Long-time methods for MD simulations: Markov State Models and Milestoning

Brajesh Narayan

The thesis is submitted to University College Dublin in fulfilment of the requirement for the degree of

Doctor of Philosophy

School of Physics

Principal Supervisor: Assoc. Prof. Nicolae-Viorel Buchete
Co Supervisor: Assoc. Prof. Christina Kiel

May 2022

Table of Contents

Acknowledgements

I would like to thank my advisor, Prof. Nicolae-Viorel Buchete, and my co-advisor, Prof. Christina Kiel, for giving me the great opportunity to carry out research under their supervision throughout my PhD. I would also like to thank Prof. Ron Elber for his kind guidance and supervision throughout my PhD. I am extremely grateful for the guidance and help of Prof. Martin Karplus, Mrs. Marci Karplus and Dr. Victor Ovchinnikov. Finally, I would like to thank my family, without their support my studies and this thesis would have not been possible.

For funding I thank the Thomas Preston PhD Scholarship fund and UCD.

Abstract

In this thesis, the aim is to contribute to the development of long-time methods for molecular dynamics (MD) simulations such as the Milestoning method and Markov State Model (MSM), and their application to study the conformational dynamics of systems ranging from the small piezoelectric amyloid peptide diphenylalanine (FF) to much larger cancer-related systems such as the Abl and K-Ras proteins.

Albeit all the progress in high-performance computing capacity, MD simulations of large systems with atomistic detail is computationally very expensive and requires advanced methods that can aid probing slow kinetics. Markov State Modelling and the Milestoning method are two fast developing novel and powerful computational approaches that can accurately capture molecular kinetics besides thermodynamics, and yet are relatively simple and easy to implement. The MSM approach relies of mapping the complex underlying dynamics of complex biomolecular systems on relatively simple networks with nodes corresponding to stable conformational states that are interconnected through Markovian transitions. The Milestoning method is useful where the sampling problem can be stated in terms of estimating the thermodynamics and kinetics along transition pathways that connect two known metastable states of the underlying molecular system. One could target the sampling in order to identify and characterize the most probable pathway(s) and the corresponding intermediate states that are relevant to the underlying reaction mechanism.

Firstly, I showed that the thermodynamics and kinetics of the ensemble of conformations adopted by amyloid FF peptides solvated in explicit water molecules can be analysed in detail by using an efficient enhanced sampling method, replica exchange molecular dynamics (REMD), while simultaneously applying external electric fields and probing a range of temperatures. I also showed that even for such a small system, there could be possible artifacts arising from due to the coupling the exchange with external fields, and I proposed how to overcome these artifacts in our simulations.

Next, I combined a reaction path algorithm with the theory and algorithm of Milestoning to study kinetics of the DFG flip and disassociation of Gleevec from the Abelson murine leukaemia viral oncogene homolog (Abl) kinase. This allowed me to probe the detailed mechanism for the unbinding transition, at a timescale longer than accessibly by conventional MD studies. This also allowed the accurate calculation of the slow underlying kinetic timescales from our sets of short atomistic MD trajectories, while sampling the unbinding pathway of Gleevec from Abl and providing detailed insight into the corresponding dissociation kinetics.

Finally, using also sets of appropriately initialized yet relatively short trajectories, I analysed the underlying free energy landscape of K-Ras4B and unveiled new information on its underlying conformational states, and sheds new light on the activation/inactivation mechanism. This new MSM study of K-Ras, based on sets of short trajectories approach, unveils details underlying its equilibrium conformational kinetics, including the role of cancer-relevant mutations and the corresponding changes in activation/inactivation propensities.

Statement of original authorship

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere based on the research presented in this submitted work.

Brajesh Narayan

Collaborations

In my work presented in chapter 3, I have been fortunate to collaborate with Colm Herbert, Prof. Brian Rodriguez (School of Physics, Dublin, Ireland), and Dr. Bernard R. Brooks (Laboratory of Computational Biology, National Institutes of Health, Bethesda, Maryland 20892, United States). Chapter 4 and chapter 5 is in collaboration with Prof. Ron Elber (UT Austin, USA) and Prof. Ronald M. Levy (Temple University, USA).

Author Contributions

In chapter 3, B.N., C.H., B.R, B.R.B., and N.V.B. designed research, B.N. and C.H. performed research, B.N., C.H. and N.V.B. analyzed data and B.N. and N.V.B. wrote the paper. In chapter 4, B.N., A.R., C.T., P.H., S.A., R.E., N.V.B, and R.M.L. designed research, B.N., A.R., and C.T. performed research, B.N., A.R., C.T., and R.E analyzed data and B.N., A.R., and R.E wrote the paper. In chapter 5, B.N., N.V.B., and R.E. designed research, B.N. performed research, B.N. and R.E. analyzed data and B.N., N.V.B., and R.E wrote the paper. In chapter 6, B.N., C.K., and N.V.B. designed research, B.N. performed research, B.N. and N.V.B. analyzed data and B.N. and N.V.B wrote the paper.

Abbreviations

Abl	Abelson murine leukaemia viral oncogene homolog
AMBER	Assisted Model Building with Energy Refinement
CHARRM	Chemistry at Harvard Macromolecular Mechanics
DFG	Aspartic Acid-Phenylalanine-Glycine (Asp-Phe-Gly)
FF	Diphenylalanine
GDP	Guanine diphosphate
GROMACS	Groningen Machine for Chemical Simulations
GTP	Guanine triphosphate
HPC	High-performance Computing
K-Ras	Kristen rat sarcoma viral oncogene homologue
MC	Monte Carlo
MD	Molecular Dynamics
MFPT	Mean First Passage Time
MSM	Markov State Model
NAMD	Nanoscale Molecular Dynamics
NMR	Nuclear Magnetic Resonance
RAF	Rapidly Accelerated Fibrosarcoma
REMD	Replica Exchange Molecular Dynamics
RMSD	Root Mean Square Deviation
SMD	Steered Molecular Dynamics
TBA	Transition Based Assignment
TREMD	Temperature Replica Exchange Molecular Dynamics
WT	Wild type

List of figures

aligning a trajectory to structures corresponding to representative conformations for state 1 and state 2, respectively. As expected, these are most informative only for RMSD values close to zero. (b) By combining the two signals in $RelRMSD_{12}$ (red, circles) the overall ability to discriminate between the two states is improved, and the assignment of states can be done more accurately (for example, by using the transition-based assignment method) resulting in a discretised trajectory (horizontal lines).

Figure 2.8. The RelMSD calculation can be used multi-dimensionally in the state assignment step for discriminating between multiple MSM states (in this case three conformational states).

Figure 2.9. Example of using RelRMSD for MSM analysis of the conformational dynamics of a small diphenylalanine (FF) peptide (replica exchange MD simulation with explicit water molecules).

Figure 2.10. Schematic 1d representation of two typical approaches to sampling free energy barriers in molecular simulations using (a) one or a few typically long trajectories, and (b) short trajectories initialized along the transition pathway between two end points (e.g., reactants, R, and products, P). In (b), central in the implementation of the Milestoning method, the aim is to achieve higher computational efficiency by initializing the short trajectories from multi-dimensional hypersurfaces (vertical lines, blue) that are located between anchors (red dots) along a low dimensional reaction coordinate (see also Fig. 2.7).

Figure 2.11. In Milestoning, the sampling is initiated by first defining a set of “anchors” (red) conformations that span a (typically low dimensional) transition pathway between two end points (e.g., reactants, R, and products, P). The underlying kinetic and thermodynamic information is extracted by analysing sets of short trajectories initiated in conformational space regions (blue, milestones, where M_{ij} is the milestone between anchors *i* and *j*, respectively). The short trajectories (green) are terminated as soon as they reach another milestone, different from the one at which they were initiated.

Figure 2.12. Example of using a multi-dimensional Milestoning approach in a large and more complex molecular system: the characterization of the DFG-flip dynamics in Abl kinase.

Figure 3.1. Representative conformations of FF peptides in the absence of externally applied electric fields. Values of the d_{ee} distances (i.e., distances between the CZ atoms at the ends of the two sidechains, in Å), are shown in black.

Figure 3.2. (a) Distributions of potential energy values (*U*, in kcal/mol) calculated from REMD simulations in the presence of an external electric field with an intensity of E = 30 kcal/mol·Å·e. **(b)** Illustration of the problems that could occur when attempting REMD simulations in external electric fields. The presence of the field can induce some (in this case the first two) replicas to adopt conformations that are significantly lower in energy than the corresponding initial conformational states of the other replicas. This is a serious artifact, as illustrated in (a), as it changes the expected equilibrium *U* distributions.

Figure 3.3. Distributions of potential energy values (*U*, in kcal/mol) calculated from REMD simulations (a) at E = 0 kcal/mol·Å·e, and (b) with corrected initial conditions in the presence of an external electric field with an intensity of E = 30 kcal/mol·Å·e.

Figure 3.4. Distributions of RMSD values calculated for the heavy atoms of FF peptides for conformations from REMD simulations in the presence of external electric fields with intensities of (a) E = 0 kcal/mol·Å·e, (b) E = 30 kcal/mol·Å·e, and (c) E = 45 kcal/mol·Å·e.

Figure 3.5. Distributions of the dipole moment magnitude (*m*, Debye units), calculated for FF peptides for conformations from REMD simulations in the presence of external electric fields with intensities of (a) E = 0 kcal/mol·Å·e, (b) E = 30 kcal/mol·Å·e, and (c) E = 45 kcal/mol·Å·e.

Figure 3.6. Replica exchange equilibrium distributions of sidechain-sidechain distances of FF amyloid peptides, with no external electric field applied, (a) for each replica (R-trajectories), and (b) at each temperature (T-trajectories) of the REMD simulation set.

Figure 3.7. Distributions of sidechain-to-sidechain distances, d_{ee}, for simulations with an applied electric field of 30 kcal/mol·Å·e, (a) for each replica (R-trajectories), and (b) at each temperature (T-trajectories) of the REMD simulation set. Note that, at this field intensity, the conformational dynamics is restricted to one extended structure with a most probable d_{ee} value of ~8.9 Å.

Figure 3.8. Distributions of d_{ee} values for simulations with an applied electric field of 45 kcal/mol·Å·e, (a) for each replica (R-trajectories), and (b) at each temperature (T-trajectories) of the REMD simulation set. At this field intensity, the conformational dynamics is restricted further to a single extended structure with a most probable d_{ee} value of ~10 Å.

Figure 3.9. Representative conformations of FF amyloid peptides derived by kinetic analysis of REMD simulations at different electric fields. In the absence of electric fields, the FF peptide adopts three main Markovian conformational states: S_1, S_2 and S_3 (top). The corresponding equilibrium transition rates between these states (blue arrows, see text) are shown as numbers. These REMD rates are for the data corresponding to all the replicas (all R-trajectories). Each arrow's thickness is proportional to the magnitude of its corresponding transition rate. On the bottom are shown the representative conformations, S_2' and S_2'', adopted in presence of external electric fields with intensities of E = 30 kcal/mol·Å·e, and E = 45 kcal/mol·Å·e, respectively.

Figure 4.1. A schematic representation of the Abl kinase protein. The active and inactive states of the activation loop is shown in yellow and red, respectively. The C helix is green. The magnified region shows the start of the activation loop that includes the DFG switch. The image was generated by the software VMD.[1]

Figure 4.2. The reduction in the norm of force, averaged over the entire reaction path, as a function of iteration number. The norm of the force drops rapidly in the first ten minimization steps and then decays more gradually. It seems to stabilize at around 100 iterations. The final gradient is around 0.35 kcal/mol $Å^{-1}$. It is not zero since the norm of the force along the reaction coordinate is included.

Figure 4.3. A schematic representation of the discretization of the coarse space following the transition pathway. R and P represent the reactant and product states, respectively. The black line shows the reaction pathway. The red dots are the anchors, and the blue lines are the milestones. Every milestone is numbered by its corresponding anchors. For example, milestone (*j,k*) is the boundary between cells *j* and *k*. The green arrows show 4 unbiased trajectories initiated from milestone (*i,j*). The trajectories are terminated when they hit any other milestone for the first time. Re-crossing the original milestone does not lead to trajectory termination.

Figure 4.4. Representation of all the 171 milestones considered for computing the transition matrix. Every point corresponds to one milestone. The red points are the initial milestones between the consecutive anchors along the transition path. The blue points represent the milestones discovered during the analysis of the free trajectories that are used to enrich the sampling of the pathways.

Figure 4.5. Free energy of pairs of anchors as computed from the Milestoning theory. The energy values are in Kcal/mol. The two paths with maximum flux from the reactant to the product are lines in red and in magenta. Note that the significantly off-diagonal “jumps” on the surface are a consequence of long-range connection between milestones that are not in sequence along the reaction pathway (see also Fig. 4.4).

Figure 4.6. Two optimal free energy profiles along the two max-flux pathways from active to inactive state. In panel a, the milestones are numbered from 1 to 41 for the corresponding points along the red path shown in figure 4.5 starting from active state, and from 1 to 32 for the magenta path in figure 4.5, starting from the inactive state for panel b.

Figure 4.7. Color-coding the committor function at every milestone. The committor of a milestone is the probability of a complete trajectory initiated at that milestone to reach the product before the reactant state.

Figure 4.8. A stick model of residues 381 to 386 for active (yellow), inactive (red), and a sample configuration at anchor 23 where the committor value is near 0.5 (blue). Note that Arg386 already reached its final position at the transition state, while Asp381 did not change its configuration significantly. The residue Phe382 is found at half of the way of the transition.

Figure 4.9. Changes in the salt bridge between Lys271 and Glu286 for inactive (A), an intermediate state (B), and active (C) states. The salt bridge exists in the active state and inactive states but during the transition from active to inactive state, the salt bridge breaks. The intermediate state shown is anchor 27. The DFG residues are shown in red, blue, and yellow for inactive, intermediate, and active states, respectively.

Figure 4.10. Distribution of lifetime for (a) milestone (21,22), (b) milestone (22,23), (c) milestone (24,25) and (d) milestone (40,41).

Figure 4.11. Distributions of MFPT for transition from active to inactive states (top) and the reverse process (bottom). The insets show the corresponding distributions for 1/MFPT which are estimates of the rate coefficients consistent with the simulation data and the error analysis. We quote the mean values of the rate coefficients for the forward and backward transition.

Figure 4.13. Fluctuations and systematic drifts of residues in Abl-kinase. Top panel reports the B factors of the reactant and product structures as a function of the residue index to identify flexible domains. In the lower panel we compare the structure of Anchor 27 with the reactant and product using room mean square difference (RMSD) between all the heavy atoms of the residues in the protein. The two pink arrows point to Glu286 and Lys271 that forms a blocking salt bridge. Note that the transition state differs about equally from the reactant and from the product structures. There are several spikes at Phe382, Leu384, Arg386, Met388 and Ala397 that belongs to the A loop and are included in the set of coarse variables.

Figure 4.14. The distances between atom CZ of Phe382 to atoms representing the ends of the sidechain of residues Met290, Glu286, and Lys271 are shown. For Glu286, the minimum distance two both oxygens OE1 and OE2 was measured. Lys271 and Glu286 are the salt bridge residues. Around anchor 30, the Phe382 reaches as close as possible to the salt bridge, (Lys 271) breaks it and then goes away from this residue. The distance to Glu286 remains roughly a constant throughout the transition.

Figure 5.1. The potential energy (kcal/mol) and the distance of Gleevec from the binding pocket as a function of pulling time (ns). Gleevec is pulled out of the binding pocket at a small constant velocity (0.5 Å/ns) to reduce system strain, providing a flat sampling of the underlying potential energy (top panel). The lower panel shows the center of mass distance (in Å, see main text for the definition) as the function of time (ns). 43 configurations are selected from this SMD trajectory to serve as anchors in the Milestoning calculations.

Figure 5.2. Abl kinase in complex with Gleevec. Abl kinase with Gleevec bound (reactant) is shown in red. Abl kinase with unbound Gleevec (product) is shown in green. Gleevec has been highlighted by representing it as opaque and using a translucent/shaded texture for the kinase matrix.

Figure 5.3. Schematic representation of the reaction space in two dimensions. The axes represent coarse variables. The unfilled circles represent reactant (R) and product (P) conformations. The filled blue circles are the anchors, obtained by exploratory Steered Molecular Dynamics calculation, and serving as centers of Voronoi cells. The dashed lines are the milestones or the boundaries of the Voronoi cells. Note that anchors placed on a straight reaction pathway segment appear connected to only two neighboring milestones, forward and backwards. However, in more dimensions, a milestone can be connected to more than two milestones.

Figure 5.4. Representation of the 126 milestones used in the present study. The axes are the anchor indices. A milestone between anchors *i* and *j* is represented by the point (*i, j*). The red dots are the initial milestones between the consecutive anchors along the transition path. The blue dots represent the 84 new milestones discovered during the analysis of the free trajectories from the initial set of only 42 milestones.

Figure 5.5. Network representation of the anchors' space. Each node represents an anchor, there are 43 anchors in total. The first node represents the reactant (1st anchor, shown in red). The last node represents the product (43rd anchor, is shown in green). A connection between any two anchors is represented by a straight black line. There are 126 connections (or milestones) in total.

Figure 5.6. Free energy plot for 126 milestones. The free energy of every milestone is colored according to its numerical value in kcal/mol (see color-bar). The maximum flux (Max Flux) pathway is shown in red (see text for details).

Figure 5.7. Free energy profile (kcal/mol) along the maximum weight path. To estimate the mean values (dots, red) and the errors (standard deviation, red vertical lines) we sampled the transition matrices and lifetimes from their Milestoning model distribution, using a set of 1000 samples. The committor-estimated transition state (TS) appears to be located late and broad

between milestone 20 and 30 along the reaction coordinate. The energy minimum for Gleevec unbinding is seen near milestone 12. Structural differences between the reactant (shown in yellow) and the structure at milestone 12 (shown in pink) have been illustrated in the inset figure. At this minimum, we observe the outward displacement of Gleevec (IMA), show with blue arrow. We also observe an outward rotation in the αC helix (see inset blue arrow). There is an RMSD difference of 2.6Å. The center of mass distance of Gleevec between the two structures is ~4.5Å.

Figure 5.8. Color-coded committor function at each milestone. The committor function, C_i, is defined as the probability that a trajectory initiated at milestone *i* will reach the product before the reactant. Milestones with committor values close to 0.5 are candidates for the transition state and have been highlighted with red squares.

Figure 5.9. Representative structure from the Transition State Ensemble (TSE) estimated using the committor function (called TS-1). The image was generated using the VMD software.[32] At this position, the probability to return to the bound state is equal to the probability of escaping the protein to the aqueous solution.

Figure 5.10: Transition function (defined as the logarithm of the ratio of the exit times towards the product and the reactant (Eq. (4)).[21] Milestones with similar exit times to both the product and the reactant, are close to the transition state. The region highlighted by the blue contains milestones (marked with small blue squares) near the transition state with the transition function close to zero. The R and P states are located inside the red and the green boxes, respectively, representing reactant and products.

Figure 5.11. A representative TSE structure found using the transition function (Eq. 4). By construction, the transition function value is ~0 for TSE conformations, as the exit times to the product and reactant are equal. The image was generated using VMD.[32]

Figure 5.12. Conformational changes along the Gleevec dissociation reaction pathway. GLU282-LYS274 is shown in green and Gleevec is shown in yellow. Gleevec, when inside the binding pocket blocks the direct interaction between GLU282 and LYS274, shown in (a). Panel (b) represents the configuration at the transition state 2. (c) Finally, when the Gleevec molecule is out of the kinase matrix, the distance reduces to ~2Å.

Figure 5.13. The Abl kinase sequence - comparison with the corresponding residues from Src. The two kinase sequences have 50.6% sequence identity and 69.0% sequence similarity. Identical residues are colored in red. P-loop, αC helix and A-loop are highlighted in yellow, cyan and green, respectively.

Figure 5.14. Significant differences in salt bridge interactions are observed between Abl and Src kinases. We show salt bridges that are formed in the Abl kinase, making significant contribution to the reaction pathway, and are modified in Src kinase. P-Loop, A-Loop, αC helix, salt bridges and Gleevec are shown in blue, green, red, magenta, and yellow, respectively. Inset graphs show the changes in the salt bridge distance as a function of milestone positions along GMW path. At the bottom left of each panel are shown the salt-bridges for the Abl-kinase

case (highlighted with green), and the corresponding residues for Src-kinase (cyan). The corresponding Abl and Src residues that are not similar have been boxed in red.

Figure 5.15. Highly conserved LYS271-GLU286 salt-bridge. For the kinase to be active, DFG needs to be in the 'in' conformation, LYS271-GLU286 salt bridge should be formed, the catalytic spine that involves the residues Asp421, His361, Phe382, Met290, and Leu301 needs to be formed, and the binding site should be accessible to ATP. Thus, the integrity of the LYS271-GLU286 salt bridge is central to kinase activity. Shown above is the LYS271-GLU286 salt bridge in green, Gleevec (yellow), and the kinase (cyan). During the unbinding of Gleevec, this salt bridge breaks near the transition state 2 (panel b). The distance between the two residues increases from ~3Å (panel a) to ~7.8Å (panel b). The bond is formed again when Gleevec is completely out of the kinase matrix (panel c).

Figure 5.16. Dihedral angles along the GMW path. As Gleevec moves away from the binding pocket to slide out of the Abl kinase, the steric hinderance decreases, and an increase in the movement and rotation is observed. The dihedral angle, Θ, defined by the C8, C15, C32 and C37 atoms of Gleevec, was recorded for the milestones along the GMW path from reactant to product. Larger ranges of dihedral angles at milestones outside the binding pocket suggests greater flexibility and entropy. The onset of larger flexibility is near transition state 2.

Figure 5.17: Distribution of first passage times for Gleevec dissociation from the Abl kinase. The corresponding MFPT value derived from Milestoning calculations is 0.055 s. Note the broad distribution of predicted MFPT values suggesting significant uncertainties in the calculations.

Figure 6.1. **(a)** K-Ras4B structural elements. The switch I and switch II regions are highlighted in yellow and magenta, respectively. The GTP ligand is shown in licorice and colored by atom type. The C_a (for residues T35 and G60) and P_b (for GTP) atoms are shown as blue spheres. d_1 is the distance (dashed black line) between the C_a atom of G60 and the P_b atom of GTP. d_2 is the distance between the C_a atom of T35 and the same P_b atom of GTP (dashed black line). The D33 residue is circled in red. **(b)** Schematic representation of the GTP-dependent activation of K-Ras4B and the hypothesized relationship between its active/inactive states and the d_1 and d_2 distances.

Figure 6.2. **(a)** The detailed K-Ras4B wild type (WT) free energy landscape (ΔG, kcal/mol) for its GTP-bound structure in d_1 and d_2 (Å) coordinates (see Fig. 1a). The six main conformational states of K-Ras4B WT are labelled S_1 to S_6, (yellow). **(b)** The corresponding ΔG landscape calculated for the GTP-bound K-Ras4B D33E mutant, with the new conformational basins labelled S_1' to S_6'.

Figure 6.3. Visual representation of rate matrix with rates (in ns^{-1}) shown along the lines. Relative population (in %) of each state is shown along the nodes. S2 is the most populated state with 46.648 % population and S6 is the least populated state with 3.579 %. Estimate of error in rates and population has been shown in supplementary table S6.2 and table S6.3.

Figure 6.4. (a) The positions of atoms defining the angle q (i.e., the C-C_a-C_b-C_g dihedral angle for residue D38, in degrees) used for measuring the relative orientation (w.r.t. the local

backbone) of the D38 side chain in the K-Ras4B WT with respect to the local backbone. The GTP ligand is shown as licorice and the D38 atoms as balls-and-sticks. **(b)** The corresponding overall distribution of θ_{38} values for K-Ras4B WT. **(c)** Dihedral angle (θ_{38}) distributions in each of the six states of K-Ras4B WT. **(d)** The corresponding dihedral angle (θ_{38}) distributions in each of the six states of K-Ras4B D33E.

Figure 6.5. RAF-RAS binding interface. **(a)** Crystal structure of K-Ras4B-GNP in complex with RAF1 (from PDB ID 6XI7)[2]. The binding interface for RAF1is shown with surface representation and colored with residue type and binding interface residues of K-Ras are shown with sticks and colored with residue type. Acidic residues of K-Ras like residue 37 and 38 (shown in red and circled) binds with the basic interface residues (shown in blue) on RAF1 binding surface. **(b)** RAF1 docked to S2' peak1 and peak2 structures. RAF1 shown in cyan is docked to S2' peak 1 and RAF1 shown in purple is docked to S2' peak 2. The two structures share only a small part of binding interface. RAF1bound to peak 2 structure is closer to the crystal structure RAF1.

Figure S5.1. Thermodynamic cycle for alchemical free energy calculations. We compare the free energy changes of the bound (left) and transition states (right) for Gleevec interactions with the wild-type Abl Kinase and Y253F mutant. The mutated residue and the ligand (green) are shown using a CPK representation, in color.

Figure S5.2. Initial and final Abl (ribbon) and Gleevec (licorice) structures and relative positions for the three inbound (i.e., moving towards the binding pocket) trajectories.

Figure S6.1. Representative structures for the six conformational states of K-Ras4B WT, S_1 to S_6, evidenced by the corresponding free energy map of GTP-bound K-Ras4B in the d_1-d_2 coordinates (in Å, see Fig. 6.2). Note differences in the relative positions of the switch I and switch II regions highlighted with red and blue arrows, respectively (see also Fig. 6.1). The sets of (d_1, d_2) coordinates of the representative structures selected here as centers of the S_1 to S_6 regions are (6.07, 6.45), (6.1, 8.58), (8.16, 10.45), (8.66, 8.82), (11.86, 8.57) and (9.01, 6.52), respectively.

Figure S6.2. Distributions of the distances from switch I to GTP. Shown is the distribution of distance between alpha carbon of residues 32-40 and beta phosphate of GTP. Clearly, the d_2 distance (i.e., using the alpha carbon of T35) is the best reaction coordinate as it can discriminate more states.

Figure S6.3. **(a)** Corresponding locations of the d_1 and d_2 values from experimental crystal structures overlapped on the free energy map of GTP-bound K-Ras4B WT (see also Fig. 2a). The positions of crystal structures of K-Ras and H-Ras bound to GTP (or GTP analogue), and to GDP (or GDP analogue) are highlighted in black and blue, respectively. The corresponding PDB codes for these structures are shown in the legend, using superscript K or H to distinguish between K-Ras and H-Ras structures, respectively. Wild type structures are marked with *. **(b)** Values of the corresponding d_1 and d_2 distances (in Å) and of the angle θ_{38} (the C-C_a-C_b-C_g dihedral angle for residue D38, in degrees) for experimental PDB structures.

Figure S6.4. Slowest relaxation time with respect to change in window length. A sliding window was used to build the transition probability matrix and slowest relaxation time was estimated using the second eigenvalue. For final analysis, window length of 20 ns was used. Inset is the error in slowest relaxation time, for 20 ns window, with change in the diameter of for-sure-zone.

Figure S6.5. Distribution of docking scores, using PatchDock, obtained for docking the K-Ras4B representative structures S1 and S6 (see Fig. 2) to the CRAF1 (from PDB ID 6XI7). Note that PatchDock is appropriate as it successfully at predicts only a few complex structures with high scores, including structures of the binding interface that have a small RMSD from the experimental interface (PDB ID 6XI7).

List of tables

1. Introduction

Experimentally studying biological systems and reactions in atomistic detail, is often impossible due to their intrinsic complexity. Both conceptual and computational challenges arise due to the large number of degrees of freedom involved. However, molecular dynamics (MD) simulation-based methods offer a promising alternative by allowing researcher to perform simulations of many biological systems at atomistic level of detail. It can help us to connect microscopic interactions and structures with thermodynamic properties and, unveil the kinetics and mechanisms of molecular processes. Using MD methods, we can calculate physics-based parameters such as the underlying free energy landscapes, mean first passage times (MFPT), relative populations, etc., which could have been difficult or, often, impossible to obtain experimentally. Atomically detailed simulations can also supplement experimental data to obtain a more complete understanding of the underlying kinetics.

The development of computational methods has progressed greatly in last few decades. The first MD simulations were performed in 1957.[4] Simulation of protein was first carried out in 1977.[5] In 2013, the Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Ariel Warshel for development of multiscale models for complex chemical systems, groundwork that enabled these simulations.

Despite all the progress, MD simulations come with the significant limitation that their time step (typically one or a few femtoseconds) and the corresponding total simulations lengths feasible on current computers (up to milliseconds) are much shorter than the relaxation times relevant to most biomolecular processes. This limitation makes MD simulations, especially of systems as big as the Abelson murine leukaemia viral oncogene homolog (ABL) protein, computationally very challenging and expensive. Although the availability of high-performance computing (HPC) hardware capacity has grown considerably and is likely to continue its remarkable growth, the high complexity of accurate kinetic and thermodynamic studies based on

molecular dynamics simulations of biomolecular systems requires novel modelling methods that can be accurate yet simple and relatively easy to implement and, advanced methods to extend the time scale of the simulations and produce trajectories probing slow kinetics.

One approach, that takes advantage of the intrinsically parallel architecture HPC facilities or the highly distributed computing projects such as Folding@Home,[6] is the use Markov State Models (MSMs). It relies on mapping the complex underlying dynamics of otherwise complex biomolecular systems on relatively simple networks with nodes corresponding to stable conformational states that are interconnected through Markovian transitions. Thus, both the design and the goals of MD studies are targeted towards identifying the regions of the corresponding configuration space that are characterized by a sufficient conformational stability such that subsequent inter-state transition are independent from each other (i.e., Markov states). This approach has strong roots in the statistical mechanics of biomolecules being related to the projection operator formalism. [7]

A second approach, which takes advantage of parallel architecture of HPC and which does not require states to be Markovian, is the Milestoning method. This method can be used for systems where the sampling problem can be stated in terms of estimating the thermodynamics and kinetics along transition pathways that connect two known metastable states (e.g., reactant and product) one could target the sampling in order to identify and characterize the pathway(s) (typically with the maximum reaction flux) and the corresponding intermediate states that are relevant to the underlying reaction mechanism.

In this thesis, I present the development and proper application of the two master-equation-based methods for MD simulations, on two very important proteins, namely K-Ras4B and ABL kinase. Kristen rat sarcoma viral oncogene homologue (K-Ras) is a GTPase that controls cellular proliferation by playing an important role in the signal transduction pathway.[8] It acts as a molecular switch, flipping between inactive guanine diphosphate (GDP) bound state and active form of guanine triphosphate (GTP) bound state. K-Ras is one of the most mutated oncogenes and has been associated with many fatal cancers like colorectal cancer, pancreatic ductal

adenocarcinoma and lung cancer.[9-13] Many computational and experimental efforts have been made to understand the conformational dynamics of K-Ras, effects of mutations and to find mutation specific drugs. [8, 14, 15] Just like K-Ras, mutations in ABL, a kinase, is also associated with certain cancers like chronic myelogenous leukaemia. Kinases are enzymes that catalyse the transfer of the γ-phosphate group from an ATP molecule to the hydroxyl group of the serine, threonine or tyrosine residue. Thus, they act as effective switches along cellular transduction pathways, because of their ability to alternate between catalytically active and inactive state in response to specific molecular signals. Hence, kinases play important roles in cell growth, proliferation and differentiation. In cancer, uncontrolled division of cells and malignant transformations are direct consequences of kinase deregulation. Abelson murine leukaemia viral oncogene homolog 1 (ABL) is a kinase protein that, in humans, is encoded by the ABL gene located on chromosome9.[16] ABL encodes cytoplasmic and nuclear protein tyrosine kinase, which is involved in the process of cell division, adhesion, differentiation and response to stress. ABL's activity is controlled by its SH3 domain. Absence of SH3 domain makes ABL oncogene. Such mutation in the ABL gene has been associated with chronic myelogenous leukaemia. [17]

In **chapter 2**, I introduce the theory and methods underpinning the work in this thesis. In the first section, the theory guiding Molecular dynamic simulations is briefly described. In the next section of the chapter, replica exchange molecular dynamics (REMD) is discussed. REMD is used to improve the sampling. In the next two sections, theory of the two approaches, MSM and the Milestoning method, is discussed.

In **chapter 3**, I study the temperature-dependent conformational dynamics of FF peptides solvated in explicit water molecules, an environment relevant to biomedical applications, by using an enhanced sampling method, REMD, in conjunction with applied electric fields. Simulations highlight and overcome possible artifacts that may occur during the setup of REMD simulations of explicitly solvated peptides in the presence of external electric fields, a problem particularly important in the case of short peptides such as FF. The presence of the external fields could over-stabilize certain conformational states in one or more REMD replicas, leading to distortions of the underlying potential energy distributions observed at each temperature.

In **chapter 4**, I combine a reaction path algorithm with the theory and algorithm of Milestoning to study kinetics of the three residue motif Aspartic acid-Phenylalanine-Glycine (DFG) flip and compute the mechanism and the rate of the transition in ABL kinase. The activation of kinases includes a conformational transition of the DFG motif that is important for enzyme activity but is not accessible to conventional MD. I propose a detailed mechanism for the transition, at a timescale longer than conventional MD, using a combination of reaction path and Milestoning algorithms. The mechanism includes local structural adjustments near the binding site as well as collective interactions with more remote residues.

In **chapter 5**, I use atomically detailed simulations within the Milestoning framework to study the molecular dissociation mechanism of Gleevec from Abl Kinase. I compute the dissociation free energy profile, the mean first passage time for unbinding, and explore the transition state ensemble of conformations. The milestones form a multidimensional network with average connectivity of about 2.93, which is significantly higher than the connectivity for a one-dimensional reaction coordinate (RC). I examined the transition state conformations using both, the committor and transition function. I show that near the transition state the highly conserved salt bridge of K217 and E286 is transiently broken. Together with the calculated free energy profile, these calculations can advance the understanding of the molecular interaction mechanisms between Gleevec and Abl kinase and play a role in future drug design and optimization studies.

In **chapter 6**, I probe the equilibrium conformations adopted by GTP-bound K-Ras4B proteins using long-time atomistic molecular dynamics (MD) simulations. I analyse the underlying free energy landscape of wildtype K-Ras4B projected on two important distances, labelled d_1 and d_2 (i.e., coordinating the P_β atom of the GTP ligand with two key residues, T35 and G60), that are useful reaction coordinates for discussing the K-Ras4B activation/inactivation mechanism. However, the detailed inspection of the K-Ras4B conformational landscape reveals a more complex network of underlying equilibrium states. I show that including a new reaction coordinate to account for the orientation of acidic K-Ras4B sidechains such as D38, with respect to the interface with binding effectors such as RAF1, is needed to rationalize the

activation/inactivation propensities. I also show that a relatively minor mutation, D33E, in the switch 1 region can lead to significantly different activation propensities of monomeric K-Ras4B. This study shades new light on the role of residues located at the K-Ras4B – RAF1 interface on its underlying GTP-dependent activation/inactivation mechanism.

Finally, **chapter 7**, provides a concise summary of all the work and the main results presented in this thesis and published or prepared for publication in peer-reviewed articles. [18-21]

2. Theoretical and Computational Methods[1]

2.1 Overview

This chapter introduces the theory and methods underpinning the work in this thesis. In the first section, the theory guiding Molecular Dynamics, which is the main tool used for simulations in this work, is briefly described. In the next section of the chapter, replica exchange molecular dynamics (REMD) is discussed. REMD is used to improve the sampling. In the next two sections, theory of the two approaches, Markov State Modelling and the Milestoning method, is discussed.

2.2 Molecular Dynamics

Molecules like lysozyme, alanine, etc. are normally represented as static structures, but in fact these are dynamic. Most experimental properties for example, measure a time or an ensemble average over the range of possible confirmations the molecule can adopt. One way to study the range of possible configurations is to simulate the motions of the molecule. Molecular Dynamics is one of the main simulation technique, other being Monte Carlo (MC). Advantage of MD over MC is that MD is deterministic and gives a route to dynamical properties of the system. MD simulations consists of step by step numerical solution of Newton's equations of motion. For a simple atomic system, we can write

$$f_i = m_i \ddot{r}_i$$

$$f_i = -\frac{\partial}{\partial r_i} U$$

where U = potential energy, m_i = mass, and r_i = position of atom i.

Factors like, degrees of freedom, solvation effects, boundary conditions, treatment of temperature and pressure, force field parameters, etc. govern the outcome of any MD simulation. Force field and force field parameters can be considered as one of the most important factors.

1. This chapter has been adapted from reference [19].

Force Fields:

In molecular dynamics, a molecule is considered as a series of charged points where points represent atoms linked by spring, where springs represents bonds. To find the time evolution of bond lengths, bond angles and torsions, along with non-bonding van der Waals, electrostatic, etc. interactions between atoms, an atomistic force field [22, 23] is used. Such a force field is collection of equations and associated parameters designed to reproduce molecular geometry and selected properties of tested structures. In other words, it is a mathematical expression describing energy dependence molecule on the coordinates of atoms in it. Force Field is made up of two components:

1)Set of equations called potential function

2)Parameters used in potential function.

The simplest form of potential function can be written as

$$U(\vec{R}) = \sum_{bonds} k_i^{bond}(r_i - r_0)^2 + \sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2 + \sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i \emptyset_i + \delta_i)] + \sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\pi D r_{ij}}$$

First term is due to all the bonds. It describes linear spring like terms for every bond. k^{bond} is the spring constant and r_0 is the equilibrium bond length. Second term describes potential due to all the angle (e.g. CCC, OCH, COH, …). θ_0 is the equilibrium angle and k^{angle} is the force constant. Similarly, the third term describes torsional or dihedral motion. $\emptyset$ is torsional angle, n is the number of maxima or minima in between 0 to 2π and δ is the phase. The last two terms are due to non-bonded interactions (van der Waals and Coulombic interactions respectively). ϵ is van der Waals's Lennard-Jones (LJ) well-depth, σ is LJ radius, q is partial atomic charge and D is dielectric constant. Apart from these terms there can be many other terms, for example due to dipole-dipole interactions, hydrogen bonds, etc.

All interatomic potentials are derived from various experimental data and have many approximations. These are empirical. With advances in the experimental

techniques and in our understanding of the behaviour of atoms, there is improvement in the force fields (ff), as well.

Force fields can be of several different types, such as:

1. All atom - parameters given for every atom in the system, including hydrogen
2. Coarse grained - atoms are grouped into super atoms and molecule is represented by these
3. United atom – except non-polar hydrogen, parameters given for all atoms in the system. In united atom ff, hydrogen and carbon atoms in methyl groups and methylene is considered as a single interaction centre.

Some commonly used force fields for molecular dynamics of macromolecules are AMBER [24] force field, CHARMM [25] force field and GROMOS [26]. These are also commonly used for minimizing the energy. Detailed discussion of all available force fields is beyond the scope of this thesis. Some good reviews on this topic can be found in refs. [24, 27, 28]. For all research work presented in this thesis CHARMM force field was used.

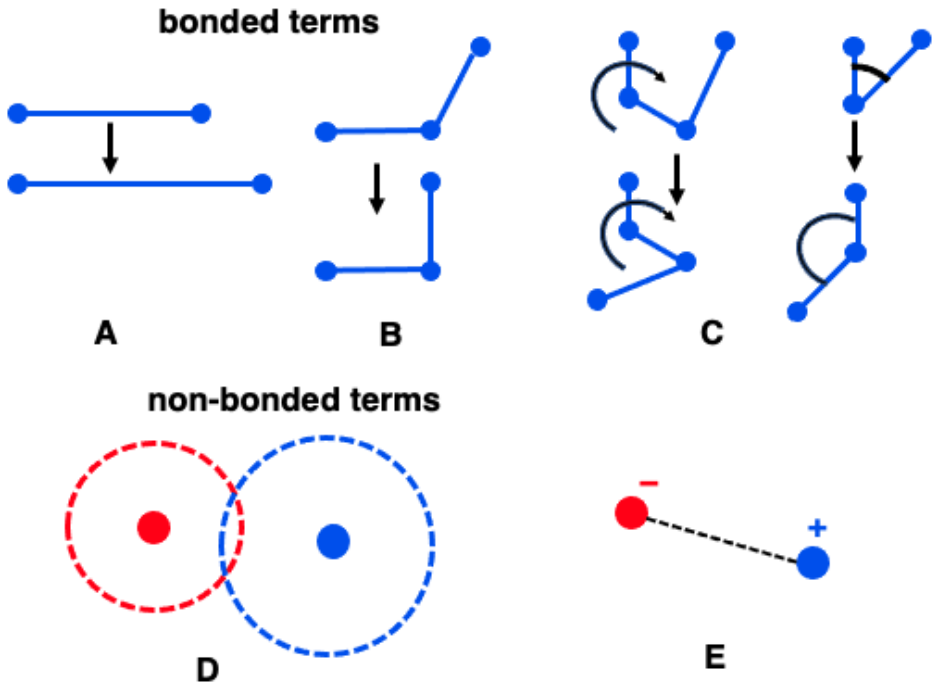


Figure 2.1. Visual representation of bonded and non-bonded terms in the potential function defined above. A represents bond length; B represents bond angle; dihedral term is represented by C; non-bonded terms Vander Waals interaction and electrostatic interaction is represented by D and E respectively.

Numerical Integration:

Once we know the potential function, U(R), and the corresponding force field parameters, we can obtain the dynamical trajectory of a system of N atoms. We integrate Newton's equations of motions, that is, solve the classical equation of motion.

$$m_i \frac{d^2 r_i}{dt^2} = f_i = -\frac{\partial}{\partial r_i} U(r_1, r_2, \dots, r_N)$$

where $U(r_1, r_2, \dots, r_N)$ is the potential energy depending on the coordinates of the N particles. This set of second order non-linear differential equations is solved numerically, step by step using integration algorithm. By using numerical integrator, we can generate an approximate solution trajectory given a time-step and initial positions and velocities. Various numerical integration algorithms used with MD are-

1. Verlet algorithm [29],
2. Leap-frog method and
3. velocity-Verlet algorithm.

All the methods mentioned above produce deterministic dynamical systems, that is, no stochastic element is there. Leap Frog algorithm is essentially the same as velocity-Verlet and give equivalent trajectories. The difference between the two algorithm is that the velocities are not calculated at the same time as positions and, the leap frog and the velocity-Verlet have different restart files. These algorithms have been summarized in ref. [30].

Thermodynamic Ensembles

Statistical Ensemble is an idealization consisting of a large number of copies of a system, each of these copies represents a possible state in which the real system might be in. A statistical ensemble that is in statistical equilibrium is a thermodynamic ensemble. Before initializing the MD simulation, one must select a thermodynamic ensemble, which will depend on the properties one need to study. For the purpose of MD simulation three main ensembles can be considered, the microcanonical, canonical and isothermal-isobaric ensembles.

1. Microcanonical Ensemble: It is a system that is completely isolated from its surrounding, such that there is no transfer of energy or matter between the system

and the surroundings. Here the total number of particles (N), the total volume (V) and the total energy (E) are constant, thus this ensemble is abbreviated as NVE.

2. Canonical Ensemble: In canonical ensemble, energy can transfer across the boundary between system and surroundings but matter cannot. Also the volume of the system is fixed. The system is immersed in a heat bath at a temperature (T), and temperature here is constant. Canonical ensemble is abbreviated as NVT.
3. Isothermal-Isobaric Ensemble: Similar to canonical ensemble, in isothermal-isobaric ensemble, energy can transfer across the boundary but matter cannot. Also, similar to canonical, the system is immersed in a heat bath at a temperature (T), and temperature is constant. Unlike canonical, here volume is not constant and it changes such that the internal pressure of the system matches the pressure on the system by surroundings and pressure is constant. This ensemble is abbreviated as NPT.

A more detailed discussion can be found in ref. [31]

It is relatively easy to achieve the microcanonical (NVE) ensemble in a simulation. To achieve NVT ensemble, where we need to make temperature constant, we need to use a temperature control algorithm (also called thermostat). In case of NPT ensemble, in addition to thermostat for temperature control, a barostat (pressure control algorithm) is also required.

Thermostats maintain a constant temperature by modifying velocities of subsets of particles in the system. Some of the thermostats used in MD simulations are discussed briefly below.

1. Berendsen thermostat: It rescales the velocities of all particles to remove a predefined fraction of the difference from the predefined temperature. This analogous to coupling the system to a heat bath kept at a constant temperature.
2. Andersen thermostat: It controls the temperature by assigning a subset of atoms new velocities that are randomly selected from the Maxwell-Boltzmann distribution. This is similar to every atom, on average, experiencing a stochastic collision with a virtual particle every time step.
3. Langevin thermostat: This thermostat mimics the viscous aspect of a solvent and interaction with the environment by adding a frictional force and a random force

to the equation of motion. The amount of friction is controlled by the damping coefficient.

4. Nosé-Hoover thermostat: An artificial variable associated with a fictional heat bath mass is introduced to the equations of motion. This thermostat can control temperature without involving random numbers, thus correlated motions are not impaired. The drawback of using this thermostat is that it imparts the canonical distribution and ergodicity.

These are discussed in more detained in ref. [32-35]

Barostats regulate pressure by adjusting volume of the system. This is achieved by scaling coordinates of each atom in the system by a small factor and thus changing the size of the system. Some commonly used barostats are mentioned here.

1. Berendsen barostat: Conceptually similar to Berendsen thermostat. It changes the volume by an increment proportional to the difference between the internal pressure and pressure in a weakly coupled bath.
2. Andersen barostat: Pressure is controlled by introducing an additional degree of freedom corresponding to the volume of a simulation box which adjusts itself to equalize the internal and external pressure. Other barostats like Parrinello-Rahman barostat, the Nosé-Hoover barostat, etc, are all based on this.
3. Langevin piston Barostat: This is based on Langevin thermostat.

These are discussed in more detail in ref. [32-38]

General steps involved in MD simulation

The standard steps followed in any MD simulation are as follows:

1. Generate Topology: To generate topology, initial structure of desired protein/ molecule is required. Atomic coordinate file of the molecule, obtained through X ray crystallography or NMR measurement, is available in .pdb, .gro, etc. format. In case crystal or NMR structure of desired biomolecule is not available, homology modelling can be used to obtain the initial structure.
2. Solvate: Once we have a structure to start with, we need to define the simulation box of required size. The box should be large enough that the protein does not interact with its image in case of periodic boundary conditions. Box and the molecule is solvated using suitable water model eg. TIP3P water model. To

replicate physiological conditions certain ions like, NaCl, can also be added to the system. If required molecule can be simulated without solvation and addition of ions.

3. Energy minimization: Since the atomic coordinates of protein and water box are obtained from different source there may be steric clashes. To remove any such clashes, energy minimization is attempted, usually using steepest descent algorithm.
4. Equilibration: After minimization step, the system is equilibrated with position restraint placed on protein, so that water molecules and protein atoms can relax and reach an equilibrium state. In this step, the system is heated to desired and equilibrated. Heating is done with small increments in temperature and followed by equilibration step. Equilibration step is required to re-equilibrate out the energies that have been introduced to the system during heating. For high temperature simulations, this is an important step for reaching a stable initial structure. Normally for high-temperature simulations heating and equilibration are done in several steps until the desired temperature is reached. Gradual heating is used to avoid the simulation from crashing or the system from degenerating, which is a possibility when trying to heat a system in one step.
5. MD simulation: After our system is equilibrated, we are ready to run MD simulation for required time. Alternately, we can also run replica exchange molecular dynamics simulation.

2.3 Replica Exchange MD

Sugita and Okamoto in 1999 [39] were the first to formulate parallel tempering for molecular dynamics called replica-exchange molecular dynamics. Now, it has become one of the most widely used tool for molecular dynamics simulation [40-46].

The algorithm helps to run multiple MD simulations in parallel at a sequence of increasing temperatures (the distribution of temperature can be uniform or exponential; it has been shown that exponential distribution is better and more efficient see ref. [47]). The initial conditions and structures are same, apart from temperature. Each of these structures at different temperature are called replicas. After every n-steps, we

try to swap simulations between neighbouring replicas, say j and k= j+1. n should be large enough for replicas to relax after a swap [48, 49] and small enough [50, 51] to keep computational cost low and improve convergence. Typically in most REMD studies exchange is attempted after 1ps to 3 ps [44] (1000 to 3000 steps in 1 fs integration time step).

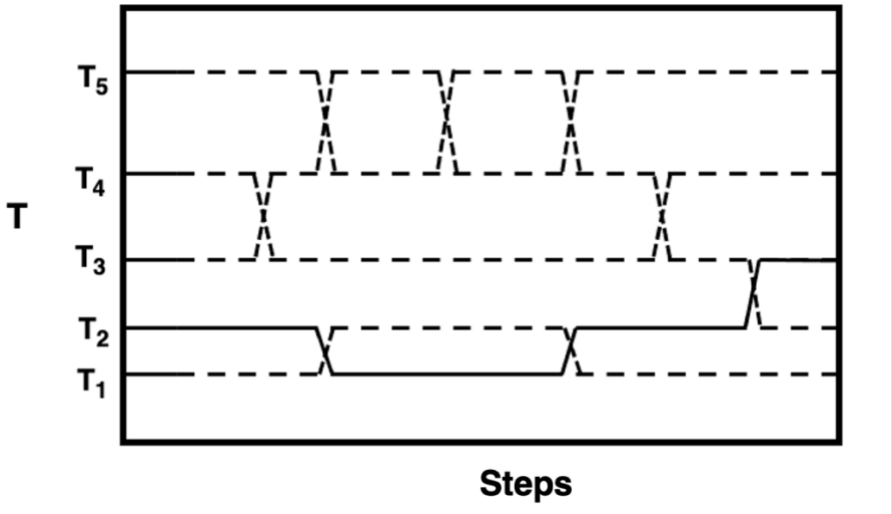


Figure 2.2. Plot of temperature replica exchange molecular dynamics (TREMD) trajectory. Trajectory originating from temperature T_2 has been represented by solid line to show how it visits other temperatures, as well (T_1, T_3).

We accept the swap with the probability given by Metropolis ratio [39]:

$$min\left\{1, \frac{\pi_{T_k}(q_j,p_j)\pi_{T_j}(q_k,p_k)}{\pi_{T_k}(q_k,p_k)\pi_{T_j}(q_j,p_j)}\right\}$$

The distribution $\pi_{T_k}(q_j, p_j)$ is the Boltzmann distribution for replica j at temperature T_i. From above ratio, we can get probability to visit from state x to x' as

$$P(x \to x') = \begin{cases} 1 & for \quad \Delta \leq 0 \\ \exp(-\Delta) & for \quad \Delta > 0 \end{cases}$$

where Δ can is defined as

$$\Delta = \left(\frac{1}{k_B T} - \frac{1}{k_B T'}\right)[U(x') - U(x)]$$

Temperature for replicas are selected such that there is at least 20 percent overlap of probability distribution curves of two neighbouring replicas. If the temperatures are far apart then there won't be overlap and Δ would be very large thus the probability of exchange will be negligible. For the same reason, we attempt exchange between the neighbouring replicas and not any two replicas.

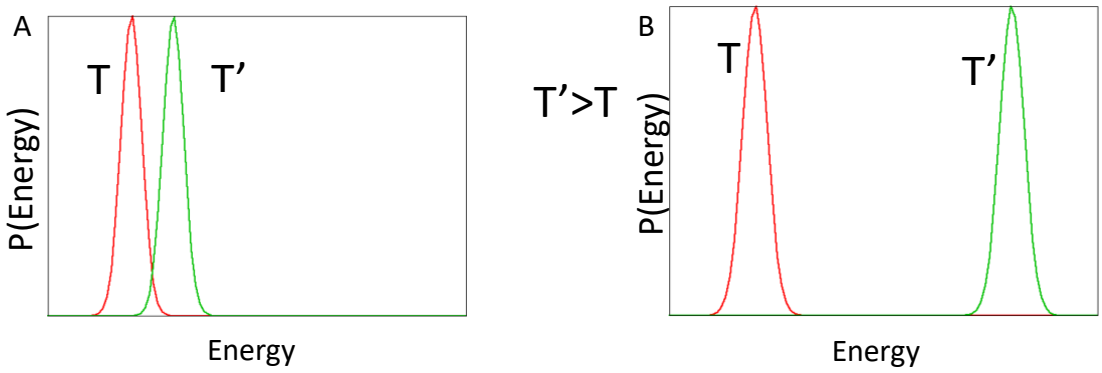


Figure 2.3. Probability distribution of energy at temperature T and T', where T' > T has been shown in above figures. In part (A), the difference between temperatures is not large, thus, there is an overlap between the plots. This overlap is required for exchange to be accepted, in REMD. In part (B), the difference between T and T' is very large, i.e., the replicas as very far from each other and hence there is no overlap. In such a case, Δ becomes very large and the probability of acceptance tends to zero. In general, the distribution of temperature is selected such that there is about 20% overlap between the neighbouring replicas.

This stochastic dynamical system on $X = \mathbb{R}^{2dn}$ has enabled the crossing of large energy barriers and the efficient exploration of the corresponding energy landscape. The output of REMD can be represented in form of continuous and discontinuous trajectory. Continuous trajectory is called R-trajectory while the discontinuous trajectory is called T-trajectory (as shown in figure 2.5).

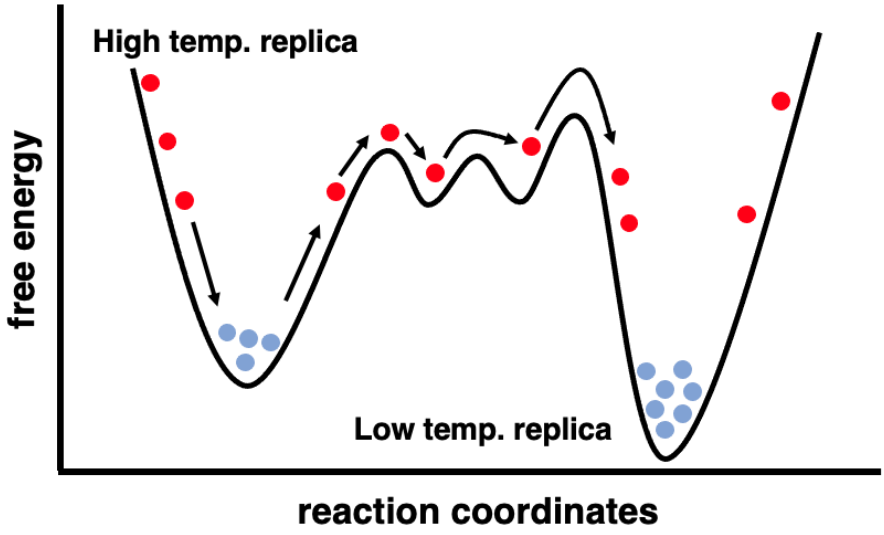


Figure 2.4. Possible energy landscape of a molecule. Energy landscape of a molecule is very complex, with many valleys and hills. If we run simulation at only one temperature then we might get to see some states at and nearby local minima, as shown in figure with cyan colour dots. Instead, if we run simulation at a range of temperatures using REMD, we can visit many other possible states as well (red dots form run at high temperature and blue dots from run at low temperature). Thus, we get to study the whole landscape. In REMD, a better sampling is achieved for every temperature used. This makes REMD a much better tool to study a system.

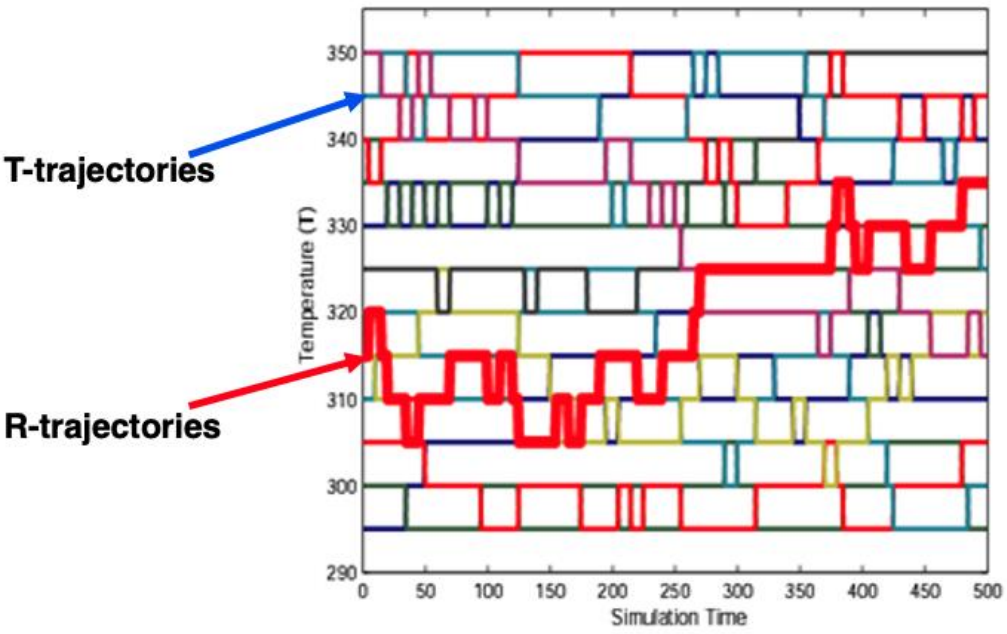


Figure 2.5. Horizontal lines of different colours at a temperature is called T-trajectory. It is a discontinuous trajectory. R-trajectory is a continuous trajectory which visits other temperatures as well. Here, it has been shown with a solid red line. If the REMD is run long enough, it visits all temperatures used for the simulation.[52]

Typically, in any REMD simulation, the atomistic coordinates are saved at each temperature, and these are termed as T-trajectories. History of exchange events is also saved in a separate exchange data file. Using information in both the files we can generate R-trajectories. These corresponds to a replica R as it progresses to different temperature after accepted exchange. These trajectories are continuous, unlike T-trajectories. It can be used to assign states with TBA method[53]. It is only after this step, by using again the exchange history data, that states can be also assigned accurately along the more typical REMD *T*-trajectories, enabling thus the temperature-dependent investigation of the dynamics.

2.4 Markov State Modelling

Although the availability of high-performance computing hardware capacity is likely to continue its remarkable growth, the high complexity of accurate kinetic and thermodynamic studies based on molecular dynamics simulations of biomolecular systems requires novel modelling methods that can be accurate yet simple and relatively easy to implement. One approach, that takes advantage of the intrinsically parallel architecture HPC facilities or the highly distributed computing projects such as Folding@Home[6], relies of mapping the complex underlying dynamics of otherwise complex biomolecular systems on relatively simple networks with nodes corresponding to stable conformational states that are interconnected through Markovian transitions. Thus, both the design and the goals of MD studies are targeted towards identifying the regions of the corresponding configuration space that are characterized by a sufficient conformational stability such that subsequent inter-state transition are independent from each other (i.e., Markov states). This approach has strong roots in the statistical mechanics of biomolecules being related to the projection operator formalism.[7]

A Markov State Model (MSM) can be used to describe the dynamics of the system. MSM is a square matrix, generally made of transition probabilities between two states. The whole space is divided into states and by determining the state of MD simulations, we can track the dynamical progress of the system in state space by writing down which state the trajectory is at time points separated by τ, referred to as the lag time or window length. To build a MSM, a Markovian lag time is selected. In

such case the system is memoryless meaning that the probability of going a state *j*, after an increment of τ, given it is in state *i* does not depend on where the system was before it entered state *i*. The goal is to choose the n states such that they best capture the dynamics of the system and are interpretable, and a lag time that is long enough to be Markovian but short enough to resolve the dynamics and keep computational cost low.

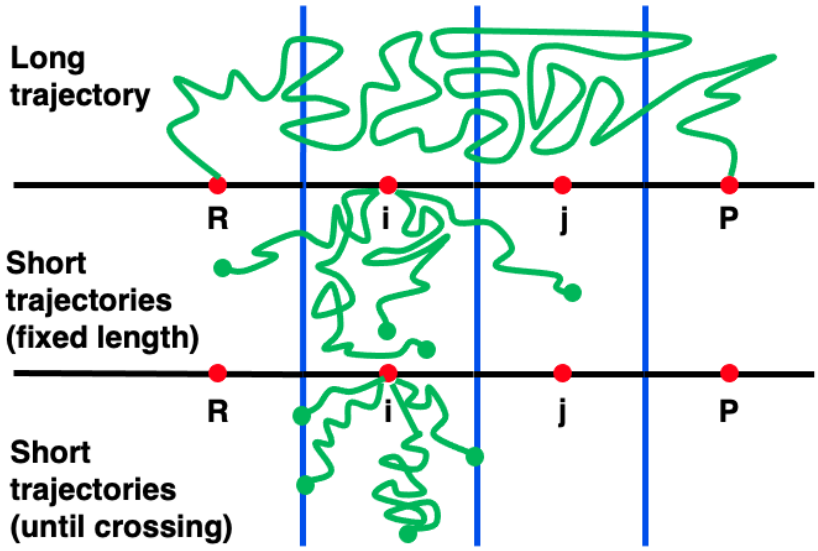


Figure 2.6. Markov State Models aim to sample accurately the underlying free energy landscape by use of either long or short equilibrium trajectories. Consider the sampling of the conformational space region containing a typical transition pathway between two end points (e.g., reactants, R, and products, P). Using short, rather than long equilibrium trajectories brings major sampling advantages. In typical MSM simulations, sets of short trajectories are initialized from intermediate conformations between R and P regions that are candidates for Markovian states. During analysis, transition probabilities between candidate states are extracted at different time intervals (also called lag times) and the states can be either divided or combined (clustered) until the transitions between them can be shown to be truly Markovian and a proper MSM is built. Short trajectories can be simulated either for the same duration ("fixed length") or only until crossing into nearby regions (macro-states).

Once the Markovian transition probabilities between different states are inferred from either long or short trajectories (see Fig. 2.6), they can be used to build transition probability matrices corresponding to different lag times and, ultimately, time-independent rate matrices that contain a complete representation (in the limits of the underlying Markovian model) of the thermodynamics and kinetics of the corresponding

biomolecular system.[21, 54, 55] Due to their popularity and usefulness,[56] several packages are available to assist both the implementation and the analysis of MSMs from molecular trajectories as well as several useful reviews.[56-61] Here, we highlight the basic theoretical considerations behind extracting Markovian transition probabilities and rate matrices from molecular simulations.

Markovian approach: Markov State Models for MD Simulations

Markov State Models are becoming increasingly popular as they have proven to be a useful approach to both generating and analysing the results of a broad range of molecular dynamics simulations, from folding/unfolding of proteins and studies of conformational dynamics under applied forces,[6, 7] to binding/unbinding of peptides.[62, 63] MSM-based studies allow for the convenient combination of several MD trajectories in to a single model of the underlying network of kinetic transitions between Markovian states from which experimental observables can be estimated, often to a high degree of accuracy. [54, 56-58, 60, 61]

Master equation and rate matrix:

Assuming ideal and Markovian states, a master equation accurately describes the rate of change of the population of the states in terms of fluxes in and out of the states. In a real system the states are highly dependent on the reaction coordinate chosen and on the lag time used to observe the system, and as such the system, for very short lag times, cannot be seen as Markovian. However, beyond certain lag times, most complex systems become “memoryless” and therefore the Markovian assumption holds, and the master equation can accurately represent the kinetics of the system. Say the system has *N* states and the probability of a particular state *m* is given by p_m. The rate constant, k_{mn}, is the rate of transition from state *m* to state *n* and thus the master equation of the system is given by:

$$\frac{dp_m(t)}{dt} = \sum_{n=0}^{N-1} [k_{mn} p_n(t) - k_{nm} p_m(t)]$$

This can be written in simpler terms by using matrix notation:

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{K}(t)\mathbf{p}(t) \qquad (1)$$

where **p**(t) is the probability matrix and **K**(t) is the $N \times N$ rate matrix of the system.

At equilibrium the rate of change of the population is zero and therefore the equilibrium population, $\mathbf{p}^o$, can be defined as:

$$\mathbf{K}\mathbf{p}^o = 0 \tag{2}$$

which is normalized according to the relation $\sum_{n=0}^{N-1} p_n^o = 1$ and positive $p_n^o > 0, n = \{0,1,...,N-1\}$.

The components of the rate matrix **K** are given by:

$$k_{nm} = \begin{cases} -\sum_{i=0\,(i\neq n)}^{N-1} k_{in}, & n = m \\ k_{mn} p_n^o / p_m^o, & n > m \end{cases}.$$

This allows detailed balance to be maintained, which is a requirement for a rate matrix when the system is at equilibrium

$$k_{nm} p_m^o = k_{mn} p_n^o .$$

Interestingly, one can express the likelihood Λ of a system of *N* states, being represented by the rate matrix **K**. Propagators, or the corresponding Green's functions $G(n, \Delta t \mid m, 0)$ are the probability of being in state *n* at time Δt after having been in state *m* at time 0. These propagators are weighted by the number of transitions from *m* to *n* observed in a trajectory, during an interval of Δt, and cumulating the observed numbers of these transitions in a transition matrix $T_{nm}(\Delta t)$ such that

$$\Lambda = \prod_{n=0}^{N-1} \prod_{m=1}^{N_{int}} [G(n, \Delta t | m, 0)]^{T_{nm}(\Delta t)} \tag{3}$$

where N_{int} is the number of time intervals of equal time length Δt such that the total length of the trajectory is $t_{total} = N_{\text{int}} \Delta t$ [54]. Equation 3 makes the connection between observed transitions and the likelihood of these transitions. This allows the extraction of the rate matrix of the system by maximizing the likelihood, for example by

using the method of simulated annealing of the free parameters of the rate matrix until convergence upon the maximum likelihood. In practice, the minimum of the log-likelihood is determined.

Eigen-spectrum properties of rate matrices:

The eigenvectors of the rate matrix **K** satisfy the following eigenvalue equations

$$\mathbf{K}\varphi_i = \lambda_i\varphi_i, \text{ and } \chi_i\mathbf{K} = \lambda_i\chi_i$$

where φ_i and χ_i are the left and right eigenvectors **K** respectively and the eigenvalues λ_i can be sorted such that $\lambda_0 = 0 > \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{N-1}$.

The symmetrized rate matrix, **H**, is given by

$$\mathbf{H} = \mathbf{P}_0^{-1/2}\mathbf{K}\mathbf{P}_0^{1/2}$$

where $\mathbf{P}_0 = diag\left[p_0^o, ..., p_{N-1}^o\right]$ is the equilibrium population matrix consisting of the equilibrium populations of each state and such that the trace is unity, i.e. $tr(\mathbf{P}_o) = 1$.

The components of **H** are given by

$$h_{nm} = \begin{cases} k_{nn} = -\sum_{i=0\,(i\neq n)}^{N-1} k_{in}, & n = m \\ \sqrt{k_{nm}k_{mn}}, & n \neq m \end{cases}.$$

The eigenvalues, λ_i, of **H** are the same as the eigenvalues of **K** and they satisfy the following equation

$$\mathbf{H}\psi_i = \lambda_i\psi_i$$

such that ψ_i are the orthonormal eigenvectors of **H**. We note that in many practical applications it is better to work with the symmetrize rate matrix **H** rather than with **K**, to avoid confusion, and for numerical accuracy. The eigenvectors of both matrices are connected analytically to each other and to the vector of equilibrium population of states, as described below.

Autocorrelation functions:

The autocorrelation function of a time-dependent observable ***a(t)***, which depends on the state, *s(t)* of the system at time *t* such that $s(t) \in \{0,1,...,N-1\}$, can be written using spectral decomposition

$$\langle \mathbf{a}(t)\mathbf{a}(0) \rangle = \sum_{i=0}^{N-1} \left[\sum_{i=0}^{N-1} a_n \psi_0(n) \psi_i(n) \right]^2 exp(\lambda_i t)$$

A useful identity can be derived if the case of $\mathbf{a}(t) = \psi_i(s(t)) / \psi_0(s(t)) = \chi_i(s(t))$ is calculated. This gives

$$\left\langle \frac{\psi_i(s(t))}{\psi_0(s(t))} . \frac{\psi_i(s(0))}{\psi_0(s(0))} \right\rangle = \exp(\lambda_i t) \tag{4}$$

Eq. (4) is used to validate the extracted master equation against actual simulation trajectories of Markovian dynamic systems.

Relation between symmetrized and non-symmetrized rate matrices:

The right and left eigenvectors of the original rate matrix **K** can be recovered from the eigenvectors of **H** in the manner below:

$$\varphi_i^2(n) = p_n^o . \psi_i^2(n),$$
$$\chi_i^2(n) = \psi_i^2(n) / p_n^o$$

and also $\varphi_i^2(n) = (p_n^o)^2 . \chi_i^2(n)$

The first right eigenvector, φ_0, of **K**, corresponding to the eigenvalue $\lambda_0 = 0$ is found to be the equilibrium population from Eq. (2). Therefore, $\varphi_0 = p_n^o$. From this it can be seen that $\chi_0(n) = 1$ for all $n = \{0,1,...,N-1\}$. Therefore Eq. (4) shows that

$$p_n^o = \psi_0^2(n), \quad \forall n \in \{0,1,...,N-1\}$$

Further relations can be extracted by using the orthonormality of the eigenvectors $\psi_i(n)$ of **H**.

$$\sum_{n=0}^{N-1}\psi_i(n)\psi_j(n)=\sum_{n=0}^{N-1}\chi_i(n)\chi_j(n)p_n^o=\sum_{n=0}^{N-1}\varphi_i(n)\varphi_j(n)\,/\,p_n^o=\sum_{n=0}^{N-1}\varphi_i(n)\chi_j(n)=\delta_{ij}$$

where δ_{ij} is the Kronecker delta. Filling in for $j=0$ results in:

$$\sum_{n=0}^{N-1}\psi_i(n)\sqrt{p_n^o}=\sum_{n=0}^{N-1}\chi_i(n)p_n^o=\sum_{n=0}^{N-1}\varphi_i(n)=\delta_{i0}$$

Deriving likelihood function:

We can write an analogous equation to Equation 1 for the symmetrized rate matrix **H** if the substitution $\pi=\mathbf{P}_0^{-1/2}\mathbf{p}$ is used.

$$\frac{d\pi(t)}{dt}=\mathbf{H}(t)\pi(t) \tag{5}$$

The solution of Eq. (5) is exponential along with the eigenvectors and eigenvalues of **H** and has the form

$$p_n(t)=p_n(0)\sum_{i=0}^{N-1}\psi_i^2(n)\exp(\lambda_i t)$$

The propagators of the system (i.e., the conditional probability of being in state *n* at time *t* given that the system was in state *m* at time $t_0=0$) are the Green's functions

$$G(n,t\mid m,0)=\left[e^{\mathbf{K}t}\right]_{nm}=\frac{\psi_0(n)}{\psi_0(m)}\sum_{i=0}^{N-1}\psi_i(n)\psi_i(m)\exp(\lambda_i t)\,.$$

The propagators of the system are used to form the likelihood function Λ of the Markovian trajectory such that

$$\Lambda=\prod_{i=1}^{N_{\text{int}}}G\left(s(i\Delta t),\Delta t\mid s((i-1)\Delta t),0\right)$$

where N_{int} is the number of time intervals of equal time length Δt such that the total length of the trajectory is $t_{total}=N_{\text{int}}\Delta t$. The likelihood function can be factorized into a product of products given the Markovian nature of the system leading to the Eq. (3).

In typical Markov state modelling, as illustrated in Figure 2.6, a set of either long or short[64] trajectories are used to sample the underlying free energy landscape. The trajectories are first processed to assign Markovian states, using a variety of possible approaches such as, for example, the transition-based assignment (TBA, see Ref. [65]). Secondly, the transition probabilities, or propagators described above are used to estimate either transition matrices at certain lag times or, the time-independent rate matrix **K** (e.g., by likelihood maximization using Eq. 3).[65]

Relative Root-Mean-Square Deviation (RMSD) for state assignment:

Here, we illustrate a simple yet effective method, the use of relative RMSD (RelRMSD or $RMSD^{rel}$) measure to assign the configurations to correct conformational states in biomolecular simulations. Let $RMSD_i$ be the RMSD value calculated along a trajectory after aligning it with the set of atomic coordinates for the molecular conformation *i*. Similarly, $RMSD_j$ is the RMSD value calculated along a trajectory after aligning it with the set of atomic coordinates for the molecular conformation j (see fig. 2.7a). Using the $RMSD_i$ and $RMSD_j$ values, we can define the time dependent relative RMSD ($RMSD^{rel}$).

$$RMSD_{ij}^{rel}(t) = \frac{RMSD_i(t)}{RMSD_i(t) + RMSD_j(t)}$$

The $RMSD^{rel}$ will always have values bound between 0 and 1. When the trajectory is close to conformation "*i*", the $RMSD^{rel}$ function takes positive values close to zero, and when the trajectory is close to the conformation "*j*", RelRMSD takes close to one but always less than one (see Fig. 2.7b). Combination of pairwise RelRMSD can be used in cases with more than two main conformations. We refer the readers to Ref. [66] for a more detailed description.

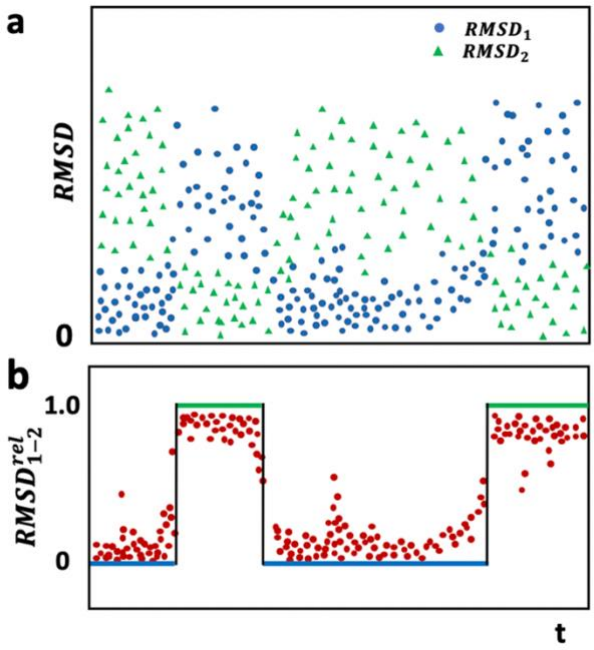


Figure 2.7. Using the relative RMSD (RelRMSD) reaction coordinate to assign trajectories. (a) $RMSD_1$ (blue, circles) and $RMSD_2$ (green, triangles) are two illustrative values obtained by aligning a trajectory to structures corresponding to representative conformations for state 1 and state 2, respectively. As expected, these are most informative only for RMSD values close to zero. (b) By combining the two signals in $RelRMSD_{12}$ (red, circles) the overall ability to discriminate between the two states is improved, and the assignment of states can be done more accurately (for example, by using the transition-based assignment method) resulting in a discretised trajectory (horizontal lines).

In Fig. 2.7 it is illustrated the use of the relative RMSD (RelRMSD) reaction coordinate to assign trajectories. Here, $RMSD_1$ (blue, circles) and $RMSD_2$ (green, triangles) are two illustrative values obtained by aligning a trajectory to structures corresponding to representative conformations for state 1 and state 2, respectively (Fig. 2.7a). As expected, these are only informative for RMSD values close to zero. In Fig. 2.7b it is shown how by combining the two signals in $RelRMSD_{12}$ (red, circles) the signal is improved, and the assignment of states can be done more accurately (for example, by using transition-based assignment) resulting in a discretised trajectory (lines).

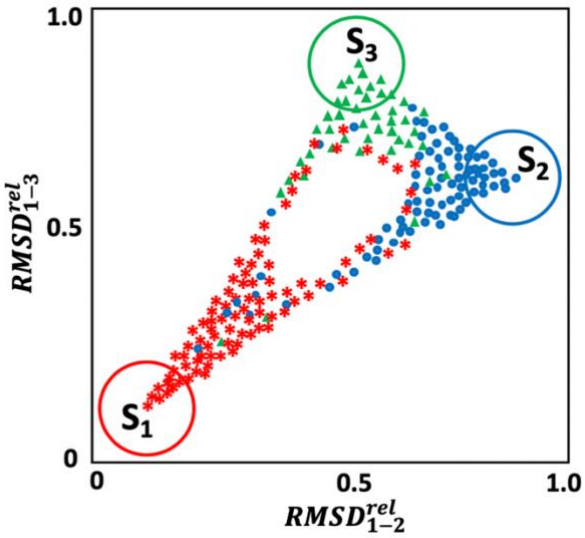


Figure 2.8. The RelMSD calculation can be used multi-dimensionally in the state assignment step for discriminating between multiple MSM states (in this case three conformational states). Structures assigned to S_1, S_2 and S_3 states are shown in red, blue and green, respectively.

Interestingly, the RelRMSD calculation can be used multi-dimensionally in the state assignment step for discriminating between multiple MSM states (in this case 3 conformational states), as illustrated in Figs. 2.8 and 2.9. Using the RelRMSD to assign states (e.g., in conjunction with the TBA method) is particularly useful in cases in which well-defined representative biomolecular conformations exist for the corresponding Markov states and can be used to make the thermodynamic and kinetic MSM analysis of the underlying conformational dynamics more automatic.

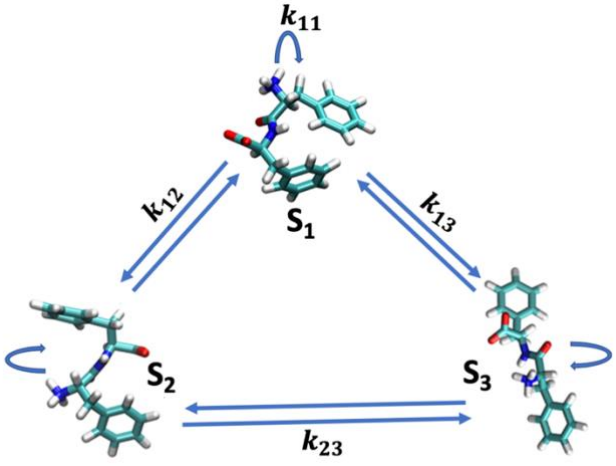


Figure 2.9. Example of using RelRMSD for MSM analysis of the conformational dynamics of a small diphenylalanine (FF) peptide (replica exchange MD simulation with explicit water molecules).

Figure 2.9 shows an example of using RelRMSD for MSM analysis of the conformational dynamics of a diphenylalanine peptide (replica exchange MD simulation with explicit water molecules), which is shown to be essentially 3-state in Ref. [66].

2.5 Milestoning

For systems where the sampling problem can be stated in terms of estimating the thermodynamics and kinetics along transition pathways that connect two known metastable states (e.g., “reactant”, R, and “product”, P, see Figures 2.6 and 2.7) one could target the sampling in order to identify and characterize the pathway(s) (typically with the maximum reaction flux) and the corresponding intermediate states that are relevant to the underlying reaction mechanism.

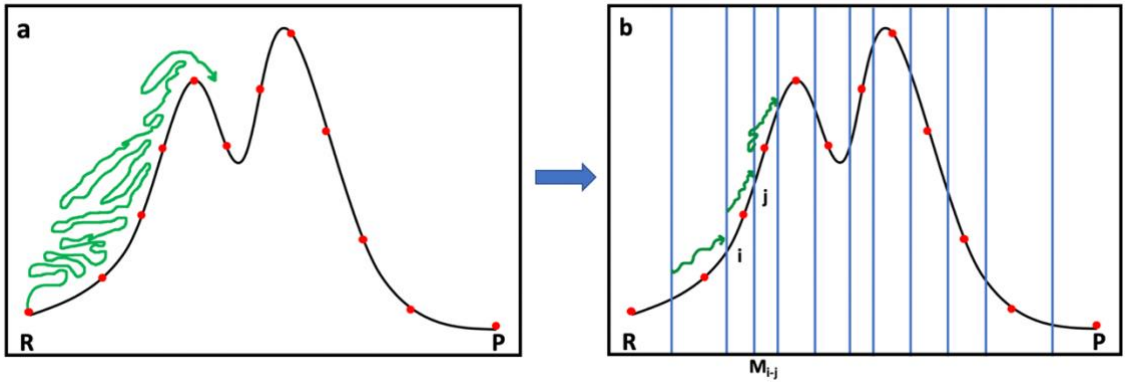


Figure 2.10. Schematic 1d representation of two typical approaches to sampling free energy barriers in molecular simulations using (a) one or a few typically long trajectories, and (b) short trajectories initialized along the transition pathway between two end points (e.g., reactants, R, and products, P). In (b), central in the implementation of the Milestoning method, the aim is to achieve higher computational efficiency by initializing the short trajectories from multi-dimensional hypersurfaces (vertical lines, blue) that are located between anchors (red dots) along a low dimensional reaction coordinate (see also Fig. 2.7).

In Milestoning, an initial pathway between R and P conformations can be used to first select a set of intermediate conformations along any candidate reaction path that may be available (e.g., from a high-temperature simulation, steered MD, etc.). The

intermediate representative conformations (Fig. 2.10, red points) are referred to as anchors. Milestones are further introduced as hypersurfaces in the free energy landscape that have an equal minimal distance from at least two anchors (blue lines in Figs. 2.10 and 2.11). Finally, once the energy landscape is divided along the reaction coordinate using milestones, one can run sets of short trajectories from each milestone (shown in Figs. 2.10 and 2.11). These short trajectories are terminated as soon as they reach a new milestone. The statistical information on local transitions between each pair of milestones (e.g., first passage times), can be combined to obtain the global statistical information on possible R-P transition pathways, free energy profiles and relative transition times.

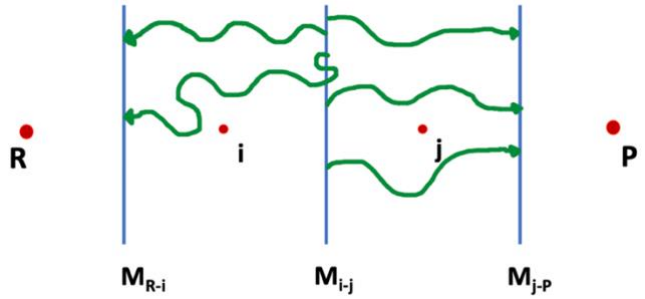


Figure 2.11. In Milestoning, the sampling is initiated by first defining a set of "anchors" (red) conformations that span a (typically low dimensional) transition pathway between two end points (e.g., reactants, R, and products, P). The underlying kinetic and thermodynamic information is extracted by analysing sets of short trajectories initiated in conformational space regions (blue, milestones, where M_{ij} is the milestone between anchors *i* and *j*, respectively). The short trajectories (green) are terminated as soon as they reach another milestone, different from the one at which they were initiated.

Here we briefly illustrate the Milestoning method. To initiate Milestoning, we need a set of sample configurations, that we call anchors, from the coarse space between R and P (red dots in Figs. 2.10 and 2.11). These anchors are the centres of Voronoi cells in space of several course variables. A milestone between two anchors, *i* and *j*, is thus a set of points with equal distance from anchors *i* and *j*. This distance is smaller than the separation from all other anchors. Milestone between anchors *i* and *j* is denoted by M_{ij} (shown in blue in Fig. 2.10) or M_{α}.

At this stage, one needs to sample configurations on each milestone (e.g., by using MD simulations constrained to corresponding milestone coordinates). Finally, we launch unbiased MD trajectories from the sampled configurations on each milestone. Distance from all the anchors is measures along these unbiased trajectories. If the

closest anchor is other than the two anchors used to define the milestone from which the unbiased trajectory was initiated, then we know that trajectory hit a new milestone. These trajectories are terminated. We record the starting and terminating milestone index and the time of termination. Using this information, we can estimate transition probability between milestones, also called kernel, **W**, and the corresponding lifetime of the milestone. Let n_α be the total number of trajectories initiated at the milestone M_α. Let $n_{\alpha\beta}$ be the number of trajectories which started at milestone M_α and terminated at milestone M_β. The transition probability, $W_{\alpha\beta}$, is estimated as $n_{\alpha\beta}$ / n_α. The lifetime of the milestone, M_α, is $t_\alpha = \frac{1}{n_\alpha}\sum_{l=1,\ldots,n_\alpha} t_l$, where *l* is the index of the trajectory and t_l is the time length of the trajectory *l.*

Using the transition probability or kernel matrix, **W**, and the milestone lifetime, t, we can compute important kinetic and thermodynamic observables such as the flux through a milestone, free energy, mean first passage time, etc. [67] The stationary flux of trajectories through a milestone is denoted by the eigenvector, **q**, of the matrix **W**, with an eigenvalue of one. It can be shown that the free energy of a milestone a is $F_a = -k_B T \log\left[q_a t_a\right]$ and the mean first passage time (MFPT, $\langle\tau\rangle$) can be written as $\langle\tau\rangle = \mathbf{p_0}(\mathbf{I} - W)^{-1}\mathbf{t}$, where $\mathbf{p}_0$ is the vector of the initial distribution and **I** is the identity matrix.[67]

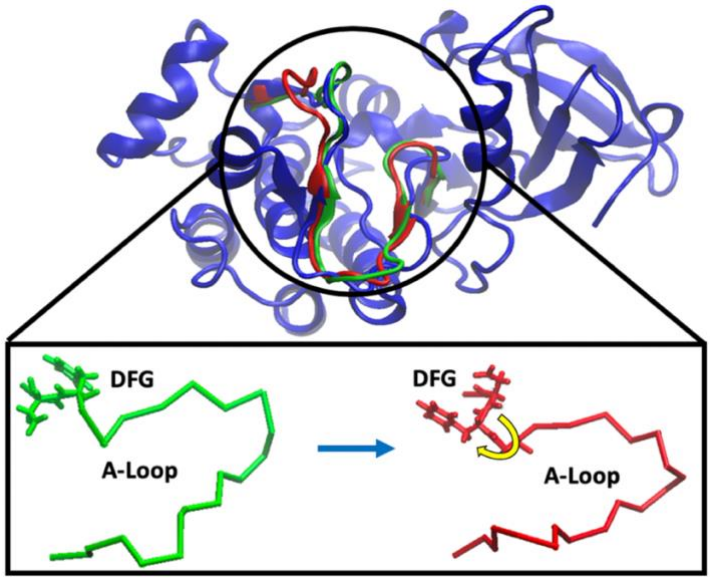


Figure 2.12. Example of using a multi-dimensional Milestoning approach in a large and more complex molecular system: the characterization of the DFG-flip dynamics in Abl kinase.

Figure 2.12 illustrates the use of a multi-dimensional Milestoning approach in a large and more complex molecular system: the characterization of the DFG-flip dynamics in Abl kinase.[18] Here the Milestoning method to study the transition between active and inactive conformations of Abl kinase (shown in Fig. 2.12). Note that it would have been impractical to study a complex system like this with regular MD simulations using comparable HPC resources.[18, 20]

2.6 Conclusions

In spite of significant and sustained advances in computational hardware allowing for larger size and longer scale computations, molecular dynamics studies of a large majority of biomolecular systems remain outside the reach of traditional simulation methods. A main problem springs from the intrinsically high-dimensional and complex nature of the underlying free energy landscape of most systems, and the necessity to sample accurately such landscapes for identifying kinetic and thermodynamic states in the configurations space, and for accurate calculations of both free energy differences and the corresponding transition rates between states. Here, we reviewed two modern methods that allow longer-time MD studies of biomolecular systems that can open a broad spectrum of applications. A first approach Markov State Models, relies on identifying a set of configuration states in which the system resides sufficiently long to relax and loose the memory of previous transitions, and on using simulations for mapping the underlying complex energy landscape on a network of Markovian transitions. The independence of the underlying transition probabilities creates the opportunity to increase the sampling efficiency by using sets of appropriately initialized sets of short simulations rather than more typical long MD trajectories, which leads to both enhanced sampling and higher accuracy. This allows MSM studies to unveil bio-molecular mechanisms and to estimate free energy barriers with high accuracy, in a manner that is both systematic and relatively automatic, which accounts for their increasing popularity. The second approach, Milestoning, is focused on accurate studies of the ensemble of pathways connecting two specific end-states (e.g., reactants and products) in a similarly systematic and highly automatic and highly accurate manner. Conceptually, both methods are theoretically identical for transition paths between Markovian states, however Milestoning can be generalized and applied

to studies of non-Markovian transitions as well. It has been shown in ref. [68] that in general milestoning procedure the dynamics is not of a continuous time Markov chain. Despite the basic idea in MSM and the milestoning method being same, there are other two main differences. First difference is in the definition of states. In MSM, states form a partition of phase space, whereas in the Milestoning states are hypersurfaces in the phase space. In MSM, one needs to select a lag time at which transition probability matrix should to be computed. It is not very obvious at which lag time the results of MSM will be accurate. In milestoning, provided that optimal milestones are being used, exact mean first passage time can be obtained. More detailed comparison can be found in ref. [68-71] As highlighted by the increasing number of studies using both methods, we anticipate that they will open new avenues for the investigation of systematic sampling of reactions pathways and mechanisms occurring on longer time scales than currently accessible by purely computational hardware and parallelization-related advances.

3. Replica Exchange Molecular Dynamics of FF Amyloid Peptides in Electric Fields[2]

3.1 Overview

Here, I study the temperature-dependent conformational dynamics of FF peptides solvated in explicit water molecules, an environment relevant to biomedical applications, by using an enhanced sampling method, replica exchange molecular dynamics, in conjunction with applied electric fields. Simulations highlight and overcome possible artifacts that may occur during the setup of REMD simulations of explicitly solvated peptides in the presence of external electric fields, a problem particularly important in the case of short peptides such as FF. The presence of the external fields could over-stabilize certain conformational states in one or more REMD replicas, leading to distortions of the underlying potential energy distributions observed at each temperature. This can be overcome by correcting the REMD initial conditions to include the lower energy conformations induced by the external field.

3.2 Introduction

Small, biocompatible peptides, such as amyloid-forming diphenylalanine (FF) have raised an increasing interest in both theoretical[66, 72-75] and experimental[76-80] nanoscience studies for almost two decades. This success is due both to their intrinsic propensity to self-assemble in a hierarchic manner from FF monomers into diverse nanostructures, and to the interesting emerging biophysical properties of these nanostructures (e.g., piezoelectric, optical and mechanical strength properties).[77, 78, 81] FF is one of the smallest, naturally occurring amyloid peptides, found commonly in the hydrophobic structural core the amyloid beta (Aβ) protein, which

2. This chapter has been adapted from reference [21].

allowed its identification as one of the smallest peptides capable of self-assembly leading to the formation of ordered fibrillar amyloid nanostructures.[81]

Amyloid FF peptides and their bioinspired nano-scale structures such as FF and nanotubes, nanospheres, or even nanorods[82] have led to a multitude of applications in biomedicine, nanoscience and nanotechnology.[75, 77, 78, 83] However, there are also significant limitations to using FF-based nanomaterials, one of the main factors being the instability of FF nanotubes in solution (e.g., a major limitation hindering the development of FF nanotube-based biosensors or drug delivery systems) and the relative heterogeneity of the local, nm-scale structures formed by self-assembly of the FF peptides under various conditions, including but not limited to temperature, pH and solvation.[79, 80] To overcome such barriers it becomes important to understand and control the peptide self-assembly process. Innovative approaches such as directed self-assembly have been developed, such as subjecting a system under the influence of an externally applied stimuli, including mechanical mixing, temperature or pH variations. Thus, different degrees of control are achieved by enabling the tuning of desired interactions, structure and properties of the final self-assembled nanomaterials. Recent experimental methods of directed self-assembly such as di-electrophoresis, rely on applying an external electric field on the entire ensemble of assembling peptides and have been used to modulate the alignment of FF nanotubes.[84-88] However, a main challenge with directed self-assembly remains the need for predictive models that bridge the detailed conformational behaviour of a single FF molecule under an electric field and the properties of the resulting nanoscale self-assembled structures.

In this study, we use atomistic molecular dynamics simulations to study the combined effect of applied electric fields and temperature dependence on the detailed conformational dynamics of FF peptides solvated in explicit water molecules, an environment relevant to biomedical applications. In order to capture the temperature effect on the FF thermodynamics and kinetic properties, our simulations rely on an enhanced sampling method, temperature replica exchange molecular dynamics. Here, we first highlight and overcome a possible problem that may lead to artifacts during the setup of REMD simulations of explicitly solvated peptides in the presence of external electric fields, a problem particularly important in the case of short peptides such as

FF. We show how to overcome this problem (i.e., by correcting the REMD initial conditions to include the lower energy conformations induced by the external field, and we analyse the converged REMD data using a Markovian description of conformational states of the simulated system. Finally, we discuss the observed temperature, and electric field-dependent thermodynamic and kinetic properties of small FF amyloid peptides, which may be useful in understanding and devising new methods to control their aggregation-prone biophysical properties and, possibly, the structural and biophysical properties of FF molecular nanostructures.

3.3 Methods

REMD Simulations in external electric fields

We use atomistic REMD simulations of FF peptides, following a similar procedure to our previous study described in Ref. [66] (though, in that case we did not use external electric fields), with the MD package Gromacs (version 5.1.4),[89, 90] using Langevin dynamics with a friction coefficient of 0.1 ps^{-1}.[91] These REMD simulations used the particle-mesh Ewald implementation with a switching distance for the van der Waals interactions and nonbonded electrostatics of 8.5 Å and a cut-off distance of 12 Å, and an integration time step of 2 fs. The runs were performed in the NPT ensemble, using an improved Berendsen-type weak coupling method for temperature coupling,[92] Parrinello-Rahman isotropic pressure coupling,[93] the recent CHARMM[25] 36 all-atom protein force field parameters (C36),[94] and using explicit TIP3P[95] water molecules. The FF peptide was included in a simulation box containing 1112 water molecules. To enhance the sampling, REMD is performed with 12 replicas running in parallel at temperature values chosen according to an optimized protocol[3] (Table 3.1) in the range of 310.00 K to 373.45 K.[63]

For the REMD simulations, we prepared the system including the FF amyloid peptide and water molecules using VMD's[1] Molefacture Plugin protein builder tool, followed by minimization, heating and equilibration stages, at each electric field value. The system was simulated using the Gromacs REMD implementation,[3] with an average acceptance probability for the replica exchanges of ~20%. The atomic velocities and coordinates were saved every 100 fs and, after simulation, the REMD

per-replica trajectory data (i.e., referred to as R-trajectories) was also transformed for analysis into per-temperature data (i.e., referred to as T-trajectories) using the Gormacs *demux* command. The Gromacs *trajconv* command was used to select system conformations only every 1 ps (i.e., every 500th MD frame, with a 2 fs integration timestep) for our detailed thermodynamic and kinetic analysis.

E = 0 kcal/mol·Å·e												
Replica no	1	2	3	4	5	6	7	8	9	10	11	12
Temp [K]	310.00	315.38	320.82	326.35	331.96	337.64	343.41	349.26	355.19	361.20	367.30	373.45
Time [ns]	126	126	126	126	126	126	126	126	126	126	126	126
E = 30 kcal/mol·Å·e												
Replica no	1	2	3	4	5	6	7	8	9	10	11	12
Temp [K]	310.00	315.38	320.82	326.35	331.96	337.64	343.41	349.26	355.19	361.20	367.30	373.45
Time [ns]	100	100	100	100	100	100	100	100	100	100	100	100
E = 45 kcal/mol·Å·e												
Replica no	1	2	3	4	5	6	7	8	9	10	11	12
Temp [K]	310.00	315.38	320.82	326.35	331.96	337.64	343.41	349.26	355.19	361.20	367.30	373.45
Time [ns]	98	98	98	98	98	98	98	98	98	98	98	98

Table 3.1. The three main sets of REMD data analysed here correspond to simulations with explicit water molecules, performed in the presence of external electric fields with intensities of (a) 0, (b) 30, and (c) 45 kcal/mol·Å·e, respectively. Each run used 12 replicas, at temperatures spaced according to an optimized protocol,[3] as indicated in the table together with the corresponding run times. The total simulation time is ~4 ms (i.e., also including the initial setup and testing runs at 30 kcal/mol·Å·e)

For the first run, in the absence of electric fields, the production simulations were done for 126 ns for each of the 12 replicas, giving a total REMD simulation time of 1.512 μs, which was sufficient for achieving convergence of all the relevant thermodynamic and kinetic quantities. As an additional test for convergence we also checked the “equal occupancy rule” of replicas at each temperature,[96] which is a very useful method for assessing quickly the performance of parallel tempering simulations.[52, 96] Subsequently, kinetic data on the identified conformational Markov states and the corresponding transition probabilities was calculated from the REMD trajectories as discussed below.

Extracting transition probabilities and rates from REMD data

To identify and test the Markovian conformational states and the corresponding transition states for REMD trajectories, we analysed the temperature-dependent FF data by following the workflow that we developed and presented in our previous study,[66] and the corresponding transition probabilities were extracted and compared. Relative RMSD was extracted using representative structures for the three peaks of d_{ee} distribution. Using TBA method, states were assigned and then transition between these states were counted to extract transition probability matrix (see Ref 59). As highlighted in Ref. [63], the replica R-trajectories are continuous, even though they travel at various temperatures during the REMD as exchange attempts are accepted (e.g., Fig. 1 in Ref. [63]), while the data captured as T-trajectories is actually discontinue, being interrupted at time steps when exchange attempts are accepted. Note that, unlike other REMD analysis methods that are focused on T-trajectories, due to their well-defined temperatures, we showed that it is convenient to start by analysing R-trajectories in order to take advantage by their time-continuity both in the initial assignment of states, and, importantly, in assessing convergence.[62, 63] As demonstrated in Ref. [63], there is an analytical relation that connects both R-trajectories and T-trajectories. The propagators (i.e., conditional probabilities) for transitions along R-trajectories were shown to be in effect the weighted geometric means of propagators extracted for the corresponding transitions in T-trajectories. This observation enables a powerful direct application of kinetic analysis along R-trajectories, on which state assignment is easier due to their continuous nature, rather than performing directly a more laborious (and thus more prone to errors) kinetic analysis of the discontinuous T-trajectories.[62, 63]

Following the procedure detailed for REMD data of FF peptides in Ref. [62], here we assume that the conformational space of a system can be discretized into *N* distinct states that obey a master equation, which can be expressed in matrix notation as $\frac{d\mathbf{p}(t)}{dt} = \mathbf{K}(t)\mathbf{p}(t)$, with $\mathbf{p}(t)$ being the time dependent column vector of probabilities with elements such that $p_n(t) > 0, n \in \{1, \ldots, N\}$. Here, **K**(*t*) is the $N \times N$ rate matrix, the **K** element k_{nm} is the rate of transition from state *m* to state *n*, and p_m is the probability of the state labelled *m,* at time *t*.[7, 54, 64, 97-105] At thermodynamic and kinetic

equilibrium, we have $\mathbf{Kp}^{o} \equiv 0$, with $\mathbf{p}^{o}$ being thus the vector of equilibrium populations that has positive elements, $p_n^o(t) > 0, n \in \{1, \ldots, N\}$, and it is properly normalized ($\sum_{n=1}^{N} p_n^o = 1$).Therefore, $\mathbf{p}^{o}$ appears as the first right eigenvector of **K**, corresponding to the first eigenvalue $\lambda_1 = 0$.

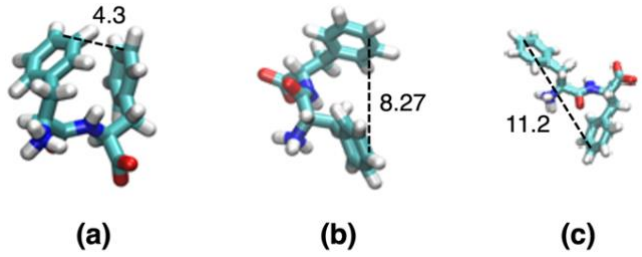


Figure 3.1. Representative conformations of FF peptides in the absence of externally applied electric fields. Values of the d_{ee} distances (i.e., distances between the CZ atoms at the ends of the two sidechains, in Å), are shown in black.

Similarly to previous studies,[62, 66] we use the Markov-based DTC method[62, 106] for extracting transition rates from REMD trajectories (in this case, for different values of an externally applied electric field), which requires the initial assignment of conformational states of the system. The conformational states of the peptide are assigned by following each replica using both T-trajectories and R-trajectories, using the transition based assignment (TBA) method described and used in previous studies.[54, 57, 63] We use the TBA method of assignment of Markov states for biomolecular MD trajectories introduced in Ref. [54], and reviewed in detail subsequently in Ref. [57]. The TBA method requires initially a reasonable choice of reaction coordinates that allow a good discrimination between the different conformational Markov states. However, though these reaction coordinates need to be reasonably good, the subsequent state assignment step does not depend entirely on their absolute quality, as the TBA method also uses additional, more specific information from analyzing the actual transition paths (i.e., time sequence of transition events) to the state assignment process.[57] As described next here, we use the d_{ee} distances (i.e., distances between the CZ atoms at the ends of the two sidechains, in Å), illustrated in Fig. 3.1 in black, as a useful choice for initiating the TBA analysis step.

Tests of REMD convergence

Following previous studies,[66] we tested initially the REMD data convergence by investigating the “equal occupancy” rule of replicas (if converged then a trajectory would spend equal time at each replica) at each temperature,[96] which is a fast and useful to assess the performance of parallel tempering simulations.[52, 96] Additionally, we have also analyzed and compared data from both R- and T-trajectories to show that our extracted quantities are converged (e.g., as shown in the probability distributions of different relevant observables illustrated in Figs. 3.5 to 3.8). As discussed in Ref. [66], it is important to note that, in cases with several Markovian states present (here, for the FF dynamics in the absence of electric fields), the transition probabilities extracted from REMD data after applying the TBA method to project the R- and T-trajectories to states and performing the kinetic analysis can also serve as the ”ultimate” test of the convergence of the REMD simulations performed. In practice, we can also use subset of REMD data to estimate statistical errors for the extracted transition probabilities, as errors of the means for each data set. The analysis of the errors in the extracted intrinsic parameters, like, transition probabilities, of the Markovian kinetics offers a reliable assessment of the convergence of the data in the MD trajectories generated [66]. Figure 3.1. Representative conformations of FF peptides in the absence of externally applied electric fields. Values of the d_{ee} distances (i.e., distances between the CZ atoms at the ends of the two sidechains, in Å), are shown in black.

3.4 Results

We generate and use new data from REMD simulations performed in the presence of external electric fields to probe the combined Temperature and E-field dependent conformational dynamics of FF peptides (Fig. 3.1).[66, 74] However, while the REMD simulations without electric fields were rather straight forward, the presence of the external field allowed us to unveil interesting artifacts. These that may occur in general during the setup of any REMD simulations of explicitly solvated peptides in the presence of external electric fields, though they have a particularly high likelihood in the case of short peptides such as FF. In this case, the presence of the external fields can induce rapidly (i.e., on the order of tens of picoseconds) and over-stabilize (i.e., as compared to conformational dynamics in the absence of external fields) a low-energy

conformational state in one or more REMD replicas, leading to distortions of the underlying potential energy distributions observed at each temperature.

The issue is illustrated in Fig. 3.2 that shows the potential energy distributions for our initial replica exchange FF simulation with explicit water molecules in an electric field of intensity E = 30 kcal/mol·Å·e. The non-Gaussian shape of the potential energy distribution is evidenced for the first two lowest temperatures as shown in Fig. 3.2a, while the induced transitions to field-stabilized low-energy conformations is illustrated schematically in Fig. 3.2b. The REMD implementation in most software packages, and the underlying replica exchange attempts, are designed to preserve detailed balance when sampling from canonical distributions. Thus, the REMD exchange protocol relies on accurate dynamics that preservers the Gaussian shape of the underlying potential energy distributions. Parameters of the REMD simulations such as the number of replicas and the exact values of the temperatures selected depend directly on the correct shape of the underlying energy distributions and on their overlap (e.g., which controls the acceptance/rejection exchange probabilities for a simulation of a system with a certain number of atoms and the corresponding thermodynamic conditions). Thus, replica potential energy distributions with non-Gaussian shapes due to, in this case, the presence of external fields can easily lead to serious artifacts. We note that this cause is different from REMD artifacts due to modified underlying energy distributions, for example, due to the use of weak-coupling thermostats which were highlighted before [107]. Earlier studies have shown that REMD simulations of other small peptides, such as dialanine [107] and pentaalanine [108] using explicit TIP3P water molecules,[95] the choice of weak-coupling thermostats can significantly affect the outcome of REMD simulations, though in that case through a narrowing of the underlying potential energy distribution for each replica.

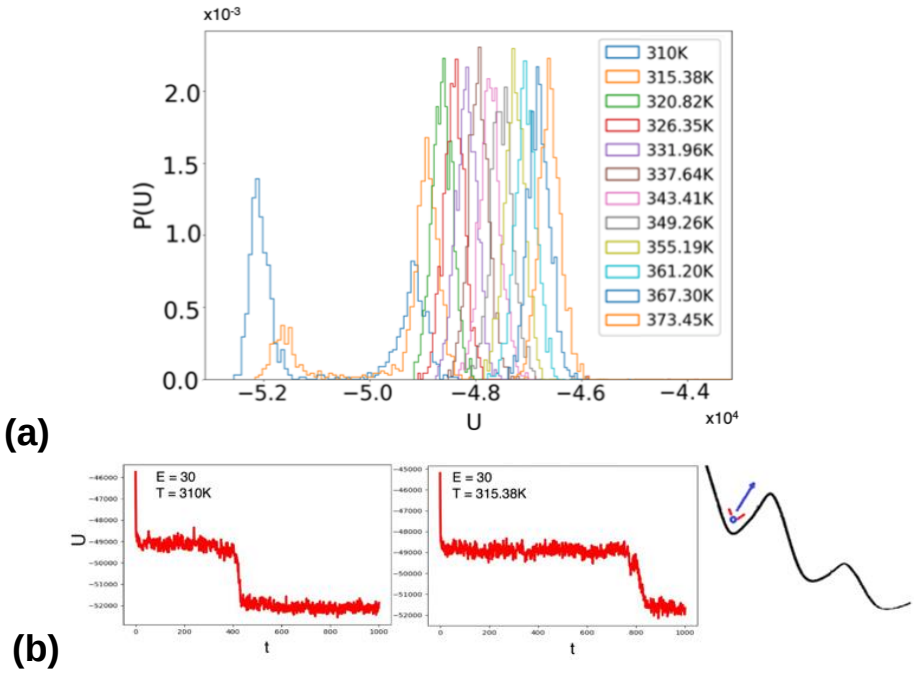


Figure 3.2. (a) Distributions of potential energy values (*U*, in kcal/mol) calculated from REMD simulations in the presence of an external electric field with an intensity of *E* = 30 kcal/mol·Å·e. **(b)** Illustration of the problems that could occur when attempting REMD simulations in external electric fields. The presence of the field can induce some (in this case the first two) replicas to adopt conformations that are significantly lower in energy than the corresponding initial conformational states of the other replicas. This is a serious artifact, as illustrated in (a), as it changes the expected equilibrium *U* distributions.

The artifacts due to external electric fields can be overcome, as demonstrated here, by correcting the REMD initial conditions to include the lower energy conformations induced by the external field for all replicas. When transitioning from properly equilibrated initial conditions to simulations when an additional field is present, it is thus crucial to not only re-equilibrate but also monitor the underlying energy distributions for all replicas (see Fig. 3.3), at all temperatures, and re-initialize the REMD protocol to include the lower energy conformations that may be induced. Subsequently the REMD protocol can proceed to achieve enhanced sampling by use of replicas running at higher temperatures, in parallel, while preserving the correct underlying dynamic and thermodynamic behavior of the system at all temperatures. Fig. 3.3a shows the corrected REMD distributions of potential energy values (*U*, in kcal/mol) calculated from REMD simulations at *E* = 0 kcal/mol·Å·e (Fig. 3.3a), and also with the new, corrected initial conditions in the presence of an external electric field with an intensity of *E* = 30 kcal/mol·Å·e (Fig. 3.3b).

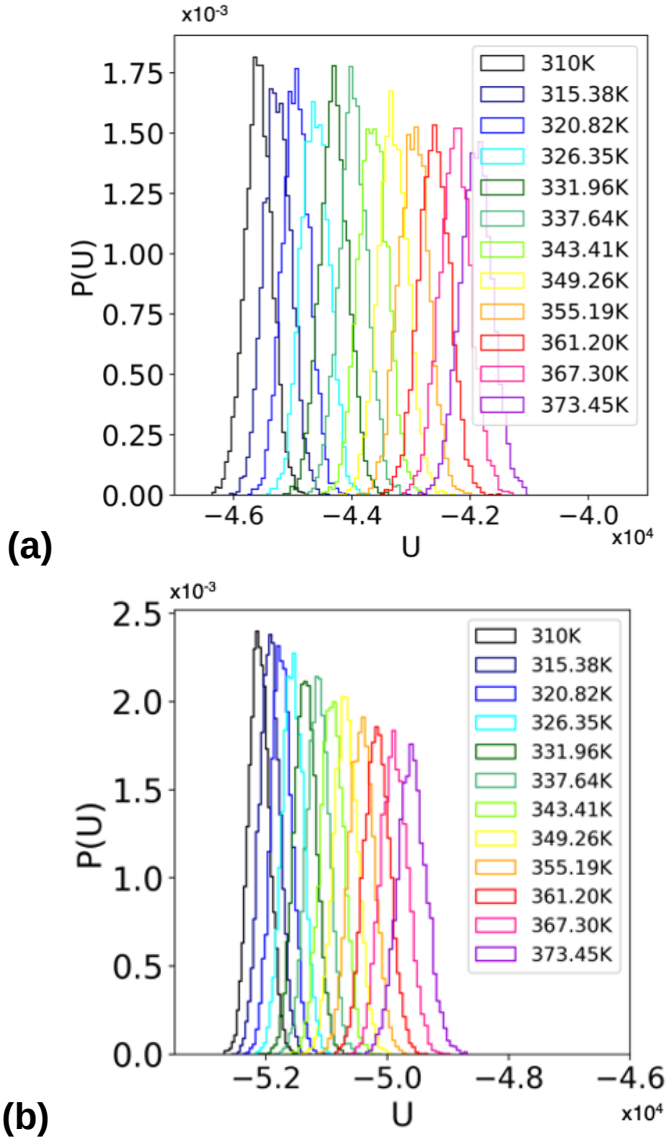


Figure 3.3. Distributions of potential energy values (*U*, in kcal/mol) calculated from REMD simulations (a) at *E* = 0 kcal/mol·Å·e, and (b) with corrected initial conditions in the presence of an external electric field with an intensity of *E* = 30 kcal/mol·Å·e.

Figure 3.4 shows the distributions of root-mean-square deviation of atomic positions (RMSD) values calculated for the heavy atoms of FF peptides for conformations from REMD simulations, with respect to initial configuration, in the presence of external electric fields with intensities in the three cases studied here and detailed in Table 1: *E* = 0, 30, and 45 kcal/mol·Å·e, respectively. These distributions shows clearly that the complexity of the conformational dynamics of the FF amyloid peptides is dramatically reduced in the presence of external fields, in agreement with earlier studies that, however used much reduced sampling in simple MD simulations.[74]

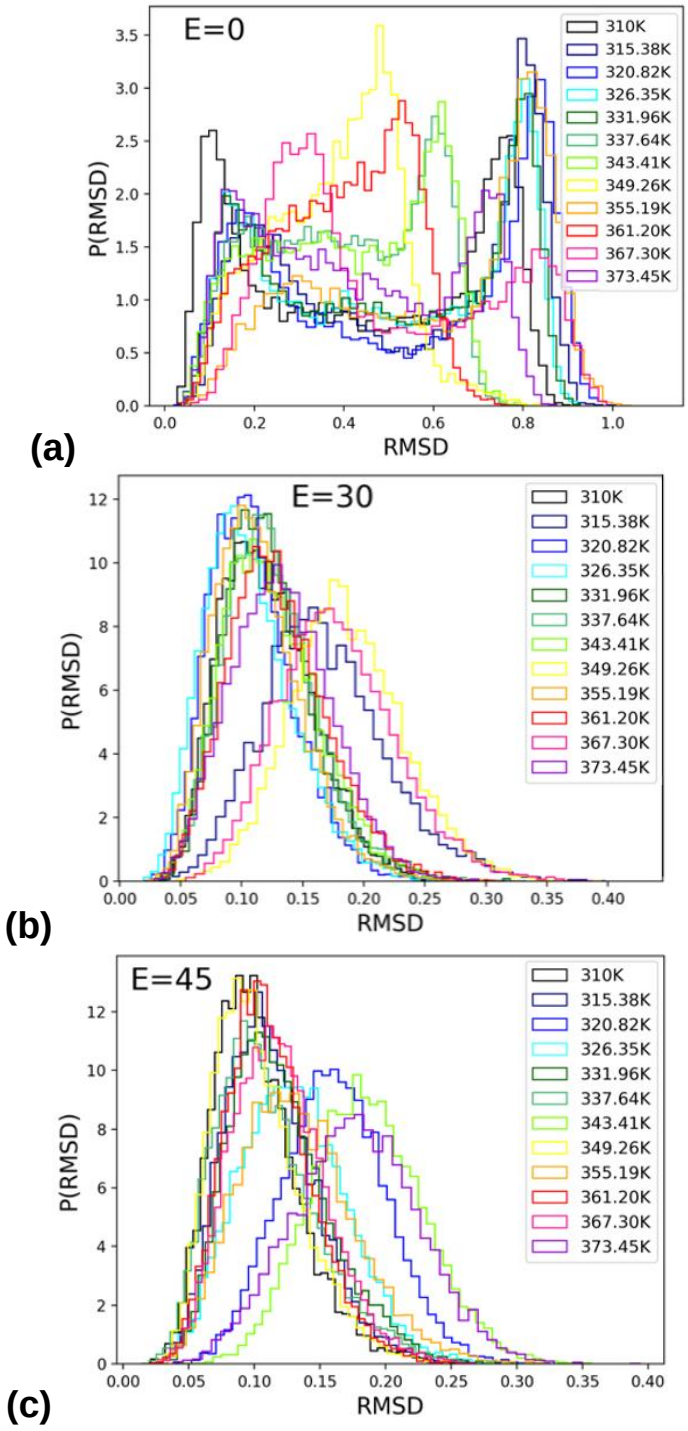


Figure 3.4. Distributions of RMSD values calculated for the heavy atoms of FF peptides for conformations from REMD simulations in the presence of external electric fields with intensities of (a) *E* = 0 kcal/mol·Å·e, (b) *E* = 30 kcal/mol·Å·e, and (c) *E* = 45 kcal/mol·Å·e.

In relation to the piezoelectric behaviour of FF amyloid peptides, in Figure 3.5 are shown the distributions of the dipole moment magnitude (*m*, Debye units), calculated for FF peptides for conformations from our three sets of REMD simulations

in the presence of different external electric. In agreement with earlier observations, there is a noticeable effect on the magnitude of the dipole moment which increases systematically with larger *E* values, showing less complexity and fluctuations at all temperatures, as the peptide adopts more extended conformations.

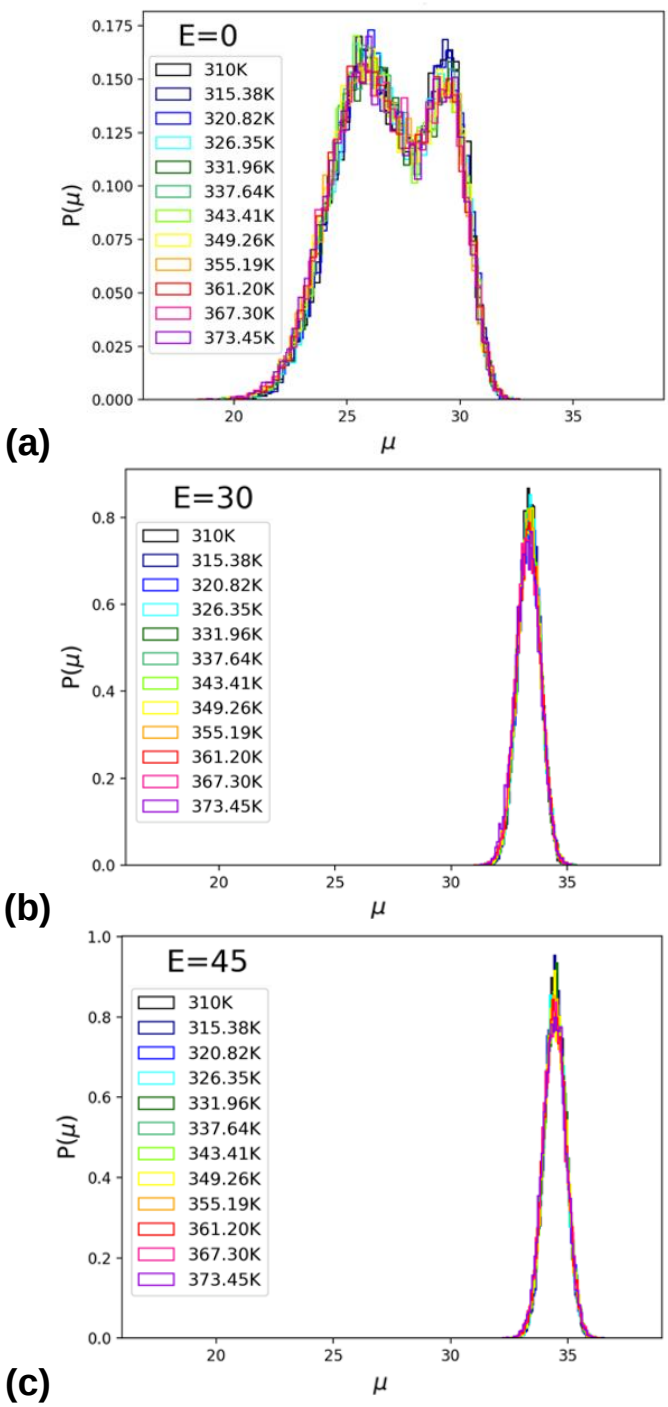


Figure 3.5. Distributions of the dipole moment magnitude (*m*, Debye units), calculated for FF peptides for conformations from REMD simulations in the presence of external electric fields with intensities of (a) *E* = 0 kcal/mol·Å·e, (b) *E* = 30 kcal/mol·Å·e, and (c) *E* = 45 kcal/mol·Å·e.

However, while both RMSD and the dipole moment magnitude are useful collective variables utilized in MD analysis of FF peptides, Figs. 3.4 and 3.5 also illustrate their intrinsic limitations in allowing us to identify and discuss the detailed dynamics. Thus, here we choose to focus on a different measure, the d_{ee} distances (i.e., distances between the CZ atoms at the ends of the two sidechains, in Å, shown in Fig. 3.1 in black) as a useful choice for our more detailed kinetic and thermodynamic analysis. Fig. 3.6 shows REMD equilibrium distributions of d_{ee} values for FF amyloid peptides, in the case where no external electric field is applied, for each replica (R-trajectories, Fig. 3.6a), and at each temperature (T-trajectories, Fig. 3.6b) of the REMD trajectories. We note the clear presence of three conformational peaks.

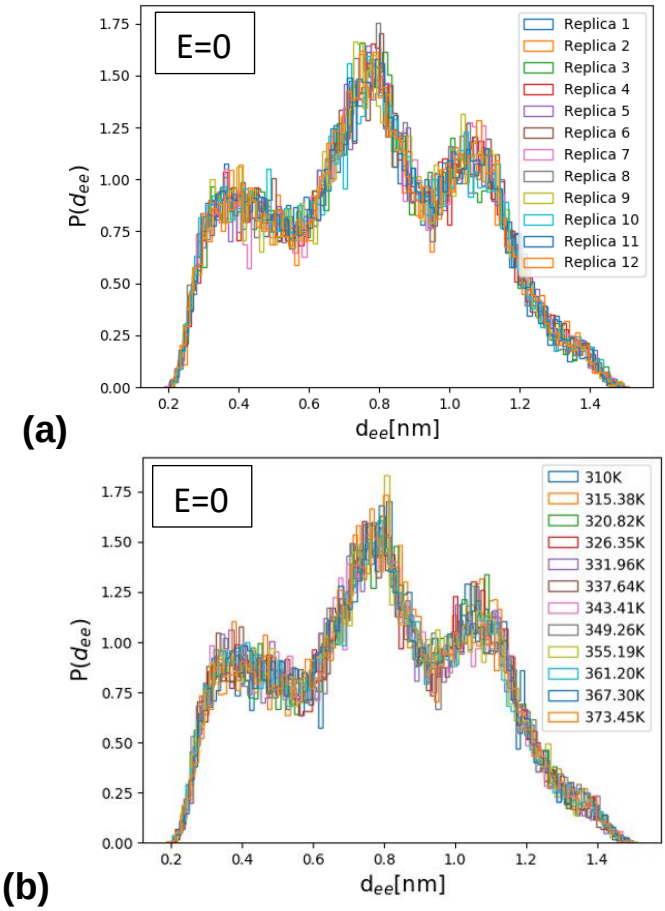


Figure 3.6. Replica exchange equilibrium distributions of sidechain-sidechain distances of FF amyloid peptides, with no external electric field applied, (a) for each replica (R-trajectories), and (b) at each temperature (T-trajectories) of the REMD simulation set.

The corresponding distributions of sidechain-to-sidechain distances for simulations with an applied electric field of 30 kcal/mol·Å·e, are shown in Fig. 3.7 for

each replica (R-trajectories, Fig. 3.7a), and at each temperature (T-trajectories, Fig 3.7b) of the REMD simulation set. We note that, at a field intensity of 30 kcal/mol·Å·e, the conformational dynamics is restricted to one extended structure with a most probable d_{ee} value of ~8.9 Å.

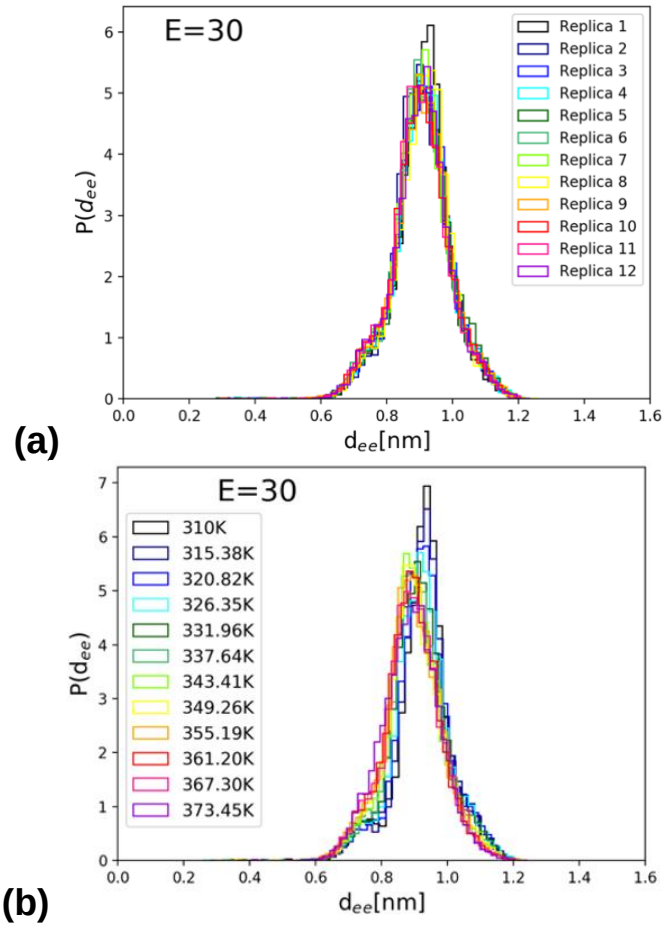


Figure 3.7. Distributions of sidechain-to-sidechain distances, d_{ee}, for simulations with an applied electric field of 30 kcal/mol·Å·e, (a) for each replica (R-trajectories), and (b) at each temperature (T-trajectories) of the REMD simulation set. Note that, at this field intensity, the conformational dynamics is restricted to one extended structure with a most probable d_{ee} value of ~8.9 Å.

Finally, in Fig. 3.8 are shown the measured distributions of d_{ee} values for simulations with an applied electric field of 45 kcal/mol·Å·e, for each replica (R-trajectories, Fig. 3.8a) and, once again, at each temperature (T-trajectories, Fig. 3.8b) of the REMD simulation set. At this field intensity, the conformational dynamics is restricted further to a single extended structure with a most probable d_{ee} value of ~10 Å which, as expected is a bit higher than in the previous case for an applied electric field of only 30 kcal/mol·Å·e. As shown by data in Figs. 3.5 to 3.8, our choice or electric

field intensities, in agreement with earlier studies with less sampling,[74] allows us the monitor the entire expected range of conformational changes that can occur when using a classical MD simulation force field. While the results are intrinsically limited by the classical nature of our MD simulations, they capture, nevertheless the expected overall behaviour or the FF system and allows us to study the response conformational dynamics electric field and temperature perturbations.

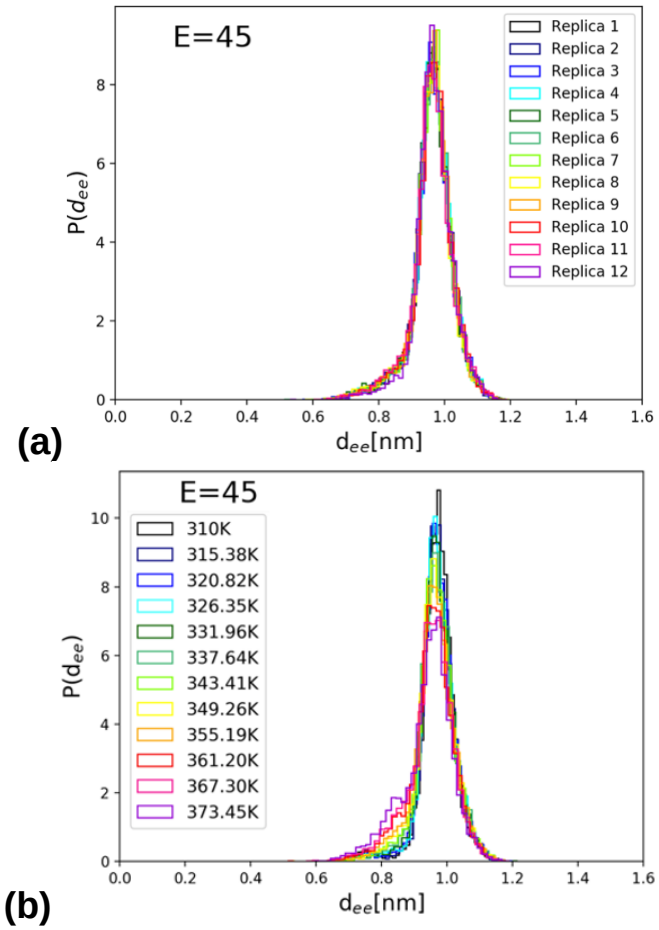


Figure 3.8. Distributions of d_{ee} values for simulations with an applied electric field of 45 kcal/mol·Å·e, (a) for each replica (R-trajectories), and (b) at each temperature (T-trajectories) of the REMD simulation set. At this field intensity, the conformational dynamics is restricted further to a single extended structure with a most probable d_{ee} value of ~10 Å.

The main results of our conformational and kinetic analysis are summarized in Figure 3.9. Here, are illustrated the temperature-dependent Markov kinetic network estimated from our new REMD simulation trajectories for FF peptides in the absence (top) and the presence of representative electric field intensities. Fig. 3.9 shows the

relative transition probabilities (blue arrows) between the three major conformational Markovian states (denoted as S_1, S_2 and $S_{3)}$ and their corresponding probabilities of occurrence (or state populations in percentages). Note that, in the absence of electric fields (Fig. 3.9 top), the FF peptide adopts three different main Markovian conformational states: S_1, S_2 and S_3. In Fig. 3.9, the corresponding equilibrium transition rates between these states (blue arrows, see text) are shown as numbers. The REMD transition rates, were extracted for the data corresponding to transitions occurring in all trajectories, cumulated for all the replicas (all R-trajectories). As shown in earlier works on analysing and extracting kinetic information from REMD data from different atomistic systems (e.g., FF [66], pentaalanine[109] and NNQQ[62, 63] peptides), while the data from all the R-trajectories correspond to dynamics at an intermediate temperature that is not exactly defined, they are nevertheless representative for the entire set of REMD replicas, at all temperatures. Moreover, the propagators for transitions along R-trajectories can be calculated analytically as weighted geometric means of propagator values extracted for the corresponding transitions in T-trajectories.[63] In Fig. 3.9, each arrow's thickness is proportional to the magnitude of its corresponding transition rate. On the bottom are illustrated the representative FF conformations, denoted here as S_2' and S_2'', adopted by the peptide in presence of external electric fields with intensities of 30 and 45 kcal/mol·Å·e, respectively. As illustrated in Fig 3.9 (and as suggested by the notation), our REMD simulations show that the S_2' and S_2'' conformations induced by the external electric field, at different field magnitudes, are part of the same conformational ensemble as the S_2 conformations adopted intrinsically by the FF peptide even in the absence of an externally applied electric field, but with a probability of only ~42 %. The S_2-type of molecular conformations, shown in Fig. 3.9, results from the peptide backbone stretching effect due to the presence of the external field, and results in a more direct exposure of the hydrophobic aromatic rings of the phenyl sidechains to peptide-peptide interactions facilitating FF aggregation. The interplay between increased backbone dipolar moments and stronger side chain-side chain interactions could be particularly important in understanding the dependence of FF-peptide aggregation propensities on physical parameters such as temperature and external electric fields.

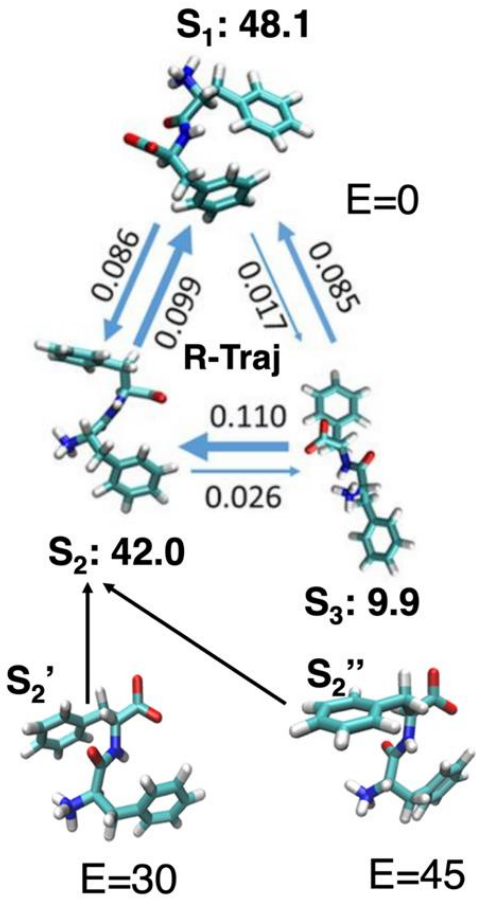


Figure 3.9. Representative conformations of FF amyloid peptides derived by kinetic analysis of REMD simulations at different electric fields. In the absence of electric fields, the FF peptide adopts three main Markovian conformational states: S_1, S_2 and S_3 (top). The corresponding equilibrium transition rates between these states (blue arrows, see text) are shown as numbers. These REMD rates are for the data corresponding to all the replicas (all R-trajectories). Each arrow's thickness is proportional to the magnitude of its corresponding transition rate. On the bottom are shown the representative conformations, S_2' and S_2'', adopted in presence of external electric fields with intensities of E = 30 kcal/mol·Å·e, and E = 45 kcal/mol·Å·e, respectively.

3.5 Conclusions

In summary, we show that replica exchange molecular dynamics (REMD) trajectories of explicitly solvated FF peptides can be used to probe in detail the interplay between temperature and electric field effects on the detailed thermodynamic and kinetic properties of the conformational dynamics of FF peptides in the presence of explicit water molecules.[66, 74] While their well-documented piezoelectric properties allow FF molecules and their aggregates (e.g., FF nanotubes) to be aligned in a controlled way by application of external electric fields, the detailed response of individual peptides to both temperature and electric fields are not fully understood. Here, we show that the thermodynamics and kinetics of the ensemble of conformations adopted by amyloid FF peptides solvated in explicit water molecules - an environment relevant to biomedical applications - can be analysed in detail by using REMD to enhance sampling, while simultaneously applying external electric fields and probing temperature ranges relevant to earlier studies.[74, 79, 80, 84, 85]

Methodologically important, our simulations highlight and overcome possible artifacts that may occur during the setup of REMD simulations of explicitly solvated peptides in the presence of external electric fields, a problem particularly important in the case of short peptides such as FF. The effect of an external electric field on the dipole moment due to the charged ends should be larger for relatively short peptides which have more extended backbone conformations and can respond easily to external perturbations. On the other hand, larger peptides and proteins may have more complex folds, backbone conformations and, also more charged residues in their composition. The effect of the external electric fields on conformations of such large systems may be more complex and harder to quantify than for shorter peptides.

The presence of the external fields could over-stabilize certain conformational states in one or more REMD replicas, leading to distortions of the underlying potential energy distributions observed at each temperature. This cause is different from REMD artifacts reported and documented by earlier studies, which were due to modified underlying energy distributions caused, for example, by the use of weak-coupling thermostats.[107, 108] In our case, we show that the resulting artifacts can be

overcome by correcting the REMD initial conditions to include the lower energy conformations induced by the external field. This is illustrated by the initial energy distributions shown in Fig. 3.2 and the corrected ones from Fig. 3.3. Such corrections could be also important in other replica-exchange simulations (e.g., using methods such as REST2 [110, 111]) that enable the use of broader range of temperatures in atomistic MD studies of amyloid peptide aggregation. [112]

Subsequently, we show that the corrected and converged REMD data can be analysed using a Markovian description of conformational states and show that a rather complex, 3-state, temperature-dependent conformational dynamics in the absence of electric fields collapses to only one of these states in the presence of the electric fields. As illustrated in Figure 3.9, we can study and analyse the detailed interplay between temperature and electric field on the thermodynamic and kinetic properties of solvated FF peptides. In particular, we identify and characterize the ensemble of S_2-type of molecular conformations, illustrated in Fig. 3.9, which are expected to play a particularly important role in understanding the dependence of FF-peptide aggregation propensities on physical parameters such as temperature and external electric fields. The mechanistic details behind the temperature-, and electric field-dependent thermodynamic and kinetic properties of small FF amyloid peptides can be useful in understanding and devising new methods to control their aggregation-prone biophysical properties and, possibly, the structural and biophysical properties of FF molecular nanostructures.

4. Transition Between Active and Inactive Conformation of Abl Kinase Studied by Milestoning[3]

4.1 Overview

Here, I combine a reaction path algorithm with the theory and algorithm of Milestoning to study kinetics of the DFG flip and compute the mechanism and the rate of the transition in ABL kinase. The activation of kinases includes a conformational transition of the DFG motif that is important for enzyme activity but is not accessible to conventional Molecular Dynamics. I propose a detailed mechanism for the transition, at a timescale longer than conventional MD, using a combination of reaction path and Milestoning algorithms. The mechanism includes local structural adjustments near the binding site as well as collective interactions with more remote residues.

4.2 Introduction

Kinases form one of the largest family of enzymes. In the human genome, there are about 500 predicted protein kinases. They catalyze the transfer of the γ-phosphate group from ATP to the hydroxyl group of a serine, threonine or tyrosine residue, a type of transfer that is found in many biological processes. Malfunctioning kinases are involved in many major human health-related problems such as cardiovascular diseases, diabetes, and cancer. Despite their diversity of function, the structure of their catalytic domain is shared across the kinase family. Roughly, the kinase domains consist of an N-lobe, and a C-lobe connected by a flexible hinge region (Fig. 4.1). The active sites consist of three conserved structural elements: the activation loop (A-loop), the Asp-Phe-Gly (DFG) motif, and the αC helix (which is part of the N-lobe).

3. This chapter has been adapted from reference [18].

These conserved structural elements make it challenging to design a drug that would be specific to only one kinase. Nevertheless, such a design is desirable to minimize unwanted, yet likely side effects, given the high structural similarity of members of this family.

Nevertheless, the drug imatinib[113] was found to be selective and inhibit BCR-Abl but not c-Src. Understanding the origin of this selectivity is of significant interest and potential for enhancing drug design efforts.[114] An intriguing proposal[115] explains the selectivity using variation in active site flexibility and binding of the drug to the inactive kinase conformation. However, a recent experiment suggests that the selectivity of imatinib towards Abl-kinase and not Src-kinase is a result of a slow conformational change that occurs after ligand binding.[116] The importance of the DFG flip to selectivity is therefore in doubt. Nevertheless, the DFG must change a structure at some step along the reaction to allow the entrance of the inhibitor to the active site.

Only kinases that are able to form DFG-out (inactive) conformation can open up a pocket to facilitate binding to the imatinib. Supporting evidence for a high activation loop conformational flexibility emerged from X-ray crystallography (significant variations were observed in many structures of kinase proteins).[117] Also illustrating diversity are NMR spectroscopy,[118] and molecular simulations.[119-123]

There have been numerous experimental studies to estimate the rate of transition between DFG-in and DFG-out for different members of the kinase family. It was argued that the operating mechanism of the enzyme is influenced by the rate of the DFG flip. The early evidence for this mechanism came from the experimental observations of the significant differences between the rates of binding of inhibitors to the DFG-out and the DFG-in states in P38 kinase. While the DFG-in inhibitors binding is quite fast and within the diffusion-controlled regime, the rate for DFG-out inhibitors were orders of magnitude slower.[118],[124-126] This observation suggests that DFG-in states are more populated in equilibrium. Kinetic may also play a role. If the rate of transition between the protein conformations is fast, the drug can always find a ready conformation to bind, and the ratio of the unbound population: [DFG-in]/[DFG-out] remains the same. If the rate is slow a shift in the conformation of the unbound protein

will be observed and the binding of the inhibitors to one of the states may slow down if saturation of proteins bound to ligands is reached.

In an NMR study,[127] it was shown that the binding of different inhibitors can influence the dynamics of the DFG motif in Abl-kinase. A wide range of timescales was suggested for the dynamics of different residues in presence of different inhibitors. However, determining the accurate time scale of the DFG flip remains challenging due to the lack of signal for some key residues from DFG motif and activation loop. In another study,[118] probing the kinetics of the DFG-flip in NMR measurements for P38 kinase, the observed experimental line-shape is very broad and indicates an intermediate time scale on the NMR time scale (milliseconds). In a combined experimental and computational paper, the time scale of imatinib binding, under a variety of perturbations, was estimated to be in the range of ten milliseconds.[121] A recent study combines NMR and tryptophan fluorescence on imatinib binding to Abl-kinase and observes two time-scales for the binding event.[116] One time-scale is at, or below a few milliseconds and a second time is of 100 milliseconds to seconds.[116] The process with a slower time scale was identified as a large conformational transition. The shorter time scale is assigned to a local binding event. If we assume that the local and fast event of imatinib binding to Abl kinase is associated with the DFG flip, we can obtain an indirect estimate of the time scale ~2ms, as we illustrate in the discussion.

Summarizing experimental results of particular relevance to the present study, these studies shared the following observations (i) DFG-in state is more populated than DFG-out state in Abl kinases for unprotonated Asp381and (ii) Some indirect evidence is available that the timescale for the flip DFG-in to DFG-out in kinases is in the millisecond time scale.

Computational studies are in general agreement for the slightly larger stability of the DFG-in active conformation compared to DFG-out state in Abl kinase with Asp 381 unprotonated.[122, 128, 129] However, the study of kinetics is more restricted due to limitations on conventional MD. A recent study of Abl kinase, using MD and the Markov State Models[120] suggest a time scale for the transition of milliseconds.[116] However, the statistics of transitions was small. This estimate is consistent with

conventional MD simulations that were not able to sample the DFG-in and DFG-out transition on hundreds of nanoseconds simulation time without an assisting mutation (M290A).[121] The kinetic of the DFG-in to DFG-out transition is an important component of the Abl kinase activity. It is therefore of interest to further investigate the mechanism of the transition and quantify the time scale of the process.

4.3 Method

Choice of reactant and product structures

The reaction path approach that we use requires as input the conformations of the two end states, the reactant and the product.[130] Given the richness of crystallographic structures of kinases and specifically of Abl kinases the choice of the end structures requires discussion.

The "classical DFG-out" cluster of structures in the PDB (with ~200 structures) is the most prevalent inactive state observed among kinase structures with DFG-out, in which the activation loop(A-loop) is fully folded/closed.[131] The second most populated cluster of DFG-out inactive structures is a cluster we refer to as "DFG-out minimally perturbed", and 2G2F is a member of this cluster. The root-mean-square deviation (RMSD) of the activation loop over all DFG-out inactive structures with respect to 2F4J which is a representative of an active kinase with the A-loop extended in the active conformation, indicates that the A-loop in the DFG_{out}-$A\text{-}loop_{minimally\text{-}perturbed}$ cluster differs by 1-5 Å from 2F4J, in contrast to the large range variation of between 11-19 Å in the classical DFG-out cluster (Ref structure: 2F4J chain A). Representatives of the DFG_{out}-$A\text{-}loop_{minimally\text{-}perturbed}$ cluster have been observed for 11 kinase families, and Abl (with ~10 unique pdbs) is the most observed family among them.

The Protein Data Bank (PDB) structures of the Abl kinase with accession codes of 2F4J[132] and 2G2F[133] were used as active (reactant) and inactive (product) conformations, respectively. We picked two conformations that are not profoundly different with the exception of the activation loop, making it possible for us to focus on the transition of the DFG motif. For example, there is no significant shift of the αC helix between the two states. In previous studies it was also shown that the inactivated set

of structures is highly flexible (see for instance[120, 131]). A significant number of structures are sampled with conventional Molecular Dynamics in the flexible state and it is likely that a single pathway calculation assisted with unbiased trajectory sampling of Milestoning will probe the transition network.

The structures were solvated separately with TIP3P water molecules and salt concentration of 0.15M (NaCl). The systems consist of ~45,000 atoms. All the simulations have been conducted with the NAMD program[134] and the CHARMM36[135] forcefield has been used. Periodic boundary conditions were used, and the system was minimized using conjugate gradient algorithm for 10,000 steps. Equilibration followed in the NPT ensemble with Nose-Hoover Langevin piston pressure control for 5 ns at pressure of 1atm and temperature of 310 K.[136, 137] Then the system was equilibrated in the NVT ensemble at 310 K using Langevin thermostat for additional 10 ns. Water molecules were kept rigid with the SETTLE algorithm[138] and all other bonds with hydrogen atoms were kept fixed with the SHAKE algorithm.[139] The cut off distance for non-bonded interactions was 12 Å and the Particle Meshed Ewald method was used to sum the electrostatic interactions.[140] The timestep was 1 fs. The final configurations of the equilibrated structures were used as the reactant and product states for pathway generation, with root mean square distances of the equilibrated structures relative to the initial structure were 0.9 Å and 1.1 Å, respectively.

Generation and optimization of pathway

Examining the active and inactive conformations (Fig. 4.1) we realize that the major differences between the two conformations are concentrated at the activation loop (from residue number 380 to 400).

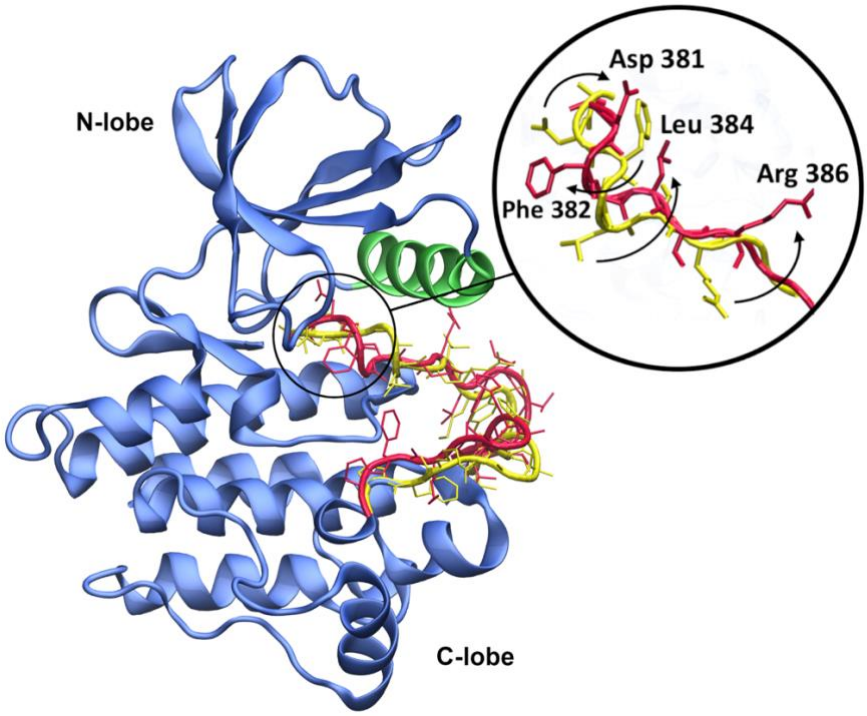


Figure 4.1. A schematic representation of the Abl kinase protein. The active and inactive states of the activation loop is shown in yellow and red, respectively. The C helix is green. The magnified region shows the start of the activation loop that includes the DFG switch. The image was generated by the software VMD.[1]

Therefore, the reaction space or the coarse variables that guide the reaction path calculations were selected from this region. First, the backbones of the two structures were aligned for a best overlap for the entire structures excluding the activation loop. Then 24 atoms along the loop were selected to represent the coarse space. The selected 24 atoms include both alpha carbon and atoms from the side chains of the residues with the highest RMSD values between the reactant and product. These atoms are listed in Table 4.1. The actual number of degrees of freedom in the coarse space is smaller than $3 \times 24 = 72$ since bond lengths and bond angles do not vary significantly in the calculations and may be considered fixed. We define an active torsion in the reaction space if at least one atom from the coarse space is included in the definition of the torsion, and the torsion changes along the reaction coordinate by at least 60 degree. The number of such torsions that have significant contribution to the coarse space is 22. The reaction space that we considered is therefore quite large compared to other studies in the field.

Val379(CA)	Val379(CB)	Ala380(CA)	Ala380(CB)
Asp381(CA)	**Asp381 (CG)**	**Phe382 (CA)**	**Phe382 (CG)**
Gly383(CA)	**Leu384(CA)**	**Leu384(CG)**	Ser385(CA)
Arg386(CA)	**Arg386(CZ)**	Leu387(CA)	**Leu387(CG)**
Met388(CA)	**Met388(CE)**	Thr389(CA)	Thr389(CB)
Asp391(CA)	Tyr393(CA)	His396(CA)	**Ala397(CA)**

Table 4.1. List of the 24 atoms used to define the coarse space in the calculations of the pathway. The final 12 atoms that are used in the Milestoning calculations are indicated in red. See text for more details about the selection.

The method of Milestoning is used to compute kinetic and thermodynamic observables. As a first step in a Milestoning calculation we require a rough sample of the space linking the reactant and product. The number of intermediate configurations that we use as a sample varies depending on the system characteristics. It is between a few tens to several thousand structures. These configurations form centers of Voronoi cells and their interfaces are used as Milestones, either in the Markovian Milestoning approach,[70] or in other variants of Milestoning.[141] The choice of the Voronoi cells impact the rate of convergence of the calculation but should not impact the final results if the system is close to equilibrium, or if iterations of the exact Milestoning approach are used.[67]

One way of generating centers of Voronoi cells that cover the relevant space is by a reaction path calculation. There are multiple studies of generating reaction coordinates in complex systems, starting with the study of a conformational transition in myoglobin [142] and continuing to a number of other complex systems [143, 144]. In these approaches, and variants of them [145, 146], a guess is generated for the path using methods like, self-penalty walk method [147], and then optimized. Other Path sampling algorithms have be summarized in ref [148, 149]. Unfortunately, the end result may be biased by the initial guess, especially on rough energy landscapes in which multiple pathways exist. This problem was discussed in Ref. [128] in the context of kinase conformational transitions. The activation loop is highly flexible and transitions in a space of several dimensions. Nevertheless, our current interest is

focused on the DFG transition, which is spatially small. The transition is also activated, and is expected to be a rare event. Determining the pathway is challenging for conventional MD, but it is more straightforward to compute with reaction path calculations.

Recently we introduced a new method for computing reaction coordinate in complex systems that does not require an initial guess (Rock Climbing[130]). The local optimization of the path were carried with implicit solvent, using a Generalized Born method[150] with ion concentration of 0.15M and solvent dielectric of 78.5 while the global optimization was done in explicit solvent. During the pathway generation only two regions of the protein are allowed to move. The first region is of residues 376 to 405 that includes the activation loop. The second region is of residues 278 to 299 that contains the αC helix. The backbone of the rest of the structure was restrained with harmonic potentials to their initial positions during the pathway calculations.

We divide the path calculation into two steps. First, using a local and greedy algorithm we generate a pathway from the coordinates of the reactant to the product as follows: The generation of the pathway starts with adding a displacement vector, δ, along the vector connecting reactant to product. By providing information about the end points we made the process global. However, we do not provide an initial guess for the entire path and the path generation follows a local procedure. To begin with, each atom of the coarse space of the reactant is shifted by $\delta = 0.25\,A$ toward the product (equilibrated inactive kinase). Then harmonic restraints with force constant of 2000 kcal/mol A^2 were applied on the selected atoms of the coarse space and the rest of the system was minimized for 500 steps followed by 5000 steps of MD at 310 K. Finally, the harmonic restraints were released, and the system was minimized for 50 steps.

The last configuration is used to generate a new displacement vector, δ, toward the product configuration. The displacement is added to the current coordinate set and the relaxation process described above follows. The process is repeated until the product is reached. The value of δ is tuned during the process to allow for faster convergence or better accuracy. We obtained 850 structures interpolating between the

reactant and product, with an average displacement length of about 0.1 Å. Out of these 850 structures, 50 structures were selected that are approximately equidistant in the coarse space. These structures are used in the next step of path refinement.

In the second step, the path is globally optimized. All the 50 structures including the reactant and product, are solvated and equilibrated for 1ns in the NPT ensemble using Nose-Hoover Langevin piston[136, 137] followed by 5 ns in the NVT at 310 K. The coarse variables are restrained to their corresponding positions along the path using harmonic force constants of 150 kcal/mol. After equilibration of the bath coordinates (coordinates that are not included in the coarse space), we refine the pathway. Each configuration along the pathway in the coarse space is represented by the vector x_i. The following target function is optimized:

$$T = \frac{1}{2}\sum_{i=1}^{n}[(\,|\,\nabla U(x_i)|\,+\,|\,\nabla U(x_{i-1})\,|)\Delta l_{i-1,i} + k\left(\Delta l_{i-1,i} - <\Delta l>\right)^2] + \sum_{i=2}^{n} H\left(\theta_{i,i-1,i-2} - \theta_0\right)k'(\theta_{i,i-1,i-2} - \theta_0)^2 \quad (1)$$

The heavy side function is denoted by $H(y)$.

$$H(y) = \left\{ \begin{array}{ll} 0 & y < 0 \\ 1 & y \geq 0 \end{array} \right\}$$

The norm of the force in coarse space, $|\,\nabla U(x_i)|$, is an average over all coordinates of the bath, i.e. the coordinates that are not included in the coarse space. The first term in Eq. (1) is a discrete approximation to a functional of the path:

$$S\left[x(l)\right] = \int_{x(0)}^{x(L)} \left|\nabla U\left(x(l)\right)\right| dl \quad (2)$$

The path that minimizes $S\left[x(l)\right]$ is the Steepest Descent Path (SDP) in the free energy landscape of the coarse variables, *x*.[151]

Eq. (1) is a discrete version of Eq. (2) in which we must ensure that the configurations are distributed uniformly along the path. Therefore, we added the second term. The norm of the distance vector between the structures x_i and x_{i-1}, is $\Delta l_{i-1,i}$ and $<\Delta l>$ is

the average distance over all the neighbor distances of the path ($\langle \Delta l \rangle = \frac{1}{n}\sum_{i=1,\ldots,n} \Delta l_{i-1,i}$). Using $< \Delta l >$ allows the path to gradually expand or shrink in every iteration, if required, and keeping the points equally spaced along the pathway. The force constant *k* is *200 kcal/mol.A*^{2}.

Another difference between the discrete and continuous paths is that the discrete path can makes sharp turns of high curvature that reduce the value of the discrete functional. The third term prevents large path curvatures. It is applied only when the angle, $q_{i-1,i,i+1}$, between three sequential configurations of the path is larger than a threshold value, $\cos(\theta_0)$. $\left(\cos\left(q_{i-1,i,i+1}\right) = \left(x_{i+1} - x_i\right) \times \left(x_i - x_{i-1}\right) / \left\| \left(x_{i+1} - x_i\right) \times \left(x_i - x_{i-1}\right) \right\|\right)$. Here we use $\theta_0 = 60^o$ and *k'=5 kcal/mol.deg*^{2}*.*

With the above definition of the target function of the path, the optimization is conducted in iterations. In every iteration, the 48 intermediate structures are simulated for 1ns to compute the average force. Harmonic restraints with force constants of 150 kcal/mol/Å^2 are applied on the atoms of the coarse space during the 1ns simulations and the mean forces at each structure are computed. The coordinates of the coarse variables of the path are then adjusted by a small step (typically 0.01-0.03Å) to minimize the function of Eq. (1) and to move toward the minimum free energy path. This procedure is repeated one hundred times. Convergence is assumed when the norm of the force gradient does not vary significantly. Only the force component parallel to the reaction path remains when the SDP is reached (Fig. 4.2).

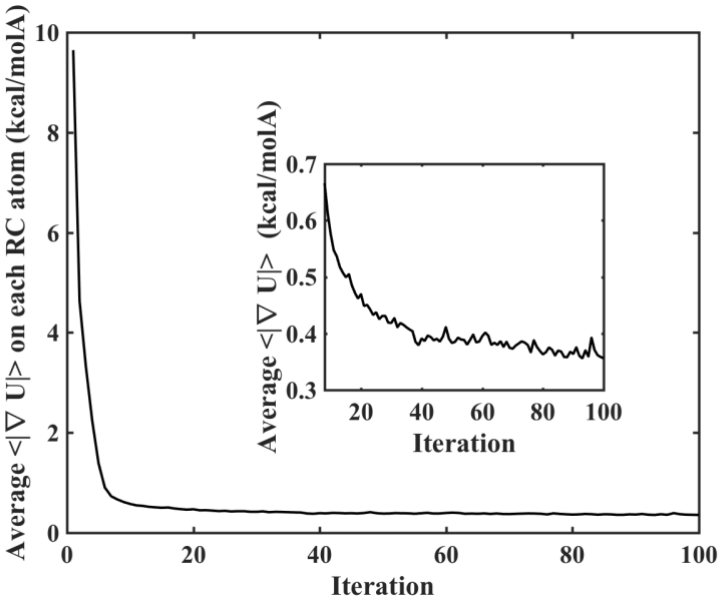


Figure 4.2. The reduction in the norm of force, averaged over the entire reaction path, as a function of iteration number. The norm of the force drops rapidly in the first ten minimization steps and then decays more gradually. It seems to stabilize at around 100 iterations. The final gradient is around 0.35 kcal/mol $Å^{-1}$. It is not zero since the norm of the force along the reaction coordinate is included.

After the calculations of the optimized path were complete, we found that some of the atoms included in the coarse space were hardly moving. We therefore adjusted the coarse space to include only coordinates that were modified significantly along the optimal reaction pathway. This reduces the number of coarse variables that define the coarse space from 72 to 36. The final list of atoms that determine the coarse space is given in Table 4.1.

Milestoning

Milestoning is a versatile theory and algorithm to compute thermodynamics and kinetics of complex systems using a large number of short trajectories. It was discussed extensively in the literature[67, 152, 153] and a review article is available.[154] We therefore described it below only briefly.

In the first step of a Milestoning application we provide a sample of configurations from the coarse space between the reactants and products (Fig. 4.3).

In the present study the sample consists of the configurations along the reaction coordinate, $\{x_i\}$, described in the previous section. Each configuration is called an anchor and is a center of a Voronoi cell in coarse space. The boundaries between the Voronoi cells (say cells i and j) are called milestones $M_{ij} \equiv M_\alpha$. Milestone M_{ij} is the set of points with equal minimal distances to anchors i and j and larger distances from all other anchors. For brevity we index milestones also by a single Greek letter, e.g. M_a .

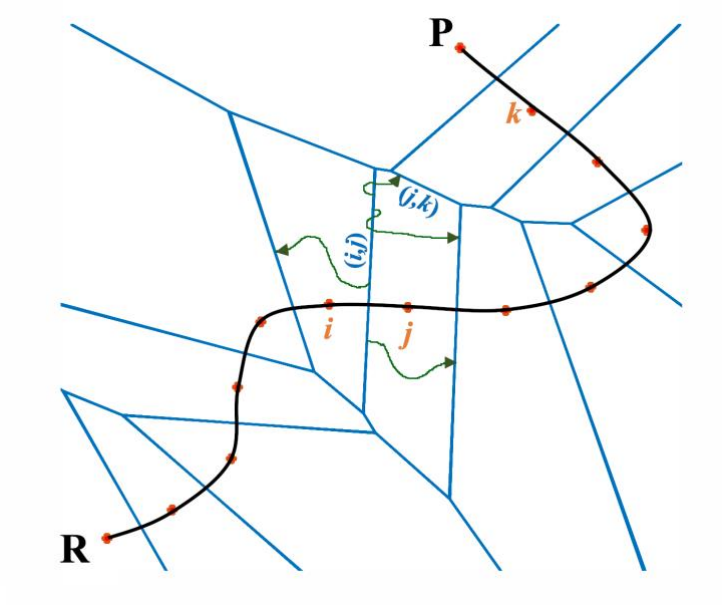


Figure 4.3. A schematic representation of the discretization of the coarse space following the transition pathway. R and P represent the reactant and product states, respectively. The black line shows the reaction pathway. The red dots are the anchors, and the blue lines are the milestones. Every milestone is numbered by its corresponding anchors. For example, milestone (*j,k*) is the boundary between cells *j* and *k*. The green arrows show 4 unbiased trajectories initiated from milestone (*i,j*). The trajectories are terminated when they hit any other milestone for the first time. Re-crossing the original milestone does not lead to trajectory termination.

In the second step of Milestoning we sample configurations from the Boltzmann distribution using MD simulations constrained to the milestones. We generate configurations from the conditional probability density

$p(X|M_{ij}) = [\exp(-\beta H(X))\, X \in M_{ij}]$ where X denotes the phase points constrained to milestone M_{ij}.

In the third step we launch unbiased MD trajectories from the sampled configurations at the milestone generated in the second step. We terminate these trajectories when they hit for the first time a milestone different from the milestone they started from. We recorded the identities of the milestones and the time of termination. In one variant of Milestoning, which is not used here, (exact Milestoning[67]) we also retain the terminating phase space configuration at the b milestone.

In the fourth step we use the information gathered in the first step to compute two functions $K_{\alpha\beta}$, the probability that a trajectory initiated in milestone α will terminate at milestone β and t_a, the lifetime of milestone α. Let n_a be the number of trajectories initiated at milestone α. Let n_{ab} be the number of trajectories that were initiated at milestone α and were terminated at milestone β. We estimate the transition probability also called the kernel as $K_{\alpha\beta} \cong n_{\alpha\beta} / n_\alpha$ and the lifetime $t_\alpha = \frac{1}{n_\alpha}\sum_{l=1,\ldots,n_\alpha} t_l$, where l is the index of the trajectory, and the time length of trajectory l is t_l .

In the final and fifth step we compute the thermodynamic and kinetic observables. The stationary flux of trajectories through a milestone is the eigenvector, $\mathbf{q}$, of the matrix $\mathbf{K}$ with an eigenvalue of one: $\mathbf{q}^t\mathbf{K} = \mathbf{q}^t$ where we used bold face for vectors and matrices. The free energy of a milestone a is given by $F_a = - k_B T \log\left[q_a t_a\right]$ and the mean first passage time (MFPT, $\langle t \rangle$) is given by $\langle t \rangle = \mathbf{p}_0\left(\mathbf{I} - \mathbf{K}'\right)^{-1}\mathbf{t}$, $\mathbf{p}_0$ is the vector of the initial distribution and $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{K}'$ is an adjusted $\mathbf{K}$ matrix in which absorbing boundaries are placed at the product state.[67]

More computational details on the implementation of the Milestoning method to the kinase problem are given below.
Step 1: We use 50 structures equally distributed along the path to define the anchors. Sequential anchors are separated by a distance of 0.5 Å along the reaction coordinate (RMSD of the coarse variables) which is a typical value for a Milestoning calculation.

Step 2: Since the anchors are sampled along a one-dimensional reaction coordinate, it is suggestive to place the initial milestones between sequential anchors. We

therefore initiate trajectory sampling for 49 milestones between each of two consecutive anchors along the path. We conduct 1ns MD simulation at a constant temperature while two restraints were applied to keep the sampling trajectory restrained to the milestone. The first restraint is a harmonic term $k(d_i-d_j)^2=0$ to keep the distances d_i and d_j (in coarse space) between the current configuration and the two anchors *i* and *j* equal. The restraining force constant is 2000 kcal/mol.A^2. A second set of half-harmonic restraints are $k'.(d_m-d_l)^2$ when $d_m<d_l$, $l=i,j$ and zero otherwise with k'=1000 and *m* being any milestone other than *i* or *j.* This restraint prevents the system from getting closer to any other milestone, *m*. 100 samples were kept from the final 0.5ns of the restrained simulation at each milestone.

Step 3: We release all the restraints and conduct unbiased short MD trajectories at the NVE ensemble starting from the configurations at each milestone that were sampled in step 2. The trajectories are terminated when they hit a milestone different from the one that they were initiated at. The typical length of each of the unbiased trajectories is 10 ps. The trajectories could either reach one of the initial 49 milestones we started from or they may reach a new milestone. For example, in Fig. 4.3 one of the trajectories initiated from the milestone (*i*,*j*) reaches a new milestone (*j*,*k*) that connects non-sequential milestones along the path. We therefore add the new milestone to the list and sample configurations at the newly discovered milestones. Finally, we launch unbiased trajectories from these new samples. The process is continued until we do not visit new milestones, or a connected network is observed in which the MFPT value for the transition between reactant and product converges to a finite value. The final number of milestones we ended up with in the current project was 171 (Fig. 4.4).

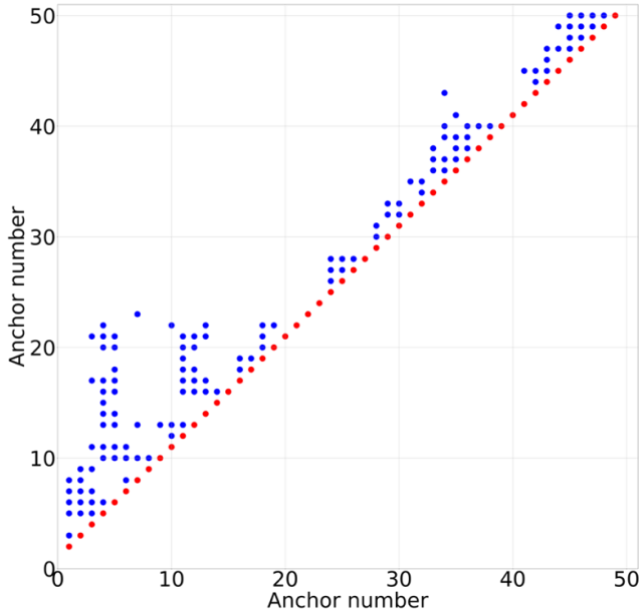


Figure 4.4. Representation of all the 171 milestones considered for computing the transition matrix. Every point corresponds to one milestone. The red points are the initial milestones between the consecutive anchors along the transition path. The blue points represent the milestones discovered during the analysis of the free trajectories that are used to enrich the sampling of the pathways.

4.4 Results

Fig. 4.5 shows the free energy landscape of all the milestones in two dimensions where the dimensions are the anchor indices. To achieve a better qualitative understanding of plausible transition paths between the active and inactive states, we determine optimal pathways in the milestones space. Every milestone is a node in a network and we assign weights to the network's edges as rate coefficients for transitions between milestones. Rate coefficients are computed from the transition matrix **K** and average lifetimes of the milestone, **t**. The rate coefficient for a transition between a milestone pair (*i*,*j*) is given by

$$k(i,j) = \frac{K_{ij}}{\boldsymbol{t_i}} \tag{3}$$

It was shown that choosing rate coefficients between the milestones in this way leads to a Master Equation where the milestones are the states. It was also shown that the exact MFPT is obtained with this Master Equation (but not higher moments of the first passage time).[155] Using the rate coefficients, global maximum weight paths

(GMWP) from reactant to product and vice-versa, were obtained from the network, shown in Fig 4.4, using recursive Dijkstra's algorithm[156] (Fig. 4.5). In other words, these are the pathways on the network with the fastest rates. 1D free energy profiles are shown in Fig. 4.6 for the two GMWPs between the active and inactive states. According to these plots, the free energy landscape consists of multiple barriers with a maximum height of ~15.7 $\pm$ 2.5 kcal/mol with respect to the active state. Also, the free energy of the inactive state is 2.8$\pm$2.0 kcal/mol higher than the active state, suggesting that the active state is more stable.

The local minima and maxima that are observed along the paths follow bond rotations of different loop segments. These rotations are roughly independent, and they are observed at different positions along the reaction coordinate. For example, the rotation of the DFG motif contributes to the first barrier along the pathway near position 13 in Fig 4.6a. The local minima between positions 16 and 24 correspond to states where the rotation of DFG and residue 385 are complete. Finally, the rotations of Leu387 and Met388 contribute to the free energy hump from positions 24 to 31.

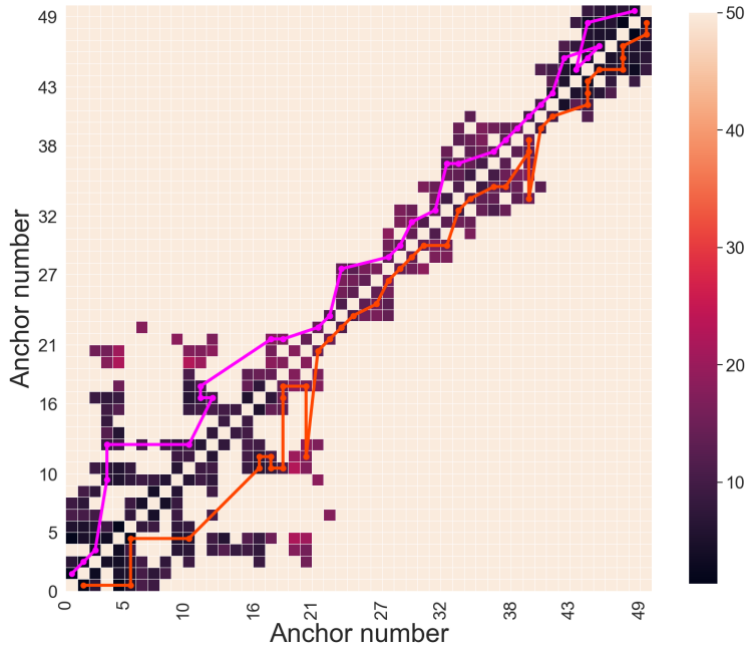


Figure 4.5. Free energy of pairs of anchors as computed from the Milestoning theory. The energy values are in Kcal/mol. The two paths with maximum flux from the reactant to the product are lines in red and in magenta. Note that the significantly off-diagonal "jumps" on the surface are a consequence of long-range connection between milestones that are not in sequence along the reaction pathway (see also Fig. 4.4).

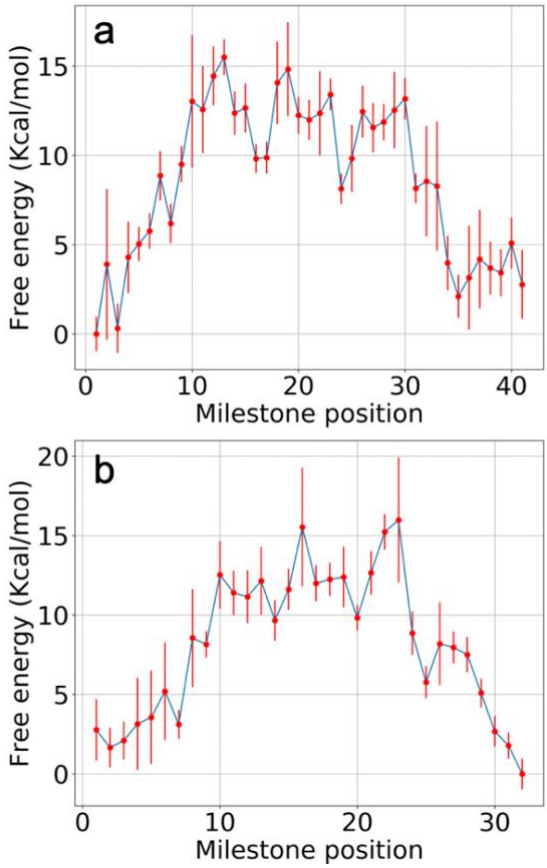


Figure 4.6. Two optimal free energy profiles along the two max-flux pathways from active to inactive state. In panel a, the milestones are numbered from 1 to 41 for the corresponding points along the red path shown in figure 4.5 starting from active state, and from 1 to 32 for the magenta path in figure 4.5, starting from the inactive state for panel b.

Note that Fig. 4.6 is plotted as a function of the index of the milestones, which is different from the anchor number. This is because we consider nodes and edges on a network that are labelled separately. The mapping between anchor pairs and milestones can be found in Fig. 4.5. The mapping is not trivial since the path is not monotonic in the index of the anchor and is monotonic in the milestone index.

The committor function is another useful quantity that can be calculated directly from Milestoning.[157] The committor function, C, at every milestone, is the probability that a complete trajectory initiated at that milestone will reach the product before reactant. Interesting milestones are those with values close to $C \approx 0.5$ that can serve as a definition of the transition state. According to Fig. 4.7 this occurs near milestones (21,22), (22,23), and (23,24). Fig. 4.8 shows the conformations of the first residues of the activation loop for anchor 23, and the active and inactive states. At anchor 23 residue Arg386 already moved to its final state while Asp381 did not change its

configuration significantly. The sidechains of Phe382 from the DFG switch and of Leu384 are rotated approximately half of the way through. Their significant motions at the transition state and their intermediate structures suggest them as a bottleneck for the transition.

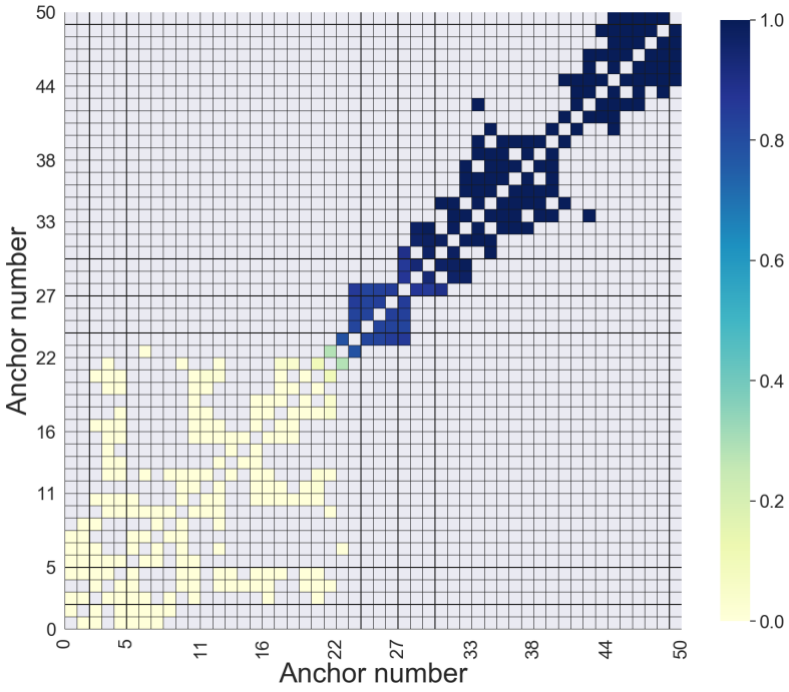


Figure 4.7. Color-coding the committor function at every milestone. The committor of a milestone is the probability of a complete trajectory initiated at that milestone to reach the product before the reactant state.

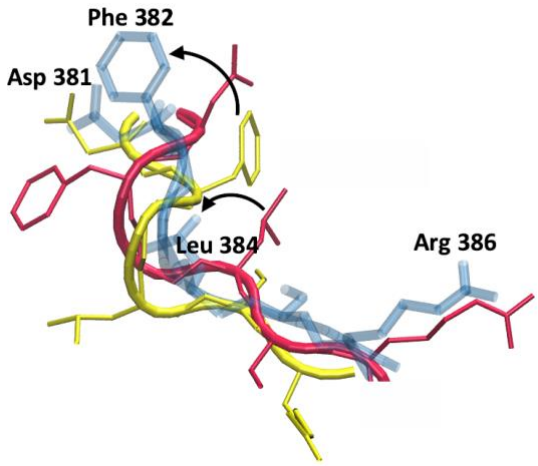


Figure 4.8. A stick model of residues 381 to 386 for active (yellow), inactive (red), and a sample configuration at anchor 23 where the committor value is near 0.5 (blue). Note that Arg386 already reached its final position at the transition state, while Asp381 did not change its configuration significantly. The residue Phe382 is found at half of the way of the transition.

Although the DFG-in and extended conformations of the A-loop are essential elements for a kinase to be active, active conformations are also associated with other features such as $\beta3_{Lys271}$ to $\alpha\text{-Chelix}_{Glu286}$ salt-bridge. The orientation of Lys interacting with both Glu and the DFG-Asp with the assistance of ions provides a proper network for coordinating the ATP phosphate in active kinase structures.[133] Detailed structural investigation of the DFG_{out}-$A\text{-loop}_{minimally\text{-}perturbed}$ structures indicates that this salt-bridge is not always maintained due to differences in the orientation of the DFG-Phe side chain. There are three possibilities: 1) DFG-Phe is located between the Lys and Glu, in a way that Glu is free to interact with the HRD-Arg or DFG-Gly amide groups. Therefore, the salt-bridge is broken. 2) The DFG-Phe points to the back-pocket and the salt-bridge is maintained. 3) The DFG-Phe points out to the solvent and the salt-bridge is maintained. Abl kinase structures are observed in all three groups, 2G2F chain B belongs to the second group for which the salt bridge is present in the inactive state. For the structures studied in this paper, this salt bridge exists at both, active and inactive states. With the reaction coordinates between the two states at hand, we are able to probe the status of this salt bridge at different steps. Our results show that the rotation of the DFG residues and more specifically, Phe382, requires breakage of this salt bridge. This was pointed out already in Ref.[120]. The salt bridge breaks between anchors 22 to 28 along the reaction pathways, which the committor analysis suggests to be close to C~0.5. Fig 4.9 shows a sample configuration from anchor 27 in which the distance between the charged groups of Lys271 and Glu286 reaches ~7 Angstroms. When the Phe382 reaches its final destination, the salt bridge reforms.

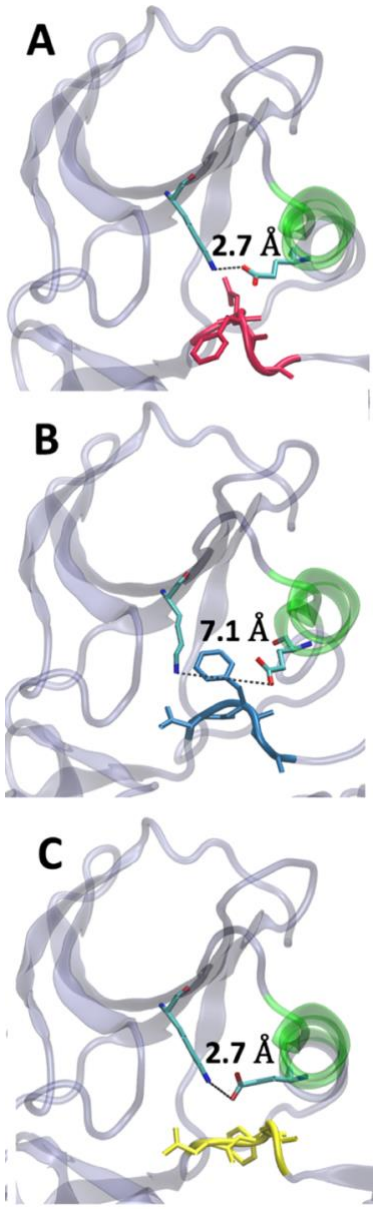


Figure 4.9. Changes in the salt bridge between Lys271 and Glu286 for inactive (A), an intermediate state (B), and active (C) states. The salt bridge exists in the active state and inactive states but during the transition from active to inactive state, the salt bridge breaks. The intermediate state shown is anchor 27. The DFG residues are shown in red, blue, and yellow for inactive, intermediate, and active states, respectively.

Non-Markovianity and MFPT

One of the advantages of the Milestoning method is that the description of the dynamics is not required to be Markovian. The assumption of Markovianity fails in numerous biological processes. As an illustration, Fig. 4.10 shows the distribution of the lifetimes for a few milestones. If the kinetics is Markovian, then the probability

distribution of the transition time between two states (in our case milestones) must follow an exponential behavior of the type:

$$P(t) = Ae^{-t/<t>} \tag{4}$$

Where *P(t)* is the distribution of the lifetime, *A* is a constant, and <*t*> is the average lifetime of a milestone. In Fig. 4.10, we first calculated the average lifetime from the distribution and then fitted an exponential curve to the computed histograms. The R^2 values are shown in every window. The distributions deviate significantly from a Markovian behaviour. A deviation from a single exponential behaviour for the binding of imatinib was pointed out experimentally.[116]

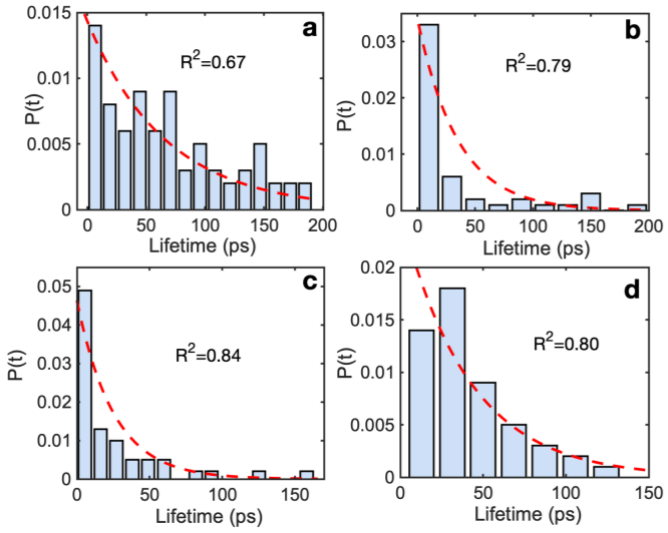


Figure 4.10. Distribution of lifetime for (a) milestone (21,22), (b) milestone (22,23), (c) milestone (24,25) and (d) milestone (40,41).

Finally, the MFPT for the transition can be calculated from Milestoning by sampling transition matrices and lifetimes from their known distributions. The sampling procedure was discussed in details in reference [158]. Fig. 4.11 shows the distribution of the MFPTs obtained from 1,000 sample transition matrices and lifetimes. The averaged MFPTs for forward and backward processes are 1.0 ± 0.9 s and 0.2 ± 0.1 s, respectively. These values are obtained by ignoring the bins containing values less than 5 % of the bin with the maximum population.

It can be seen that the mean transition time for going from active to inactive state is slightly higher than the reverse process since the active state is located approximately 2.8 kcal/mol lower than the inactive state.

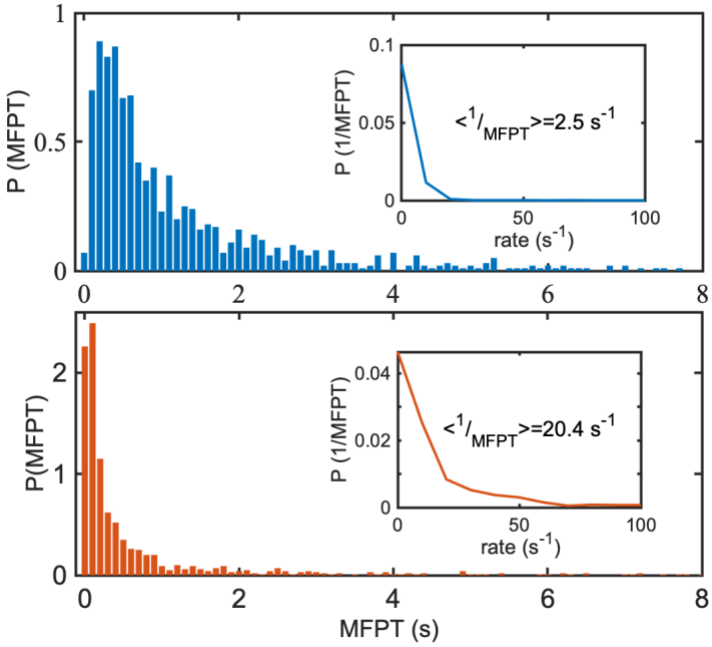


Figure 4.11. Distributions of MFPT for transition from active to inactive states (top) and the reverse process (bottom). The insets show the corresponding distributions for 1/MFPT which are estimates of the rate coefficients consistent with the simulation data and the error analysis. We quote the mean values of the rate coefficients for the forward and backward transition.

In the present study, we investigate computationally the kinetics of the DFG-in to DFG-out transition. There are two main observables that we discuss below. The first is the time scale for the process and the second is the mechanism or the structural features of the reaction. Since significant variations were observed in the measured and computed rates for different kinases, there is a significant uncertainty in the comparisons.

The time-scale for the DFG-in to DFG-out transition in Abl-kinase.

Milestoning simulations provide an estimate of the Mean First Passage Time without assuming the existence or the need to identify a bottleneck or a transition state. The only assumption invoked in the present version of Milestoning is that the system remains close to equilibrium.[154] Even though we have used a set of coarse variables to guide the calculations, the calculations are exact provided that the equilibrium

assumption is satisfied and the coarse variables are able to differentiate between reactants and products. This theory is different from other approaches in the field to study kinetics[120] and it is interesting to compare its results to those of related calculations and experiments.

Experimentally, there are a few indirect observations about the rate of the DFG flip that we discuss below.[116, 121] One observation is an NMR measurement of the dynamics of p38 kinase.[118] The NMR spectrum of the DFG motif was found to be very broad suggesting an intermediate NMR time scale (milliseconds). Fast processes in NMR are characterized by a single averaged peak while multiple peaks of different conformations correspond to a slow process by NMR scale.

Combined NMR and flow experiments on Abl kinase provide an alternative picture.[116] The first is fast and was associated with the local binding of imatinib, and the second, a slower process. The second process was interpreted as an induced fit that follows the binding event. The experiment does not provide a direct structural information on the DFG dynamics during those events, however, a reasonable assumption is that the fast process includes the DFG flip. The measurements that include the insertion of the inhibitor are significantly different from our study that focuses only on the DFG flip. We therefore declined to analyse our results in the context of reference [116].

From computational point of view, a Markov State Model was used to estimate the transition rate of the DFG flip at milliseconds.[120] The model was based on samples of significantly shorter MD trajectories (similar to Milestoning). A small number of transitions was detected making the uncertainty of the longer time scale significant, as noted by the authors. The estimates of the kinetics that use unbiased trajectories, MSM and Milestoning, provide answers close to each other and close to the experimental finding.

Structural features of the reaction pathways from DFG-in to DFG-out in Abl

In this section we discuss in more details the structural features of the transition from DFG-in to DFG-out. In Fig. 4.13 we show the residue root mean square difference (RMSD) between the reactant, and several structures along the reaction pathway. We

include a comparison of the reactant and of the product to anchor 27, which is near a committor value of 0.5 (Fig. 4.7).

We overlap optimally all the heavy atoms (atoms that are not hydrogens) of the entire structures and compute the heavy-atom RMSD between the individual residues. We plot the residue RMSD as a function of the residue number. For comparison we also plot the B factor (also called temperature factor) extracted from the PDB coordinates of the reactant (PDB 2F4J) and the product (PDB 2G2F). The B factor measures the displacement of the atomic positions from its mean position, thus makes it possible for us to identify residues that are flexible, and/or are present in multiple static conformational states. In contrast, residues that contribute to the spatial progress of the reaction may have displacements not detected by the B factors which are coupled to the DFG flip. Several of the peaks at residues F382, L384, R386, M388 and A397 are part of the A loop.

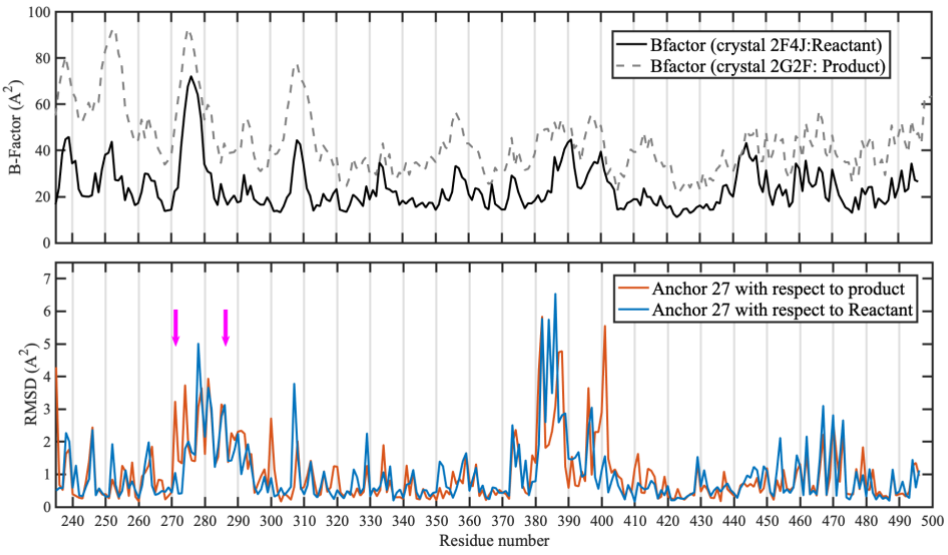


Figure 4.13. Fluctuations and systematic drifts of residues in Abl-kinase. Top panel reports the B factors of the reactant and product structures as a function of the residue index to identify flexible domains. In the lower panel we compare the structure of Anchor 27 with the reactant and product using RMSD between all the heavy atoms of the residues in the protein. The two pink arrows point to Glu286 and Lys271 that forms a blocking salt bridge. Note that the transition state differs about equally from the reactant and from the product structures. There are several spikes at Phe382, Leu384, Arg386, Met388 and Ala397 that belongs to the A loop and are included in the set of coarse variables.

We mark with pink arrows the locations of the salt bridge residues: Glu286 shows significant deviation from both, the reactant and the product, but Lys271 is closer to the position of the product state.

There are several residues that interact strongly with the DFG motif. Two of them form the salt bridge that we already discussed. The side chain of Phe401 is highly flexible. It flips between rotational states several times along the reaction coordinate. It is not coupled to the reaction coordinate. Therefore, the sharp spike we observe in Fig. 4.13 has no impact on the progress of the reaction. Another important residue is Met290 that was discussed extensively in reference [121]. Flips of the DFG were observed using conventional MD only after the mutation of the Met290 residue to alanine, reducing the barrier to enter the binding site. To appreciate the coupling between different residues and Phe382 and the order of events during the transition we plot the distance between the Phe382 ring (the CZ atom) and the edges of the side chains of the other residues (Glu286 min(Oe1, OE2), Met290 SD, and Lys271 NZ) in Fig 4.14. The distance between Phe382 and Glu286 shows significant fluctuations between 6 and 8Å but not overall drift. The distance between Phe382 and Lys271 is decreasing near anchors 20 to 32, which is close to the committor value of 0.5 (Fig. 4.7) and hence to a transition state. The motions of Met290 come late in the process and significant displacement is observed after the system is leaving the transition state and continue forward. Hence, the Met 290 transition is a late event in the DFG flip.

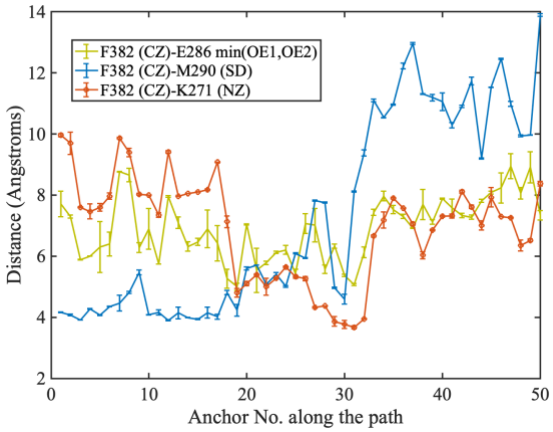


Figure 4.14. The distances between atom CZ of Phe382 to atoms representing the ends of the sidechain of residues Met290, Glu286, and Lys271 are shown. For Glu286, the minimum distance two both oxygens OE1 and OE2 was measured. Lys271 and Glu286 are the salt bridge residues. Around anchor 30, the Phe382 reaches as close as possible to the salt bridge, (Lys 271) breaks it and then goes away from this residue. The distance to Glu286 remains roughly a constant throughout the transition.

4.5 Conclusions

The activation loop (A-loop) in kinases attracted considerable experimental and theoretical attention, being highly flexible and able to adopt multiple conformations, which hinders its accurate computational investigation with commonly used sampling methods. Here we investigated in detail a transition pathway in Abl-kinase from its active to its inactive form. While we investigated the transition of the entire A-loop, which includes multiple rotational events, we focus our analysis on the kinetics of the activated transition of the DFG motif from a DFG-in to a DFG-out state. Quantifying accurately the kinetics of this transition is, however, difficult from both experimental and computational viewpoints. The long transition time (estimated by NMR[118]) makes the direct atomistic simulations of these rare events challenging. Enhanced sampling techniques such as meta dynamics[122] and the string method[128] were applied to the system and were used to primarily investigate its equilibrium conformations. Simulations with Milestoning, [154] suggest that the time scales for the DFG flip are milliseconds to seconds. The extensive information that we gather from the reaction coordinate calculations as well as from the short Milestoning trajectories, allows us to propose a detailed molecular mechanism for the events of the reaction and their coupling to different residues.

We also note that the calculations reported in this paper are conducted with unprotonated Asp381. Previous simulations highlighted the importance of the protonation state of this residue for the conformational transition.[121] Our study will enable future work to examine the impact of the protonation state on the transition pathway.

One should keep in mind the significant conceptual and sampling challenges that calculation of kinetics in the Abl kinase molecular system poses. We cannot be sure that our set of coarse variables is complete, and we are uncertain if all significant transition pathways were sampled. It is expected that the impact of missing alternative pathways will be to reduce the estimated time scale. Nevertheless, the overall agreement of this calculation and experimental estimates is encouraging and it opens

the way for quantitative comparisons between simulations and experimental measurements of long-time events in complex and activated biomolecular systems.

5. Dissociation Mechanism of Gleevec from Abl Kinase using Milestoning[4]

5.1 Overview

Here, I use atomically detailed simulations within the Milestoning framework to study the molecular dissociation mechanism of Gleevec from Abl Kinase. I compute the dissociation free energy profile, the mean first passage time for unbinding, and explore the transition state ensemble of conformations. The milestones form a multidimensional network with average connectivity of about 2.93, which is significantly higher than the connectivity for a one-dimensional reaction coordinate. I examined the transition state conformations using both, the committor and transition function. I show that near the transition state the highly conserved salt bridge of K217 and E286 is transiently broken. Together with the calculated free energy profile, these calculations can advance the understanding of the molecular interaction mechanisms between Gleevec and Abl kinase and play a role in future drug design and optimization studies.

5.2 Introduction

Kinases are a family of enzymes that catalyze the transfer of the γ-phosphate group from ATP to the hydroxyl group of a serine, threonine, or tyrosine residue.[159] They act effectively as switches along cellular transduction pathways due to their ability to alternate between catalytically active and inactive state in response to specific

4. This chapter has been adapted from reference [20].

signals. Hence kinases play an important role in cell growth, proliferation and differentiation.

Uncontrolled division of cells and malignant transformations are a direct consequence of kinase deregulation. A mutation in Abl has been associated with chronic myelogenous leukaemia.[160] The use of small molecular inhibitors for selective inhibition of a kinase is an effective first-line therapeutic method for treatment of several cancers, including leukemia. Finding a specific drug for kinase inhibition is, however, challenging since this protein family serves diverse functions while retaining high structural similarity. Gleevec (a.k.a., imatinib) is a successful drug which is highly specific to the Abl kinase. Despite the ~54% sequence identity between the Abl to the Src kinases, and the presence of highly similar binding pockets in both kinases, Gleevec has almost 3000 times stronger affinity towards Abl.[161] This anomaly has puzzled researchers for many years and a number of computational[162-165] and experimental[116, 118, 127] studies have been made to understand the underlying mechanism. Comprehensive studies by Lin et al.[163, 166] provided significant insight to the thermodynamics of binding to Abl and Src. Emphasis was made on the free energy of Gleevec binding and its dependence on the conformation of the DFG motif. Here, we focus, however, on the unbinding kinetics of Gleevec from the Abl kinase.

Selectivity in biology is sometimes controlled by intermediate states and kinetics, and not only by thermodynamic considerations. The suggestion for kinetic selectivity is, of course, valid in non-equilibrium states, but selectivity by kinetics plays a significant role also in equilibrium. A trivial example is an enzymatic reaction that reduces the barrier for the correct substrate. In more general terms, the binding constant, L, is determined by the ratio of the forward and the backward rates, k_{on}/k_{off}. If the backward rates of the same reaction by two different proteins (or ligands) are different, while the forward rates are the same, the selectivity is determined by the kinetics of the backward reaction. In this case, an alternative and potentially more detailed picture of the selectivity is provided by kinetics.

In the past, Elber et al. illustrated that the protein HIV reverse transcriptase selects a correct substrate by variation in the backward rate of an induced fit transition.[167] Here we simulate the dissociation pathway of Gleevec from Abl kinase

at atomistic detail using the Milestoning method. During the unbinding process, while Gleevec is in the protein matrix, we expect the DFG motif to remain fixed in the DFG-out inactive conformation. Differences in the rates of the unbinding of Gleevec to Abl or to Src kinases in that case are unlikely to originate from the DFG flip. Therefore, this calculation presents a test to the DFG hypothesis and may be able to point out alternative sources of selectivity. For example, formation of a transient barrier.

Kinetic calculations are, however, more computationally expensive than calculations of thermodynamic properties of the binding-unbinding process.[18, 19, 168-172] Due to these limitations, and the special significance of the off-rate discussed in the previous paragraph, in this manuscript we focus only on the dissociation pathway. We compute first passage times as well as free energies for the unbinding pathway of Gleevec from the Abl kinase.

In a recent study we investigated the DFG-in to DFG-out transition in Abl kinase.[18] This prior study motivated the present investigation, exploring further the mode of action of this protein. We emphasize however, that the DFG transition does not play a significant role in Gleevec dissociation as we show in the present paper and was also argued experimentally.[116] Therefore the present paper supplements our previous study and is not a direct extension.

Molecular Dynamics is an increasingly useful tool to study kinetics of complex biological systems at atomistic level. However, the MD integration time step (on the order of 10^{-15} s) is much shorter than observation times in many biological processes and it makes the calculation of events at the milliseconds or longer computationally difficult. As the kinetics of kinases extend to milliseconds and seconds,[116] methods for enhanced sampling of kinetics are desired. One such method that has been designed to bridge this gap is Milestoning, used already extensively to investigate long-time processes.[168]

In the Milestoning approach,[154] the conformational space between the reactant and product is partitioned by a set of dividing hypersurfaces, or milestones. An ensemble of initial conditions is prepared at each milestone. From each of these initial points trajectories are simulated until another milestone is reached. These

trajectories are shorter compared to a reactive trajectory of the overall process (picoseconds versus milliseconds). Moreover, these Milestoning trajectories are easily parallelized. For every short trajectory, we record the starting milestone, the milestone where the trajectory terminates and the time lengths of the trajectories. With this information we construct a transition probability matrix (**K**), where K_{ij} is the probability of transition from milestone *i* to milestone *j*. K_{ij} is estimated from the short trajectories initiated at milestone *i*: $K_{ij} = n_{ij}/n_i$ where n_{ij} is the number of trajectories initiated at milestone *i* that terminate at milestone *j* and n_i is the total number of trajectories launched from milestone *i*. The average lifetime of a milestone *i*, t_i, is the average time length of all the n_i trajectories launched from the milestone *i*: $t_i = \sum_{1,..,n_i} t_{il}/n_i$, where t_{il} is the time length of the *l* trajectory initiated at milestone *i* and terminated at any other milestone. Observables of interest are listed below. The Mean First Passage time (MFPT) (Eq. (1)), the free energy of the trajectories passing milestone *i* (Eq. (2)), the committor (Eq. (3)), and the transition function (Eq. (4)) are given below. The expression for MFPT or $\langle t \rangle$ is

$$< \tau > = p_0(\mathbf{I} - \mathbf{K}')^{-1}\boldsymbol{t} \quad (1)$$

Here p_0 is the initial population, **I** is the identity matrix, **K'** is the transition probability matrix with absorbing boundary at the product and reflecting boundary at the reactant. **t** is the lifetime vector.[67]

The free energy of a trajectory passing milestone *i* is given by

$$F_i = -k_B T \log(q_i t_i) \quad (2)$$

Here q_i is the steady state flux at milestone *i*, which is obtained from the linear equation

$q^t \boldsymbol{K} = q^t$ and t_i is the lifetime of milestone i.[67]

The committor value can be calculated as

$$\boldsymbol{C} = \lim_{n\to\infty} (\mathbf{K}^c)^n \boldsymbol{e_p} \quad (3)$$

Here $\mathbf{K}^c$ is the adjusted transition probability matrix with terminating boundaries at the reactant (r) (every trajectory that reaches the reactant disappears) and absorbing boundaries at the product (p) (every trajectory that reaches the product remains there), $K_{rr}=0$ & $K_{ri}=0$; $K_{pp}=1$ & $K_{pi}=0$; $\boldsymbol{e_p} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{1})$.[173] At the transition state, $\boldsymbol{C_{TSE}}$=0.5.

Finally, as an alternative estimate for the transition state conformations, we also use the following expression for the transition function, T_i^e

$$T_i^e = \log[\tau_{i\to P}^e / \tau_{i\to R}^e] \quad (4)$$

here $\tau_{i\to P}^e$ is the exit time from milestone *i* to the product and $\tau_{i\to R}^e$ is the exit time to the reactant.[174] The exit times are computed from any milestone in reaction space to terminating boundaries at the product or the reactant. For conformational states in the transition state ensemble (TSE), $T_{TSE}^e \to 0$.

5.3 Method

The crystal structure 2HYY[175] of human Abl kinase with the drug Gleevec bound was used to initiate the present study. The structure was solvated with TIP3P water molecules. Ions were added to obtain 0.15 M NaCl solution. The system was minimized for 10000 steps, followed by an NPT run for 5ns and an NVT run for 20ns using periodic boundary conditions at 300K. The integration timestep was 1fs. All simulations have been done using the program NAMD[134] with the CHARMM36 forcefield.[135]

To initiate a Milestoning calculation we need a rough sampling of the space between the reactant and product. This sample is provided here by a pulling trajectory. Seventy-six nanoseconds of constant-velocity Steered Molecular Dynamics (SMD) simulation pulled the center of mass distance of the ligand out of the final structure of the NVT run. The distance was the only restraint and therefore the bias was not directional. Three C_α protein atoms of residues GLU 238, ASP 325, and Pro 465 were constrained to fix the overall translations and rotations of the protein. The distance between the initial and final centers of mass of the Gleevec is 38Å (Figure 5.1 and 5.2). The constant velocity pulling was slow enough to allow the system to relax to local equilibrium and avoid significant stresses due to the external pulling force. In Fig. 5.1, we show the potential energy during the steering simulation to illustrate that the pulling perturbation is small. The average and the fluctuations of the potential energy are

roughly constant during the entire trajectory. Note, however, that the distance to the binding pocket is growing non-linearly in time, indicating that the trajectory is strongly influenced by the specific molecular forces (as compared to the external pulling).

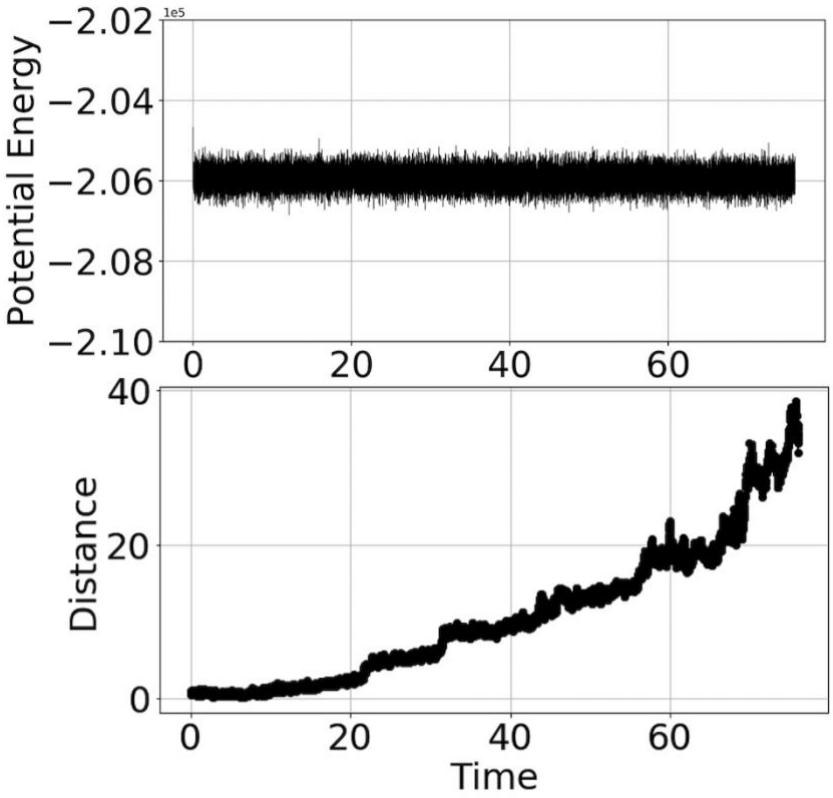


Figure 5.1. The potential energy (kcal/mol) and the distance of Gleevec from the binding pocket as a function of pulling time (ns). Gleevec is pulled out of the binding pocket at a small constant velocity (0.5 Å/ns) to reduce system strain, providing a flat sampling of the underlying potential energy (top panel). The lower panel shows the center of mass distance (in Å, see main text for the definition) as the function of time (ns). 43 configurations are selected from this SMD trajectory to serve as anchors in the Milestoning calculations.

From the SMD trajectory, 43 configurations, including reactant and product, were selected as anchors for the Milestoning run. These configurations were selected such that (i) they are separated sequentially by ~0.1 Å and (ii) that the Gleevec's center of mass distance from the binding pocket is monotonically increasing.

Gleevec is bound to the kinase in the first anchor, and the last anchor represents the unbound state with Gleevec entirely exposed to the aqueous solution (Figure 5.2). In Milestoning the anchors form centers of Voronoi cells. The interface between the

Voronoi cells (say cells *i* and *j*) is called a milestone - M_{ij}. Milestone, M_{ij}, is the set of points with equal distances to anchors *i* and *j* and larger distances from all the other anchors. To commence, we defined initially 42 milestones, each between two consecutive anchors (Fig. 5.3).

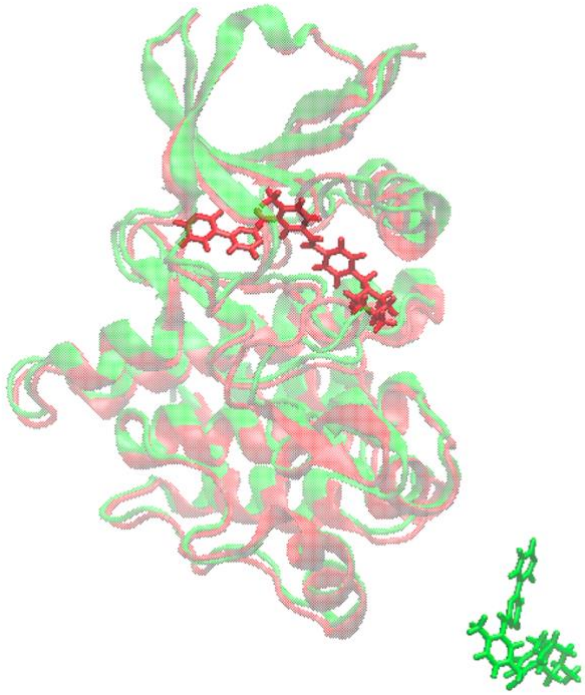


Figure 5.2. Abl kinase in complex with Gleevec. Abl kinase with Gleevec bound (reactant) is shown in red. Abl kinase with unbound Gleevec (product) is shown in green. Gleevec has been highlighted by representing it as opaque and using a translucent/shaded texture for the kinase matrix.

In the following step, we sample configurations on the milestones using 1ns Molecular Dynamics simulations. The trajectories were restrained to the milestone. We use two restraints. The first restraint is a harmonic term $k(d_i - d_j)^2 = 0$ to keep equal the distances d_i and d_j (in coarse space) between the current configuration and the two anchors *i* and *j*. Here, d_i is the difference between the center of mass distance of Gleevec at current configuration and the center of mass distance of Gleevec at anchor *i*. The restraining force constant is 1500 (kcal/mol)/Å^2. A second set of half-harmonic restraints are $k'(d_m-d_l)^2$ when $d_m<d_l$, $l=i,j$ and zero otherwise with $k'=1000$ and *m* being any milestone other than *i* or *j*.

100 samples were kept from the final 0.5ns of the restrained simulation at each milestone.

Then all restraints were removed, and unbiased MD trajectories were initiated from configurations sampled on each milestone. The trajectories terminate when they reach a milestone other than the one that they were initiated on. The average time length of these trajectories is about 60 ps, a time easily accessible by conventional MD.

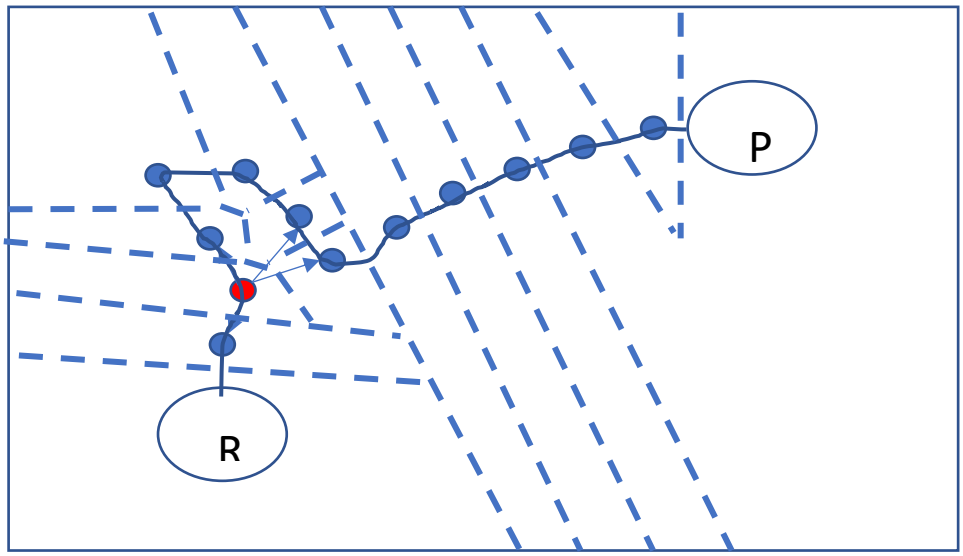


Figure 5.3. Schematic representation of the reaction space in two dimensions. The axes represent coarse variables. The unfilled circles represent reactant (R) and product (P) conformations. The filled blue circles are the anchors, obtained by exploratory Steered Molecular Dynamics calculation, and serving as centers of Voronoi cells. The dashed lines are the milestones or the boundaries of the Voronoi cells. Note that anchors placed on a straight reaction pathway segment appear connected to only two neighboring milestones, forward and backwards. However, in more dimensions, a milestone can be connected to more than two milestones.

If a trajectory is terminated on a milestone that was not sampled before, we add it to the list, sample at the new milestones and launch a new set of trajectories. This process of launching new trajectories and adding to the list newly discovered milestones is repeated until we obtain a connected transition matrix and a finite MFPT. An alternative is to continue the iterations until no new milestone is found. The last process is considerably more expensive and was not used in the present study. We started with a set of 42 milestones. After sampling the first set of milestones, we

discovered 84 new milestones. In the next round, we sampled these 84 new milestones. Thus, in total we sampled 126 milestones.

The overall computational cost for a system of this complexity is relatively modest. In summary, the system was minimized for 10,000 steps, followed by an NPT run for 5 ns and an NVT run for 20 ns. SMD pulling was completed in 76 ns. Sampling on all the milestones required 126 ns (1 ns per milestone). Total simulation time for all the free trajectories from all the milestones is 0.816 μs. Thus, a total simulation time of 1.043 μs MD was required to complete the calculations. We also note that the unbiased trajectories are independent and are trivially parallelized.

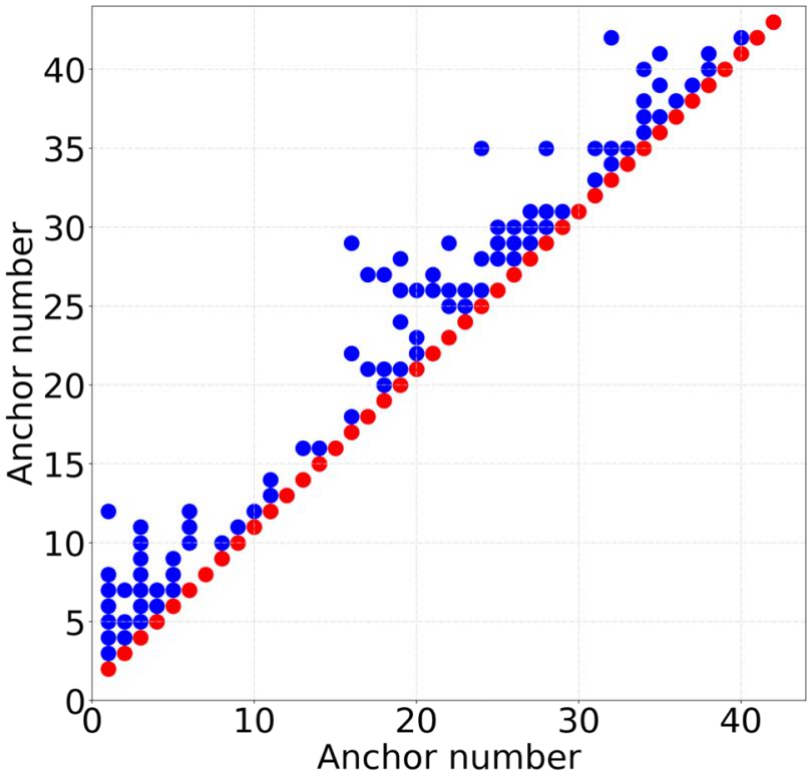


Figure 5.4. Representation of the 126 milestones used in the present study. The axes are the anchor indices. A milestone between anchors *i* and *j* is represented by the point (*i, j*). The red dots are the initial milestones between the consecutive anchors along the transition path. The blue dots represent the 84 new milestones discovered during the analysis of the free trajectories from the initial set of only 42 milestones.

5.4 Results

Fig. 5.4 suggests that the reaction is better described with a network than in one dimension. In Fig. 5.5 we show a typical graph representation of the milestones' space. If the system was exactly one dimensional then the number of edges of a node (also called the degree of a node) would be exactly 2, excluding the nodes at the boundaries. If the system is embedded in a square two-dimensional lattice, we expect four edges per node. The average of the degrees of all the nodes in Fig. 5.5 is about 3. The reaction space dimensionality is, therefore, between one and two if we use the square lattice example.

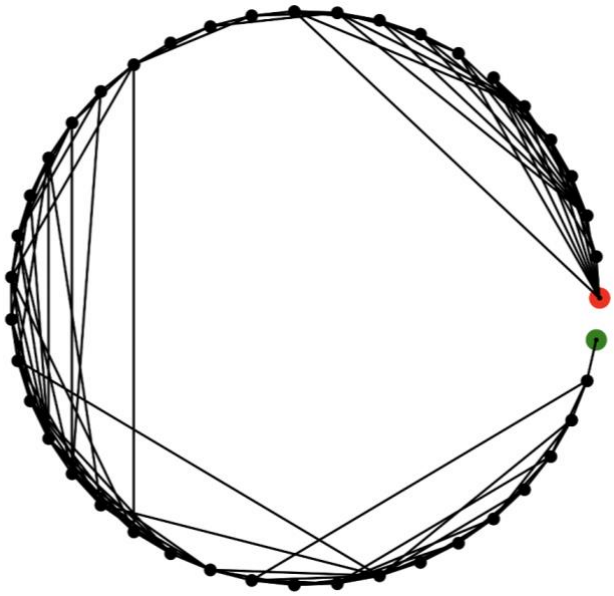


Figure 5.5. Network representation of the anchors' space. Each node represents an anchor, there are 43 anchors in total. The first node represents the reactant (1st anchor, shown in red). The last node represents the product (43rd anchor, is shown in green). A connection between any two anchors is represented by a straight black line. There are 126 connections (or milestones) in total.

The free energy landscape of all the milestones is shown in Fig. 5.6. The two axes are the anchor indices, and the free energy of the milestone is color coded (Eq. 2). Hence the free energy landscape is illustrated in two dimensions. The zero is taken to be the lowest energy value. The diagonal lattice points are not defined since an anchor cannot connect directly to itself. Therefore, the diagonal elements are colored

with the highest energy values. There are two domains that have significant width near the diagonal. One domain is near the minimum (anchor 1-8) and the second is between 19-25. The bound Gleevec is in a deep minimum, but the location of the transition state and its characteristics are difficult to grasp. We therefore conduct further analyses as described below.

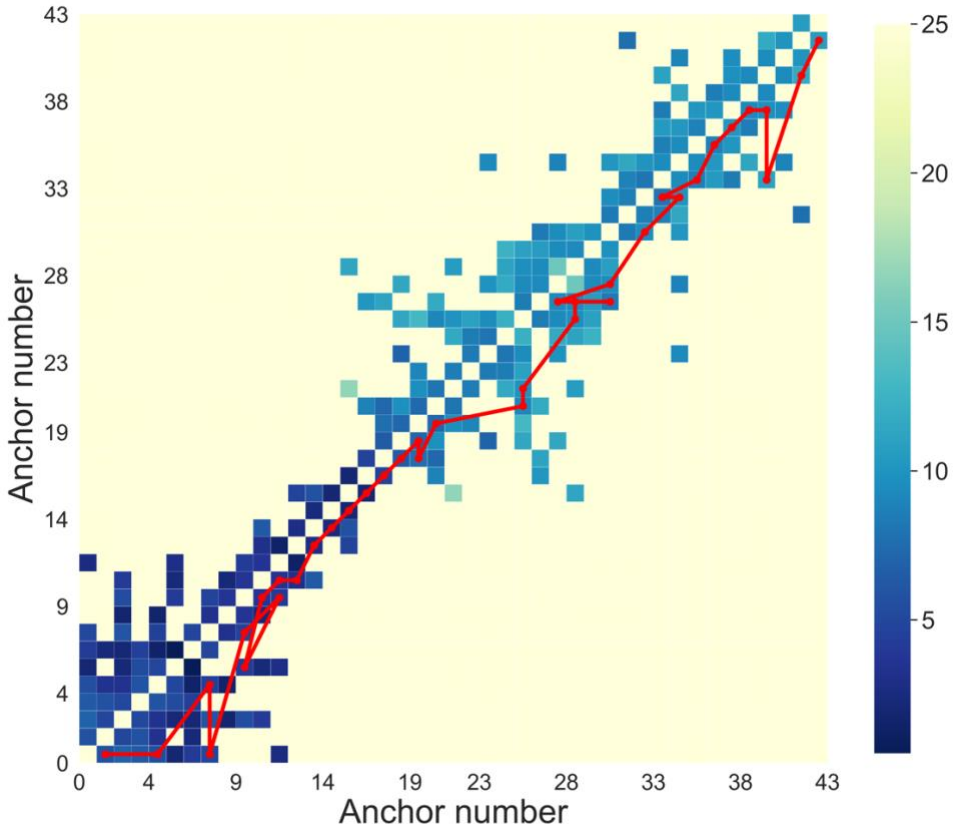


Figure 5.6. Free energy plot for 126 milestones. The free energy of every milestone is colored according to its numerical value in kcal/mol (see color-bar). The maximum flux (Max Flux) pathway is shown in red (see text for details).

Identifying a significant and representative one-dimensional reaction coordinate in a larger reaction space is a useful analysis tool and a common practice. One possible definition of a reaction coordinate is the Max Flux Pathway (MFP).[176] We have used the MFP to estimate important coordinates in the milestone space, represented on a network.[156] To determine the MFP we need to assign weights to different edges in a network. We consider a network in which every milestone is a node (note that the milestones' network is different from the anchors' network of Fig. 5.5).

We assign weights to the network's edges as the rate coefficients for transitions between milestones.

Rate coefficients are computed from the transition matrix **K** and average lifetimes of the milestones, **t**. The rate coefficient for a transition between milestones *i* and *j* is given by K_{ij} / t_i.[177] We used these rate coefficients to obtain the global maximum weight path (GMW)[156] from the Gleevec bound state (reactant) to the Gleevec unbound state (product). To compute the GMW a recursive Dijkstra's algorithm was used, which has been discussed in the context of Milestoning in Refs. [156, 178, 179].

The free energies of the milestones, on the GMW path (Eq. (2)), are shown in Figure 5.6. The free energies of the reactant and product are 6.05 ± 0.91 and 10.99 ± 0.94 kcal/mol, respectively. Clearly, the reactant and the nearby metastable state are favored over other intermediate configurations, that is, the inactive conformation of Abl kinase with Gleevec bound is more stable with respect to other conformations in the reaction space. Note, however, that we do not have a reliable estimate of the free energy of the unbound Gleevec that requires an estimate of the entropy in the aqueous solution. As stated in the introduction, we focus on the off rate. A combination of a reduced model and atomically detailed Milestoning can be used to provide this estimate.[1] After the initial displacement of Gleevec, from the binding site of the crystal structure, the free energy drops by about 4 kcal/mol with a minimum near milestone 12 (Figure 5.7).

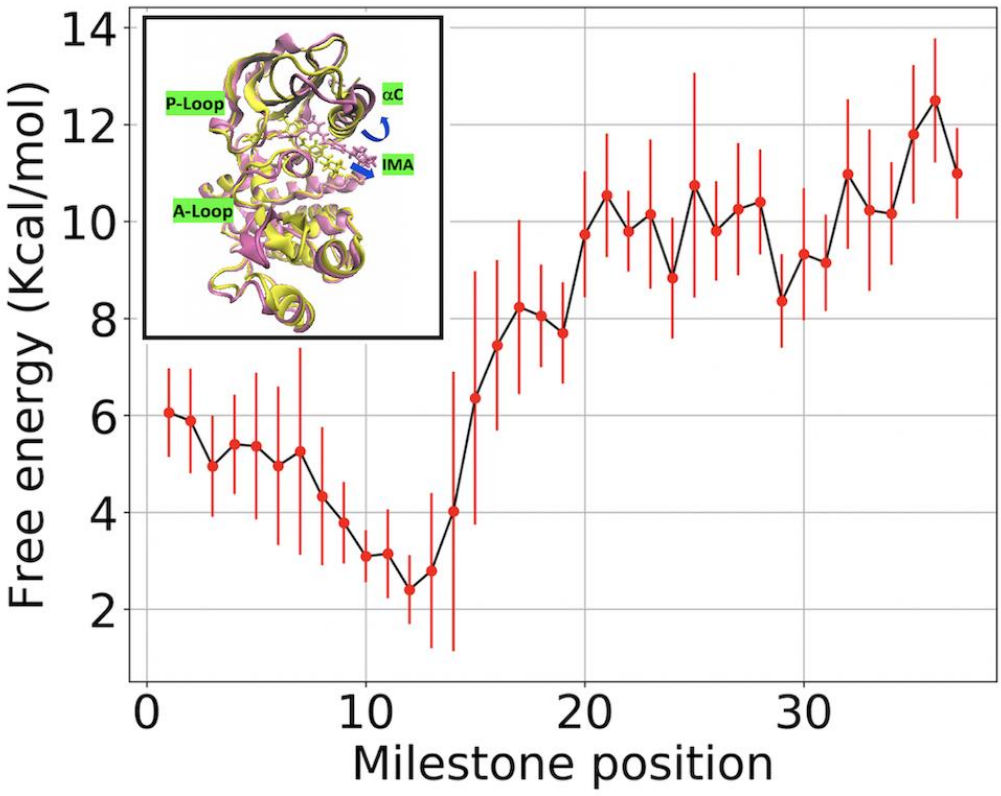


Figure 5.7. Free energy profile (kcal/mol) along the maximum weight path. To estimate the mean values (dots, red) and the errors (standard deviation, red vertical lines) we sampled the transition matrices and lifetimes from their Milestoning model distribution, using a set of 1000 samples. The committor-estimated transition state (TS) appears to be located late and broad between milestone 20 and 30 along the reaction coordinate. The energy minimum for Gleevec unbinding is seen near milestone 12. Structural differences between the reactant (shown in yellow) and the structure at milestone 12 (shown in pink) have been illustrated in the inset figure. At this minimum, we observe the outward displacement of Gleevec (IMA), show with blue arrow. We also observe an outward rotation in the αC helix (see inset blue arrow). There is an RMSD difference of 2.6Å. The center of mass distance of Gleevec between the two structures is ~4.5Å.

Another approach to estimate reaction coordination uses a manifold of hyper surfaces. The committor function, C_i, at a milestone *i*, is the probability that a complete trajectory initiated at that milestone will reach first the product before the reactant. With the transition matrix **K** at hand the committor function is computed in a closed form (Eq. (3)) with no need to run additional trajectories.[20] The committor is a function of all coarse variables and the hypersurface with the same value of *C* is the reaction coordinate. The transition state (hypersurface) is when *C* is equal 0.5. Milestones between the anchor pairs (27,28), (27,30), (28,30), (29,30) and (29,31) have a committor value close to 0.5 (Fig. 5.8). A representative structure from the ensemble

of trajectories initiated from a milestone with a value of C ~ 0.5, is shown in Fig. 5.9. The location of the iso-committor surface was further tested by generating an additional set of unbiased trajectories initialized at the transition state of the ABl-Gleevec complex. We label this transition state conformations as transition state 1 (TS1).

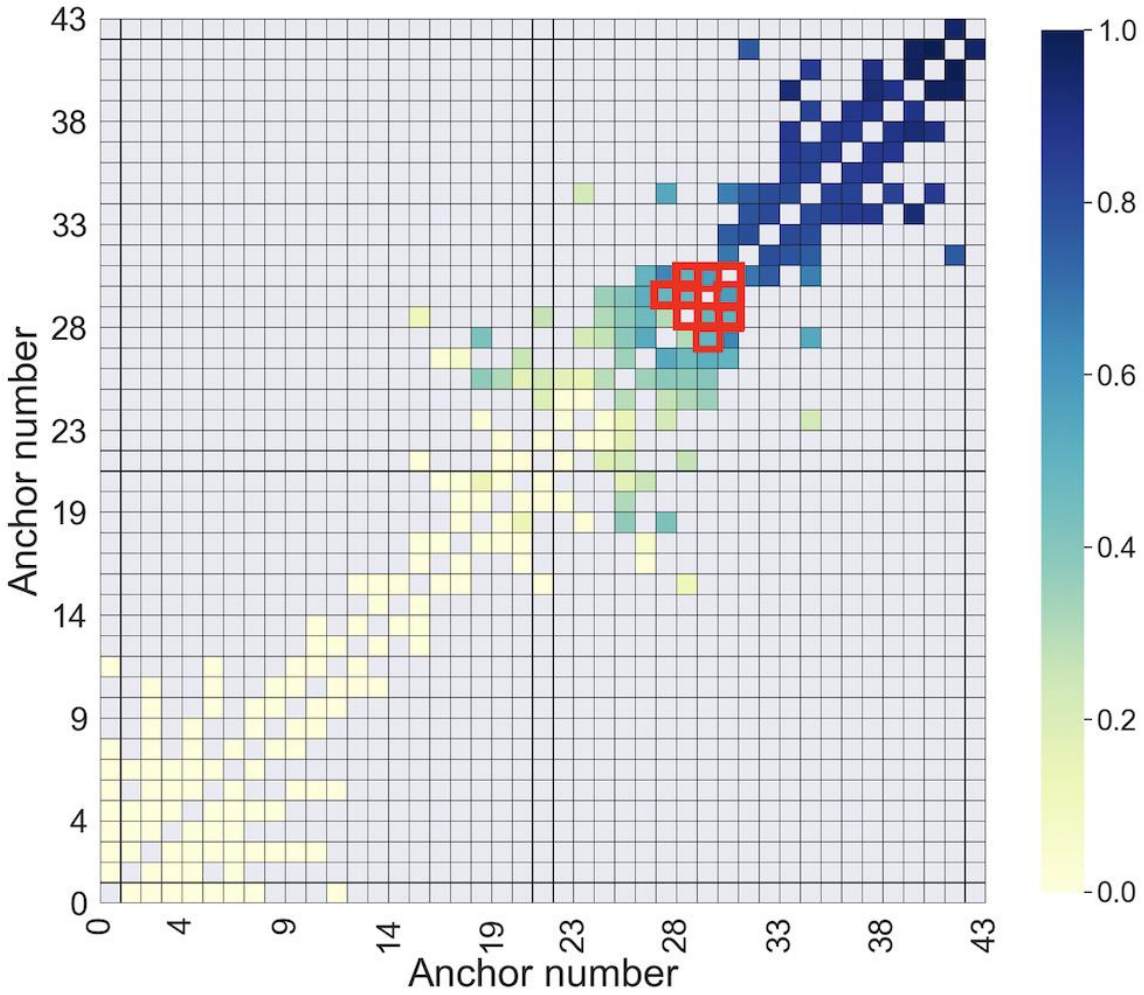


Figure 5.8. Color-coded committor function at each milestone. The committor function, C_i, is defined as the probability that a trajectory initiated at milestone *i* will reach the product before the reactant. Milestones with committor values close to 0.5 are candidates for the transition state and have been highlighted with red squares.

Testing the iso-committor surface predicted by Milestoning. We initiated unbiased trajectories starting from the transitions state ensemble in order to test how many of these trajectories made progress towards the reactant versus the product. 10 unbiased trajectories, 4ns-long each, were initiated from the TS1. As shown in Table 5.1, we observed 4 trajectories going towards the binding pocket (inbound) and 6 trajectories

moving away from the binding pocket. Thus, approximately 40% of the trajectories moved towards the binding pocket. Anchor 1 is the bound state.

INBOUND		
	Starting Anchor	Final Anchor
1	27	16
2	27	19
3	27	22
4	27	25
OUTBOUND		
1	27	29
2	27	43
3	27	31
4	27	30
5	27	43
6	27	32

Table 5.1. Starting and final anchors for 10 unbiased test MD trajectories launched from the TS1 conformation.

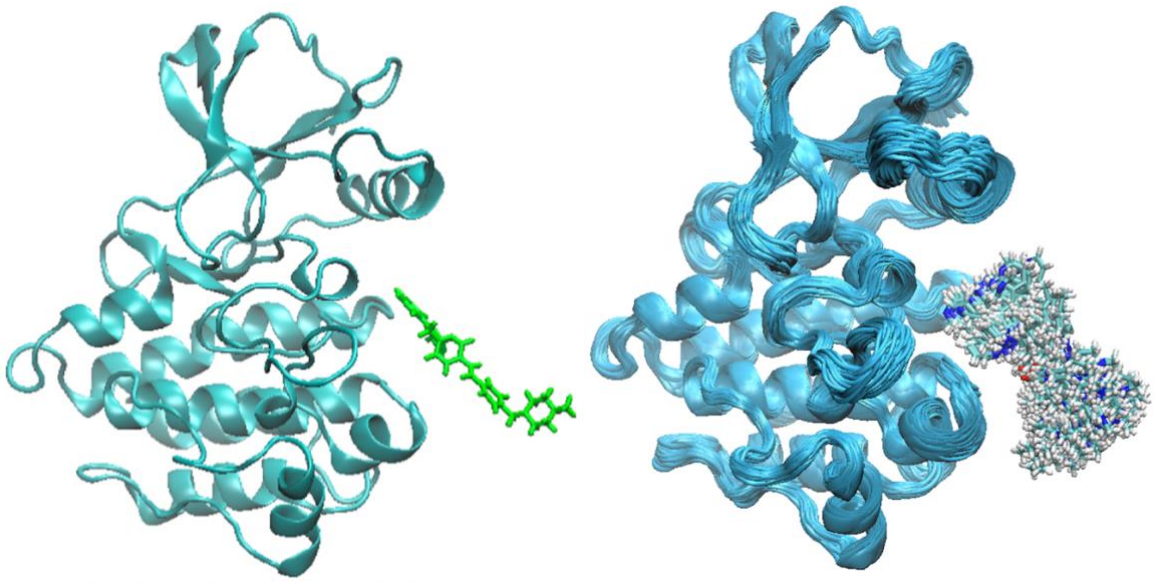


Figure 5.9. Representative structure from the Transition State Ensemble (TSE) estimated using the committor function (called TS-1). The image was generated using the VMD software.[32] At this position, the probability to return to the bound state is equal to the probability of escaping the protein to the aqueous solution.

Another useful approach to define a transition state is by using the transition function (Eq. (4)).[174] The transition state estimated using the committor function is based on probability and is time independent. In contrast the transition state of the transition function is based on escape times. The transition state is defined as the set of points that have the same exit times to the reactant and the product states. The transition function (Eq. (4)) is zero at the transition state. Milestones (14,15), (15,16), (16,17), (16,18) and (17,18) have transition function value close to zero, that is, $T^{e}_{i} \approx 0$, see Figure 5.10. A representative structure for transition state estimated from transition function is shown in Figure 5.11. This candidate of transition state is referred to as transition state 2. The striking difference between transition state 1 and 2 is a result of the high asymmetry of the transition state with respect to the reactant and product.

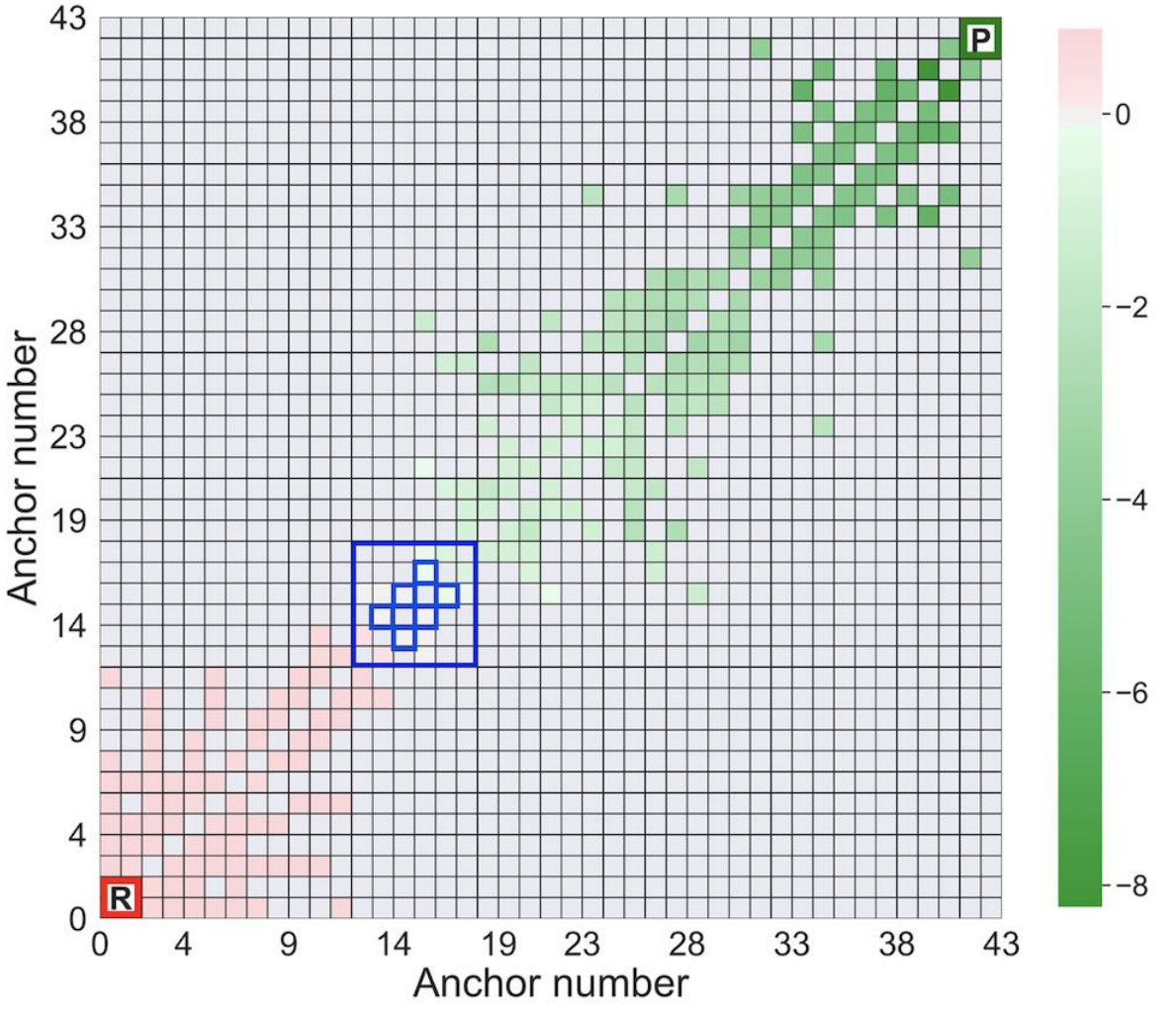


Figure 5.10: Transition function (defined as the logarithm of the ratio of the exit times towards the product and the reactant (Eq. (4)).[174] Milestones with similar exit times to both the product and the reactant, are close to the transition state. The region highlighted by the blue contains milestones (marked with small blue squares) near the transition state with the transition function close to zero. The R and P states are located inside the red and the green boxes, respectively, representing reactant and products.

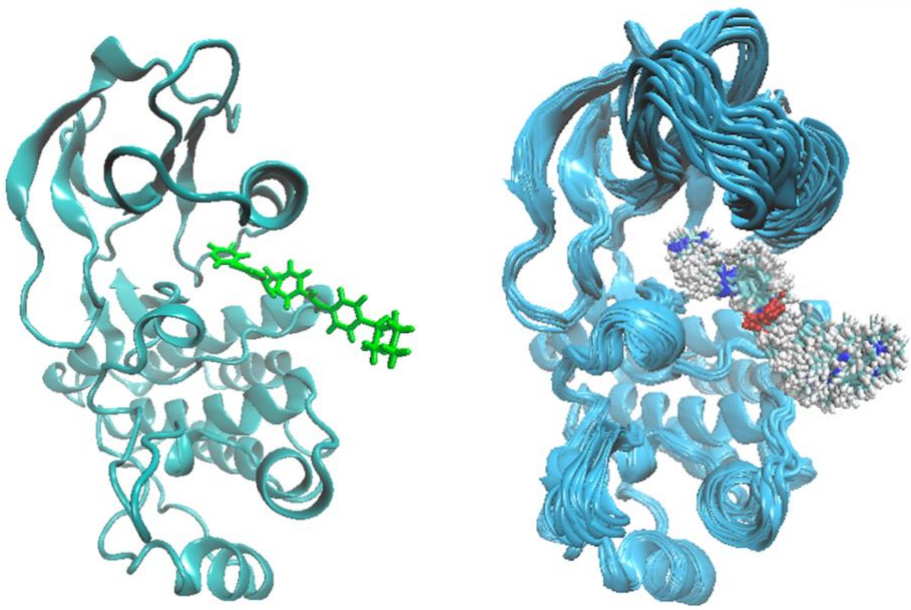


Figure 5.11. A representative TSE structure found using the transition function (Eq. 4). By construction, the transition function value is ~0 for TSE conformations, as the exit times to the product and reactant are equal. The image was generated using VMD.[1]

During the process of Gleevec unbinding, we observe several conformational changes in Abl Kinase close and far from the binding pocket (P-loop). For example, close to binding pocket, the distance between GLU282-LYS274 is reduced from ~17Å to only ~2Å (Fig. 5.12); far from the binding pocket (P-loop) the distance between GLU466-ARG460 increases from ~3Å to ~8Å and then decreases again to ~3Å.

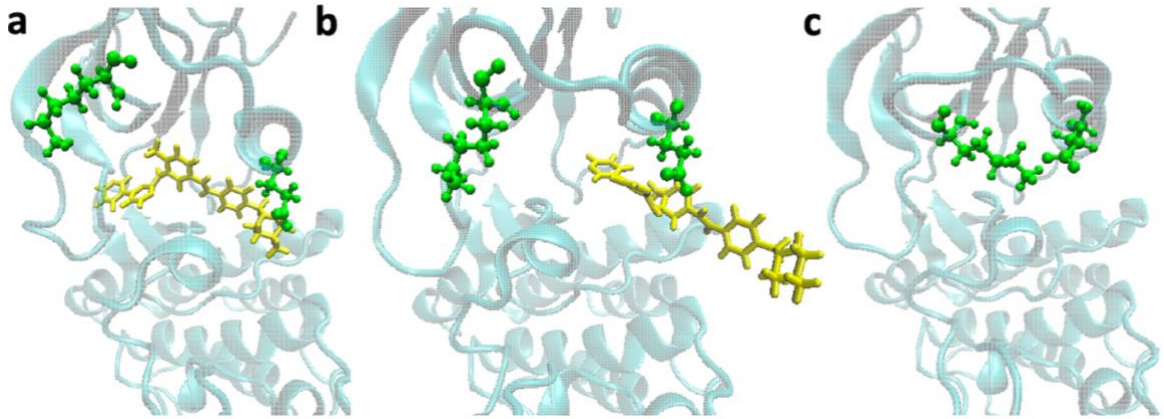


Figure 5.12. Conformational changes along the Gleevec dissociation reaction pathway. GLU282-LYS274 is shown in green and Gleevec is shown in yellow. Gleevec, when inside the binding pocket blocks the direct interaction between GLU282 and LYS274, shown in (a). Panel (b) represents the configuration at the transition state 2. (c) Finally, when the Gleevec molecule is out of the kinase matrix, the distance reduces to ~2Å.

It is interesting to compare the pathways for Gleevec's escape in Abl and Src kinase. We search for critical residues along the pathway of Gleevec escape which are significantly different in Src kinase and therefore may contribute to alternative binding properties. The Abl and Src sequences were aligned using the Smith-Waterman algorithm[180] and BLOSUM62 substitution matrix.[181] In Fig. 5.13 we show the Abl kinase sequence compared to corresponding Src sequence (presenting 50.6% sequence identity and 69.0% sequence similarity).

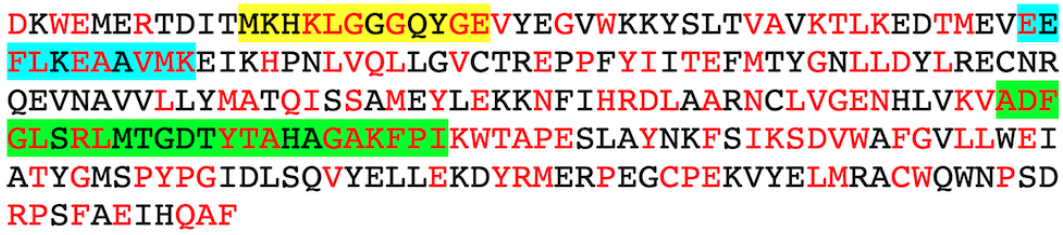


Figure 5.13. The Abl kinase sequence - comparison with the corresponding residues from Src. The two kinase sequences have 50.6% sequence identity and 69.0% sequence similarity. Identical residues are colored in red. P-loop, αC helix and A-loop are highlighted in yellow, cyan and green, respectively.

Four salt-bridges that are not conserved and eight conserved salt-bridges were identified. The spatial changes of the non-conserved salt-bridges along the GMW path is in Fig. 5.14 in the top insets, whiles in lower insets of each part list the corresponding mutations in Src kinase. These residues illustrate significant variations along the pathway and are therefore likely to impact the kinetics.

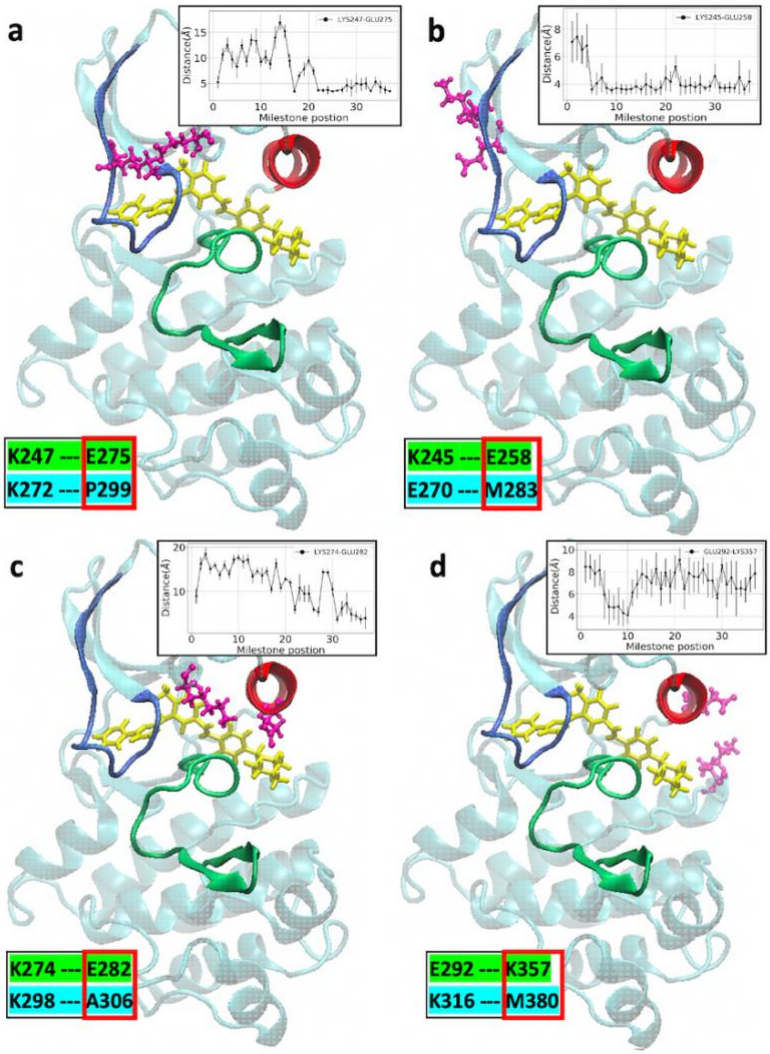


Figure 5.14. Significant differences in salt bridge interactions are observed between Abl and Src kinases. We show salt bridges that are formed in the Abl kinase, making significant contribution to the reaction pathway, and are modified in Src kinase. P-Loop, A-Loop, αC helix, salt bridges and Gleevec are shown in blue, green, red, magenta, and yellow, respectively. Inset graphs show the changes in the salt bridge distance as a function of milestone positions along GMW path. At the bottom left of each panel are shown the salt-bridges for the Abl-kinase case (highlighted with green), and the corresponding residues for Src-kinase (cyan). The corresponding Abl and Src residues that are not similar have been boxed in red.

One of the highly conserved salt-bridge identified is LYS271-GLU286. The LYS271-GLU286 salt-bridge breaks close to the 2nd transition state and forms again, as a result of the change in alpha helix (Fig. 5.15). The αC helix goes from inward rotated conformation to outward rotated conformation and finally back to inward rotated conformation. The distance between the two residues increases from ~3Å (Fig. 5.15a)

to ~7.8Å (Fig. 5.15b). The bond is formed again when Gleevec is completely out of the kinase (Fig. 5.15c).

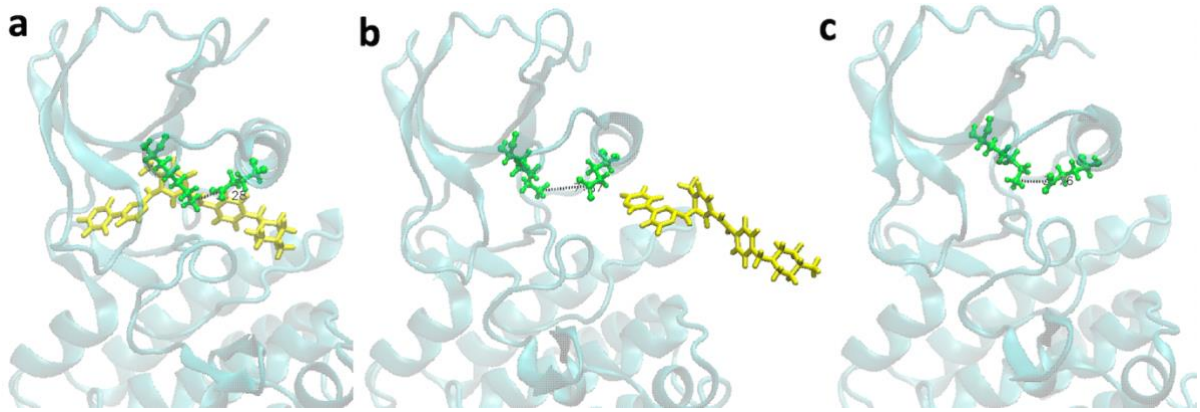


Figure 5.15. Highly conserved LYS271-GLU286 salt-bridge. For the kinase to be active, DFG needs to be in the 'in' conformation, LYS271-GLU286 salt bridge should be formed, the catalytic spine that involves the residues Asp421, His361, Phe382, Met290, and Leu301 needs to be formed, and the binding site should be accessible to ATP. Thus, the integrity of the LYS271-GLU286 salt bridge is central to kinase activity. Shown above is the LYS271-GLU286 salt bridge in green, Gleevec (yellow), and the kinase (cyan). During the unbinding of Gleevec, this salt bridge breaks near the transition state 2 (panel b). The distance between the two residues increases from ~3Å (panel a) to ~7.8Å (panel b). The bond is formed again when Gleevec is completely out of the kinase matrix (panel c).

Along the GMW path, the DFG motif remains in out-conformation. This suggests that Gleveec inhibits the activity of Abl kinase even after getting displaced from the initial binding position. Inside the kinase matrix, the average end-to-end distance (d_{EE}), that is the distance between C8 and C37 atom of Gleevec, is ~20Å. A cis-type conformation, with d_{EE} of ~16Å, is observed outside the kinase. Significant structural fluctuations of Gleevec are observed after transition state 2. The dihedral angle as a function of the center of mass distance of Gleevec and the binding pocket (which is equivalent to the dihedral angle as the function of milestone position) shows, not surprisingly, an increase in the flexibility of Gleevec outside of the kinase (shown with red bars in Fig. 5.16).

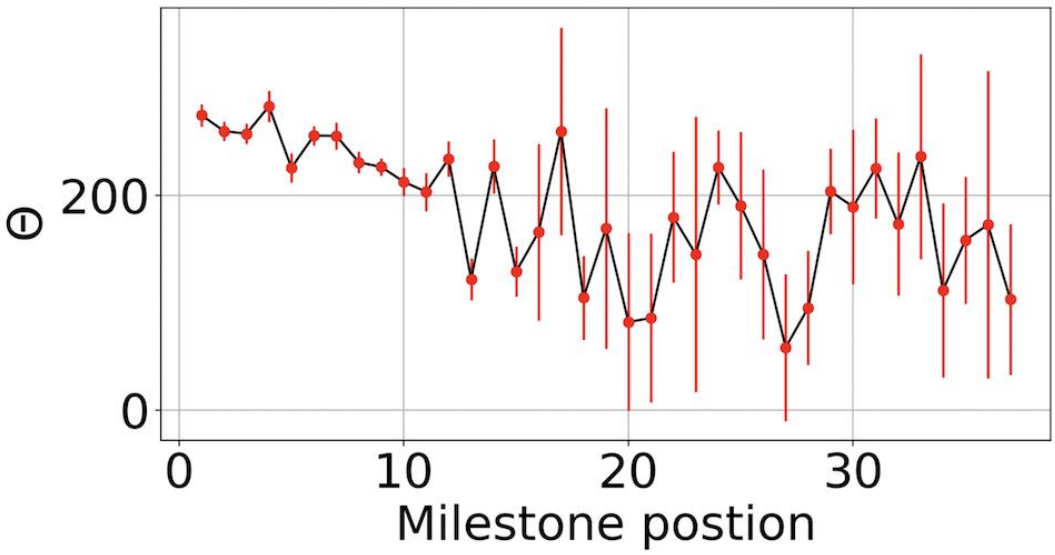


Figure 5.16. Dihedral angles along the GMW path. As Gleevec moves away from the binding pocket to slide out of the Abl kinase, the steric hinderance decreases, and an increase in the movement and rotation is observed. The dihedral angle, Θ, defined by the C8, C15, C32 and C37 atoms of Gleevec, was recorded for the milestones along the GMW path from reactant to product. Larger ranges of dihedral angles at milestones outside the binding pocket suggests greater flexibility and entropy. The onset of larger flexibility is near transition state 2.

We compute the Mean First Passage Time (MFPT), the residence time in the binding site, or the inverse of the off rate $1/k_{off}$, using Eq. (1) and the data from the Milestoning runs. To estimate the mean and the errors, we sampled transition matrices and lifetimes from their model distribution.[158] The distribution of MFPTs from a set of 1000 samples is shown in Fig. 5.17. The averaged MFPT is 0.055s.

Agafonov et al. in their experimental studies[116] suggest the process of Gleevec binding and unbinding to be a two-step process.

$$E_{out} + I \Leftrightarrow E_{out}.I \Leftrightarrow E^{*}_{out}.I \quad (5)$$

The first step is the binding which is relatively fast. The second step, which is slower, has been attributed to either being an induced fit step, or to the presence of a significant number of binding intermediate states along the transition patheway.[116, 182] Here we are studying the unbinding process to unveil mechanistic details about these kinetically different steps. To find the experimental residence time, we use the k_{off} reported by Agafonov et al. (Table S5.1 in the supplementary information) which is $1/k_{off} = 1/(25 \pm 6) = 0.04 \pm 0.01$ s. Our results are consistent with the experimental

value. To further assess the time scale, we also evaluated the kinetics in the mutant Y253F[182] and found it to be similar to the wild type.

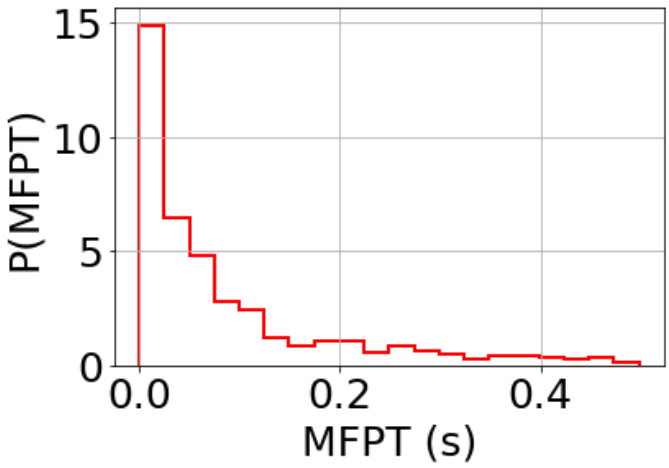


Figure 5.17: Distribution of first passage times for Gleevec dissociation from the Abl kinase. The corresponding MFPT value derived from Milestoning calculations is 0.055 s. Note the broad distribution of predicted MFPT values suggesting significant uncertainties in the calculations.

After the simulation work of the present study was essentially complete, a related manuscript was published.[183] This study used multiple long molecular dynamics trajectories in conjunction with Markovian modelling to depict the mechanism of Gleevec escape from the protein matrix. Their approach generates and examines a broader range of pathways than in our case. [183] Here, we have used a single SMD pathway as a starting point, due to the focus on estimating kinetics with Milestoning, while they used several. Their study is therefore more comprehensive, and the sampling of alternative pathways is more complete. However, our statistics for transition along the single pathway is larger and was obtained at lower computational cost. Here, we choose to focus on a single pathway determined by unbiased pulling which is also one of the main pathways examined in their studies (below the C helix), and we provide detailed kinetics and analysis of the underlying reaction mechanism along this pathway.

The current study has a number of new contributions. Our analysis of the kinetics is based on the Milestoning approach, which is an alternative, computationally efficient, procedure to study slow kinetics. The time courses in Milestoning do not have a time lag like in the Markov State Model. Since the Milestoning simulations rely on

short trajectories they are considerably more efficient than long trajectories. Our statistics for the non-directional path are better than what one can obtain from a few long trajectories and are obtained at less computational resources. With a total simulation of only 1.043 μs, we are able to estimate an exit-time on the order of tens of ms. The pathway under the C helix, and the overall time scale of milliseconds agrees quite well with the Markov modelling-based investigations. Also, the observation of long-lived intermediate is consistent with the prediction of the Milestoning calculations of a broad intermediate free energy minimum.

Our analysis of the committor and transition functions adds additional insight to a particular important dissociation pathway that does not involve the DFG motif directly and opens the possibility of comparison with drug interactions in other kinases, such as the Src (Fig. 5.13). Our Milestoning-based study allows the efficient identification and analysis of TSE conformations, as well as the calculation of detailed kinetic properties along the dissociation pathway, unveiling the underlying distribution of first passage times (Fig. 5.17).

5.5 Conclusions

Using the Milestoning method,[154, 168] we computed detailed kinetic and thermodynamic information on the molecular dissociation of Gleevec from the Abl kinase protein matrix. Building on previous work,[18, 168] Milestoning simulations allow us to focus on calculating slow kinetic timescales (i.e., associated slow processes on the order of tens to hundreds of milliseconds) from atomistic MD trajectories, while sampling the unbinding pathway of Gleevec from Abl, currently beyond the reach of conventional MD-based simulations. [18, 19, 168] We focus on the off rate as a critical measure of drug activity. Our study unveils significant insight into the dissociation kinetics. Interestingly, we found that the transition state ensemble appears to be late and broad according to the committor function approach, while being located rather early (i.e., closer to the Gleevec bound “reactant” state) according to the transition function (i.e., equal escape times towards both ends of the reaction pathway).

The reaction mechanism shows that Gleevec slides under the αC helix during dissociation. Subsequently, critical salt bridges are broken along the pathway in dissociation intermediates and reform once Gleevec has exited the protein matrix (Fig. 5.14). The LYS271-GLU286 salt bridge breaks at the transition state 2 (Fig. 5.11). During dissociation, Gleevec is held tightly in the binding pocket but as it is starting to emerge from it, its flexibility significantly increases. The entire pathway is conducted with the DFG motif in the DFG-out, inactive conformation. This observation provides the opportunity to examine kinetic effects of selectivity. Our results suggest that future investigations should study the details of the corresponding Gleevec escape pathway in different cases, such as the Src kinase, which presents significant sequence identity and similarity with Abl (Fig. 5.13). These calculations, conducted with a DFG-out conformation, can serve as a further test for the impact of the DFG flip on the selectivity for Gleevec binding.

Our calculations indicate that Milestoning can play a central role in studies that facilitate the rational design of specific kinase inhibitors, by unveiling both kinetic and thermodynamic details of the rather complex kinase-drug molecular interactions. This is expected to be particularly important for relatively large and complex drug molecules, such as Gleevec, that can adopt several relatively long-lived distinct intermediate conformations in their TSE, in agreement with other recent experimental and computational studies.[18, 116, 182, 183]

6. K-Ras4B GTP-dependent activation/inactivation mechanistic reaction coordinates*

6.1 Overview

I probe the equilibrium conformations adopted by GTP-bound K-Ras4B proteins using long-time atomistic molecular dynamics (MD) simulations. I analyse the underlying free energy landscape of wildtype K-Ras4B projected on two important distances, labelled d_1 and d_2 (i.e., coordinating the P_β atom of the GTP ligand with two key residues, T35 and G60), that are useful reaction coordinates for discussing the K-Ras4B activation/inactivation mechanism. However, the detailed inspection of the K-Ras4B conformational landscape reveals a more complex network of underlying equilibrium states. I show that including a new reaction coordinate to account for the orientation of acidic K-Ras4B sidechains such as D38, with respect to the interface with binding effectors such as RAF1, is needed to rationalize the activation/inactivation propensities. I also show that a relatively minor mutation, D33E, in the switch 1 region can lead to significantly different activation propensities of monomeric K-Ras4B. This study shades new light on the role of residues located at the K-Ras4B–RAF1 interface on its underlying GTP-dependent activation/inactivation mechanism.

6.2 Introduction

Kristen rat sarcoma viral oncogene homologue (K-Ras), known as an oncogene for almost four decades,[184] is a small GTPase that controls cellular proliferation by playing an important role in many cancer signal transduction pathways.[8] It acts as a molecular switch, flipping between an inactive guanine diphosphate (GDP) bound

* This chapter has been adapted from submitted work; Narayan B., Buchete N.V., Kiel C., 2022

state and an active form of guanine triphosphate (GTP) bound state.[8, 9] There are many known K-Ras mutations that have been associated with many fatal cancers such as colorectal cancer, pancreatic ductal adenocarcinoma, and lung cancer (Fig. 1b).[9-13] Many computational and experimental efforts have been made to understand the conformational dynamics of K-RAS, the effects of its mutations, and to find mutation-specific drugs.[8, 14, 15] Understanding the detailed thermodynamic and equilibrium kinetic behavior of wildtype K-Ras4B proteins remains an important, outstanding aim due to its involvement in a variety of cancers.[8, 9]

K-Ras has two splice variants, namely K-Ras4A and K-Ras4B. Both variants are oncogenic but have distinct mechanism of membrane trafficking. K-Ras4A has 189 amino acids, like H-Ras and N-Ras. However, K-Ras4B has 188 amino acids. The two variants of K-Ras have a high degree of sequence identity of their catalytic domains, but a very low degree of sequence identity of their hypervariable region. In this paper, focus of our study is K-Ras4B. In this paper, K-Ras4B (the focus of this study) and K-Ras notations are used interchangeably, both referring to the K-Ras4B splice variant. K-Ras when bound to GDP is inactive as it makes an unstable complex with downstream effectors such as RAF1.[185] However, GTP-bound K-Ras can bind to downstream effectors, leading to signal transduction via pathways like Ras-Raf-MEK-ERK pathway.[2, 186] A major difference between GDP-bound K-Ras and GTP-bound K-Ras is in the regions switch I (residues 30-38) and switch II (residues 59-76) (Fig. 6.1a). In GDP-bound K-Ras, switches are more flexible and their distance from the GDP ligand is large. Experimental observations suggest that switch regions in the GTP-bound K-Ras structures are less flexible and tighter packed to the ligand. Computational studies of K-Ras have shown existence of GDP-bound K-Ras like conformations in the GTP-bound structures.[8] This suggests that even when K-Ras is bound to GTP, it can still adopt conformations which are not favorable for RAF1 binding, suggesting existence of both active and inactive states for GTP-bound K-Ras. Here, we computationally probe the detailed atomistic conformational dynamics of GTP-bound K-Ras4B proteins solvated with explicit water molecules, and the GTP-dependent activation and inactivation mechanism or K-Ras4B (WT) and K-Ras4B (D33E). The unexpected large effect of D33E (i.e., since both the charge are size properties are preserved to a good approximation) was also observed in Ref. [187].

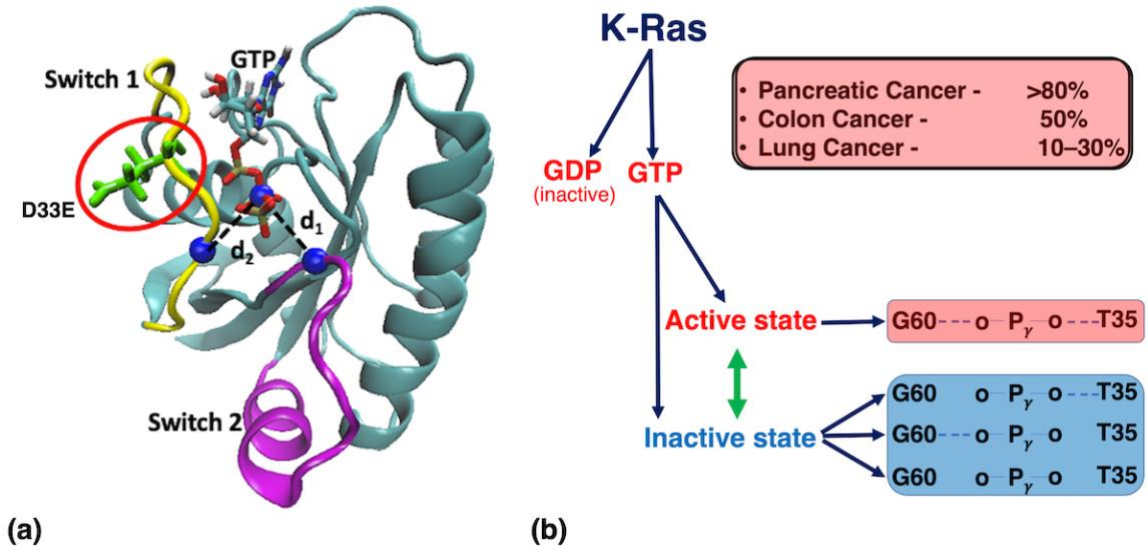


Figure 6.1. **(a)** K-Ras4B structural elements. The switch I and switch II regions are highlighted in yellow and magenta, respectively. The GTP ligand is shown in licorice and coloured by atom type. The C_a (for residues T35 and G60) and P_b (for GTP) atoms are shown as blue spheres. d_1 is the distance (dashed black line) between the C_a atom of G60 and the P_b atom of GTP. d_2 is the distance between the C_a atom of T35 and the same P_b atom of GTP (dashed black line). The D33 residue is circled in red. **(b)** Schematic representation of the GTP-dependent activation of K-Ras4B and the hypothesized relationship between its active/inactive states and the d_1 and d_2 distances.

6.3 Method

To model wild type K-Ras4B bound to GTP, the structure of human K-Ras (Q61H) in complex with the GTP analogue GNP (PDB ID 3GFT) was used, and residue 61 was mutated to GLN and GNP was replaced by GTP. The system was solvated with 16274 transferable intermolecular potential with 3 points (TIP3P) water molecules. NaCl ions were to neutralize and set the ion concentration to physiological conditions (0.15 mol/L). The system was minimized for 10,000 steps, followed by an isothermal-isobaric run (NPT) of 5 ns and a canonical ensemble run (NVT) of 20 ns using periodic boundary conditions at 300 K with 1 fs timestep. For data collection, 40 ns long 40 MD trajectories were launched from S1, S4 and S6 initial conformations (see figure). S1 is the crystal structure state with small d_1 and d_2 value. S4 and S6 are states already observed in first 25 ns run. A total of 4.8 ms of K-Ras4B(WT) MD simulation data was used for final analysis. D33E model was prepared by mutating the K-Ras(WT).GTP

prepared above. Same minimization and equilibration protocol was followed as explained for K-Ras(WT).GTP.

For data collection, 14 trajectories 100 ns long each were launched from the high energy area between S2 and S4 of WT free energy map (see Fig. 6.2a). This is a different sampling approach than previous MD studies of K-Ras that allows us to extract a detailed and fully converged landscape in (d_1,d_2) coordinates.[8, 15] A total of 1.4 ms of MD simulation data was used for analysis. The CHARMM 36 force field was used for all simulations and NAMD (2.13 multicore CUDA version) was used for MD simulations.

For the docking experiment, the PatchDock program[188] was used. RAF1 from the crystal structure[2] (the Ras binding and cysteine rich domains; PDB ID 6XI7) was used to dock to MD-generated K-Ras4B-GTP structures. A list (not exhaustive) of potential binding sites residues on RAF1 and Ras were provided to the program (see supplementary material). This list included residues from both Ras binding domain (RBD) and cysteine rich domain (CRD), as CRD increases the binding affinity to K-Ras and CRD K-Ras interaction is necessary for RAF1 activation, thus downstream signaling. However, RBD is more significant for RAF-RAS binding. Unlink RBD which primarily binds with switch region of K-Ras, CRD interacts with inter-switch region of K-Ras and C-terminal alpha helix via nonbonded interactions and hydrogen bonds. Binding interface (alpha carbons on RAF1 which are within 5 Å of K-Ras atoms) RMSD for top 5 structures, generated by the docking program, was computed. RMSD was calculated with respect to the 6XI7 crystal structure,[2] after aligning K-Ras. The minimum RMSD value out of 5 values in each case has been reported (see Table 6.1). VMD was used for generating representative structures. Python was used for data analysis and for generating plots.

6.4 Results

To check the suitability of using d_2 (see Figs. 6.1 and 6.2) as an reaction coordinate, we computed the corresponding distances between alpha carbon of several residues in the 32-40 sequence range and the P_b atom of GTP from our long

MD trajectories (Fig. S6.2). The distributions for of residues 32-34, and 36-40 clearly show a single peak (located at approximately 6.5, 8.5, 8, 10.2, 12, 14, 15 and 15 Å, respectively, see Fig. S6.2). However, the distribution for d_2 (residue T35) captures clearly the most structural states (i.e., three peaks at approximately 6.5, 8.2 and 12 Å, see Fig. S6.2). Clearly, the d_2 distance can capture more conformational metastable states as compared to other options.

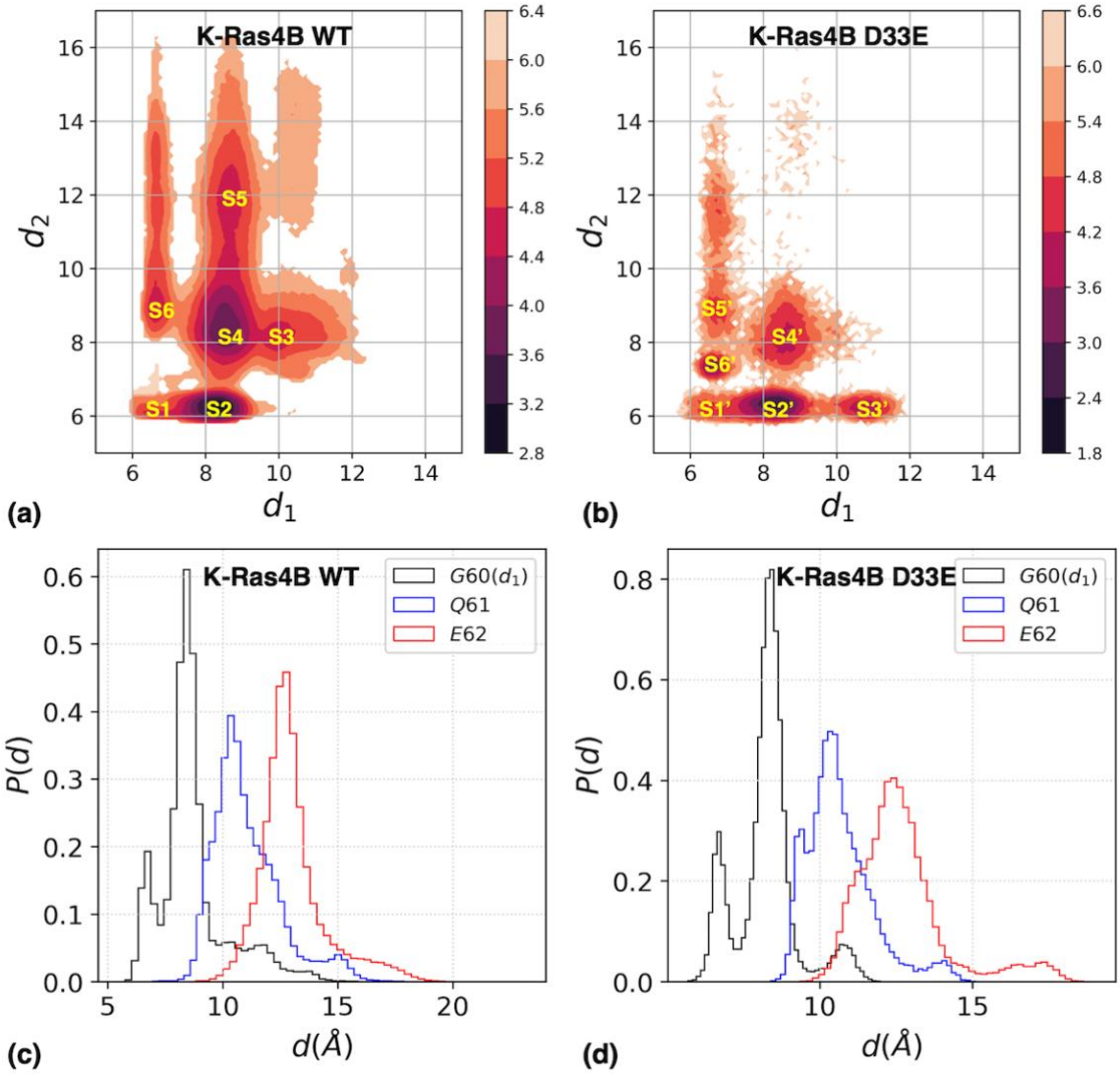


Figure 6.2. **(a)** The detailed K-Ras4B wild type (WT) free energy landscape (ΔG, kcal/mol) for its GTP-bound structure in d_1 and d_2 (Å) coordinates (see Fig. 1a). The six main conformational states of K-Ras4B WT are labelled S_1 to S_6, (yellow). **(b)** The corresponding ΔG landscape calculated for the GTP-bound K-Ras4B D33E mutant, with the new conformational basins labelled S_1' to S_6'.

Similarly, to check the suitability of using d_1 as a reaction coordinate, residues G60, Q61 and E62 located on switch 2 were selected, and the corresponding distance, (defined as the distance between the C_a atom of G60 and the P_b atom of GTP, dashed black line, Fig. 6.1a) is computed. The corresponding distance distributions for residues Q61 and E62 have only a single peak (at ~10.5 and ~13 Å, respectively, see Fig. 6.2). However, the distance distribution for residue G60 shows at least two clearly defined peaks at ~6.5 and ~8.5 Å (see Fig. 6.2c). Once again, the choice of using the d_1 distance as an RC seems to be the best choice.

Convinced that, for the force field use here, d_1 and d_2 are reasonable (Figs. 6.2c and S6.2), we extracted the underlying free energy landscape by measuring the population density in (d_1, d_2) space (Fig. 6.2a). We identified six representative conformations as the centers of free energy minima (highest population density), illustrated in Fig. S6.1. These six centers of energy minima represent states labeled, S1, S2, S3, S4, S5 and S6 (Fig. 6.2a). This more detailed free energy map suggests that GTP bound K-Ras (WT) has higher resolution than previously described. d_1 and d_2 distances were measured for various GTP analogue/ GTP bound K-Ras and GDP analogue/GDP bound K-Ras crystal structures (see Fig. S6.3). These corresponding distances for crystal structures show that most GTP bound K-Ras WT and mutant structures cluster closer to the S1 state on our computational free-energy map.

Using the six centers of free energy minima, identified above, Markov state model was built. Area within a radius of 0.5 Å around all six centers of free energy minima was assigned as the for-sure-zone for the respective state. To assign states to the conformations outside for-sure-zone, we used transition-based assignment (TBA) method. TBA method has been described in detail in ref. [54] A window length of 20 ns was selected to ensure Markovianity (see supplementary fig. S6.4). Using the sliding window of carefully selected window length, we estimated both populations and transition rates between the active and inactive states unveiled by this study, which suggest a more complex equilibrium molecular dynamics than previously reported (see figure 6.3). Detailed transition rate matrix and relative population of each state has been reported in figure 6.3. Our calculations showed that S2 is the most populous state with a population of about 46.7%. S6 is the least populated state (~3.6%) and S1 is the second least populated state (~3.8%).

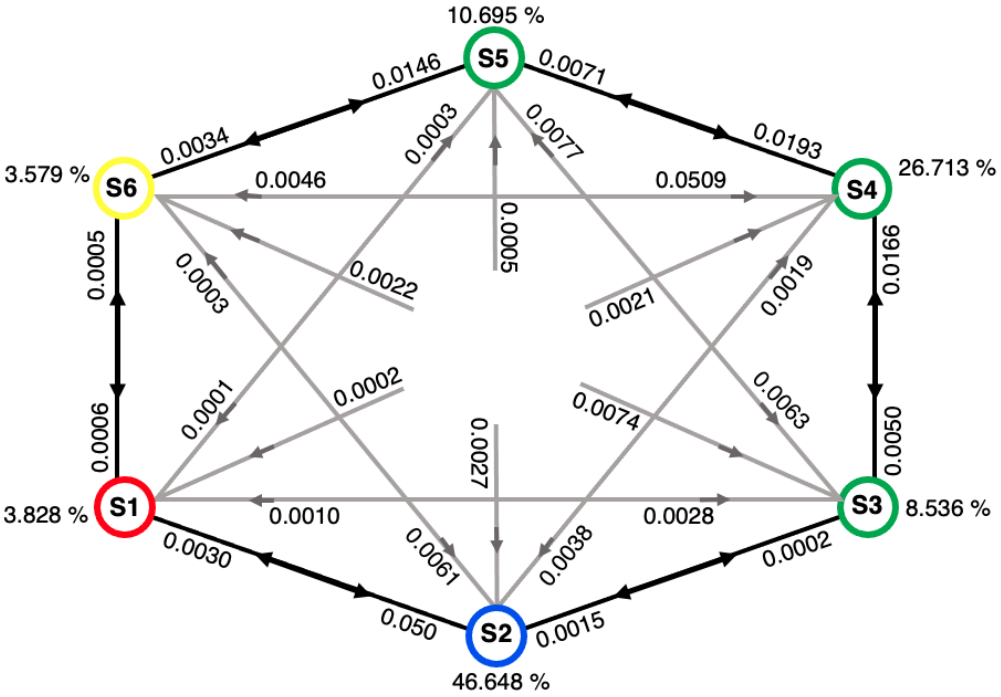


Figure 6.3. Visual representation of the rate matrix with rates (ns^{-1}) shown. The relative population (%) of each state is shown near the nodes. S2 is the most populated state with 46.648 % population and S6 is the least populated state with 3.579 %. Estimates of errors for rates and populations are shown in supplementary Table S6.1.

Structural analysis of RAF1 and K-Ras complex from the crystal structure PDB ID 6XI7,[2] shows that the acidic residues of K-Ras (e.g., residue D38) interacts with the basic residues of RAF1 creating strong salt bridges. To measure the orientation of D38, we computed the distribution of values of the dihedral angle θ_{38} (i.e., defined as the C-C_a-C_b-C_g dihedral angle for residue D38, in degrees, see Fig 6.4). For the 6XI7 crystal structure[2] dihedral angle for residue D38 is 151.63°. The D38 dihedral angle distribution for MD generated trajectories shows two clear peaks (Fig 6.4). Distribution of dihedral angle, θ_{38}, in states S1, S2, S3, S4, S5 and S6 shows the presence of both peaks in all six (d_1, d_2) conformational states.

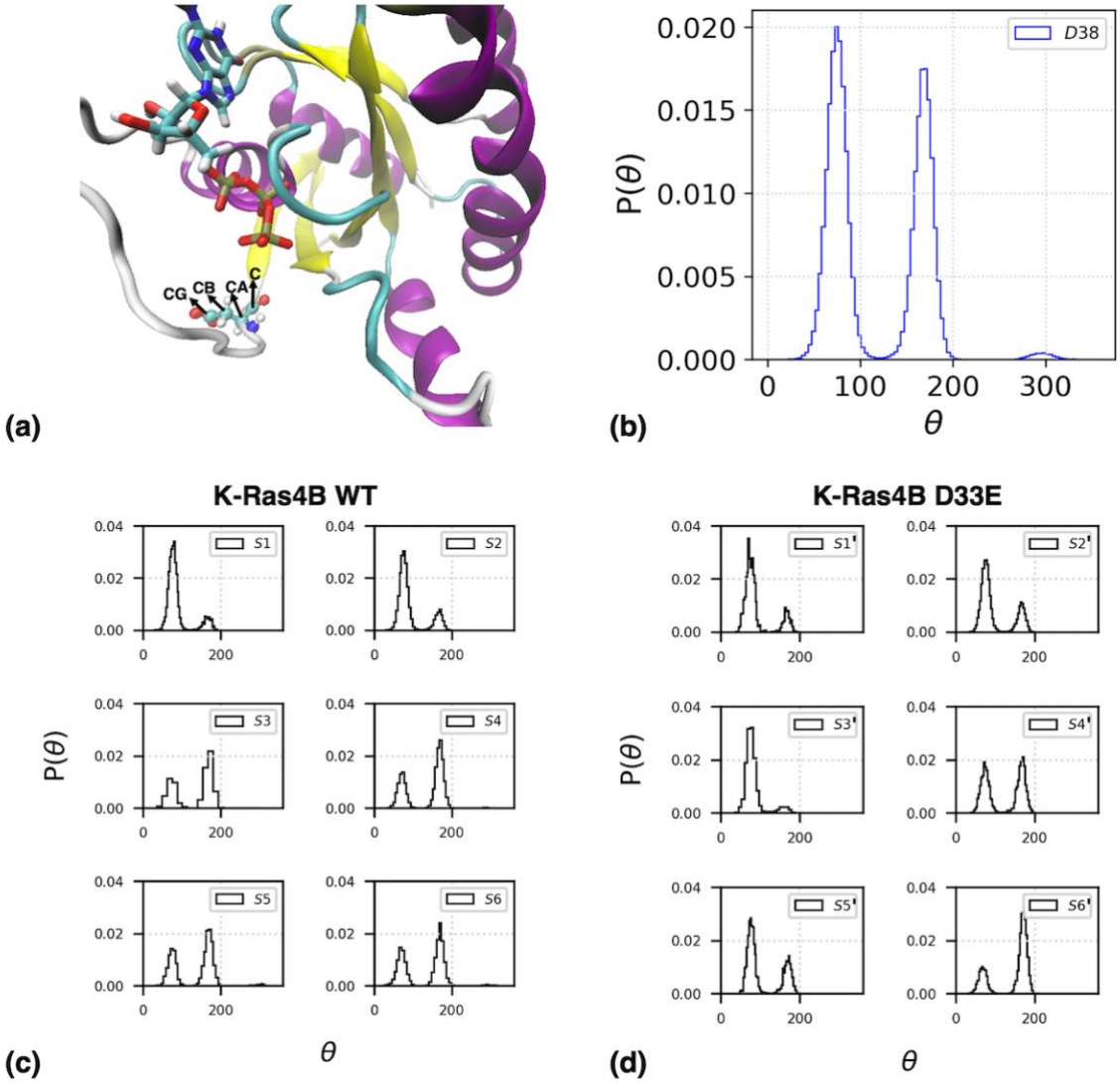


Figure 6.4. (a) The positions of atoms defining the angle q (i.e., the C-C_a-C_b-C_g dihedral angle for residue D38, in degrees) used for measuring the relative orientation (w.r.t. the local backbone) of the D38 side chain in the K-Ras4B WT with respect to the local backbone. The GTP ligand is shown as licorice and the D38 atoms as balls-and-sticks. **(b)** The corresponding overall distribution of θ_{38} values for K-Ras4B WT. **(c)** Dihedral angle (θ_{38}) distributions in each of the six states of K-Ras4B WT. **(d)** The corresponding dihedral angle (θ_{38}) distributions in each of the six states of K-Ras4B D33E.

Based on these observations, which are supported by other studies that suggested that the orientations of interface residues such as are important descriptors of K-Ras binding propensity to effector molecules,[189, 190] we hypothesize that for K-Ras4B GTP bound to be active, it should be able (i.e., have a high propensity) to bind to RAF1. To determine which possible combination of d_1, d_2 and theta is good for binding we dock RAF1, extracted from the 6XI7 structure, to peak 1 and peak 2

representative structures for all 6 (S1 to S6) states. We can safely assume that if the high scoring docked structure is close to the crystal structure orientation, then the binding is more likely and thus the K-Ras4B GTP conformation is active. Binding interface RMSD for top 5 structures, generated by the docking program, was computed. RMSD was calculated with respect to 6XI7 crystal structure, after aligning K-Ras. Low RMSD value means closer to crystal structure and higher propensity of RAF1-Ras binding. Smallest of the 5 RMSD values has been reported in Table 6.1. Clearly, S1 state conformation with large dihedral angle (i.e., peak 2) binds to RAF1 better than S1 state conformation with small dihedral angle (i.e., peak 1). Previous studies showed that if d_1 and d_2 distances are small then K-Ras-GTP is active and binds to RAF1.[8] Our study, shows that even for small values for d_1 and d_2, not all structures (including S1) are always active. We also note that theta alone is not sufficient for establishing RAF1 binding propensity. For example, states, like S2, large theta doesn't give a low RMSD. Importantly, from the docking calculations reported in Table 6.1, we can conclude that structures with small d_1 and d_2, and large theta values have the highest propensity to bind to RAF1. However, other reaction coordinates, besides d_1, d_2 and θ_{38} may be able to also correlate with the K-Ras(GTP)-RAF1 binding propensity. Interestingly, θ_{38} has the advantages over d_1 and d_2 that (i) it presents a largely bimodal distribution overall. and (ii) one of its two modes (peak 2) is much more compatible with RAF1 binding, as it promotes interfacial salt bridges between the acidic D37 and D38 residues of K-Ras and the basic residues of RAF1such as R89.

K-Ras4B WT			K-Ras4B D33E		
	θ_1	θ_2		θ_1	θ_2
S1	5.0	2.8	S1'	19.4	2.0
S2	5.1	20.1	S2'	21.3	3.0
S3	6.8	4.9		-	
	-		S3'	20.1	11.6
S4	21.4	5.4	S4'	9.6	19.8
S5	7.5	7.6		-	
S6	20.8	3.1	S5'	20.7	8.9
	-		S6'	4.2	20.3

Table 6.1. Results of docking-based modelling. RMSD values (in Å) obtained for comparing the docking interface of K-Ras4B and RAF1 obtained experimentally (PDB code 6XI7 with the dimer structures corresponding to the two peaks (denoted here by q_1 and q_2, respectively) of the angle θ_{38} (i.e., the C-C_a-C_b-C_g dihedral angle for residue D38, in degrees) used for measuring the relative orientation (w.r.t. the local backbone) of the D38 side chain in the K-Ras4B WT with respect to the local backbone. See text for discussion.

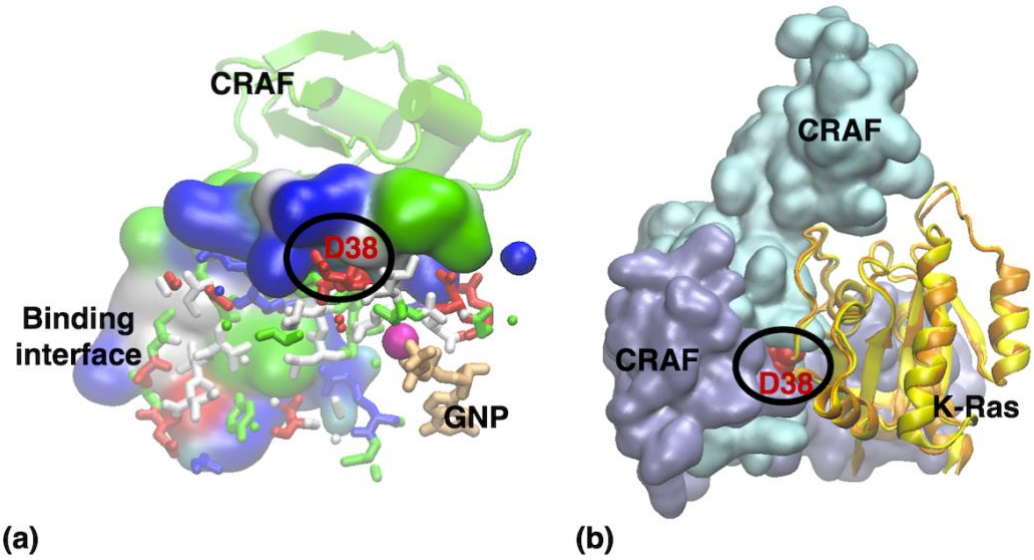


Figure 6.5. RAF-RAS binding interface. **(a)** Crystal structure of K-Ras4B-GNP in complex with RAF1 (from PDB ID 6XI7)[2]. The binding interface for RAF1is shown with surface representation and colored with residue type and binding interface residues of K-Ras are shown with sticks and colored with residue type. Acidic residues of K-Ras like residue 37 and 38 (shown in red and circled) binds with the basic interface residues (shown in blue) on RAF1 binding surface. **(b)** RAF1 docked to S2' peak1 and peak2 structures. RAF1 shown in cyan is docked to S2' peak 1 and RAF1 shown in purple is docked to S2' peak 2. The two structures share only a small part of binding interface. RAF1bound to peak 2 structure is closer to the crystal structure RAF1.

In Fig. 6.5 are illustrated important residues located at the binding interface between K-Ras4B and its RAF1 effector (from PDB ID 6XI7).[2] Our analysis shows that a main stabilizing factor is the network of complementary acidic (red) and basic (blue) amino acids located at this interface (Fig. 6.5a). Neutralizing mutations such as acting on D38 would have a destabilizing effect. However, our analysis of the θ_{38} angle shows that mutations such as D33E, can in fact increase the K-Ras affinity to RAF1 by allowing the D33 side chain (and thus the switch 1 region) to interact more closely to RAF1.

6.5 Conclusions

To sum up, we have designed and used a new set of MD trajectories, initialized from various regions of the (d_1, d_2) conformational space of K-Ras4B to extract high-resolution, converged free energy maps of both the K-Ras4B WT protein and of its K-Ras4B WT D33E mutant (Fig. 6.2a and 6.2b).

For the K-Ras4B WT molecule, the new map showcases six distinct (d_1, d_2) states that cannot be fully correlated with activation/inactivation propensities of K-Ras4B WT and mutated structures, demonstrating the need of additional reaction coordinates. The basins of the new ΔG map in (d_1, d_2) coordinates appears to correlate well with experimental information available on the structures of crystal structures that are either likely to be active (i.e., GTP or GTP-analogue bound) or known as inactive (GDP or GDP-analogue bound) as shown in Fig. S6.3. This is not unexpected as the d_1, and d_2 RCs were constructed to capture active-like conformations in regions such as S1 and S2. However, we find that some inactive-like conformation can also correspond to small (d_1, d_2) coordinates (e.g., labeled 8 in Fig. S6.3) while some active-like cases (e.g., labeled 7 and 9 in Fig. S6.3) correspond to rather large (d_1, d_2) values. To address this issue, we showed that the dihedral angle, θ_{38}, of an acidic interface residue, D38, has a rather bimodal distribution in all the six main conformational states evidenced in the (d_1, d_2) coordinates, with conformations corresponding to its two peaks correlating well to activity propensity (see data in Fig. S6.3b). This aspect was also supported by our docking results that showed that both high scores and small RMSD values for the docking interface are obtained by S1 and S2 structures that are

docked to RAF1. Interestingly, including the θ_{38} conformations in the analysis showed that experimental-like (i.e., small RMSD values from the experimental interface structure in 6XI7, in the range of ~2 to ~3 Å) can be obtained for states S1 and S2 if the θ_{38} conformations are also similar to active-like conformations (i.e., peak 2, Fig. 6.4). These observations highlight a new paradigm in analyzing the effect of K-Ras (WT) conformations and its mutant for inferring propensities of activation/inactivation, namely that a more complex collective variable such as obtained by monitoring all three RCs discussed (d_1, d_2, and θ_{38}) simultaneously, is necessary. This ability would be particularly useful in quantifying the effects of mutations on K-Ras propensities of activation/inactivation. As shown here, using the (d_1, d_2, θ_{38}) triad, we can explain the otherwise puzzling observation that a rather minor mutation, D33E, can have a rather dramatic effect by significantly increasing the K-Ras activation propensities for its S1 and S2 confirmational states. The unexpected large effect of D33E was also observed in Ref. [187]. Our MD and docking-based analysis approach can help both identify and validate in silico assessment of activation/inactivation propensities and shed new light on the underlying molecular mechanisms (i.e., in this case, the modulation of the network of salt bridges at the binding interface with the RAF1 downstream effector).

Our new and more complex mechanistic reaction coordinates provide an explanation as to how Ras uses the same binding site to engage with multiple effectors forming diverse binding interfaces through modulation of intermolecular interactions, such as salt bridges. Indeed, binding affinities between Ras and effectors vary a lot,[191] as do contributions of individual amino acid contacts (e.g. ‘hot-spots’) in the different Ras-effector interface.[191, 192] Our analysis also paves the way for a better mechanistic understanding of Ras oncogenic mutations that are suggested to differentially impact binding to effectors.[193]

7. Conclusions

The unifying theme of the work presented in this thesis is that we aim to contribute to the development and application of new advanced statistical methods that can be used to extend current kinetic and thermodynamic analysis models to larger and more complex cancer related proteins and, thus, paving the way to a better understanding of cancer-related systems (i.e., including K-Ras4B, Abl and Src studied here).

The basic methods and concepts that are employed throughout this work are summarized in Chapter 2. I summarize the theory guiding MD simulations and also discuss our application of the replica exchange molecular dynamics (REMD) method to studying the conformational dynamics of piezoelectric amyloid peptides in water. I also discuss how two reaction coordinates can be combined in a simple yet effective way to give a better collective reaction coordinate. Finally, the theory of the two approaches, Markov State Modelling (MSM) and the Milestoning method, is discussed. The first approach, MSMs, relies on identifying a set of configuration states in which the system resides sufficiently long to relax and loose the memory of previous transitions, and on using simulations for mapping the underlying complex energy landscape on a network of Markovian transitions. The independence of the underlying transition probabilities creates the opportunity to increase the sampling efficiency by using sets of appropriately initialized sets of short simulations rather than more typical long MD trajectories, which leads to both enhanced sampling and higher accuracy. This allows MSM studies to unveil bio-molecular mechanisms and to estimate free energy barriers with high accuracy, in a manner that is both systematic and relatively automatic, which accounts for their increasing popularity. The second approach, Milestoning, is focused on accurate studies of the ensemble of pathways connecting two specific end-states (e.g., reactants and products) in a similarly systematic and highly automatic and highly accurate manner. Conceptually, both methods are theoretically identical for transition paths between Markovian states, however

Milestoning can be generalized and applied to studies of non-Markovian transitions as well.

In chapter 3, I showed that replica exchange molecular dynamics (REMD) trajectories of explicitly solvated FF peptides can be used to probe in detail the interplay between temperature and electric field effects on the detailed thermodynamic and kinetic properties of the conformational dynamics of FF peptides in the presence of explicit water molecules.[66, 74] I showed that the thermodynamics and kinetics of the ensemble of conformations adopted by amyloid FF peptides solvated in explicit water molecules - an environment relevant to biomedical applications - can be analysed in detail by using REMD to enhance sampling, while simultaneously applying external electric fields and probing temperature ranges relevant to earlier studies.[74, 79, 80, 84, 85] Here I highlighted possible artifacts and how to overcome these artifacts that may occur during the setup of REMD simulations of explicitly solvated peptides in the presence of external electric fields, a problem particularly important in the case of short peptides such as FF. The presence of the external fields could over-stabilize certain conformational states in one or more REMD replicas, leading to distortions of the underlying potential energy distributions observed at each temperature. This cause is different from REMD artifacts reported and documented by earlier studies, which were due to modified underlying energy distributions caused, for example, by the use of weak-coupling thermostats.[107, 108] In this case, I showed that the resulting artifacts can be overcome by correcting the REMD initial conditions to include the lower energy conformations induced by the external field.

In chapter 4 and chapter 5, I combined a reaction path algorithm with the theory and algorithm of Milestoning to study kinetics of the DFG flip and disassociation of Gleevec from ABL kinase. I computed the mechanism, the rate of the transition in ABL kinase and MFPT for unbinding of Gleevec. The activation of kinases includes a conformational transition of the DFG motif that is important for enzyme activity but is not accessible to conventional Molecular Dynamics. I proposed a detailed mechanism for the transition, at a timescale longer than conventional MD, using a combination of reaction path and Milestoning algorithms. The mechanism includes local structural adjustments near the binding site as well as collective interactions with more remote residues. Milestoning simulations allowed to calculated slow kinetic timescales (i.e.,

associated slow processes on the order of tens to hundreds of milliseconds) from atomistic MD trajectories, while sampling the unbinding pathway of Gleevec from Abl, currently beyond the reach of conventional MD-based simulations. 13-15 We focus on the off rate as a critical measure of drug activity. This study unveiled significant insight into the dissociation kinetics. Interestingly, I found that the transition state ensemble appears to be late and broad according to the committor function approach, while being located rather early (i.e., closer to the Gleevec bound "reactant" state) according to the transition function (i.e., equal escape times towards both ends of the reaction pathway). These two studies indicate that Milestoning can play a central role in studies that facilitate the rational design of specific kinase inhibitors, by unveiling both kinetic and thermodynamic details of the rather complex kinase-drug molecular interactions.

Finally in chapter 6, I use the new short-trajectory approach to analyze the underlying free energy landscape of K-Ras4B and obtained states never discovered before and then using Markov State Modelling, I obtain the kinetic insight into the system. The new map showcases six distinct (d_1, d_2) states that cannot be fully correlated with activation/inactivation propensities of K-Ras4B WT and mutated structures, demonstrating the need of additional reaction coordinates. The basins of the new ΔG map in (d_1, d_2) coordinates appears to correlate well with experimental information available on the structures of crystal structures that are either likely to be active (i.e., GTP or GTP-analogue bound) or known as inactive (GDP or GDP-analogue bound). However, I found that some inactive-like conformation can also correspond to small (d_1, d_2) coordinates (e.g., labeled 8 in Fig. S6.3) while some active-like cases (e.g., labeled 7 and 9 in Fig. S6.3) correspond to rather large (d_1, d_2) values. To address this issue, I showed that the dihedral angle, θ_{38}, of an acidic interface residue, D38, has a rather bimodal distribution in all the six main conformational states evidenced in the (d_1, d_2) coordinates, with conformations corresponding to its two peaks correlating well to activity propensity (see data in Fig. S6.3b). This aspect was also supported by the docking results that showed that both high scores and small RMSD values for the docking interface are obtained by S1 and S2 structures that are docked to RAF1. Interestingly, including the θ_{38} conformations in the analysis showed that experimental-like (i.e., small RMSD values from the experimental interface structure in 6XI7, in the range of ~2 to ~3 Å) can be obtained for states S1 and S2 if the θ_{38} conformations are also similar to active type conformations (i.e., peak 2, Fig.

6.4). These observations highlight a new paradigm in analyzing the effect of K-Ras (WT) conformations and its mutant for inferring propensities of activation/inactivation, namely that a more complex collective variable such as obtained by monitoring all three RCs discussed (d_1, d_2, and θ_{38}) simultaneously, is necessary. This ability would be particularly useful in quantifying the effects of mutations on K-Ras propensities of activation/inactivation. By using the (d_1, d_2, θ_{38}) triad, one can explain the otherwise puzzling observation that a rather minor mutation, D33E, can have a rather dramatic effect by significantly increasing the K-Ras activation propensities for its S1 and S2 confirmational states. The unexpected large effect of D33E was also observed in Ref. [187]. The MD and docking-based analysis approach presented in chapter 6 can help both identify and validate *in silico* assessment of activation/inactivation propensities and shed new light on the underlying molecular mechanisms (i.e., in this case, the modulation of the network of salt bridges at the binding interface with the RAF1 downstream effector). This new and more complex mechanistic reaction coordinates provide an explanation as to how K-Ras can use the same binding site to engage with multiple effectors forming diverse binding interfaces through modulation of intermolecular interactions, such as salt bridges.

The methods developed and presented in this thesis could be extended to even larger systems and to more complex problems of biological relevance. When the problem involves estimating the thermodynamics and kinetics along transition pathways that connect two known metastable states of complex and large system, the multi-dimensional Milestoning approach, as introduced, discussed and applied in chapters 4 and 5, could be used. If the problem involves sampling the underlying dynamics of a very complex system, with no knowledge of the transition end points or a corresponding connecting path, then the second approach developed in chapter 6 could be used. With the new short trajectory-based approach, we were able to obtain converged and more detailed energy landscape and kinetics for a relatively large cancer-relevant system such as K-Ras4B.

Appendix 1- Supplementary material for Chapter 5

In this Supplementary Material we provide additional tests of our results. To assess the prediction of the Gleevec off rate, we compute the changes in the free energy barrier upon mutation. In Figure S5.1, we illustrate the thermodynamic cycle that we consider, and in Table S5.1 we report the corresponding results.

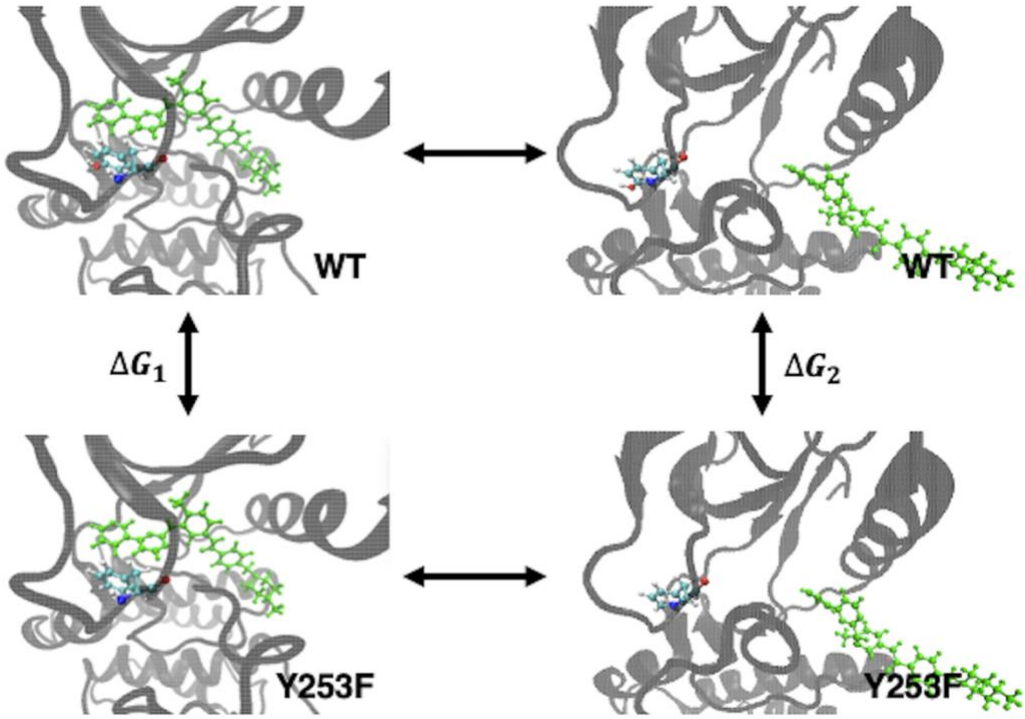


Figure S5.1. Thermodynamic cycle for alchemical free energy calculations. We compare the free energy changes of the bound (left) and transition states (right) for Gleevec interactions with the wild-type Abl Kinase and Y253F mutant. The mutated residue and the ligand (green) are shown using a CPK representation, in color.

Table S5.1. Alchemical free energy differences for the transformation from wild-type (WT) to Y253F, at bound (ΔG_1) and transition (ΔG_2) state conformations.

	VALUE (kcal/mol)	ERROR (kcal/mol)
$\boldsymbol{\Delta G_1}$	-1.20	0.09
$\boldsymbol{\Delta G_2}$	-1.64	0.07

All alchemical calculations were implemented in NAMD using a single-dual topology approach[1]. For the wild type to Y253F transformation, 20 windows (window length = 0.05) were used, $\lambda = 0$ represents the wild type and $\lambda = 1$ represents the fully mutated state. In each window, 50,000 steps were used for equilibration and 150,000 steps were used for FEP data collection. The same strategy was used for the backward transformation. Thus, the total simulation time for forward and backward transformations at the bound state was 8 ns. The BAR estimator was used to obtain the final $\boldsymbol{\Delta G_1}$ and $\boldsymbol{\Delta G_2}$ values. A similar procedure was followed for calculation at the transition state.

Testing the iso-committor surface predicted by Milestoning. We initiated unbiased trajectories starting from the transitions state ensemble in order to test how many of these trajectories made progress towards the reactant versus the product. 10 unbiased trajectories, 4ns-long each, were initiated from the TS1.

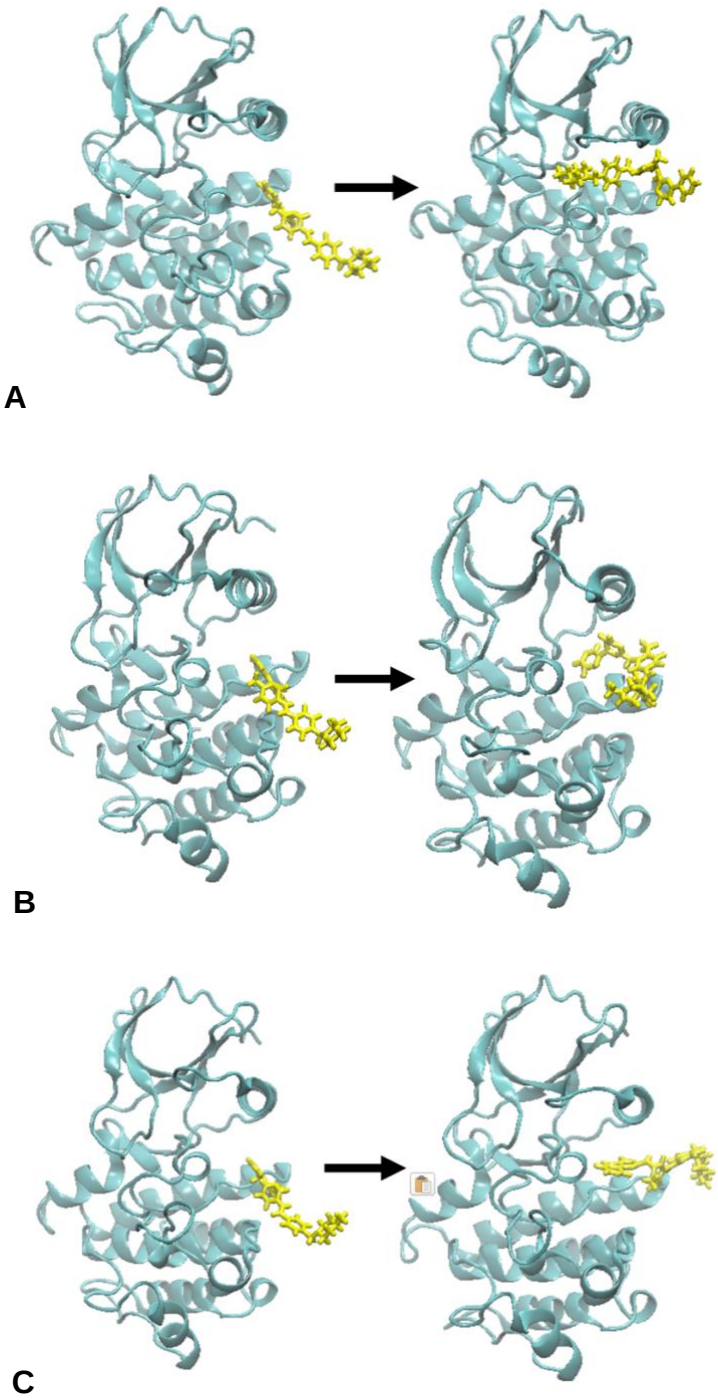


Figure S5.2. Initial and final Abl (ribbon) and Gleevec (licorice) structures and relative positions for the three inbound (i.e., moving towards the binding pocket) trajectories.

Appendix 2- Supplementary material for Chapter 6

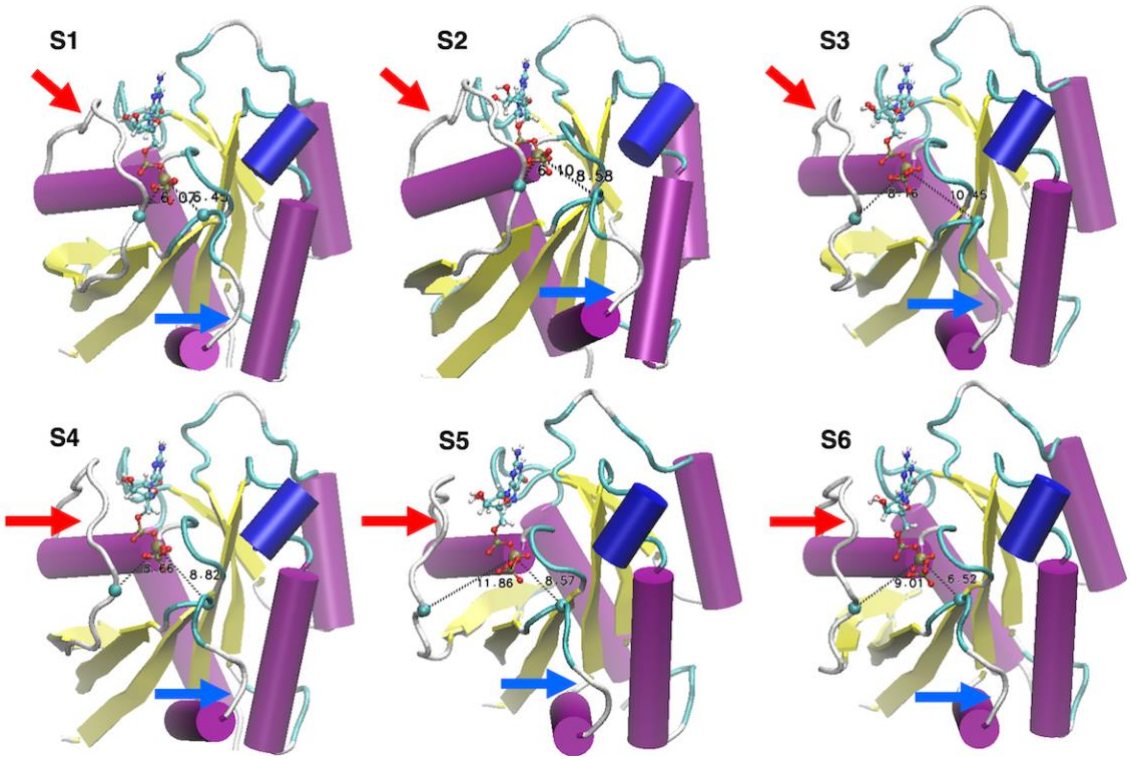


Figure S6.1. Representative structures for the six conformational states of K-Ras4B WT, S_1 to S_6, evidenced by the corresponding free energy map of GTP-bound K-Ras4B in the d_1-d_2 coordinates (in Å, see Fig. 6.2). Note differences in the relative positions of the switch I and switch II regions highlighted with red and blue arrows, respectively (see also Fig. 6.1). The sets of (d_1, d_2) coordinates of the representative structures selected here as centers of the S_1 to S_6 regions are (6.07, 6.45), (6.1, 8.58), (8.16, 10.45), (8.66, 8.82), (11.86, 8.57) and (9.01, 6.52), respectively.

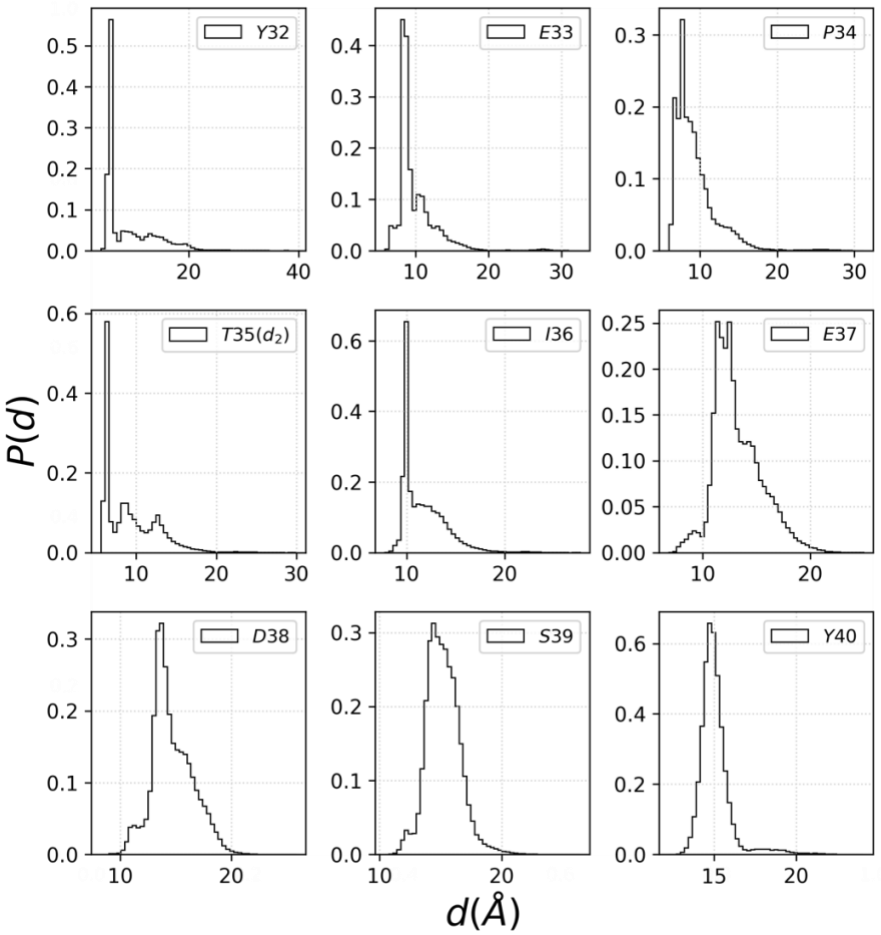
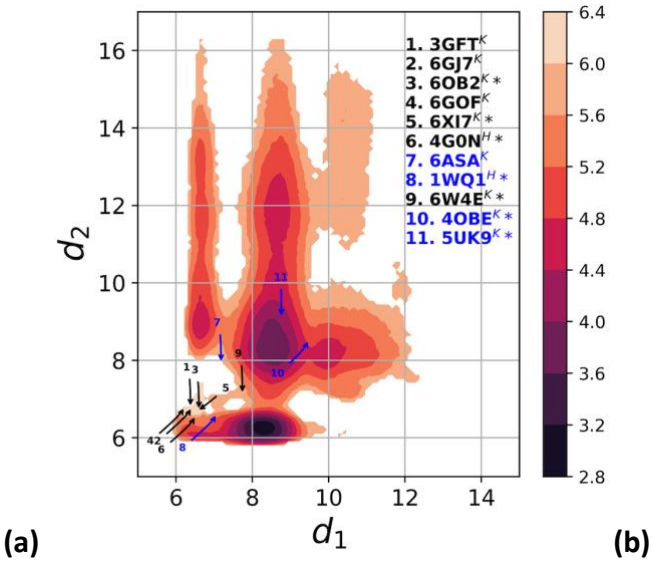


Figure S6.2. Distributions of the distances from switch I to GTP. Shown is the distribution of distance between alpha carbon of residues 32-40 and beta phosphate of GTP. Clearly, the d_2 distance (i.e., using the alpha carbon of T35) is the best reaction coordinate as it can discriminate more states.



(a) (b)

	PDB CODE	STRUCTURE	d_1	d_2	θ_{38}
1	3GFT	K-Ras(Q61H)-GNP	6.38	6.74	170.42
2	6GJ7	K-Ras(G12D)-GPPCP-22	6.42	6.83	179.27
3	6OB2	K-Ras(WT)-GMPPNP-GAP(GRD)	6.60	6.67	174.28
4	6GOF	K-Ras(G12D)-GPPNHP	6.23	6.84	170.63
5	6GOD	K-Ras(WT)-GNP	6.34	6.74	171.75
6	6XI7	K-Ras(WT)-GNP-Raf	6.51	6.70	151.63
7	4G0N	H-Ras(WT)-GppNHp-RBD of Raf complex	6.52	6.61	152.10
8	6ASA	K-Ras(D33E)-GDP	7.18	7.89	103.22
9	1WQ1	H-Ras(WT)-GDP-AlF3-GAP complex	7.07	6.66	183.02
10	6W4E	K-Ras(WT)-GSP-Lipid bilayer	7.74	7.09	118.84
11	4OBE	K-Ras(WT)-GDP	9.47	8.58	83.81
12	5UK9	K-Ras(WT)-GDP	8.76	9.06	84.93
13	6XHB	K-Ras(WT)-GNP-Raf	6.55	6.74	150.95
14	6WGN	K-Ras(G12D)-GPPNHP-KD2	9.32	n/a	188.84
15	5VQ6	K-Ras(WT)-GTP-gamma-S	n/a	n/a	127.02
16	4WA7	K-Ras(Q61L)-GDP	n/a	8.86	85.24
17	1BKD	H-Ras-GEF SOS	n/a	n/a	141.29

Figure S6.3. **(a)** Corresponding locations of the d_1 and d_2 values from experimental crystal structures overlapped on the free energy map of GTP-bound K-Ras4B WT (see also Fig. 2a). The positions of crystal structures of K-Ras and H-Ras bound to GTP (or GTP analogue), and to GDP (or GDP analogue) are highlighted in black and blue, respectively. The corresponding PDB codes for these structures are shown in the legend, using superscript K or H to distinguish between K-Ras and H-Ras structures, respectively. Wild type structures are marked with *. **(b)** Values of the corresponding d_1 and d_2 distances (in Å) and of the angle θ_{38} (the C-C_a-C_b-C_g dihedral angle for residue D38, in degrees) for experimental PDB structures.

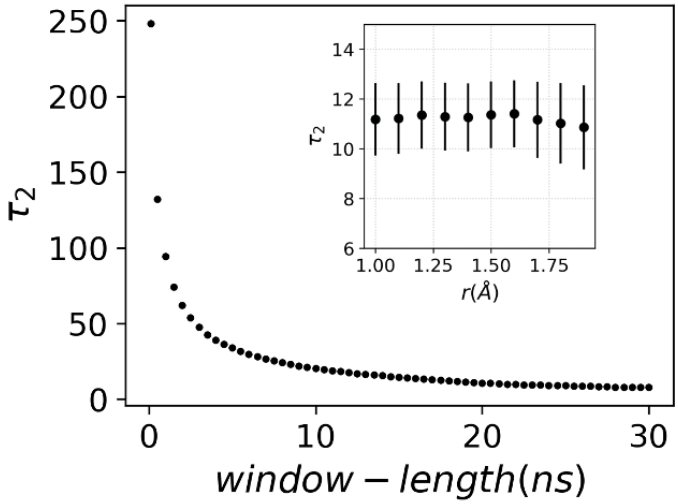


Figure S6.4. Slowest relaxation time with respect to change in window length. A sliding window was used to build the transition probability matrix and slowest relaxation time was estimated using the second eigenvalue. For final analysis, window length of 20 ns was used. Inset is the error in slowest relaxation time, for 20 ns window, with change in the diameter of for-sure-zone.

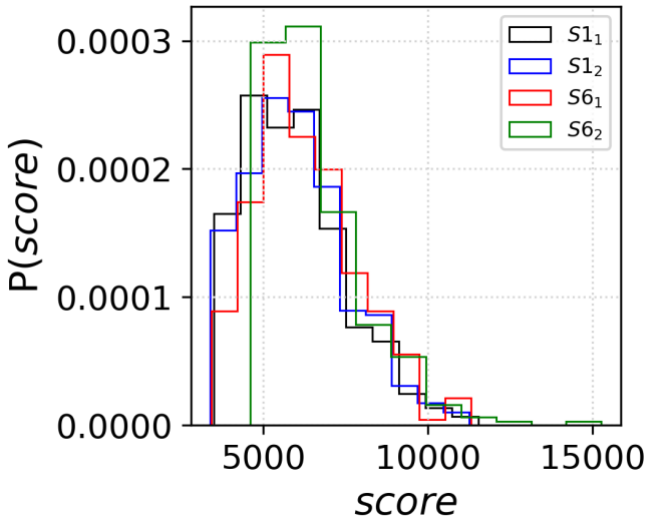


Figure S6.5. Distribution of docking scores, using PatchDock,[188] obtained for docking the K-Ras4B representative structures S1 and S6 (see Fig. 2) to the CRAF1 (from PDB ID 6XI7). Note that PatchDock is appropriate as it successfully at predicts only a few complex structures with high scores, including structures of the binding interface that have a small RMSD from the experimental interface (PDB ID 6XI7).

(a)

	S1	S2	S3	S4	S5	S6
S1	0.00393	0.00157	0.00167	0.00033	0.00018	0.00017
S2	0.00008	0.00029	0.00007	0.00004	0.00031	0.00004
S3	0.00067	0.00037	0.00684	0.00527	0.00069	0.00042
S4	0.00004	0.00110	0.00047	0.00294	0.00190	0.00037
S5	0.00007	0.00156	0.00099	0.00200	0.00543	0.00082
S6	0.00002	0.00287	0.00389	0.00792	0.00272	0.01742

(b)

	Error
S1	0.653
S2	4.281
S3	0.793
S4	6.345
S5	1.906
S6	1.286

Table S6.1. Error estimated as standard deviation. Data was split in four equal part and rate and population was calculated for each data set and the whole data set. Error reported here is the standard deviation of calculated (a) rate and (b) population values.

(a)

RAF1	
RBD	N64
	T68
	V69
	R89
CRD	139
	143
	178

(b)

K-Ras4B
R41
D38
E37
S39
Q43
G48
R149

Table S6.2. Binding-cite residues. (a) RAF1 residues on RBD and CRD provide to the docking software PatchDock. (b) K-Ras residues provided to the docking software.

References

1. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.
2. Tran, T.H., et al., *KRAS interaction with RAF1 RAS-binding domain and cysteine-rich domain provides insights into RAS-mediated RAF activation.* Nature Communications, 2021. **12**(1): p. 1176.
3. Patriksson, A. and D. van der Spoel, *A temperature predictor for parallel tempering simulations.* Phys. Chem. Chem. Phys., 2008. **10**(15): p. 2073-2077.
4. Alder, B.J. and T.E. Wainwright, *Phase transition for a hard sphere system.* The Journal of chemical physics, 1957. **27**(5): p. 1208-1209.
5. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-90.
6. Shirts, M. and V.S. Pande, *Computing - Screen savers of the world unite!* Science, 2000. **290**(5498): p. 1903-1904.
7. Zwanzig, R., *From classical dynamics to continuous time random walks.* Journal of Statistical Physics, 1983. **30**(2): p. 255-262.
8. Lu, S., et al., *Ras Conformational Ensembles, Allostery, and Signaling.* Chem Rev, 2016. **116**(11): p. 6607-65.
9. Fernandez-Medarde, A. and E. Santos, *Ras in cancer and developmental diseases.* Genes Cancer, 2011. **2**(3): p. 344-58.
10. Cefalì, M., et al., *Research progress on KRAS mutations in colorectal cancer.* Journal of Cancer Metastasis and Treatment, 2021. **7**: p. 26.
11. Umetani, N., et al., *Involvement of APC and K-ras mutation in non-polypoid colorectal tumorigenesis.* Br J Cancer, 2000. **82**(1): p. 9-15.
12. Ando, M., et al., *Higher frequency of point mutations in the c-K-ras 2 gene in human colorectal adenomas with severe atypia than in carcinomas.* Jpn J Cancer Res, 1991. **82**(3): p. 245-9.
13. Gerber, M., S. Goel, and R. Maitra, *In silico comparative analysis of KRAS mutations at codons 12 and 13: Structural modifications of P-Loop, switch I&II regions preventing GTP hydrolysis.* Computers in Biology and Medicine, 2022. **141**: p. 105110.
14. Kempf, E., et al., *KRAS oncogene in lung cancer: focus on molecularly driven clinical trials.* Eur Respir Rev, 2016. **25**(139): p. 71-6.
15. Wang, X., *Conformational Fluctuations in GTP-Bound K-Ras: A Metadynamics Perspective with Harmonic Linear Discriminant Analysis.* Journal of Chemical Information and Modeling, 2021. **61**(10): p. 5212-5222.
16. Hagemeijer, A., *Chromosome abnormalities in CML.* Baillieres Clin Haematol, 1987. **1**(4): p. 963-81.
17. Pane, F., et al., *Neutrophilic-chronic myeloid leukemia: a distinct disease with a specific molecular marker (BCR/ABL with C3/A2 junction).* Blood, 1996. **88**(7): p. 2410-4.
18. Narayan, B., et al., *The transition between active and inactive conformations of Abl kinase studied by rock climbing and milestoning.* 2020.
19. Narayan, B., et al., *Long-time methods for molecular dynamics simulations: Markov State Models and Milestoning.* Prog Mol Biol Transl Sci, 2020. **170**: p. 215-237.

20. Narayan, B., N.V. Buchete, and R. Elber, *Computer Simulations of the Dissociation Mechanism of Gleevec from Abl Kinase with Milestoning.* J Phys Chem B, 2021. **125**(22): p. 5706-5715.
21. Narayan, B., et al., *Replica Exchange Molecular Dynamics of Diphenylalanine Amyloid Peptides in Electric Fields.* J Phys Chem B, 2021. **125**(20): p. 5233-5242.
22. Levitt, M., *The birth of computational structural biology.* Nat Struct Biol, 2001. **8**(5): p. 392-3.
23. Monticelli, L. and D.P. Tieleman, *Force fields for classical molecular dynamics.* Methods Mol Biol, 2013. **924**: p. 197-213.
24. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Adv Protein Chem, 2003. **66**: p. 27-85.
25. Brooks, B.R., et al., *CHARMM: The biomolecular simulation program.* J. Comput. Chem., 2009. **30**(10): p. 1545-1614.
26. Scott, W.R.P., et al., *The GROMOS Biomolecular Simulation Program Package.* The Journal of Physical Chemistry A, 1999. **103**(19): p. 3596-3607.
27. Harrison, J.A., et al., *Review of force fields and intermolecular potentials used in atomistic computational materials research.* Applied Physics Reviews, 2018. **5**(3): p. 031104.
28. Martin-Garcia, F., et al., *Comparing molecular dynamics force fields in the essential subspace.* PLoS One, 2015. **10**(3): p. e0121114.
29. Verlet, L., *Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules.* Physical Review, 1967. **159**(1): p. 98-103.
30. Chen, J., *The Development and Comparison of Molecular Dynamics Simulation and Monte Carlo Simulation.* IOP Conference Series: Earth and Environmental Science, 2018. **128**: p. 012110.
31. Hünenberger, P.H., *Thermostat Algorithms for Molecular Dynamics Simulations*, in *Advanced Computer Simulation: Approaches for Soft Matter Sciences I*, C. Dr. Holm and K. Prof. Dr. Kremer, Editors. 2005, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 105-149.
32. Nosé, S., *A unified formulation of the constant temperature molecular dynamics methods.* The Journal of Chemical Physics, 1984. **81**(1): p. 511-519.
33. Hoover, W.G., *Canonical dynamics: Equilibrium phase-space distributions.* Physical Review A, 1985. **31**(3): p. 1695-1697.
34. Pastor, R.W., B.R. Brooks, and A. Szabo, *An analysis of the accuracy of Langevin and molecular dynamics algorithms.* Molecular Physics, 1988. **65**(6): p. 1409-1419.
35. Andersen, H.C., *Molecular dynamics simulations at constant pressure and/or temperature.* The Journal of Chemical Physics, 1980. **72**(4): p. 2384-2393.
36. Parrinello, M. and A. Rahman, *Crystal Structure and Pair Potentials: A Molecular-Dynamics Study.* Physical Review Letters, 1980. **45**(14): p. 1196-1199.
37. Nosé, S. and M.L. Klein, *Constant pressure molecular dynamics for molecular systems.* Molecular Physics, 1983. **50**(5): p. 1055-1076.
38. Åqvist, J., et al., *Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm.* Chemical Physics Letters, 2004. **384**(4): p. 288-294.
39. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1): p. 141-151.

40. Gnanakaran, S., et al., *Peptide folding simulations.* Curr Opin Struct Biol, 2003. **13**(2): p. 168-74.
41. Nguyen, P.H., et al., *Free energy landscape and folding mechanism of a beta-hairpin in explicit water: a replica exchange molecular dynamics study.* Proteins, 2005. **61**(4): p. 795-808.
42. Beck, D.A., G.W. White, and V. Daggett, *Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations.* J Struct Biol, 2007. **157**(3): p. 514-23.
43. Lockhart, C. and D.K. Klimov, *Alzheimer's Abeta10-40 peptide binds and penetrates DMPC bilayer: an isobaric-isothermal replica exchange molecular dynamics study.* J Phys Chem B, 2014. **118**(10): p. 2638-48.
44. Qi, R., et al., *Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example.* Methods Mol Biol, 2018. **1777**: p. 101-119.
45. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhanced sampling techniques in molecular dynamics simulations of biological systems.* Biochim Biophys Acta, 2015. **1850**(5): p. 872-877.
46. Abrams, C. and G. Bussi, *Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration.* Entropy, 2014. **16**(1): p. 163-199.
47. Kofke, D.A., *On the acceptance probability of replica-exchange Monte Carlo trials.* The Journal of Chemical Physics, 2002. **117**(15): p. 6911-6914.
48. Periole, X. and A.E. Mark, *Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent.* J Chem Phys, 2007. **126**(1): p. 014903.
49. Abraham, M.J. and J.E. Gready, *Ensuring Mixing Efficiency of Replica-Exchange Molecular Dynamics Simulations.* J Chem Theory Comput, 2008. **4**(7): p. 1119-28.
50. Sindhikara, D., Y. Meng, and A.E. Roitberg, *Exchange frequency in replica exchange molecular dynamics.* J Chem Phys, 2008. **128**(2): p. 024103.
51. Sindhikara, D.J., D.J. Emerson, and A.E. Roitberg, *Exchange Often and Properly in Replica Exchange Molecular Dynamics.* J Chem Theory Comput, 2010. **6**(9): p. 2804-8.
52. Doll, J.D., et al., *Rare-event sampling: Occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods.* The Journal of Chemical Physics, 2012. **137**(20): p. 204112.
53. Rosta, E. and G. Hummer, *Error and efficiency of replica exchange molecular dynamics simulations.* Journal of Chemical Physics, 2009. **131**(16): p. 12.
54. Buchete, N.-V. and G. Hummer, *Coarse Master Equations for Peptide Folding Dynamics.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 6057-6069.
55. Buchete, N.V. and G. Hummer, *Coarse master equations for peptide folding dynamics.* Journal of Physical Chemistry B, 2008. **112**(19): p. 6057-6069.
56. Husic, B.E. and V.S. Pande, *Markov State Models: From an Art to a Science.* Journal of the American Chemical Society, 2018. **140**(7): p. 2386-2396.
57. Buchner, G.S., et al., *Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory.* Biochimica Et Biophysica Acta-Proteins and Proteomics, 2011. **1814**(8): p. 1001-1020.
58. Chodera, J.D. and F. Noé, *Markov state models of biomolecular conformational dynamics.* Current Opinion in Structural Biology, 2014. **25**: p. 135-144.

59. Noé, F. and S. Fischer, *Transition networks for modeling the kinetics of conformational change in macromolecules.* Current Opinion in Structural Biology, 2008. **18**(2): p. 154-162.
60. Wang, W., et al., *Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules.* WIREs Computational Molecular Science, 2018. **8**(1): p. e1343.
61. Bowman, G.R., et al., *Progress and challenges in the automated construction of Markov state models for full protein systems.* The Journal of Chemical Physics, 2009. **131**(12): p. 124101.
62. Leahy, C.T., et al., *Peptide dimerization-dissociation rates from replica exchange molecular dynamics.* The Journal of Chemical Physics, 2017. **147**(15): p. 152725.
63. Leahy, C.T., et al., *Coarse Master Equations for Binding Kinetics of Amyloid Peptide Dimers.* The Journal of Physical Chemistry Letters, 2016. **7**(14): p. 2676-2682.
64. Chodera, J.D., et al., *Long-time protein folding dynamics from short-time molecular dynamics simulations.* Multiscale Modeling & Simulation, 2006. **5**(4): p. 1214-1226.
65. Buchete, N.V., R. Tycko, and G. Hummer, *Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments.* Journal of Molecular Biology, 2005. **353**(4): p. 804-821.
66. Narayan, B., et al., *Conformational analysis of replica exchange MD: Temperature-dependent Markov networks for FF amyloid peptides.* The Journal of Chemical Physics, 2018. **149**(7): p. 072323.
67. Bello-Rivas, J.M. and R. Elber, *Exact milestoning.* Journal of Chemical Physics, 2015. **142**(9).
68. Vanden-Eijnden, E., et al., *On the assumptions underlying milestoning.* J Chem Phys, 2008. **129**(17): p. 174102.
69. Berezhkovskii, A.M. and A. Szabo, *Committors, first-passage times, fluxes, Markov states, milestones, and all that.* J Chem Phys, 2019. **150**(5): p. 054106.
70. Vanden-Eijnden, E. and M. Venturoli, *Markovian milestoning with Voronoi tessellations.* Journal of Chemical Physics, 2009. **130**(19): p. 13.
71. Elber, R. and A. West, *Atomically detailed simulation of the recovery stroke in myosin by Milestoning.* Proc Natl Acad Sci U S A, 2010. **107**(11): p. 5001-5.
72. Tsai, C.-J., J. Zheng, and R. Nussinov, *Designing a Nanotube Using Naturally Occurring Protein Building Blocks.* PLoS Comput. Biol., 2006. **2**(4): p. e42.
73. Tsai, C.-J., et al., *Structure by design: from single proteins and their building blocks to nanostructures.* Trends Biotechnol., 2006. **24**(10): p. 449-454.
74. Kelly, C.M., et al., *Conformational dynamics and aggregation behavior of piezoelectric diphenylalanine peptides in an external electric field.* Biophys. Chem., 2015. **196**: p. 16-24.
75. Almohammed, S., et al., *Enhanced photocatalysis and biomolecular sensing with field-activated nanotube-nanoparticle templates.* Nat. Commun., 2019. **10**(1): p. 2496.
76. Görbitz, C.H., *Nanotube Formation by Hydrophobic Dipeptides.* Chem. - Eur. J., 2001. **7**(23): p. 5153-5159.
77. Gazit, E., *Self-assembled peptide nanostructures: the design of molecular building blocks and their technological utilization.* Chem. Soc. Rev., 2007. **36**(8): p. 1263-1269.
78. Kholkin, A., et al., *Strong Piezoelectricity in Bioinspired Peptide Nanotubes.* ACS Nano, 2010. **4**(2): p. 610-614.

79. Ryan, K., et al., *Nanoscale Piezoelectric Properties of Self-Assembled Fmoc–FF Peptide Fibrous Networks.* ACS Appl. Mater. Interfaces, 2015. **7**(23): p. 12702-12707.
80. Ryan, K., et al., *Thermal and aqueous stability improvement of graphene oxide enhanced diphenylalanine nanocomposites.* Sci. Technol. Adv. Mater., 2017. **18**(1): p. 172-179.
81. Reches, M. and E. Gazit, *Casting Metal Nanowires Within Discrete Self-Assembled Peptide Nanotubes.* Science, 2003. **300**: p. 625-627.
82. Guo, C., et al., *Triphenylalanine peptides self-assemble into nanospheres and nanorods that are different from the nanovesicles and nanotubes formed by diphenylalanine peptides.* Nanoscale, 2014. **6**(5): p. 2800-2811.
83. Seabra, A.B. and N. Durán, *Biological applications of peptides nanotubes: An overview.* Peptides, 2013. **39**: p. 47-54.
84. Castillo, J., et al., *Manipulation of self-assembly amyloid peptide nanotubes by dielectrophoresis.* Electrophoresis, 2008. **29**(24): p. 5026-5032.
85. Nakano, A. and A. Ros, *Protein dielectrophoresis: Advances, challenges, and applications.* Electrophoresis, 2013. **34**(7): p. 1085-1096.
86. Domigan, L., et al., *Dielectrophoretic manipulation and solubility of protein nanofibrils formed from crude crystallins.* Electrophoresis, 2013. **34**(7): p. 1105-1112.
87. Wang, M., et al., *Charged Diphenylalanine Nanotubes and Controlled Hierarchical Self-Assembly.* ACS Nano, 2011. **5**(6): p. 4448-4454.
88. Wang, X., et al., *Electric-field-enhanced oriented cobalt coordinated peptide monolayer and its electrochemical properties.* J. Colloid Interface Sci., 2013. **390**(1): p. 54-61.
89. Van der Spoel, D., et al., *GROMACS: Fast, flexible, and free.* J. Comput. Chem., 2005. **26**(16): p. 1701-1718.
90. Hess, B., et al., *GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation.* J. Chem. Theory Comput., 2008. **4**(3): p. 435-447.
91. Pastor, R.W., B.R. Brooks, and A. Szabo, *An analysis of the accuracy of Langevin and molecular dynamics algorithms.* Mol. Phys., 1988. **65**(6): p. 1409-1419.
92. Bussi, G., D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling.* J. Chem. Phys., 2007. **126**(1): p. 014101.
93. Parrinello, M. and A. Rahman, *Study of an F center in molten KCl.* J. Chem. Phys., 1984. **80**(2): p. 860-867.
94. Best, R.B., et al., *Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles.* J. Chem. Theory Comput., 2012. **8**(9): p. 3257-3273.
95. Jorgensen, W.L., et al., *COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER.* J. Chem. Phys., 1983. **79**(2): p. 926-935.
96. Doll, J.D. and P. Dupuis, *On performance measures for infinite swapping Monte Carlo methods.* J. Chem. Phys., 2015. **142**(2): p. 024111.
97. Schütte, C., et al., *A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo.* J. Comput. Phys., 1999. **151**(1): p. 146-168.
98. De Groot, B.L., et al., *Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds.* J. Mol. Biol., 2001. **309**(1): p. 299-313.

99. Levy, Y., J. Jortner, and R.S. Berry, *Eigenvalue spectrum of the master equation for hierarchical dynamics of complex systems.* Phys. Chem. Chem. Phys., 2002. **4**(20): p. 5052-5058.
100. Swope, W.C., et al., *Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β-hairpin peptide.* J. Phys. Chem. B, 2004. **108**(21): p. 6582-6594.
101. Chekmarev, D.S., T. Ishida, and R.M. Levy, *Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models.* J. Phys. Chem. B, 2004. **108**(50): p. 19487-19495.
102. Sriraman, S., I.G. Kevrekidis, and G. Hummer, *Coarse master equation from Bayesian analysis of replica molecular dynamics simulations.* J. Phys. Chem. B, 2005. **109**(14): p. 6479-84.
103. Zheng, W., et al., *Recovering Kinetics from a Simplified Protein Folding Model Using Replica Exchange Simulations: A Kinetic Network and Effective Stochastic Dynamics.* J. Phys. Chem. B, 2009. **113**(34): p. 11702-11709.
104. Berezhkovskii, A.M., F. Tofoleanu, and N.-V. Buchete, *Are Peptides Good Two-State Folders?* J. Chem. Theory Comput., 2011. **7**(8): p. 2370-2375.
105. Berezhkovskii, A.M., R.D. Murphy, and N.-V. Buchete, *Note: Network random walk model of two-state protein folding: Test of the theory.* J. Chem. Phys., 2013. **138**(3): p. 036101.
106. Stelzl, L.S. and G. Hummer, *Kinetics from Replica Exchange Molecular Dynamics Simulations.* J. Chem. Theory Comput., 2017. **13**(8): p. 3927-3935.
107. Cooke, B. and S.C. Schmidler, *Preserving the Boltzmann ensemble in replica-exchange molecular dynamics.* J. Chem. Phys., 2008. **129**(16): p. 164112.
108. Rosta, E., N.-V. Buchete, and G. Hummer, *Thermostat Artifacts in Replica Exchange Molecular Dynamics Simulations.* J. Chem. Theory Comput., 2009. **5**(5): p. 1393-1399.
109. Buchete, N.V. and G. Hummer, *Peptide folding kinetics from replica exchange molecular dynamics.* Phys. Rev. E, 2008. **77**(3): p. 4.
110. Jo, S. and W. Jiang, *A generic implementation of replica exchange with solute tempering (REST2) algorithm in NAMD for complex biophysical simulations.* Comput. Phys. Commun., 2015. **197**: p. 304-311.
111. Wang, L., R.A. Friesner, and B.J. Berne, *Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2).* J. Phys. Chem. B, 2011. **115**(30): p. 9431-9438.
112. Smith, A.K. and D.K. Klimov, *De novo aggregation of Alzheimer's Aβ25-35 peptides in a lipid bilayer.* Sci. Rep., 2019. **9**(1): p. 7161.
113. Druker, B.J., et al., *Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia.* New England Journal of Medicine, 2001. **344**(14): p. 1031-1037.
114. Cohen, P., *Protein kinases - the major drug targets of the twenty-first century?* Nature Reviews Drug Discovery, 2002. **1**(4): p. 309-315.
115. Wang, Z.L., et al., *Structural basis of inhibitor selectivity in MAP kinases.* Structure, 1998. **6**(9): p. 1117-1128.
116. Agafonov, R.V., et al., *Energetic dissection of Gleevec's selectivity toward human tyrosine kinases.* Nature Structural & Molecular Biology, 2014. **21**(10): p. 848-853.
117. Huse, M. and J. Kuriyan, *The conformational plasticity of protein kinases.* Cell, 2002. **109**(3): p. 275-282.

118. Vogtherr, M., et al., *NMR characterization of kinase p38 dynamics in free and ligand-bound forms.* Angewandte Chemie-International Edition, 2006. **45**(6): p. 993-997.
119. He, P., et al., *Conformational Free Energy Changes via an Alchemical Path without Reaction Coordinates.* Journal of Physical Chemistry Letters, 2018. **9**(15): p. 4428-4435.
120. Meng, Y.L., et al., *Predicting the Conformational Variability of Abl Tyrosine Kinase using Molecular Dynamics Simulations and Markov State Models.* Journal of Chemical Theory and Computation, 2018. **14**(5): p. 2721-2732.
121. Shan, Y.B., et al., *A conserved protonation-dependent switch controls drug binding in the Abl kinase.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(1): p. 139-144.
122. Lovera, S., et al., *The Different Flexibility of c-Src and c-Abl Kinases Regulates the Accessibility of a Druggable Inactive Conformation.* Journal of the American Chemical Society, 2012. **134**(5): p. 2496-2499.
123. Sultan, M.M., G. Kiss, and V.S. Pande, *Towards simple kinetic models of functional dynamics for a kinase subfamily.* Nature Chemistry, 2018. **10**(9): p. 903-909.
124. Regan, J., et al., *The kinetics of binding to p38 MAP kinase by analogues of BIRB 796.* Bioorganic & Medicinal Chemistry Letters, 2003. **13**(18): p. 3101-3104.
125. Casper, D., M. Bukhtiyarova, and E.B. Springman, *A Biacore biosensor method for detailed kinetic binding analysis of small molecule inhibitors of p38 alpha mitogen-activated protein kinase.* Analytical Biochemistry, 2004. **325**(1): p. 126-136.
126. Regan, J., et al., *Pyrazole urea-based inhibitors of p38 MAP kinase: From lead compound to clinical candidate.* Journal of Medicinal Chemistry, 2002. **45**(14): p. 2994-3008.
127. Vajpai, N., et al., *Solution conformations and dynamics of ABL kinase-inhibitor complexes determined by NMR substantiate the different binding modes of imatinib/nilotinib and dasatinib.* Journal of Biological Chemistry, 2008. **283**(26): p. 18292-18302.
128. Meng, Y.L., Y.L. Lin, and B. Roux, *Computational Study of the "DFG-Flip" Conformational Transition in c-Abl and c-Src Tyrosine Kinases.* Journal of Physical Chemistry B, 2015. **119**(4): p. 1443-1456.
129. Meng, Y.L., M.P. Pond, and B. Roux, *Tyrosine Kinase Activation and Conformational Flexibility: Lessons from Src-Family Tyrosine Kinases.* Accounts of Chemical Research, 2017. **50**(5): p. 1193-1201.
130. Templeton, C., et al., *Rock climbing: A local-global algorithm to compute minimum energy and minimum free energy pathways.* Journal of Chemical Physics, 2017. **147**(15): p. 10.
131. Vijayan, R.S.K., et al., *Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors.* Journal of Medicinal Chemistry, 2015. **58**(1): p. 466-479.
132. Young, M.A., et al., *Structure of the kinase domain of an imatinib-resistant Abl mutant in complex with the aurora kinase inhibitor VX-680.* Cancer Research, 2006. **66**(2): p. 1007-1014.
133. Levinson, N.M., et al., *A Src-like inactive conformation in the Abl tyrosine kinase domain.* Plos Biology, 2006. **4**(5): p. 753-767.
134. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.

135. Huang, J., et al., *CHARMM36: An Improved Force Field for Folded and Intrinsically Disordered Proteins.* Biophysical Journal, 2017. **112**(3): p. 175A-176A.
136. Martyna, G.J., D.J. Tobias, and M.L. Klein, *CONSTANT-PRESSURE MOLECULAR-DYNAMICS ALGORITHMS.* Journal of Chemical Physics, 1994. **101**(5): p. 4177-4189.
137. Feller, S.E., et al., *CONSTANT-PRESSURE MOLECULAR-DYNAMICS SIMULATION - THE LANGEVIN PISTON METHOD.* Journal of Chemical Physics, 1995. **103**(11): p. 4613-4621.
138. Miyamoto, S. and P.A. Kollman, *SETTLE - AN ANALYTICAL VERSION OF THE SHAKE AND RATTLE ALGORITHM FOR RIGID WATER MODELS.* Journal of Computational Chemistry, 1992. **13**(8): p. 952-962.
139. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of cartesian equations of m otion of a system with constraints - molecular dynamics of N-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
140. Essmann, U., et al., *A SMOOTH PARTICLE MESH EWALD METHOD.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.
141. Majek, P. and R. Elber, *Milestoning without a Reaction Coordinate.* Journal of Chemical Theory and Computation, 2010. **6**(6): p. 1805-1817.
142. Elber, R. and M. Karplus, *A method for determining reaction paths in large molecules - application to myoglobin.* Chemical Physics Letters, 1987. **139**(5): p. 375-380.
143. Elber, R., *A milestoning study of the kinetics of an allosteric transition: Atomically detailed simulations of deoxy Scapharca hemoglobin.* Biophysical Journal, 2007. **92**(9): p. L85-L87.
144. Atis, M., K.A. Johnson, and R. Elber, *Pyrophosphate Release in the Protein HIV Reverse Transcriptase.* Journal of Physical Chemistry B, 2017. **121**(41): p. 9557-9565.
145. Jonsson, H., G. Mills, and K.W. Jacobsen, *Nudged elastic band method for finding minimum energy paths of transitions*, in *Classical and quantum dynamics in condensed phase simulations*, B.J. Berne, G. Ciccotti, and D.F. Coker, Editors. 1998, World Scientific: Singapore. p. 385-403.
146. E, W.N., W.Q. Ren, and E. Vanden-Eijnden, *String method for the study of rare events.* Physical Review B, 2002. **66**(5): p. 4.
147. Czerminski, R. and R. Elber, *Self-avoiding walk between two fixed points as a tool to calculate reaction paths in large molecular systems.* International journal of quantum chemistry, 1990. **38**(S24): p. 167-185.
148. E, W. and E. Vanden-Eijnden, *Transition-path theory and path-finding algorithms for the study of rare events.* Annu Rev Phys Chem, 2010. **61**: p. 391-420.
149. Chong, L.T., A.S. Saglam, and D.M. Zuckerman, *Path-sampling strategies for simulating rare events in biomolecular systems.* Curr Opin Struct Biol, 2017. **43**: p. 88-94.
150. Tanner, D.E., et al., *Parallel Generalized Born Implicit Solvent Calculations with NAMD.* Journal of Chemical Theory and Computation, 2011. **7**(11): p. 3635-3642.
151. Olender, R. and R. Elber, *Yet another look at the steepest descent path.* Theochem-Journal of Molecular Structure, 1997. **398**: p. 63-71.
152. Faradjian, A.K. and R. Elber, *Computing time scales from reaction coordinates by milestoning.* Journal of Chemical Physics, 2004. **120**(23): p. 10880-10889.
153. Kirmizialtin, S. and R. Elber, *Revisiting and Computing Reaction Coordinates with Directional Milestoning.* Journal of Physical Chemistry A, 2011. **115**(23): p. 6137-6148.

154. Elber, R., *A new paradigm for atomically detailed simulations of kinetics in biophysical systems.* Quarterly Reviews of Biophysics, 2017. **50**.
155. West, A.M.A., R. Elber, and D. Shalloway, *Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide.* Journal of Chemical Physics, 2007. **126**(14).
156. Viswanath, S., et al., *Analyzing milestoning networks for molecular kinetics: Definitions, algorithms, and examples.* Journal of Chemical Physics, 2013. **139**(17).
157. Elber, R., et al., *Calculating Iso-Committor Surfaces as Optimal Reaction Coordinates with Milestoning.* Entropy, 2017. **19**: p. 219.
158. Ma, P., et al., *The Impact of Protonation on Early Translocation of Anthrax Lethal Factor: Kinetics from Molecular Dynamics Simulations and Milestoning Theory.* Journal of the American Chemical Society, 2017. **139**: p. 14837-14840.
159. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
160. Hantschel, O., *Structure, regulation, signaling, and targeting of abl kinases in cancer.* Genes Cancer, 2012. **3**(5-6): p. 436-46.
161. Seeliger, M.A., et al., *c-Src binds to the cancer drug imatinib with an inactive Abl/c-Kit conformation and a distributed thermodynamic penalty.* Structure, 2007. **15**(3): p. 299-311.
162. Yang, L.J., et al., *Steered molecular dynamics simulations reveal the likelier dissociation pathway of imatinib from its targeting kinases c-Kit and Abl.* PLoS One, 2009. **4**(12): p. e8470.
163. Lin, Y.L. and B. Roux, *Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases.* J Am Chem Soc, 2013. **135**(39): p. 14741-53.
164. Lin, Y.L., et al., *Computational study of Gleevec and G6G reveals molecular determinants of kinase inhibitor selectivity.* J Am Chem Soc, 2014. **136**(42): p. 14753-62.
165. Lovera, S., et al., *Towards a Molecular Understanding of the Link between Imatinib Resistance and Kinase Conformational Dynamics.* PLoS Comput Biol, 2015. **11**(11): p. e1004578.
166. Lin, Y.L., et al., *Explaining why Gleevec is a specific and potent inhibitor of Abl kinase.* Proc Natl Acad Sci U S A, 2013. **110**(5): p. 1664-9.
167. Kirmizialtin, S., et al., *How conformational dynamics of DNA polymerase select correct substrates: experiments and simulations.* Structure, 2012. **20**(4): p. 618-27.
168. Elber, R., *Milestoning: An Efficient Approach for Atomically Detailed Simulations of Kinetics in Biophysics.* Annu Rev Biophys, 2020. **49**: p. 69-85.
169. Ray, D., et al., *Kinetics and free energy of ligand dissociation using weighted ensemble milestoning.* J Chem Phys, 2020. **153**(15): p. 154117.
170. Ray, D. and I. Andricioaei, *Weighted ensemble milestoning (WEM): A combined approach for rare event simulations.* J Chem Phys, 2020. **152**(23): p. 234114.
171. Grazioli, G. and I. Andricioaei, *Advances in milestoning. II. Calculating time-correlation functions from milestoning using stochastic path integrals.* J Chem Phys, 2018. **149**(8): p. 084104.
172. Grazioli, G. and I. Andricioaei, *Advances in milestoning. I. Enhanced sampling via wind-assisted reweighted milestoning (WARM).* J Chem Phys, 2018. **149**(8): p. 084103.

173. Elber, R., et al., *Calculating Iso-Committor Surfaces as Optimal Reaction Coordinates with Milestoning.* Entropy (Basel), 2017. **19**(5).
174. Ma, P., R. Elber, and D.E. Makarov, *Value of Temporal Information When Analyzing Reaction Coordinates.* J Chem Theory Comput, 2020. **16**(10): p. 6077-6090.
175. Cowan-Jacob, S.W., et al., *Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia.* Acta Crystallogr D Biol Crystallogr, 2007. **63**(Pt 1): p. 80-93.
176. Baron, R., et al., *Multiple pathways guide oxygen diffusion into flavoenzyme active sites.* Proc Natl Acad Sci U S A, 2009. **106**(26): p. 10603-8.
177. West, A.M., R. Elber, and D. Shalloway, *Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide.* J Chem Phys, 2007. **126**(14): p. 145104.
178. Kreuzer, S.M., R. Elber, and T.J. Moon, *Early events in helix unfolding under external forces: a milestoning analysis.* J Phys Chem B, 2012. **116**(29): p. 8662-91.
179. Kreuzer, S.M., T.J. Moon, and R. Elber, *Catch bond-like kinetics of helix cracking: network analysis by molecular dynamics and milestoning.* J Chem Phys, 2013. **139**(12): p. 121902.
180. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
181. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
182. Hoemberger, M., W. Pitsawong, and D. Kern, *Cumulative mechanism of several major imatinib-resistant mutations in Abl kinase.* Proc Natl Acad Sci U S A, 2020. **117**(32): p. 19221-19227.
183. Paul, F., T. Thomas, and B. Roux, *Diversity of Long-Lived Intermediates along the Binding Pathway of Imatinib to Abl Kinase Revealed by MD Simulations.* J Chem Theory Comput, 2020. **16**(12): p. 7852-7865.
184. Tsuchida, N., T. Ryder, and E. Ohtsubo, *Nucleotide Sequence of the Oncogene Encoding the p21 Transforming Protein of Kirsten Murine Sarcoma Virus.* Science, 1982. **217**(4563): p. 937-939.
185. Kiel, C., et al., *Improved Binding of Raf to Ras·GDP Is Correlated with Biological Activity*.* Journal of Biological Chemistry, 2009. **284**(46): p. 31893-31902.
186. Huang, L., et al., *KRAS mutation: from undruggable to druggable in cancer.* Signal Transduction and Targeted Therapy, 2021. **6**(1): p. 386.
187. Chen, J., et al., *Mutation-Induced Impacts on the Switch Transformations of the GDP- and GTP-Bound K-Ras: Insights from Multiple Replica Gaussian Accelerated Molecular Dynamics and Free Energy Analysis.* Journal of Chemical Information and Modeling, 2021. **61**(4): p. 1954-1969.
188. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W363-7.
189. Li, Y., et al., *Specific Substates of Ras To Interact with GAPs and Effectors: Revealed by Theoretical Simulations and FTIR Experiments.* The Journal of Physical Chemistry Letters, 2018. **9**(6): p. 1312-1317.
190. Vatansever, S., B. Erman, and Z.H. Gümüş, *Oncogenic G12D mutation alters local conformations and dynamics of K-Ras.* Scientific Reports, 2019. **9**(1): p. 11730.
191. Kiel, C., D. Matallanas, and W. Kolch, *The Ins and Outs of RAS Effector Complexes.* Biomolecules, 2021. **11**(2).

192. Kiel, C., L. Serrano, and C. Herrmann, *A detailed thermodynamic analysis of ras/effector complex interfaces.* J Mol Biol, 2004. **340**(5): p. 1039-58.
193. Yuan, T.L., et al., *Differential Effector Engagement by Oncogenic KRAS.* Cell Rep, 2018. **22**(7): p. 1889-1902.