Title
Identification of Allosteric Mechanisms in Thrombin through Molecular Dynamics Simulations

Permalink
https://escholarship.org/uc/item/0js141rb

Author
Gasper, Paul M.

Publication Date
2013

Peer reviewed|Thesis/dissertation
Identification of Allosteric Mechanisms in Thrombin through Molecular Dynamics Simulations

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry

by

Paul M. Gasper

Committee in charge:

Professor J. Andrew McCammon, Chair
Professor Rommie E. Amaro
Professor Philip E. Bourne
Professor John E. Johnson
Professor Elizabeth A. Komives

2013
The dissertation of Paul M. Gasper is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____________________________________________

_____________________________________________

_____________________________________________

_____________________________________________

_____________________________________________

Chair

University of California, San Diego

2013
# Table of Contents

Table of Contents ................................................................. iv
List of Figures ........................................................................ vi
List of Tables ......................................................................... viii
List of Abbreviations ................................................................ ix
Acknowledgements .................................................................. x
Vita ...................................................................................... xii
Abstract of the Dissertation ..................................................... xiii

Chapter 1: A Molecular Dynamics Approach to Understanding Thrombin Allostery 1
  Introduction .......................................................................... 1
  Molecular dynamics simulations ............................................. 2
  Accelerated molecular dynamics simulations .......................... 3
  Thrombin and its regulation by thrombomodulin ..................... 4
  Allostery in thrombin TM regulation ..................................... 6

Chapter 2: The Dynamic Structure of Thrombin in Solution .......... 7
  Abstract ............................................................................... 7
  Introduction ......................................................................... 8
  Methods ............................................................................. 12
  Protein purification and sample preparation ............................ 12
  NMR experiments and data analysis .................................... 13
  Accelerated molecular dynamics calculations ........................ 13
  Residual local frustration analysis ..................................... 15
  Results and Discussion ...................................................... 16
  Backbone assignments of active site inhibited human thrombin... 16
  Residual dipolar couplings .................................................. 17
  Backbone dynamics experiments ....................................... 18
  Variable temperature experiments ...................................... 21
  Relation between dynamics and local residual frustration ....... 22
  Molecular dynamics simulations ....................................... 22
  Conclusion .......................................................................... 27
  Acknowledgments .............................................................. 29
  Supplemental Information .................................................. 30
  Details of the thrombin preparation and purification ............... 30
  Details of the NMR experimental methods ............................ 31
  Details of the AMD protocol ............................................. 34
List of Figures

Figure 1-1. Thrombin crystal structures apo (grey) and TM456 bound (purple) cartoon format overlay shows remarkably similar conformations. Interaction with TM456 shown by domain TM4 (violet), TM5 (tan), TM6 (blue) surface format. .......................... 5

Figure 1-2. Cartoon illustration comparing a) enthalpic allostery in which the binding of ligand elicits changes in average protein structure resulting in altered activity to b) entropic allostery in which the binding of ligand elicits changes in protein dynamics resulting in altered activity. ........................................................................... 6

Figure 1-3. Hypothetical two-dimensional representations of AMD potential energy surfaces for increasing values of $\alpha$ with the same cutoff. ................................................................. 3

Figure 2-1. TROSY 2D $[^1H,^{15}N]$ correlation spectrum of PPACK-inhibited $[^2H,^{15}N]$-labeled human $\alpha$-thrombin at 800 MHz and 25°C. Newly assigned resonances are labeled. b) Sequence of human $\alpha$-thrombin showing both commonly used numbering schemes. .................................................................................................................. 17

Figure 2-2. a) Residual dipolar coupling (RDC) values measured by TROSY-antiTROSY experiment. A schematic of the secondary structure is depicted above the plot. b) Plot of measured RDCs vs. those calculated from the x-ray structure (PDB code 1PPB) using PALES. ........................................................................................................... 18

Figure 2-3. Backbone relaxation data, a) $R_1$, b) $R_2$ and c) heteronuclear NOE, collected at 600 MHz, 25°C for $[^2H,^{15}N]$ PPACK-inhibited human $\alpha$-thrombin. d) $S^2$ values derived from the model-free analysis of the relaxation data using TENSOR2........................................................................................................................................ 19

Figure 2-4. a) Plot of the $S^2$ values calculated from the relaxation data using TENSOR2 (black), the CMD ensemble (blue) and the AMD ensemble (red). b) $R_{ex}$ values calculated from the TROSY Hahn-echo experiment. c) Plot of the intensity of the TROSY peak for each resonance at different temperatures. ....... 21

Figure 2-5. a) Plot of measured RDCs vs. those calculated from the CMD ensemble ($R^2=0.80$). b) Plot of measured RDCs vs. the RDCs calculated from the best-fit AMD ensemble ($R^2=0.92$). c) Plot of the “RDC improvement” as defined by $\frac{|(RDC_{expt}-RDC_{xray})|}{|(RDC_{expt}-RDC_{AMD})|}$.. ......................................................................................... 24

Figure 2-6. a) Structure of $\alpha$-thrombin depicting the ps-ns motions determined from the backbone relaxation experiments. b) Minimally-frustrated contacts (green) and highly-frustrated contacts (red); c) Structural ensemble using AMD. d) Residues with measured $R_{ex} > 6$ s$^{-1}$ (red), predicted supra-$\tau_c$ motions (orange). 26

Figure 3-1. The structure of thrombin:TM456. Top: Looking directly into the active site. Bottom: The domain architecture of thrombin:TM456. ......................................................... 41
Figure 3-2. Community analysis results (left) and corresponding structures colored by community (right) for a) thrombin b) thrombin:TM56 and c) thrombin:TM456. The communities are represented by spheres and the edge width is proportional to the cumulative betweenness of intercommunity edges. ................................. 44

Figure 3-3. Thrombin residue NH order parameters for isolated thrombin (top), thrombin:TM56 (middle) and thrombin:TM456 systems. b) Slow timescale (level 2 acceleration) differential NH backbone order parameters for thrombin:TM56 (black) and thrombin:TM456 (red)................................................................. 46

Figure 3-4. Normalized comparative analysis of observed thrombin cross-correlated motions in a) isolated thrombin, b) thrombin:TM56 (upper triangle) and thrombin:TM456 (lower triangle) obtained from representative free energy weighted (level 2) AMD molecular ensembles................................................ 49

Figure 3-5. Slow timescale cross-correlated molecular motions for thrombin in the thrombin:TM456 system projected onto a random snapshot taken from the thrombin:TM456 level 2 AMD trajectory................................................................. 50

Figure 3-6. Extended cross-correlated dynamic map for thrombin:TM456 obtained from a representative free energy weighted (level 2) AMD trajectory. The TM domains are defined as TM4 (residues 345 to 389), TM5 (residues 390 to 426) and TM6 (residues 427 to 462)................................................................. 51

Figure 4-1. a) and c) community analysis of apo-thrombin, and b) and d) PPACK-thrombin. Two dimensional view of communities in panels a) and b) depicts relative size of communities Panels c) and d) are structural representations of communities........................................................................................................... 75

Figure 4-2. Ensemble of the 20 lowest energy structures from the RDC-calibrated AMD for a) apo-thrombin and b) PPACK-thrombin................................................................. 76

Figure 4-3. a) Order parameters calculated from AMD simulations of apo-thrombin (red) and PPACK-thrombin (black). b) Differences in order parameters between the forms: \( \Delta S^2 = S^2_{\text{apo}} - S^2_{\text{PPACK}} \) from AMD order parameters. c) Those residues with \( \Delta S^2 > +0.1 \) are marked with red spheres on the structure of thrombin....... 78

Figure 4-4. Analysis of correlated motions performed on the AMD trajectories of PPACK-thrombin (top triangle) and apo-thrombin (bottom triangle). The motions range from 0.0 (no correlation, white) to 1.0 (completely correlated, black) ...... 79

Figure 4-5. Analysis of residual local frustration in the lowest energy structure from the RDC-calibrated AMD ensemble of apo-thrombin (left) and PPACK-thrombin (right). The contacts that are minimally frustrated are shown in green and the contacts that are highly frustrated are shown in red........................................ 80

Figure 4-6. Comparison of the order parameters reflecting ns timescale motions vs. longer timescale motions (\( S^2_{\text{AMD}} \)) with the average per residue fraction of highly frustrated contacts for the three lowest energy structures from the AMD simulation.................................................. 81
List of Tables

Table 3-1. Chymotrypsin and sequential residue numbering schemes for key regions of thrombin. .......................................................... 60
List of Abbreviations

ABE1 – Anion binding exosite 1
AMD – Accelerated molecular dynamics
CMD – Conventional molecular dynamics
CT – Chymotrypsin
EGF – Epidermal growth factor
H/D – Hydrogen-deuterium
HSQC – Heteronuclear single quantum correlation
ITC – Isothermal titration calorimetry
MALDI-TOF – Matrix-assisted laser desorption/ionization – time of flight
MD – Molecular dynamics
NASC – Number of atoms in the simulation cell
NMR – Nuclear magnetic resonance
NOE – Nuclear Overhauser effect
NOESY – Nuclear Overhauser effect spectroscopy
NSR – Number of solute residues
PDB – Protein data bank
PES – Potential energy surface
PPACK – D-Phe-Pro-Arg chloromethylketone
PME – Particle mesh Ewald
RDC – Residual dipolar coupling
\( R_{\text{ex}} \) – Relaxation due to chemical exchange
RMSD – Root mean square deviation
\( S^2 \) – Order parameter
SVD – Singular value decomposition
TM – Thrombomodulin
TM56 – Thrombomodulin fragment containing EGF-like domains 5 and 6
TM456 – Thrombomodulin fragment containing EGF-like domains 4, 5 and 6
TROSY – Transverse relaxation optimized spectroscopy
Acknowledgements

I would like to begin by thanking Professor J. Andrew McCammon, my committee chair, for his support and direction over the course of my doctoral work. I would also like to thank Dr. Mikolai Fajer, Dr. Morgan Lawrenz, Dr. Sara Nichols, Dr. Robert Swift, Professor Yi Wang, Professor Jeff Wereszczynski and the entire McCammon group past and present for helpful discussions during this time.

I must also thank my collaborators, whose contributions were critical to the success of this work. Professor Elizabeth A. Komives and Dr. Brian Fuglestad contributed invaluable NMR and other experimental results as well as their expertise on thrombin and allostery, and Dr. Phineus R. L. Markwick provided foundational work relating NMR observables to AMD ensembles as well as indispensible advice and efforts along the way.

I am also grateful to all of the agencies that contributed funding towards this work. I was supported by a Cell and Molecular Genetics (CMG) training grant (National Institutes of Health Grant T32 GM007240). Additional support was provided by a grant from the Center for Theoretical Biophysics (CTBP) (National Science Foundation Grant PHY-0822283), the National Science Foundation, the Howard Hughes Medical Institute, the National Biomedical Computation Resource (NBCR), and the San Diego Supercomputer Center.

Chapter 2 is a minimally modified reprint of the material as it appears in Brian Fuglestad, Paul M. Gasper, Marco Tonelli, J. Andrew McCammon, Phineus R. L. Markwick, and Elizabeth A. Komives, The Dynamic Structure of Thrombin in
Solution, 2012. I was the primary investigator and author of the computational aspects of this paper.

Chapter 3 is a minimally modified reprint of the material as it appears in Paul M. Gasper, Brian Fuglestad, Elizabeth A. Komives, Phineus R. L. Markwick and J. Andrew McCammon, Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities, Proceedings of the National Academy of Science 2012. I was the primary investigator and author of this paper.

Chapter 4 is a minimally modified reprint of the material as it appears in Brian Fuglestad, Paul M. Gasper, J. Andrew McCammon, Phineus R. L. Markwick, and Elizabeth Ann Komives, Correlated Motions and Residual Frustration in Thrombin, The Journal of Physical Chemistry B 2013. I was the primary investigator and author of the computational aspects of this paper.
Vita

2003 B. S. in Biochemistry
University of Wisconsin-Madison

2004-2007 Advanced Quality Technician
Abbott Laboratories

2009 M.S. in Chemistry
University of California, San Diego

2013 Ph.D. in Chemistry
University of California, San Diego

Publications

* These authors contributed equally to this work


** This author was the primary contributor of the computational aspects of this work
Abstract of the Dissertation

Identification of Allosteric Mechanisms in Thrombin through Molecular Dynamics Simulations

by

Paul M. Gasper

Doctor of Philosophy in Chemistry

University of California San Diego, 2013

Professor J. Andrew McCammon, Chair

Molecular dynamics (MD) simulation, an established method for investigating the internal motions of biomolecules, is applied to thrombin protein, a critical blood coagulation cascade protease with complex, not yet fully understood regulatory mechanisms. Accelerated MD (AMD) is employed to achieve enhanced conformational sampling of more biologically relevant timescales, and nuclear magnetic resonance (NMR) experiments are used to validate and tune AMD parameters. In chapter 1, the thrombin system and its interaction with cofactor thrombomodulin (TM) is introduced. The potential contribution of dynamics to
thrombin allostery is discussed, and methods for MD and AMD simulations are
described. In chapter 2, the combined use of NMR and AMD to examine the dynamics
of thrombin is detailed. AMD generated ensembles are shown to recapitulate NMR
residual dipolar couplings (RDCs), observables that report on ms timescale dynamics.
The resulting picture of thrombin depicts a stable core surrounded by highly dynamic
surface loops. In chapter 3, a computational study comparing isolated thrombin with
TM bound forms is reported. Community network analysis identifies two allosteric
pathways from the TM binding exosite to the thrombin active site. The presence of the
fourth EGF-like domain of TM, known to be essential for thrombin regulation, is
shown to establish and strengthen these allosteric pathways. This essential domain is
also observed to elicit enhanced dynamics and cross-correlated motion in the distal
active site loops. In chapter 4, these analyses are extended to compare apo thrombin
with active site inhibitor bound thrombin. Allosteric pathways between the active and
TM binding exosite site are again observed and are altered by the presence of
inhibitor. Residual local frustration analysis reveals a minimally frustrated thrombin
core surrounded by highly frustrated surface loops. The highly frustrated contacts
regions show significant overlap with regions undergoing slow timescale dynamics. In
chapter 5, overarching implications for a dynamical mechanism to thrombin:TM
allostery are discussed. The collective observations of thrombin active site surface
loops undergoing concerted, long timescale dynamics in response to TM binding at
exosite 1 strongly suggest a dynamic, allosteric mechanism in thrombin regulation.
Chapter 1:

A Molecular Dynamics Approach to Understanding Thrombin Allostery

Introduction

Knowledge of molecular dynamics is essential to understanding protein function. For over three decades now, it has been appreciated that constant changes in protein structure, resulting from thermal fluctuations of individual atoms, facilitate fundamental biomolecular processes such as enzyme catalysis, molecular transport and allosteric regulation. The ensemble of structures traversed by a protein over biologically relevant temperatures and timescales is not random but has been finely tuned and selected for optimizing its biochemical role. The thermodynamics and kinetics of protein conformational transitions are governed by the underlying potential energy surface (PES), which arises from a complex array of intra- and intermolecular forces. Classical force fields have been carefully constructed to model these forces from atomic positions. Molecular dynamics (MD) simulations in turn model protein dynamics through iterative application of Newton’s equations of motion from force field energies. The resulting molecular trajectories provide a wealth of information on
the internal dynamics of biomolecules and thus valuable insight on protein function. In this work, advanced MD methods are used in conjunction with experimental NMR to examine thrombin, a critical protein in the blood coagulation cascade.

**Molecular dynamics simulations**

Molecular dynamics (MD) simulations are a well-established computational technique. Beginning with atomic structures, typically generated by x-ray crystallography or NMR, Newton’s equations of motion are propagated for each atom over fs time-steps on which forces do not change substantially. The force active on each atom $i$ is calculated from the gradient of the potential energy, $F_i = -\nabla_i V(r)$, as determined by a molecular force field. Typical force fields model deviations from equilibrium bond lengths and angles as perturbations of harmonic springs. A periodicity term is introduced to account for forces resulting from rotation about dihedral angles. Non-bonded van der Waals forces are treated by a Lennard-Jones 6-12 potential and electrostatic interactions are described by a Columbic term. The summation of these energetic terms results in the total potential, $V(r)$.

$$V(r) = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi \cos(n\phi - \gamma)$$

$$+ \sum_{\text{van der Waals}} \varepsilon_{ij} \left( \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^6} \right) + \sum_{\text{electrostatic}} \frac{q_i q_j}{\epsilon_D r_{ij}}$$

When sufficient sampling is achieved, the trajectory is said to be ergodic and representative of the ensemble of structures under the specified conditions.
The atomic resolution on timescales of ns or less provided by MD simulations allows a wide range of analyses for investigating protein behavior. The use of simulation facilitates rapid replication with fine control over input parameters and avoids many of the costs and potential complications associated with experiment. The description of protein dynamics gained from MD simulations is invaluable toward the understanding of the mechanisms of protein function.

**Accelerated molecular dynamics simulations**

Conformational sampling in conventional MD simulations is typically limited by modern computational resources to timescales on the order of ns. As many biologically interesting events occur on the order of ms or longer, numerous methods for enhanced conformational sampling have been developed. One method that has proven successful is accelerated MD (AMD) (1). In AMD, the potential energy at each conformation sampled, V(r), below a cutoff energy E, is modified by the addition of a boost energy ΔV(r) (Fig. 1-1).

\[
V^*(r) = \begin{cases} 
V(r), & V(r) \geq E \\
V(r) + \Delta V(r), & V(r) < E 
\end{cases}
\]

\[
\Delta V(r) = \frac{(E - V(r))^2}{\alpha + (E - V(r))}
\]

Figure 1-1. Hypothetical two-dimensional representations of AMD potential energy surfaces for increasing values of α with the same cutoff, E (2).
The modified potential, $V^*(r)$, is raised and flattened, increasing the probability of transitions between low energy states. The strength of the boost is determined by the depth of an energy well, $E-V$, and the tuning parameter $\alpha$, which adjusts the overall extent of flattening. The corrected canonical ensemble average of the system can in theory be obtained by reweighting each point by the strength of the Boltzmann factor of the bias energy, $\exp[\beta \Delta V(r)]$.

While increasing levels of boost energy increase conformational sampling, over-acceleration may lead to poor statistics after reweighing as high energy conformations are over-sampled. Additionally, the timescale sampled by AMD simulations cannot be determined. To address these concerns, AMD ensembles can be calibrated to NMR observables. Residual dipolar couplings (RDCs), average spin-spin couplings measured in alignment media, report on dynamics up into the ms timescale (3). These values can also be back-calculated from AMD simulations using singular value decomposition (4). This process was used, as described in chapter 2, to identify optimal parameters for sampling of ms timescales in thrombin systems, and AMD generated ensembles at these levels gave good agreement with experimental NMR results.

**Thrombin and its regulation by thrombomodulin**

Thrombin is a serine protease that plays a pivotal role in the blood coagulation cascade. In the absence of its cofactor thrombomodulin (TM), activated thrombin acts as a procoagulant, cleaving fibrinogen to form fibrin, activating platelets and up-regulating its own activation. In contrast, when in complex with TM, thrombin acts as
an anticoagulant, changing its substrate specificity from fibrinogen to instead cleave protein C to its active form. Activated protein C inactivates coagulation cascade factors fVa and fVIIIa, leading to cessation of clotting (5; 6). However, the mechanism of thrombin regulation by TM is not yet fully understood. The TM binding site, anion binding exosite 1 (ABE1), is distal to the active site where fibrinogen and protein C cleavage occurs, and the crystal structure of thrombin is essentially unchanged by its interaction with TM (Fig. 1-2) (7). Molecular modeling studies have shown that TM binding blocks fibrinogen binding and provides an extended binding surface for protein C, accounting in part for enhanced thrombin activity for protein C (7). However, fluorescence (8), H/D exchange (9) and isothermal titration calorimetry (ITC) (10) studies have revealed changes in the thrombin active site for TM binding in the absence of protein C, indicating TM additionally has a direct effect on active site thermodynamics.

Figure 1-2. Thrombin crystal structures apo (grey) and TM456 bound (purple) cartoon format overlay shows remarkably similar conformations. Interaction with TM456 shown by domain TM4 (violet), TM5 (tan), TM6 (blue) surface format.
Allostery in thrombin TM regulation

The changes in thrombin active site thermodynamics that are observed on TM binding may be the result of allosteric signaling between the binding and active sites. While in the traditional view of allostery (11), such a signal is thought to be propagated via a series of changes in the average protein structure, enthalpic allostery, (Fig. 1-3a), the high degree of similarity between the apo and TM bound crystal structures makes this description appear inapplicable to thrombin binding TM. However, as has long been theorized (12) and more recently demonstrated (13), allostery may also be propagated via changes in protein dynamics, entropic allostery, (Fig. 1-3b). In chapters 3 and 4 extensive evidence is presented that this is indeed the case for the thrombin:TM system.

Figure 1-3. Cartoon illustration comparing forms of allostery. In both cases, the allosteric effector molecule (yellow, oval) binds protein resulting in enhanced affinity for substrate (cyan, hexagon) at the distal active site. However in a) enthalpic allostery, this is the result of changes in the average protein structure whereas in b) entropic allostery, this is the result of changes in protein dynamics. Figure modified from Tzeng and Kalodimos (13).
Chapter 2:
The Dynamic Structure of Thrombin in Solution

Abstract

The backbone dynamics of human α-thrombin inhibited at the active site serine were analyzed using $R_1$, $R_2$, and heteronuclear NOE experiments, variable temperature TROSY 2D [$^1$H-$^{15}$N] correlation spectra and $R_{ex}$ measurements. The N-terminus of the heavy chain, which is formed upon zymogen activation and inserts into the protein core, is highly ordered, as is much of the double beta barrel core. Some of the surface loops, by contrast, remain very dynamic with order parameters as low as 0.5, indicating significant motions on the ps-ns timescale. Regions of the protein that were thought to be dynamic in the zymogen and to become rigid upon activation, in particular the γ-loop the 180s loop and the Na$^+$ binding site, have order parameters below 0.8. Significant $R_{ex}$ was observed in most of the γ-loop, in regions proximal to the light chain, and in the β-sheet core. Accelerated molecular dynamics simulations yielded a molecular ensemble consistent with measured residual dipolar
couplings that revealed dynamic motions up to ms. Several regions including the light chain and two proximal loops did not appear highly dynamic on the ps-ns timescale, but had significant motions on slower timescales.

**Introduction**

In blood coagulation, a cascade of proteolytic events converts zymogens to active proteases. Ultimately, prothrombin is cleaved twice removing the lipid-binding domain, and activating it for catalysis (14; 15). The resulting α-thrombin cleaves twelve known substrates, including fibrinogen and factor (f) XIII (responsible for forming and cross-linking the fibrin clot) and the protease-activated receptors (responsible for platelet activation and many other cell signaling events). Thrombin up-regulates its own formation by activating essential cofactors fV and fVIII and the protease fXI. Thrombin also down-regulates its own formation when, in complex with thrombomodulin, it activates protein C, which is responsible for inactivating fVa and fVIIIa. The allosteric regulation of thrombin specificity towards multiple substrates is not yet understood (16; 17).

The coagulation proteases all resemble trypsin, but whereas trypsin has a broad substrate specificity cleaving after any exposed Lys or Arg, the coagulation proteases have exquisite substrate specificity. Sequence alignment analysis revealed that the coagulation proteases have non-conserved surface insertions at several loops (18) and it has been suggested that only the trypsin-like catalytic core is conserved (19).

The structure of thrombin has mainly been analyzed by x-ray crystallography, and the many available structures reveal thrombin surface loops “trapped” in different
conformations under crystallographic conditions. Di Cera’s group recently captured the same mutant form of thrombin in two different conformational states in the same crystals (20). Hydrogen-deuterium (H/D) exchange followed by mass spectrometry suggested that most of the surface loops are dynamic based on their rapid amide exchange behavior (9). In the first report of NMR resonance assignments of variously liganded forms the resonances for exosite 1 (both 70s and 30s loops) were missing unless a ligand was bound directly to this loop and resonances for the γ-loop were not observed under any conditions (21). Some resonances for the light chain were also missing. Although the authors of this work asserted that the absence of resonances is indicative of μs-ms dynamics, many other reasons including proteolysis of surface loops, or weak association of the proteins at high NMR concentrations could also account for such observations.

In this study, we provide a detailed quantitative analysis of the dynamic properties of thrombin across a broad range of timescales using a combination of NMR experiments and state-of-the-art molecular dynamics simulation. In the past two decades, NMR has emerged as the method of choice for investigating protein dynamics across a broad spectrum of timescales (22; 23). In the present study, we have employed nuclear spin relaxation (R₁, R₂, heteronuclear NOE), residual dipolar coupling (RDC) experiments and relaxation due to chemical exchange (Rex) measurements to study the dynamic properties of PPACK-inhibited thrombin on multiple timescales. Although both nuclear spin relaxation and RDCs are associated with magnetic dipolar coupling, they probe protein dynamics on very different
timescales. Nuclear spin relaxation results from the time dependent stochastic modulation of physical interactions between spin-active nuclei, inducing random field fluctuations, which relax the excited spin state back to equilibrium. These experiments are now routinely employed to probe local dynamic fluctuations up to the characteristic rotational correlation time of the system, which is nearly 20 ns for thrombin. In the framework of the well-known Lipari-Szabo model-free analysis, the temporal reorientation fluctuation of an inter-nuclear vector is described using three parameters: the internal and total correlation times and an order parameter, $S^2$, which characterizes the spatial restriction of the inter-nuclear vector.

Residual Dipolar Couplings (RDCs) have long been recognized as very powerful tools for structure determination, particularly when combined with other distance and dihedral angle restraints derived from NOEs and scalar J-couplings respectively (24). However, it is now well accepted that RDCs can also be applied to the study of protein dynamics (3). RDCs are averaged over all orientations of the magnetic dipolar interaction vector sampled up to a timescale defined by the inverse of the alignment-induced coupling, and therefore report on time-averaged motions up to the ms range. This early work focused on rigid body domain motions (25; 26). More recently attention has focused on the study of slow backbone conformational dynamics primarily in ubiquitin using both motional models and MD simulation strategies: The methods include Self-Consistent RDC-based Model free approach (SCRM), (27), the Direct Interpretation of Dipolar Couplings (DIDC) method (28), the Structure-Free Gaussian Axial Fluctuation model (SF-GAF) approach (29), Ensemble Refinement
with Orientational Restraints (EROS) (30) and the Accelerated Molecular Dynamics (AMD) approach (4) employed in the present work. The agreement between the RDCs back-calculated from the AMD ensemble was statistically significantly better than the match to any crystal structure. Even more dramatic improvements were observed when the same approach was applied to the first four ankyrin repeats of IκBα (31).

Measurements of $R_{ex}$ are an indicator of structural fluctuations on the 100 µs to ms timescales (23). These measurements are a result of the nuclear probe (backbone nitrogens in this study) undergoing exchange between states of distinct chemical shifts on the NMR timescale. The effect of temperature on line broadening was also measured. With increasing temperature, peak intensities are expected to increase as the tumbling of the molecule becomes faster. Deviations from this expected behavior are indicative of exchange processes and may give insight into the characteristics of these exchange processes.

The experimental NMR studies were complemented by a theoretical investigation using both conventional molecular dynamics (CMD) simulations and AMD, an enhanced conformational space sampling algorithm. We report here the first direct measurements of the backbone dynamics of D-Phe-Pro-Arg chloromethylketone (PPACK)-inactivated human α-thrombin. The results reveal an ensemble in which much of the surface of thrombin remains surprisingly dynamic even in the active site liganded form.
Methods

Protein purification and sample preparation

Isotopically labeled wild-type human thrombin ($^2$H$^{15}$N or $^2$H$^{15}$N$^{13}$C) was produced by a procedure modified from the previously published method (32). The procedure involves expression of “prethrombin-2” in *E. coli* from a modified pET23a+ vector containing the sequence of prethrombin-2 plus residues 310-327 of prothrombin at the N-terminus, which were critical for proper refolding. Details of the protein preparation are found in the SI.

The fully active α-thrombin was confirmed by fibrinogen clotting times (activity of >3500 U/mg), chromogenic substrate assays, and MALDI-TOF mass spectrometry of reduced and oxidized samples (revealing the correct molecular weight of the oxidized form of 34kD and of the heavy chain upon reduction of 30.5 kD (including PPACK and seven acetamido groups on the cysteines)).

Covalent inhibition of thrombin was achieved by incubation with a 10x molar excess of H-D-Phe-Pro-Arg-chlormethylketone (PPACK, Haematologic Technologies Inc.) at room temperature for 1 hour. To separate excess PPACK from thrombin and exchange the protein into NMR buffer, the protein solution was concentrated and loaded onto a Superdex S75 size exclusion column (GE Healthcare Life Sciences) eluted with 25 mM sodium phosphate pH 6.5, 150 mM NaCl, 0.05% sodium azide. A final PPACK-thrombin concentration of 150 µM and 10% (v/v) D$_2$O was used for the NMR experiments at a final volume of 300 µL in a Shigemi tube (Shigemi Inc. cat. # BMS-005).
NMR experiments and data analysis

NMR resonance assignment experiments were performed at 298°K unless otherwise specified on either a Bruker Avance III 600 MHz (UCSD Pharmacology), Varian VS 800 MHz (UCSD Chemistry), or a Varian VNS 800 MHz (NMRFAM) all equipped with cryoprobes. Triple resonance assignment experiments performed were HNCO, HN(CA)CO, HN(CO)CA, HN(COCA)CB at 600 MHz as well as NOESY-$^1$H-$^{15}$N-HSQC ($\tau_{\text{mix}}$=200 ms), TROSY-HNCA and TROSY-HN(CA)CB at 800 MHz (33-36). The details of the NMR experiments used here are found in the SI.

Accelerated molecular dynamics calculations

The details of the AMD protocol have been discussed previously in the literature (1; 37) and specifics of the protocol used here are found in the SI. In previous studies (4), we assumed that a single alignment tensor was sufficient to describe the orientation of the molecule in the alignment medium. In the present work, we employed a multiple alignment tensor model as the PPACK-thrombin system exhibits considerable flexibility, particularly on slower timescales. Whereas for the previous study of ubiquitin the backbone RMSD across the AMD trajectory with respect to a randomly chosen structure varied up to 1.4 Å; in this study the backbone RMSD for PPACK-thrombin varied up to 2.9 Å. Large backbone RMSD values may represent a significant change in the preferential orientation of the molecule in the alignment medium. As the exchange between conformational states occurs on timescales considerably slower than the rotation diffusion coefficient (~20 ns according to Stokes theory), the system has ample time to adopt a different
preferential orientation as the system evolves from one conformational state to the next. The introduction of a multiple alignment tensor analysis presented in this work essentially couples the slow intra-molecular conformational dynamics of the system to the preferential orientation of the molecule, while still maintaining the appropriate time and ensemble averaging properties of the RDC observables. The agreement between the theoretical and experimental RDCs was monitored using the R-factor:

\[
R_{\text{factor}} = \sqrt{\frac{\sum_{i=1}^{N} (X(i)_{\text{theory}} - X(i)_{\text{exp}})^2}{2 \sum_{i=1}^{N} (X(i)_{\text{exp}})^2}}
\]

where \(X(i)_{\text{theory}}\) and \(X(i)_{\text{exp}}\) are the theoretically determined and experimental observables respectively. The “optimal” torsional acceleration level (and hence the “optimal” conformational space sampling) for the reproduction of the experimental RDCs was found to be at \([E_b(\text{dih})-V_0(\text{dih})=960 \text{ kcal/mol}, \alpha(\text{dih})= 240 \text{ kcal/mol}]\). The R-factors for the agreement of the experimental RDCs were 0.37 to the x-ray structure, 0.32 to the CMD ensemble, and 0.20 to the optimal AMD ensemble. Without the use of the multiple alignment tensor approach, the agreement was 0.23.

The internal dynamics present in the different CMD and AMD simulations of PPACK-thrombin were assessed by calculating the \(S^2\) values relevant for the Lipari-Szabo type analysis (38) of \(^{15}\text{N}\) auto-relaxation data and for the \(S^2\) values derived from the AMD ensemble that best fits the RDC data, respectively. In all cases, molecular ensembles generated from the standard CMD simulations and the free energy weighted molecular ensembles generated from the AMD trajectories were superposed...
onto the backbone atoms (N, Cα, C’) of all heavy chain residues for the appropriate average structure. Order parameters, \( S^2 \), were calculated as:

\[
S^2 = \frac{1}{2} \left[ 3 \sum_{i=1}^{3} \sum_{j=1}^{3} < \mu_i \mu_j >^2 - 1 \right]
\]

where \( \mu_i \) are the Cartesian coordinates of the normalized inter-nuclear vector of interest. Others have shown that \( S^2 \) values calculated from standard MD simulations in this way were in excellent agreement with experimental \( S^2 \) values calculated using the Lipari-Szabo autocorrelation function approach (39).

**Residual local frustration analysis**

To analyze the crystal structure of PPACK-inhibited thrombin, we used an algorithm for determining residual local frustration, *i.e.* whether a contact between amino acid residues is energetically favorable or not in the folded state (40). This algorithm assesses residue-residue interactions by systematically perturbing the identity of individual residues and evaluating the resulting total energy change. The amino acids forming a particular contact are changed to other amino acids generating a set of decoys for which the total energy of the protein is recomputed. After constructing a histogram of the energy of the decoys and comparing the distribution to the native energy, cut-offs are implemented to identify minimally frustrated or highly frustrated residues. Energetically favorable contacts between residues are depicted by green lines and highly frustrated or energetically unfavorable contacts are red.
Results and Discussion

Backbone assignments of active site inhibited human thrombin

In an attempt to access loop regions of thrombin that appeared highly dynamic, we collected a suite of three-dimensional experiments for resonance assignments of human α-thrombin with PPACK bound at the active site. Of the 278 possible backbone amides, 245 were assigned for 88% coverage of the backbone. Backbone resonances (H\(^N\), N, C’, C\(^\alpha\), C\(^\beta\)) matched well with previously published assignments of the protein at 37°C (21), with some differences (Fig. 2-1). At 25°C, the C-terminus of the light chain including resonances G1f through D1a were visible as was N60g, R126, D178, and K236. Remarkably, the resonances for the entire γ-loop, T147 through Q151, which were missing from the previously published assignments were observed and assigned in our spectra. Resonances V31 through K36, R67 through G69, I79 and E80, I88, W141, and Q156, which were assigned at 37°C, were not observed at 25°C. The temperature dependence of the line broadening is most likely due to chemical exchange in these residues.
Residual dipolar couplings

RDCs were measured using spin-state selected TROSY 2D [$^1$H-$^{15}$N] correlation experiments to measure chemical shift differences between the TROSY and $^1$H anti-TROSY peaks in both isotropic and Pf1 phage aligned samples (42). Surprisingly, PPACK-thrombin aligned in only 3 mg/ml Pf1 phage and RDC values under these alignment conditions ranged from -39.07 to 48.68 (Fig. 2-2a). Molecular dipole calculations indicate that the thrombin molecule is highly polarized (43), and it is likely that the alignment is electrostatically-driven. RDCs could be measured for 209 of the 245 assigned residues. The RDCs calculated from the x-ray crystal structure (PDB ID: 1PPB, (44)) using the program, PALES (45), matched poorly with the
experimental values ($R^2=0.72$, Fig. 2-2b). Specifically, of 209 measured RDCs, 55 were more than 10 Hz different from those calculated for the best-fit alignment tensor for the crystal structure. The poor agreement was substantially worse than was observed for the more rigid protein, ubiquitin (4) and somewhat worse than was observed for IκBα (31). The largest discrepancies were found in the loop regions (Fig. 2-2b).

Figure 2-2. a) Residual dipolar coupling (RDC) values measured by TROSY-antiTROSY experiment collected at 800 MHz and 25°C using 3 mg/ml Pf1 phage for alignment. A schematic of the secondary structure is depicted above the plot. A break in the secondary structure diagram indicates where the light chain ends and the heavy chain starts. b) Plot of measured RDCs vs. those calculated from the x-ray structure (PDB code 1PPB) using PALES (45).

**Backbone dynamics experiments**

$R_1$, and $R_2$ relaxation rates and $^{15}\text{N}-^{1}\text{H}$NOEs were measured for PPACK-$^{2}\text{H}, ^{15}\text{N}$ thrombin at 600 MHz (Fig. 2-3). Much of the protein appeared rigid as indicated by high $^{15}\text{N}-^{1}\text{H}$NOEs and consistent $R_2/R_1$ ratios, but some regions had remarkably lower $^{15}\text{N}-^{1}\text{H}$NOE values (Fig. 2-3c). In particular, the C-terminus of the light chain (residues 14G-14M; 29-36) the γ-loop (residues 149-152; 185-193), and the 180s loop (residues 186A-186D; 229-233) had $^{15}\text{N}-^{1}\text{H}$NOE values lower than 0.4, similar to the disordered termini of the protein. Such low $^{15}\text{N}-^{1}\text{H}$NOE values
strongly indicate motions on ps-ns timescales. These regions also had lower than expected \( R_2 \) values (Fig. 2-3b) indicative of heterogeneous dynamics.

![Figure 2-3. Backbone relaxation data, a) \( R_1 \), b) \( R_2 \) and c) heteronuclear NOE, collected at 600 MHz, 25°C for \([^{1}H, ^{15}N]\) PPACK-inhibited human \( \alpha \)-thrombin as described in the SI section. d) \( S^2 \) values derived from the model-free analysis of the relaxation data using TENSOR2. The schematic of the secondary structure is depicted above the plot.](image)

The relaxation data was interpreted within the framework of the Lipari-Szabo model-free analysis using the program, TENSOR2. Initially, the \( R_1 \) and \( R_2 \) relaxation rates were used to calculate the overall rotational correlation time. A comparative analysis of isotropic, axially symmetric and fully anisotropic diffusion tensors, revealed that the axially symmetric model provided the best representation of the rotational diffusion as identified by a statistical Monte Carlo analysis (46; 47). The rotational correlation time was 16.3 ns, consistent with the molecular weight of thrombin, and \( D_\parallel/D_\perp \) of 0.868. Order parameters (\( S^2 \)) were calculated for 192 of the
thrombin resonances (Fig. 2-3d). The program TENSOR2 offers a variety of motional models of increasing complexity to fit the relaxation data (48). Over half of the residues (a total of 101 residues) could be fit with the simplest motional model (model 1) and nearly all of these had $S^2$ values above 0.8 and were in the core of the structure. Many residues (a total of 40) required fitting with a more complex motional model (model 5) and most of these had $S^2$ values below 0.8. These were located in the N-terminus and C-terminus of the light chain, the C-terminus of the heavy chain, the 30s loop (residues 30 and 36A; 51 and 58), and the $\gamma$-loop (residues 149A-150; 186-191). Residues in the 60s loop (60-60A and 60I; 82-83 and 91), the 90s loop (96 and 97A; 128 and 130), and the 180s loop (residues 184,185, 186D and 191; 225, 227, 232 and 237) also had $S^2$ values below 0.8 (Fig. 2-6a). The fact that so many resonances required model 5 fitting was strong evidence of heterogeneous fast timescale dynamics. In order to extract the contribution of chemical exchange to spin-spin relaxation, TROSY Hahn-echo experiments were performed (49). Of the 225 residues, only 15 had significant line broadening due to chemical exchange ($R_{ex} > 6$ s$^{-1}$, Fig. 2-4b). These residues were: 1F and 3 (3 and 11) of the light chain; 23 and 27 (44 and 48); 66 (97) leading up to the 70s loop; 87 (119) in the surface strand that connects the 70s loop to the 90s loop; 144, 147, 149a, 149c, and 151 (180, 183, 186, 188 and 192) of the $\gamma$-loop; 138, 157, and 199 (174, 198, and 245) in the core $\beta$-sheet; and 233 (281) that forms a kink in the C-terminal helix. Of these residues, only 23 (44) and 138 (174) required an $R_{ex}$ term to properly fit a Lipari-Szabo model to the relaxation data.
These direct measurements of $R_{ex}$ are much more accurate than the estimates obtained from the TENSOR2 analysis of the $R_1$, $R_2$, and $^{15}$N-$^{1}$H NOE data.

Figure 2-4. a) Plot of the $S^2$ values calculated from the relaxation data using TENSOR2 (black), compared to the $S^2$ values calculated from the CMD ensemble (blue) and compared to the $S^2_{RDC}$ values calculated from the RDC-optimized AMD ensemble (red). Regions of the heavy chain comprising the activation domain that were thought to become less dynamic in $\alpha$-thrombin as compared to prethrombin-2 are marked with horizontal bars. Regions with significant $S^2_{RDC}$ values that show little motional contribution from relaxation calculated $S^2$ values and measured $R_{ex}$ are boxed in gray. b) $R_{ex}$ values calculated from the TROSY Hahn-echo experiment. Horizontal line at $R_{ex}$=6 s$^{-1}$ is the threshold for significant contributions to relaxation from chemical exchange. c) Plot of the intensity of the TROSY peak for each resonance at different temperatures from 35°C down to 10°C. For most of the protein, the resonance intensity increases with temperature, however for some regions (boxed and enlarged) the relationship between intensity and temperature is indicative of complex slower timescale dynamics.

**Variable temperature experiments**

Resonances for most of the 30s loop and the 70s loop were neither observed in previous experiments (at 37°C) nor in our experiments (at 25°C) (Fig. 2-1). To further probe the temperature dependence of the thrombin backbone resonance intensities, we carried out variable temperature $^1$H-$^{15}$N TROSY-HSQC experiments. These
experiments revealed complex motions for several of the thrombin surface loops. For example, resonances in the γ-loop were not observed at 37°C (21), but are observed at lower temperatures (Fig. 2-4c). Resonances at the C-terminus of the light chain show similar behavior. In both cases, the decreased signal intensity at higher temperature indicates the region is undergoing exchange processes that are contributing to relaxation (50).

Relation between dynamics and local residual frustration

Over half of the residues in thrombin have high $S^2$ values, and formed a rigid core. This is consistent with analysis of the results of a residual local frustration analysis (40), which shows a strong network of minimally frustrated contacts throughout the core of the protein (Fig. 2-6b). It should be noted that this part of the molecule is also the part that is evolutionally conserved (18). Adjacent to the active site is Ile 16 (36), which makes a very large number of minimally frustrated contacts, and this corresponds to the new N-terminus generated by proteolytic activation that inserts into the core of the protein. Cross-peaks for the oxyanion hole, residues 193-196 (239-242), are not observed, and it is interesting that these residues are engaged in highly frustrated contacts. This is the only region of thrombin with frustrated contacts located in the core of the protein. Conversely, the dynamic surface loops have many highly frustrated contacts (Fig. 2-6b).

Molecular dynamics simulations

CMD simulations were carried out for 20 ns to reach the same timescale as the rotational diffusion time of the thrombin molecule (51; 23). The $S^2$ values calculated
from the CMD simulations matched well with those determined from the $R_1$, $R_2$ and $^{15}\text{N}-${\textsuperscript{1}$\text{H}}$NOE experiments (Fig. 2-4a). This result is a testament to the quality of contemporary force field used in this study, ff99SB, and is particularly remarkable in light of the many assumptions that have to be made when interpreting spin relaxation data (52; 53).

The excellent agreement between experimental and simulated $S^2$ values encouraged us to extend our theoretical study to investigate slower timescale motions using an enhanced conformational space sampling algorithm; AMD (1). AMD is an extended biased potential MD approach that allows for the efficient study of biomolecular systems up to timescales several orders of magnitude greater than accessible using standard CMD approaches, yet maintains a fully atomistic representation of the system. AMD has already been employed with great success to study the dynamics and conformational behavior of a variety of biomolecular systems including polypeptides, folded and natively unstructured proteins (37).

A series of five dual-boost AMD simulations were performed at increasing levels of acceleration (see SI). For each AMD simulation, a corrected canonical ensemble was obtained by performing a Boltzmann free energy reweighting protocol. Successively larger acceleration levels sample greater conformational space and the optimal conformational space sampling for the reproduction of the RDC data was obtained using singular value decomposition (SVD), following a protocol similar to that previously applied to ubiquitin and IκBα (31; 4). A multiple alignment tensor model was implemented as the PPACK-thrombin system exhibited considerable
flexibility, particularly on slower timescales (see Methods). The slower motions of the surface loops resulted in slight changes in the molecular alignment tensor such that the use of the multiple alignment tensors improved the agreement of the RDCs calculated from the AMD ensemble with the measured RDCs. This should not be confused with the molecular tumbling analysis implemented in TENSOR2 that determines whether the residues with relaxation parameters within the trimmed mean fit a particular tumbling model. Using this procedure, the optimal acceleration level, and hence the optimal conformational space sampling for the reproduction of the experimental RDCs was determined (see Methods). This optimal AMD ensemble recapitulated the measured RDC values much better ($R^2 = 0.92$) (Fig. 2-5b) than both the x-ray crystal structure ($R^2 = 0.72$) (Fig. 2-2b) and the CMD ensemble ($R^2 = 0.80$) (Fig. 2-5a). The most dramatic improvement in the agreement with the experimental RDCs was found for residues in the loop regions (Fig. 2-5c).

Figure 2-5. a) Plot of measured RDCs vs. those calculated from the CMD ensemble ($R^2 = 0.80$). b) Plot of measured RDCs vs. the RDCs calculated from the best-fit AMD ensemble ($R^2 = 0.92$). c) Plot of the “RDC improvement” as defined by $|\text{RDC}_{\exp} - \text{RDC}_{\text{AMD}}| - |\text{RDC}_{\exp} - \text{RDC}_{\text{xray}}|$. 
$S^2$ values were calculated from the standard MD simulation (see Methods), and the results from the CMD ensemble were in excellent agreement with the experimental values calculated from the Lipari-Szabo autocorrelation function analysis, with a few small deviations for the most flexible residues (Fig. 2-4a). A temporally heterogeneous distribution of dynamics is revealed in the AMD ensemble (as well as in the experiments), particularly in the light chain (residues 14G-14M; 29-36), the $\gamma$-loop, residues 149A-150 (186-191), the 180s loop, residues 184, 185, 186D and 191 (225, 227, 232 and 237), and the Na$^+$ binding site, residues 216–223 (264-271). The loops do not appear to hinge, but rather they appear to be appendages that extend from the rigid core in such a way that within the span of a few residues the $S^2$ values transition from completely ordered to highly disordered with motions on multiple timescales (Figs. 2-6b and 2-6c).

The $\gamma$-loop remains remarkably dynamic in the inhibited form. The resonances of this loop were missing from all liganded states in previous studies that were carried out at 37°C (21), and at first we were convinced that the previous studies had been inadvertently carried out with proteolyzed protein. The entire $\gamma$-loop was visible in our studies carried out at 25°C, but it showed remarkably low $S^2$ values, some of the lowest reported for a region other than the terminus of a protein this size. These results help explain the proteolytic sensitivity of the $\gamma$-loop, which is also called the autolysis loop, because it can be cleaved by thrombin itself. The fact that the $\gamma$-loop does not become ordered even in the active site inhibited form is remarkable.
Resonances in the 70s loop in anion binding exosite 1 were only observed in previous work when a ligand was bound at this exosite and it was thus thought to be dynamic (21). We also did not observe resonances for this loop at 25°C. Surprisingly, however, the $S^2_{\text{RDC}}$ values calculated from the AMD ensemble (which represents dynamics out to ms) revealed that the 70s loop is no more dynamic than other visible dynamic regions of the protein. Typically the reason that resonances are not observed is that they are transitioning between conformational states with different chemical shifts on an intermediate timescale in the NMR experiment. It is interesting that the AMD simulations do not show any evidence of large conformational fluctuations in this loop, in fact, it appears uniformly structured (Fig. 2-6c, 70s loop shown in brown). The results of the AMD simulations demonstrate that the absence of peaks in the 70s loop is not due wholly to dynamics as previously suggested (21) and illustrates the necessity of quantitative study of dynamics in thrombin.

Figure 2-6. a) Structure of PPACK-inhibited human α-thrombin depicting the ps-ns motions as determined from the backbone relaxation experiments. $S^2$ values are depicted as a backbone sausage model where the scale was red ($S^2 \leq 0.5$) to blue ($S^2 \geq 0.8$) with unobserved residues in gray. b) Structure of PPACK-inhibited human α-thrombin showing the minimally-frustrated contacts (green) and the highly-frustrated contacts (red); the highly frustrated oxyanion hole and the minimally frustrated N-terminus of the heavy chain are circled. c) Structural ensemble of PPACK-inhibited human α-thrombin calculated using AMD that best matched the RDC data. The dynamic regions are colored: light chain (light green), 60s loop (cyan), 70s loop (brown), γ-loop (red), 180s loop (purple) Na$^+$ binding site (blue). d) Structure of PPACK-thrombin depicting residues with measured $R_{ex} > 6 \, \text{s}^{-1}$ (red), those with predicted supra-τc motions (orange) and unobservable residues (gray).
Conclusion

Due to its size and complex folding pathway, direct measurements of backbone dynamics in thrombin have not been reported previously. In this first report, we wished to explore the solution structure of active human α-thrombin with a small inhibitor bound at the active site. Parts of the light chain, the entire γ-loop, and tips of the 60s and 90s loops and the Na$^+$ binding site had order parameters lower than 0.8 indicative of substantial dynamics on the ps-ns timescale. TROSY Hahn-echo experiments revealed a substantial number of residues undergoing relaxation due to chemical exchange including residues in the light chain, the beginning of the heavy chain, one residue in the 90s loop, one residue in the strand that connects the 70s to the 90s loop, much of the γ-loop, two residues in regions proximal to the light chain, and three residues in the β-sheet core. The emerging picture was thus of a very dynamic molecule.

To obtain a more complete quantitative description of the dynamic properties of the system across multiple timescales, MD simulations were employed. AMD simulations were calibrated using RDC data and these extended MD simulations provided a more complete picture of temporally heterogeneous dynamic behavior of observed residues. Together the experiments and simulations yield an unprecedented molecular ensemble view of thrombin that truly represents the solution structure, revealing that PPACK-thrombin possesses a largely rigid core corresponding to the evolutionarily conserved protease with highly dynamic and less conserved loop appendages. The order parameters obtained from the AMD simulations ($S^2_{RDC}$ values)
were often much lower than the $S^2$ values calculated from the relaxation data revealing the true extent of thrombin’s motions on longer timescales. Many residues had high $S^2$ values and much lower $S^2_{RDC}$ values, but did not appear to be undergoing R$_{ex}$ (Fig. 2-6d). These residues may be undergoing motions slower than the slow limit of the relaxation calculated $S^2$ values ($\tau_c$, ~20ns) but faster than the fast limit of the chemical exchange regime (~100ns). This is strong evidence of motions in the supra-$\tau_c$ regime (30). In addition, the AMD simulations revealed the dynamic behavior of the residues that were invisible in the NMR experiments.

Previous dynamic studies of enzymes have mostly observed that loops surrounding the active site display a hinging motion on the ms timescale between an open and closed state. Our combined relaxation experiments and simulations revealed multiple timescale dynamics of the thrombin loops, thus differing from this paradigm. The complex temperature dependence of resonance peak intensities in these regions is further evidence of the temporal heterogeneity of the loop motions. The loops do not appear to hinge, but rather they appear to be flexible appendages that extend from the rigid core. One could say that thrombin resembles an anemone in the sense that its active site is surrounded by loops that exhibit both fast and long timescale dynamics.

It has long been proposed that one feature of zymogen activation is the further structuring of four structural segments that were observed to be deformed and flexible in crystal structures of the zymogen form: residues light chain up to residue 19, the $\gamma$-loop (residues 142–152; 178-195), the 180s loop (residues 184–193; 225-239), and the Na$^+$ binding site (residues 216–223; 264-271), which together form the activation
domain (54). These regions also display temporally heterogeneous dynamics. Our results show that, in fact, zymogen activation does not result in ordering of these regions (marked with horizontal bars in Fig. 2-4) since they remain highly dynamic even in the active form albeit at longer timescales. The tracks of residues undergoing motions on the supra-$\tau_c$ timescale appear to surround a track of residues undergoing $R_{ex}$ (Fig. 2-6d). These apparent pathways of dynamic residues may be particularly important in the allosteric communication between the anion binding exosites and the active site (9; 10).

Acknowledgments

We thank Tracy Handel and Michael Massiah for helpful discussions. Financial support was from R01 HL070999. Some NMR experiments were performed at NMRFAM. B.F. acknowledges training support from the Molecular Biophysics training grant (T32 GM08326), and from the American Heart Association. P.G. acknowledges training support from the Cell and Molecular Genetics training program. The computational work was funded in part by the NIH, NSF, and HHMI, the Center for Theoretical Biological Physics, the National Biomolecular Computing Resource, and the San Diego Supercomputer Center.

Chapter 2 is a minimally modified reprint of the material as it appears in Brian Fuglestad, Paul M. Gasper, Marco Tonelli, J. Andrew McCammon, Phineus R. L. Markwick, and Elizabeth A. Komives, The Dynamic Structure of Thrombin in Solution, Biophysical Journal 2012. The dissertation author was the primary investigator and author of the computational aspects of this paper.
Supplemental Information

Details of the thrombin preparation and purification

Freshly transformed *E. coli* were acclimated to D\textsubscript{2}O by sequential growths in M9ZB (10 ml, 12 h), M9 (10 ml, 1 h), 50% D\textsubscript{2}O-M9 (2 ml, 1 h), and 90% D\textsubscript{2}O-M9 (10 ml, 12 h). Typically, 1 L of D\textsubscript{2}O minimal M9 media with \(\text{^{15}NH}_4\text{Cl} (2 \text{ g/L})\) or \(\text{^{15}NH}_4\text{Cl} (2 \text{ g/L})\) and \(\text{^{13}C}\text{-glucose} (8 \text{ g/L})\) was grown for 24 hr to O.D.\textsubscript{600} of 0.8 and protein production was initiated by addition of 1 mM IPTG. Cells were grown for an additional 24 hrs and pellets were frozen at -80°C. Inclusion bodies were isolated from thawed cell pellets, solubilized, purified, reduced and dialyzed as previously described (32).

Refolding of “prethrombin-2” was initiated by drop wise rapid dilution into a 50x volume of buffer containing 50 mM Tris/HCl pH 8.5, 0.6 M arginine HCl, 20 mM CaCl\textsubscript{2}, 10% (v/v) glycerol, 0.2% (w/v) Brij58, 50 mM NaCl followed by rapid stirring for 3 days. The protein solution was then concentrated to 500 mL using a Millipore Pellicon 10 kDa MWCO unit, 1 mg/ml PEG8000 was added, then the solution was dialyzed twice with 4 L of 25 mM Tris/HCl pH 7.4, 2 mM EDTA, and 0.1% (w/v) PEG 8000.

Purification of the refolded “prethrombin-2” was achieved by ion exchange chromatography of the dialyzed solution diluted 1:1 with Tris/HCl pH 7.4, 1 mg/ml PEG 8000, on a SP Sepharose Fast Flow column (GE Healthcare Life Sciences) with a gradient of Tris/HCl pH7.4, 1 mg/ml PEG8000 from 0.0 to 1.0 M NaCl. Typical yields of properly folded “prethrombin-2” ranged from about 6-12 mg. Activation to
α-thrombin was achieved by incubation of the “prethrombin-2” diluted to 50 mL with 50 mM Tris/HCl pH 7.4, 1 mL 1 M CaCl₂, 1 mg/ml PEG8000, 5% (v/v) glycerol with 250 mg AEBSF pretreated E. carinatus venom (Sigma-Aldrich Cat. # V8250) at room temperature. Optimal activation times varied from 2-10 hr, as predetermined by a time dependant thrombin activation test. At the end of the activation period, the protein solution was diluted 1:1 with 50 mM Tris/HCl pH 7.4, 1 mg/ml PEG 8000 and loaded onto a MonoS cation exchange column (GE Healthcare Life Sciences). Separation of active α-thrombin from preactive (prethrombin-2, meizothrombin) and auto-protealyzed forms (β- and γ-thrombin) is efficiently achieved using MonoS chromatography with a gradient of 100 mM to 500 mM NaCl in 25 mM phosphate pH 6.5.

Details of the NMR experimental methods

¹H chemical shifts were referenced against DSS methyl protons at 0 ppm with ¹⁵N and ¹³C chemical shifts calculated from ¹H chemical shift (55). All data were processed with the NMRPipe suite (56) and visualized with Sparky (57). Assignments were performed manually and independently of those previously published (21).

Variable temperature experiments were recorded with ¹H-¹⁵N TROSY 2D [¹H-¹⁵N] correlation experiments (64 scans and 2048x256 complex data points) at 283°K, 293°K, 298°K, 303°K, and 308°K on a sample of 100 □M ¹⁵N, ²H-labeled PPACK-thrombin. Peak intensities were measured using Sparky as described by others working with similarly large proteins (58).
Relaxation experiments were performed at 600 MHz on a $^{15}$N, $^2$H-labeled thrombin sample at 150 μM. $^{15}$N–{$^1$H}NOEs were measured with presaturated and unsaturated experiments that were recorded interleaved and in duplicate to estimate error based on standard deviation between each duplicate $^{15}$N–{$^1$H}NOE value. Experiments were collected with 256 scans in a 1024x120 data matrix with a recycle delay of 4.5 s. The $^{15}$N–{$^1$H}NOE values were calculated from the ratio of peak intensities from the presaturated to the unsaturated experiments determined in SPARKY. $^{15}$N longitudinal ($T_1$) and transverse ($T_2$) relaxation experiments were performed at 600 MHz in an interleaved manner with 2 duplicate points for error estimation. Experiments were collected with 64 scans, 1024x200 complex data points, $T_1$ delays of 1, 50, 100(x2), 150, 350, 700(x2), 1000 and 1700 ms, and $T_2$ delays of 0, 17.6, 35.2(x2), 52.8, 70.4, 88.0(x2), 123.2, and 176.0 ms. Peak intensities were measured in Sparky and errors were linearly interpolated from duplicate measurement errors by the perl script sparky2rate (http://xbeams.chem.yale.edu/~loria/sparky2rate, P. J. Loria, Yale University) before non-linear least squares fitting with the program CurveFit (http://www.palmer.hs.columbia.edu/software/curvefit.html, A. G. Palmer III, Columbia University).

PPACK-thrombin dynamic analysis was performed using the program TENSOR2 (46). To estimate the overall rotational diffusion, we used only those residues with an $R_2/R_1$ ratio within 1 standard deviation of the average, and low errors. These data were best fit by the axially symmetric tensorial model for rotational diffusion of PPACK-thrombin and yielded a $\tau_c$ value of 16.3 ns, consistent with the
molecular weight of thrombin and $D_\parallel/D_\perp$ of 0.868. $S^2$ values were calculated using the
Lipari-Szabo model free approach (38; 48). When necessary, extended models
accounting for chemical exchange and motions on two timescales were employed
(59).

$^1$H-$^{15}$N backbone Residual Dipolar Couplings (RDCs) were measured by
comparing changes in the chemical shift differences between the $^1$H TROSY and anti-
TROSY peaks in an interleaved spin-state selected TROSY-anti-TROSY experiment
(42). Aligned samples of 100 $\mu$M $^{15}$N, $^2$H-labeled thrombin were prepared by using
ultracentrifugation to exchange Pf1 phage (Asla Biotech) into the buffer described
above (including 10% D$_2$O), then adding Pf1 to a total concentration of 3mg/ml.
Experiments on the aligned (isotropic) sample at 800 MHz had 128 scans (64 scans),
with 2048 x 256 complex data points collected over about 20 hours (10 hours).

$R_{ex}$ was measured at 800 MHz by using the TROSY based Hahn-echo pulse
sequence as previously described (49) with a spin-echo delay of $2\tau = 2/J_{NH} = 21.6$ ms.
The experiment was performed interleaved and in triplicate with 64 scans and
1024x256 complex data points. Peak intensities of each spin state relaxation
experiment ($I^\alpha$, $I^\beta$ and $I^{2HzNz}$) were used to calculate $R_{ex}$ according to the relations
(49):

$$R_{ex} = R_2^\alpha - R_1^{2HzNc} / 2 - \eta_{xy}(\kappa - 1) + R_1^N / 2$$

$$R_2^\alpha - R_1^{2HzNc} / 2 = \frac{\ln[I^\alpha(2\tau)/I^{2HzNc}(\tau)]}{-2\tau}$$

$$-\eta_{xy} = \frac{\ln[I^\beta(2\tau)/I^\alpha(2\tau)]}{4\tau}$$
with \( \tau \) calculated as the trimmed mean of \( 1+(R_2^n-R_1^{2HzN/2})/\tau_{xy} \) for all non-exchanging residues. \( R_1^N \) is negligible for the 34 kDa thrombin molecule.

**Details of the AMD protocol**

In the AMD approach a reference or 'boost energy', \( E_b \), is defined which lies above the minimum of the potential energy surface (PES). At each step in simulation, if the instantaneous potential energy, \( V(r) \), lies below the boost energy a continuous, non-negative bias potential, \( \Delta V(r) \), is added to the actual potential. If the potential energy is greater than the boost energy, it remains unaltered. The application of the bias potential results in a raising and flattening of the PES, decreasing the magnitude of the energy barriers and thereby accelerating the exchange between low energy conformational states, while still maintaining the essential details of the potential energy landscape. Explicitly, the modified potential, \( V^*(r) \), on which the system evolves during an AMD simulation is given by:

\[
V^*(r) = V(r), \quad V(r) \geq E_b
\]

\[
V^*(r) = V(r) + \Delta V(r), \quad V(r) < E_b
\]

where the bias potential, \( \Delta V(r) \), is defined as:

\[
\Delta V(r) = \frac{(E_b - V(r))^2}{E_b - V(r) + \alpha}
\]

The extent of acceleration (i.e., how aggressively the conformational space is sampled) is determined by the choice of the boost energy, \( E_b \), and the acceleration parameter, \( \alpha \). Conformational space sampling can be enhanced by either increasing the boost energy, or decreasing \( \alpha \). During the course of the AMD simulation, if the
potential energy is modified, the forces on the atoms are recalculated for the modified potential and the use of the bias potential defined above ensures that the derivative of the modified potential will not become discontinuous at points where \( V(r) = E_b \). In the present work, we have implemented a “dual boost” AMD approach (60), in which two acceleration potentials are applied simultaneously to the system: The first acceleration potential is applied to the torsional terms only, and a second, weaker acceleration is applied across the entire potential. The background, total acceleration potential enhances the translational-rotational diffusion properties of the solvent molecules, thereby facilitating slow diffusive motions in the solute. As such, the dual boost AMD protocol represents a unified approach to efficiently sample both the torsional degrees of freedom and the diffusive motions of system.

One of the favorable characteristics of the AMD approach is that it yields a canonical average of an ensemble, so that thermodynamic and other equilibrium properties can be accurately determined. The corrected canonical ensemble average of the system is obtained by reweighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, \( \exp[\beta \Delta V(r_{t(i)})] \) at that particular point.

To calculate residual dipolar couplings (RDCs) from the AMD ensembles, we used an approach similar to those detailed in previous studies on ubiquitin (4) and IB (31). Atomic coordinates for PPACK-thrombin were obtained from the x-ray crystal structure [PDB ID: 1PPB (44)] and the system was placed in a periodically repeating box with 11495 water molecules and three Cl- counter-ions. Initially a set of
five standard conventional MD (CMD) simulations was performed using an in-house modified version of the AMBER 10 code (61). For each of these simulations, a different random seed generator for the Maxwellian distribution of atomic velocities was employed and, after standard energy minimization and equilibration procedures, a 20 ns production run MD simulation was performed under periodic boundary conditions with a time-step of 2 fs. Bonds involving protons were constrained using the SHAKE algorithm. Electrostatic interactions were treated using the Particle Mesh Ewald (PME) method with a direct space sum limit of 10 Å. With the exception of the PPACK inhibitor, for which an in-house gaff force field was generated, the ff99SB force field was used for the solute residues and the TIP3P water force field was employed for the solvent molecules. These initial five 20 ns CMD simulations acted as a control set and were used as the starting point for the AMD simulations. These simulations also provided the average (unbiased) dihedral angle energy, $<V_0^{\text{dih}}>$ and average total potential energy, $<V_0^{\text{tot}}>$, used to define the acceleration parameters in the AMD simulations.

A series of five “dual boost” AMD simulations (which also used the AMBER 10 code) were performed for 10,000,000 steps (the equivalent of 20 ns) at increasing levels of acceleration. For the torsional acceleration, the acceleration parameter, $\alpha^{\text{dih}}$ was fixed at 240 kcal/mol and the torsional boost energy, $E_b^{\text{dih}}$ for the four acceleration levels was set at 480, 720, 960 and 1200 kcal/mol above the average dihedral angle energy, $<V_0^{\text{dih}}>$ estimated from the unbiased 20 ns CMD simulations. In all AMD simulations, a fixed background total acceleration potential
was employed with a boost energy, $E_b(tot)$, defined such that $[E_b(tot) - V_0(tot)]$ was equal to 0.16 kcal/mol times the number of atoms in the simulations cell (NASC), and the acceleration parameter, $\alpha(tot)$ was also set to 0.16 kcal/mol*(NASC). The physical conditions, force fields and all other simulation parameters employed in the AMD simulations were identical to those described for the CMD trajectories.

For each AMD simulation, a corrected canonical ensemble was obtained by performing the Boltzmann free energy reweighting protocol described above using the bias potential block averaging method (37) to remove statistical noise errors. In this way, five representative free energy weighted molecular ensembles were generated at each of the four acceleration levels, along with the five unbiased CMD simulations. Each free energy weighted molecular ensemble was divided into molecular clusters using an in-house RMSD-based QR factorization clustering algorithm. In a series of initial tests, we varied the clustering parameters until optimal convergence in the computed RDC values was obtained. Under optimal clustering conditions, we found that each member of a given cluster had a backbone RMSD ranging up to 0.6-0.7 Å with respect to the average structure for the cluster. A Singular Value Decomposition (SVD) analysis (62) was performed to determine the optimal alignment tensor for each cluster, and N-H$^N$ RDCs were calculated. A weighted average of the calculated RDC values was obtained using the relative number of structures in each cluster. Finally the RDCs for each free energy weighted ensemble associated with a given acceleration level were averaged.
Chapter 3:

Allosteric Networks in Thrombin Distinguish Procoagulant vs. Anticoagulant Activities

Abstract

The serine protease α-thrombin is a dual action protein that mediates the blood-clotting cascade. Thrombin alone is a procoagulant, cleaving fibrinogen to make the fibrin clot, but the thrombin-thrombomodulin complex initiates the anticoagulant pathway by cleaving protein C. A thrombomodulin fragment consisting of only the 5th and 6th EGF-like domains (TM56) is sufficient to bind thrombin, but the presence of the 4th EGF-like domain (TM456) is critical to induce the anticoagulant activity of thrombin. Crystallography of the thrombin:TM456 complex revealed no significant structural changes in thrombin, suggesting that TM4 may only provide a scaffold for optimal alignment of protein C, for its cleavage by thrombin. However, a variety of experimental data have suggested that the presence of TM4 may affect the dynamic properties of the active site loops. In the present work we have used both conventional and accelerated molecular dynamics simulation to study the structural dynamic
properties of thrombin, thrombin:TM56 and thrombin:TM456 across a broad range of timescales. Two distinct, yet interrelated allosteric pathways are identified that mediate both the pro- and anticoagulant activities of thrombin. One allosteric pathway, which is present in both thrombin:TM56 and thrombin:TM456 directly links the TM5 domain to the thrombin active site. The other allosteric pathway, which is only present on slow timescales in the presence of the TM4 domain, involves an extended network of correlated motions linking the TM4 and TM5 domains and the active site loops of thrombin.

**Introduction**

Thrombin is a dual-action serine protease that plays a pivotal role in the blood-clotting cascade. The specific functional activity of thrombin is mediated by its interaction with the co-factor thrombomodulin (TM). In the absence of TM, thrombin acts as a procoagulant, cleaving fibrinogen to fibrin. However, when in complex with TM, thrombin acts as an anticoagulant by cleaving and thereby activating protein C (5; 6). TM contains six EGF-like domains; the 5\textsuperscript{th} domain interacts directly with thrombin at the fibrinogen binding site, anion binding exosite 1 (ABE1). It has been shown that a TM fragment consisting of only the 5\textsuperscript{th} and 6\textsuperscript{th} EGF-like domains (TM56) is sufficient to bind thrombin and inhibit fibrinogen cleavage, but the additional presence of the 4\textsuperscript{th} EGF-like domain of TM (TM456) is critical to induce the anticoagulant activity of thrombin (63; 64). TM456 binding to thrombin significantly increases the association rate, $k_{a}$, of a variety of active site directed inhibitors of thrombin (65-68),
and the $k_a$ for protein C binding is 1000-fold higher for the thrombin:TM456 complex compared to thrombin alone (69).

The mechanism by which TM456 enhances protein C cleavage is not yet fully understood and is particularly intriguing because ABE1 is distal to the active site and the essential TM4 domain makes no direct contact with thrombin whatsoever (Fig. 3-1) (7). The simplest explanation for the dramatically increased association rates is that TM456 alters the structure of thrombin, most notably the conformation of the loops that surround the active site, a process generally referred to as structural- or enthalpic-allostery (11). However, a comparative analysis of the x-ray crystal structures of thrombin (1PPB) (70) and thrombin:TM456 (1DX5) (7) revealed no significant structural differences in the thrombin active site loops. It should be noted that in these x-ray crystal structures the active site is occupied by an inhibitor, potentially stabilizing the loops in a closed conformation.

To date, the most robust argument for the role of TM4 in the activation of protein C comes from a recent molecular modeling study (7) which suggests that TM4 forms an extended binding surface for protein C, providing optimal alignment for insertion into the active site and subsequent cleavage. Critical electrostatic interactions between protein C and residues in TM4, including Glu382 in the $\beta4$-$\beta5$ loop, Asp398 and Glu357 as well as hydrophobic interactions with the aromatic residues Tyr358 and Phe376 were identified. The importance of these residues for the activation of protein C has been empirically validated by alanine-scanning mutagenesis and H/D exchange experiments (71; 72). However, the proposed ‘docking and optimal alignment’
mechanism may be insufficient to explain the 1000-fold increase in observed association rates, particularly for smaller substrates and other inhibitors, which may enter the active site of thrombin without directly interacting with the TM4 domain.

Figure 3-1. The structure of thrombin:TM456. Top: The structure of thrombin looking directly into the active site. Important regions are highlighted by spheres: 90\textsubscript{CT} loop (red), 148\textsubscript{CT} (gamma) loop (cyan), 170\textsubscript{CT} loop (green), 186\textsubscript{CT} loop (purple), 220\textsubscript{CT} loop (blue) and the 60\textsubscript{CT}-insertion (yellow). The ABE1 binding site comprises the 30\textsubscript{CT} loop (orange) and the 70\textsubscript{CT} loop (magenta). Chymotrypsin and sequential residue numbering for these regions are provided in the SI. Bottom: The domain architecture of thrombin:TM456: Thrombin is depicted as ribbon: light chain (coral), heavy chain (gray) and TM as surface: TM4 (violet), TM5 (tan), TM6 (blue).
Several studies including fluorescence (8), H/D exchange (9) and isothermal titration calorimetry (ITC) (10) have identified changes in the thrombin active site region that occur on binding to different constructs of TM in the absence of protein C. These studies suggest that the presence of TM4 may affect the dynamic properties of both the active site and the loops surrounding it, a process referred to as entropic allostery (12; 73; 13). A very recent NMR/molecular dynamics simulation study on thrombin bound to the inhibitor D-Phe-Pro-Arg-chloromethylketone (PPACK) identified significant dynamic motions in the active site loops across timescales ranging from ps to tens of ms even with inhibitor bound (74).

In the present work we employ both conventional molecular dynamics (CMD) simulation and an enhanced conformational space sampling algorithm, accelerated molecular dynamics (AMD) (1) to study the conformational behavior and potential allosteric pathways in thrombin, thrombin:TM56 and thrombin:TM456. AMD is an extended biased potential MD approach that allows for the efficient study of biomolecular systems up to timescales several orders of magnitude greater than those accessible using CMD, while maintaining a fully atomistic representation of the system. AMD has already been employed with great success to study the dynamics and conformational behavior of a variety of biomolecular systems including polypeptides and both natively unstructured and folded proteins (37). Community network models (75) have been employed to identify potential allosteric pathways. This information is supplemented by both a generalized correlated motion analysis
(76) and the calculation of backbone NH$^N$ order parameters to compare the dynamic behavior of thrombin in the three different systems across a variety of timescales.

**Results**

We initially performed a set of six independent 20 ns CMD simulations for each of the three systems: thrombin, thrombin:TM56 and thrombin:TM456. For all three systems a rather large conformational transition in the active sites loops was observed during the equilibration procedure. Central to this conformational relaxation was a reorientation of both the extended 148$_{\text{CT}}$ γ-loop (residues 178 to 192) and the 220$_{\text{CT}}$ loop (residues 265 to 274) forming a more ‘open’ active site pocket. Notably, the extended 148$_{\text{CT}}$ γ-loop moves closer to the ABE1 70$_{\text{CT}}$ loop (residues 93 to 109) (see SI for residue numbering details). This observation confirms the hypothesis that the presence of the inhibitor in the available x-ray crystal structures, which was removed before performing the CMD simulations, promotes a closed loop conformation. After this initial structural relaxation, the conformational dynamics of each of the three systems rapidly stabilized and the backbone root mean square distance (RMSD) to the average structure for the thrombin molecule was found to be 1.4 Å, 1.2 Å and 1.2 Å for thrombin, thrombin:TM56 and thrombin:TM456 respectively.

The molecular ensembles generated from the CMD simulations were subjected to a community network analysis, which identifies communities (clusters of highly connected residues) based on residue-by-residue correlation and proximity (75). A measure of ‘betweenness’ of the intercommunity edges (see Methods) affords the
identification and characterization of potential allosteric pathways. Community network analysis results for thrombin, thrombin:TM56 and thrombin:TM456 are shown in Fig. 3-2.

Figure 3-2. Community analysis results (left) and corresponding structures colored by community (right) for a) thrombin b) thrombin:TM56 and c) thrombin:TM456. The communities are represented by spheres and the edge width is proportional to the cumulative betweenness of intercommunity edges, which is a measure of the strength of the potential allosteric interaction between communities.

Several interesting and highly reproducible communities of residues were observed in the analyses in the majority of the CMD trajectories. In all cases, TM4, TM5 and TM6 are each represented by a single community. By contrast, the thrombin molecule in all systems consists of multiple interconnected communities divided between most notably, the TM binding site (ABE1) and the proteolytic active site, including the catalytic triad (H57\textsubscript{CT}, D102\textsubscript{CT} and S195\textsubscript{CT}), the 60\textsubscript{CT} insertion and the
active site loops (90_{CT}, 148_{CT} (gamma), 170_{CT}, 186_{CT} and 220_{CT}). However, the number of thrombin communities differs from one system to the next: On average, isolated thrombin ostensibly comprised 8 distinct large communities, whereas the thrombin molecule in thrombin:TM56 and thrombin:TM456 comprised only 5-6 large communities on average. Similarly, the metric of cumulative betweenness of inter-community edges, which is a direct measure of the strength of the potential allosteric interaction between communities, is greater in the two thrombin:TM complexes compared to thrombin alone. In the thrombin:TM56 system, a strong allosteric pathway is consistently observed linking the TM5, ABE1 and active site communities. The TM4 domain, which forms a strong allosteric connection to TM5, acts to further enhance the strength of this allosteric pathway. A second and very important difference between the community networks for thrombin:TM56 and thrombin:TM456 is the presence of a strong communication between TM5 and the 148_{CT} gamma-loop of thrombin, which is only observed in the thrombin:TM456 construct.

The fast timescale dynamics of thrombin in the isolated thrombin, thrombin:TM56 and thrombin:TM456 systems were analyzed by calculating backbone NH order parameters, which provide a quantitative measure of the extent of (internal) reorientation dynamics of the N-H bond vector per residue (Fig. 3-3a). Similar to the recent NMR/MD simulation study on thrombin:PPACK, a strongly heterogeneous distribution of fast timescale motions was observed in all three systems (74). In the isolated thrombin system, order parameters in the 30_{CT} loop (residues 53 to 61), which forms part of ABE1, and in the 60_{CT} insertion (residues 83 to 91) were as low as 0.66
and 0.64 respectively. By contrast in both the thrombin:TM complexes these order parameters were substantially higher with the lowest value being 0.76. Reduced fast timescale dynamics (higher order parameters) in the 30CT loop are readily explained as this region makes direct contact with the TM5 domain. However, the same cannot be said for the 60CT insertion. Significantly, the community network analysis models for both thrombin:TM56 and thrombin:TM456 (Fig. 3-2) identified a strong allosteric pathway connecting TM5 to the active site of thrombin via the 60CT insertion.

Figure 3-3. a) Thrombin residue NH\textsuperscript{N} order parameters for isolated thrombin (top), thrombin:TM56 (middle) and thrombin:TM456 (bottom) systems. The black line depicts fast timescale order parameters obtained from a series of 20 ns CMD simulations. The red line depicts order parameters obtained from a set of free energy weighted AMD trajectories at acceleration level 2, probing dynamics on timescales up to several tens of µs. Order parameters for acceleration level 1 were intermediate and are not shown on the figure for clarity. b) Slow timescale (level 2 acceleration) differential NH\textsuperscript{N} backbone order parameters for thrombin:TM56 (black) and thrombin:TM456 (red) compared to the isolated thrombin system. Positive values indicate an increase in the amount of reorientation dynamics relative to the isolated thrombin system, negative values indicate less dynamics.
Higher order parameters in these same loops were also observed experimentally (74). The dynamic communication observed between the active site, the $30_{CT}$ loop and the $60_{CT}$ insertion is a prime example of entropic allostery. No other significant differences in the fast timescale NH$^N$ order parameters for the three systems were observed. Calculated backbone root mean square atomic fluctuations also revealed no other significant differences and the average structures of thrombin for the three systems were remarkably similar (backbone RMSD $< 0.5$ Å).

In order to access $\mu$s-ms timescale motions more typical of allosteric regulation, a series of accelerated molecular dynamic (AMD) simulations were performed at two acceleration levels: A moderate acceleration level (level 1), which samples configurational dynamics up to timescales of several hundred ns, and an aggressive acceleration level (level 2), probing dynamics up to several tens of $\mu$s (see SI). After free energy weighting of the AMD trajectories, a representative ensemble of structures was obtained. Whereas no substantial domain reorientation was observed for the TM EGF domains in the CMD, the AMD trajectories showed some domain reorientation though with limited extent of motion. From the free energy weighted trajectories, averaged thrombin NH$^N$ order parameters were calculated (Fig. 3-3a). By contrast to the fast timescale dynamics, marked differences in the order parameters were observed for the three systems, particularly at the most aggressive acceleration level (Fig. 3-3b). Both TM-bound systems exhibit considerably less dynamics in the $30_{CT}$ loop (residues 53 to 61) and the $60_{CT}$ insertion (residues 83 to 91) consistent with what was observed for the fast timescale order parameters. However, on slow
timescales thrombin:TM456 exhibits remarkably more reorientation dynamics in the active site loop regions; the \(90_{CT}\) loop (residues 122 to 132), the \(148_{CT}\) loop (residues 184 to 192), the \(170_{CT}\) loop (residues 213 to 221), the \(186_{CT}\) loop (residues 228 to 234) and the \(220_{CT}\) loop (residues 265 to 274). By contrast, thrombin:TM56 shows either no more or in some regions even less slow timescale dynamics compared to the isolated thrombin system.

A residue-by-residue cross-correlation analysis (76) (see Methods) applied to the free energy weighted molecular ensembles obtained at the most aggressive acceleration level also revealed striking differences between the three systems (Fig. 3-4). In all three systems we observed cross-correlated motion between the \(30_{CT}\) loop and the \(60_{CT}\) insertion, and the strength of this cross-correlation was markedly enhanced in the thrombin:TM456 system. Additionally, the thrombin:TM456 system exhibited a complex pattern of strongly correlated motions across the active site loops (Fig. 3-5) directly coinciding with the observation of enhanced slow timescale dynamics in the level 2 AMD trajectories of the thrombin:TM456 complex (Fig. 3-3b). In contrast, correlations between the corresponding regions in both the thrombin:TM56 and the isolated thrombin systems were very weak or absent.
Figure 3-4. Normalized comparative analysis of observed thrombin cross-correlated motions in a) isolated thrombin b), thrombin:TM56 (upper triangle) and b) thrombin:TM456 (lower triangle) obtained from representative free energy weighted (level 2) AMD molecular ensembles. Blue boxes have been inserted to highlight the most significant cross-correlated motions common to all three systems. In the case of thrombin:TM456, significant cross-correlated motions are observed between the active site loops (black boxes) which are either absent or very weak in the case of thrombin and thrombin:TM56 (green boxes shown for comparative purposes). Correlated motions extend into the light chain. The functional consequences of this phenomenon are not clear but are in agreement with mutagenesis experiments indicating a potential allosteric role for the light chain (77).
Figure 3-5. Slow timescale cross-correlated molecular motions for thrombin in the thrombin:TM456 system projected onto a random snap-shot taken from the thrombin:TM456 level 2 AMD trajectory. For reasons of clarity, the TM domains are not shown. The catalytic triad [residues His$^{79\text{CT}}$, Asp$^{135\text{CT}}$ and Ser$^{241\text{CT}}$] are in yellow. The active site loops that exhibit enhanced slow timescale dynamics and strongly inter-connected correlated motions (cf. Fig. 4b) are shown using a ‘ball-and-stick’ representation: Blue: extended 90$_{\text{CT}}$ loop (residues 125 to 140), red: extended 148$_{\text{CT}}$ loop (residues 175 to 192), green: 220$_{\text{CT}}$ loop (residues 263 to 274), orange: residues 103 to 110 which form part of the 70$_{\text{CT}}$ loop. Regions in and neighboring ABE1, the 30$_{\text{CT}}$ loop (residues 53 to 61, tan) and 60$_{\text{CT}}$-insertion (residues 83-91, black) respectively shown in cartoon format are more rigid than in the isolated thrombin system but also show extensive correlated motions on both fast and slow timescales.

To confirm that the pattern of strongly correlated motions observed in thrombin:TM456 complex were due to TM4, we extended our analysis to include the TM domains. To circumvent complications from the re-orientation dynamics of the TM domains, multiple cross-correlation analyses were performed after superposing the molecular ensemble on different regions of the molecule, and then the complete cross-correlated dynamics map (Fig. 3-6) was reconstructed. The results clearly show that the complex network of correlated motions observed in thrombin:TM456 propagate from the TM4 domain through the TM5 domain and into the thrombin molecule (Fig. 3-6). Global reorientation dynamics of TM4 contribute significantly to
the entire network of correlated motion by inducing global reorientation dynamics of TM5 and hence the specific orientation and interactions of TM5 with thrombin. Additionally, inter-domain, residue-by-residue correlations are observed between regions of TM4 (residues 359 to 369) and TM5 (residues 404 to 414). Global domain correlations and specific residue correlations in TM together extend into ABE1 of thrombin, including the 30\textsubscript{CT} loop (residues 53-61), the 60\textsubscript{CT} insert (83 to 91) and the 70\textsubscript{CT} loop (93 to 109). TM5 residues (404 to 414) are strongly correlated because they lie in close proximity to the 30\textsubscript{CT} loop. In addition, Asp416 and Asp417 from TM5, make a significant contact with Thr105/Arg106 in thrombin. All of these TM residues were identified previously as critical by alanine scanning mutagenesis (72). The network of correlated motion continues from ABE1 into the thrombin active site loops and remains especially strong in the 148\textsubscript{CT} loop (residues 180-190) and includes the 90\textsubscript{CT} loop (residues 122-132) and the 170\textsubscript{CT} loop (residues 210 to 220).

Figure 3-6. Extended cross-correlated dynamic map for thrombin:TM456 obtained from a representative free energy weighted (level 2) AMD trajectory. The TM domains are defined as TM4 (residues 345 to 389), TM5 (residues 390 to 426) and TM6 (residues 427 to 462). Magenta boxes are drawn to highlight the most significant correlations between TM domains and TM domains with thrombin. Significant correlations within thrombin are indicated by black boxes as in Fig. 3-4.
Discussion

The work presented in this paper provides a unique and detailed insight into the functional dynamics of the thrombin:TM system, which not only rationalizes the available experimental data, but also explicitly identifies and defines the important role played by the TM4 domain in the activation of protein C. The results show two distinct, but inter-related allosteric pathways that mediate the anticoagulant activity of the thrombin:TM complexes. One pathway connects the TM5 domain with ABE1, the 60_{CT} insertion and the active site. This allosteric network, which exists in both thrombin:TM56 and thrombin:TM456 systems was previously identified using H/D exchange (9). Community network models further indicate that TM4 strengthens this allosteric pathway (Fig. 3-2) leading to a significant enhancement in the cross-correlated motion between the 30_{CT} loop and the 60_{CT} insertion in the thrombin:TM456 system compared to the isolated thrombin and thrombin:TM56 systems (Fig. 3-4). Indeed, binding of TM56 to thrombin stabilizes some thrombin active site loops on slow timescales, as shown by the backbone NH$_N$ order parameters calculated from AMD simulations (Fig 3-3b).

The addition of the 4th EGF-like domain (TM456) activates slow timescale dynamics in the active site loops of thrombin via another allosteric pathway. Global reorientation and intra-domain configurational dynamics of TM4 are strongly coupled to both the global and intra-domain dynamics of TM5. The importance of several of the most highly cross-correlated TM residues in these regions was previously characterized in an NMR study which found that backbone dynamics of Tyr358 and
Gln359 are correlated with anticoagulant activity and that backbone dynamics of Tyr413 and Ile414 are inversely correlated with thrombin binding (78). This tight coupling between TM4 and TM5 was previously shown experimentally to depend on M388 in the TM4-TM5 linker (79; 80). The specific orientation and interactions of TM5 induced via coupling to TM4 mediate specific interactions with ABE1. The dynamics of ABE1 are in turn highly correlated with the dynamics of the thrombin active site loops only in the TM4 containing construct. Through this complex pattern of extended correlated motions, slow timescale configurational dynamics of TM4 are directly linked to the thrombin active site, providing evidence for entropic allostery between TM4 and thrombin (Figs. 3-5 and 3-6). The community network analysis also reveals that when TM456 is bound, the information transfer between the loop regions of thrombin is significantly enhanced, resulting in the observation of fewer, yet larger communities, and increased potential for communication between communities. This enhanced communication is particularly strong for the path connecting the active site loops to TM4 via ABE1 and TM5 and interestingly includes a direct allosteric interaction between the 148$_{CT}$ loop and TM5. The highly correlated motions between the active site loops on slower timescales observed in thrombin:TM456 overlap directly with the different communities identified in the community network model. In particular, the 90$_{CT}$ loop coalesces with the active site into a single community consistent with H/D exchange experiments that identified subtle changes in the 90$_{CT}$ loop only in the presence of TM456 (15). Our observation of different slow timescale dynamic behavior of the active site loops of thrombin in thrombin:TM456 vs.
thrombin:TM56 provides a mechanism for these heretofore inexplicable experimental results.

The slow conformational loop dynamics, mainly of the active site loops of thrombin, which only occur in thrombin:TM456 probably facilitate the association of small substrates and inhibitors (that may enter the active site without interacting directly with TM4), thereby increasing the association rate by several orders of magnitude. For the specific case of the significantly enhanced $k_a$ of protein C, two complementary mechanisms exist: First, following the work of Fuentes-Prior et al., the presence of TM4 forms an extended binding surface for protein C, providing optimal alignment for insertion into the active site and subsequent cleavage (7). This ‘docking and optimal alignment’ mechanism probably works together with the allosteric mechanism identified in this study, in which the altered dynamics of the active site loops caused by TM binding increases the protein C $k_a$. The binding of protein C to TM4 may also affect dynamics within TM4 to enhance the allosteric network and further promoting Protein C cleavage.

**Methods**

**Conventional molecular dynamics (CMD)**

Atomic coordinates for thrombin were obtained from the protein data bank 1.9Å x-ray crystal structure [PDB ID: 1PPB] (12) and the initial coordinates for thrombin:TM56 and thrombin:TM456 were taken from the 2.3Å x-ray crystal structure [PDB ID: 1DX5, chains A,M and I] (7). The active site inhibitor was removed from all structures. For thrombin:TM56 and thrombin:TM456 residues
Arg456 and His475, which had been mutated to facilitate crystallization, were restored to wild-type. Each system was placed at the center of a periodically repeating box and the simulation cell size was defined such that the distance between the edge of the simulation box and the surface of the solute was at least 12Å. All simulations were performed in explicit solvent and an appropriate number of Na\(^+\) or Cl\(^-\) counter-ions were introduced to obtain cell neutrality. A set of six standard classical MD (CMD) simulations was performed for each system. For each of these simulations, a different random seed generator for the Maxwellian distribution of atomic velocities was employed and, after standard energy minimization and equilibration procedures, a 20 ns production run CMD simulation was performed under periodic boundary conditions with a time-step of 2 fs. Bonds involving protons were constrained using the SHAKE algorithm. Electrostatic interactions were treated using the Particle Mesh Ewald (PME) method (81) with a direct space sum limit of 10Å. The ff99SB force field (82) was used for the solute residues and the TIP3P water force field (83) was employed for the solvent molecules. These initial six 20 ns CMD simulations acted as a control set and were used as the starting point for the AMD simulations. These simulations also provided the average (unbiased) dihedral angle energy, \( <V_0(\text{dih})> \) and total energy \( <V_0(\text{tot})> \) values used to define the acceleration parameters in the AMD simulations described below.

**Accelerated molecular dynamics (AMD)**

The details of the accelerated molecular dynamics (AMD) protocol have been discussed previously (1; 37) and only a brief summary is provided here. In AMD, a
reference or 'boost energy', $E_b$, is defined which lies above the minimum of the potential energy surface. At each step in simulation, if the instantaneous potential energy, $V(r)$, lies below the boost energy a continuous, non-negative bias potential, $\Delta V(r)$, is added to the actual potential. If the potential energy is greater than the boost energy, it remains unaltered. The application of the bias potential results in a raising and flattening of the potential energy surface (PES), decreasing the magnitude of the energy barriers and thereby accelerating the exchange between low energy conformational states, while still maintaining the essential details of the potential energy landscape. Explicitly, the modified potential, $V^*(r)$, on which the system evolves during an AMD simulation is given by (1):

$$V^*(r) = V(r), \quad V(r) \geq E_b$$

$$V^*(r) = V(r) + \Delta V(r), \quad V(r) < E_b$$

where the bias potential, $\Delta V(r)$, is defined as:

$$\Delta V(r) = \frac{(E_b - V(r))^2}{E_b - V(r) + \alpha}$$

The extent of acceleration is determined by the choice of the boost energy, $E_b$, and the acceleration parameter, $\alpha$. Conformational space sampling can be enhanced by either increasing the boost energy or decreasing $\alpha$. In the present work, we have implemented a “dual boost” AMD approach (60), in which two acceleration potentials are applied simultaneously to the system: The first acceleration potential is applied to the torsional terms only, and a second, weaker acceleration is applied across the entire potential. For each of the three systems, thrombin, thrombin:TM56 and
thrombin:TM456, dual boost AMD simulations were performed at two (torsional) acceleration levels. For the most aggressive AMD simulations (level 2), the specific acceleration parameters were defined as $E_b(dih)-<V_0(dih)> = [4 \text{ kcal/mol} \times \text{NSR}]$, and the acceleration parameter, $\alpha(dih)$, was set to one fifth of this value. The specific choice of these AMD parameters was based on the recent NMR/AMD study of thrombin:PPACK (see SI for more details) (74). For the second, moderate AMD simulations (level 1), the $\alpha(dih)$ parameter was kept the same and $E_b(dih)-<V_0(dih)>$ was reduced to $[2 \text{ kcal/mol} \times \text{NSR}]$. In all AMD simulations, the total background acceleration parameters were fixed at $E_b(tot)-<V_0(tot)> = \alpha(tot) = [0.16 \text{ kcal/mol} \times \text{NASC}]$. For each of the three systems, six AMD simulations were performed at both accelerations levels for $10,000,000$ steps (the equivalent of $20$ ns standard MD). The physical conditions and computational parameters employed in the AMD simulations were identical to those defined above for the CMD simulations and all MD simulations were performed using an in-house modified version of the AMBER10 simulations suite (61). For each AMD trajectory, a corrected canonical ensemble was obtained by performing a Boltzmann free energy reweighting protocol using the bias potential block averaging method to remove statistical noise errors (see SI). In this way, six representative free energy weighted molecular ensembles were generated at both acceleration levels, along with the six unbiased $20$ ns CMD simulations for all three systems.
**Trajectory analysis**

Allosteric networks were characterized using a community network analysis approach previously applied to investigate allostery in tRNA:protein complexes and other protein systems (84; 75; 85). This approach constructs a dynamic contact map consisting of a network graph in which each residue is treated as a ‘node’, connected by edges to other nodes when two residues are deemed to be “in contact”. The dynamic contact map is subsequently decomposed into communities (clusters of residues) of highly intra-connected, but loosely inter-connected nodes using the Girvan-Newman algorithm (86). Central to this method is calculation of edge “betweenness”, the number of shortest paths that cross an edge. The edge betweenness is calculated for all edges and the edge with the greatest betweenness is removed. This process is repeated and a modularity score is tracked to identify the division that results in the optimal community structure. Network graph calculations were performed using the python module NetworkX (86).

Residue-by-residue cross-correlations were calculated using the generalized cross-correlation approach applied to all backbone Cα atomic coordinates based on the mutual information method developed by the Grubmüller group (76) using the g_correlation module in GROMACS 3.3.3 (87).

The internal dynamics were monitored by calculating backbone NH bond order parameters \( S^2 \) from the different CMD and AMD simulations, which provide a quantitative measure of the extent of reorientational motion of the given bond vector (38). In all cases, molecular ensembles generated from the standard CMD simulations
and the free energy weighted AMD trajectories were superposed onto the heavy backbone atoms of all residues for the appropriate average structure. Order parameters were calculated as (88):

\[
S^2 = \frac{1}{2} \left[ 3 \sum_{i=1}^{3} \sum_{j=1}^{3} \left( \mu_i \mu_j \right)^2 - 1 \right]
\]

where \( \mu_i \) are the Cartesian coordinates of the normalized inter-nuclear vector of interest. The resulting order parameters were then averaged over all MD/AMD trajectories.

**Acknowledgments**

PG gratefully acknowledges a CMG training grant [NIH- T32 GM007240] and this work was co-supported by the Center for Theoretical Biophysics [NSF PHY-0822283]. BF and EAK acknowledge support from NIH-HL070999. Additional support was provided by the NIH, NSF, CTBP, HHMI, NBCR and the San Diego Supercomputer Center.

Chapter 3 is a minimally modified reprint of the material as it appears in Paul M. Gasper, Brian Fuglestad, Elizabeth A. Komives, Phineus R. L. Markwick and J. Andrew McCammon, Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities, Proceedings of the National Academy of Science 2012. The dissertation author was the primary investigator and author of this paper.
Supplemental Information

Chymotrypsin and sequential residue numbering schemes

To accommodate readers who use one of several different numbering schemes for thrombin, we report here residue labeling in the both the chymotrypsin and sequential numbering schemes. Thrombin possesses both a light chain (residues 1H_{CT}-15_{CT} in the CT labeling scheme, residues 1-36 in sequential numbering) and a heavy chain that comprises the active site and active site loops (see Table 3-1.) In the sequential residue numbering scheme, thrombin is defined as residues 1-295, and the thrombomodulin domains are defined as TM4 (residues 296-340), TM5 (residues 341-377) and TM6 (residues 378-413). Notably the anion binding exosite site 1 (ABE1) consists of two regions, the 30_{CT} loop (33_{CT}-39_{CT}), comprising residues 54-61 in the sequential residue numbering scheme and the “ABE1 large loop”, residues 66_{CT}-81_{CT} in the chymotrypsin numbering scheme, residues 93-109 in the sequential residue numbering scheme.

Table 3-1. Chymotrypsin and sequential residue numbering schemes for key regions of thrombin.

<table>
<thead>
<tr>
<th>Name</th>
<th>Chymotrypsin Numbering</th>
<th>Sequential Numbering</th>
</tr>
</thead>
<tbody>
<tr>
<td>60_{CT}-insertion</td>
<td>60A_{CT}-60I_{CT}</td>
<td>83-91</td>
</tr>
<tr>
<td>ABE1</td>
<td>33_{CT}-39_{CT} and 66_{CT}-81_{CT}</td>
<td>54-61 and 93-109</td>
</tr>
<tr>
<td>90_{CT} loop</td>
<td>90_{CT}-99_{CT}</td>
<td>122-132</td>
</tr>
<tr>
<td>148_{CT} (gamma) loop</td>
<td>142_{CT}-151_{CT}</td>
<td>178-192</td>
</tr>
<tr>
<td>170_{CT} loop</td>
<td>172_{CT}-180_{CT}</td>
<td>213-221</td>
</tr>
<tr>
<td>186_{CT} loop</td>
<td>186_{CT}-188_{CT}</td>
<td>228-234</td>
</tr>
<tr>
<td>220_{CT} loop</td>
<td>217_{CT}-226_{CT}</td>
<td>265-275</td>
</tr>
</tbody>
</table>
Accelerated molecular dynamics (AMD)

Accelerated molecular dynamics (AMD) (1; 37) is an extended biased potential molecular dynamics approach that allows for the efficient study of biomolecular systems up to timescales several orders of magnitude greater than those accessible using conventional MD methods, while still maintaining a fully atomistic representation of the system. In the AMD algorithm a reference or 'boost energy', $E_b$, is defined which lies above the minimum of the potential energy surface (PES). At each step in simulation, if the instantaneous potential energy, $V(r)$, lies below the boost energy a continuous, non-negative bias potential, $\Delta V(r)$, is added to the actual potential. If the potential energy is greater than the boost energy, it remains unaltered. The application of the bias potential results in a raising and flattening of the PES, decreasing the magnitude of the energy barriers and thereby accelerating the exchange between low energy conformational states, while still maintaining the essential details of the potential energy landscape. Explicitly, the modified potential, $V^*(r)$, on which the system evolves during an AMD simulation is given by:

$$V^*(r) = V(r), \quad V(r) \geq E_b$$

$$V^*(r) = V(r) + \Delta V(r), \quad V(r) < E_b$$

and the bias potential, $\Delta V(r)$, is defined as:

$$\Delta V(r) = \frac{(E_b - V(r))^2}{E_b - V(r) + \alpha}$$

The extent of acceleration is determined by the choice of the boost energy, $E_b$, and the acceleration parameter, $\alpha$. Broadly speaking, the magnitude of $[E_b - V(r)]$
defines the strength of the bias, while the acceleration parameter, \( \alpha \), controls the curvature (i.e. extent of flattening) of the modified PES. As such, conformational space sampling can be enhanced by either increasing the boost energy while holding \( \alpha \) fixed, or decreasing \( \alpha \) for a fixed value of \( E_b \). Care must be taken not to ‘over-accelerate’ the system: For example, setting \( E_b \) too large or \( \alpha \) too small leads to the generation of an iso-energetic modified PES, resulting in a random walk through configurational space causing the system to spend a large proportion of time sampling energetically unfavorable regions. During the course of the AMD simulation, if the potential energy is modified, the forces on the atoms are recalculated for the modified potential and the use of the bias potential defined above ensures that the derivative of the modified potential will not become discontinuous at points where \( V(r) = E_b \).

One of the favorable characteristics of the AMD approach is that it yields a canonical average of an ensemble, so that thermodynamic and other equilibrium properties can be accurately determined. The corrected canonical ensemble average of the system is obtained by reweighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, \( \exp[\beta \Delta V(r^{(i)})] \) at that particular point. AMD has already been employed with great success to study the dynamics and conformational behavior of a variety of biomolecular systems including polypeptides, folded and natively unstructured proteins. (37).

Central to the success of an AMD simulation is the choice of the acceleration parameters, \( E_b \) and \( \alpha \). In principal, for every system there exists a suitable or ‘optimal’
set of acceleration parameters that efficiently enhance the conformational space sampling without generating instabilities in the trajectory and avoiding a random walk. While the choice of optimal acceleration parameters is certainly system specific and depends most notably on the size of the system, comparative analysis of a set of successful AMD studies revealed that for torsional acceleration, the optimal value of \([E_b(dih)-V(dih)]\) is equal to 3-5 kcal/mol times the number of solute residues (NSR) in the system and the acceleration parameter, \(\alpha\), should be approximately one fifth of this value. (37) In line with this general observation, in a very recent NMR/”dual boost” AMD study of thrombin:PPACK, (74) it was found that the optimal torsional acceleration parameters for the recapitulation of experimental NMR-based N-H\(^{\text{N}}\) Residual Dipolar Couplings (RDCs) were \([E_b(dih)-V(dih)]=4 \text{ kcal/mol } \times \text{NSR}\) and \(\alpha=[0.8 \text{ kcal/mol } \times \text{NSR}]\). In this study, the acceleration parameters for the background, total acceleration \([E_b(tot)-V(tot)]\) and \(\alpha(tot)\) were both set to 0.16 kcal/mol times the number of atoms in the simulation cell (NASC), which were obtained from a detailed study of the effect of acceleration on bulk water (89). RDCs are averaged over all orientations of the magnetic dipolar interaction vector sampled up to a timescale defined by the inverse of the alignment induced coupling, thereby reporting on temporal motions up to the ms range. However, an analysis of the R1/R2/hetNOE data revealed no significant exchange relaxation in the thrombin:PPACK system. As exchange relaxation specifically reports on motions in a time window spanning ~100-µs to several ms, we concluded that the dynamic motions that were being accessed in the AMD simulations were occurring on timescales up to several tens of µs. In the
present study on thrombin, thrombin:TM56 and thrombin:TM456, we have specifically chosen the same acceleration parameters \{E_b(dih)-V(dih)=4\, \text{kcal/mol}\, \text{NSR}, \alpha(dih)=0.8\, \text{NSR}\} and \{[E_b(tot)-V(tot)]=\alpha(tot)=0.16\, \text{NASC}\} for the most aggressive acceleration level (level 2) and, by analogy to our previous study on thrombin:PPACK, conclude that the associated timescale of the configurational space sampling in these AMD trajectories is approximately several tens of \(\mu s\). For each AMD simulation, a corrected canonical ensemble was obtained by performing the Boltzmann free energy reweighting protocol described above using the bias potential block averaging method (37) to remove statistical noise errors.

**Community network analysis models**

Potential allosteric pathways may be examined through community network analysis, which identifies communities of highly intra-connected residues based on residue-by-residue correlation and proximity. A comparison of the distribution and connectivity of communities between similar biomolecular systems highlights differences in potential allosteric pathways. This approach has previously been applied to investigate allostery in tRNA:protein complexes by Luthey-Schulten et al. (75) and employs the community network analysis algorithm developed by Girvan and Newman (90).

A dynamic contact map was prepared for thrombin in each of the three systems (isolated thrombin, thrombin:TM56 and thrombin:TM456) from the CMD simulations. The dynamic contact map consists of a network graph in which each residue was treated as a node. Edges were added to the network by connecting pairs of
nodes whose residues were “in contact”, which was defined as having any heavy atom within 5.0Å for greater than 75% of the simulation. Each edge was assigned a weight based on the cross-correlation between the residues it joined ($W_{ij} = -\ln|C_{ij}|$). As such, weaker cross-correlations result in longer edges, representing less potential for information transfer. The length of the path between two nodes ($D_{ij}$) is given as the summation of the edge weights for a series of edges connecting ‘i’ to ‘j’. The path with the shortest length, $D^0_{ij}$, represents the most efficient route for information transfer and was determined for each node pair using the Floyd-Warshall algorithm.

The Girvan-Newman algorithm (90) was used to coarse-grain the dynamic contact maps into communities of highly intra-connected, but loosely inter-connected nodes. Central to this algorithm is the calculation of edge “betweenness”, or the number of shortest paths that cross a given edge. The edge betweenness is calculated for all edges and the edge with the greatest betweenness is removed. This process is repeated and a modularity score is tracked to identify the division that results in the optimal community structure. Calculations involving network graphs were performed using the python module NetworkX (86).
Chapter 4:

Correlated Motions and Residual Frustration in Thrombin


**Abstract**

Thrombin is the central protease in the cascade of blood coagulation proteases. The structure of thrombin consists of a double β-barrel core surrounded by connecting loops and helices. Compared to chymotrypsin, thrombin has more extended loops that are thought to have arisen from insertions in the serine protease that evolved to impart greater specificity. Previous experiments showed thermodynamic coupling between ligand binding at the active site and distal exosites. We present a combined approach of molecular dynamics (MD), accelerated molecular dynamics (AMD), and analysis of the residual local frustration of apo-thrombin and active site bound (PPACK-thrombin). Community analysis of the MD ensembles identified changes upon active site occupation in groups of residues linked through correlated motions and physical contacts. AMD simulations, calibrated on measured residual dipolar couplings, reveal
that upon active site ligation correlated loop motions are quenched, but new ones connecting the active site with distal sites where allosteric regulators bind, emerge. Residual local frustration analysis reveals a striking correlation between frustrated contacts and regions undergoing slow timescale dynamics. The results elucidate a motional network that probably evolved through retention of frustrated contacts to provide facile conversion between ensembles of states.

**Introduction**

Thrombin is the central protease in the cascade of blood coagulation proteases. Structurally, thrombin consists of a double β-barrel core surrounded by connecting loops and helices. Genetic analysis of the clotting factor genes demonstrates that the clotting proteases of the chymotrypsinogen superfamily have evolved as a result of several gene duplications, exon shuffling, and intron sliding events. Prothrombin has a unique exon organization and is thought to be the ancestral gene of the clotting factor family (91). The extended active site loops in thrombin are thought to have arisen from insertions in the serine protease that evolved to impart greater specificity (91; 18). Thrombin is produced in a low activity, zymogen form that requires proteolytic cleavage to attain full activity. This cleavage event results in small overall changes to the molecular architecture but results in a large change in dynamics wherein one β-barrel becomes more dynamic and the other becomes less dynamic (92). The result is a more perfectly formed active site for rapid proteolytic cleavage activity. Despite the highly specific nature of thrombin activity, in association with allosteric modulators, the substrate specificity is tuned to activate either pro-coagulant or anti-coagulant
substrates (93). In addition, allostery is key to thrombin regulation (6; 21) and misregulation can lead to bleeding disorders or thrombosis. Although traditionally allostery was defined as occurring among subunits in a multisubunit system such as hemoglobin, (94) the phenomenon of altered activity resulting from binding of a regulatory molecule on the opposite side of a monomeric enzyme is now also recognized as a form of allostery (95; 96).

Several experimental and computational approaches have hinted that the solution structure of thrombin is a broad and malleable dynamic ensemble (74; 97). H/D exchange mass spectrometry showed D-Phe-Pro-Arg chloromethylketone (PPACK) occupation of the active site not only protected the active site loops, but also propagated to decreased exchange in several regions of the protein distant from the active site (98). Thrombin has two binding sites distal to the active site, exosite 1 is where thrombomodulin binds, and exosite 2 is where heparin binds. Isothermal titration calorimetry (ITC) experiments showed alteration of thermodynamic parameters of ligands binding to thrombin exosites when the active site was occupied (99). Binding of active site ligands altered the balance of enthalpic and entropic contributions to binding of exosite 1 ligands and vice versa (10). Such thermodynamic compensation phenomena are more likely if the allosteric mechanism is entropic rather than enthalpic suggesting that differences in the dynamic properties of the system affect the ligand binding mechanism (12; 13). Indeed, x-ray crystallography shows no significant changes in the thrombin structure upon ligand binding providing further evidence that it may exist as a malleable dynamic ensemble.
NMR studies and MD simulations remain the most direct approaches to investigating protein dynamics (22). A recent NMR/MD study on PPACK-thrombin revealed a large degree of dynamic motions, particularly in the active site loops, spanning timescales from ps to ms (74). A computational exploration revealed a strong dynamic component in the allostERIC regulation of thrombin by TM (97). In the work presented here, we combine conventional MD, NMR calibrated accelerated MD (AMD), and analysis of the residual local frustration to further explore the dynamics of thrombin with particular interest in changes that occur upon active site ligation. The combined approach allows us to analyze a broad range of motional timescales.

**Methods**

**Molecular dynamics and community analysis**

Atomic coordinates for thrombin were obtained from the protein data bank 1.9Å x-ray crystal structure [PDB ID: 1PPB] (70). The active site inhibitor was removed for the apo-thrombin calculations. Both systems were placed at the center of a periodically repeating box and the simulation cell size was defined such that the distance between the edge of the simulation box and the surface of the solute was at least 12Å. All simulations were performed in explicit solvent and three Cl⁻ counterions were introduced to obtain cell neutrality. Six 20 ns conventional MD simulations were performed using a different random seed generator for the Maxwellian distribution of atomic velocities following standard energy minimization and equilibration procedures. Periodic boundary conditions and a time-step of 2 fs were employed. Bonds involving protons were constrained using the SHAKE algorithm.
Electrostatic interactions were treated using the Particle Mesh Ewald (PME) method (81) with a direct space sum limit of 10Å. The ff99SB force field (82) was used for the solute residues and the TIP3P water force field (83) was employed for the solvent molecules. In the case of PPACK-thrombin simulations, an in-house gaff force field was generated for the PPACK inhibitor. The conventional MD simulations were analyzed in the community analysis and also constituted the starting point for the AMD simulations. These simulations also provided the average (unbiased) dihedral angle energy, \( <V_0(\text{dih})> \) and total energy \( <V_0(\text{tot})> \) values used to define the acceleration parameters in the AMD simulations described below.

Allosteric networks were characterized using a community network analysis approach previously applied to investigate allostery in tRNA:protein complexes and other protein systems (84; 75; 85). This approach constructs a dynamic contact map consisting of a network graph in which each residue is treated as a ‘node’, connected by edges to other nodes when two residues are deemed to be “in contact” throughout the majority of the simulation. The dynamic contact map is subsequently decomposed into communities (clusters of residues) of highly intra-connected, but loosely inter-connected nodes using the Girvan-Newman algorithm (90). Central to this method is calculation of edge “betweenness”, the number of shortest paths that cross an edge. The edge betweenness is calculated for all edges and the edge with the greatest betweenness is removed. This process is repeated and a modularity score is tracked to identify the division that results in the optimal community structure. Network graph calculations were performed using the python module NetworkX (86).
**AMD simulations and analysis**

AMD simulations were performed as described previously using an in-house modified version of the AMBER 10 code (74; 97). A “dual boost” AMD approach (60), in which two acceleration potentials are applied simultaneously to the system: The first acceleration potential is applied to the torsional terms only, and a second, weaker acceleration is applied across the entire potential. This dual boost AMD protocol represents a unified approach facilitating the efficient sampling of both the torsional degrees of freedom and slow diffusive motions in the solute. In total, six independent dual boost AMD simulations were performed for 10,000,000 steps (the equivalent of 20 ns MD) for each system. The physical conditions, force fields and all other simulation parameters employed were identical to those described for the conventional MD simulations. The specific acceleration parameters used in this study were $E_b(dih) - \langle V_0(dih) \rangle = [4 \text{ kcal/mol} \times \text{NSR}]$ and $\alpha(dih) = [0.8 \text{ kcal/mol} \times \text{NSR}]$ for the torsional acceleration and $E_b(tot) - \langle V_0(tot) \rangle = \alpha(tot) = [0.16 \text{ kcal/mol} \times \text{NASC}]$ for the background, total acceleration. These acceleration parameters had been previously identified as the optimal acceleration parameters for the reproduction of experimental RDCs in PPACK-thrombin, accessing configurational dynamics on timescales up to 10-100 µs. (74). For each AMD simulation, a corrected canonical ensemble was obtained by reweighting each point in the configuration space on the modified potential by the strength of the Boltzmann factor of the bias energy, $\exp[\beta \Delta V(r_{ii})]$ at that particular point and the bias potential block averaging method was employed to remove statistical noise errors (37).
The internal dynamics present in the different AMD simulations of apo-thrombin and PPACK-thrombin were assessed by calculating order parameters, $S^2$, from the free energy weighted AMD ensembles: Members of each ensemble were superimposed onto the backbone atoms (N, C\textsuperscript{\alpha}, C') of all heavy chain residues for the appropriate average structure and order parameters, $S^2$ were calculated as:

$$S^2 = \frac{1}{2} \left[ \frac{3}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \langle \mu_i, \mu_j \rangle - 1 \right]$$

where $\mu_i$ are the Cartesian coordinates of the normalized inter-nuclear vector of interest. Others have shown that $S^2$ values calculated from standard MD simulations in this way were in excellent agreement with experimental $S^2$ values calculated using the Lipari-Szabo autocorrelation function approach (39). The order parameters presented here are averaged over all six AMD trajectories for each system.

Residue-by-residue cross-correlations for the free energy weighted AMD ensembles were calculated using the generalized cross-correlation approach applied to all backbone C\textsuperscript{\alpha} atomic coordinates based on the mutual information method developed by the Grubmüller group (76) using the g\_correlation module in GROMACS 3.3.3 (87).

**Residual local frustration analysis**

An algorithm for determining residual local frustration, i.e. whether a contact between amino acid residues is energetically optimized or not in the folded state was developed by the Wolynes and Komives groups some time ago (40). This algorithm assesses residue-residue interactions by systematically perturbing the identity of
individual residues and evaluating the resulting total energy change. For the work presented here, we used the “configurational frustration” index, in which the decoy set involves randomizing not just the identities but also the distance and densities of the interacting amino acids i, j. This scheme effectively evaluates the native pair with respect to a set of structural decoys that might be encountered in the folding process. After constructing a histogram of the energy of the decoys and comparing the distribution to the native energy, cut-offs are implemented to identify minimally frustrated or highly frustrated residues. Energetically favorable contacts between residues are minimally frustrated whereas highly frustrated contacts are energetically unfavorable in the native state. Depictions of the contacts on structural models typically show minimally frustrated contacts in green and highly frustrated contacts in red. The average of the frustration scores over all the contacts made by a particular residue are also plotted in a per-residue format. A webserver is now available for performing these computations (100). For the work presented here, the minimally and highly frustrated contacts are depicted on the lowest energy structure from the Boltzmann-rewighted ensemble of structures from the AMD simulations. To compute the average per-residue frustration, we averaged the frustration scores of all contacts made by each residue in the three lowest energy structures from the Boltzmann-rewighted ensemble of structures from the AMD simulations. It is interesting to note that the residual local frustration varied between members of the ensemble, and the error bars on the residual local frustration plots represent one standard deviation.
Results

Community network analysis

We performed a set of six independent 20 ns conventional MD simulations for both apo-thrombin and PPACK-thrombin. During the equilibration procedure, a rather large conformational transition in the active sites loops was observed for apo-thrombin that involved a reorientation of both the γ-loop (178-195) and the Na⁺-binding loop (264-271) forming a more ‘open’ active site pocket.

A community network analysis approach (75; 85) was applied to identify groups of residues undergoing correlated motions in PPACK-thrombin and apo-thrombin. Representative community network analyses obtained from conventional MD simulations are shown in Fig. 4-1. The flow of information in the physical network of the protein was measured by the edge betweenness, defined as the number of shortest paths that pass through the edge in the network, and is a direct measure of the strength of intercommunity communication within the network (black lines in Figs. 4-1a and 4-1b). PPACK ligation causes consolidation of the community structure including the two active site communities most proximal to the PPACK binding site (green and brown, Fig. 4-1). The community that includes part of the Na⁺ binding loop in apo-thrombin (Fig. 4-1c, brown) consolidates with the active site serine 195\textsubscript{CT} (241), the 70s loop (98-113), and residues 191-194\textsubscript{CT} of the γ-loop (237-240) and the N-terminal residues 17-19\textsubscript{CT} (38-40) in the PPACK-liganded form (brown, Fig. 4-1d). PPACK ligation also acts to consolidate the N-terminal β-barrel, which is formed by two separate communities in apo-thrombin and form a large
community that also contains the 30s loop (55-62) and part of the 60s loop (82-94) (orange, Fig. 4-1d). In summary, the community analysis revealed consolidation of the Na\(^+\) binding site, the base of the \(\gamma\)-loop, and the N-terminus of the heavy chain into one community and most of the active site loops into a second community upon PPACK binding. These two communities, which are strongly connected in PPACK-thrombin, unite the residues required for proteolytic catalysis. The light chain community also becomes more strongly connected to the active site and the community containing the 70s loop. Based on the substantial consolidation observed in the community analysis upon PPACK-binding, we set out to examine whether there were concomitant changes in dynamics.

Figure 4-1. a) and c) community analysis of apo-thrombin and b) and d) PPACK-thrombin. The two dimensional view of communities in panels a) and b) depicts relative size of communities (based on number of residues) as colored circles of varying sizes with the thickness of the connecting lines representing the relative interconnectivity among communities. Panels c) and d) are structural representations of communities.
AMD simulations

Residual dipolar couplings (RDCs), which report on an ensemble average over all orientations of the magnetic dipole interaction vector up to the chemical shift coalescence limit, provide useful experimental data for determining the ensemble of structures that best represents the dynamic properties of a protein (31; 29). When the experimentally derived RDCs are compared to RDCs that are back-calculated from ensembles of structures generated from AMD simulations, the acceleration level that provides the most realistic representative structural ensemble can be identified. We previously demonstrated that the RDCs measured on PPACK-thrombin did not agree well with the available crystal structures ($R^2=0.72$). Agreement was only marginally improved ($R^2=0.80$) when the RDCs were back-calculated from an ensemble of structures obtained by conventional MD (74). However, remarkable agreement was obtained between the experimental RDCs and the RDCs back-calculated from the ensemble of structures obtained from an AMD simulation at the optimal acceleration level ($R^2=0.92$) (74). The ensembles obtained from such AMD simulations of both PPACK-thrombin and apo-thrombin are shown in Fig. 4-2.

Figure 4-2. Ensemble of the 20 lowest energy structures from the RDC-calibrated AMD for a) apo-thrombin and b) PPACK-thrombin. The loops are colored as in to the schematic under the structures.
For PPACK-thrombin, order parameters \( (S^2) \) from conventional MD simulations agreed extremely well with those measured by NMR relaxation experiments that are limited to the ps-ns time regime by the molecular tumbling time (~17 ns) (74). However, the fact that ensembles obtained from AMD were required for good agreement with the RDC measurements suggested that motions on longer timescales are contributing to the solution structure. Therefore order parameters for the N-H bond vectors were calculated from the AMD ensembles \( (S^2_{AMD}) \). A comparison of \( S^2_{AMD} \) for apo-thrombin (97) to those obtained previously for PPACK-thrombin (74) is shown in Fig. 4-3a. Most of the active site loops in both forms are highly flexible, yet a marked decrease in flexibility is observed upon active site ligation with PPACK (Fig. 4-3b). As expected, PPACK ligation caused significant ordering \( (\Delta S^2 > 0.1) \) of residues that directly contact the PPACK Arg side chain. The loops that surround the active site also experience significant ordering including the 60s loop (82-94), the \( \gamma \)-loop, and the 180s loop (225-239). Some regions distal to the PPACK also showed significant ordering upon PPACK ligation including the light chain, the 30s loop, residue 221\text{CT} \ (269) \) of the Na\textsuperscript{+} binding loop, and residues in the C-terminal helix.
Correlated motion analysis

To identify residues undergoing correlated motions on longer timescales, AMD simulations were performed that were optimized based on previous work comparing experimental RDCs to those back-calculated from AMD simulations carried out at different acceleration levels (74). The analysis of apo-thrombin revealed correlated motions between the active site loops, exosite 1, and other distal sites. In particular, the entire γ-loop appears to undergo strongly correlated motions with the light chain residues 1H-1D_CT and 12-14C_CT (1-5 and 20-25), the catalytic triad; H57_CT, D102_CT, S195_CT (79, 135, 241), the 60s loop, the 70s loop, the 90s loop (127-133), the surface strand under the 70s loop (145-151), the 170s loop (204-219), and the 180s loop (Fig. 4-4, lower triangle). Whereas these correlated motions appeared to involve the entire γ-loop in apo-thrombin, only the tip of the γ-loop (residues 146-149E_CT (182-190)) appears to be undergoing the same set of motions in PPACK-thrombin (Fig. 4-4 upper triangle). In apo-thrombin, the 170s and to a lesser extent the 180s loop, which are strongly correlated to the γ-loop, are also correlated with the light
chain and catalytic residues. Upon PPACK ligation, most of these correlated motions are lost. In apo-thrombin, the 30s loop and the 60s loop are weakly correlated, but the 30s loop is not correlated to the 70s and 90s loops. Correlated motions between the 30s and 60s loops are stronger in PPACK-thrombin and these extend to the 70s and 90s loops (Fig. 4-4).

![Figure 4-4](image.png)

Figure 4-4. Analysis of correlated motions performed on the AMD trajectories of PPACK-thrombin (top triangle) and apo-thrombin (bottom triangle). The motions range from 0.0 (no correlation, white) to 1.0 (completely correlated, black). The schematic diagram indicating the location of surface loops is inserted above the correlated motions plot. The black boxes indicate correlations of the 170s and 180s loops with the light chain and surface loops that are stronger in apo-thrombin than in PPACK-thrombin. The blue boxes indicate correlations of the 30s loop with the 60s, 70s, and 90s loops that are stronger in PPACK-thrombin than in apo-thrombin.

**Residual Frustration correlates with longer timescale dynamics**

We applied a previously derived algorithm to discover the residual local frustration in representative structures of RMSD clusters of the Boltzmann reweighted AMD simulation results (40). According to the principle of minimal frustration, (101) contacts made in the folded native state should be minimally frustrated meaning that they are energetically favorable. We previously showed that while most contacts made
in the native state are, indeed, minimally frustrated, some 10 – 15% of contacts are energetically unfavorable (ie. highly frustrated) in the native state. These highly frustrated contacts map to functional sites, and are thought to have been preserved in evolution. Both apo- and PPACK-thrombin show regions of high frustration in many of the surface loops (Fig. 4-5).

Figure 4-5. Analysis of residual local frustration (102) in the lowest energy structure from the RDC-calibrated AMD ensemble of a) apo-thrombin and b) PPACK-thrombin. The contacts that are minimally frustrated are shown in green and the contacts that are highly frustrated are shown in red. Thin lines represent water-mediated contacts. The active site catalytic residues are shown in magenta.

To discover whether regions of high frustration also map to dynamic regions, we compared the average residual frustration across the representative structures from the three most populated RMSD clusters of the AMD simulations to order parameters. The order parameters derived from conventional MD simulations ($S^2_{ns}$) agree very well with order parameters derived from NMR relaxation experiments on thrombin, (74) and reveal the disorder resulting from motions in the ns time regime. The order parameters derived from the RDC-calibrated AMD simulations ($S^2_{AMD}$) reveal the disorder resulting from motions on longer timescales. The $S^2_{ns}$ did not correspond well to the regions of high residual frustration, however the correspondence with the $S^2_{AMD}$ is remarkable (Fig. 4-6). These results highlight that regions of high residual
frustration may have evolved to allow slow timescale motions to occur with relative energetic ease as previously suggested by Wolynes and colleagues (103).

![Graph](image)

Figure 4-6. Comparison of the order parameters reflecting ns timescale motions vs. longer timescale motions ($S^2_{\text{AMD}}$) with the average per residue fraction of highly frustrated contacts for the three lowest energy structures from the AMD simulation. 

a) The $S^2_{\text{ns}}$ (grey) and $S^2_{\text{AMD}}$ (red) for apo-thrombin are compared to the average fraction of highly frustrated contacts (cyan). 

b) The $S^2_{\text{ns}}$ (grey) and $S^2_{\text{AMD}}$ (black) for PPACK-thrombin are compared to the average fraction of highly frustrated contacts (blue). The schematic of important surface loops is provided above the graph.

**Discussion**

We used a combination of community network analysis, RDC-calibrated AMD simulations and analysis of residual frustration to explore the dynamic ensemble of thrombin in solution from the ns into the $\mu$s time regime. Comparative analysis of apo-thrombin and the active site liganded, PPACK-thrombin, systems identified differences in the dynamic fluctuations and changes in correlated motions upon active site ligation.

Analysis of thrombin complexed with the relatively small substrate analog PPACK, showed a substantial rearrangement of the community structure. PPACK
binding consolidated the two active site communities most proximal to the PPACK binding site, as well as the Na\(^+\) binding loop with the active site serine, the 70s loop, part of the \(\gamma\)-loop and part of the light chain. Thus, substrate binding is predicted to consolidate the catalytic residues. Interestingly, this consolidation resulted in more residual local frustration near the active site (Fig. 4-5).

The generalized cross-correlation analysis predicted that many of the thrombin surface loops are undergoing correlated motions (Fig. 4-4). The analysis of residual local frustration revealed that all of these loops are highly frustrated in the native structure of thrombin (Fig 4-5). The striking correlation between the long timescale dynamics and the residual local frustration suggests that evolution has selected for energetically unfavorable contacts within the surface loops in order to facilitate such longer timescale, presumably larger amplitude motions. Upon PPACK-binding, many of the loops retain dynamics and also remain highly frustrated.

The surface loops undergoing correlated motions are not necessarily nearby in the structure, hinting at how allosteric regulation by binding of ligands at sites distal to the active site would readily occur. Upon PPACK-binding the active site community consolidates with the 70s loop where fibrinogen binds to stimulate its own cleavage and where thrombomodulin binds to alter substrate specificity. Indeed, previous experiments have demonstrated a thermodynamic coupling between the 70s loop and the active site (10). Another loop that is distal to the active site but that differs between apo- and PPACK-thrombin, is the \(\gamma\)-loop. Residues leading up to the \(\gamma\)-loop are highly frustrated and highly dynamic in apo-thrombin but become much less dynamic in the
PPACK-bound form with only the tip of the \(\gamma\)-loop remaining dynamic. Although the strongly correlated motions in apo-thrombin between the \(\gamma\)-loop and the 170s loop almost completely disappear upon PPACK ligation, correlations are strengthened between the tip of the loop and nearly every other loop in thrombin including the 30s loop, the 60s loop, the 70s loop, the \(\beta\)-strand connecting the 60s and 70s loops, and the 90s loop. Thus, we can speculate that the reduced dynamics result from the binding of PPACK causing these regions to sample a more restricted conformational space and that this reduced sampling leads to stronger correlated motions extending between the active site and the surface loops involved in allosteric regulation. The order parameter analysis revealed, as expected, that active site occupation decreased the \(\mu\)s motions of the active site loops (Fig. 4-3) consistent with amide H/D exchange experiments that showed that binding of PPACK dampened exchange in all of these surface loops that are distal to the active site (98). The shuffling of motions in the \(\mu\)s-ms timescale could explain how active site occupation lowers the entropic penalty for binding an exosite 1 allosteric modulator (10). Thermodynamic measurements of binding show that the allosteric pathway between these two sites operate in both directions (10) and changes in order parameters in the thrombin surface loops upon PPACK binding to the active site recapitulate those seen upon TM56 binding to exosite 1 (97).

The serine protease architecture is that of a double \(\beta\)-barrel fold in which the \(\beta\)-strands are connected by surface loops. The active site is between the two \(\beta\)-barrels, each of which has several \(\beta\)-strands connected by the surface loops. Such a \(\beta\)-strand architecture provides a direct connection between distal surface loops allowing
transmission of allosteric information by subtle redistribution of the dynamic ensemble (104). Such local unfolding or cracking was predicted some time ago to be the fundamental mechanism of allostery, and residual local frustration is indeed found at such sites (105). It will be interesting to see if dynamic allostery is present in other proteins with mostly β-strand structures similar to thrombin.

Dedications

This work was dedicated to Peter Wolynes, on the occasion of his 60th birthday.

Acknowledgements

BF and EAK gratefully acknowledge support from NIH-HL070999. BF was supported by the Molecular Biophysics training grant (T32 GM08326), and a predoctoral fellowship from the American Heart Association, Western States Affiliate. PG acknowledges a CMG training grant [NIH-T32 GM007240] and this work was co-supported by the Center for Theoretical Biophysics [NSF PHY-0822283]. Additional support was provided by the NIH, NSF, CTBP, HHMI, NBCR and the San Diego Supercomputer Center.

Chapter 4 is a minimally modified reprint of the material as in appears in Brian Fuglestad, Paul M Gasper, J. Andrew McCammon, Phineus R. L. Markwick, and Elizabeth Ann Komives, Correlated Motions and Residual Frustration in Thrombin, The Journal of Physical Chemistry B 2013. The dissertation author was the primary investigator and author of the computational aspects of this paper.
Chapter 5:

Remarks on Thrombin Allostery as revealed by Molecular Dynamics

Conclusions

In this work, MD simulations revealed that thrombin consists of a rigid core, corresponding to evolutionarily conserved regions, surrounded by highly dynamic, less conserved surface loops. MD ensemble derived order parameters, reporting on internal N-H bond vector reorientation dynamics, were found to be in good agreement with those determined by experimental NMR. Together these data show that thrombin activation does not result in an ordering of the surface loops as previously thought. AMD ensemble derived order parameters revealed that these loops undergo a significantly greater range of motion over longer, μs to ms timescales. The boost parameters determining the extent of AMD acceleration were successfully optimized to sample conformations on these timescales by comparison with experimental NMR RDCs. The resulting back-calculated AMD RDCs show considerable improvement with experiment over RDCs back-calculated from conventional MD simulations and
from the crystal structure. The heterogeneity of loop dynamics observed over multiple
timescales is not indicative of a conventional hinging motion, but rather resembles an
anemone with appendages moving on multiple timescales.

The comparison of MD and AMD simulations among three thrombin systems
(isolated thrombin, thrombin:TM56 and thrombin:TM456) identified dynamical
changes specific to thrombin:TM456 which propagated between the TM binding
exosite, ABE1, and the distal active site. Community network analysis revealed two
allosteric pathways connecting TM to the active site region. While the first pathway,
via ABE1 and the 60ct insertion, is present for both thrombin:TM56 and
thrombin:TM456, it is significantly strengthened by the presence of the TM4 domain.
The second pathway, which connects the TM5 domain directly with and the active
site, is unique to thrombin:TM456. In addition, generalized cross-correlations show
significant concerted motions along this second pathway only in the thrombin:TM456
system, and AMD ensemble derived order parameters show increased active site
dynamics only in this system. Taken together, these results provide a strong indication
of a dynamic, entropic allostery between TM4 and the thrombin active site.

The comparison of MD and AMD simulations between apo-thrombin and
inhibitor bound PPACK-thrombin systems also identified allostery between ABE1 and
the active site. Generalized cross-correlations revealed differences in active site and
exosite dynamics upon active site ligation. Additionally, PPACK-thrombin showed a
substantial rearrangement of the community structure. Residual frustration analysis
revealed highly frustrated regions in both apo- and PPACK-thrombin that are correlated with the highly dynamics surface loops.

Throughout this work, thrombin is shown to have highly dynamic surface loops, which undergo significant, slow timescale motions. These loops contribute to an entropic, allosteric network connecting the active site and ABE1. Both PPACK binding at the active site and TM binding at ABE1 were observed to alter the dynamics and community structure of this network. The presence of the critical TM4 domain resulted in significantly increased correlation and dynamics in the active site loops as well as an increased potential for communication, suggesting that an entropic, allosteric mechanism contributes to the enhanced cleavage of protein C by TM bound thrombin.
### References


