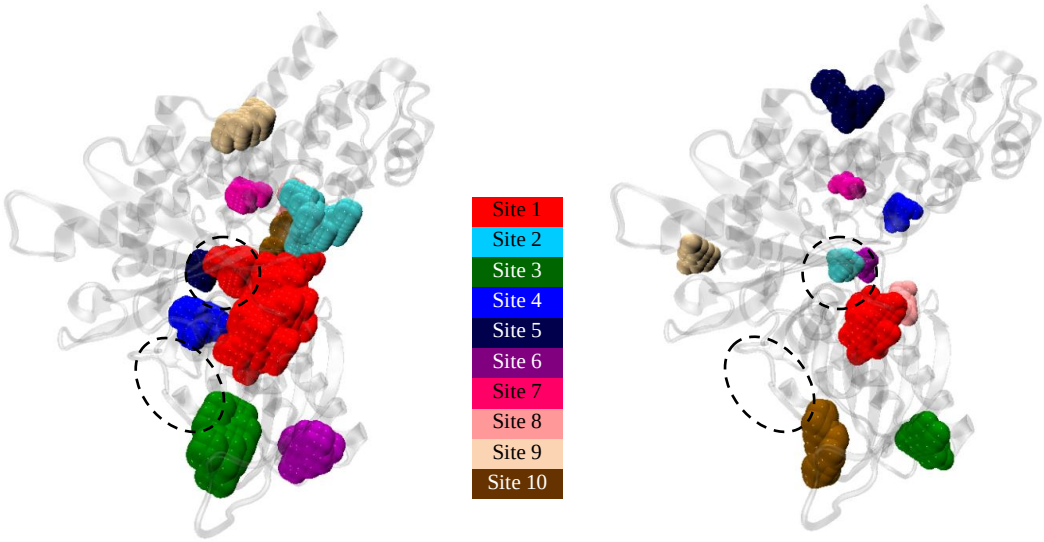


Figure 6.4: On the left, top 10 sites predicted for GLK_AZ by Pocket-Finder, and on the right the corresponding by Q-SiteFinder. In the centre is the colour code for the sites in order of rank. The dotted circle and oval indicate the location of the glucose and allosteric sites, in the original X-ray structure (PDB ID: 1v4s), respectively.

Looking at the crystal structure, both binding site search methods correctly identify the glucose binding site, although as discussed earlier, Q-SiteFinder appears to split sites

that Pocket-Finder would identify as one large site and the ranking is thereby affected. The allosteric binding site is not identified by either method. The 3rd site identified by Pocket-Finder, and site 10 by Q-SiteFinder correspond to a region that larger allosteric ligands may extend in to as seen partially in one of the latest publicly X-ray structure of GLK bound to glucose and an allosteric activator (PDB ID: 3h1v) (*263*). The region is the space created by the very flexible β-turn (residues 91-100).

To establish if the allosteric binding site opened at any stage of either simulation, Pocket-Finder and Q-SiteFinder were utilised to search 25 frames from each simulation (GLK_AZ_apo and GLK_AZ_apo$_{pseudo}$) at 4 ns intervals.

In this case, to identify the active or allosteric sites in the 10 sites predicted by either method, residues that were identified to point into either binding sites in the original X-ray structure (PDB ID: 1v4s), identified by Pocket-Finder and Q-SiteFinder, were combined (Table 5.2). A combination of the residues, rather than separate lists for the two methods were used, for ease of analysis; as the residues are just a means by which the locations of the sites can be identified without the need for visual inspection. The residues are listed in table 5.2, and consist of 24 residues for the active site and 19 for the allosteric site. These residues were then used to work out which predicted site in GLK_AZ corresponds to the active or allosteric site, with a minimum of at least 3 residues required for the site to be classed as a relevant allosteric or active site pocket, at or near either binding site location in comparison with 1v4s. The search was carried out with the aid of a purpose-written perl script, which searches through the text files returned from the server and can report the number of relevant residues as a counter for each pocket, and the actual residue name/number involved. The active, glucose, binding site was fully predicted by both binding site search methods for the GLK_AZ X-ray structure, as seen in figure 6.4.

Throughout the GLK_AZ_holo simulation glucose remained in the active site, and therefore the cavity should be identified throughout the simulation. Pocket-Finder consistently identifies the active site as the 1st rank site, similar to the starting structure (figure 6.5). Q-SiteFinder does identify the active site in all frames but does not rank it as the top binding site. This is partly due to the fact that Q-SiteFinder splits a large site

into precise individual sites; therefore, it can be seen in the plot (figure 6.5), more than one site would be identified in the region, affecting the ranking.

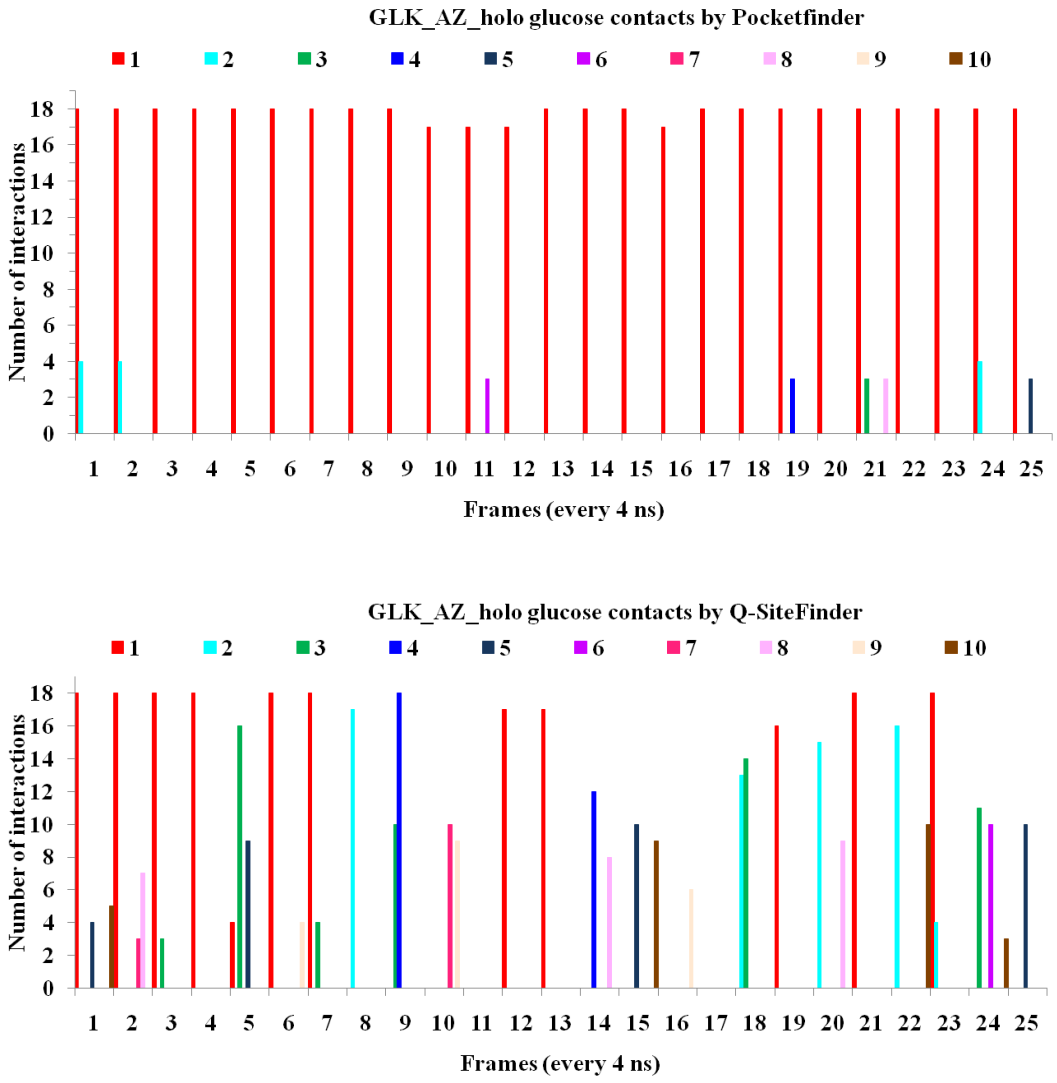


Figure 6.5: Active site binding-site search profile on frames from the 100 ns simulation of GLK_AZ_holo. On top, binding sites identified by Pocket-Finder and below by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site for it to be counted as a site at the active or allosteric site in GLK_AZ. Numbers 1 to 10 refer to the rank of the binding site.

In a few of the frames, including frames 11 and 17, despite the presence of glucose in the binding site, Q-SiteFinder does not predict a binding site in the active site region. In these two frames there is a slight move of the small domain away from the large domain.

The local conformational changes has led to small pockets appearing in less populated regions, and as only 10 sites are reported, the glucose binding site does not get identified. The conformation is such that the active site is not identified as optimum for binding. Interestingly, in these frames, strong allosteric sites are predicted (figure 6.6).

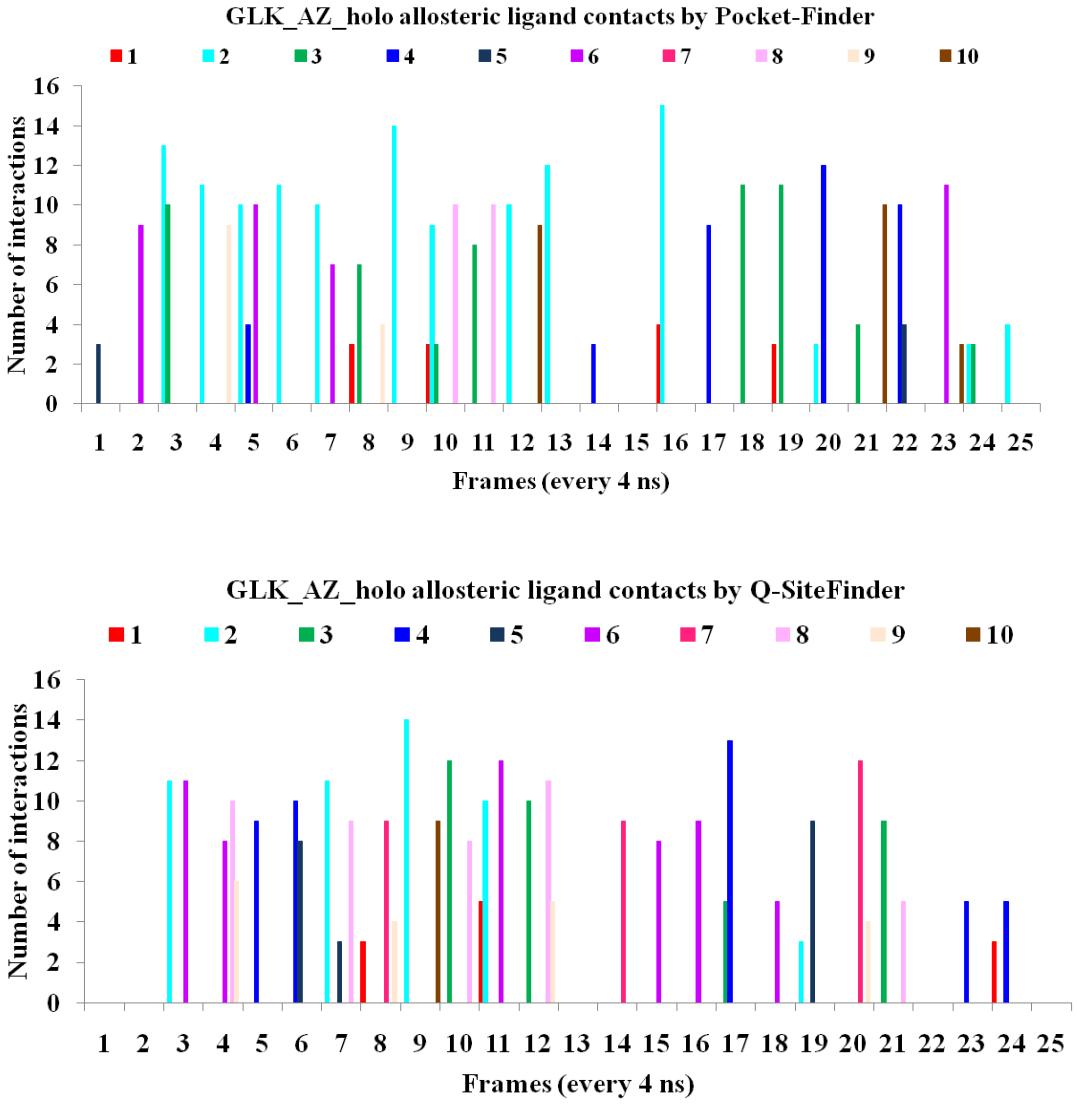


Figure 6.6: Allosteric binding site search profile for frames from the 100 ns simulation of GLK_AZ_holo. On top, binding sites identified by Pocket-Finder and below by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site for it to be counted as a site at the active or allosteric site in GLK_AZ. Numbers 1 to 10 refer to the rank of the binding site.

During the simulation of GLK_AZ_holo, the allosteric site has also been analysed (figure 6.6). From the 3rd frame (~9-12 ns) in the simulation, Pocket-Finder identifies the allosteric site as the 2nd site in 13 of 25 frames, which is promising.

In frames where the allosteric site has not been as ranked 2nd, the site is still predicted within the top 6, in most cases as 3rd or 4th ranked.

Q-SiteFinder also identifies the allosteric site within the top ten sites for most frames; however, the method again suffers from splitting, which ultimately influences the ranking. For the active site, Q-SiteFinder picks more than one site in many frames (figure 6.5), in comparison with Pocket-Finder that predominately picks just one large site in the region, which allows the ranking to remain fairly insensitive to small local conformational changes. The sensitivity of Q-SiteFinder, which is an attempt for accuracy, does lead to lower ranking of the allosteric binding site.

The location of sites 2nd ranked by Pocket-Finder has been depicted in figure 6.7, to demonstrate in what location on the protein surface the site has been identified, when not exactly at the allosteric site for simulation GL_AZ_holo. This is an overlay of 13 frames, where the predicted 2nd site is not precisely at the allosteric site; sites are grouped into 5 locations, and coloured accordingly. Therefore, each location may contain more than one predicted binding site, indicated by the frame number in figure 6.7.

When the 2nd predicted binding site is not at the allosteric site (12 of 25), there are three locations that could be either an extension to the active or the allosteric sites (figure 6.7). There are two locations that cannot be categorised, but only occur once in the 25 frames, which can be regarded as fairly rare.

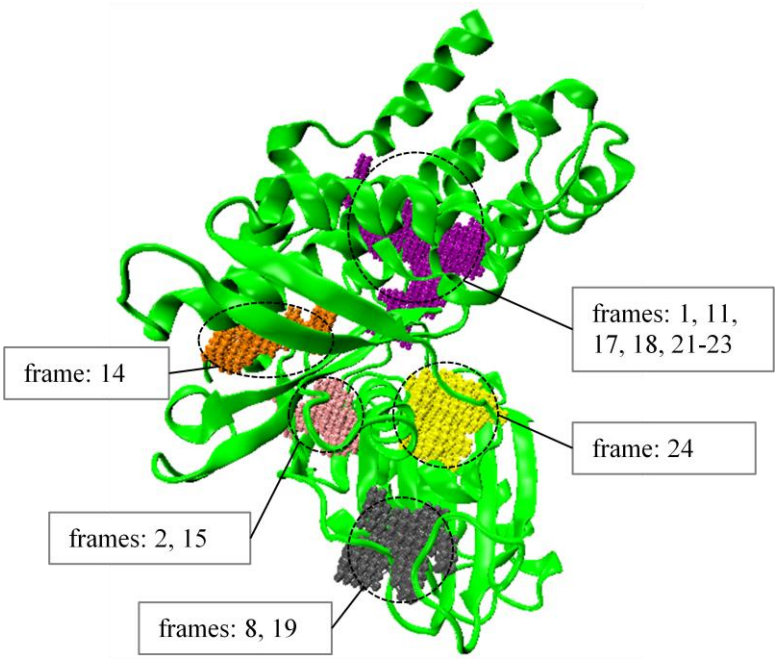


Figure 6.7: 2nd ranked sites in simulation GLK_AZ_holo, when not directly at the allosteric site, identified by Pocket-Finder. The number of occurrence in the vicinity of each site is indicated by the frame number. There are five areas in which the 2nd site appears. Of those, the purple corresponds to a region that is very close to the active site. The pink is at a position close to the allosteric site, but involving residues preceding the connecting loop. The gray could be an extension to the allosteric binding site. The location of the other site cannot be categorised.

In frames 8 and 19 the 2nd site has been predicted between the connecting loop and the flexible β-sheet (residues 91 to 100). This loop has been demonstrated to adopt many different conformations in the presence of different activators (figure 4.4). We propose that this region may be used as an extension to the allosteric binding site for larger ligands. Frame 24 is one that has been depicted in the plot (figure 6.6) but the visualisation shows that the pocket is clearly not in the allosteric region (figure 6.7). This is due the interaction that the site has with two residues (D205 & T209) of the α-helix between the active and allosteric site (residues 205-219, appendix B) but not with the side-chain atoms of those residues that normally point into the allosteric pocket. In addition the pocket interacts with backbone atoms of N204 preceding this α-helix, side-chains of which point into the allosteric binding site.

Similarly, the 2nd site predicted by Q-SiteFinder is depicted in figure 6.8. Here more locations are observed, mainly due to the splitting of sites in Q-SiteFinder.

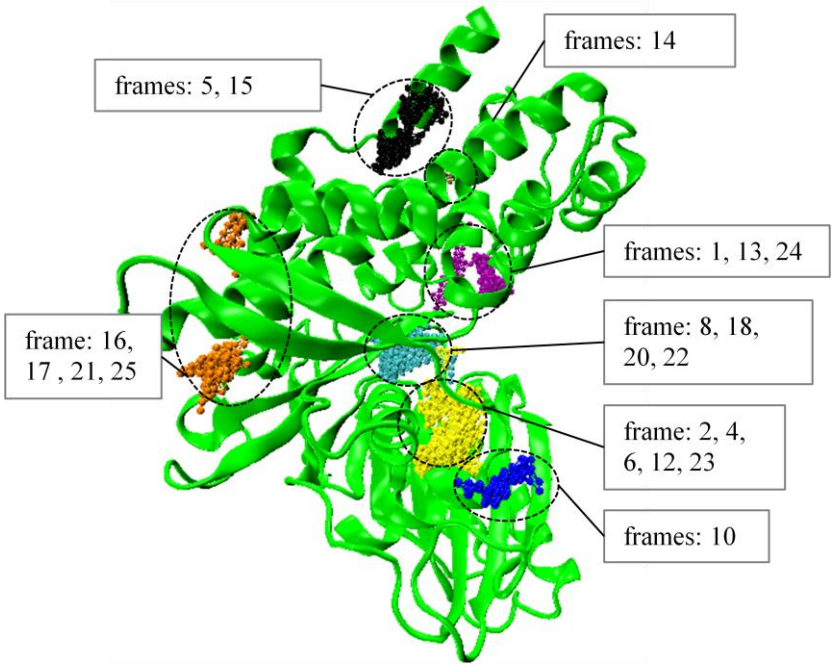


Figure 6.8: 2nd ranked sites in simulation GLK_AZ_holo, when not at the allosteric site, identified by Q-SiteFinder. The number of occurrence in the vicinity of each site is indicated by the frame number. Here, predicted 2nd ranked sites have been grouped into 7 locations, when not at the precise allosteric site. Three sites in purple, cyan and yellow could be imagined as an extension to the active site.

The allosteric site is ranked as the 2nd site only in 5 frames out of 25 with Q-SiteFinder (figure 6.6). In figure 6.8, locations indicated by purple, cyan and yellow could be thought of as an extension to the active site. In orange, two locations have been ranked as the 2nd site in frames 5, 15, 16, 17, 21 and 25. The significance of this site is not known.

The binding site search profile with Q-SiteFinder, demonstrates that the allosteric binding site is ranked among the top 5 sites in 14 frames (of 25) during the 100 ns simulation trajectory (GK_AZ_holo). Although, not always highly ranked using both methods, which is to be expected due to the flexibility of the binding site in the absence of a ligand, in combination with the knowledge of the location of the activating mutations (see section 4.4) at or near the allosteric site, the binding site search along the MD trajectory could offer very useful insight into the allosteric region. The conformational ensemble from the MD simulation could have also been used as a good

starting point in flexible docking of allosteric activators in the absence of the allosteric bound structure (PDB ID: 1v4s).

If we ignore the binding site ranking predicted by Pocket-Finder and Q-SiteFinder and combine all binding pocket grid points predicted for every frame, one can use this as an ensemble of grid point populations, which can give an indication of regions of high occupancy. Below in figure 6.9, this has been demonstrated for sites predicted by both Pocket-Finder and Q-SiteFinder.

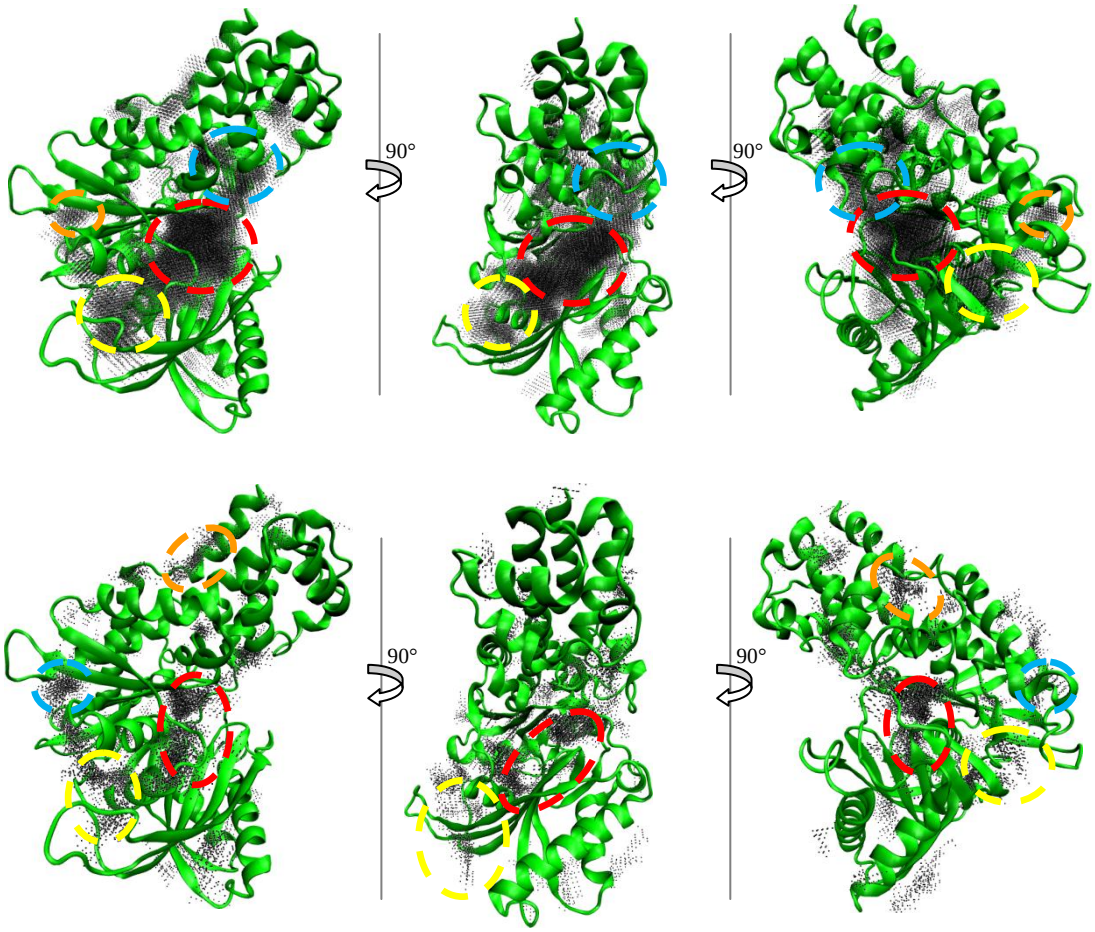


Figure 6.9: All grid points for all predicted binding sites by Pocket-Finder on top and by Q-SiteFinder below, overlaid for all 25 frames at 4 ns interval in the simulation GLK_AZ_holo. For each frame, 10 sites were predicted, including all 25 frames leads to 250 sites in each case. For clarity, for each method, three orientations of the sites on the protein surface have been shown, rotated by 90°. Regions of high population have been circled in different colours, and repeated when rotated.

In the case of grid point predicted by Pocket-Finder, without prior knowledge of the system, the active site would clearly be noticed as a highly populated region, which is consistent with reality. The region circled in blue in figure 6.9 (top) is very close to the active site; and therefore we could imagine this as an extension to this site. The allosteric region is also noticeably populated, which may be targeted without any prior knowledge. Although the grid points are quite spread out over the entire protein surface, the active and allosteric sites do appear to be the highly populated areas.

For Q-SiteFinder it is harder to judge the population (figure 6.9, bottom). In this case more frames from the simulation may be necessary. The active site is noticeably populated, which is promising, however it is harder to differentiate the grid population in the allosteric region from others. In comparison with all other sites, the allosteric site does appear as one large patch, relatively buried in the protein that may favour binding in comparison with other small patches on the surface of the protein. When viewed in 3D, the circled areas are the most populated and the allosteric site is at least included in this group.

In the simulation of GLK_AZ_apo$_{pseudo}$, during which glucose was removed from the binding site, one would expect more freedom in the system. Although a simulation length of 100 ns is not sufficiently long to observe the full opening of the two domains in the absence of glucose, this simulation has been surprisingly stable. When both glucose and the allosteric activator were removed from the simulation of the original X-ray structure (PDB ID: 1v4s), there was sufficient fluctuation to observe a small opening of the two domains. It may be possible that the mutations in the GLK_AZ sequence have influenced this extended stability.

The glucose binding site remains fairly stable during this simulation (GLK_AZ_apo$_{pseudo}$). Both Pocket-Finder and Q-SiteFinder identify the glucose binding site and in the case of Pocket-Finder, rank it as the 1st site, and in the case of Q-SiteFinder, if not ranked as the 1st site, it is ranked as the 2nd and 4th site (figure 6.10).

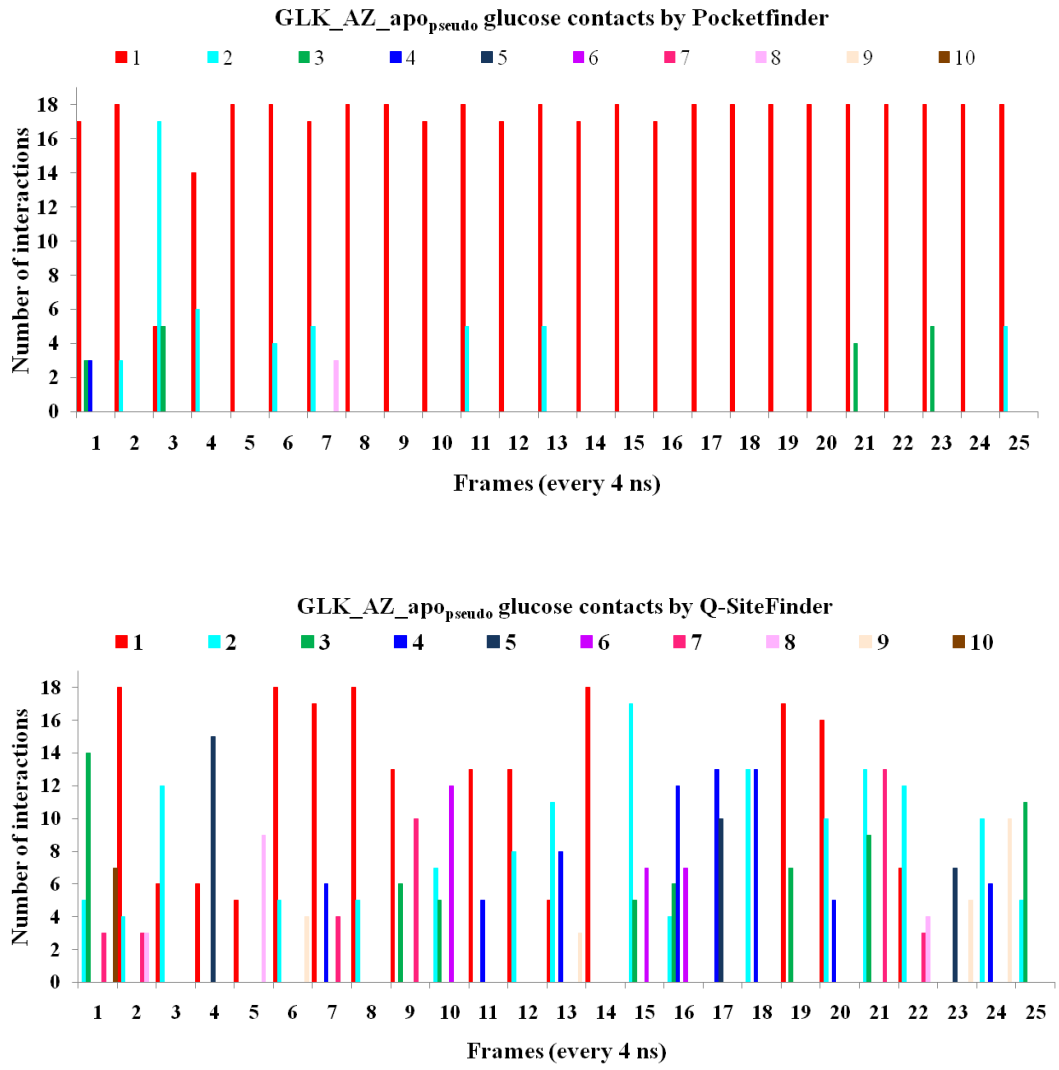


Figure 6.10: Active binding site search profile for frames from the 100 ns simulation of GLK_AZ_apo$_{pseudo}$. On top, binding sites identified by Pocket-Finder and below by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site to be counted as a site at the active or allosteric site in GLK_AZ. Numbers 1 to 10 refer to the rank of the binding sites.

In this simulation (GLK_AZ_apo$_{pseudo}$), the allosteric binding site profile (figure 6.12) is quite different to that of GLK_AZ_holo (figure 6.6). In the first 4 ns of the simulation Pocket-Finder identifies a similar site to that of GLK_AZ_holo. However, as the simulation progresses, in comparison, the rank of the allosteric site deteriorates (figure 6.11) in comparison with the holo simulation (figure 6.6). In GLK_AZ_holo, the allosteric binding site was predicted among the top 5 sites in 13 frames (out of 25), for which at least 8 interactions were identified. Here, in the GLK_AZ_apo$_{pseudo}$ simulation,

the number of predicted interactions have been considerably reduced in most frames in comparison with GLK_AZ_holo simulation (figure 6.6), as the allosteric binding site does not expand as much as the holo simulation.

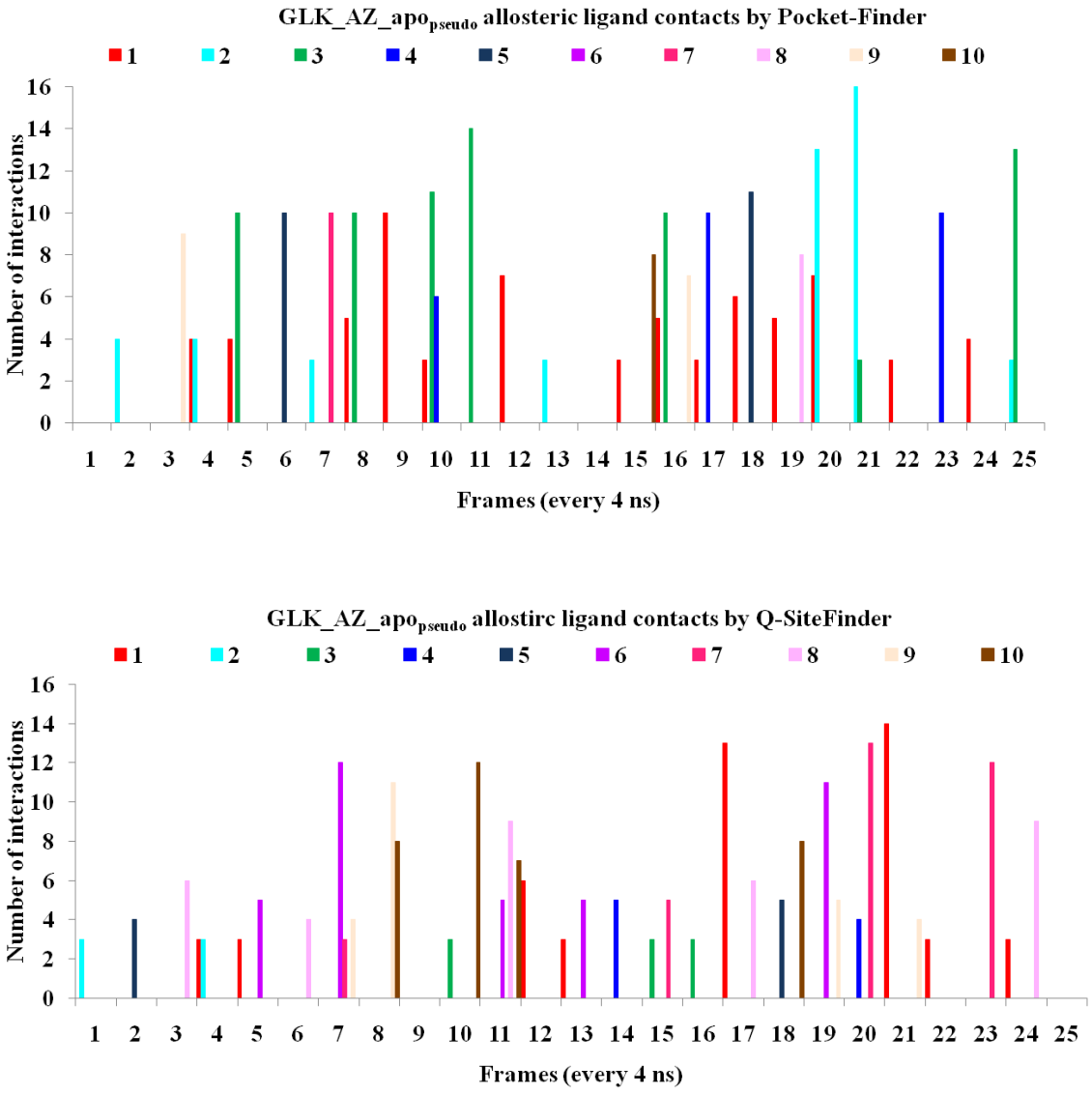


Figure 6.11: Allosteric binding site search profile for frames from the 100 ns simulation of GLK_AZ_apo$_{pseudo}$. On top, binding sites identified by Pocket-Finder and below by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site to be counted as a site at the active or allosteric site in GLK_AZ. Numbers 1 to 10 refer to the rank of the binding site.

Q-SiteFinder does not identify strong allosteric sites in most frames of the trajectory. As the RMSF of the simulation suggested, during the simulation the structure has been fairly rigid. This is a mutant structure, specifically designed to remain stable in the absence of an activator and used to co-crystallise candidate allosteric activators. It is highly likely that the mutations are the reason for the highly rigid nature of the structure. Interestingly, the simulations of 1v4s in chapter 5 suggested that glucose at the active contributes to the flexibility in the allosteric region, including the loop connecting the two domains and the β-turn (residues 91-100) close to the allosteric site. This theory is further supported by the highly rigid nature of the GLK_AZ_apo$_{pseudo}$.

The profile of all grid points identified in all 25 frames along the simulation of the GLK_AZ_apo$_{pseudo}$ does highlight the active site as a highly populated area with Pocket-Finder (figure 6.12, top) but the same cannot be said for the allosteric region. As seen in the individual frame binding site profile in figure 6.11, in many frames only a small number of allosteric site interactions have been identified in the top 10 binding sites predicted. The loop connecting the two domains is not as mobile in this simulation relative to the GLK_AZ_holo. The rigidity of this loop leads to a small site at the allosteric site that does not involve all allosteric site residues (figure 6.12, yellow circle). Overall, the allosteric site is not completely unpopulated and would be noticed as a populated site, but a smaller binding site is observed in comparison with the GLK_AZ_holo simulation.

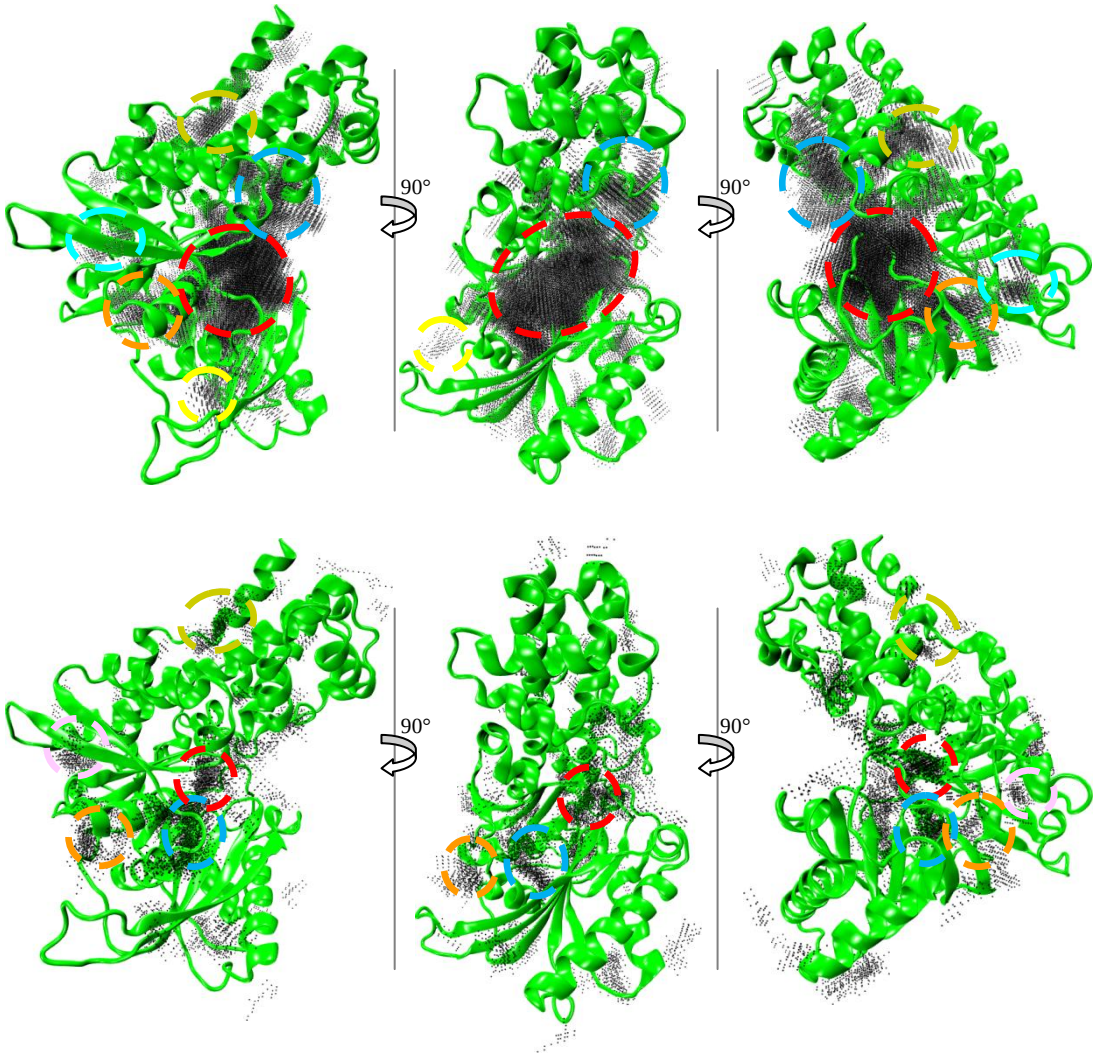


Figure 6.12: All grid points for all predicted binding sites by Pocket-Finder on top and by Q-SiteFinder below, overlaid for all 25 frames at 4 ns interval in the simulation GLK_AZ_apo$_{pseudo}$. For each frame, 10 sites were predicted. Including all 25 frames leads to 250 sites in each case. For clarity, for each method, three orientations of the sites on the protein surface have been show, rotated by 90°. Regions of high population have been circled in different colours, and repeated when rotated.

In the following section the structure of GLK_holo will be subjected to PCA to establish if the conformational changes around the allosteric binding site take place along the slow global motions of the simulation.

6.3.2 Principal component analysis of GLK_AZ_holo simulation

The allosteric binding site has been successfully identified in frames along the simulation trajectory. It is useful to establish if the conformational change required around the allosteric binding site, in particular in residues connecting the two domains, takes places along the principal modes of the simulation. This concept could be useful in reducing the binding site search space to the principal modes. In addition, in the future, essential dynamics sampling could be used to avoid the long time-scale MD simulations (*148, 264*), if this is true.

The principal components were generated by GROMACS (*181*) for the 100 ns simulation trajectory of GLK_AZ_holo, containing 5000 frames at 20 ps intervals. The simulation frames were then projected onto the principal components for further analysis.

A scree plot in figure 6.13 for the Cα PCA, demonstrates a few principal components (PCs) are sufficient to capture most of the large conformational changes in the simulation in the backbone.

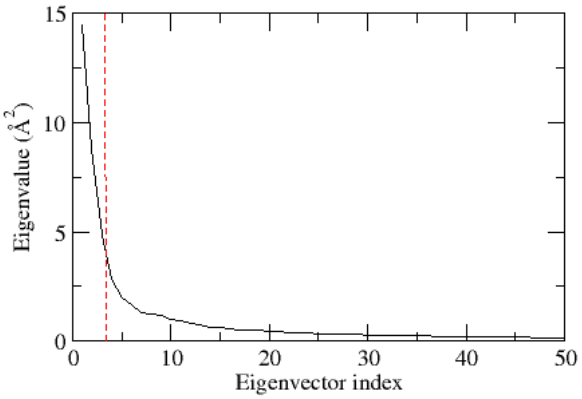


Figure 6.13: Scree plot of the first 50 principal modes for GLK_AZ_holo (PCA on Cα). The line is depicts the eigenvalues of the first three eigenvectors.

In GLK_AZ, for the allosteric binding site to open it is necessary for the connecting loop to move. Figure 6.14 depicts the global motions of the entire protein in the first three principal modes. The two extreme conformations sampled for each principal mode are illustrated with arbitrary frames in between.

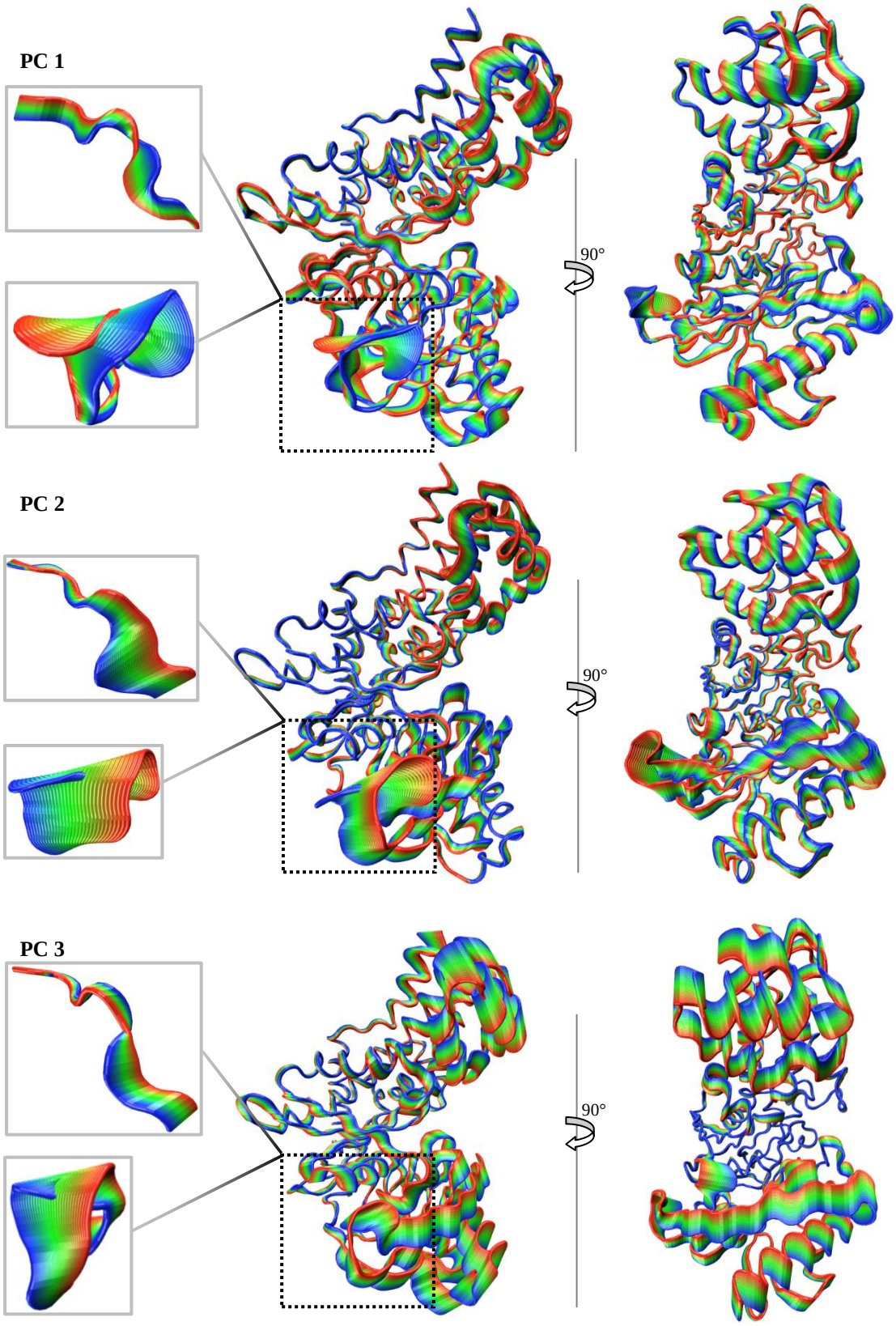


Figure 6.14: Global motions along the top three PCs. The colouring refers to the degree of sampling along the particular PC, with the extremes at red and blue. On the left, the connecting loop (on top) and the β-turn (bottom) close to the allosteric site. In the centre, the side view of the entire protein and on the right rotated by 90°.

All three modes capture significant motions of the two domains, in an opening/closing or twisting motion. The opening/closing and twisting motion in PC 1 & 2 are less than that in the 3rd PC. This may be due to the large conformational flexibility that the flexible β-turn samples along the 1st and 2nd PCs in the presence of glucose at the active site. Consistent with previous results in chapter 5 and with the RMSF plots of the GLK_AZ structure, glucose presence at the active site increases the mobility in the allosteric region.

Below (figure 6.14) the RMSF along the 1st three PCs highlights the degree of mobility sampled along the three PCs and demonstrates the degree of reduction of mobility in the backbone even beyond the 2nd PC.

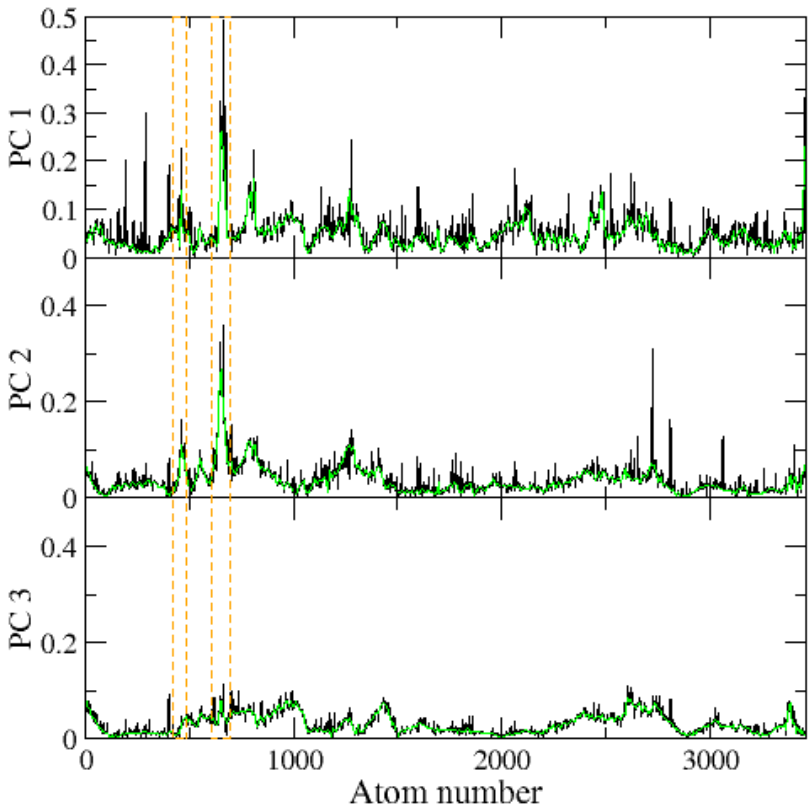


Figure 6.15: RMSF (nm) for the first 3 principal components in simulation GLK_AZ_holo for all atoms in black and Cα in green. The orange boxed regions refer (in order) to the loop connecting the two domains and β-turn, highlighted in figure 6.14.

6.3.3 Identification of binding pocket from PCA of GLK_AZ_holo simulation

The first few principal components demonstrate a considerable mobility of the allosteric region. For this reason, it is interesting to establish if the allosteric binding site would be identified considering only the first few PCs. This would allow the reduction of the binding site search space. Further advantages in the possibility of identifying the allosteric site along principal modes, arise if one could make use of principal modes for essential dynamics by improving efficiency by driving the MD simulation along the principal modes to avoid long time-scale simulations.

Before investigating the PCs, the average structure generated for the simulation, which is a prerequisite of PCA, was considered. Owing to true Cartesian averaging which causes unrealistic bond lengths and orientations of the sidechains, the average structure was subjected to 1000 steps of steepest descent minimisation, following by a 1000 steps of Conjugate gradient, while restraining the backbone (Cα, C, N, O) atoms and the C_β of the sidechains with a force constant of 500 $kcalmol^{-1}Å^{-1}$. This procedure was repeated for every trajectory frame projected onto the principal modes, used for binding site search, in order to optimise the unrealistic bond lengths and orientations.

Binding site search using Pocket-Finder on the average structure reveals two small pockets in the allosteric site region, ranked as the 2nd and 3rd. Similarly Q-SiteFinder finds the 3rd site predicted by Pocket-Finder, but it does not predict a similar site to the 2nd ranked site by Pocket-Finder (figure 6.15).

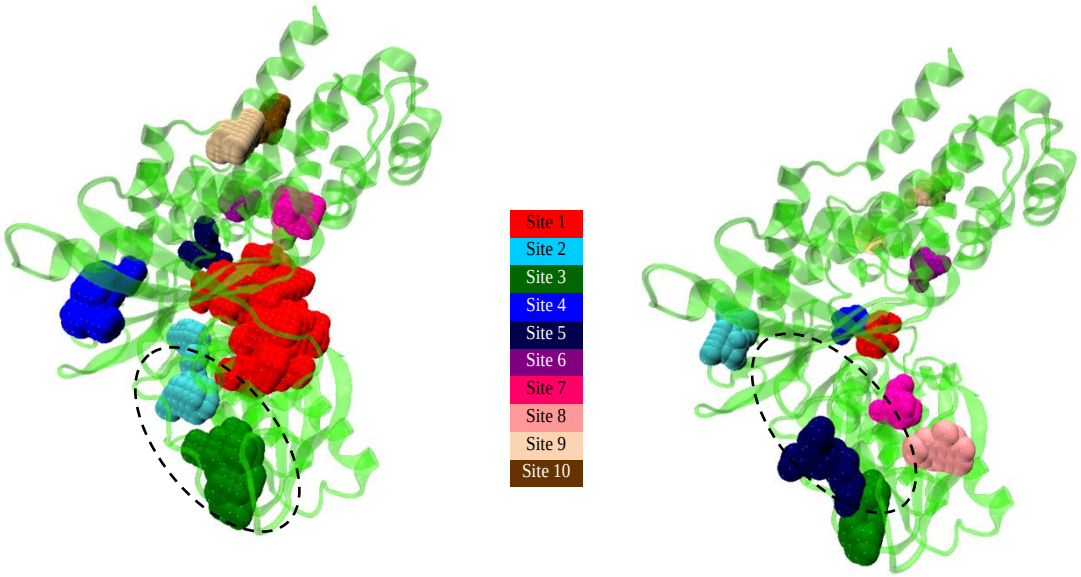


Figure 6.15: Binding sites predicted by Pocket-Finder on the left and Q-SiteFinder on the right, for the average structure in GLK_AZ_holo simulation. The binding site rank is indicated by the table in the centre. The allosteric site relevant pockets have been circled.

Both tools predict a large binding pocket, ranked 3rd in the region between the connecting loop and the flexible β-turn previously discussed. The allosteric binding site is still relatively tight in the average structure. In figures 6.16-6.17, the binding site profiles of snapshots along the 1st three principal modes are presented. Similar analysis along the 1st PC does improve the allosteric binding site size and rank (data not shown), however, the profile improves further with the inclusion of three PCs.

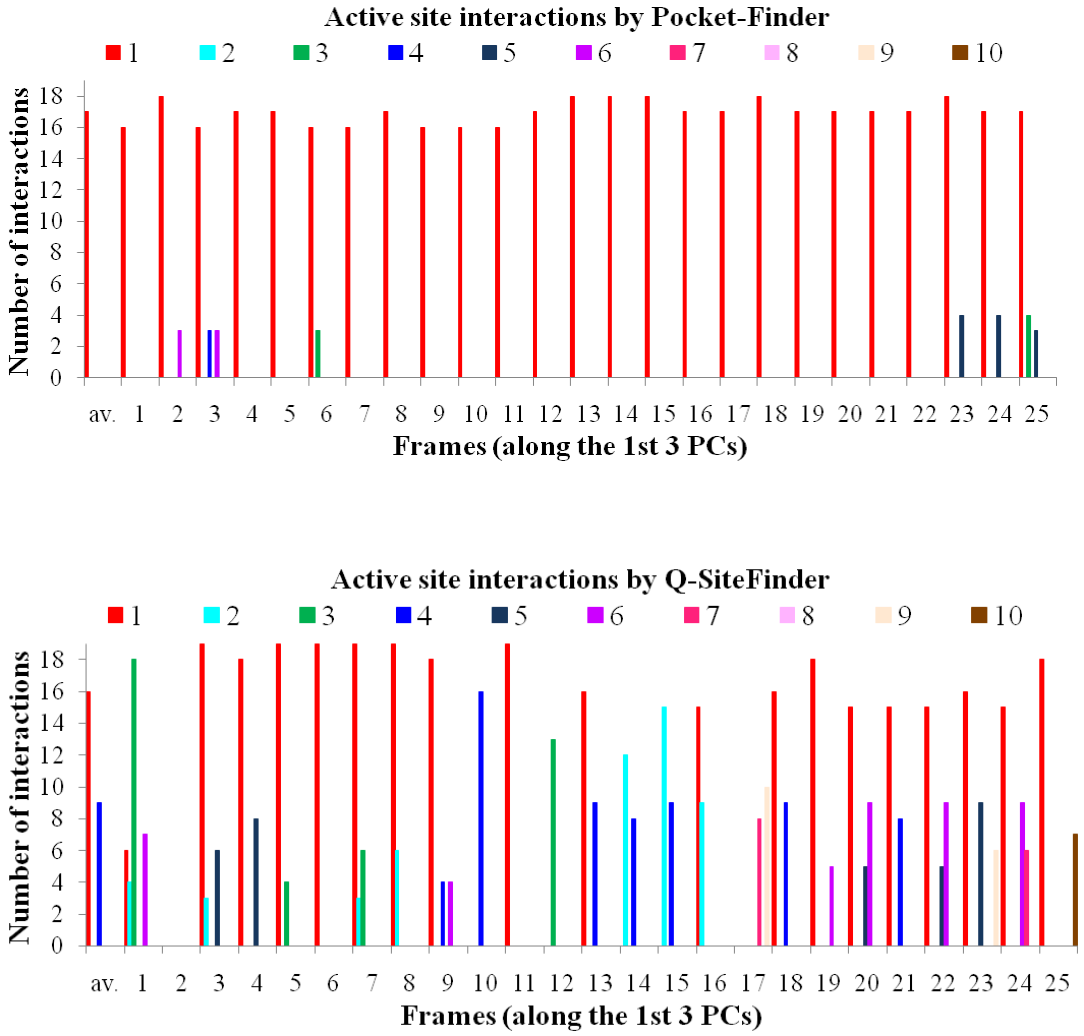


Figure 6.16: Active site binding pocket profiles for GLK_AZ_holo, PCA average structure and 25 frames along the projection of simulation trajectory onto the 1st 3 PCs. On top Pocket-Finder and bottom with Q-SiteFinder.

The active site is top ranked among the top three PCs, consistent with the average structure, with Pocket-Finder (figure 6.16), which started as a top ranked binding site by Pocket-Finder in the starting structure (figure 6.4).

The active site rank is already improved in the average structure, predicted by Q-SiteFinder. In the starting structure, the active site was ranked 2nd with Q-SiteFinder (figure 6.4). In addition, the binding site profile along the trajectory is improved along the principal modes, in comparison with the frames along the trajectory. Here, residues

of the active site have been ranked 1st 20 times (out of 25) with Q-SiteFinder, in comparison with 12 (out of 25) along the trajectory (figures 6.5 and 6.16).

Similarly, below the allosteric binding site profile has been depicted along the top three PCs (figure 6.17).

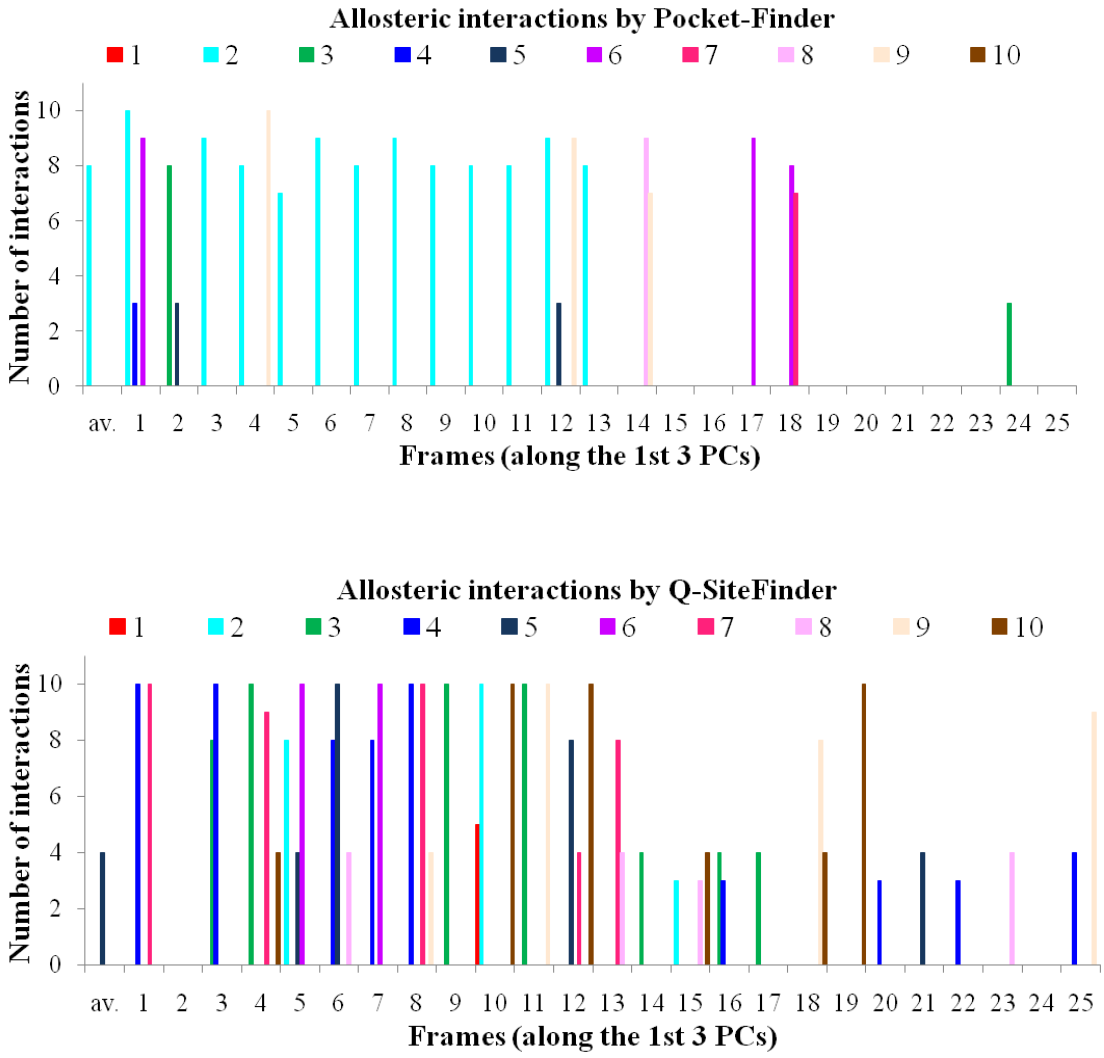


Figure 6.17: Allosteric site binding pocket profiles for GLK_AZ_holo, PCA average structure and 25 frames along the projection of simulation trajectory onto the 1st 3 PCs. On top, Pocket-Finder and bottom with Q-SiteFinder.

The allosteric binding site is not as large as seen along the standard trajectory (figure 6.6) with Pocket-Finder. However, the rank has improved and fewer segmentations of the pocket are observed here, in comparison with the binding site search along the trajectory frames, where along with a potential small site at the allosteric region, additional sites were observed as an extension to the allosteric site. Later frames 14-25, either do not demonstrate a binding site in the region or have low rank. PCA captures the major motion; at one end of the principal modes suitable conformation for binding are observed but also conformations of the allosteric region exist in the conformational ensemble, in particular the loop connecting the two domains that are not suitable for allosteric activator binding.

In comparison, Q-SiteFinder identifies small pockets at the allosteric site in these frames (14-25). Q-SiteFinder demonstrates an improvement to the allosteric site prediction, in comparison with average structure, for which a small site ranked 5th was predicted. In comparison with the allosteric binding site profile along the trajectory (figure 6.6), consistently smaller binding sites are predicted along the principal modes. This may be due to the conformational restriction imposed by reducing the trajectory to the 3 principal modes, which sample less of the side-chain flexibilities in comparison with the entire trajectory and the sensitivity of Q-Site finder to small changes in the structure.

Interestingly, binding site searching along only the principal modes does not appear to significantly deteriorate the binding site profiles on either binding sites, and in fact with Pocket-Finder, binding site search along the 1st three PCs appears to improve the prediction by focusing the trajectory on the important conformational changes in the trajectory.

6.4 Normal mode analysis

An identical protocol described to that in section 5.3 was applied to the GLK_AZ structure. Two starting structures were prepared, summarised in table 6.1.

Normal mode analysis starting structures	
NMA A	GLK_AZ in complex with glucose
NMA B	GLK_AZ glucose removed (pseudo apo)

Table 6.1: Normal mode analysis starting structures for the GLK_AZ X-ray structure (unpublished).

Despite the similarity of the protocol to that of 1v4s, in the GLK_AZ structure, the normal mode analysis leads to an RMSF plot (figures 6.19-6.20) that appears to demonstrate higher degree of residue mobility. In the holo form, NMA (A), hardly any difference is observed in the inclusion of different number of normal mode RMSFs, in comparison with that of 1v4s (figure 5.15 & 5.17), indicating that the first 5 normal modes capture all the backbone mobility in this structure in the presence of glucose.

The GLK_AZ normal modes (NMA A & B) do not capture the mobility of the residues at the beginning of the connecting loop (residues 64-71) which is required for the allosteric pocket opening in this structure.

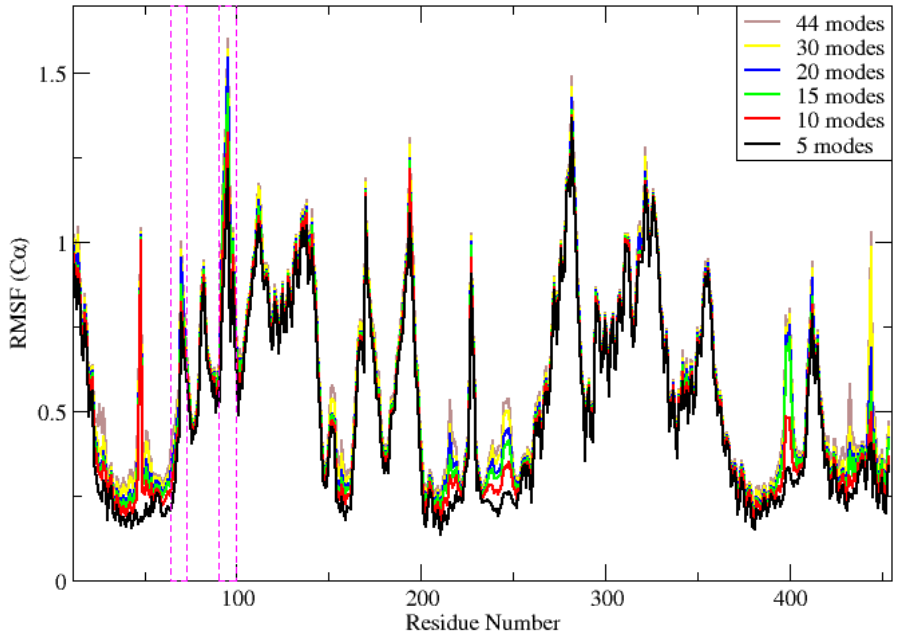


Figure 6.19: An overlay of RMSFs for normal modes for NMA (A), GLK_AZ_holo, in the presence of glucose. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site.

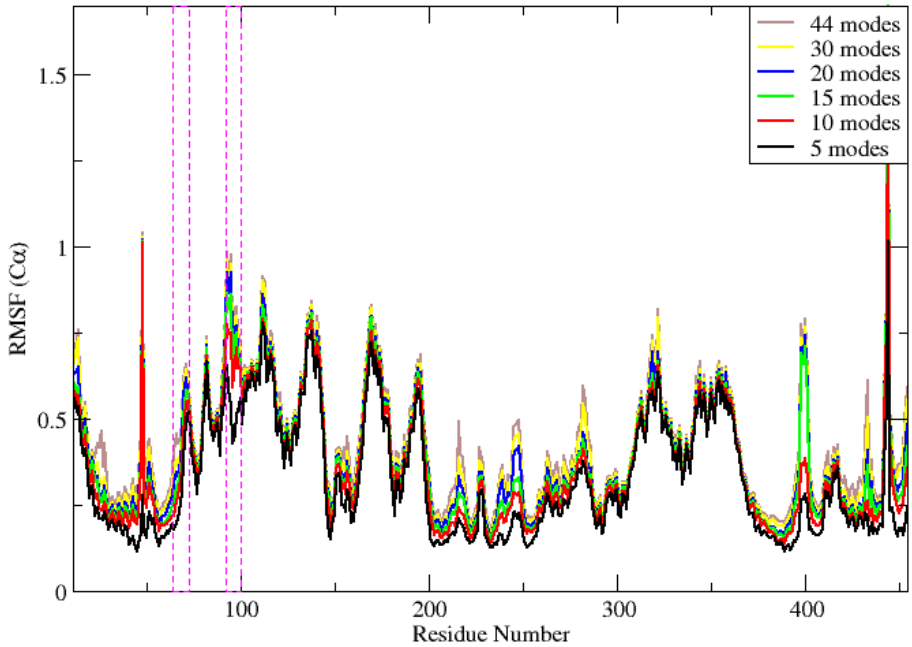


Figure 6.20: An overlay of RMSFs for normal modes for NMA (B), GLK_AZ_apo$_{pseudo}$, in the presence of glucose. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site.

The unusually high fluctuation in the NMA (A) where GLK_AZ is in complex with glucose is difficult to explain. As NMA is heavily dependent on the starting structure, small variations in conformation can influence the motion. Figures 6.21-6.22 depict the conformations of the GLK_AZ structures before and after simulation.

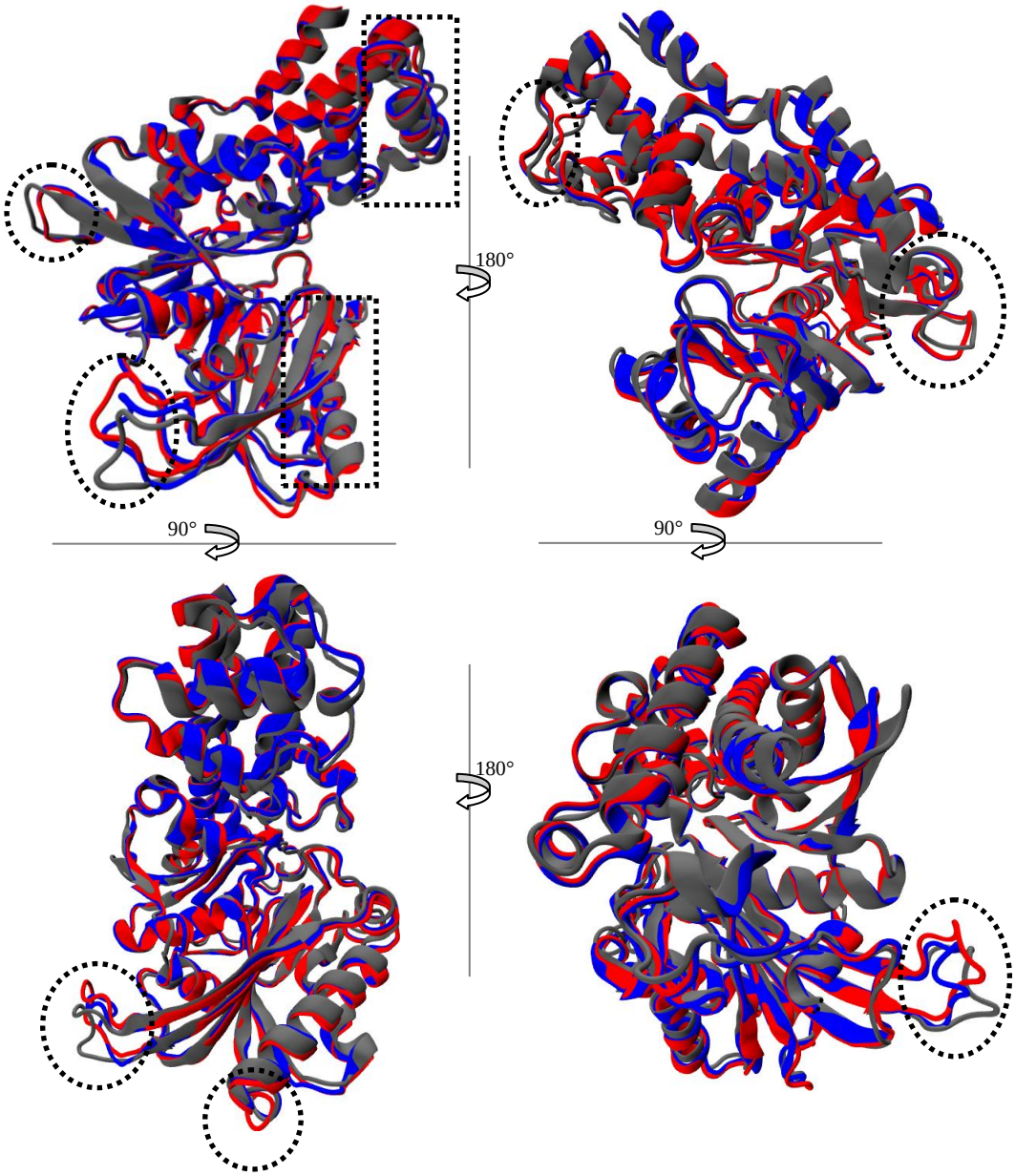


Figure 6.21: An overlay of the mutant NMA (A), GLK_AZ in complex with glucose, before (blue) and after minimisation (red) with 1v4s (gray), rotated in different orientations. Circled in black are regions with highest difference. The rectangles includes region referred to as the tip of the large and small domains in the text.

Figure 6.21: An overlay of the GLK_AZ in complex with glucose, after minimisation (red) and GLK_AZ in the absence of glucose after minimisation (green), rotated in different orientations. Circled in black are regions with highest difference.

Small conformational variations are observed in the GLK_AZ minimised structures for NMA (A) and (B). In interestingly these variations are very close to the allosteric binding site (figure 6.20-6.21).

In contrast to NMA (A), NMA (B) demonstrates less intensity in the residue RMSFs, in particular in residues at the tip of the large and small domains (see figure 6.20). The RMSF profile of NMA (B), in the absence of glucose, demonstrates similar fluctuations to that of 1v4s (figure 5.17).

The visual representation (figure 6.22) of the residue fluctuations for NMA (A), GLK_AZ in complex with glucose, highlights the degree of mobility in the entire protein. Despite higher residue mobility of the GLK_AZ structure in comparison with 1v4s in chapter 5 (figure 5.19), the loop connecting the two domains does not displace sufficiently to open the allosteric binding site.

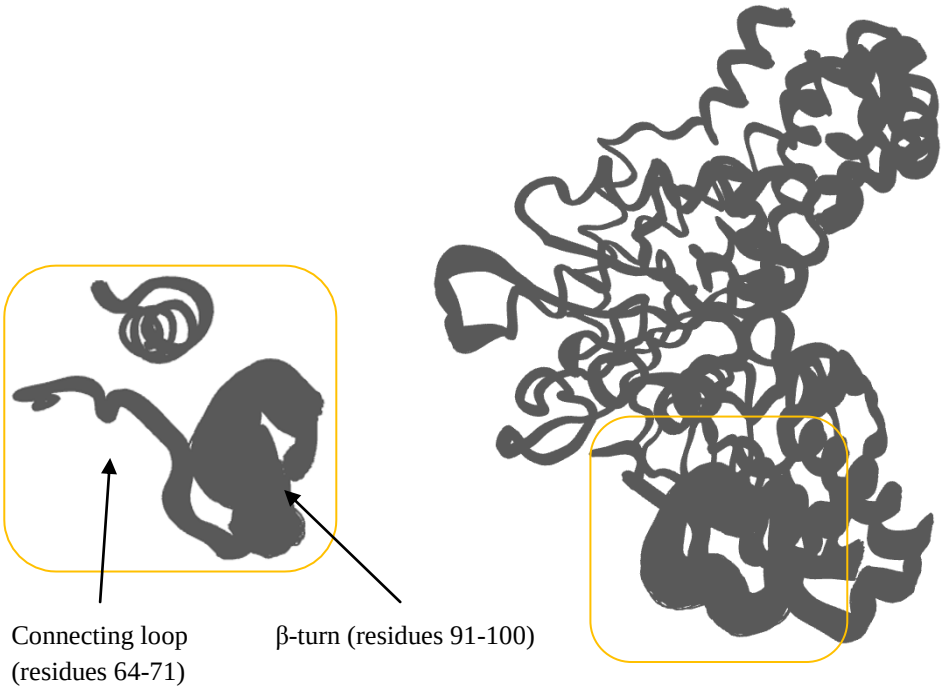


Figure 6.22: Residue mobility of the top 20 normal modes in NMA (A), GLK_AZ in complex with glucose. On the left, the allosteric region zoomed in. On the right, the entire protein.

A binding site search on the minimised structure of GLK_AZ_holo, in complex with glucose leads to the identification of a small site in the allosteric region, ranked as 6th by Pocket-Finder and no pocket predicted at that site by Q-SiteFinder, apart from small pockets close to the flexible β-turn near the allosteric site. This is in contrast to the starting structure with no pocket predicted in the allosteric site by either method (figure

6.4). Additionally, the small pocket near the flexible β-turn is reduced in rank in the minimised structure, predicted by Pocket-Finder.

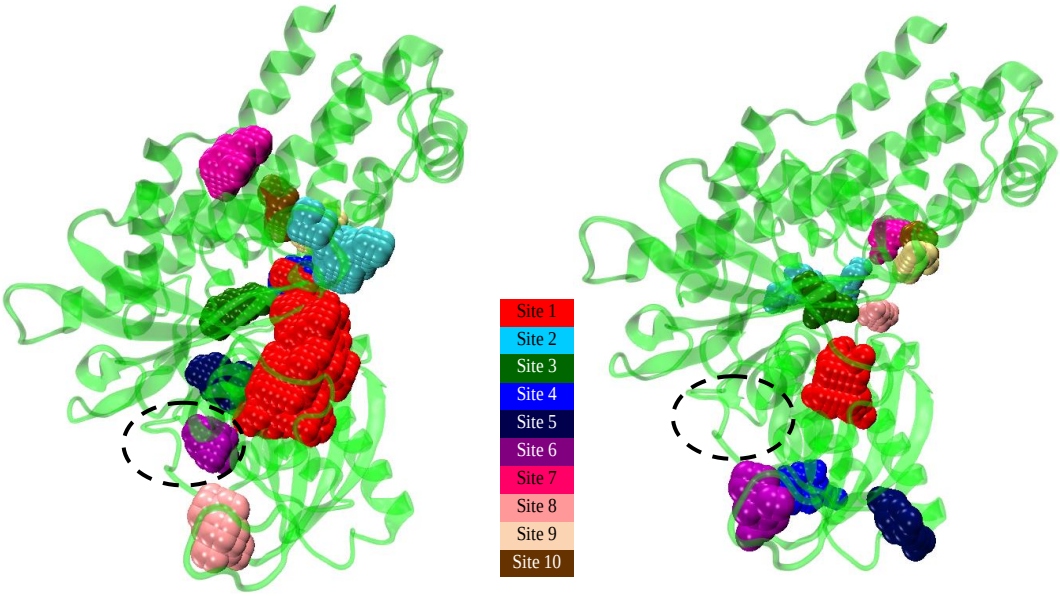


Figure 6.23: Binding pockets predicted by Pocket-Finder on the left and Q-SiteFinder on the right for the minimised structure of GLK_AZ in the presence of glucose (NMA A), prior to normal mode analysis. The binding site rank is indicated by the table in the centre. The hypothetical allosteric site is circled in black.

NMA on this system does not demonstrate sufficient mobility of the loop connecting the two domains to find the allosteric binding pocket. The wild-type sequence of this structure will be discussed in the following chapter to establish the influence of the mutations in the GLK_AZ structure on the NMA outcome.

6.5 Summary

The GLK_AZ structure, a mutant X-ray structure (unpublished) with a 97% sequence identity with the publicly available closed state GLK X-ray structure (PDB ID: 1v4s) was studied in this chapter. The major difference in conformation between the GLK_AZ structure and 1v4s lies in the residues connecting the two domains, the β-turn

close to the allosteric site and the c-terminus α-helix. Three residues in GK_AZ X-ray structure were not resolved, highlighting a degree of flexibility.

The RMSF plots of the two simulations of this structure, the GLK_AZ_holo and GLK_AZ_apo$_{pseudo}$ (figure 6.3) demonstrated the degree of flexibility of the allosteric region in the presence of glucose at the active site. In the absence of glucose, residues in the simulation sampled similar mobility to that in the holo, with the exception of the connecting loop (residues 64-71) and the β-turn (91-100) close to the allosteric site, which demonstrated lower fluctuations. This observation is consistent with the simulations of 1v4s in the previous chapter (figure 5.5).

Binding site search on the starting X-ray structure demonstrated that a pocket could not have been identified at the allosteric site, in the static X-ray structure. Yet, this is an X-ray structure that has been used to co-crystallise with the allosteric activators at AstraZeneca (unpublished data). In this structure the 3rd ranked site identified by Pocket-Finder and 10th site by Q-SiteFinder correspond to a region in between the allosteric site and the flexible β-turn nearby. This region may be an extension that may be exploited by the larger allosteric activators.

Binding site search by both methods on MD simulation frames demonstrated that in both simulations the active site remains fairly stable. The allosteric binding site is predicted more consistently in the holo simulation. This is to be expected as glucose binding does shift the equilibrium to the closed state. Simulations in chapter 5 also suggested that the allosteric binding site becomes more flexible in the presence of glucose at the active site. The ranking of the allosteric site however does not always remain high in the GLK_AZ_holo simulation. Although the pocket may have not been reliably predicted purely based on the binding pocket profiles, the ensemble of all the grid points from the 25 frames in each simulation, would give a strong indication of a potential site. In combination with the knowledge of the activating mutations, the allosteric may have been identified as a potential region to explore.

Interestingly, in the average structure of the GLK_AZ holo simulation, Pocket-Finder predicts a pocket at the allosteric site, ranked 2nd despite the relatively tight conformation of the allosteric region in this structure. Q-SiteFinder does not predict a site directly at the allosteric site, but does predict the extension region, ranked 3rd.

PCA analysis on the GLK_AZ_holo simulation highlighted that the first three PCs should be sufficient to sample the majority of the global conformational fluctuations. Visualisation of the three PCs, also demonstrated high mobility in the allosteric region (figure 6.14). Binding site search along the filtered trajectory of the first 3 principal modes, showed an improvement on the average structure in terms of number of interactions identified at the allosteric site, i.e. the size of the binding pocket, for which the rank remained as 2nd when predicted by Pocket-Finder (figure 6.18) . A considerable improvement was observed with the Q-SiteFinder prediction in comparison with the average structure, in terms of the size of the pocket; however, the rank does not remain consistently high. In both plots, the allosteric binding site is either not predicted or weakly predicted by both methods, after frame 14. This is to be expected as the conformational ensemble along the principal modes will include conformations of the allosteric region, in particular the connecting loop that will not be suitable for binding.

The binding site prediction along PCs demonstrates that it may be possible to compress the trajectory to the principal modes, owing to the significant storage challenges that are associated with the large data generated by MD simulations (*265*). In addition, the use of essential dynamics sampling may be another possibility in focusing on the principal motions of the systems, hence gaining more sampling (*149*).

Normal mode analysis did not capture sufficient mobility of the allosteric region to reveal the allosteric binding. Vibrations around a highly stable minimised structure in this case were not sufficient to reveal a binding site that requires a relatively large displacement of the connecting loop at the allosteric site.

In the following chapter a comparison of the GLK_AZ structure is made to the wild-type homology model, to highlight the potential influences of the mutations on the observations seen here.

Chapter 7

GLK_AZ wild-type sequence homology

7.1 Aim

In the previous chapter the structure of GLK_AZ, a mutant structure, was studied. The mutations were made to aid crystallisation due the low stability of the active closed state GLK. The allosteric binding site was revealed by MD, but not by NMA. In this chapter, the influence of the mutations are discussed by generating a homology model of the sequence of the publicly available structure (PDB ID: 1v4s) with respect to the secondary structure conformation of GLK_AZ, which will hereafter be referred to as the $GLK\_AZ_{\text{wild-type homology}}$ to ensure that the results were not biased by the mutations. For sequence comparison of GLK_AZ and 1v4s refer to appendix A and for the exact locations of the mutations in GLK_AZ to figure 6.1.

Once again molecular dynamics and normal mode analysis were applied to study the dynamics of this homology model, in the presence and absence of the active site ligand, glucose. This was followed by binding sites search.

7.2 System preparation and MD parameters for the simulation of $GLK\_AZ_{\text{wild-type homology}}$

For the $GLK\_AZ_{\text{wild-type homology}}$ structure, the sequence of isoform 2 also in (PDB ID: 1v4s) was used to generate a homology model based on the secondary structure of

GLK_AZ. This was carried out externally (*219*) using PRIME, a Schrödinger tool (*261*). For direct comparison with the GLK_AZ simulations, two simulation starting structures were prepared; one in the presence of glucose at the active site ($GLK\_AZ_{wild\text{-}type\ homology}$-holo hereafter) and another, in the absence of glucose from the active site, as a pseudo apo ($GLK\_AZ_{wild\text{-}type\ homology}$-$apo_{pseudo}$ hereafter).

Similar to the other normal modes set-ups in previous chapters, for each simulation, if applicable the glucose ligand was protonated via the PRODRG (*247*) web server, and the protein in WHATIF (*248*).

Similar to the mutant structure both simulations have been carried out using the AMBER (*114*) molecular dynamics package using the AMBER03 force field for the protein and gaff force field, AM1-BCC charge method (*249, 250*), for the ligand, with Particle-Mesh-Ewald (PME) boundary condition limited to a 10 Å cutoff.

Amber package, tool XLEAP was used to solvate the system with TIP3P water models, with a minimum distance of 12 Å from the protein (19287 water molecules). A total of 22 sodium ions were added to both $GLK\_AZ_{wild\text{-}type\ homology}$-holo and $GLK\_AZ_{wild\text{-}type\ homology}$-$apo_{pseudo}$ to neutralise the overall charge in each case.

The minimisation was carried out in stages, starting with the minimisation of solvent only while applying a restraint force constant of 500 kcal mol^{-1} $Å^{-2}$ on the rest of the particles, followed by the minimisation of protein with restraint on other particles, then the ligands, if applicable, and finally the entire system, removing all restraint; in each stage, reaching an energy value below the set root-mean squared deviation of 0.001 kcal-$mol^{-1}Å^{-1}$. In each case, the minimisation commenced with 100 steps of steepest descent, followed by conjugate gradient for the remainder.

For each, the minimised system was heated gradually to 300 K in the NVT ensemble, in six 50 K blocks, allowing 15000 MD step for each block, using a Langevin thermostat with a 1 ps^{-1} collision frequency.

The production MD simulations were run in the NPT ensemble, using Langevin thermostat with a 1 ps^{-1} damping parameter at 300 K and a time-step of 2 fs. The

pressure was controlled by isotropic position scaling with a relaxation time of 2 ps. All bonds containing hydrogen were constrained using the SHAKE algorithm. The non-bonded cut-off of 10 Å was employed.

For each simulation, 100 ns of MD production was collected after removing the initial ~1.5 ns assuming equilibration at the beginning of the NPT production run, based on stabilisation of properties such as energy, volume, density and RMSD.

7.3 MD analysis for GLK_AZ$_{\text{wild-type homology}}$

There is no backbone conformational variation in the homology starting structure in comparison with GLK_AZ, for which an overlay has been depicted in figure 7.1.

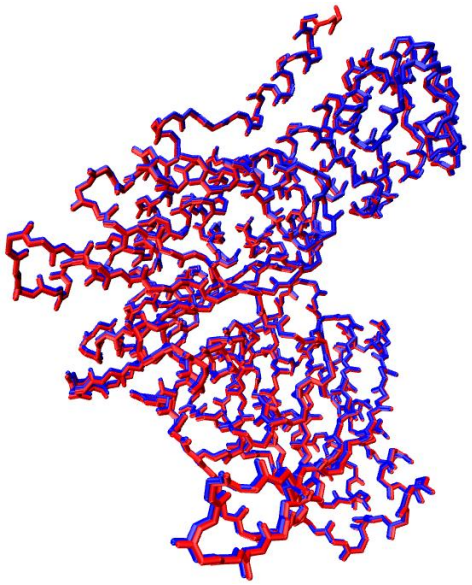


Figure 7.1: An overlay of the backbone atoms of GLK_AZ in red and the GLK_AZ$_{\text{wild-type homology}}$ in blue, starting structures.

The major differences in conformation between the GLK_AZ structure and that of the publicly available closed state structure (PDB ID: 1v4s) have been highlighted in the previous chapter (figure 6.2). Here both GLK_AZ$_{\text{wild-type homology}}$ starting structures will have the same conformational differences with respect to 1v4s, which are residues in the connecting loop (residues 64-71), the β-turn close to the allosteric region (residues 91-100) and the c-terminus α-helix and the loop leading to it (residues 441-461).

Results from MD simulations, GLK_AZ$_{\text{wild-type-homology}}$-holo and GLK_AZ$_{\text{wild-type homology}}$-apo, are presented here. A Comparison of the RMSF plots (figure 7.2) demonstrates the structural mobility in the two simulations.

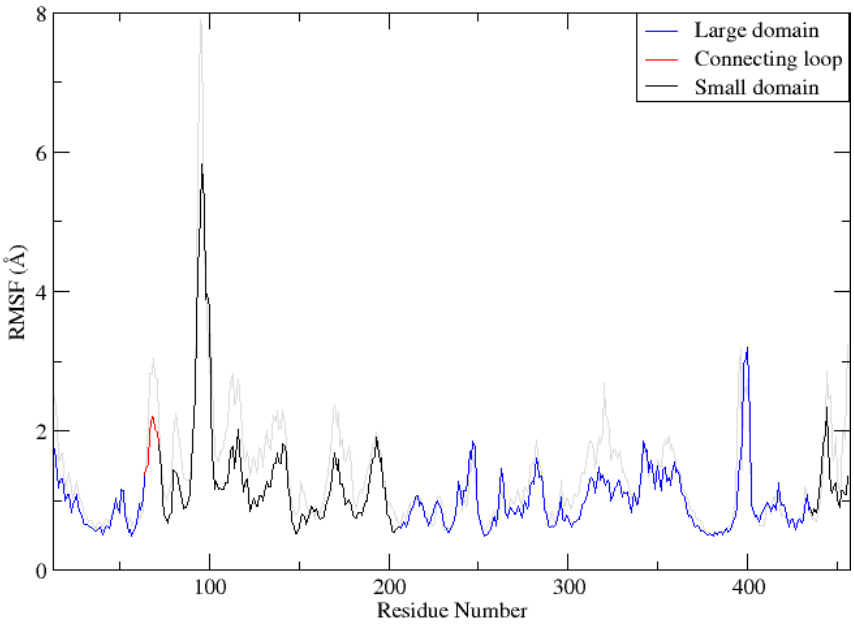


Figure 7.2: RMSF (Cα) GLK_AZ$_{\text{wild-type homology}}$-holo and GLK_AZ$_{\text{wild-type homology}}$-apo (gray).

A greater degree of mobility is observed in simulation GLK_AZ$_{\text{wild-type homology}}$-apo in comparison with GLK_AZ$_{\text{wild-type homology}}$-holo (figure 7.2), in particular in residues of the small domain. The connecting loop (residues 64-71) and the β-turn (residues 91-100) are more mobile in the apo simulation, which is opposite to the mutant simulations, discussed in the previous chapter (figure 6.3), and is consistent with the simulation of the original closed state structure (PDB ID: 1v4s) (figure 5.5).

A comparison of the RMSF of the two simulations here with respect to the mutant and the original closed state structures (PDB ID: 1v4s) is given in figure 7.3. It is noteworthy that the RMSFs of the 1v4s simulations are based on 25 ns simulations, in comparison with that of both GLK_AZ and GLK_AZ$_{\text{wild-type homology}}$ simulations which are based on 100 ns simulations.

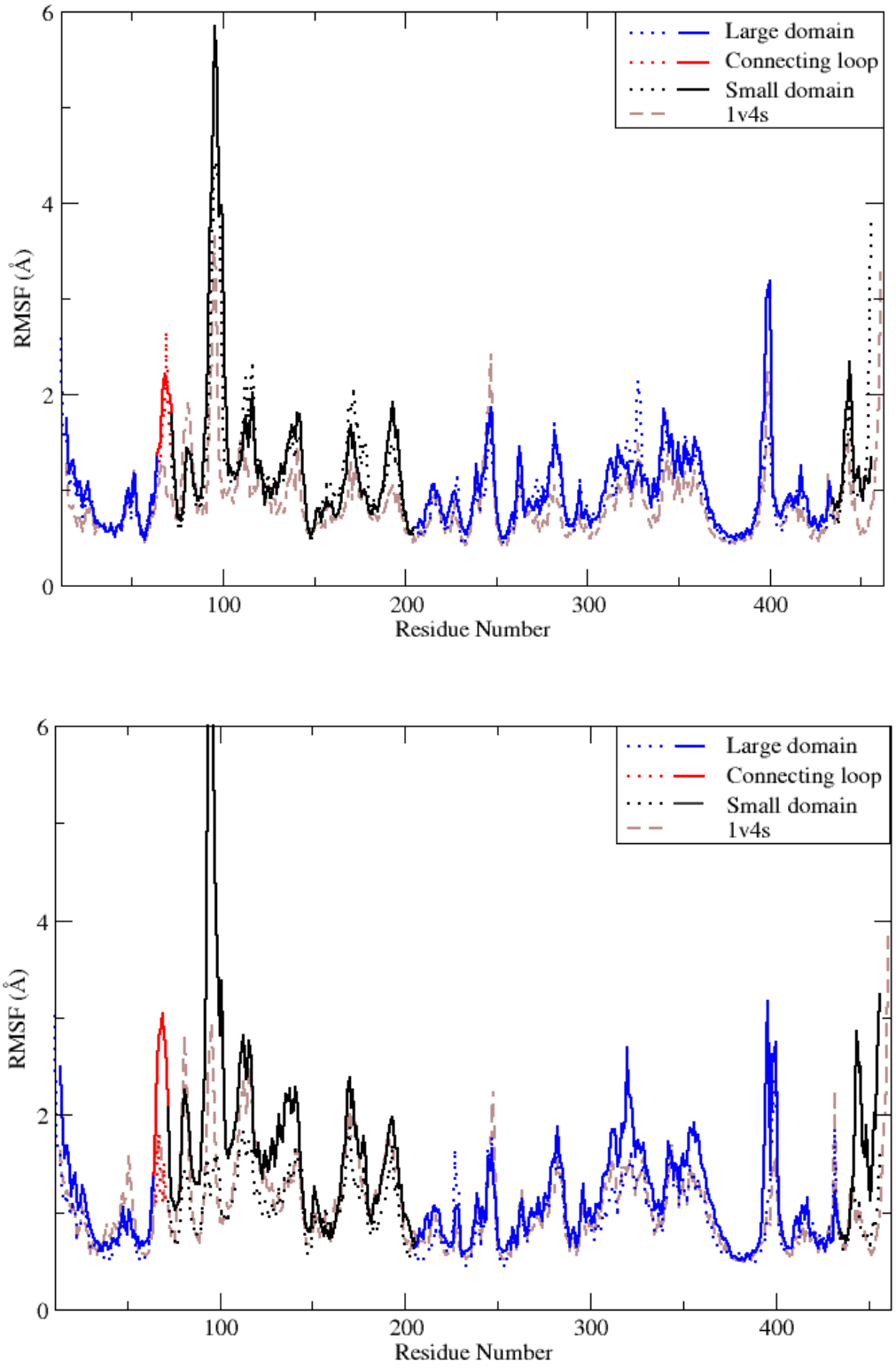


Figure 7.3: On top, an overlay of the RMSF (Cα) for simulations GLK_AZ$_{\text{wild-type homology}}$-holo (solid line), GLK_AZ_holo (dotted line) and 1v4s, simulation 'B' in presence of glucose (dashed brown line). On the bottom, the corresponding for GLK_AZ$_{\text{wild-typehomology}}$-apo, GLK_AZ_apo and 1v4s, simulation 'D', in the absence of both ligands. For the purpose of direct comparison, the plots have been limited to 6 Å on the y-axis.

The RMSFs of the holo simulations, GLK_AZ$_{\text{wild-type homology}}$-holo, GLK_AZ_holo and 1v4s (simulation 'B') (figure 7.3, on top), are fairly similar. Overall, a slightly lower mobility is observed in 1v4s, which may be due to the shorter simulation time.

In contrast to GLK_AZ_apo where mobility of the small domain residues are reduced in comparison with the holo form (figure 6.3), GLK_AZ$_{\text{wild-type homology}}$-apo demonstrates higher degrees of mobility in the apo form, particularly in the connecting loop (residue 64-71) and the β-turn (residues 91-100). This further flexibility is also observed when compared with the RMSF of 1v4s 'simulation D', in which both glucose and the allosteric activator are removed from the structure (figure 5.5). In addition, the central residues of the connecting loop (64-71), the β-turn residues (residues 91-100) and residues of a loop in the large domain (residues 398-402, see appendix B) demonstrate greater flexibility (~ 0.5-2.5 Å) in comparison with 1v4s. This may be due to longer simulation times in this chapter, or the conformation of the starting structure which is based on the GLK_AZ. In most residues of the small domain the fluctuations are more similar between the GLK_AZ_apo simulation and 1v4s (simulation 'D') than for GLK_AZ$_{\text{wild-type homology}}$-apo, but the flexible β-turn (residues 91-100) is among the residues that differ significantly between the two (GLK_AZ_apo and 1v4s (simulation 'D'), which may influence the binding site search profile in the allosteric region. Overall, we observe that in particular in the absence of glucose, the mutations have lead to a reduced mobility in the system (figure 7.3). The simulation of the wide-type sequence is highly mobile, even more so in the absence of glucose, which is consistent with a lower stability active closed state structure that does not crystallise in the absence of an activator.

For the purpose of the allosteric binding site identification, the main emphasis is on the holo simulation, in the presence of glucose, and therefore the similarity of the RMSF profile between the mutant GLK_AZ_holo and GLK_AZ$_{\text{wild-type homology}}$-holo and 1v4s (simulation 'B') is promising and should lead to a similar results in terms of allosteric binding site prediction.

In the following sections, the binding site search for frames from both the apo and holo simulations of the GLK_AZ$_{\text{wild-type homology}}$ will be discussed.

7.4 Binding site profiles in the starting structure GLK_AZ$_{\text{wild-type homology}}$

Although there are not any backbone conformational changes after the homology modelling in comparison with the GLK_AZ structure, small local variations in the side chains from the mutated structure back to the wild-type sequence demonstrate a subtle change in the binding site prediction for the starting structure (figure 7.4) in contrast to GLK_AZ starting structure (figure 6.4).

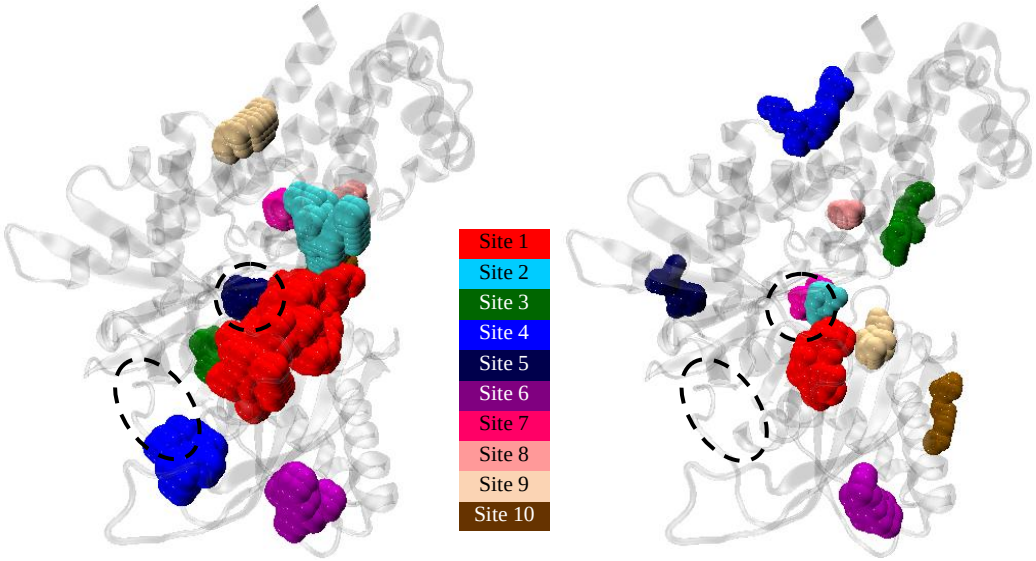


Figure 7.4: On the left, top 10 sites predicted for GLK_AZ$_{\text{wild-type homology}}$ by Pocket-Finder, and on the right the corresponding by Q-SiteFinder. In the centre is the colour code for the sites in order of rank. The dotted circle and oval indicate the location of the glucose and allosteric sites, in the original X-ray structure (PDB ID: 1v4s), respectively.

In the GLK_AZ, Pocket-Finder, predict a small site between the flexible β-turn (residues 91-100) near the allosteric site and connecting loop (residues 64-71), ranked 3rd (figure 6.4). Here, the rank for the same site has been lowered to 4th and the size of the site is reduced. Essentially, the location of 3rd and 4th ranked sites are swapped here. With Q-SiteFinder the location of predicted sites are relatively different, as Q-SiteFinder is more discriminating, no site has been predicted near the β-turn, which had been in the GLK_AZ, ranked as 10th (figure 6.4).

In the following section, the binding site search profile through the simulation will be discussed.

7.4.1 Binding site profiles in the GLK_AZ$_{\text{wild-type homology}}$-holo simulation

As expected, in the presence of glucose at the active site in GLK_AZ$_{\text{wild-type homology}}$-holo simulation, the active site remains as the 1st ranked site predicted by both methods (figure 7. 5).

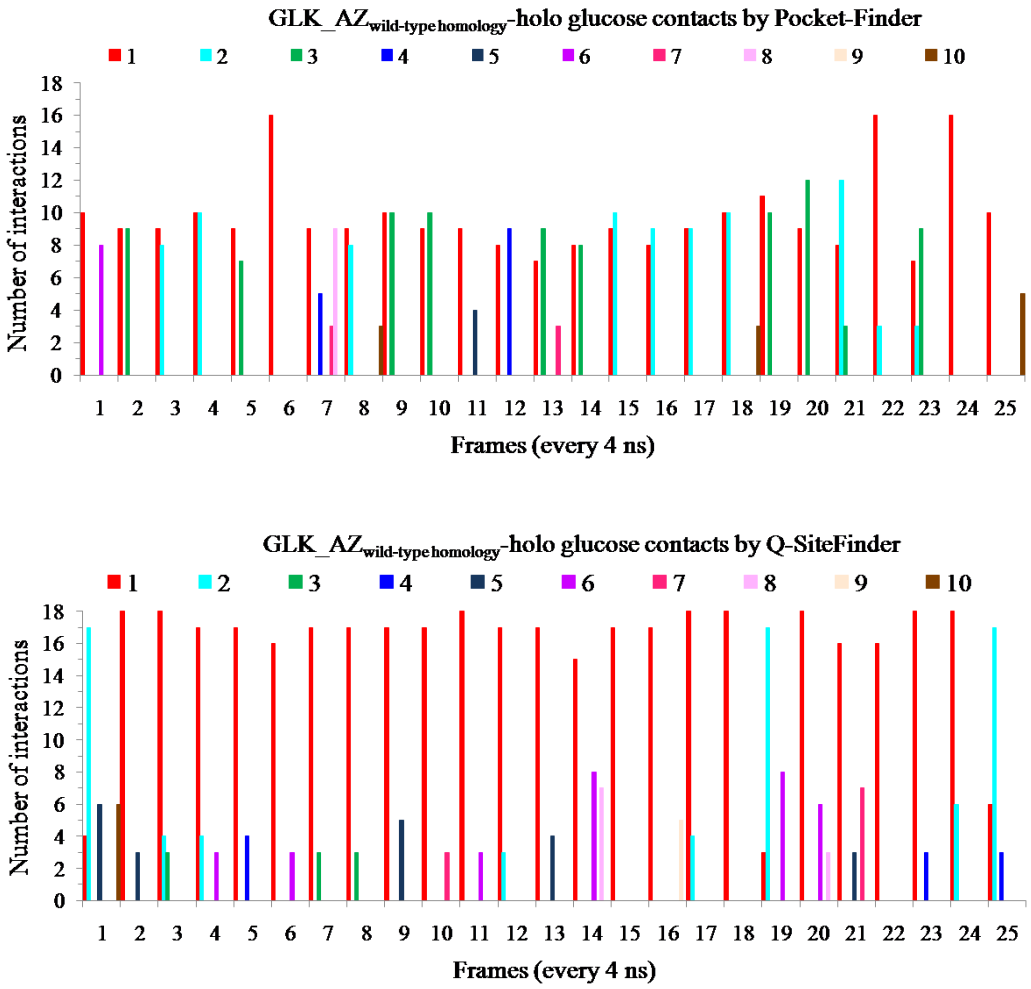


Figure 7.5: Active binding site search profile for frames from the 100 ns simulation of GLK_AZ$_{\text{wild-type homology}}$-holo. On top, binding sites identified by Pocket-Finder and on the bottom by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site to be counted as a site at the active site or allosteric site in GLK_AZ. Numbers 1 to 10 refer to the rank of the binding site.

The active site binding site profile demonstrates an improvement to that predicted for the GLK_AZ_holo simulation seen in chapter 6 (figure 6.5). Pocket-Finder does not directly map glucose at the binding site, and hence the interaction residues are split between two sites. In contrast, Q-SiteFinder perfectly maps glucose in the predictions, and hence a stronger prediction profile is observed. This is surprising since, in all previous observations, Pocket-Finder had performed with lower sensitivity and although the predicted pocket would be much larger than the actual ligand at the binding site, the ligand would not be partially outside the largest, highest ranked site.

In the case of the allosteric binding site, Pocket-Finder ranks the site as the 2^{nd} only in 7 out of 25 frames (figure 7.6) , in contrast to 13 in GLK_AZ_holo, in the previous chapter (figure 6.6). In a number of frames, allosteric interacting residues have been identified as part of the first ranked site, which also includes the active site (figure 7.5).

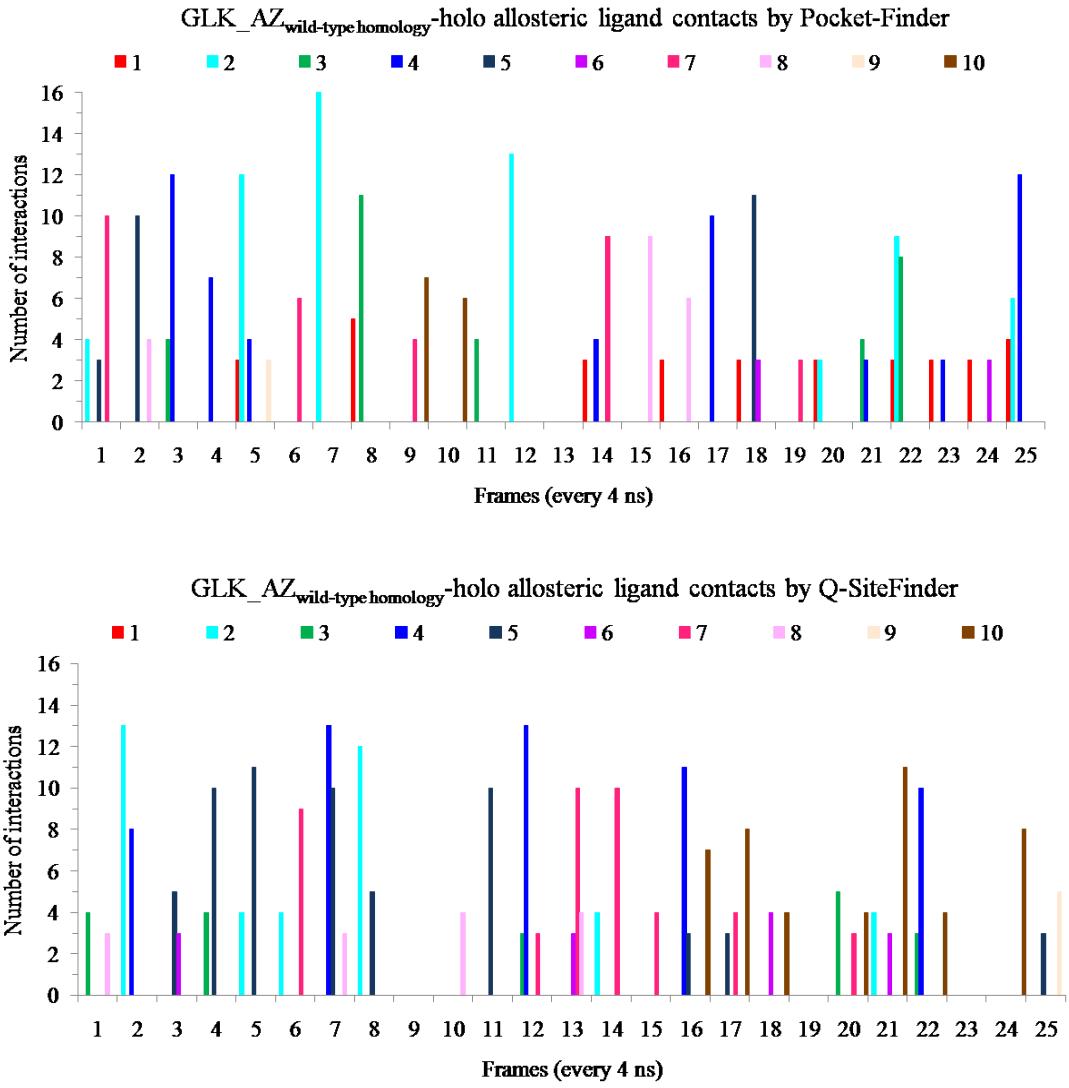


Figure 7.6: Allosteric binding site search profile for frames from the 100 ns simulation of GLK_AZ$_{\text{wild-type homology}}$-holo. On top, binding sites identified by Pocket-Finder and on the bottom by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site to be counted as a site at the active site or allosteric site in GLK_AZ. Numbers 1 to 10 refer to the rank of the binding site.

An overlay of all pocket grids points for each frame from the simulation (GLK_ AZ$_{\text{wild-type homology}}$-holo), predicted by both methods, gives an indication of the population of grid points for the 25 frames (figure 7.7). The active site is populated by a high number of grid points; it is therefore noteable, although in the case of the prediction by Pocket-Finder the size of the spread of the grid points in the region may be misleading with respect to the exact location of the glucose at the active site. With Q-SiteFinder the assembly of the grid points at the active site are more precise.

The allosteric site is occupied with grid point using both methods, but not necessarily much more noticeable than other predicted site locations (figure 7.8). In combination with the knowledge of the allosteric activators, it can be demonstrated that in comparison with static X-ray structure or in this case, the homology model, that the allosteric region becomes populated by grid points.

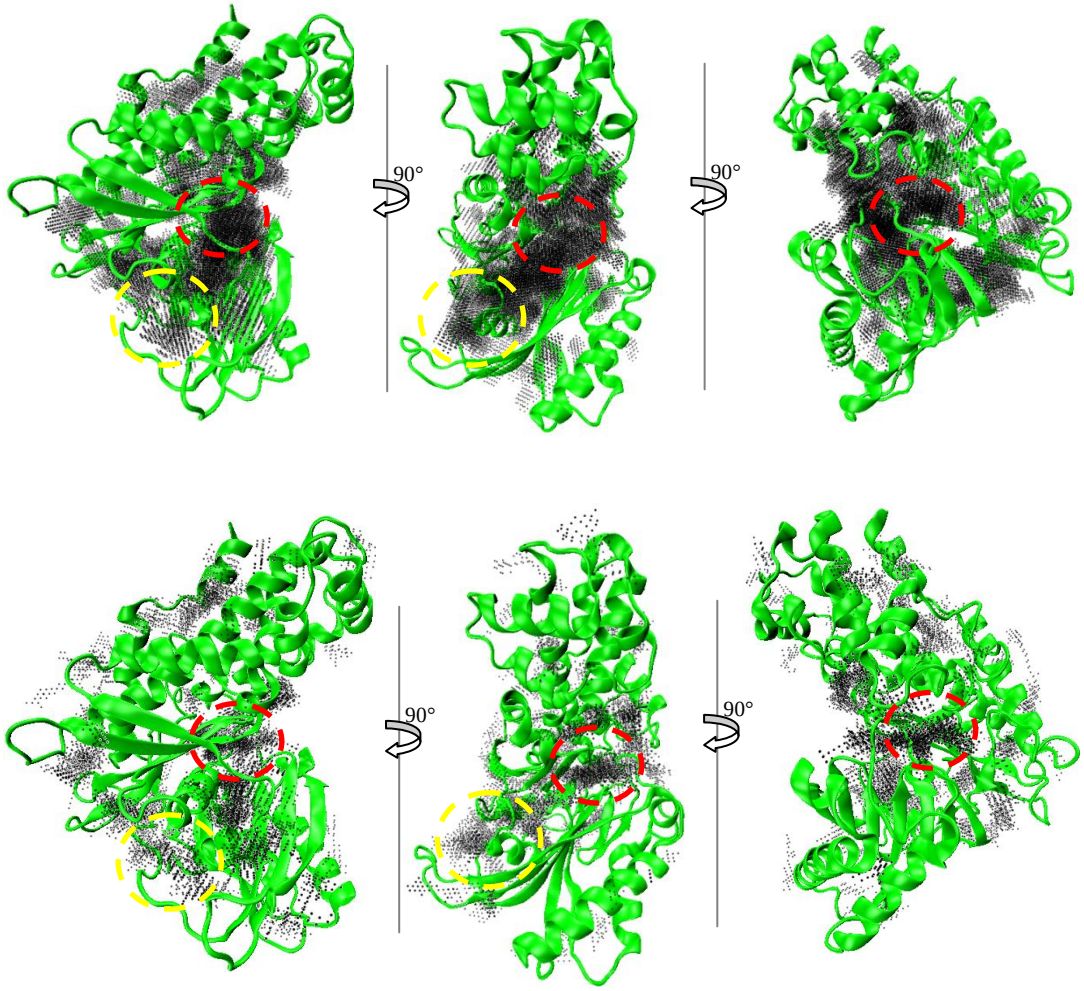


Figure 7.7: All grid points for all predicted binding sites by Pocket-Finder on top and by Q-SiteFinder below, overlaid for all 25 frames at 4 ns interval in the simulation GLK_AZ$_{\text{wild-type homology}}$-holo. For each frame, 10 sites were predicted, including all 25 frames leads to 250 sites in each case. For clarity, for each method, three orientations of the sites on the protein surface have been shown, rotated by 90°. The active and allosteric sites have been circled. The allosteric site has not been circled in the rotated orientation on the right, as the site is not visible in this orientation.

In comparison with the GLK_AZ_holo simulation, the loop connecting the two domains in this simulation (GLK_$_{AZwild\text{-}type\ homology}$-holo) does not demonstrate as much mobility, in the frames selected for analysis. The inclusion of more frames in the analysis may have a positive influence on the results; however, the processing becomes time consuming. The connecting loop mobility in the RMSF plot for this simulation (figure 7.2) is slightly less than that of the GLK_AZ_holo simulation (figure 6.3), which further supports the lower flexibility in accommodating an allosteric activator. There may be a lower population of suitable conformations of the connecting loop in this simulation ensemble in comparison to that of the mutant GLK_AZ_holo. In this simulation however, the β-turn (residues 91-100) demonstrates higher mobility in comparison with that of GLK_AZ_holo. The starting conformation of this β-sheet, is based on the conformation of the mutant X-ray structure with the wild-type sequence, which is opposite to that adopted in 1v4s (figure 7.8).

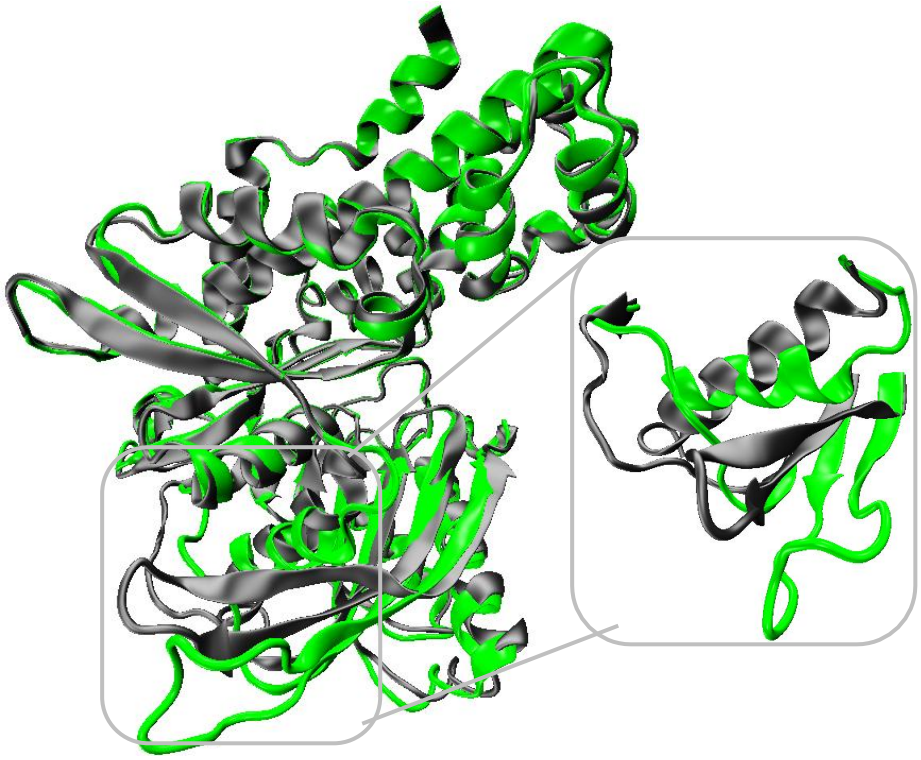


Figure 7.8: An overlay of GLK_AZ$_{wild\text{-}type\ homology}$ and the original closed state GLK (grey) (PDB ID: 1v4s). On the right, a close-up view of the allosteric region.

7.4.2 Binding site profiles in the GLK_AZ$_{\text{wild-type homology}}$-apo simulation

In the apo form, the active site is strongly ranked as the first site in all frames, apart from the first frame where a smaller site has been identified, which may result from minimisation of the protein in the absence of any ligands (figure 7.10).

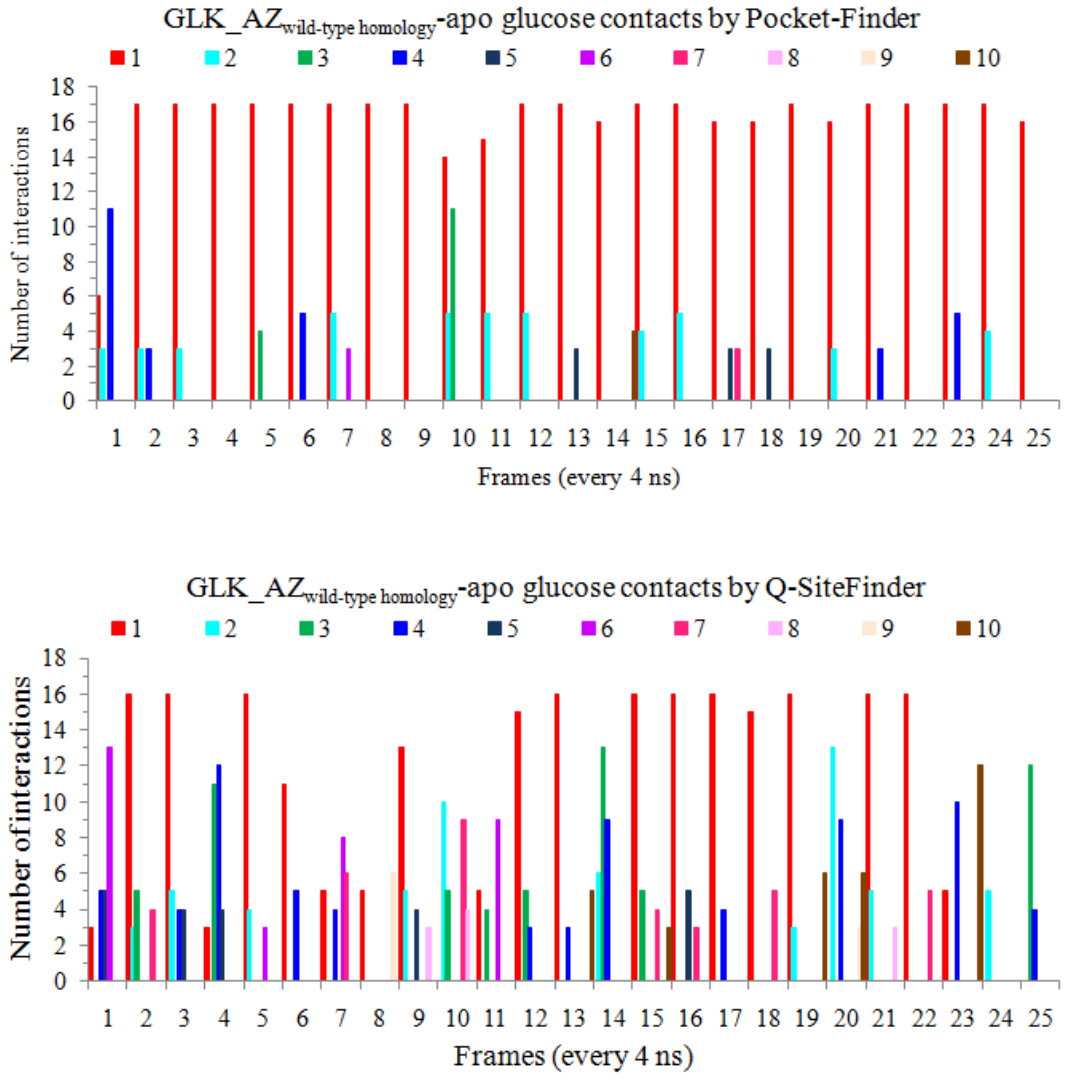


Figure 7.10: Active binding site search profile for frames from the 100 ns simulation of GLK_AZ$_{\text{wild-type homology}}$-apo. On top, binding sites identified by Pocket-Finder and on the bottom by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site to be counted as a site at the active site or allosteric site. Numbers 1 to 10 refer to the rank of the binding site.

The first frame in plots (figure 7.10), is the first frames of the production trajectory, after minimisation, heating and equilibration. The pockets identified by Pocket-Finder span a large area outside, in addition to, the actual active site, which may be as a result of the extended freedom of the protein in the apo form. The active site is additionally strongly predicted by Q-SiteFinder, in the absence of both ligands from the simulation.

Prediction of the allosteric binding site (figure 7.11) considerably deteriorates in this apo wild-type sequence, in comparison to the GLK_AZ_apo simulation (figure 6.11).

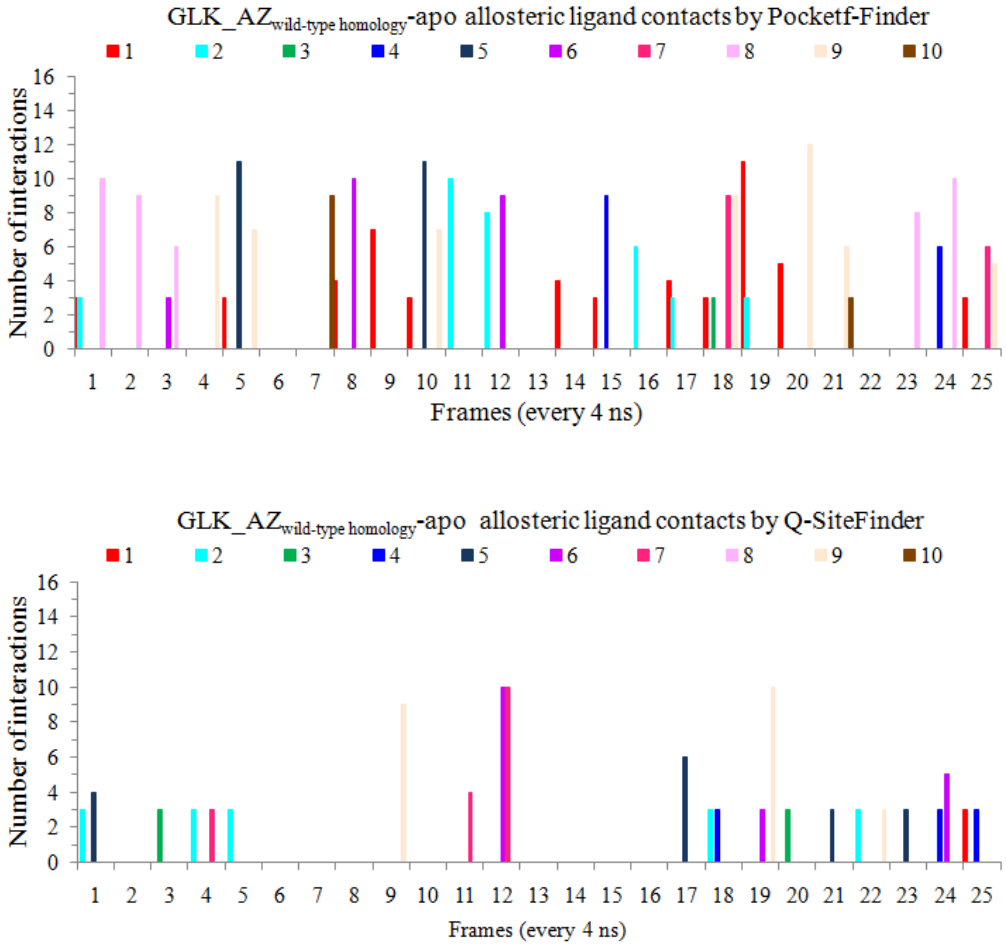


Figure 7.11: Allosteric binding site search profile for frames from the 100 ns simulation of GLK_AZ$_{\text{wild-type homology}}$-apo. On top, binding sites identified by Pocket-Finder and on the bottom by Q-SiteFinder. A minimum of 3 relevant interaction residues were required to be present in the predicted site to be counted as a site at the active site or allosteric site. Numbers 1 to 10 refer to the rank of the binding site.

In the absence of glucose at the active site, although the loop connecting the two domains (residues 64-71) and the β-turn (residues 91-100) near the allosteric site demonstrates relatively high mobility in the RMSF plot (figure 7.2), the prediction of the allosteric site remains fairly poor, in particular with Q-SiteFinder. Visual inspection suggests that the C-terminus α-helix is fairly mobile which is also seen in the RMSF plot (figure 7.2) and occupies the space required for the allosteric binding site. In addition, the domains move further apart in comparison with the starting conformation. An RMSD plot (figure 7.12) highlights the move away from the starting structure, which occurs at ~14 ns into the simulation. Interestingly, a similar timescale was also observed for the opening of the two domains in the simulation of the original closed state X-ray structure (PDB ID: 1v4s), in the absence of both ligands (simulation D) (figure 5.6).

The starting conformation of the α-helix (residues 444-456) and the loop leading to the α-helix (residues 441 to 443), is very different to that of 1v4s (figure 7.9). Additionally, owing to the shorter resolved C-terminus in the GLK_AZ structure, the homology model is 5 residues shorter than that of the simulations of 1v4s. It is possible that in the absence of glucose at the active site, with longer simulation times, this helix would rotate out, towards the conformation of the C-terminus α-helix adopted in the super-open state. It is likely, that in this simulation, we have sampled intermediate conformations towards the full-opening of the two domains. Although this is not directly relevant the aims of this thesis, it would be interesting to take a closer look at the trajectory frames, to increase our understanding of the allosteric transition in GLK. Targeted MD simulations of the closed and super-open states of GLK are the only computational study (*151*), in which the transition between the two states has been monitored, but then the target conformation is provided and the simulation trajectory is constrained by the end target. Here, we have a relatively long trajectory (100 ns), where the system has been allowed to freely evolve in the absence of glucose, which as expected has led to a slight opening of the two domains (figure 7.12). In comparison, in the mutant simulation, GLK_AZ_apo, this behaviour was not observed. This highlights that, the crystallography mutations stabilise the structure, even in the absence of glucose at the active site.

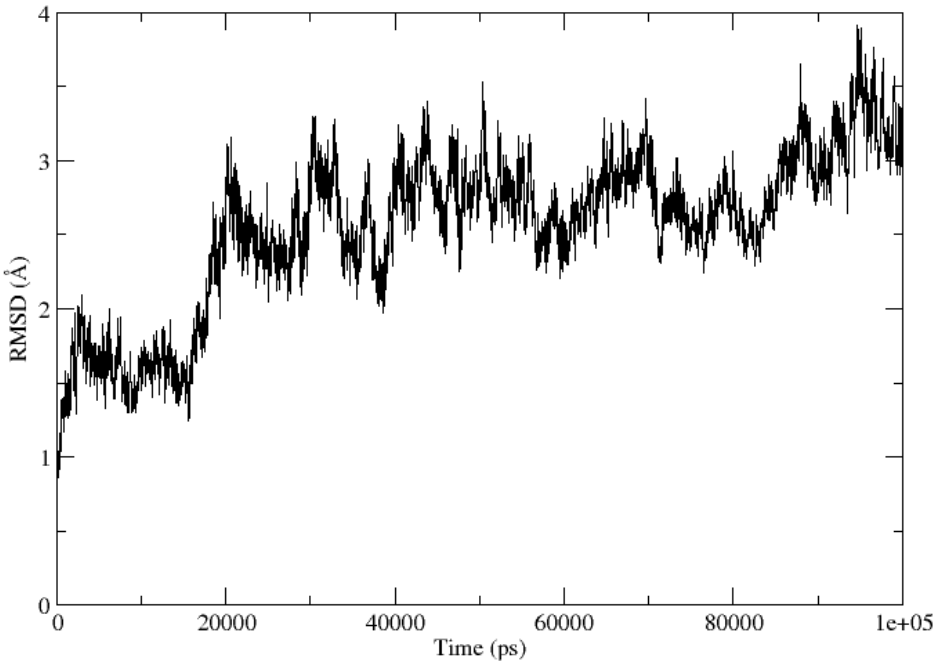


Figure 7.12: Cα RMSD plot of GLK_$AZ_{\text{wild-type homology}}$-apo, with respect to the first frame of the trajectory.

Figure 7.13 depicts an overlay of the grid points collected for all 25 frames included in the binding site search study for the apo simulation, GLK_ $AZ_{\text{wild-type homology}}$-apo. The active site is highly occupied, however the allosteric region is barely occupied, which further confirms that the 25 frames selected at every 4 ns interval do not have suitable conformations for the allosteric ligand pocket. However, this may improve if more frames are selected for analysis.

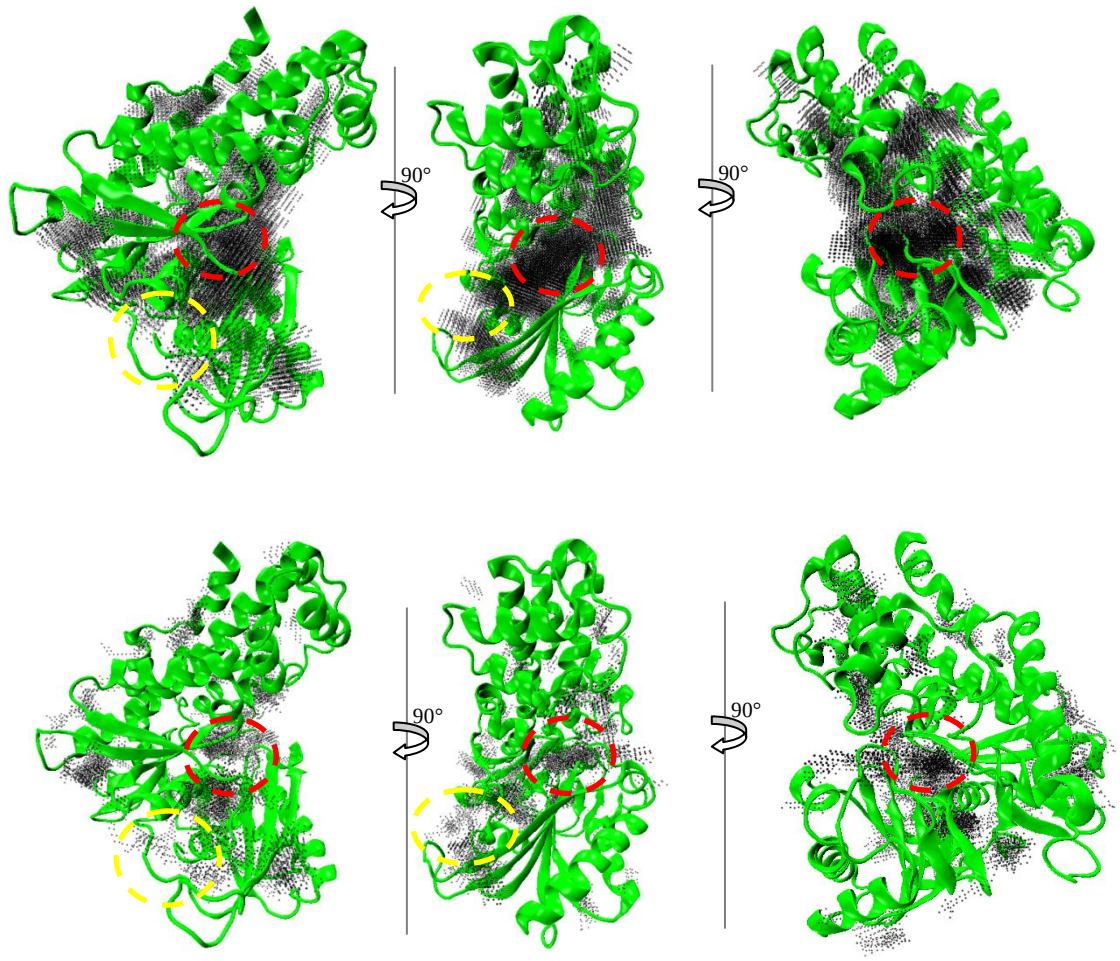


Figure 7.13: All grid points for all predicted binding sites by Pocket-Finder on top and by Q-SiteFinder below, overlaid for all 25 frames at 4 ns interval in the simulation GLK_AZ$_{\text{wild-type homology}}$-apo. For each frame, 10 sites were predicted, including all 25 frames leads to 250 sites in each case. For clarity, for each method, three orientations of the sites on the protein surface have been shown, rotated by 90°. The active and allosteric sites have been circled. The allosteric site has not been circled in the rotated orientation on the right, as the site is not visible in this orientation.

The simulations of the holo state, both with the mutated sequence, GLK_AZ_holo and the wild-type, GLK_ AZ$_{\text{wild-type homology}}$-holo, are relatively similar with respect to the RMSF profile but the apo simulations, GLK_AZ_apo$_{\text{pseudo}}$ and GLK_AZ$_{\text{wild-type homology}}$-apo are relatively different, in particular in the residues of the small domain (figure 7.3),. The allosteric binding site prediction in the GLK_AZ$_{\text{wild-type homology}}$-apo simulation (figures 7.11 & 7.13), is weaker than that of the GLK_AZ_apo$_{\text{pseudo}}$ simulation (figures

6.11 & 6.12), which further highlights that the simulation of the mutated structure is biased to remain in the closed state even in the absence of glucose, where the mutations may be also contributing to the flexibility of the allosteric region.

In the following section, the normal mode analysis of the homology structure will be discussed and compared to results obtained for the mutated structure.

7.5 Normal mode analysis

An identical protocol described to that in section 5.3 was applied to the GLK_AZ$_{\text{wild-type homology}}$ structure for comparison with the mutant GLK_AZ results. Two atomistic starting structures were prepared, summarised in table 7.1.

Normal mode analysis starting structures	
NMA A	GLK_AZ$_{\text{wild-type homology}}$ in complex with glucose
NMA B	GLK_AZ$_{\text{wild-type homology}}$ glucose removed (pseudo apo)

Table 7.1: Normal mode analysis starting structures for the GLK_AZ$_{\text{wild-type homology}}$ structure.

Interestingly, the overall normal modes RMSF profiles of the homology structure (figures 7.14), is very similar to the original 1v4s (figures 5.15 & 5.17), but with higher fluctuation in the β-turn residue (res. 91-100), where ~20 modes are sufficient to capture most of the backbone conformational mobility. The profile is however, different to that of GLK_AZ, in particular the glucose-bound structure (figures 6.19 & 6.20). The observation here confirms that the RMSF profile for the GLK_AZ, in complex with glucose (NMA A) in the previous chapter is an unusual behaviour.

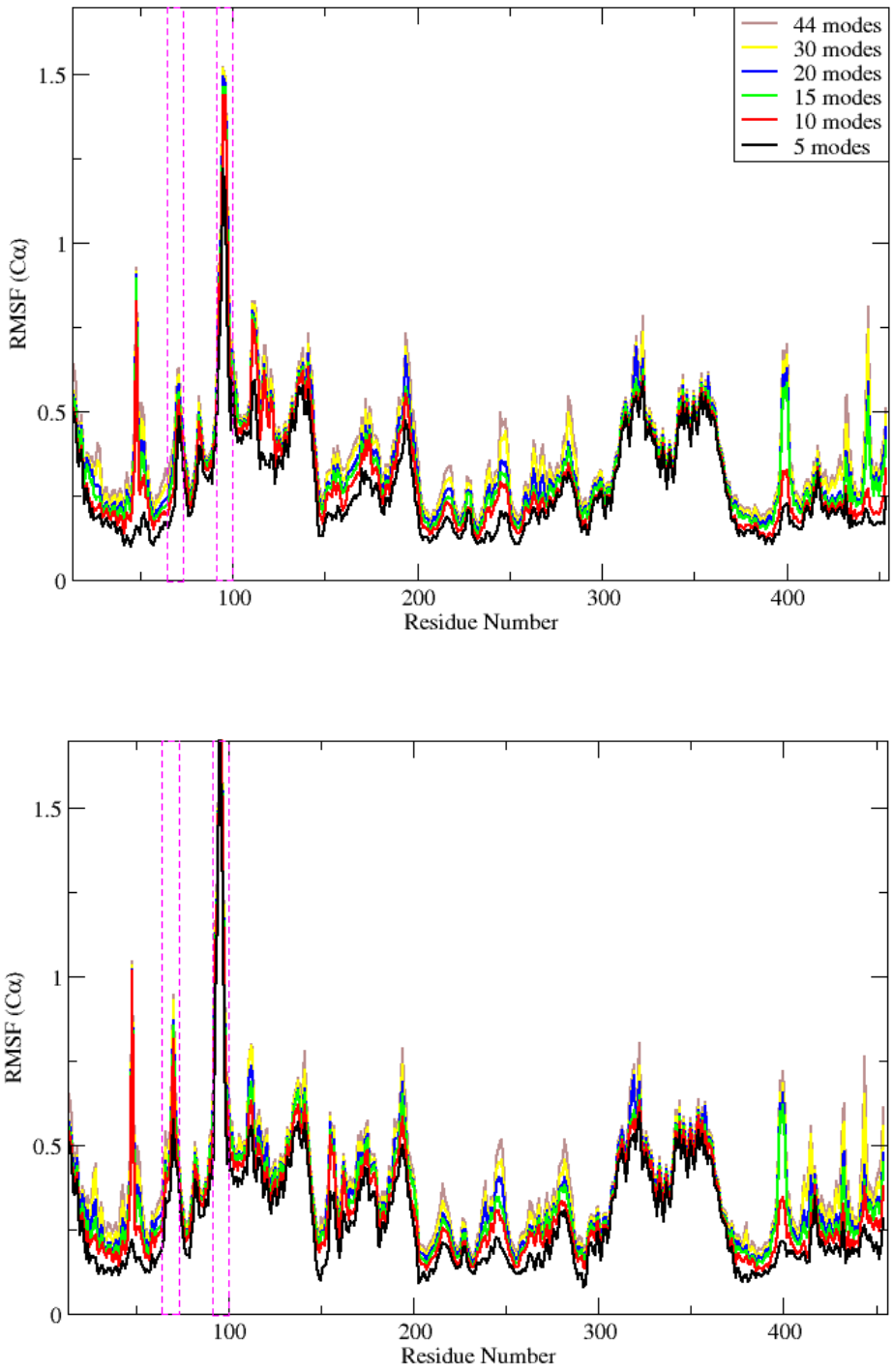


Figure 7.14: On top, an overlay of RMSFs for normal modes for NMA (A), GLK_AZ$_{\text{wild-type-homology}}$ in the presence of glucose. The magenta boxed areas refer to the connecting loop (residues 65-71) and the flexible β-sheet (91-100) close to the allosteric binding site. On the bottom, the corresponding for the NMA (B), GLK_AZ$_{\text{wild-type homology}}$ in the absence of glucose.

The minimised conformations of the GLK_AZ and GLK_AZ$_{\text{wild-type homology}}$ structures prior to NMA are visualised in figure 7.15. The major conformational difference lies in the residues of the flexible β-turn (residues 91-100), close to the allosteric site.

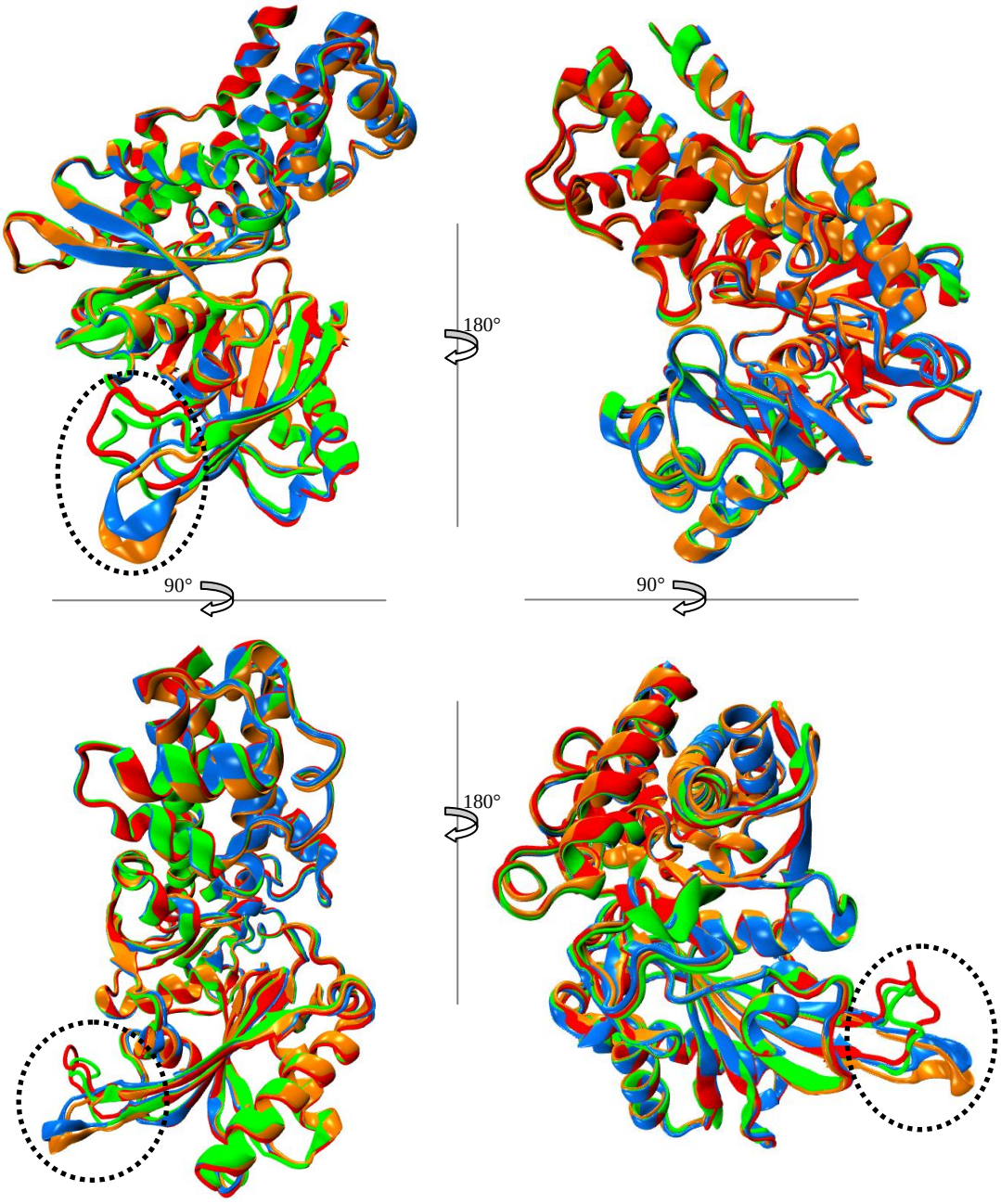


Figure 7.15: An overlay of the minimised structures of mutant GLK_AZ, in complex with glucose (red), in the absence of glucose (green) and the homology structure GLK_ AZ$_{\text{wild-type-homology}}$, in complex with glucose (orange), in the absence of glucose (blue), rotated in different orientations. Circled in black are regions with highest difference which are residues in the flexible β-turn (res. 91-100).

It is plausible that the stabilising influence of the surface mutations in the GLK_AZ structure seen in the previous chapter (figure 6.20), cause unusual vibrations to compensate for the restriction that is imposed on the system to remain in a closed conformation, which lead to high residue fluctuations at the tip of the two domains in the GLK_AZ, glucose-bound structure in chapter 6 (figure 6.21). In the absence of any other obvious structural reason contributing to this unusual RMSF profile in the GLK_AZ structure, a re-run of the normal modes calculation for this starting structure, in the future may clarify this issue.

Visual inspection demonstrates that in GLK_ $AZ_{\text{wild-type homology}}$ in complex with glucose (NMA A), while he connecting loop demonstrates mobility in the RMSF plot (figure 7.14), the direction in which it does so is not that required for the allosteric site opening (figure 7.16). The β-turn (residues 91-100) close to allosteric site, demonstrates a large degree of mobility, in addition to the opening/closing motion of the two domains. The significant mobility of the β-turn residues may stem from the unfavourable starting conformation of this region, which is based on the GLK_AZ mutant conformation.

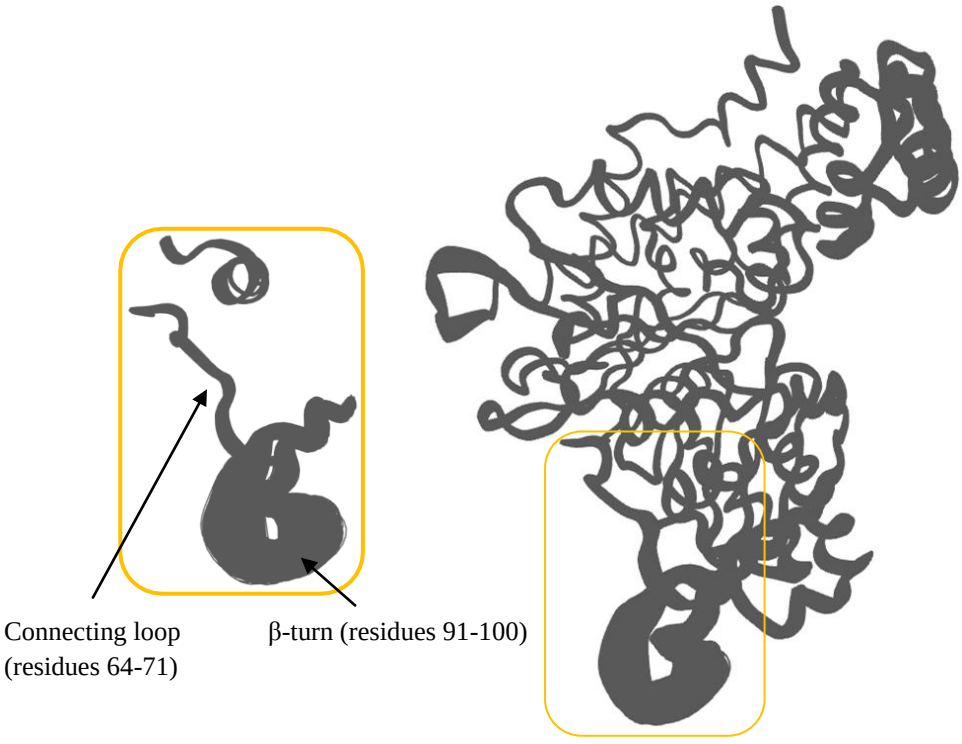


Figure 7.16: Top 20 non-zero modes for NMA (A), GLK_ in the presence of glucose. On the left, the allosteric region, zoomed in.

A binding site search on frames from NMA(A) will not lead to the identification of the allosteric binding site, as the connecting loop occupying the allosteric region is not displaced in the course of the vibrational modes calculated here.

This observation is consistent with results obtained for the mutant structure, GLK_AZ (section 6.4). The starting conformation of allosteric region is not sufficiently displaced with the normal modes to reveal the allosteric binding site.

7.6 Summary

The aim of this chapter was to monitor to influence of the crystallography mutations in the GLK_AZ structure on results observed in chapter 6. The RMSF plots (figure 7.3) for the 100 ns MD simulation for both systems, GLK_AZ and GLK_ $AZ_{wild\text{-}type\text{-}homology}$ demonstrates that in the holo form, the mobility of residues are very similar. However, in the apo form, the GLK_AZ mutation simulation, demonstrates an unusually stable and restricted mobility pattern in the absence of glucose from the active site during the MD simulation. It is plausible that the crystallographic mutations make considerable contribution to stabilising the closed-state conformation, even in the absence of glucose, hence aiding the crystallisation of the system which has proved difficult in the absence of allosteric activator. The reduced mobility of the allosteric region in the apo GLK_AZ MD simulation, highlights that glucose binding at the active site increases the mobility of the allosteric region, leading to a higher number of frames in which the allosteric binding site is first identified but also better ranked (figures 6.6 & 6.11).

In the homology model MD simulation, $GLK\_AZ_{wild\text{-}type\ homology}$, the allosteric binding site is identified in the holo simulation (figure 7.6), but the ranking is not as stable as that in the mutation simulation, GLK_AZ_holo (figure 6.6). The highly stable nature of the mutated structure may influence this outcome. The analysis here has been carried out at 4 ns intervals; the inclusion of more frames along the trajectory may improve the prediction, although it is more time-consuming. However, it can be expected that in the absence of an allosteric activator, the stable mutated simulation would sample more favourable conformations for the binding of the allosteric ligands, whereas the wild-type homology model, has the freedom to sample a wider range of conformations during the

100 ns simulation, and hence the population of suitable allosteric binding site conformations may be slightly reduced.

In both cases, the mutated and wild-type homology model, normal mode analysis does not reveal the allosteric binding site. This is disappointing, owing to the considerable shorter computational time of the method in comparison with MD simulations. However, it may not come as a surprise, considering the limitations of the method in calculating vibrations of the system around a local minimum and the inability for potential-well hopping. The starting conformation of the GLK_AZ structure demonstrated significant conformational differences to that of the original closed state X-ray structure (PDB: 1v4s), bound to glucose and an allosteric activator (figure 4.3). With MD, the simulation trajectory has the opportunity to visit several conformational states on the energy potential surface, which in the case of GLK_AZ, has led to the opening of the allosteric binding site, with the displacement of the loop connecting the two domains; however, in normal mode analysis, although the two domains, move apart, there is not sufficient mobility in the allosteric region to reveal the allosteric site.

Chapter 8

Conclusions

Allosteric regulation, and in particular the identification of allosteric binding sites, is of great interest due to the advantages associated with allosteric drugs. Most allosteric drugs to date have been identified serendipitously through high-throughput screening. Although experimental methods can be successful in identifying such binding sites, owing to the cost and time associated with these experiments, it would be useful to be able to aid the prediction of the location of such alternative binding sites with computational methods.

The starting point of a structure-based computational protocol is usually an X-ray or NMR structure. In this static structure, the challenge lies in identifying the allosteric site location in the absence of a modulator, as often the presence of the allosteric modulator stabilises a rare conformation in the conformational ensemble. A suitable computational method is then required to capture the dynamics of the system. Based on the pre-existing conformational equilibrium believed to be associated with allosteric transitions, molecular dynamics (MD) simulations can be used as a plausible route to model the dynamic nature of the protein, although if suitable conformations for allosteric modulator binding are rare, then long simulation times may be required. A large conformational change between the two states of a system, may be suitably modelled by normal mode analysis (NMA). We assessed the plausibility of revealing the allosteric binding site along the large transitions of the protein.

Human glucokinase (GLK) was used as a test-case, owing to structural availability and the knowledge of the location of the allosteric binding site with known modulators. In

recent years this target has gained a considerable therapeutic attention after the discovery of the first allosteric activators, and it is now an important target for the treatment of diabetes.

GLK is a monomeric enzyme, comprising two domains, the large and the small. Structurally there are two conformational states of this enzyme for which there are publicly available X-ray structures. One in the closed active state, bound to glucose at the active site and the allosteric activator at the allosteric site, which is about 20 Å away from the active site in this state. This is the less stable conformation. The allosteric activator shifts the conformational equilibrium towards the closed active state, hence enhancing glucose binding. In the unbound state, the two domains move apart and the small domain undergoes drastic conformational change, where both binding sites collapse.

As the apo form of this enzyme, the super-open, is so drastically different in conformation, and the transition of the super-open to an open or closed state may be a very slow process, we did not expect to be able predict the allosteric binding site from this state. However, AstraZeneca crystallographers provided an X-ray structure, only bound to glucose. This allowed us to have a real case scenario, in which the allosteric activator, if present, would have been able to bind to this structure, but in the absence of an effector the cavity necessary for allosteric effector binding is absent, owing to local conformational rearrangements in the allosteric site region.

Four structures of this enzyme were studied in this thesis. The closed-state X-ray structure bound to glucose and an allosteric activator (PDB ID: 1v4s), the super-open apo state (PDB ID:1v4t), the glucose only bound closed-state X-ray structure, containing several crystallographic mutations, and finally a homology structure, with the wild-type sequence based on the conformation of the AstraZeneca structure.

In chapter 5, the study of the allosteric-bound structure (PDB ID: 1v4s) was treated as a benchmark to ensure a suitable simulation protocol for this system and to gain insight into the dynamic nature of the enzyme in the presence and absence of either/both ligands. The study of this structure also allowed us to ensure that we could predict the binding site in a structure which we would expect to do so, considering the presence of the allosteric activator in this structure. Even when the allosteric ligand was removed from the binding site during the 25 ns MD simulation, the binding site did not collapse.

In the simulation where both ligands were removed, a decrease in the size and rank of the allosteric binding site was observed, demonstrating the key role of glucose at the active site.

In the super-open state, despite the drastic conformational change, a small pocket was identified in the allosteric region involving a few of the allosteric site interacting residues. The binding site profile observed is consistent with kinetic studies, indicating that the allosteric activator could bind the enzyme in the absence of glucose, although with very low affinity (*218*).

The MD simulation of the AstraZeneca structure in chapter 6, GLK_AZ led to the prediction of the allosteric binding site, although not always highly ranked due to the highly flexible nature of the site in the absence of an activator. A high number of frames from the trajectory demonstrated suitable conformations suitable for binding a ligand at the allosteric site. A population analysis of the locations of predicted sites in all selected frames demonstrated a highly populated region at the allosteric site. The knowledge of the activating mutations in the allosteric region may be utilised in this particular case to guide the prediction with confidence. The apo simulation of the GK_AZ structure, GLK_AZ_apo$_{pseudo}$, demonstrated a small decrease in the prediction of the allosteric binding site, but this was not as much as expected considering the long 100 ns simulation in the absence of glucose. An apo simulation of the homology structure in chapter 7, GLK_AZ$_{wild\text{-}type\ homology}$-apo based on the wild-type sequence and the 3D structure of GLK_AZ, demonstrated that the prediction of the allosteric binding site in the GLK_AZ_apo$_{pseudo}$ simulation may have been influenced by the surface mutations in the system. The closed state conformation is fairly stable in the absence of glucose in the GK_AZ_apo$_{pseudo}$ simulation, and hence the allosteric binding site remains fairly strong. In contrast, the apo simulation of the homology structure demonstrated a weak prediction of the allosteric binding site in the absence of glucose from the simulation. The glucose-bound simulations of both GLK_AZ and the homology structure were similar in terms of allosteric binding site prediction, which suggests that the mutations in the GLK_AZ structure did not have a biasing influence on the outcome of the results.

Normal mode analysis in both the GLK_AZ and homology model structures did not reveal the allosteric binding site. Despite capturing the opening and closing of the two

domains, the loop connecting the two domains did not sufficiently move in the correct direction to create a space required by an allosteric modulator.

In the absence of an allosteric-bound X-ray structure, snapshots from MD simulations in this study could have also provided suitable conformations to be used in ensemble docking of ligands or fragments.

The simulations here have also highlighted the flexibility of the allosteric region, and the adjacent regions which may not currently be explored by the publicly available activator-bound GLK X-ray structures. This information can be helpful when extending the ligand size or improving particular structural properties of current activators at the binding site.

Future work in the quest to predict allosteric binding sites will require the inclusion of more allosteric protein systems with varying degrees of conformational change. The allosteric classification by Tsai et al. (*1*) is a useful guide in potentially grouping proteins in terms of the degree of conformational change that they may undergo to reveal an allosteric binding site. Additional knowledge, such as the location of naturally occurring mutations in a protein may also guide computational efforts.

Overall, however, the work presented in this thesis has demonstrated that, for GLK, computer simulations can be used to identify the presence of an allosteric binding site, from a structure without an allosteric modulator bound.

Appendix A

Sequence alignment between 1v4s and GLK_AZ

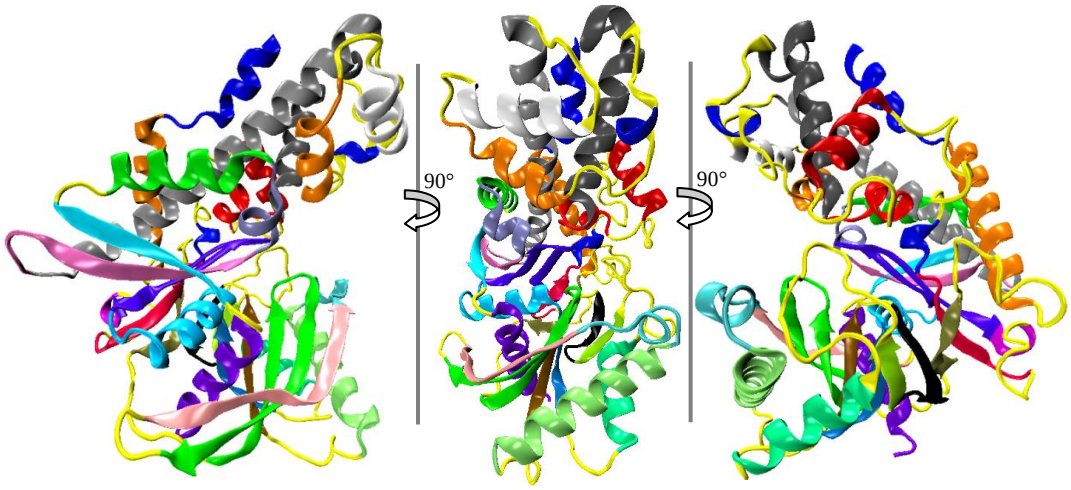
Sequence alignment between the GLK_AZ mutant structure (isoform 3) and publicly available closed-state X-ray structure (PDB ID: 1v4s), which is an isoform 2, both which are expressed in the liver. The sequence alignment was carried out with the online tool ClustalW2 with default parameters provided on the server (*266*, *267*). There is a 97% sequence identity between the two sequences.

```
CLUSTAL 2.0.10 multiple sequence alignment

mutant              SSNSQVEQILAEFQLQAADLKKVMRRMQKEMDRGLRLETHAAASVKMLPT 50
1V4S_A|PDBID|CHAIN  MALTLVEQILAEFQLQEEDLKKVMRRMQKEMDRGLRLETHEEASVKMLPT 50
                     : : ***********  **********************  ********

mutant              YVRSTPEGSEVGDFLSLDLGGTNFRVMLVKVGAGAAGQWSVKTKHQMYSI 100
1V4S_A|PDBID|CHAIN  YVRSTPEGSEVGDFLSLDLGGTNFRVMLVKVGEGEEGQWSVKTKHQMYSI 100
                    ********************************* *  **************

mutant              PEDAMTGTAEMLFDYISECISDFLDKHQMKHKKLPLGFTFSFPVRHEDID 150
1V4S_A|PDBID|CHAIN  PEDAMTGTAEMLFDYISECISDFLDKHQMKHKKLPLGFTFSFPVRHEDID 150
                    **************************************************

mutant              KGILLNWTKGFKASGAEGNNVVGLLRDAIKRRGDFEMDVVAMVNDTVATM 200
1V4S_A|PDBID|CHAIN  KGILLNWTKGFKASGAEGNNVVGLLRDAIKRRGDFEMDVVAMVNDTVATM 200
                    **************************************************

mutant              ISCYYEDHQCEVGMIVGTGCNACYMEEMQNVELVEGDEGRMCVNTEWGAF 250
1V4S_A|PDBID|CHAIN  ISCYYEDHQCEVGMIVGTGCNACYMEEMQNVELVEGDEGRMCVNTEWGAF 250
                    **************************************************

mutant              GDSGELDEFLLEYDRLVDESSANPGQQLYEKLIGGKYMGELVRLVLLRLV 300
1V4S_A|PDBID|CHAIN  GDSGELDEFLLEYDRLVDESSANPGQQLYEKLIGGKYMGELVRLVLLRLV 300
                    **************************************************

mutant              DENLLFHGEASEQLRTRGAFETRFVSQVESDTGDRKQIYNILSTLGLRPS 350
1V4S_A|PDBID|CHAIN  DENLLFHGEASEQLRTRGAFETRFVSQVESDTGDRKQIYNILSTLGLRPS 350
                    **************************************************

mutant              TTDCDIVRRACESVSTRAAHMCSAGLAGVINRMRESRSEDVMRITVGVDG 400
1V4S_A|PDBID|CHAIN  TTDCDIVRRACESVSTRAAHMCSAGLAGVINRMRESRSEDVMRITVGVDG 400
                    **************************************************

mutant              SVYKLHPSFKERFHASVRRLTPSCEITFIESEEGSGRGAALVSAVA---- 446
1V4S_A|PDBID|CHAIN  SVYKLHPSFKERFHASVRRLTPSCEITFIESEEGSGRGAALVSAVACKKA 450
                    *******************************************

mutant              -----
1V4S_A|PDBID|CHAIN  CMLGQ 455
```

Appendix B

GLK structure colour-code

Below, the GLK active closed-state structure has been colour-coded for the reader.



14-26.....27-46.....47-57....58-64.....65-71.....72-94 .....

.....95-112.....113-118.....119-138..139-144...145-151...152-160.....161-165..

166-180..181-192.....193-204.....205-219.....220-237.....238-247...

248-256.....257-271...271-281..282-287.288-296...297-319..320-33...339-345...

346-360.....361-398.....399-410....411-417..418-431.....432-443.....444-460

Appendix C

NAB normal mode analysis routine

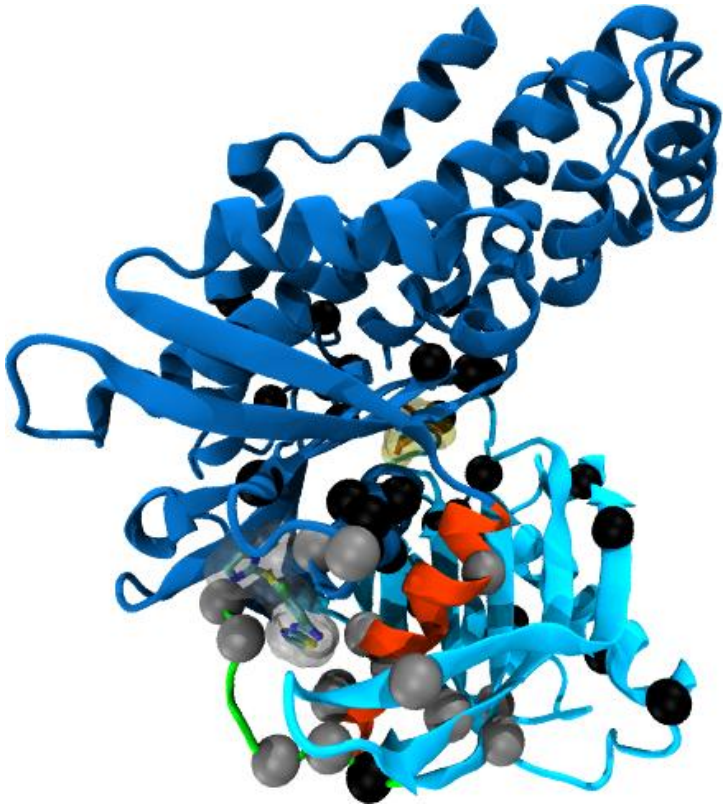
```
// carry out molecular mechanics minimization and some simple
dynamics

molecule m;
int  ier;
float  m_xyz[ dynamic ], f_xyz[ dynamic ], v[ dynamic ];
float  dgrad, fret, elapsed, t1, t2;
string i, time;

//-------------------------------------------------------------

time=timeofday();
printf("start time %s\n\n", time);
t1=second();

//-------------------------------------------------------------
m = getpdb("final-1v4s-glu-mrk.pdb");
readparm(m, "1v4s-glu-mrk.top");

allocate m_xyz[ 3*m.natoms ];  allocate f_xyz[ 3*m.natoms ];
allocate v[ 3*m.natoms ];

setxyz_from_mol( m, NULL, m_xyz );

//-------------------------------------------------------------

mm_options( "cut=99.0, rgbmax=99.0, ntpr=500, nsnb=25, gb=1, diel=C"
);
mme_init( m, NULL, ":::ZZZ", m_xyz, NULL );

fret = mme( m_xyz, f_xyz, 1 );
printf( "\nInitial energy is %8.3f\n", fret );

//-------------------------------------------------------------

//conjugate gradient minimisation

dgrad = 0.00001;
ier = conjgrad( m_xyz, 3*m.natoms, fret, mme, dgrad, 10.0, 1000000
);
printf("\nconj-grad returns: %d\n", ier );

printf("\nGonj-grad is done\n", i);

t2=second();
elapsed=t2-t1;
printf("\nelapsed after conj-grad:  %f\n\n", elapsed);
time=timeofday();
printf("\nTime after conj-grad: %s\n\n", time);

//-------------------------------------------------------------
```

```
//Newton-Rhapson minimisation

mm_options( "cut=99.0, rgbmax=99, ntpr=1, nsnb=25, gb=1, diel=C" );
dgrad =1.e-11;

ier= newton(m_xyz, 3*m.natoms, fret, mme,mme2, dgrad, 0.0, 200);

printf("\nNewton-Rhapson is done\n", i);

t2=second();
elapsed=t2-t1;
printf("\nelapsed after Newton-rhap:  %f\n\n", elapsed);

time=timeofday();
printf("\nTime after Newt-RhaphL %s\n\n", time);

setmol_from_xyz(m, NULL, m_xyz);
putpdb("minimised.pdb", m);


//-----------------------------------------------------------------

// get the normal modes:

ier = nmode( m_xyz, 3*m.natoms,mme2, 50);
printf("nmode returns %d\n", ier );

t2=second();
elapsed=t2-t1;
printf("\nelapsed after normal modes: %f\n\n", elapsed);


time=timeofday();
printf("end time %s\n\n", time);
```

Appendix D

Location of the inactivating and activating mutations in MODY and PHHI, respectively.

The close state structure of GLK (PDB ID: 1v4s), bound to glucose in yellow transparent surface and the allosteric activator in grey transparent surface. In black the spheres highlight the locations of the inactivating mutations associated with MODY which are quite spread around the protein but a large number are near the active site. The activating mutations, grey spheres, on the other hand heavily populate the allosteric region.



Bibliography

1. Tsai, C. J., del Sol, A., and Nussinov, R. (2008) Allostery: Absence of a change in shape does not imply that allostery is not at play, *J. Mol. Biol. 378*, 1-11.
2. Hardy, J. A., and Wells, J. A. (2004) Searching for new allosteric sites in enzymes, *Curr. Opin. Struct. Biol. 14*, 706-715.
3. Vionnet, N., Stoffel, M., Takeda, J., Yasuda, K., Bell, G. I., Zouali, H., Lesage, S., Velho, G., Iris, F., Passa, P., Froguel, P., and Cohen, D. (1992) Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes-mellitus, *Nature 356*, 721-722.
4. Cuesta-Munoz, A. L., Huopio, H., Otonkoski, T., Gomez-Zumaquero, J. M., Nanto-Salonen, K., Rahier, J., Lopez-Enriquez, S., Garcia-Gimeno, M. A., Sanz, P., Soriguer, F. C., and Laakso, M. (2004) Severe persistent hyperinsulinemic hypoglycemia due to a de novo glucokinase mutation, *Diabetes 53*, 2164-2168.
5. Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J. I., and Nagata, Y. (2004) Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase, *Structure 12*, 429-438.
6. Conn, P. J., Christopoulos, A., and Lindsley, C. W. (2009) Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders, *Nat. Rev. Drug Discov. 8*, 41-54.
7. Demerdash, O. N. A., Daily, M. D., and Mitchell, J. C. (2009) Structure-based predictive models for allosteric hot spots, *PLoS Comput. Biol. 5*, e1000531.
8. Lemieux, R. U., and Spohr, U. (1994) How fischer,emil was led to the lock and key concept for enzyme specificity, *Adv. Carbohydr. Chem. Biochem. 50*, 1-20.
9. Stryer, L., Berg, J. M., and Tymoczko, J. L. (2006) *Biochemistry*, 6 ed., W. H. Freeman and Company.
10. Bosshard, H. R. (2001) Molecular recognition by induced fit: how fit is the concept?, *News Physiol. Sci. 16*, 171-173.
11. Koshland, D. E. (1958) Application of a theory of enzyme specificity to protein synthesis, *Proc. Natl. Acad. Sci. U.S.A. 44*, 98-104.
12. Gunasekaran, K., Ma, B. Y., and Nussinov, R. (2004) Is allostery an intrinsic property of all dynamic proteins?, *Proteins 57*, 433-443.
13. Fetler, L., Kantrowitz, E. R., and Vachette, P. (2007) Direct observation in solution of a preexisting structural equilibrium for a mutant of the allosteric aspartate transcarbamoylase, *Proc. Natl. Acad. Sci. U.S.A. 104*, 495-500.
14. Monod, J., Wyman, J., and Changeux, J. (1965) On the nature of allosteric transitions: A plausible model, *Journal of molecular biology 12*, 88-118.
15. Koshland, D. E., Némethy, G., and Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits, *Biochemistry 5*, 365-385.
16. Laskowski, R. A., Gerick, F., and Thornton, J. M. (2009) The structural basis of allosteric regulation in proteins, *FEBS Lett. 583*, 1692-1698.
17. Jaffe, E. K. (2005) Morpheeins: A new structural paradigm for allosteric regulation, *Trends in Biochemical Sciences 30*, 490-497.

18. Goodey, N. M., and Benkovic, S. J. (2008) Allosteric regulation and catalysis emerge via a common route, *Nat. Chem. Biol. 4*, 474-482.
19. Wand, A. J. (2001) Dynamic activation of protein function: A view emerging from NMR spectroscopy, *Nat. Struct. Biol. 8*, 926-931.
20. Hawkins, R. J., and McLeish, T. C. B. (2004) Coarse-grained model of entropic allostery, *Phys. Rev. Lett. 93*, 098104.
21. Homans, S. W. (2005) Probing the binding entropy of ligand-protein interactions by NMR, *Chembiochem 6*, 1585-1591.
22. Popovych, N., Sun, S. J., Ebright, R. H., and Kalodimos, C. G. (2006) Dynamically driven protein allostery, *Nat. Struct. Mol. Biol. 13*, 831-838.
23. Bowery, N. G. (2006) *Allosteric receptor modulation in drug targeting By N. G. Bowery* 1ed., Informa HealthCare.
24. May, L. T., Leach, K., Sexton, P. M., and Christopoulos, A. (2007) Allosteric modulation of G protein-coupled receptors, *Annu. Rev. Pharmacol. Toxicol. 47*, 1-51.
25. Kenakin, T. (2006) *A pharmacology primer: Theory, applications, and methods*, 1 ed., Elsevier.
26. Jensen, A. A., and Spalding, T. A. (2004) Allosteric modulation of G-protein coupled receptors, *Eur. J. Pharm. Sci. 21*, 407-420.
27. Johnson, V. A., Brun-Vézinet, F., Clotet, B., Günthard, H. F., Kuritzkes, D. R., Pillay , D., Schapiro, J. M., and Richman, D. D. (2008) Update of the drug resistance mutations in HIV-1: December 2008, *International AIDS Society-USA, Topics in HIV Medicine 16*, 138-145.
28. Oikonomakos, N. G., Skamnaki, V. T., Tsitsanou, K. E., Gavalas, N. G., and Johnson, L. N. (2000) A new allosteric site in glycogen phosphorylase b as a target for drug interactions, *Structure 8*, 575-584.
29. Yan, Y. W., Sardana, V., Xu, B., Homnick, C., Halczenko, W., Buser, C. A., Schaber, M., Hartman, G. D., Huber, H. E., and Kuo, L. C. (2004) Inhibition of a mitotic motor protein: Where, how, and conformational consequences, *J. Mol. Biol. 335*, 547-554.
30. Wright, S. W., Carlo, A. A., Carty, M. D., Danley, D. E., Hageman, D. L., Karam, G. A., Levy, C. B., Mansour, M. N., Mathiowetz, A. M., McClure, L. D., Nestor, N. B., McPherson, R. K., Pandit, J., Pustilnik, L. R., Schulte, G. K., Soeller, W. C., Treadway, J. L., Wang, I. K., and Bauer, P. H. (2002) Anilinoquinazoline inhibitors of fructose 1,6-bisphosphatase bind at a novel allosteric site: Synthesis, in vitro characterization, and X-ray crystallography, *J. Med. Chem. 45*, 3865-3877.
31. Wrobel, J., Sredy, J., Moxham, C., Dietrich, A., Li, Z. N., Sawicki, D. R., Seestaller, L., Wu, L., Katz, A., Sullivan, D., Tio, C., and Zhang, Z. Y. (1999) PTP1B inhibition and antihyperglycemic activity in the ob/ob mouse model of novel 11-arylbenzo[b] naphtho[2,3-d]furans and 11-arylbenzo[b] naphtho [2,3-d] thiophenes, *J. Med. Chem. 42*, 3199-3202.
32. Wiesmann, C., Barr, K. J., Kung, J., Zhu, J., Erlanson, D. A., Shen, W., Fahr, B. J., Zhong, M., Taylor, L., Randal, M., McDowell, R. S., and Hansen, S. K. (2004) Allosteric inhibition of protein tyrosine phosphatase 1B, *Nat. Struct. Mol. Biol. 11*, 730-737.
33. Horn, J. R., and Shoichet, B. K. (2004) Allosteric inhibition through core disruption, *J. Mol. Biol. 336*, 1283-1291.

34. Pargellis, C., Tong, L., Churchill, L., Cirillo, P. F., Gilmore, T., Graham, A. G., Grob, P. M., Hickey, E. R., Moss, N., Pav, S., and Regan, J. (2002) Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site, *Nat. Struct. Biol. 9*, 268-272.
35. Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992) Crystal-structure at 3.5 angstrom resoltuion of HIV-1 reverse-transcriptase complexed with an inhibitor, *Science 256*, 1783-1790.
36. Dennis, M. S., Roberge, M., Quan, C., and Lazarus, R. A. (2001) Selection and characterization of a new class of peptide exosite inhibitors of coagulation factor VIIa, *Biochemistry 40*, 9513-9521.
37. Hardy, J. A., Lam, J., Nguyen, J. T., O'Brien, T., and Wells, J. A. (2004) Discovery of an allosteric site in the caspases, *Proc. Natl. Acad. Sci. U.S.A. 101*, 12461-12466.
38. Macarrón, R., and Hertzberg, R. P. (2002) Design and implementation of high throughput screening assays, in *High Throughput Screening*, pp 1-29, Humana Press.
39. Smith, G. P. (1985) Filamentous fusion phage: novel expression vectors that display closed antigens on the vision surface, *Science 228*, 1315-1317.
40. Website. [Last accessed: September 2009], © Dyax Corp 2004. http://www.dyax.com/discovery/phagedisplay.html.
41. Dennis, M. S., Eigenbrot, C., Skelton, N. J., Ultsch, M. H., Santell, L., Dwyer, M. A., O'Connell, M. P., and Lazarus, R. A. (2000) Peptide exosite inhibitors of factor VIIa as anticoagulants, *Nature 404*, 465-470.
42. Eigenbrot, C., Kirchhofer, D., Dennis, M. S., Santell, L., Lazarus, R. A., Stamos, J., and Ultsch, M. H. (2001) The factor VII zymogen structure reveals reregistration of beta strands during activation, *Structure 9*, 627-636.
43. Erlanson, D. A., Wells, J. A., and Braisted, A. C. (2004) Tethering: Fragment-based drug discovery, *Annu. Rev. Biophys. Biomol. Struct. 33*, 199-223.
44. Ishima, R., and Torchia, D. A. (2000) Protein dynamics from NMR, *Nat. Struct. Biol. 7*, 740-743.
45. Pervushin, K. (2000) Impact of transverse relaxation optmized spectroscopy (TROSY) on NMR as a technique in structural biology, *Quarterly Reviews of Biophysics 33*, 161-197.
46. Riek, R., Pervushin, K., and Wuthrich, K. (2000) TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution, *Trends Biochem. Sci. 25*, 462-468.
47. Zhu, G., and Yao, X. (2008) TROSY-based NMR experiments for NMR studies of large biomolecules, *Prog. Nucl. Magn Reson. Spectrosc. 52*, 49-68.
48. McElroy, C., Manfredo, A., Wendt, A., Gollnick, P., and Foster, M. (2002) TROSY-NMR studies of the 91 kDa TRAP protein reveal allosteric control of a gene regulatory protein by ligand-altered flexibility, *J. Mol. Biol. 323*, 463-473.
49. Hajduk, P. J., Huth, J. R., and Fesik, S. W. (2005) Druggability indices for protein targets derived from NMR-based screening data, *J. Med. Chem. 48*, 2518-2525.
50. Jahnke, W., Blommers, M. J. J., Fernandez, C., Zwingelstein, C., and Amstutz, R. (2005) Strategies for the NMR-based identification and

optimization of allosteric protein kinase inhibitors, *Chembiochem 6*, 1607-1610.
51. Jahnke, W., Rudisser, S., and Zurini, M. (2001) Spin label enhanced NMR screening, *J. Am. Chem. Soc. 123*, 3149-3150.
52. Hilser, V. J., and Freire, E. (1996) Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors, *Journal of molecular biology 262*, 756-772.
53. Vertrees, J., Barritt, P., Whitten, S., and Hilser, V. J. (2005) COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures, *Bioinformatics 21*, 3318-3319.
54. Freire, E. (2000) Can allosteric regulation be predicted from structure?, *Proc. Natl. Acad. Sci. U.S.A. 97*, 11680-11682.
55. Pan, H., Lee, J. C., and Hilser, V. J. (2000) Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble, *Proc. Natl. Acad. Sci. U.S.A. 97*, 12020-12025.
56. Shulman, A. I., Larson, C., Mangelsdorf, D. J., and Ranganathan, R. (2004) Structural determinants of allosteric ligand activation in RXR heterodimers, *Cell 116*, 417-429.
57. Suel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins, *Nat. Struct. Biol. 10*, 59-69.
58. McClendon, C. L., Friedland, G., Mobley, D. L., Amirkhani, H., and Jacobson, M. P. (2009) Quantifying correlations between allosteric sites in thermodynamic ensembles, *J. Chem. Theory Comput. 5*, 2486-2502.
59. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families, *J. Mol. Biol. 257*, 342-358.
60. Zhang, D., and McCammon, J. A. (2005) The association of tetrameric acetylcholinesterase with ColQ tail: a block normal mode analysis, *PLoS Comput. Biol. 1*, e62.
61. Chennubhotla, C., Yang, Z., and Bahar, I. (2008) Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL, *Mol. Biosyst. 4*, 287-292.
62. Hinsen, K., Thomas, A., and Field, M. J. (1999) Analysis of domain motions in large proteins, *Proteins 34*, 369-382.
63. Marques, O., and Sanejouand, Y. H. (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations, *Proteins 23*, 557-560.
64. Tama, F., and Sanejouand, Y. H. (2001) Conformational change of proteins arising from normal mode calculations, *Protein Eng. 14*, 1-6.
65. Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, N., Yu, H. Y., and Gerstein, M. (2002) Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic, *Proteins 48*, 682-695.
66. Li, L. W., Uversky, V. N., Dunker, A. K., and Meroueh, S. O. (2007) A computational investigation of allostery in the catabolite activator protein, *J. Am. Chem. Soc. 129*, 15668-15676.
67. Bradley, M. J., Chivers, P. T., and Baker, N. A. (2008) Molecular dynamics simulation of the Escherichia coli NikR protein: Equilibrium conformational

fluctuations reveal interdomain allosteric communication pathways, *Journal of molecular biology 378*, 1155-1173.
68. Sayar, K., Ugur, O., Liu, T., Hilser, V. J., and Onaran, O. (2008) Exploring allosteric coupling in the alpha-subunit of Heterotrimeric G proteins using evolutionary and ensemble-based approaches, *BMC Structural Biology 8*, 1-14.
69. Lange, O. F., and Grubmuller, H. (2006) Generalized correlation for biomolecular dynamics, *Proteins-Structure Function and Bioinformatics 62*, 1053-1061.
70. Lange, O. F., Grubmuller, H., and de Groot, B. L. (2005) Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions, *Angewandte Chemie-International Edition 44*, 3394-3399.
71. Ming, D., Cohn, J. D., and Wall, M. E. (2008) Fast dynamics perturbation analysis for prediction of protein functional sites, *Bmc Structural Biology 8*, 1-11.
72. Ho, B. K., and Agard, D. A. (2009) Probing the flexibility of large conformational changes in protein structures through local perturbations, *PLoS Comput. Biol. 5*, e10000343.
73. Lenaerts, T., Ferkinghoff-Borg, J., Stricher, F., Serrano, L., Schymkowitz, J. W. H., and Rousseau, F. (2008) Quantifying information transfer by protein domains: Analysis of the Fyn SH2 domain structure, *Bmc Structural Biology 8*, 1-15.
74. Vajda, S., and Guarnieri, F. (2006) Characterization of protein-ligand interaction sites using experimental and computational methods, *Curr. Opin. Drug Discov. Dev. 9*, 354-362.
75. Teague, S. J. (2003) Implications of protein flexibility for drug discovery, *Nature Reviews Drug Discovery 2*, 527-541.
76. Joseph, D., Petsko, G. A., and Karplus, M. (1990) Anatomy of a conformational change - Hinged lid motion of the triosephosphate isomerase loop, *Science 249*, 1425-1428.
77. Carlson, H. A. (2002) Protein flexibility is an important component of structure-based drug discovery, *Curr. Pharm. Des. 8*, 1571-1578.
78. Yon, J. M., Perahia, D., and Ghelis, C. (1998) Conformational dynamics and enzyme activity, *Biochimie 80*, 33-42.
79. Gutteridge, A., and Thornton, J. (2004) Conformational change in substrate binding, catalysis and product release: an open and shut case?, *FEBS Lett. 567*, 67-73.
80. Kempner, E. S. (1993) Movable lobes and flexible loops in proteins structural deformations that control biochemical activity, *FEBS Lett. 326*, 4-10.
81. Hammes, G. G. (2002) Multiple conformational changes in enzyme catalysis *Biochemistry 41*, 8221-8228.
82. Ringe, D. (1995) What makes a binding site a binding site?, *Current Opinion in Structural Biology 5*, 825-829.
83. Dennis, S., Kortvelyesi, T., and Vajda, S. (2002) Computational mapping identifies the binding sites of organic solvents on proteins, *Proc. Natl. Acad. Sci. U.S.A. 99*, 4290-4295.

84. Kortvelyesi, T., Silberstein, M., Dennis, M. S., and Vajda, S. (2003) Improved mapping of protein binding sites, *Journal of Computer-Aided Molecular Design 17*, 173-186.
85. Ruppert, J., Welch, W., and Jain, A. (1997) Automatic identification and representation of protein binding sites for molecular docking, *Protein Science 6*, 524-533.
86. Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V., and Willett, P. (2001) SuperStar: Improved knowledge-based interaction fields for protein binding sites, *Journal of molecular biology 307*, 841-859.
87. An, J. H., Totrov, M., and Abagyan, R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes, *Molecular & Cellular Proteomics 4*, 752-761.
88. Laskowski, R. A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions, *J. Mol. Graph. 13*, 323-330.
89. Hendlich, M., Rippmann, F., and Barnickel, G. (1997) LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins, *J. Mol. Graph. Model. 15*, 359-+.
90. Brady, G. P., and Stouten, P. F. W. (2000) Fast prediction and visualization of protein binding pockets with PASS, *J. Comput. Aided Mol. Des. 14*, 383-401.
91. Binkowski, T. A., Naghibzadeh, S., and Liang, J. (2003) CASTp: Computed atlas of surface topography of proteins, *Nucleic Acids Research 31*, 3352-3355.
92. Weisel, M., Proschak, E., and Schneider, G. (2007) PocketPicker: Analysis of ligand binding-sites with shape descriptors, *Chemistry Central Journal 1*.
93. Sotriffer, C., and Klebe, G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design, *Farmaco 57*, 243-251.
94. Campbell, S. J., Gold, N. D., Jackson, R. M., and Westhead, D. R. (2003) Ligand binding: functional site location, similarity and docking, *Current Opinion in Structural Biology 13*, 389-395.
95. Laurie, A. T. R., and Jackson, R. M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites, *Bioinformatics 21*, 1908-1916.
96. McCarthy, D. J., Hogle, J. H., and Karplus, M. (1997) Use of the multiple copy simultaneous search (MCSS) method to design a new class of picornavirus capsid binding drugs, *Proteins Structure Function and Genetics 29*, 32-58.
97. Leach, A. R. (2001) *Molecular modelling, principles and applications*, 2 ed., Pearson Education Ltd.
98. York, D. M., Darden, T. A., Pedersen, L. G., and Anderson, M. W. (2002) Molecular dynamics simulation of HIV-1 protease in a crystalline environment and in solution, *Biochemistry 32*, 1443-1453.
99. MacKerell, A. D., Bashford, D., Bellott, M., 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. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins, *Journal of Physical Chemistry B 102*, 3586-3616.

100. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, *Journal of Computational Chemistry 24*, 1999-2012.
101. Hermans, J., Berendsen, H. J. C., Gunsteren, W. F. V., and Postma, J. P. M. (1984) A consistent empirical potential for water-protein interactions, *Biopolymers 23*, 1513-1518.
102. Ott, K. H., and Meyer, B. (1996) Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations, *Journal of Computational Chemistry 17*, 1968-1084.
103. MacKerell, A. D.
104. Jorgensen, W. L., and Tirado-Rives, J. (2002) The OPLS optimized potentials for liquid simulations: Potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, *Journal of the American Chemical Society 110*, 1657-1666.
105. Lopes, P., Roux, B., and MacKerell, A. (2009) Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications, *Theor. Chem. Acc. 124*, 11-28.
106. van Duin, A. C. T., Dasgupta, S., Lorant, F., and Goddard, W. A. (2001) ReaxFF: A reactive force field for hydrocarbons, *The Journal of Physical Chemistry A 105*, 9396-9409.
107. Go, N., Noguti, T., and Nishikawa, T. (1983) Dynamics of a small globular protien in terms of low-frequency vibrational-modes, *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 80*, 3696-3700.
108. Brooks, B., and Karplus, M. (1983) Harmonic dynamics of protiens: normal modes and fluctuations in bovine pacreatic trypsin-inhibitor, *Proc. Natl. Acad. Sci. USA 80*, 6571-6575.
109. Brooks, B., and Karplus, M. (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme, *Proc. Natl. Acad. Sci. U.S.A. 82*, 4995-4999.
110. Levitt, M., Sander, C., and Stern, P. S. (1985) Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme, *Journal of molecular biology 181*, 423-447.
111. Kukol, A. (2008) *Molecular modelling of proteins*, Vol. 443.
112. Ma, J. P. (2004) New advances in normal mode analysis of supermolecular complexes and applications to structural refinement, *Curr. Protein Pept. Sci. 5*, 119-123.
113. Cui, Q., and Bahar, I. (2005) *Normal mode analysis: Theory and application to biological and chemical systems (Mathematical and Computational Biology)*, 1 ed., Chapman & Hall.
114. Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., Walker, R. G., Zhang, W., Wang, B., Hayik, S., roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C.,

Mongan, J., Hornak, V., Cui, G., Beroza, P., Mathews, D. H., Schafmeister, C., Ross, W. S., and Kollman, P. A. (2006) AMBER 9, University of Califonial, San Francisco.

115. Durand, P., Trinquier, G., and Sanejouand, Y. H. (1994) New approach for determining low-frequency norma modes in macromolecules, *Biopolymers 34*, 759-771.
116. Tirion, M. M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis, *Physical Review Letters 77*, 1905-1908.
117. Bahar, I., Atilgan, A. R., and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, *Folding & Design 2*, 173-181.
118. Hinsen, K. (1998) Analysis of domain motions by approximate normal mode calculations, *Proteins-Structure Function and Genetics 33*, 417-429.
119. van Vlijmen, H. W. T., and Karplus, M. (2005) Normal mode calculations of icosahedral viruses with full dihedral flexibility by use of molecular symmetry, *Journal of molecular biology 350*, 528-542.
120. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem. 4*, 187-217.
121. Mouawad, L., and Perahia, D. (1993) Diagonalization in a mixed basis: A method to compute low-frequency normal modes for large macromolecules, *Biopolymers 33*, 599-611.
122. Perahia, D., and Mouawad, L. (1995) Computation of low-frequency normal modes in macromolecules: Improvements to the method of diagonalization in a mixed basis and application to hemoglobin, *Computers & Chemistry 19*, 241-246.
123. Thomas, A., Field, M. J., Mouawad, L., and Perahia, D. (1996) Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase, *Journal of molecular biology 257*, 1070-1087.
124. Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000) Building-block approach for determining low-frequency normal modes of macromolecules, *Proteins-Structure Function and Genetics 41*, 1-7.
125. Li, G. H., and Cui, Q. (2002) A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca^{2+}-ATPase, *Biophysical Journal 83*, 2457-2474.
126. Bertaccini, E. J., Trudell, J. R., and Lindahl, E. (2007) Normal-mode analysis of the glycine alpha1 receptor by three separate methods, *Journal of Chemical Information and Modeling 47*, 1572-1579.
127. Brooks, B., and Karplus, M. (1985) Normal modes for specific motions of macromolecules: Application to the hinge-bending mode of lysozyme, *Proceedings of the National Academy of Sciences of the United States of America 82*, 4995-4999.
128. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model, *Biophysical Journal 80*, 505-515.
129. Yang, L., and Chng, C. (2008) Coarse-grained models reveal functional dynamics - I. Elastic network models - Theories, comparisons and perspectives., *Bioinformatics and Biology Insights 2*, 25-45.

130. Wang, Y., Rader, A. J., Bahar, I., and Jernigan, R. L. (2004) Global ribosome motions revealed with elastic network model, *Journal of Structural Biology 147*, 302-314.
131. Yang, L. J., Tan, C. H., Hsieh, M. J., Wang, J. M., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P. A., and Luo, R. (2006) New-generation amber united-atom force field, *Journal of Physical Chemistry B 110*, 13166-13176.
132. Van Wynsberghe, A. W., and Cui, Q. (2006) Interpreting correlated motions using normal mode analysis, *Structure 14*, 1647-1653.
133. Ma, J. P., and Karplus, M. (1997) Ligand-induced conformational changes in ras p21: A normal mode and energy minimization analysis, *Journal of molecular biology 274*, 114-131.
134. Ma, J. P. (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes, *Structure 13*, 373-380.
135. Bahar, I., Lezon, T. R., Bakan, A., and Shrivastava, I. H. (2009) Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins, *Chem. Rev.*, Article ASAP.
136. Sanejouand, Y. H. (1996) Normal-mode analysis suggests important flexibility between the two N-terminal domains of CD4 and supports the hypothesis of a conformational change in CD4 upon HIV binding, *Protein Engineering 9*, 671-677.
137. Field, M. J. (2005) *A Practical Introduction to the Simulation of Molecular Systems*, Cambridge University Press.
138. Karplus, M., and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules, *Nature Structural Biology 9*, 646-652.
139. Kern, D., and Zuiderweg, E. R. P. (2003) The role of dynamics in allosteric regulation, *Current Opinion in Structural Biology 13*, 748-757.
140. Volkman, B. F., Lipson, D., Wemmer, D. E., and Kern, D. (2001) Two-state allosteric behavior in a single-domain signaling protein, *Science 291*, 2429-2433.
141. Stevens, S. Y., Sanker, S., Kent, C., and Zuiderweg, E. R. P. (2001) Delineation of the allosteric mechanism of a cytidylyltransferase exhibiting negative cooperativity, *Nat Struct Mol Biol 8*, 947-952.
142. Xu, Y., Shen, J., Luo, X., Shen, X., Chen, K., and Jiang, H. (2004) Steered molecular dynamics simulations of protein-ligand interactions, *Science in China Series B: Chemistry 47*, 355-366.
143. Isralewitz, B., Baudry, J., Gullingsrud, J., Kosztin, D., and Schulten, K. (2001) Steered molecular dynamics investigations of protein function, *Journal of Molecular Graphics and Modelling 19*, 13-25.
144. Isralewitz, B., Gao, M., and Schulten, K. (2001) Steered molecular dynamics and mechanical functions of proteins, *Current Opinion in Structural Biology 11*, 224-230.
145. Schlitter, J., Engels, M., Kruger, P., Jacoby, E., and Wollmer, A. (1993) Targeted molecular dynamics simulation of conformational change: Application to the T ↔ R transition in insulin, *Molecular Simulation 10*, 291-308.
146. Krüger, P., Verheyden, S., Declerck, P. J., and Engelborghs, Y. (2001) Extending the capabilities of targeted molecular dynamics: Simulation of a large conformational transition in plasminogen activator inhibitor 1, *Protein Science 10*, 798-808.

147. Grubmüller, H. (1995) Predicting slow structural transitions in macromolecular systems: Conformational flooding, *Physical Review E 52*, 2893-2906.
148. Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. (1993) Essential dynamics of proteins, *Proteins-Structure Function and Genetics 17*, 412-425.
149. Groot, B. L. d., Amadei, A., Scheek, R. M., Nuland, N. A. J. v., and Berendsen, H. J. C. (1996) An extended sampling of the configurational space of HPr from E. coli, *Proteins: Structure, Function, and Genetics 26*, 314-322.
150. Lu, B., Wong, C. F., and McCammon, J. A. (2005) Release of ADP from the catalytic subunit of protein kinase A: A molecular dynamics simulation study, *Protein Science 14*, 159-168.
151. Zhang, J., Lu, C. J., Chen, K., Zhu, W. L., Shen, X., and Jiang, H. L. (2006) Conformational transition pathway in the allosteric process of human glucokinase, *Proc. Natl. Acad. Sci. U.S.A. 103*, 13368-13373.
152. Tai, K. (2004) Conformational sampling for the impatient, *Biophysical Chemistry 107*, 213-220.
153. Schulze, B. G., Grubmuller, H., and Evanseck, J. D. (2000) Functional significance of hierarchical tiers in carbonmonoxy myoglobin: Conformational substates and transitions studied by conformational flooding simulations, *Journal of the American Chemical Society 122*, 8700-8711.
154. Phillips, S. C., Essex, J. W., and Edge, C. M. (2000) Digitally filtered molecular dynamics: The frequency specific control of molecular dynamics simulations, *Journal of Chemical Physics 112*, 2586-2597.
155. Verlet, L. (1967) Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules, *Physical Review 159*, 98-103.
156. Swope, W. C., Andersen, H. C., Berens, P. H., and Wilson, K. R. (1982) A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters, *The Journal of Chemical Physics 76*, 637-649.
157. Hockney, R. W. (1970) The potential calculation and some applications, *Methods Comput. Phys*, 135-211.
158. Beeman, D. (1976) Some multistep methods for use in molecular dynamics calculations, *Journal of Computational Physics 20*, 130-139.
159. Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numertical integration of cartesian equations of motion of a system with constraints - molecular dynamics of N-Alkanes, *Journal of Computational Physics 23*, 327-341.
160. Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A., and Haak, J. R. (1984) Molecular dynamcis with coupling to an external bath, *Journal of Chemical Physics 81*, 3684-3690.
161. Hoover, W. G. (1985) Canonical dynamics: Equilibrium phase-space distributions, *Physical Review A 31*, 1695-1697.
162. Andersen, H. C. (1980) Molecular dynamics simulations at constant pressure and-or temperature, *Journal of Chemical Physics 72*, 2384-2393.
163. Grest, G. S., and Kremer, K. (1986) Molecular dynamics simulation for polymers in the presence of a heath bath, *Physical Review A 33*, 3628-3631.

164. Rühle, V. W. [Last accessed: September 2009]. http://www.mpip-mainz.mpg.de/~andrienk/journal_club/thermostats.pdf.
165. Shell, M. S. (2009) Advanced molecular dynamics techniques.
166. Koopman, E. A., and Lowe, C. P. (2006) Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations, *The Journal of Chemical Physics 124*, 204103-204105.
167. Adelman, S. A., and Doll, J. D. (1976) Generalized langevin equation approach for atom-solid-surface scattering: general formulation for classical scattering off harmonic solids, *Journal of Chemical Physics 64*, 2375-2388.
168. Fadrna, E., Hladeckova, K., and Koca, J. (2005) Long-range electrostatic interactions in molecular dynamics: An Endothelin-1 case study, *J. Biomol. Struct. Dyn. 23*, 151-162.
169. Ewald, P. P. (1921) Die berechnung optischer und elektrostatischer gitterpotentiale, *Annalen der Physik 369*, 253-287.
170. Barker, J. A., and Watts, R. O. (1973) Monte-carlo studies of dielectric properties of water-like models, *Molecular Physics 26*, 789-792.
171. Allen, M. B., and Tildesley, D. J. (1989) *Computer simulation of liquids*, Oxford University Press.
172. Greengard, L., and Rokhlin, V. (1987) A fast algorithm for particle simulations, *Journal of Computational Physics 73*, 325-348.
173. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh ewald: An n.log(n) method for ewald sums in large systems, *Journal of Chemical Physics 98*, 10089-10092.
174. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method, *J. Chem. Phys. 103*, 8577-8593.
175. Sagui, C., and Darden, T. A. (1999) Molecular dynamics simulations of biomolecules: Long-range electrostatic effects, *Annual Review of Biophysics and Biomolecular Structure 28*, 155-179.
176. Cramer, C. J. (2002) *Essentials of computational chemistry: Theories and models.*, John Wiley & Sons, Ltd.
177. Bizzarri, A. R., and Cannistraro, S. (2002) Molecular dynamics of water at the protein-solvent interface, *Journal of Physical Chemistry B 106*, 6617-6633.
178. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics, *Journal of the American Chemical Society 112*, 6127-6129.
179. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water, *Journal of Chemical Physics 79*, 926-935.
180. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1981) Intermolecular forces, in *Interaction models for water in relation to protein hydration*, pp 331-342, Reidel, Dordrecht, Holland.
181. Lindahl, E., Hess, B., and van der Spoel, D. (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis, *Journal of Molecular Modeling 7*, 306-317.
182. Kaiser, H. F. (1960) The application of electronic computers to factor analysis, *Educational and Psychological Measurement 20*, 141-151.

183. Cattell, R. B. (1966) The scree test for the number of factors, *Multivariate Behav. Res. 1*, 245-276.
184. Website. [Last accessed: September 2009]. http://www.statsoft.com/textbook/stfacan.html.
185. Merlino, A., Vitagliano, L., Ceruso, M. A., and Mazzarella, L. (2003) Subtle functional collective motions in pancreatic-like ribonucleases: From ribonuclease A to angiogenin, *Proteins-Structure Function and Genetics 53*, 101-110.
186. vanAalten, D. M. F., Conn, D. A., deGroot, B. L., Berendsen, H. J. C., Findlay, J. B. C., and Amadei, A. (1997) Protein dynamics derived from clusters of crystal structures, *Biophysical Journal 73*, 2891-2896.
187. de Groot, B. L., Hayward, S., van Aalten, D. M. F., Amadei, A., and Berendsen, H. J. C. (1998) Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data, *Proteins-Structure Function and Genetics 31*, 116-127.
188. Ceruso, M. A., Amadei, A., and Di Nola, A. (1999) Mechanics and dynamics of B1 domain of protein G: Role of packing and surface hydrophobic residues, *Protein Science 8*, 147-160.
189. Tatsis, V., Tsoulos, I., and Stavrakoudis, A. (2009) Molecular dynamics simulations of the TSSPSAD peptide antigen in free and bound with CAMPATH-1H fab antibody states: The importance of the β -turn conformation, *International Journal of Peptide Research and Therapeutics 15*, 1-9.
190. Levitt, D. G., and Banaszak, L. J. (1992) POCKET - A computer-graphics method for identifying and displaying protein cavities and their surrounding amino-acids, *J. Mol. Graph. 10*, 229-234.
191. Huang, B. D., and Schroeder, M. (2006) LIGSITE(csc): predicting ligand binding sites using the connolly surface and degree of conservation, *Bmc Structural Biology 6*, 1-11.
192. Jackson, R. M. (2002) Q-fit: A probabilistic method for docking molecular fragments by sampling low energy conformational space, *Journal of Computer-Aided Molecular Design 16*, 43-57.
193. Website., Jackson, R. M., Laurie, A. T. R., and Fuller, J. (2004) Pocket-Finder. [Last accessed: September 2009]. http://www.modelling.leeds.ac.uk/pocketfinder/.
194. Website., Jackson, R. M., Laurie, A. T. R., and Fuller, J. (2004) Q-SiteFinder. [Last accessed: September 2009]. http://www.modelling.leeds.ac.uk/qsitefinder/.
195. Magnuson, M. A., and Matschinsky, F. M. (2004) Glucokinase as a glucose sensor: Past, present and future, in *Glucokinase and glycemic disease: From basics to novel therapeutics* (Matschinsky, F. M., and Magnuson, M. A., Eds.), pp 1-17, Frontiers in diabetes. Basel, Karger.
196. Iynedjian, P. (2009) Molecular physiology of mammalian glucokinase, *Cell. Mol. Life Sci. 66*, 27-42.
197. Matschinsky, F. M. (1990) Gluockinase as glucose sensor and metabolic signal generation in pancreatic beta-cells and hepatocytes, *Diabetes 39*, 647-652.
198. Matschinsky, F., Liang, Y., Kesavan, P., Wang, L. Q., Froguel, P., Velho, G., Cohen, D., Permutt, M. A., Tanizawa, Y., Jetton, T. L., Niswender, K., and

Magnuson, M. A. (1993) Glucokinase as pancreatic beta-cell glucose sensor and diabetes gene, *Journal of Clinical Investigation 92*, 2092-2098.
199. Christesen, H. B. T., Jacobsen, B. B., Odili, S., Buettger, C., Cuesta-Munoz, A., Hansen, T., Brusgaard, K., Massa, O., Magnuson, M. A., Shiota, C., Matschinsky, F. M., and Barbetti, F. (2002) The second activating glucokinase mutation (A456V) - Implications for glucose homeostasis and diabetes therapy, *Diabetes 51*, 1240-1246.
200. Grimsby, J., Sarabu, R., Corbett, W. L., Haynes, N. E., Bizzarro, F. T., Coffey, J. W., Guertin, K. R., Hilliard, D. W., Kester, R. F., Mahaney, P. E., Marcus, L., Qi, L. D., Spence, C. L., Tengi, J., Magnuson, M. A., Chu, C. A., Dvorozniak, M. T., Matschinsky, F. M., and Grippo, J. F. (2003) Allosteric activators of glucokinase: Potential role in diabetes therapy, *Science 301*, 370-373.
201. Matschinsky, F. M. (2009) Assessing the potential of glucokinase activators in diabetes therapy, *Nat. Rev. Drug Discov. 8*, 399-416.
202. Al-Hasani, H., Tschop, M. H., and Cushman, S. W. (2003) Two birds with one stone: Novel glucokinase activator stimulates glucose-induced pancreatic insulin secretion and augments hepatic glucose metabolism, *Mol. Interv. 3*, 367-370.
203. Barbor, M., Boyle, M., and Cassidy, M. (1997) *Biology*, Collins educational.
204. Pilkis, S. J., and Granner, D. K. (1992) Molecular physiology of the regulation of hepatic gluconeogenesis and glycolysis, *Annual Review of Physiology 54*, 885-909.
205. Porta, M., Matschinsky, F. M., and Magnuson, M. A. (2004) *Glucokinase and glycemic disease: From basics to novel therapeutics* S Karger AG.
206. Valera, A., Pujol, A., Pelegrin, M., and Bosch, F. (1994) Transgenic mice overexpressing phosphoenolpyruvate carboxykinase develop non-insulin-dependent diabetes-mellitus, *Proceedings of the National Academy of Sciences of the United States of America 91*, 9151-9154.
207. Rosella, G., Zajac, J. D., Baker, L., Kaczmarczyk, S. J., Andrikopoulos, S., Adams, T. E., and Proietto, J. (1995) Impaired glucose tolerance and increased weight-gain in transgenic rats overexpressing a non-insulin-responsive phophooenolpyruvate carboxykinase gene, *Molecular Endocrinology 9*, 1396-1404.
208. Galan, M., Vincent, O., Roncero, I., Azriel, S., Boix-Pallares, P., Delgado-Alvarez, E., Diaz-Cadorniga, F., Blazquez, E., and Navas, M. A. (2006) Effects of novel maturity-onset diabetes of the young (MODY)-associated mutations on glucokinase activity and protein stability, *Biochemical Journal 393*, 389-396.
209. VanSchaftingen, E., VeigadaCunha, M., and Niculescu, L. (1997) The regulatory protein of glucokinase, *Biochemical Society Transactions 25*, 136-140.
210. Voet, D. J., and Voet, J. G. (2004) *Biochemistry*, 3 ed., John Wiley & Sons.
211. Veiga-da-Cunha, M., and Van Schaftingen, E. (2002) Identification of fructose 6-phosphate- and fructose 1-phosphate-binding residues in the regulatory protein of glucokinase, *Journal of Biological Chemistry 277*, 8466-8473.
212. Anderka, O., Boyken, J., Aschenbach, U., Batzer, A., Boscheinen, O., and Schmoll, D. (2008) Biophysical characterization of the interaction between

hepatic glucokinase and Its regulatory protein: Impact of physiological and pharmacologival effectors, *Journal of Biological Chemistry 283*, 31333-31340.

213. Sagen, J. V., Odili, S., Bjorkhaug, L., Zelent, D., Buettger, C., Kwagh, J., Stanley, C., Dahl-Jorgensen, K., de Beaufort, C., Bell, G. I., Han, Y., Grimsby, J., Taub, R., Molven, A., Sovik, O., Njolstad, P. R., and Matschinsky, F. M. (2006) From clinicogenetic studies of maturity-onset diabetes of the young to unraveling complex mechanisms of glucokinase regulation, *Diabetes 55*, 1713-1722.
214. Website. [Last accessed: September 2009]. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=retrieve&dopt=default&rn=1&list_uids=2645#refseq.
215. Aleshin, A. E., Zeng, C. B., Bartunik, H. D., Fromm, H. J., and Honzatko, R. B. (1998) Regulation of hexokinase I: Crystal structure of recombinant human brain hexokinase complexed with glucose and phosphate, *Journal of molecular biology 282*, 345-357.
216. Jian, Z., Chenjing, L., Kaixian, C., Weiliang, Z., Xu, S., and Hualiang, J. (2006) Conformational transition pathway in the allosteric process of human glucokinase, *Proceedings of the National Academy of Sciences of the United States of America 103*, 13368-13373.
217. Zelent, B., Odili, S., Buettger, C., Shiota, C., Grimsby, J., Taub, R., Magnuson, M. A., Vanderkooi, J. M., and Matschinsky, F. M. (2008) Sugar binding to recombinant wild-type and mutant glucokinase monitored by kinetic measurement and tryptophan fluorescence, *Biochemical Journal 413*, 269-280.
218. Antoine, M., Boutin, J. A., and Ferry, G. (2009) Binding kinetics of glucose and allosteric activators to human glucokinase reveal multiple conformational states, *Biochemistry 48*, 5466-5482.
219. Ward, R. A. AstraZeneca. Alderley park, Macclesfield.
220. Website. [Last accessed: September 2009]. http://www.uniprot.org/uniprot/P35557.
221. Cardenas, M. L., Cornish-Bowden, A., and Ureta, T. (1998) Evolution and regulatory role of the hexokinase, *Biochimica Et Biophysica Acta-Molecular Cell Research 1401*, 242-264.
222. Holroyde, M. J., Allen, M. B., Storer, A. C., Warsy, A. S., Chesher, J. M. E., Trayer, I. P., Cornishbowden, A., and Walker, D. G. (1976) Purification in high-yield and characterization of rat hepatic glucokinase, *Biochemical Journal 153*, 363-373.
223. Storer, A. C., and Cornishbowden, A. (1976) Kinetics of rat-liver glucokinase - Cooperative interactions with glucose at physiologically significant concentrations, *Biochemical Journal 159*, 7-14.
224. Cárdenas, M. L., Rabajille, E., and Niemeyer, H. (1978) Maintenance of the monomeric structure of glucokinase under reacting conditions, *Arch. Biochem. Biophys. 190*, 142-148.
225. Cornish-bowden, A., and Cárdenas, M. L. (2004) Glucokinase: A monomeric enzyme with positive cooperativity, in *Glucokinase and glycemic disease: From basics to novel therapeutics.* (Matschinsky, F. M., and Magnuson, M. A., Eds.), pp 125-134, Basel, Karger.

226. Ricard, J., Buc, J., and Meunier, J. C. (1977) Enzyme memory .1. Transient kinetic study of wheat-germ hexokinase-I, *European Journal of Biochemistry 80*, 581-592.
227. Storer, A. C., and Cornishbowden, A. (1977) Kinetic evidence for a mnemonical mechanism for rat-liver glucokinase, *Biochemical Journal 165*, 61-69.
228. Ricard, J., Meunier, J. C., and Buc, J. (1974) Regulatory behavior of monomeric enzymes .1. Mnemonical enzyme concept, *European Journal of Biochemistry 49*, 195-208.
229. Meunier, J. C., Buc, J., Navarro, A., and Ricard, J. (1974) Regulatory behavior of monomeric enzymes .2. wheat-germ hexokinase as a mnemonical enzyme, *European Journal of Biochemistry 49*, 209-223.
230. Ralph, E. C., Thomson, J., Almaden, J., and Sun, S. X. (2008) Glucose modulation of glucokinase activation by small molecules, *Biochemistry 47*, 5028-5036.
231. Neet, K. E., Keenan, R. P., and Tippett, P. S. (1990) Observation of a kinetic slow transition in monomeric gclucokinase, *Biochemistry 29*, 770-777.
232. Lin, S. X., and Neet, K. E. (1990) Demonstration of a slow conformational chnage in liver glucokinase by fluorescence spectroscopy, *Journal of Biological Chemistry 265*, 9670-9675.
233. Heredia, V. V., Thomson, J., Nettleton, D., and Sun, S. X. (2006) Glucose-induced conformational changes in glucokinase mediate allosteric regulation: Transient kinetic analysis, *Biochemistry 45*, 7553-7562.
234. Kim, Y. B., Kalinowski, S. S., and Marcinkeviciene, J. (2007) A pre-steady state analysis of ligand binding to human glucokinase: Evidence for a preexisting equilibrium, *Biochemistry 46*, 1423-1431.
235. Froguel, P., Vaxillaire, M., Sun, F., Velho, G., Zouali, H., Butel, M. O., Lesage, S., Vionnet, N., Clement, K., Fougerousse, F., Tanizawa, Y., Weissenbach, J., Beckmann, J. S., Lathrop, G. M., Passa, P., Permutt, M. A., and Cohen, D. (1992) Cloase linkage of glucokinase locus on chromosome-7p to early-onset non-insulin-dependent diabetes-mellitus, *Nature 356*, 162-164.
236. Hattersley, A. T., Turner, R. C., Permutt, M. A., Patel, P., Tanizawa, Y., Chiu, K. C., Orahilly, S., Watkins, P. J., and Wainscoat, J. S. (1992) Linkage of type-2 diabetes to the glucokianse gene, *Lancet 339*, 1307-1310.
237. Njolstad, P. R., Sovik, O., Cuesta-Munoz, A., Bjorkhaug, L., Massa, O., Barbetti, F., Undlien, D. E., Shiota, C., Magnuson, M. A., Molven, A., Matschinsky, F. M., and Bell, G. I. (2001) Neonatal diabetes mellitus due to complete glucokinase deficiency, *N. Engl. J. Med. 344*, 1588-1592.
238. Njolstad, P. R., Sagen, J. V., Bjorkhaug, L., Odili, S., Shehadeh, N., Bakry, D., Sarici, S. U., Alpay, F., Molnes, J., Molven, A., Sovik, O., and Matschinsky, F. M. (2003) Permanent neonatal diabetes caused by glucokinase deficiency: inborn error of the glucose-insulin signaling pathway, *Diabetes 52*, 2854-2860.
239. Glaser, B., Kesavan, P., Heyman, M., Davis, E., Cuesta, A., Buchs, A., Stanley, C. A., Thornton, P. S., Permutt, M. A., Matschinsky, F. M., and Herold, K. C. (1998) Familial hyperinsulinism caused by an activating glucokinase mutation, *N. Engl. J. Med. 338*, 226-230.

240. Gloyn, A. L., Odili, S., Buettger, C., Njolstad, P. R., Shiota, C., Magnuson, M. A., and Matschinsky, F. M. (2004) Glucokinase and the regulation of blood sugar, in *Glucokinase and glycemic disease. From basics to novel therapeutics* (Matschinsky, F. M., and Magnuson, M. A., Eds.), pp 92-109, Karger, Basel.
241. Matschinsky, F. M., Magnuson, M. A., Zelent, D., Jetton, T. L., Doliba, N., Han, Y., Taub, R., and Grimsby, J. (2006) The network of glucokinase-expressing cells in glucose homeostasis and the potential of glucokinase activators for diabetes therapy, *Diabetes 55*, 1-12.
242. Coghlan, M., and Leighton, B. (2008) Glucokinase activators in diabetes management, *Expert Opinion on Investigational Drugs 17*, 145-167.
243. Agius, L. (2007) New hepatic targets for glycaemic control in diabetes, *Best Practice & Research Clinical Endocrinology & Metabolism 21*, 587-605.
244. Wilson, J. E. (1995) Hexokinases, *Rev. Physiol. Biochem. Pharmacol. 126*, 65-198.
245. Pal, P., and Miller, B. G. (2009) Activating mutations in the human glucokinase Gene revealed by genetic selection, *Biochemistry 48*, 814-816.
246. Leighton, B., Atkinson, A., and Coghlan, M. P. (2005) Small molecule glucokinase activators as novel anti-diabetic agents, *Biochem. Soc. Trans. 33*, 371-374.
247. Schuttelkopf, A. W., and van Aalten, D. M. F. (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes, *Acta Crystallographica Section D-Biological Crystallography 60*, 1355-1363.
248. Vriend, G. (1990) WHAT IF: a Molecular modeling and drug design program, *Journal of Molecular Graphics 8*, 52-56.
249. Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method, *Journal of Computational Chemistry 21*, 132-146.
250. Jakalian, A., Jack, D. B., and Bayly, C. I. (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation, *Journal of Computational Chemistry 23*, 1623-1641.
251. Loncharich, R. J., Brooks, B. R., and Pastor, R. W. (1992) Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-actylananyl-N'-methylamide, *Biopolymers 32*, 523-535.
252. Izaguirre, J. A., Catarello, D. P., Wozniak, J. M., and Skeel, R. D. (2001) Langevin stabilization of molecular dynamics, *Journal of Chemical Physics 114*, 2090-2098.
253. Turner, P. J. XMGRACE. Center for coastal and land-margin research oregon graduate institute of science and technology beaverton, Oregon
254. Heredia, V. V., Carlson, T. J., Garcia, E., and Sun, S. X. (2006) Biochemical basis of glucokinase activation and the regulation by glucokinase regulatory protein in naturally occurring mutations, *Journal of Biological Chemistry 281*, 40201-40207.
255. vanAalten, D. M. F., Amadei, A., Bywater, R., Findlay, J. B. C., Berendsen, H. J. C., Sander, C., and Stouten, P. F. W. (1996) Comparison of structural and dynamic properties of different simulation methods applied to SH3, *Biophysical Journal 70*, 684-692.
256. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD - Visual molecular dyanmics, *Journal of Molecular Graphics 14*, 33-38.

257. Brown, R. A., and Case, D. A. (2006) Second derivatives in generalized born theory, *J. Comput. Chem. 27*, 1662-1675.
258. Van Wynsberghe, A., Li, G., and Cui, Q. (2004) Normal-Mode Analysis Suggests Protein Flexibility Modulation throughout RNA Polymerase's Functional Cycle *Biochemistry 43*, 13083-13096.
259. Moth, C., Callahan, T., Swanson, E., and Lybrand, T. (2002) [Last accessed: September]. http://structbio.vanderbilt.edu/~cmoth/mddisplay/.
260. Callahan, T. J., Swnson, E., and Lybrand, T. P. (1996) MD Display: An interactive graphics program for visulalization of molecular dynamics trajectories, *J. Mol. Graph. 14*, 39-41.
261. Prime, version 2.0, Schrödinger, LLC, New York, NY, 2008.
262. Molnes, J., Bjorkhaug, L., Sovik, O., Njolstad, P. R., and Flatmark, T. (2008) Catalytic activation of human glucokinase by substrate binding: Residue contacts involved in the binding of D-glucose to the super-open form and conformational transitions, *FEBS J. 275*, 2467-2481.
263. Takahashi, K., Hashimoto, N., Nakama, C., Sasaki, K., Yoshimoto, R., Ohyama, S., Hosaka, H., Maruki, H., Nagata, Y., Eiki, J. I., and Nishimura, T. (2009) The design and optimization of a series of 2-(pyridin-2-yl)-1H-benzimidazole compounds as allosteric glucokinase activators, *Bioorganic & Medicinall Chemistry 17*, 7042-7051.
264. Amadei, A., Linssen, A. B. M., De Groot, B. L., van Aalten, D. M. F., and Berendsen, H. J. C. (1996) An efficient method for sampling the essential subspace of proteins, *J. Biomol. Struct. Dyn. 12*, 615-625.
265. Meyer, T., Ferrer-Costa, C., Perez, A., Rueda, M., Bidon-Chanal, A., Luque, F. J., Laughton, C. A., and Orozco, M. (2006) Essential dynamics: A tool for efficient trajectory compression and management, *Journal of Chemical Theory and Computation 2*, 251-258.
266. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research 22*, 4673-4680.
267. Website. [last accessed: September 2009]. http://www.ebi.ac.uk/Tools/clustalw2/index.html.