THE UNIVERSITY OF CHICAGO

STATISTICAL LEARNING MODELS OF T CELL RECEPTOR DYNAMICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
JAMES CROOKS

CHICAGO, ILLINOIS
DECEMBER 2015

To Ashley Lane and my family, for all your love and support.

“Would you tell me, please, which way I ought to go from here?”
“That depends a good deal on where you want to get to,” said the Cat.
“I don’t much care where–” said Alice.
“Then it doesn’t matter which way you go,” said the Cat.
“-so long as I get SOMEWHERE,” Alice added as an explanation.
“Oh, you’re sure to do that,” said the Cat, “if you only walk long enough.”
– Lewis Carrol, Alice in Wonderland

“The future is a lens,
where things appear,
clear”
– Iris, Wayseer

ABSTRACT

In this thesis we study the dynamics of the CDR3 loops of the T cell receptor (TCR). The TCR is the protein responsible for mediating the recognition of signs of infection in the T cell, a cornerstone of the adaptive immune system. The CDR loops are responsible for this recognition process, and years of crystallographic work have shed immense light on their interactions with antigens. However, the dynamics remain difficult to study, and the relationship between the flexibility of the loops, their motions, and their interaction with antigen is still poorly understood. Here, we have simulated the dynamics of two different TCR systems with molecular dynamics, and applied machine learning and signal processing technologies to pull apart the dynamics. This thesis gives a detailed background the analytic methods, and then applies them to the dynamics of the 2C and NKT15 TCR clones.

A central question of the thesis asks if the CDR3 loops are flexible in solution and whether they demonstrate stable conformations in the absence of the environment of an antigen the TCR recognizes. Our main results are that the loops demonstrate restricted, coherent motion in solution, and that there exist distinct, stable clusters of conformations, states, of the CDR3 loops. The system undergoes transitions between these distinct conformational clusters, and this transition can be described as Markov system, providing a high level view of the dynamics. Furthermore, the simulation captures known crystallographic bound states. Finally, we show evidence for more restricted and simplified CDR3 motions in the NKT15 TCR clone, which is a TCR with more 'innate-like' behavior, in contrast to the more complex motion of the 2C clone's CDR3 loops, despite their similar architecture.

ACKNOWLEDGEMENTS

In any project of substance, there are many people who contribute, whether directly or indirectly. First and foremost, I would like to thank my advisors, Erin Adams and Ridgway Scott for their support and advice over the years. Ridg convinced me to join the Biophysics program in the first place when I interviewed at UChicago, and I'm proud to have finished under his advisement. Erin has been a source of motivation, advice, and the occasional tough pep talk; she taught me to give a good talk, properly read biology papers, and respect deadlines. She was especially correct on this last point. Without her taking a risk on letting a math student join her immunology lab five years ago, this project would have never happened. The Adams lab has been an excellent family to learn biology in, and Erin has put together an incredible group of people.

Tobin Sosnick and Benoit Roux, my remaining committee members, have contributed immensely over the years. They have challenged my conclusions, pushed me to do better science, and provided excellent advice and resources when the project encountered difficulties, especially when my first simulation turned out to be an artifact. Benoit, in particularly, pushed for the Markov approach that has turned out to be a cornerstone of the project.

Graduate school is a long and difficult time. I wouldn't have made it without good friends. Thank you to Kenneth Barr, Chris Carmean, Josiah Zaynor, Greg Tietjen, Oscar and Caitlin Trejo, Moira Flanagan, Monika and Bjorn Scholz, Ryan Mork, Jennifer Thall, and Thom Pennington for being such great friends, both in graduate school, for those fellow students, and outside the University. I'd also like to thank the Magic crew for the years of games that let us get away from work for awhile.

Further out, I want to thank Rob Agnese and Brandi Phillips for being some of the best friends anyone could ask for, even separated by hundreds of miles. I also want to thank

Jonathan King and Miklos Bona, my undergraduate advisors, for encouraging me to go to graduate school and guiding me through an intellectually rich undergraduate experience.

My family has supported me through the tough times in school, and helped me to reach the point where any of this was possible. Thank you for your love and support over all the years. Finally, I want to thank my fiancee Ashley Lane for her love and support through all the late nights and long days. I couldn't have done it without her, the most incredible partner anyone could ask for.

Table of contents

List of Figures

PREFACE

The adaptive immune system faces a fundamentally difficult problem: correctly identifying signs of infection in a noisy environment, and where the antigen being presented has never been seen by the immune system before.

The T cell receptor (TCR) recognizes antigens presented to it on the surface of cells by MHC and MHC-like proteins. It must do so in an environment where most of the would-be antigens presented to the TCR are presentations of self-molecules, all of the antigens are presented in the context of a self-protein presenter, and where triggering a self-immune response could be deadly for the organism, but failure to respond to a non-self antigen could be equally deadly. Furthermore, if only for logistical reasons, T cells and therefore T cell receptors must be able to both identify different MHCs, which are polymorphic, and different peptides presented by these MHCs, referred to as cross-reactivity, unlike the specific, potent interactions of antibodies.

The difficulty of understanding the fine dynamics of the T cell receptor is not unique. It is an understatement to say that protein dynamics are complicated. Despite the simple assumptions of Newtonian mechanics, molecular dynamics simulations produce immense amounts of data that, at least naively, live a high dimensional phase space. Physical intution, such as the use of dihedral angle descriptions, and willful ignorance, ignoring the fine motions of solvent, let us reduce the phase space when we analyze the data. Nevertheless, we are still left with mere hundreds of dimensions for our phase space, down from thousands. Thus, we turn to the tools of statistical learning and model-building, creating simplified models of the data our model produces so we can extract meaning from the mess of raw numbers.

I have two major goals in this work. The first is to apply these tools to simulations of example T cell receptors with the goal of understanding the flexibility of the CDR loops. In particular, I wish to thoroughly understand one particular aspect of this flexibility: are the loops well structured? By this, I mean to ask whether the loops are flexible in the way that a rope is flexible, capable of bending and flexing essentially anywhere along it's length, or are the loops flexibile in a way more akin to a human dancer, demonstrating flexibility at key joints that together choreograph an elegant dance. The nature of the protein backbone leads us to imagine the latter, but then how is the system choreographed? Can we identify states of the system, poses that it adopts during it's dance and then holds before moving to the next position? Do such poses even exist? Or are there just hinges that swing a lever arm back forth until contact with a target is made?

I will argue for the existence of stable collections of poses, clusters of conformations that are sufficiently similar to one another and which the protein adopts for extended periods of time. These clusters constitute states that we can interpret and understand in a manner analogous to crystal structures, though rather than investigate specific interactions, I will consider the statistics over these states, getting an idea of the general behavior tendency of individual states. Modeling the system as small sets of clusters, we can also extract probable pathways between these states, finally addressing the question of flexibility in a visualizable, intuitive manner.

The second major goal of this work is put forward the T cell receptor as a challenge. In machine learning and many fields of methods development, it is common to demonstrate that a novel method works incredibly well on one or a handful of test systems, and then announce victory. Then, when it comes time for the practitioner to apply this method to his or her system, it fails, often in unexpected ways. The states of the T cell receptor loops that I will pick apart are small, subtle re-roganizations of the backbone of two small, peptide-like segments of the protein. The system is therefore ultimately small, but it's motions are subtle compared to, for example, a folding event. This makes the TCR an excellent system to explore as a test of statistical learning techniques applied to protein dynamics, and a tractable system that exhibits complex but structured, low-dimensional dynamics embedded in a high-dimensional system (as this report will demonstrate). I hope to convince the methodologically-oriented reader that the TCR would make an excellent choice of target once the usual toy and test systems are finished.

CHAPTER 1

INTRODUCTION

Broadly speaking, the vertebrate immune system consists of two primary sub-systems, the innate and adaptive immune systems. The adaptive immune system's role is to generate specific immune responses against pathogens in response to exposure to a novel pathogen, and to retain memory of the pathogen for future responses to that pathogen. The primary mediators of the adaptive immune response are B and T cells, responsible for antibody production and cell-mediated immune responses, respectively. We focus on the role of the T cell, and in particular on the T cell receptor (TCR), a membrane bound protein expressed on the surface of T cells that mediate T cell recognition of pathogens and stimulate the T cell immune response. Figure 1.1 shows the structure of the 2C TCR clone[1].

1.1 T Cells and the T cell receptor

T cells are lymphocytes derived from haematopoietic stem cells in the bone marrow that mature in the thymus (hence T cell) before release into the peripheral blood stream. T cells are distinguished from other lymphocytes by expression of the T cell receptor, a membrane bound heterodimeric protein. T cells fall into several classes each of which serve different roles in the immune system depending on the expression of either α and β or γ and δ chains of the T cell receptor and expression of CD4, CD8, or NK1.1 that differentiate $\alpha\beta$ T cell receptor roles, with $\alpha\beta$ CD4$^+$ or CD8+ forming the most common group of T cells in human peripheral blood.

[1] K.C. Garcia et al.: An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. In: Science 274 (1996), pp. 209–219.

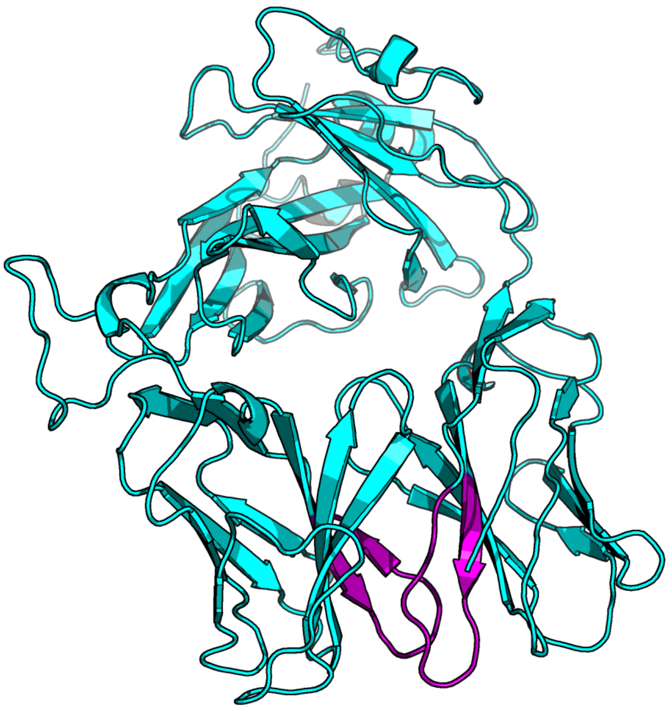

Figure 1.1: Structural view of the 2C TCR (PDB: 1TCR). The recombined CDR3α and CDR3β loops are highlighted in purple.

Similar to B cells, T cells achieve adaptive behavior through somatic recombination of the T cell receptor, analogous to recombination in the generation of antibodies. However, unlike antibodies that are specific to a particular molecular surface, TCRs must recognize the combined surface of an antigen presented in the context of a self-protein, canonically this is the presentation of a viral-genome derived peptide presented by Major Histocompatibility Complex (MHC) class I or class II, polymorphic proteins responsible for presenting peptide antigens, though non-virally derived peptides may be presented through other routes. Figure 1.2 shows a structural view of the 2C clone bound to a peptide-MHC ligandf[2]. In addition to peptide presentation by MHC, T cells recognize lipids and glycolipids presented by CD1a and CD1d, MHC-like molecules; more exotic recognition processes include specific protein

[2] M. Degano et al.: A functional hot spot for antigen recognition in a superagonist TCR/MHC complex, in: Immunity 12 (2000).

surfaces by $\gamma\delta$ T cells[3] and recognition of small molecules presented by MR1[4], though these more exotic recognition behaviors occur in T cells that are more similar to innate immune system than the adaptive. Here, we are only concerned with $\alpha\beta$ T cells, primarily with the classical class that expressed CD4 or CD8, though we will touch on the behavior of the NK T cell which expresses NK 1.1 and recognizes lipids presented by CD1d[5]. TCR recognition of pMHC or another Ag-Presenter complex occurs through binding at the Complementarity Determining Region (CDR) of the TCR, a set of six loops, three contributed by each of the α and β (or γ and δ) domains. The loops are referred to by number and domain, e.g. CDR1α, CDR1β, etc. CDR1 and CDR2 loops are directly encoded by the variable domain gene sequence, while the CDR3 loops undergo somatic recombination during T cell maturation in the thymus.

1.1.1 T cell maturation and MHC recognition

The process of T cell maturation in the thymus is intimately tied to the primary function of peptide-MHC recognition. TCRs need to recognize the peptide-MHC presentation of foreign antigen and ignore peptide-MHC presenting self-peptides; the release of self-peptide recognizing TCRs into the peripheral blood can cause autoimmune reactions. T cells maturing in the thymus undergo two selection processes, positive and negative selection, to ensure that TCRs are not autoreactive, but also sufficiently able to recognize peptide-MHC, including self-peptide presenting pMHC, that the TCRs do not ignore all potential antigens presented to them. Additionally, because of the enourmous set of possible antigenic sequences, and the polymorphism of the MHC proteins that must bind and present the antigen peptides,

[3] A. Sandstrom et al.: The B30.2 domain of Butyrophilin 3A1 binds phosphoantigens to mediate activation of human Vγ9Vδ2 T cells, in: Immunity 2014.

[4] J. Lopez-Sagaseta et al.: The molecular basis for MAIT cell recognition of MR1, in: Proceedings of the National Academy of Sciences of the United States of America 2013.

[5] J. Rossjohn et al.: Recognition of CD1d-restricted antigens by natural killer T cells, in: Nature Reviews Immunology 12 (2012), pp. 845–857.

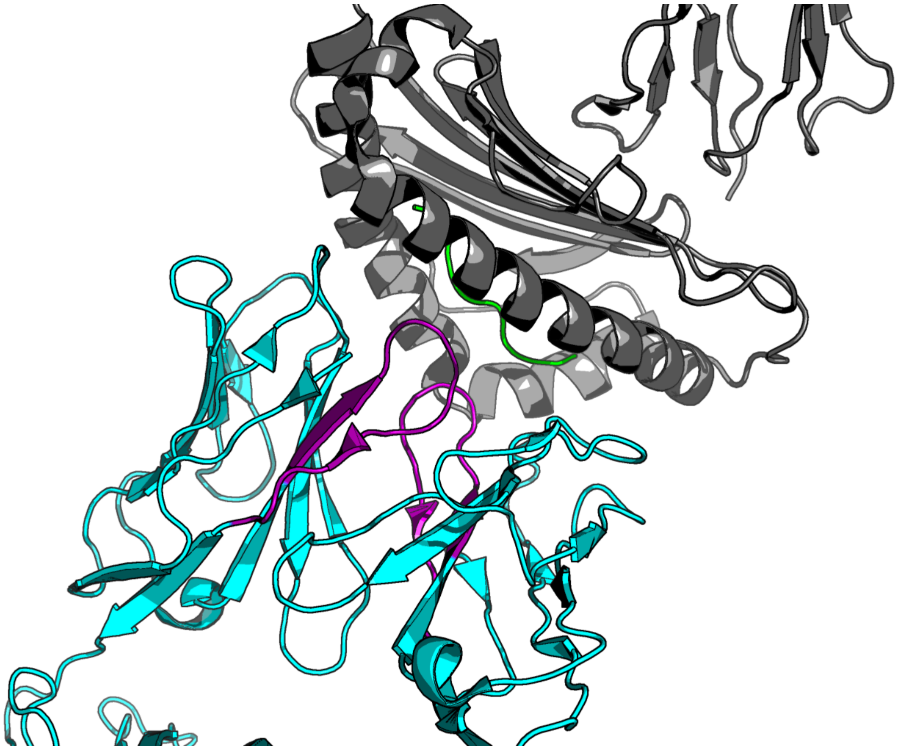

Figure 1.2: Structural view of the 2C TCR (cyan) bound to the 2-K^b MHC (grey) presenting the SIYR peptide (green) (PDB: 1G6R). CDR3 loops shown in purple.

TCRs are cross-reactive with different pMHC complexes[6]. This a major difference from the very specific interactions that define antibody target recognition. Cross-reactivity is a major reason to expect loop flexibility and alternative conformational states, as different conformations enhance interactions with different pMHC complexes.

The conventional class of $\alpha\beta$ T cells expressing either CD4 or CD8 begin in the thymus as CD4$^-$/CD8$^-$ ('double negative') T cells. These cells express a germline encoded variable domain α and β chain, and the domains selected establish the amino acid sequence of the CDR1 and CDR2 loops of both chains. Recombination leads to diversity in the CDR3 segments of both α and β chains. At the double-negative state, the process of V-D-J recombination by RAG1 and RAG2 proteins generates a repertoire of β chain sequences. Mutational studies

[6] Don Mason: A very high level of crossreactivity is an essential feature of the T-cell receptor, in: Immunology Today 19.9 (1998), pp. 395–404, URL: http://dx.doi.org/10.1016/S0167-5699(98)01299-7; A.K. Sewell: Why must T cells be cross-reactive?, in: Nature Reviews Immunology 12 (2012), pp. 669–677, URL: http://www.nature.com/nri/journal/v12/n9/abs/nri3279.html.

in mice have demonstrated that this is a necessary step, without β chain rearrangement[7], T cells do not proceed to the double-positive CD4$^+$/CD8$^+$ stage and the thymus shrinks 60-fold[8]. Successful rearrangement that results in ligand (pMHC) engagement allows progression to the next stage, ensuring that the T cells that progress are capable of engaging pMHC. This triggers differentiation into the double-positive stage, and proliferation of the cells that pass the checkpoint[9].

Double-positive T cells in the thymus then undergo negative selection, which serves to protect against autoimmunity. V-J recombination occurs in the α locus, leading to a repertoire of CDR3α sequences. The double-positive T cells with recombined α and β segments are exposed to 'ubiquitous' self-peptide antigens by professional antigen presenting cells[10], which seem to be a mixture of thymus cortical dendritic cells and cortical thymic epithelial cells[11]. Those double-positive T cells that demonstrate high-affinity for self-peptides are killed, deleting the self-recognition sequences from the T cell clonal repertoire.

Structural and Dynamic implications of Thymic Selection

The CDR3β sequence undergoes V-D-J recombination and is selected for pro-binding behavior, it reasonably follows that the CDR3β would be biased toward recognition and more promiscuous binding of peptide targets. Furthemore, recombination with diversity (D) segments is biased towards in the inclusion of glycine residues[12] which generally increase the

[7] Y. Shinkai et al.: RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement, in: Cell 1992; P. Mombaerts et al.: RAG1-deficient mice have no mature B and T lymphocytes, in: Cell 1992.

[8] E. Robey/B.J. Fowlkes: Selective events in T cell development, in: Annu. Rev. Immunol. 1994.

[9] Ibid.

[10] Ibid.

[11] L. Klein et al.: Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see), in: Nature Reviews Immunology 2014.

[12] E.Q. Roldan et al.: Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells, in: Immunogenetics 1995.

flexibility of protein regions as they become more common due to the lack of steric clashes from sidechain interactions – glycine motions are effectively only due to long-range stressors on the backbone orientation and very close hydrogen bonding interactions. On the other hand, CDR3α only undergoes V-J recombination, and so lacks the bias toward more flexible motion and faster dynamics implicated by glycine-rich regions. The negative selection process would presumably bias the CDR3α toward dynamics that reduce binding affinity, fitting with the slower dynamics and possible 'on'-'off' switch behavior observed in simulations of 2C (this work) and A6[13].

1.2 Crystallography of $\alpha\beta$ T cell receptor recognition

Over two decades of crystallograpic work have generated a large database of TCR structures, both free and bound to various foreign and self-reactive peptide-MHC complexes demonstrating significant variation in bound structure that show CDR loop flexibility as vital to TCR cross-reactivity. Reviews of the structural data over the years have concluded, with increasing conviction, that the CDR loops are flexible but in a structured manner distinctly different from the intrinsically disordered regions seen some other proteins[14]. Furthermore, general flexibility is restricted to the CDR3 loops, even under extreme changes to CDR3 loop length[15].

A large range of re-arrangements are seen in the CDR3α and CDR3β loops between the bound and unbound states of different TCRs, with the re-arrangements varying with the

[13] D.R. Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism, in: Journal of Molecular Biology 414 (2011).

[14] K.C. Garcia/E.J. Adams: How the T cell receptor see antigen - a structural view, in: Cell 122 (2005); M.G. Rudolph/R.L. Stanfield/I.A. Wilson: How TCRs bind MHCs, peptides, and coreceptors, in: Annual Review of Immunology 24 (2006); K.M. Armstrong/F.K. Insaidoo/B.M. Baker: Thermodynamics of T-cell receptor-peptide/MHC interactions: progress and opportunities, in: Journal of Molecular Recognition 104 (2008); Brian M. Baker et al.: Structural and dynamic control of T-cell receptor specificity, cross-reactivity, and binding mechanism, in: Immunological Reviews 250 (2012), URL: `http://onlinelibrary.wiley.com/doi/10.1111/j.1600-065X.2012.01165.x/full`.

[15] J.B. Resier et al.: CDR3 loop flexibility contributes to the degeneracy of TCR recognition, in: Nature Immunology 4 (2003).

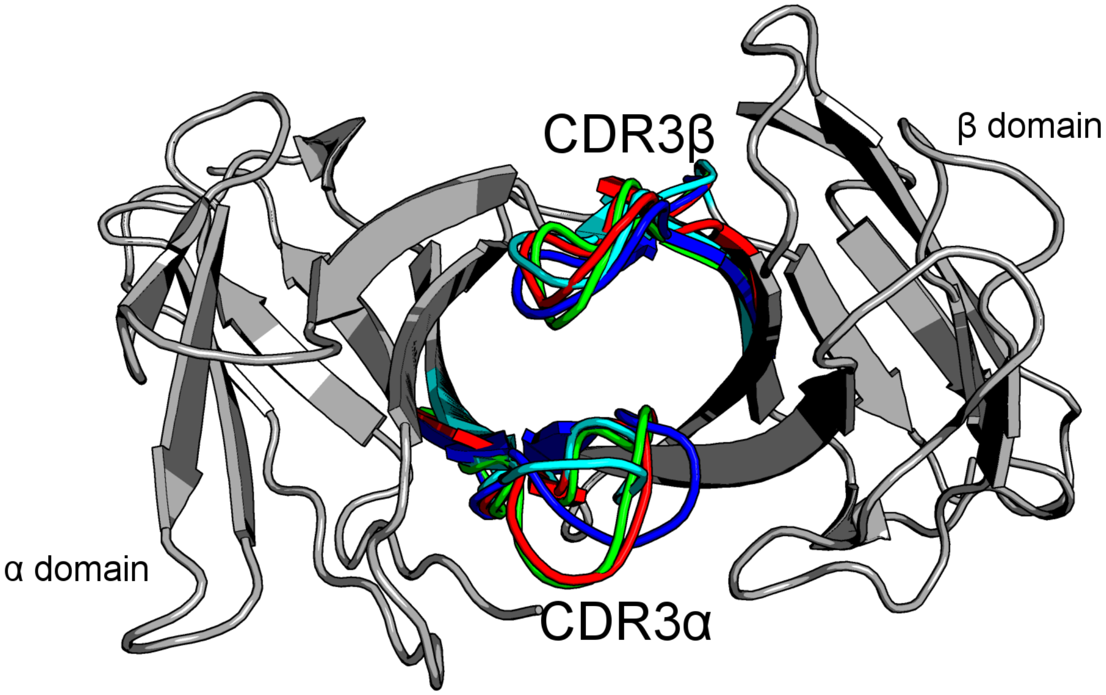

Figure 1.3: Variable domains (grey) of 2C shown from the perspective of the pMHC surface. CDR3 loops shown in color, with unbound loops (cyan, PDB: 1TCR) overlaid with CDR3α and CDR3β loops of 2C bound to MHC/peptide ligands H-2K^b/SIYR (red, PDB: 1G6R), H-2L^d (blue, PDB: 2OI9), and H-2K^b/dEV8 (green, PDB: 2CKB).

particular peptide-MHC ligand. CDR3β generally shows the largest variation in position, with up to 8 angstrom changes in Cα position of tip residue observed; on the other hand CDR1β and CDR2β generally show the least movement, suggesting the germline encoded β chain residues primarily function to bind the MHC platform itself, fitting with the positive selection process and crystallographic footprints[16]. Small changes in sequence can also induce significant changes in bound state, as several point mutants in CDR3α of 2C have demonstrated significant re-arrangement of the α loop[17].

Focusing on 2C specifically, which is the focus on the present work, Figure 1.3 shows alignment of 2C variable domains to several bound crystal structures[18]. Both CDR3 loops show variation between the bound and unbound states, as well as variation within the bound states. Note however that CDR3α makes crystal contacts due to packing in the unbound structure, making the unbound orientation of CDR3α indeterminate, though the local energy well of the apparent unbound state is broad and well-separated from the bound states in our data, and a similar difference is observed in simulations of A6[19].

1.3 TCR signaling and kinetic proofreading

The exact mechanics of TCR signaling remain an open problem, but the kinetic proofreading model provides a good phenomenological model of the signaling process, and fits with the understood biology. Kinetic proofreading (KPR) is a model originally proposed for enzymatic

[16] Armstrong/Insaidoo/Baker: Thermodynamics of T-cell receptor-peptide/MHC interactions: progress and opportunities (see n. 14).

[17] K.C. Garcia et al.: Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen, in: Science 1998.

[18] Garcia et al.: An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. (see n. 1); Degano et al.: A functional hot spot for antigen recognition in a superagonist TCR/MHC complex (see n. 2); L.A. Colf et al.: How a single T cell receptor recognizes both self and foreign MHC, in: Cell 129 (2007); Garcia et al.: Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen (see n. 17).

[19] Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism (see n. 13).

reactions that must differentiate between correct and incorrect reaction pathways. In KPR, the reaction uses a time-delay in the form of the kinetics of a multistep reaction to improve the error rate beyond what would be expected from the free energy difference of the outcomes[20]. TCR signaling proceeds by multi-step phosphorylation of the T-cell receptor ζ chain[21], acting as a signal amplifier circuit described by the KPR model, and effectively describing the basic TCR signaling process[22]. This view over-simplifies the process of TCR signaling, but to first order it shows the importance of kinetics over direct affinity measurements; biochemical experiments have shown that for many TCRs, signaling is well-correlated with the binding dwell time half-life, though this correlation is not universal, leading to debate between k_{off} and K_D being of primary importance[23]. A major contribution of more recent KPR models of TCR signaling is explaining pMHCs that act as antagonists of signaling, effectively by competing for binding but with sufficiently fast off-rates that the signaling process fails to complete and resets instead, blocking activation by slower off-rate binders[24]. The upshot of KPR as a signaling model for TCR activation is that when considering the physical dynamics of the TCR, we are interested in how the time scale of binding events and which model of TCR-pMHC interaction fits with the observed kinetics.

[20] J.J. Hopfield: Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity, in: Proceedings of the National Academy of Sciences of the United States of America 1974.

[21] R.N. Germain/I. Stefanová: The Dynamics of T Cell Receptor Signaling: Complex Orchestration and the Key Roles of Tempo and Cooperation, in: Annu. Rev. Immunol. 1999.

[22] T.W. McKeithan: Kinetic proofreading in T-cell receptor signal transduction, in: Proceedings of the National Academy of Sciences of the United States of America 1995.

[23] J.D. Stone/A.S. Chervin/D.M. Kranz: T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity, in: Immunology 2009.

[24] P. Francois et al.: Phenotypic model for early T-cell activation displaying sensitivity, specificity, and antagonism, in: PNAS 2013.

1.4 Models: Induced Fit, Conformational Selection, and Conformational Melding

Ultimately, we are interested in the flexibility of the CDR loops because we are interested in how TCRs bind to pMHC and physical mechanisms, including loop flexibility and dynamics, by which TCRs differentiate self from non-self. Several models exist, which we refer to as the Induced Fit[25], Conformational Selection[26], and Conformational Melding models[27]. The issue of flexibility is not whether it occurs at all, but rather how much is intrinsic to the CDR3 loops themselves, and how much is driven by the environment. This question is expressed in the tension between the induced fit and pre-existing equilibrium models of TCR binding.

Induced Fit Model

The induced fit model argues for initial weak binding between the TCR and MHC which allows for a conformational change to make stronger contacts resulting in a higher affinity interaction with recognized peptides. The Induced Fit model is the most well-supported from biochemical evidence; structural evidence is inconclusive as we can't know if the observed differences are due to the binding process inducing the conformational changes or if they are selected from pre-existing equilibrium states. ITC experiments have shown heat capacity and entropy changes upon binding, while binding analysis with varying temperature have shown the both association and dissociation depend on temperature, together indicating

[25] L.C. Wu et al.: Two-step binding mechanism for T-cell receptor recognition of peptide-MHC, in: Nature 418 (2002).

[26] P.D. Holler/D.M. Kranz: T cell receptors: affinities, cross-reactivities, and a conformer model, in: Molecular Immunology 40 (2004).

[27] S.J. Gagnon et al.: T cell receptor recognition via cooperative conformational plasticity, in: Journal of Molecular Biology 363 (2006); W.F. Hawse et al.: TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility, in: J Immunol 192 (2014), pp. 2885–2891, URL: http://www.jimmunolo.org/content/192/6/2885.

conformational changes upon binding and unbinding[28].

Of particular interest, structural studies have suggested a strong role for the CDR1β and CDR2β germline-encoded loops, which have shown the least rearrangement upon binding, in MHC recognition. This would provide the necessary initial bias toward MHCs required for the induced-fit model. Furthermore, experiments with ‘leaky’ negative selection mouse models where T cells were able to occasionally escape deletion despite failure at the negative selection stage have shown affinity for MHC, suggesting a germline bias toward MHC beyond that generated by the negative selection process[29].

Conformational Selection

An alternative ‘conformer’ model suggests that cross-reactivity could instead be driven by the existence of multiple CDR loop conformational states of the free TCR, which could recognize different peptide-MHC ligands so that specificity is controlled by a combination of specific contacts and the relative equilibrium populations of different conformational states[30]. These two models are difficult to distinguish biophysically as loop dynamics are difficult to capture even with techniques capable of resolving time-dependent dynamics[31], though the measurements did substantiate the use of computational methods. Computational analysis of the free A6 TCR provides strong support for the existence of distinct states in solution, where clustering of the CDR3α and CDR3β loops using RMSD as a dissimilarity metric

[28] J.J. Boniface et al.: Thermodynamics of T cell receptor binding to peptide-MHC: evidence for a general mechanism of molecular scanning, in: Proceedings of the National Academy of Sciences of the United States of America 96 (1999).

[29] S. Dai et al.: Crossreactive T Cells spotlight the germline rules for alphabeta T cell-receptor interactions with MHC molecules, in: Immunity 2008; E.S. Huseby et al.: How the T cell repertoire becomes peptide and MHC specific, in: Cell 2005.

[30] Holler/Kranz: T cell receptors: affinities, cross-reactivities, and a conformer model (see n. 26).

[31] D.R. Scott et al.: Limitations of time-resolved fluorescense suggested by molecular simulations: assessing the dynamics of T cell receptor binding loops, in: Biophysical Journal 103 (2012).

showed multiple distinct conformations of the loops[32]. Notably, CDR3α showed two distinct clusters of conformations with implied slow motions between the two conformational clusters. One cluster resembled the bound conformation of the CDR3α loop of A6, while the other cluster was distinct from the bound conformation. On the other hand, CDR3β showed multiple, smaller clusters, with much faster apparent transitions between the conformational clusters.

Conformational Melding

Conformational melding is a more recently proposed model that combines aspects of induced fit and conformational selection, and includes the role of pMHC conformation and dynamics[33]. Small changes in flexibility in both TCR and pMHC have shown effective changes in recognition[34], and small changes in MHC sequence can cause changes in peptide dynamics while bound in the MHC groove[35]. Similarly, NMR experiments have demonstrated flexibility and mobility of the CDR3β loop of 2C while bound to L^d/QL9[36], demonstrating that dynamical motion occurs even in the bound state. Conformational melding suggests that the dynamics of the pMHC and TCR may essentially match one another, so that the energy diagrams 'agree', but if such a match isn't found, the system unbinds due to an inability

[32] Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism (see n. 13).

[33] O.L. Borbulevych/K.H. Piepenbrink/B.M. Baker: Conformational melding permits a conserved binding geometry in TCR recognition of foreign and self molecular mimics, in: J. Immunol. 2011.

[34] O. Y. Borbulevych et al.: T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility, in: Immunity 31 (6 2009), URL: http://www.sciencedirect.com/science/article/pii/S1074761309004981.

[35] J.K. Archbold et al.: Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition, in: J. of Exp. Med. 2009; H. Fabian et al.: HLA-B27 subtypes differentially associated with disease exhibit conformational differences in solution. In: J. Mol. Biol. 2008.

[36] Hawse et al.: TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility (see n. 27).

to find a strong energy minima in the combined dynamics[37]. Flexibility of motion and having different sets of local dynamics available then contributes to cross-reactivity while still maintaining specificity. At the TCR-pMHC encounter, the TCR has an initial conformation selected from a set of possible states, from which a 'local search' for matching dynamics can proceed, resulting in the observed slower kinetics.

A key requirement of the conformer models and conformational melding hypotheses is the existence of distinct crystal-like states in the unbound TCR's dynamics. We have run extensive simulations of the free 2C TCR, as well as simulations of the free Natural Killer T cell receptor NKT15 to driectly address flexibility. 2C is a well studied TCR known to display significant cross-reactivity and with extensive crystallographic data available. We have generated a total of 3μs of data across 10 trajectories of 2C, providing a significantly larger data than has previously been available to study the solution state dynamics of a single TCR. Further, we have used the Markov State Model formalism, described in Chapter 3, to cluster the loop conformations in a kinetic fashion, providing crystal-like states that distinguish stable conformations from transitions and directly identifying transitions between conformational states. In accordance with previous work on A6, we observe significant flexibility in both CDR3α and CDR3β, with CDR3β showing a broad energetic well that is kinetically separated from bound-like conformations, and CDR3β showing multiple metastable states with local equilibria.

Bulk binding kinetics do not distinguish the models

SPR measurements have provided binding kinetics and affinity data for a large variety of $\alpha\beta$ TCRs. The kinetic proofreading model argues that the half-life of the interaction dominates the signaling process, though it is still debated whether the key quantity is the binding

[37] Hawse et al.: TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility (see n. 27).

affinity of the dwell-time[38]. TCRs show a large diversity of kinetic parameters, with one review showing k_{on} values ranging from a low of 633 per mole second to a high of 400000 per mole second[39] for various TCRs binding different ligands. Similarly, k_{off} varies from .009 per second to .975 per second. The on and off rates tend to move together, and generally show smaller ranges for a fixed TCR; the 2C clone under study in this work shows on rates in the 2200-22000 per mole second range and off rates in the .025-.464 per second range. The highest values for 2C both occur when binding to SIYR/K^b, and are unusually fast for 2C.

Considering the models presented above, it would seem that the models could be distinguished by using the kinetic data. This is partially made difficult by the very large range of observed rates, with on rates varying by three orders of magnitude. If we consider only 2C, this shrinks to a single order of magnitude. The question is whether we estimate a bound on the rates that would differentiate the models.

Let pMHC be in a fixed position, approximating the set-up of an SPR experiment, and assume the collision rate is diffusion limited with the TCRs diffusing via translation, so we begin with an encounter rate of $10^9 M^{-1}s^{-1}$. The binding footprint of the TCR on the pMHC interface is highly conserved across conventional $\alpha\beta$ TCRs and whether by selection processes or germline bias, the CDR1 and CDR2 loops can make a stable encounter complex with pMHC. If we consider the fastest on rate of binding as an order of magnitude slower than forming the encounter complex, then forming the complex reduces the rate by a factor of approximately 10^2, leaving an initial upper bound of $10^7 M^{-1}s{-}1$.

From the perspective of the binding energy of the TCR over time, the induced fit and conformational melding models look similar, so we treat them together here when discussing ranges of on rates. In these search-based models, the TCR forms an initial complex with the MHC via the CDR1 and CDR2 loops, shown as the initial stable encounter complex in figure 1.4. As both models presume that the solution conformation of the TCR is not the preferred

[38] Stone/Chervin/Kranz: T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity (see n. 23).

[39] Ibid.

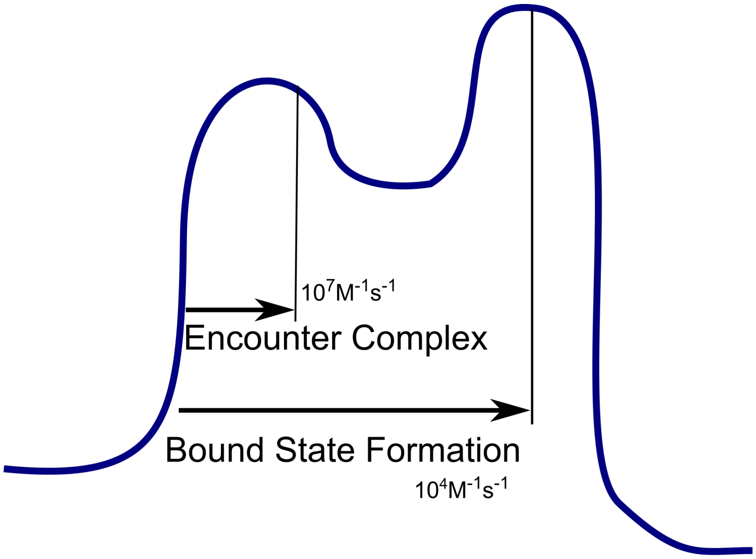

Figure 1.4: Proposed energy diagram of the TCR energy over time during the binding interaction with pMHC for the search models (induced fit and conformational melding).

binding conformation, there is an energetic barrier to re-arranging one or both of the CDR3 loops to find the final bound conformation, along with an entropic cost of freezing out the alternative conformations. This is the higher right-hand peak of the energy diagram in figure 1.4. We can estimate the bulk kinetic rate as the rate of complex formation decreased by a factor depending on the success rate of the encounter complex proceeding to the final bound state. The encounter complex proceeds to the bound state if the system finds the correct conformation before a disassociation event occurs. As a simple estimation, model both the search process and the disassociation process as independent Poisson processes. The probability that disassociation does not occur in a time segment of length t is the Poisson distribution for zero events,

$$\mathbb{P}_0(t;\lambda) = e^{-\lambda t}$$

The expected waiting time for the search to succeed is distributed according to the exponential distribution, with probability density function

$$f\left(t;\lambda\right) = \lambda e^{(-\lambda t)},$$

where λ is a parameter of the model, the rate constant, for both the Poisson and Exponential distributions. Let λ_{bound} and λ_{dis} denote the rate constants for the search finding the correct bound state and the disassociation event, respectively. We want to calculate the probabilty tha the search succeeds at time t and no disassociation events occur. This is given by

$$\int_{t=0}^{\infty} f(t;\lambda_{bound})\mathbb{P}_0(t;\lambda_{dis}) = \int_{t=0}^{\infty} \lambda_{bound} e^{-\lambda_{bound}t} \cdot e^{-\lambda_{dis}t} dt =$$
$$\int_{t=0}^{\infty} \lambda_{bound} e^{-(\lambda_{bound}+\lambda_{dis})t} dt = \frac{\lambda_{bound}}{\lambda_{dis}+\lambda_{bound}}$$

If $\lambda_{bound} \ll \lambda_{dis}$, then we can approximate $\frac{\lambda_{bound}}{\lambda_{dis}+\lambda_{bound}} \approx \frac{\lambda_{bound}}{\lambda_{dis}}$. Next, note that the λ parameter of the exponential distribution is the inverse of the mean of the distribution. So we can estimate λ_{dis} and λ_{bound} directly from the timescales of the processes. Assume the disassociation process of the encounter complex has timescale faster than the formation of the encounter complex, i.e. $10^{-8}s$, which gives $\lambda_{dis} = 10^8 s^{-1}$. We expect that the encounter complex is relatively unstable, so disassociation in absence of finding the correct bound state should be faster than the initial association. If the search process is successful on the 10 microsecond timescale, then we have $\lambda_{bound} = 10^5 s^{-1}$, which yields a success probability on the order of 10^{-3}, so we would expect on rates on the order of $10^7 M^{-1}s^{-1} \cdot 10^{-3} = 10^4 M^{-1}s^{-1}$. Faster search on the order of 1 microsecond yields a rate on the order of $10^3 M^{-1}s^{-1}$.

The CDR3α loop may or may not be able to re-arrange in this context, but if we assume it does in a pure induced fit type model, CDR3α has slower kinetics than CDR3β between states[40], and rates have not been established but both the cited study and the present work will argue that CDR3α rearrangement between bound-like and unbound-like states occurs on at least the microsecond timescale, and possibly much longer. Similarly, the present work will show the timescales of the CDR3β loop re-arrangements to occur in the range of hundreds of nanoseconds to microsecond for individual state changes. If the motions of the

[40] Scott et al.: Limitations of time-resolved fluorescense suggested by molecular simulations: assessing the dynamics of T cell receptor binding loops (see n. 31).

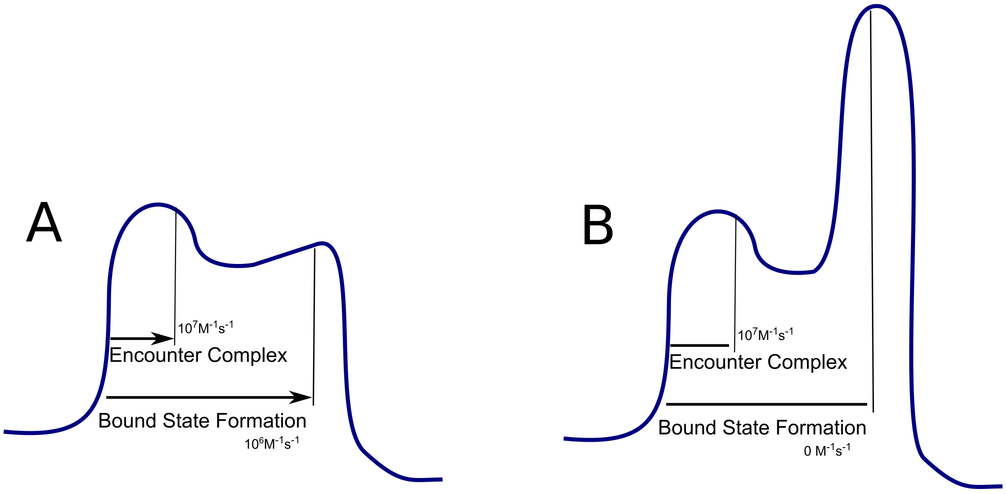

Figure 1.5: Proposed energy digrams of the TCR energy over time during the binding interaction with pMHC for the conformational selection model. (A) Proposed energy diagram when the encounter complex is in the binding-capable conformation. (B) Proposed energy diagram when the encounter complex is in a binding-incapable conformation; the energy of the second hill is expected to be sufficiently high that the reaction only reverses from the metastable state and never completes. Bulk kinetic rates of the conformational selection model are determined by the ratio of occurences of each type of diagram, rather than the behavior of a single diagram.

CDR3 loops are heavily restricted (i.e. have few degrees of freedom) and the topology of state changes is complex, in particular if it is not a fully connected graph, then the time to find the correct state can reasonably occur on the 10-100 microsecond timescale, yielding on rates of $10^3 - 10^4 M^{-1} s^{-1}$, which fits within the lower bound of experimentally observed values.

Under the conformational selection model, there are two types of energy diagrams we might expect, depending on whether the TCR is in the binding-capable conformation on collision. If the TCR is in the correct conformation, then the energetic barrier of continuing to the bound state is minimal, and possibly simply a downward slope, as shown in figure 1.5A. Assume that the energetic barrier of a collision where the TCR is in a binding-incapable conformation is sufficiently high that the probability of proceeding to the bound state is essentially 0 (figure 1.5B), so that only interactions with binding-capable conformations proceed to the bound state, and for simplicity assume that this always happens when the correct conformation is encountered. Then the bulk binding kinetics depends on the probability of the TCR being in the binding capable state. Since previous work and the data we present here show that the CDR3α loop has very slow transitions between stable bound-like and unbound-like conformations, suggesting that the CDR3α loop has a binding capable and a binding incapable conformation, we consider selection as requiring both loops to be in the proper states. Consider the system where the CDR3α loop occupies each of these states with equal probability. Furthermore, previous work shows there are a larger number of states for the CDR3β loop. If there are only five states, equally populated, and one state is binding capable for a given ligand, then there is a $\frac{1}{10}$ chance the encounter complex binds, for a net rate of $10^6 M^{-1}s^{-1}$, which is the order of magnitude of the fastest rate observed. On the other hand if the binding capable state of CDR3α is occupied with a frequency of $\frac{1}{10}$, which is not unreasonable, and the binding capable state of CDR3β is 1% of the equilibrium population, then the encounter complex binds $\frac{1}{1000}$ of the time, with a net rate around $10^4 M^{-1}s^{-1}$, near the rate of 2C's binding kinetics and in the middle of the observed rates. If there are more difficulties in forming the encounter complex due to mis-alignment during the encounter, the rates would be slower, as commonly described in models of soluble protein collisions[41]. Despite the common belief that induced fit is the better model because of the slow on rates observed, a reasonable conformational selection model can still accomodate slow bulk kinetics.

From this, we conclude that the bulk binding kinetics alone cannot distinguish between the conformational selection model on one hand, and the induced fit or conformational melding models on the other. In particular, if there exist distinct states, which is suggested by the rigid transformations observed in crystal structures, then in addition to bulk kinetics, we need to at least know what states are binding capable and what their equilibrium populations are in order to determine how much conformational selection could affect the observed bulk

[41] J. Janin: The Kinetics of Protein-Protein Recognition, in: Proteins: Structure, Function, and Genetics 1997.

rate.

Fast kinetics, Slow kinetics

An important partition exists in $\alpha\beta$ TCR recognition between fast-on/fast-off TCR binding and slower kinetics where the off-rate essentially controls the stimulation response. Following the kinetic proofreading model, we expect that recognition is effectively controlled by the off-rate; at the spatial resolution of an individual TCR-pMHC interaction, if we model unbinding as a Poisson process then a faster off-rate translates to a higher probability of unbinding over a given segment of time, and ultimately to a higher probability of unbinding before all of the KPR checks complete. This leads to a recognition failure. However, unaccounted for in standard KPR models is the time required for dephosphorylation – KPR models generally assume it occurs instantaneously if the TCR leaves, since the timescale of diffusion is faster than the on-rate, and hence the TCR diffuses away, effectively resetting the system. However, there exists a class of TCRs where the on-rate is faster than diffusion, and re-binding events occur. With simple mathematical models of the probability of rebinding rather than diffusing away, it has been shown that stimulation is well correlated by taking into account both affinity measurements and rebinding probability[42]. This mechanism elegantly explains a number of TCRs whose stimulation is poorly correlated with direct affinity measurements and heat capacity measurements that suggest a conformation-dependent mechanism. The 2C TCR, which we study here, is an exemplar of the slow kinetics category, but the conformational dynamics at play in 2C likely do not generalize to the category of TCRs that exploit rapid re-binding events to pass phosphorylation checks during signaling.

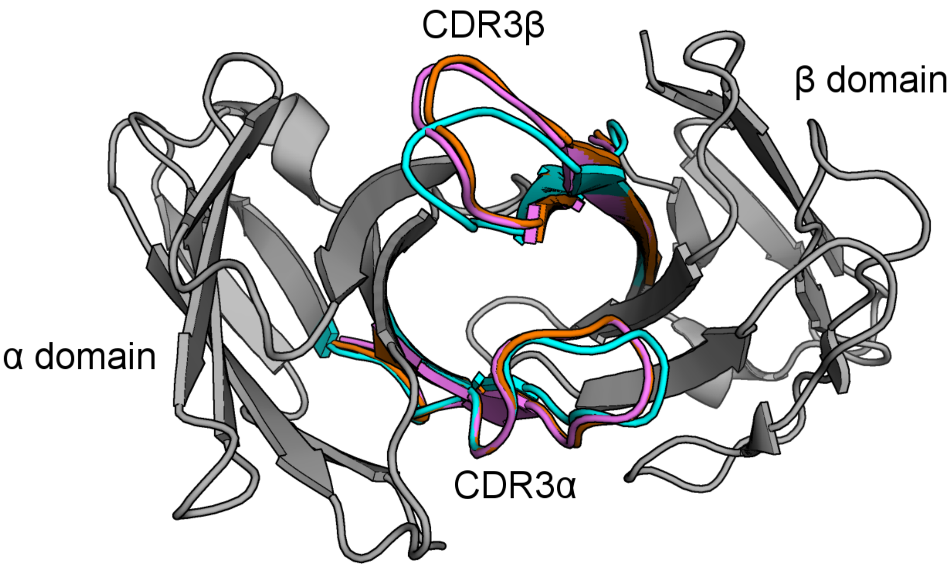

Figure 1.6: Variable domains (grey) of NKT15 shown from the perspective of the pMHC surface. CDR3 loops shown in color, with unbound loops (cyan, PDB: 2EYS) overlaid with CDR3α and CDR3β loops of NKT15 bound to CD1d with αGalCer (orange, PDB: 3HUJ) or C20:2 (pink, PDB: 3VWJ).

1.5 The Type I Natural Killer T cell receptor

In contrast to CD4$^+$ and CD8$^+$ $\alpha\beta$ T cells, type I Natural Killer T cells (NKT) recognize lipids presented by the monomorphic MHC-like molecule CD1d[43]. Type I NKT TCRs are considered to be 'semi-invariant', as they are generated through VDJ recombination as per standard $\alpha\beta$ TCRs, but use a heavily restricted Vα and Vβ chain repertoire. This restriction, along with orders of magnitude faster binding kinetics, higher affinities, and rigid binding conformations[44] in crystal structures have led them to being considered 'innate-like'[45]. Figure

[42] C. C. Govern et al.: Fast on-rates allow short dwell time ligands to activate T cells, in: Proceedings of the National Academy of Sciences of the United State of America 107 (19 2010), URL: www.pnas.org/cgi/doi/10.1073/pnas.1000966107.

[43] Rossjohn et al.: Recognition of CD1d-restricted antigens by natural killer T cells (see n. 5).

[44] Y. Li et al.: The Vα14 invariant natural killer T cell TCR forces microbial glycolipids and CD1d into a conserved binding mode, in: J. Exp. Med. 2010.

[45] Rossjohn et al.: Recognition of CD1d-restricted antigens by natural killer T cells (see n. 5).

1.6 shows the similarity of bound and unbound NKT15 structures[46]. CDR3α shows minimal rearrangement upon binding, while the CDR3β shows some re-arrangement, but the bound orientation is identical for both ligands. Importantly, mutational studies have CDR shown CDR2β and CDR3α to drive the NKT interaction with CD1d and the canonical antigen, αGalCer. The conserved binding footprint, innate-like kinetics, and stronger reliance on germline encoded interactions implies that type I NKT TCRs should demonstrate reduced flexibility and simpler dynamics behvaior, particularly in the CDR3α loop, compared to classical CD4$^+$ and CD8$^+$ $\alpha\beta$ TCRs. The restricted binding footprint in the NKT-Ag-CD1d system suggests that the NKT TCRs should serve as 'innate-like' counterpoints to CD4$^+$/CD8$^+$ $\alpha\beta$ TCRs, despite sharing the same fundamental protein architecture.

To test this model, we have simulated 1μs of free NKT15 dynamics across ten trajectories. Surprisingly, we observe flexibility and meta-stable states in both the CDR3α and CDR3β loops of NKT15. However, in contrast to 2C, the dynamic behavior of NKT15's loops is simpler in that for each loop, the major motions can be well-captured by a single degree of freedom.

1.6 Aims

We study the solution state dynamics of the class 2C $\alpha\beta$ T cell receptor. In doing so, we argue for the existence of crystal-like states of pre-existing equilibria in CD4$^+$/CD8$^+$ $\alpha\beta$ T cell receptors with slow kinetics. However, we do not find that the bound states are well represented by the local minima of the these states, rather the bound states appear on the periphery, suggesting that there is a local search for the final bound state that is seeded by a pre-existing equilibrium. States that are well-separated from the states similar

[46] L. Kjer-Nielsen et al.: A structural basis for selection and cross-species reactivity of the semi-invariant NKT cell receptor in CD1d/glycolipid recognition, in: J. Exp. Med. 2006; D.G. Pellicci et al.: Differential recognition of CD1d-alpha-galactosyl ceramide by the V beta 8.2 and V beta 7 semi-invariant NKT T cell receptors, in: Immunity 2009; K.S. Wun et al.: Ternary crystal structure of the human NKT TCR-CD1d-C20:2 complex, in: J. Biol. Chem. 2012.

to the bound structures are potentially alternative seeds for other binding targets, or binding incompetent states that reduce affinity, likely as a recombination-induced affinity regulation mechanism. Further, we take a look at NKT15 on the hypothesis that it should show simpler, and potentially less, motions than 2C due to the innate-like binding kinetics and lack of significant flexibility observed in crystal structures. Surprisingly, NKT15 shows significant loop flexibility, but the kinetics of NKT15's dynamics are simplier in our simulations than those observed in 2C.

In order to show this, we have generated a significant quantity of molecular dynamics simulation data and applied recent developments in dimensionality reduction and statistical learning specific to molecular dynamics simulations. Our data set is significant in the context of previous work, the largest collective simulation is 460ns of a single TCR in solution, with the longest trajectory at 260ns[47]; large collections of TCRs and TCR-pMHC interactions have been simulated before, but the data for any single system was limited to 100ns[48], making the data presented here among, if not the, largest available for a single TCR in solution.

Our analysis of TCR dynamics rests entirely on the application of MD-specialized signal processing and machine learning methods and interpretation of the results. Chapters 2 and 3 therefore explore these techniques. Chapter 2 explores linear dimensionality reduction and clustering techniques, specifically Principle Components Analysis, time-structured Independent Component Analysis, and the k-means and k-medoids clustering algorithms. PCA is a classical technique and needs little introduction, but we develop it here so the intimiate connection to tICA becomes apparent, and to set the stage for later connections to non-linear techniques. The clustering algorithms are again classics by this point, but

[47] Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism (see n. 13); Scott et al.: Limitations of time-resolved fluorescense suggested by molecular simulations: assessing the dynamics of T cell receptor binding loops (see n. 31).

[48] B. Knapp/J. Dunbar/Deane C.M.: Large Scale Characterization of the LC13 TCR and HLA-B8 Structural Landscape in Reaction to 172 Altered Peptide Ligands: A Molecular Dynamics Simulation Study, in: PLoS Computational Biology 10 (8 2014), URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003748.

are described because they both underpin the Markov state models described in Chapter 3, and their failure modes with respect to molecular dynamics are a major motivation for the Markov state model approach.

Chapter 3 develops the Markov state model methods to a level appropriate for following the biological results presented in Chapter 4, and should prepare the thorough reader with sufficient background to extend the results presented here. These two chapters thus serve as an introductory survey to applying machine learning methods to molecular dynamics and literature review of specific techniques. This project has drawn on a fairly diverse range of fields and knowledge, and each chapter thus mixes a light textbook background with more modern literature review.

The results are presented in Chapter 4, using the developed techniques to finally extract metastable states of the CDR loops. In doing so, we find that the CDR loops are tightly constrained, low-dimensional systems.

Finally, Chapter 5 presents a discussion of these results and the implications for the three models discussed earlier. Additionally, future experimental and *in silico* paths are suggested for extending these results. The appendicies cover additional background material: appendix A discusses stochastic simulations, appendix B covers alternative machine learning methods that showed poor results on the analyzed data set earlier in the analysis process, and appendix C covers methodological detail of the simulations and analysis for the investigator looking to reproduce or extend this work.

CHAPTER 2

DIMENSIONALITY REDUCTION AND CLUSTERING

There are two major motivating questions for this work. Do there exist discrete conformational states of the CDR loops of the T cell receptor, and how free or restrained are the motions of the CDR loops? The main tool to explore these questions is simulation of the molecular dynamics of two chosen TCR systems, the 2C TCR and the NKT15 TCR.

The fundamental idea of molecular dynamics is simple; we begin with an initial description of the positions of the protein and solvent atoms, and integrate Newton's equations of motion forward in time. In practice, MD is a complicated discipline beyond the scope of this work. The interested reader is referred to Allen and Tildesley's classic Computer Simulation of Liquids[1] for a general reference and the Amber simulation toolkit's manual for the specifics of the Amber14[2] software used to generate the data for this thesis.

For our purposes, the important facet of MD is that while the result is conceptually simple, a time-series of the atomic positions of each atom in the simulated system, interpreting and analyzing these results are far from simple. For much of the history of MD, anecdotal approaches have been common: running a few or even a single, short trajectory and inspecting it through visual analysis and measurement of a few chosen thermodynamic or reaction coordinate parameters. Increasing computer power has led to better and better sampling, and an increasing movement towards more statistical, arguably more scientific, approaches to analyzing the simulations. This is the view that we will take here, treating the simulation data as samples taken from a high-dimensional stochastic system, and thus amenable to modeling and analysis with tools from signal analysis and statistical learning.

[1] M. P. Allen/D. J. Tildesley: Computer Simulation of Liquids, 1989.

[2] D.A. Case et al.: Amber14, 2014, URL: `ambermd.org`.

Our goal is to build a simpler model of the dynamics of the CDR loops of the TCR than the simulation itself. Essentially, we aim to coarse-grain the dynamics of the system, but rather than coarse-graining the simulated model, we will coarse-grain the on the data to generate a new model that reproduces the major features of the underlying system, the original simulation, while being more amenable to human understanding. This strikes a balance between more classical statistical thermodynamic approaches, which make it possible to discuss the system broadly, but does not describe the local details we are interested in, and using ad-hoc metrics to answer specific questions about local phenomena or over-reliance on visual inspection which rely heavily on investigator intuition and interpretation.

We will use the Markov State Model formalism for building this reduced model, which will be a simple Markov model of the system that, ideally, reproduces the broad behaviors of the system. In practice, the amount of data required to build a good, quantitative Markov model of the CDR loops is beyond what we have available, and we thus use the Markov model to make qualitative, rather than quantitative, observations about the CDR loops of the 2C and NKT15 systems.

The first of our motivating questions - do there exist distinct, stable conformational states of the CDR loops - is a clustering question. We want to identify sets of conformations in the simulation data that are similar, for a certain meaning of similar, to one another and are distinct from conformations that belong to a different set. The meaning of similar is important here, and it motivates the use of the Markov state model formalism even in the absence of direct interest in the model itself. By similar, we mean kinetically similar, in the sense that two conformations are more kinetically similar the smaller the expected time for the molecule to transform between the two conformations. By building a Markov model of the system's dynamics, we will be able to cluster together states of the Markov model based on the probability of transformation between the states, yielding a clustering of conformations based on kinetic information.

Before we can cluster the data obtained from the simulation into states, we have to decide on a metric of closeness. Ultimately, we want to cluster using kinetic information, but the simulation data is made up of atomic positions, which leaves us only able to compute geometric values when comparing individual frames of data, which is the actual unit of information we have to work with.

There are two parts to dealing with this problem. First, we will use dimensionality reduction, in the form of time-structured independent component analysis (tICA), to transform the initial data into a better input space. In this transformed input space, we will use the euclidean metric as a measure of similarity and cluster the data frames into multiple clusters which we can build a Markov model on. This 'microstate' model will then provide the metric for clustering the data into the final 'macrostate' model, which yields clusters of microstates (and hence of initial data) based on the kinetics of the microstate model.

For the remainder of this chapter, we will inspect in detail each of these analytical tools and provide justification for their use and interpretation.

2.1 Toy System

It is instructive to analyze a toy system to understand the results of our analytic tools. Since the full molecular dynamics simulation we will analyze undergoes Langevin dynamics via the thermostat, we will use a single particle undergoing Brownian dynamics as an instructive toy system.

For a single particle, the equation of motion for Langevin dynamics is

$$m\frac{d^2x}{dt^2} = -\nabla U(x) - \gamma\frac{d}{dt}x + \sqrt{2\gamma k_B T m}\eta(t),$$

where m is the mass of the particle, ∇ is the gradient operator, $U(x)$ is the potential field and thus $-\nabla U(x)$ is the force acting on the particle, γ is a damping constant, k_B is Boltzmann's constant, T is the temperature, and $\eta(t)$ is a Gaussian process with zero mean and a delta kernel. In particular, $\eta(t)$ obeys

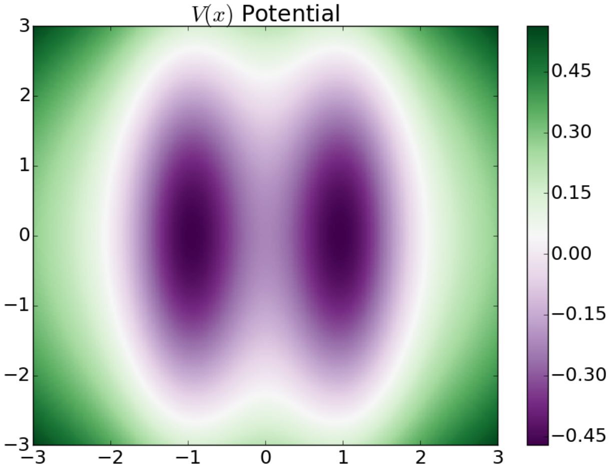

Figure 2.1: Potential energy surface of the $V(x, y)$ potential function.

- $\langle \eta(t) \rangle = 0$
- $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$

For our toy system, let $x \in \mathbb{R}^2$ and $m \to 0$. Let $\sigma^2 = \frac{k_B T}{\gamma}$ and re-arranging, we have

$$\frac{dx}{dt} = \frac{-\nabla U(x)}{\gamma} + \sqrt{2}\sigma\eta(t)$$

Choosing $\gamma = 1$ for simplicity, our toy model reduces to

$$\frac{dx}{dt} = -\nabla U(x) + \sqrt{2}\sigma\eta(t)$$

For the purposes of our toy model, σ is a free parameter of the model corresponding to the variance of the noise process. In a physical system, the variance corresponds to thermal noise that depends on temperature.

We will use two different potential systems to illustrate the coming projection techniques. Both will be anisoptropic double-well potentials, with only slight differences due to different

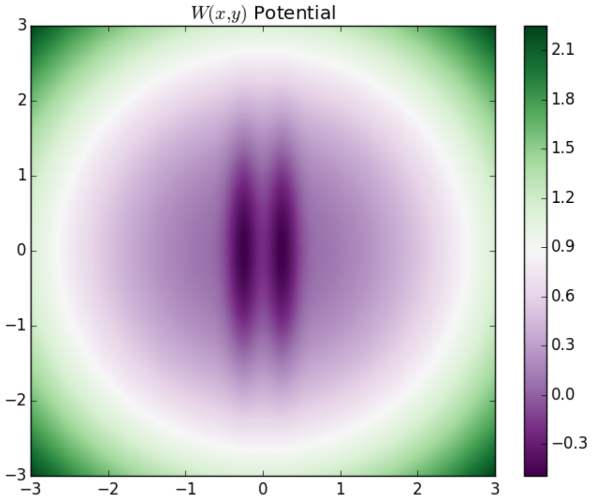

Figure 2.2: Potential energy surface of the $W(x, y)$ potential function.

parameterizations. Let U denote the general potential well, with form

$$U(x, y) = \frac{1}{2}\exp\left(-\beta(\kappa(x - x_0)^2 + y^2)\right) - \frac{1}{2}\exp\left(-\beta(\kappa(x - x_1)^2 + y^2)\right) + \alpha(x^2 + y^2),$$

α, β, and κ are parameters that control the shape of the potential, and x_0 and x_1 control the separation of the two wells on the x-axis. The last term is a harmonic potential for the purpose of constraining simulations to the region of interest.

We consider two potential fields of this form. Let $V(x, y)$ denote the U potential with parameterization $\alpha = \frac{1}{8}$, $\beta = \frac{1}{2}$, $\kappa = 16$, $x_0 = -1$, and $x_1 = 1$. The V potential is shown in Figure 2.1, as two elliptical potential wells with major axis along the y-axis and separated by a barrier along the x-axis.

The second potential is only slightly different. Denote the second potential by W, and parameterized by $\alpha = \frac{1}{8}$, $\beta = \frac{1}{2}$, $\kappa = 48$, $x_0 = -0.25$, and $x_1 = 0.25$. This potential is very nearly identical to V, with the potential wells elongated along the y-axis and brought closer together, yet still separated by a barrier along the x-axis.

Both of these potentials look like two-state systems, where the state of the system is just a matter of which potential well the particle is currently in at a given time. What we actually want to do is determine the degrees of freedom that separate these different states, so that we can use those degrees of freedom as simpler representations of the system both for direct analysis and as a reduced dimensional space to feed into further analysis techniques.

2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a classic technique commonly used as a first tool of choice when exploring new data and as a dimensionality reduction technique for complex or high-dimensional data. First, we briefly review this technique, and then demonstrate the short-comings in the context of our toy models.

PCA is an orthogonal linear transformation that transforms the data to a new basis such that the first basis element captures the maximal amount of variation in the data, and each subsequent basis element captures the maximal amount of remaining variation. PCA can be used as a dimensionality reduction method by using the first few basis elements as a projection matrix, so that a lower-dimensional view of the data can be obtained that captures the maximum variation of the data in the lower-dimensional view.

2.2.1 Derivation of PCA

There are many derivations of PCA available in the literature and there is little novelty we can add[3]. However, because the derivation is instructive to compare with the later derivation of tICA, we sketch a quick derivation here.

Consider an n-by-m data matrix X consisting of random data comprising n samples of m variables. In our toy model, we have an n-by-2 data matrix consisting of n samples of the position of the system as it evolves through the two-dimensional phase space.

Without loss of generality, we assume that the data matrix has column-mean 0. Otherwise we replace the matrix with the mean centered data matrix. We wish to find an orthogonal

[3] I.T. Jolliffe: Principal Component Analysis, 2002.

basis for the data matrix such that successive basis elements maximally capture the variance of data. Let $w_1 \in \mathbb{R}^n$ be the first such basis element. Then we want

$$w_1 = \underset{||w_1||=1}{argmax}\, Var(w^T X)$$

Since w_1 is a vector, we have that $w_1^T X$ is a linear combination of elements of X, that is, a linear combination of random variables. It follows that

$$Var(w_1^T X) = w_1^T X^T X w_1$$

Combined with the fact that we require $||w_1|| = 1$, we have

$$w_1 = \underset{||w_1||=1}{argmax}\, \frac{w_1^T X^T X w_1}{w_1^T w_1}$$

The right hand side has the form of a Rayleigh Quotient, and $X^T X$ is symmetric, so it follows that the maximizer of the right hand side is the maximal eigenvalue[4], λ_1, of $X^T X$. Thus we conclude that first principal component is exactly the first eigenvector of $X^T X$ and accounts for variance in the data proportional to the first eigenvalue.

The second principal component can be found by repeating the procedure on the new data matrix

$$\widetilde{X} = X - X w_1 w_1^T,$$

which is the original data set after removing the data corresponding to the first principal component. Repeating the previous procedure extracts the maximal eigenvalue of $\widetilde{X}^T \widetilde{X}$ and the corresponding eigenvector as the second principal component. However, this eigenvector is just the second largest eigenvector of the original $X^T X$ matrix.

In general, to find the kth principal component, we need to find the maximal eigenvalue

[4] Philippe Blanchard/Erwin Brüning: Mathematical Methods in Physics, vol. 69 (Progress in Mathematical Physics), 2015.

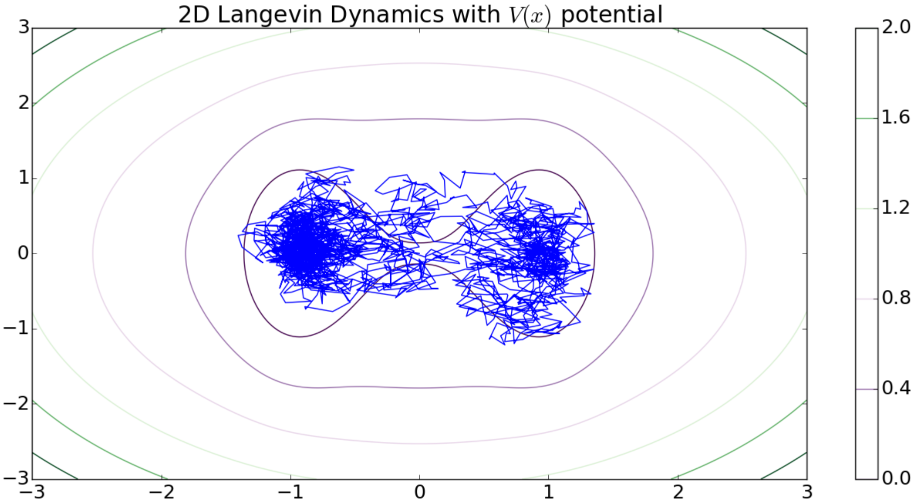

Figure 2.3: Simulated trajectory of a zero-mass particule undergoing 2D Langevin dynamics on the $V(x, y)$ potential energy surface.

of

$$\widetilde{X}_k = X - \sum_{i=1}^{k-1} X w_i w_i^T,$$

for which the maximal eigenvalue is the kth largest eigenvalue of X, and it follows that the principal components are the eigenvectors of $X^T X$, the correlation matrix of the data.

Note that the left singular vectors of X are the left eigenvectors of $X^T X$, so it is more efficient in practice to simply compute the Singular Value Decomposition of the data matrix X, which is the standard method employed in most software packages that perform PCA.

2.2.2 Applying Principal Component Analysis

To apply PCA to our toy system, we start by running a simulation of the toy model system. For this exercise, we initialize the system at the origin, and simulate for 100,000 steps. For details of simulating the toy system, see Appendix A.

We first simulate the $V(x, y)$ potential system. Figure 3 shows the trajectory of a simulation plotted against the potential energy surface. The simulation behaves as expected.

Starting from the origin, the trajectory falls into one of the wells and randomly moves around in that well, eventually the random fluctuations drive the trajectory out over the transition region into the other well. From a large scale view, this is what proteins do, too; they explore a local energy well, occasionally thermal fluctuations drive them into a significantly different conformation with a different local energy well. In our toy system, the two energy wells are separated by the line $x = 0$, and the degree of freedom along the separation is the $\mathbf{e}_x$, i.e. the x-axis. If our major concern is with transitions between the two wells, then we are primarily interested in the motion in the x dimension, and the motion in the y dimension is noise. This means we can project our data onto the $\mathbf{e}_x$ vector and analyze the system in 1-D rather than 2-D. Applying PCA yields the eigenvalue-eigenvector pairs

$$\lambda_0 = 0.582, v_0 = \begin{bmatrix} -0.99997 \\ -0.00741 \end{bmatrix}$$

$$\lambda_1 = 0.198, v_1 = \begin{bmatrix} 0.0074 \\ -0.99997 \end{bmatrix}$$

where the eigenvalues describe the variance captured along the corresponding eigenvector. The eigenvectors are, up to sign, almost exactly the axes of the coordinate system, with most of the variance along the x-axis. In this case, PCA accurately captures the degree of freedom - the x-axis - that separates the energy wells.

Next, consider the $W(x, y)$ potential system. Figure 2.4 shows the trajectory of a simulation plotted against the potential energy surface, where we see similar behavior to the $V(x, y)$ potential system. However, when applying PCA, we get different results.

$$\lambda_0 = 0.086, v_0 = \begin{bmatrix} 0.048 \\ 0.998 \end{bmatrix}$$

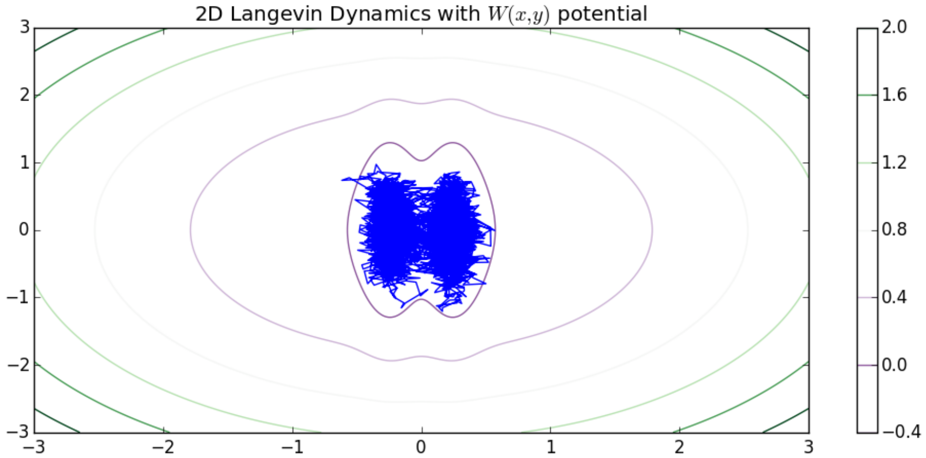

Figure 2.4: Simulated trajectory of a zero-mass particule undergoing 2D Langevin dynamics on the $W(x, y)$ potential energy surface.

$$\lambda_1 = 0.0619, v_1 = \begin{bmatrix} -0.998 \\ -0.048 \end{bmatrix}$$

The variance of the system is actually along the y-axis due to the elongation of the potential wells, but the degree of freedom we are interested in, if we want to classify the system into two states, is the location along the x-axis. Though artificial, this demonstrates the motivation behind looking to different techniques for dimensionality reduction.

2.3 Time-structured Independent Components Analysis

The major flaw of PCA for our purposes is that it optimally captures the wrong descriptive statistic. We are not interested in the maximal variance degrees of freedom; we are interested in degrees of freedom that separate locally stable conformations, that is, local energy wells. PCA accurate captures the separating degree of freedom for the $V(x, y)$ potential because the maximal variance degree of freedom happens to coincide with the degree of freedom that separates the two energy wells in the system. On the other hand, PCA fails on the $W(x, y)$ potential because the direction of maximal variance and the separating degree of freedom are distinct. The essential flaw of using PCA to study protein dynamics is that we are studying a proxy value - the variance in the data - in the hopes that it will find the degrees of freedom that separate energy wells. Intuitively, we can improve on this analysis if we can find a better proxy. Time-structures Independent Component Analysis (tICA) does precisely this, by looking for degrees of freedom that display maximal auto-correlation, rather than variance. This has the additional advantage of integrating the time component of the data, which PCA ignores in treating the data as independent draws from a distribution.

2.3.1 Independent Component Analysis

We present tICA here, which is derived from Independent Component Analysis, but shares some key differences from its originating method. Independent Component Analysis (ICA) is originally a method from the field of signal processing that attempts to linearly decompose a multivariate signal into independent non-Gaussian signals. Similar to k-means and k-medoids, ICA is properly thought of as a method with a specific outcome goal, rather than an algorithm, as there are multiple algorithms to accomplish the ICA decomposition.

Typically an ICA decomposition attempts to simultaneously minimize the mutual information of the components while maximizing the non-Gaussianity of the components. Despite the similar goal as PCA to decompose data and provide a new basis set, the problems the methods seek to solve are very different. PCA seeks an orthogonal basis set, and attempts to sequentially maximize the variance captured by each of the degrees of freedom it finds. If we fix the mean of a distribution at 0, which we can do for distributions on $\mathbb{R}^n$ by an affine transformation, then the distribution that is fully determined by its variance is the Gaussian distribution. Furthermore, when the variance is fixed, the maximal entropy distribution on $\mathbb{R}^n$ is the Gaussian. Thus, cast in a informational theoretic light, PCA finds a basis set of one-dimensional Gaussian distributions that describe the empirical data set under study. As the Gaussian is the maximal entropy distribution for fixed variance, we can argue that PCA is essentially only accounting for the variance in the data, and ignoring the information in higher-moments. On the other hand, ICA methods generally use either the kurtosis of the empirical data, or rely on mutual information (information entropy) based measurements. As stated however, ICA does not necessarily generate orthogonal components, though tICA does restrict to this condition.

2.3.2 Autocorrelation

Autocorrelation is a measure of how similar a signal or time-series is to itself shifted in time. Let $\mathbf{x}_t \in \mathbb{R}^n, t \in \mathbb{N}$ be a time-series. Under some weak assumptions on the time-series, the autocorrelation $A_\tau(\mathbf{x}_t)$ exists and is defined as

$$A_\tau(\mathbf{x}_t) = \frac{\mathbb{E}\left[(\mathbf{x}_t - \mu)^T(\mathbf{x}_{t+\tau} - \mu)\right]}{\sigma^2}$$

where τ is the time-lag of the autocorrelation we are measuring, μ is the mean and σ^2 is the variance of the time-series. Note that $\sigma = \mathbb{E}\left[(\mathbf{x}_t - \mu)^T(\mathbf{x}_t - \mu)\right]$, so $A_0(\mathbf{x}_t) = 1$.

Intuitively, autocorrelation is a better proxy statistic for finding stable sets of molecular conformations. If a Langevin system is in a local energy well, it will tend to stay near the minima of that well until thermal fluctuations force it out. Thus, the autocorrelation with a time-lag less than the average transition waiting time will be high when the system is in a local well. If we choose a linear degree of freedom of the system that maximizes the autocorrelation over the time-lag of interest, then the degree of freedom will separate regions where the system experiences high auto-correlation, that is, local energy minima. The argument that the tICA decomposition is superior to PCA for the purposes of finding conformational states of proteins rests on this concept.

2.3.3 Deriving tICA

In this section, we loosely follow the derivation of tICA presented in Schwantes and Pande, though written in the standard notation of linear algebra rather than Dirac bra-kets, but otherwise using the approach of Lagrange multipliers to transform the constrained optimization problem into a generalized eigenvalue problem. An older presentation of tICA as an eigenvalue problem can be found in Molgedey and Schuster, which most interestingly also presents tICA in the context of a recurrent neural network to compare with a neural net approach taken by Jutten and Herault to solve the blind source separation problem. This connection between neural networks and tICA is interesting in light of another line of research showing effective dimensionality reduction applied to molecular dynamics data using Autoencoder neural nets. This is explored more in Chapter 5, as a possible path to include non-linearity in the decomposition of MD data.

The goal of tICA is to find degrees of freedom with maximal autocorrelation that are orthogonal to one another. We can express this problem as a constrained optimization problem. Let $\mathbf{x}_t \in \mathbb{R}^{n}{}_{t=0}^{N-1}$ be an n-dimensional time-series, consisting of N data points. In the case of the protein system, each of these vectors in the time-series would correspond to frames of the protein simulation, or features derived from said frames. Without loss of generality, we assume the time-series has mean 0, and otherwise we can subtract the mean. As an aside, we note that this assumption implies that the time-series is generated by a stationary distribution, which is not necessarily the case, and leaves an opening for further investigation.

Our objective function is given by the autocorrelation function

$$f(\mathbf{v}) = \frac{\mathbb{E}\left[(\mathbf{v}^T\mathbf{x}_t)(\mathbf{v}^T\mathbf{x}_{t+\tau})\right]}{\mathbb{E}\left[(\mathbf{v}^T\mathbf{x}_t)(\mathbf{v}^T\mathbf{x}_t)\right]}$$

where $\mathbf{v} \in \mathbb{R}^n$. The inner product term $\mathbf{v}^T\mathbf{x}_t$ is the projection of the data vector onto the vector $\mathbf{v}$, which plays the role of a test basis element, hence we are calculating the autocorrelation of the one dimensional time-series of the data projected onto a potential basis function. The inner product is symmetric over the field of real numbers, so we have that $v^T x_t = x_t^T v$, and can re-write the objective function as

$$f(\mathbf{v}) = \frac{\mathbb{E}\left[\mathbf{v}^T \mathbf{x}_t \mathbf{x}_{t+\tau}^T \mathbf{v}\right]}{\mathbb{E}\left[\mathbf{v}^T \mathbf{x}_t \mathbf{x}_t^T \mathbf{v}\right]}$$

The expectation of the outer product $\mathbf{x}_t \mathbf{x}_t^T$ is recognizable as the covariance matrix of the data, and similarly $\mathbf{x}_t \mathbf{x}_{t+\tau}^T$ is the time-lag correlation matrix, so we can define

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{x}_t \mathbf{x}_t^T\right]$$

$$\mathbf{C}^{(\tau)} = \mathbb{E}\left[\mathbf{x}_t \mathbf{x}_{t+\tau}^T\right]$$

Assume $\boldsymbol{\Sigma} > 0$, which is true with probability one for random data vectors. The quadratic form $\mathbf{v}^T \mathbf{x}_t \mathbf{x}_t^T \mathbf{v}$ commutes with the expectation operator, which we can see as

$$\mathbb{E}\left[\mathbf{v}^T \mathbf{x}_t \mathbf{x}_t^T \mathbf{v}\right] = \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{v}^T \mathbf{x}_t \mathbf{x}_t^T \mathbf{v} = \mathbf{v}^T \left(\frac{1}{N} \sum_{t=0}^{N-1} \mathbf{x}_t \mathbf{x}_t^T \right) \mathbf{v} = \mathbf{v}^T \mathbb{E}\left[\mathbf{x}_t \mathbf{x}_t^T\right] \mathbf{v} = \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}$$

An identical argument applies to yield $\mathbb{E}\left[\mathbf{v}^T \mathbf{x}_t \mathbf{x}_{t+\tau}^T \mathbf{v}\right] = \mathbf{v}^T \mathbf{C}^{(\tau)} \mathbf{v}$. With this, we can rewrite the objective function in a more expressive form as

$$f(\mathbf{v}) = \frac{\mathbf{v}^T \mathbf{C}^{(\tau)} \mathbf{v}}{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}$$

Finally, in order to find solutions to our optimization problem, we need to constrain the solution space in some fashion so that a solution exists. In the setting of PCA, the constraint is that $\mathbf{v}^T \mathbf{v} = 1$, that is, the solutions live on the the sphere S^n. This provides an orthonormal basis for the principal component vectors, but is a poor constraint for our purposes since it means that although our choice of objective function means we optimize for autocorrelation, our solution is weighted by the variance of that degree of freedom. This is fine for PCA, which is explicitly attempting to capture variance, but worse for our purposes, as we want to measure distance between points in our space where, ideally, the distance is strongly related to kinetic similarity, locally defined by our autocorrelation objective function.

The choice of constraint is actually a free parameter of the algorithm and of interest in that we can derive a family of related algorithms by choosing different constraints. Following Schwantes and Pande, as the analysis presented in Chapter 4 assumes, we choose the constraint that our solution vector yields has unit variance, thus the problem of finding the first component is expressed as

$$\max_{\mathbf{v}} \quad f(\mathbf{v}) = \max_{\mathbf{v}} \quad \mathbf{v}^T \mathbf{C}^{(\tau)} \mathbf{v}$$

with the constraint:

$$\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = 1$$

At this point, we diverge slightly from the derivation presented in Schwantes and Pande, and note that similar to the PCA case presented earlier, we can recognize the objective function as a Rayleigh quotient, or rather, as a generalized Rayleight quotient,

$$R(\mathbf{C}^{(\tau)}, \mathbf{\Sigma}, \mathbf{v}) = \frac{\mathbf{v}^T \mathbf{C}^{(\tau)} \mathbf{v}}{\mathbf{v}^T \mathbf{\Sigma} \mathbf{v}}$$

We can rewrite this generalized Rayleigh quotient as a standard Rayleigh quotient if the Cholesky decomposition of the operator in the denomenator exists, which in this case over the field of Real numbers, requires that $\mathbf{\Sigma}$ be symmetric and positive-definite, which the correlation matrix is by construction. Let $\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^T = \mathbf{\Sigma}$ be the Cholesky decomposition of $\mathbf{\Sigma}$, and let $\widetilde{\mathbf{C}}^{(\tau)} = \widetilde{\mathbf{\Sigma}}^{-1} \mathbf{C}^{(\tau)} \widetilde{\mathbf{\Sigma}}^{-1T}$, and $\widetilde{\mathbf{v}} = \widetilde{\mathbf{\Sigma}}^T \mathbf{v}$. Then we can rewrite the expression as a standard Rayleight quotient as

$$R(\widetilde{\mathbf{C}}^{(\tau)}, \widetilde{\mathbf{v}}) = \frac{\widetilde{\mathbf{v}}^T \widetilde{\mathbf{C}}^{(\tau)} \widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^T \widetilde{\mathbf{v}}}$$

In the limit of infinite data, and using the assumption that the data matrix is sampled from a time-reversible system, we have that $\mathbf{C}^{(\tau)}$ is symmetric. Note that if $\mathbf{A}, \mathbf{B} \in \mathbb{R}^n$, with $\mathbf{B}$ a symmetric matrix, then $\left(\mathbf{A}\mathbf{B}\mathbf{A}^T\right)^T = \mathbf{A}\mathbf{B}^T\mathbf{A}^T = \mathbf{A}\mathbf{B}\mathbf{A}^T$, so it follows that $\widetilde{\mathbf{C}}^{(\tau)}$ is symmetric. We invoke the result that the Rayleight quotient is maximized by the maximal eigenvector of the operator[5], and find that our first IC is the left eigenvector of $\widetilde{\mathbf{C}}^{(\tau)}$. At this point, we can follow the same reasoning as presented in the sketch of the proof of PCA above, and we have that the independent components are the solutions to the eigenvalue problem

$$\widetilde{\mathbf{C}}^{(\tau)}\tilde{\mathbf{v}} = \lambda\tilde{\mathbf{v}}$$

Expanding this expression, we have that

$$\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{C}^{(\tau)}\widetilde{\mathbf{\Sigma}}^{-1T}\tilde{\mathbf{v}} = \lambda\tilde{\mathbf{v}}$$

Rearranging and using the fact that $\tilde{\mathbf{v}} = \widetilde{\mathbf{\Sigma}}^{T}\mathbf{v}$, we have that

$$\mathbf{C}^{(\tau)}\mathbf{v} = \lambda\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{\Sigma}}^{T}\mathbf{v} = \lambda\mathbf{\Sigma}\mathbf{v}$$

Thus we conclude that the independent components are the solution to the generalized eigenvalue problem relating the time-lag covariance matrix to the correlation matrix. To the author's knowledge, this particular proof that the generalized eigenvalue problem is the solution to the tICA problem is novel, though it should be noted that the Rayleight quotient result cited here relies on the method of Lagrangian multipliers, which is the method used by Schwantes and Pande.

2.4 Clustering

Clustering, or cluster analysis, is a fundamental task in data mining, machine learning, and, arguably, science itself. Given m items, clustering is the act of grouping the items into n different groups. We are interested in the simplest case of "hard" clustering, where each data point belongs to exactly one cluster and the clusters all live in the same space.

[5] Blanchard/Brüning: Mathematical Methods in Physics (see n. 4).

Clustering often goes hand-in-hand with dimensionality reduction, using the dimensionality reduction both as a way to combat the curse of dimensionality and as a feature extraction pre-processing step.

From the point of view of protein physics, we want to cluster together structural conformations that are 'similar', breaking the simulation data into multiple clusters of similar conformations that are distinct from other clusters. A key problem here is deciding what we mean by similar. From a thermodynamic point of view, similar conformations are those which are energetically close, in particular, those that can interconvert rapidly. Alternatively, if we think of the clusters directly, we may consider two conformations to be in the same cluster if they are part of the same local energy well. Both views accurately capture what we think of as crystallographic-like 'states' of a protein, though they deal less well with transition states. In our real problem, we are primarily interested in long-lived crystal-like states, so we'll take this to be an acceptable view for the problem at hand.

Ultimately, we will cluster the TCR simulation data into a few, large clusters using the Markov model. However, we need an initial clustering algorithm that can operate directly on the simulation data, which will be the focus of the remainder of this chapter: using the toy systems to explore two of the basic geometric clustering algorithms that will be used to find the microstate model clusters.

2.4.1 k-means

K-means is one of the oldest clustering methods, originally described in papers at Bell Labs in the 1950s. Given a data set $(x_0, x_1, ..., x_n)$, where $x_i \in \mathbb{R}^n$, and a parameter $k \in \mathbb{N}$ - the number of clusters - k-means finds sets $\{S_1, S_2, ..., S_k\}$ such that every data point belongs to exactly one set and the sets minimize the ℓ_2 distance to the cluster mean. We can give this as the problem of finding $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ such that

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||_2^2,$$

where

$$\mu_i = \frac{1}{|S_i|}\sum_{x_j \in S_i} x_j$$

This problem turns out to be quite hard. K-means is known to be NP-hard[6] for arbitrary dimension n when k is fixed to be 2, as well as NP-hard when $n \geq 2$ is fixed, but k is not. In general, K-means is solved approximately by heuristic algorithms, most commonly Lloyd's algorithm, which is detailed below. After clustering, we can classify a new data point by assigning the new point to the cluster that minimizes the distance from the cluster center to the data point. In effect, this means that K-means shatters the space into a Voronoi partition.

Lloyd's Algorithm

Lloyd's algorithm is a form of the expectation-maximization heuristic[7]. The algorithm takes an initial set of k means, $\{m_i\}_{i=1}^{k}$; the initial set can be chosen in several ways, but a common choice in standard k-means is to pick k random data points to be the initial means. After initialization, the algorithm alternates between an assignment and update step.

The **Assignment step** consists of assigning each point to exactly one set that minimizes the within cluster sum-of-squares as defined above. Since the distance is the standard Euclidean metric, the minimizing assignment is to assign each point to the nearest mean.

The **Update step** simply consists of calculating a new mean for each set, with the mean defined by μ above. For each S_i, the sum of squared distance over the data points in S_i from the new mean must be less than this sum relative to the previous mean, as the arithmetic mean is the least-squares estimator.

The algorithm halts when the assignment step doesn't change the assignment of any data point, at which point the update steps become idempotent. Since in both steps the

[6] D. Aloise et al.: NP-hardness of Euclidean sum-of-squares clustering, in: Machine Learning 2009.

[7] S.P. Lloyd: Least squares quantization in PCM, in: IEEE Transactions on Information Theory 1982.

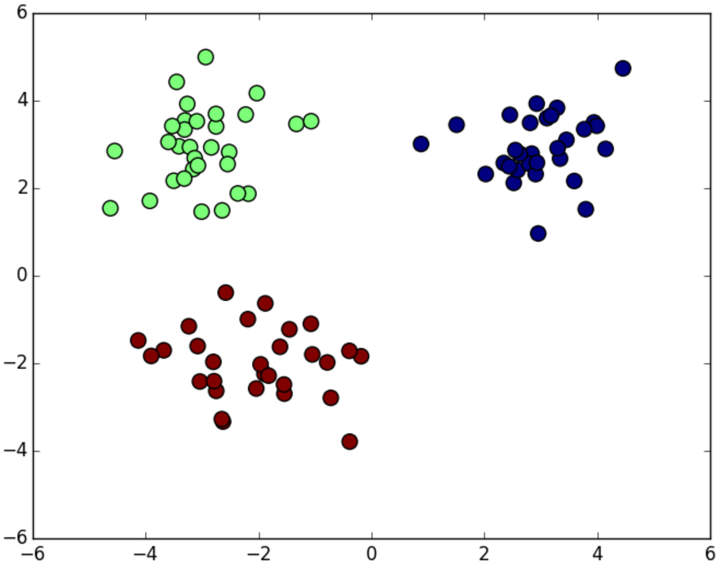

Figure 2.5: k-means clustering of well separated data drawn from three gaussian clusters with $k = 3$.

new within cluster sum-of-squares is bounded above by the current within cluster sum-of-squares, this procedure is monotonic in the cost function. Along with the fact that there are only finitely many partitions of the data points into k sets, the algorithm always halts at a local minima.

However, there is no guarentee that the global optimum is found by this procedure. In practice, Lloyd's algorithm is quite fast, so satisfactory solutions are found by re-running the algorithm with different random initial assignments for the means and keeping the best solution found.

k-means in practice

K-means is an effective algorithm for automatically clustering data when the clusters are well-separated. This is observable in figure 2.5, where the Gaussian distributions that generate the data are well separated, and so the data points are effectively associated to the right generating distribution region. Additionally, by coloring the space according to cluster

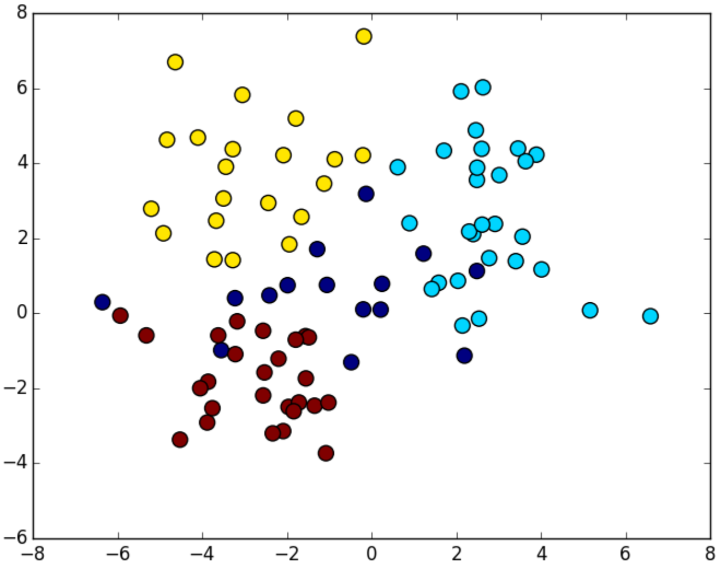

Figure 2.6: k-means clustering of poorly separated data drawn from three gaussian clusters with $k = 3$. Dark blue points are misclassified.

assignment, we can see the Voronoi partitioning of the space. However, poor separation of the data leading to poor clustering is an inherent problem of clustering when the algorithm shatters the space in a Voronoi partition. This is observable in figure 2.6, where the classification of the points begins to fail when the data sets from disparate distributions overlap. There are two common solutions to this problem, one is to find a transformation of the space that better separates disparate data points, and the other is to use a 'soft' clustering algorithm, which assigns probabilities of cluster assigments rather than strict binary assignment. K-medoids, presented next, also suffers from this problem. For this analysis, we attempt to find a projection of the data that well-separates the clusters - this is the essential role of tICA - and generally linear transformations that improve separation of data take the form of projections. However, some data cannot be linearly separated at all, and in these cases k-means and k-medoids will fail completely. For these cases, non-linear transformations are necessary. We describe some attempts to apply this approach to analyzing the TCR data in Chapter 5.

2.4.2 k-medoids

The k-medoids algorithm is related to the k-means algorithm, and is the initial clustering method we use for analyzing the protein simulation data in practice, after applying the tICA decomposition to project the data onto a lower dimensional space. Like k-means, k-medoids partitions the dataset into k sets $\{S_i\}_{i=1}^{k}$ such that each data point belongs to exactly one S_i. By inserting new data points using the same method as k-means, that is, assigning new data points to the set that minimizes the distance to the set center, k-medoids also partitions the space into a Voronoi partition. The primary difference between k-means and k-medoids is that in k-medoids, the set centers are restricted to be elements of the data set rather than arbitrary points in the data space. Unlike k-means, the restriction of k-medoids that the set center must be a data point allows for a broader range of distance functions, rather than just the standard Euclidean distance[8]. This is because with the restriction of the set centers to data points, for any given k, there are only finitely many possible choices of set centers and partitions.

In particular, a k-medoids algorithm attempts to find the sets S_k such that for all $x_i \in \mathbf{X}$, $x_i \in S_k$ for exactly one S_k, and the following cost function is minimized

$$\sum_{k=1}^{K} \sum_{x_i \in S_k} d(x_i, m_k)$$

where d is a distance metric. Note that if we choose d to be the Euclidean metric, then k-medoids solves the k-means problem with the restriction of the cluster centroids to be data points from the data set, rather than allowing the centroids to be arbitrary points of the ambient metric space.

[8] L. Kaufman/P.J. Rousseeuw: Clustering by means of Medoids, in: Statistical Data Analysis Based on the L_1-Norm and Related Methods (Statistics for Industry and Technology), 1987.

Partition around Medoids algorithm

Partitioning Around Medoids (PAM) is the most common algorithm for a k-medoids clustering, so we describe it here as with Lloyd's algorithm to get a feel for the specific operations that create a k-medoid clustering[9]. As with the k-means algorithm, we have a data set $\mathbf{X}$ and an investigator determined parameter k, the number of clusters the algorithm should output. As with k-means, we optimize the cost function defined by summing over the distances of each data point to their assigned medoid.

The algorithm begins with the **initialization** step wherein k data points are drawn at random from $\mathbf{X}$ without replacement. The k data points will be the initial set of medoids. Each data point is then assigned to same cluster as the nearest medoid.

Next the algorithm iterates the **update** step, and halts when an update does not decrease the cost function. The update step consists of: For each medoid m and data point o such that o is not a medoid:

1. Make o a medoid in place of m.
2. Compute the cost function.
3. If the cost decreases, keep the new configuration, otherwise leave o as a data point and m as the medoid.

Compared to the k-means algorithm, it is clear that the PAM algorithm is significantly more computationally expensive, as it has a quadratic complexity as presented. The major advantage of k-medoids over k-means is the ability to use a distance metric other than the standard Euclidean metric, where k-means is not guarenteed to update monotonically, and the restriction to using only data points. This restriction to only data points is useful in the case of clustering molecular dynamics data as it allows us to use a specific, realizable structural conformation as the center of the cluster, while a centroid of a k-means cluster may not even be physically realizable.

[9] Kaufman/Rousseeuw: Clustering by means of Medoids (see n. 8).

2.5 Conclusions: On the Curse of Dimensionality

The curse of dimensionality is a term that has been mentioned several times before. As a major challenge in data analysis, and one that shows up strongly in understanding protein dynamics, it deserves a brief comment before we complete this chapter.

The curse of dimensionality is a common phrase that refers to several different but related phenomena that occur in various computational disciplines, numerical analysis, statistical sampling and inference, combinatorics, machine learning, and others. The occurence in Numerical Analysis and Statistical Inference are probably the most familiar, and have to do with the super-linear scaling in the number of samples needed to produce an accurate result as the dimensionality of the problem increases. In the finite element method in numerical analysis, this takes the form of an exponential increase in the number of grid points that need to be evaluated on a mesh as the dimension increases. If, for example, we needed to evaluate a function only on the vertices of the unit hypercube, then we require four evaluations in two dimensions, but eight in three dimensions, and generally we require 2^n evaluations in n dimensions. This exponential scaling results in even simple problems quickly becoming intractable at high-dimension if the algorithm cannot scale with the dimension better than exponentially. A similar phenomena occurs in statistical inference and machine learning where the amount of data required for training quickly becomes incredibly large when the dimensionality of the problem becomes large. In protein dynamics, the curse of dimensionality takes the form of the sampling problem, and indeed, we will see the limits of sampling in analyzing the TCR simulation data in chapter 4.

The curse of dimensionality also has form in the simple problem of evaluating distance functions. The problem, essentially, is that in general the difference between ‘near’ and ‘far’ points in high dimensions becomes vanishingly small, described in more detail below.

For our purposes, this makes clustering a difficult task in high dimensional settings, such as in protein dynamics where the dimensionality of the problem is, naively, 3^n for n atoms in the simulation. Even ignoring solvent and considering only the protein, this yields a naive data dimensionality that quickly runs into the hundreds when considering small parts of proteins, and into the thousands or tens of thousands when considering even relatively small proteins.

There is a third, related, issue that is highly present in molecular dynamics data, which is irrelevant data obscuring relevant data. In MD, this has the form of thermal noise. Thermal fluctuations are necessary for protein activity, indeed we are essentially studying the effects of thermal noise on a potential energy system. In the absense of thermal noise, MD reduces to little more than gradient descent on a potential energy surface, and the dynamics of the system quite literally freeze out. However, thermal noise is also a distraction when studying the data, as not every dimension is relevant, but every dimension is continuously perturbed by thermal effects during the simulation. We are interested in studying only a subset of the degrees of freedom of the system, those that separate long-lived local energy minima - metastable states. However, most of the degrees of freedom are simply the thermal motion of atoms bouncing back and forth, with little or no long term consequences for the system as a whole. These extra, irrelevant dimensions add significant difficulty to the problem of clustering the data, above and beyond the noise in the dimensions we care about, because the distance-measurement problem means that as more noise dimensions are added to the system, the clusters become close and closer together under our distance metric, merely due to noise dimensions. Formally[10], this has the form that given a fixed distribution ρ on $\mathbb{R}$, there is an induced product distribution $\rho^{(n)}$ on $\mathbb{R}$. Let X_n denote a data vector drawn from $\rho^{(n)}$, and D_{max} and D_{min} be the maximum and minimum distances between data points in a set drawn from the distribution. Then we have that

$$\frac{D_{max} - D_{min}}{D_{min}} \to 0$$

[10] Kevin Beyer et al.: When Is ”Nearest Neighbor” Meaningful?, in: vol. 1540 (Lecture Notes in Computer Science), 1999, pp. 217–235.

under the assumption that

$$\lim_{n \to \infty} \mathrm{var}\left(\frac{||X_n||}{\mathbb{E}[||X_n||]}\right) = 0$$

This assumptions holds for a broad range of distributions and distance measures, including L_p-norms with $p \geq 1$.

Much of the analysis machinery presented and used here is about dealing with the twin distance-measuring and noise dimensions problems.

The dimensionality reduction technology, tICA for this analysis, is explictly about dealing with these problems by stripping away the excess dimensions, but the Markov models presented in the next chapter are also approaches to dimensionality reduction and 'empirical coarse-graining' that attempt to reduce the dimension of the simulation data down to where we can extract understanding. Ultimately, we use them in tandem, as the Markov models require clustering, which benefits from first passing the data through tICA to reduce the dimensionality for clustering.

CHAPTER 3
MARKOV STATE MODELS

Markov state models (MSMs) are discrete, kinetic models of protein dynamics, based on the theory of Markov models. MSMs act as a sort of coarse-graining of the system dynamics, however, unlike techniques usually referred to as coarse-graining in MD, MSMs work by building discrete models from simulated data, essentially coarse-graining the empirically observed data, rather than analytical models of the system. MSMs thus serve several purposes. The primary function of MSMs is to allow simulations to reach greater timescales than are directly accessible to molecular simulation. This is done by constructing the MSM from simulated data, and analyzing the long time-scale properties of the MSM. While this method is incapable of revealing new behavior that is not observed in the simulations, it yields a statistical understanding of what is observed, and makes it possible to understand long timescale behavior of what is observed, often this is sufficient.

In concert with this, by building a statistical model from many events, MSMs are able to integrate data from many simulations into a unified statistical model[1]. Since we analyze the statistical model for longer timescale behavior, the statistical model can describe long-term behavior in a rigorous manner that is not fully captured by any one simulation. The practical advantage is that we can run multiple simulations in parallel. As the parallel simulations are completely independent, non-communicating processes, we neatly side-step Amdahl's law[2] and obtain a linear speedup, at least in the simple case.

[1] Vijay S. Pande/Kyle Beauchamp/Gregory R. Bowman: Everything you wanted to know about Markov State Models but were afraid to ask, in: Methods 52 (1 2010), pp. 99–105.

[2] Gene M. Amdahl: Validity of the single processor approach to achieving large scale computing capabilities, in: AFIPS spring joint computer conference, 1967.

Additionally, a simultaneous advantage and drawback of MD simulation is the immense amount of data produced. The ultimate goal of the scientific process is human understanding, and no matter how much information there may be in a simulation, it is worthless unless we can extract that information into a form that can be understood and manipulated by the intuition and reason of the researcher. Two approachs to this problem was presented in the last chapter. The first, dimensionality reduction, ideally, finds the most 'information-rich' dimensions of the data, and projects onto those for interpretation. The other approach, clustering, groups similar items together, making it possible to understand the system in terms of those groups rather than taking each item, here a single frame of the simulation, as an individual. This second approach is powerful for intuition. Intuitively, a good clustering of the protein's conformations would be a 'state' of the system, where each state behaves similarly and experiences thermal fluctuations around a local energy minima. This yields an understanding of the state as very much like a collection of x-ray crystal structures. While not entirely correct, and we will ultimately want to consider the statistics of states, rather than making direct observations as in crystallography, this view is intuitive and familiar for discussion.

So if clustering is effective, why Markov state models? The problem with clustering, as presented in the previous chapter, is that it clusters on the wrong metric. In comparing data frames for clustering using k-means or k-medoids, we cluster using a similarity metric that is inevitably geometric in nature. A better approach would be to cluster using kinetic information rather than geometric, after all, what we are interested in when we describe states is collections of conformations that are nearby in the sense that they rapidly interconvert, or alternatively, that they belong to the same or nearby energy wells up to some resolution of the energy surface.

MSMs are comprised of two aspects, a discrete state space to which individual conformations of the protein are assigned, and a transition matrix that describes the kinetics of transformations between those states. We are interested in both of these aspects, however for the purposes of studying the T cell receptor, we shall primarily be interested in the process of building the discrete state space of conformational states. In doing so, we will cluster the conformations into states, not by geometric criteria, but by kinetic, and thus we can cluster the conformations without inadvertently lumping kinetically disparate conformations together and thus accidentially eliminating kinetic barriers of the energy diagram from our model. We will thus show the existence of metastable states in the T cell receptor invisible to classical x-ray crystallography.

In order to fully describe the MSM formalism, we first review the basic theory of Markov chains and Markov models. As in the previous chapter, the goal of this chapter is to introduce the reader to the analytical tools at hand, however here we rely on much more theoretical underpinnings than the previous set of algorithms. As such, we shall take a short detour through the mathematics of Markov chains before we introduce the algorithmic methods. As before, ideas are demonstrated with concrete code and figures, in addition to the theoretical results. For the non-technical reader who only needs an intuitive grasp of MSMs for reading chapter 4, the review article by Pande et al. is an excellent resource.[3]

3.1 Markov Chains

A Markov process is a stochastic process which obeys the Markov property. When the state space of the process is discrete and Markovian, it is often referred to as a Markov chain, a simple, but powerful model for many phenomena and a useful computational tool in many settings.

The Markov property, named for Andrei Markov, is simply the property that the stochastic system has no dependence on its past, but only on its current state. This is an intuitive concept, and in the deterministic setting of Newtonian mechanics it is implicitly assumed: the future behavior of a mechanical system from a point in time depends only on its current

[3] Pande/Beauchamp/Bowman: Everything you wanted to know about Markov State Models but were afraid to ask (see n. 1).

momentum and position. This is not to say that the past is irrelevant in a complete sense, the past behavior of the system is what brought it to the current moment, however, a system that is Markovian can be fully described by its current state, and thus its future can be described, either deterministically or probabilistically, without knowing how it got to its current state, only that it did, indeed, get there.

In physical systems, non-Markovian behavior is often surprising, as we have come to expect physical systems to behave in a Newtonian fashion. The classic example of history-dependence, hysteresis, is the magnetization of ferromagnetic metals, where a prior alignment of the domains causes a permant magnetic field to arise, independent of the external field applied to it. In this case, the system is history dependent in that the response to an external field is dependent on prior exposure to the field through the hysteresis mechanism. However, if we consider the system's state to include information about the alignment of the domains in the absence of a field, then it again becomes Markovian, in that we have all the information required in the current state to describe the future behavior without reference to the past. Indeed, any non-Markovian system with a dependence on only a finite length of it's history can be described as a Markovian system in a higher dimensional space where that finite window of the system's history is included in the 'current' state. Thus, Markovian systems are in fact incredibly common, and serve as excellent models for a wide variety of phenomena.

To formalize this idea[4], let Ω be a finite or countably infinite set, which we will call the **state space** of the system. A **Markov chain** is a sequence of random variables $X_0, X_1, X_2, ...$ such that $\forall i \in \mathbb{N}, X_i \in \Omega$ which has the property that

$$\mathbb{P}(X_{t+1} = i | X_t, X_{t-1}, ..., X_0) = \mathbb{P}(X_{t+1} = i | X_t)$$

That is, the Markov property is that the probability of the next random variable, drawn

[4] Daniel W. Strook: An Introduction to Markov Processes, vol. 230 (Graduate Texts in Mathematics), 2005.

from the state space, is conditionally independent of the history of previous random variables given the most recent random variable. As a trivial example, the Bernoulli process, which models repeated flips of a (fair or non-fair) coin, is Markovian. For a fair coin, the probability that the next coin flip results in a heads is 50%, independent of the results of previous flips, and also independent of the whether the last flip resulted in a heads or tails. More generally, any stochastic process that is a sequence of independent draws from a probability distribution is Markovian in a trivial sense, but these are not usually the processes of interest, rather, we want processes where the future depends on the present in a fundamental way, without depending on the entire history of the process.

3.2 Markov Models

Markov models refer to a number of stochastic processes that obey the Markov property and are used to model phenomena. Probably the two most common are the Markov chain and the Hidden Markov model, both of which describe autonomous Markov processes, with the major difference being whether we can directly see the state of the Markov process, as in a Markov chain where the data describe the state at each time step, or if the data describe a phenomena controlled by the state of the Markov chain, but the state is not directly observable. In the latter case we have a hidden Markov model, where we posit the existence of a Markov chain that describes our system of interest, but where our data consists of random variables that are draws from some probability distibution that depends on the unobserved state of the hidden Markov chain. Hidden Markov models have shown promising results for the problem we are interested in[5], identifying and modeling the state of a protein, however, we will focus on the simpler Markov chain model in which the state is visible and we can inspect the sequence of states directly, though we wish to note that we are aware of the advances made in applying HMMs to protein dynamics and believe

[5] R.T. McGibbon et al.: Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models, in: Proc. 31st Intl. Conf. on Machine Learning, 2014.

they will be an excellent tool once the methodology has matured more. In the case of protein dynamics, the difference between these two approaches essentially boils down to the issue of determining the assignment of frames to states as an independent step that precedes and possibly alternates with determining the Markov model. Building the model is greatly simplified by separating these concerns, however in doing so we are making an independence assumption about the dynamics between the states and the definition of the states themselves, introduced algorithmically by the use of k-means, k-medoids, or other clustering processes. As is often the case with machine learning techniques, the investigator must make the decision about the trade-offs inherit to any choice of one technique over another.

In all of the following, we will consider Markov models with only a finite number of states. This simplifies and makes more concrete the mathematics, essentially reducing the mechanics of the theory to a subset of finite dimensional Linear Algebra and is the only case we are interested in from a practical perspective of modeling protein dynamics in the Markov State model framework. Let us consider a simple example system.

Let $\Omega = \{A, B\}$ be the state space. Define a stochastic process $\{X_i\}, i \in \mathbb{Z}^+$ by conditional draws from a distribution given by

$$\mathbb{P}\left(X_{i+1} = A | X_i = A\right) = p_{AA}$$
$$\mathbb{P}\left(X_{i+1} = B | X_i = A\right) = p_{AB}$$
$$\mathbb{P}\left(X_{i+1} = A | X_i = B\right) = p_{BA}$$
$$\mathbb{P}\left(X_{i+1} = B | X_i = B\right) = p_{BB}$$

where $p_{AA}, p_{AB}, p_{BA}, p_{BB} \in [0, 1]$ and

$$p_{AA} + p_{AB} = 1$$
$$p_{BA} + p_{BB} = 1$$

The probability p_{XY} can be understood as the probability of transitioning to state Y at the next time step, given that the system is currently in state X. Self transitions, that is probabilities of the form p_{XX} are the probability that the system stays in the current state. We can arrange the probabilities into a **transition matrix** as

$$\begin{bmatrix} p_{AA} & p_{AB} \\ p_{BA} & p_{BB} \end{bmatrix}$$

Noting that the rows sum to one, the transition is a **right stochastic** matrix, which has the effect of taking probability vectors, non-negative vectors with L_1 norm 1, back to probability vectors under right multiplication. Specifically, let $\boldsymbol{\pi}_0$ be a probability vector. Then

$$\boldsymbol{\pi}_0\mathbf{P} = \boldsymbol{\pi}_1,$$

where $\boldsymbol{\pi}_1$ is a probability vector. Further, by induction we have that

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\mathbf{P} = \boldsymbol{\pi}_0\mathbf{P}^n,$$

where $\boldsymbol{\pi}_n$ is a probability vector. In general, the product of right stochastic matrices is again a right stochastic matrix, and as we would expect from the above, the n-th power of a right stochastic matrix is a right stochastic matrix. These same results hold for both left and doubly stochastic matrices, where left stochastic matrices are those whose columns sum to 1, and doubly stochastic matrices are both left and right stochastic.

In going from describing the stochastic process in terms of draws from a conditional distribution to the matrix notation, we have transferred from the point of view a particular trajectory or realization of the stochastic process to considering the ensemble, and in particular by studying the properties of the transition matrix, we can describe average and long-term behavior of an ensemble, much as we might in statistical mechanics. In particular, when we calculate $\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n\mathbf{P}$, we take the probability distribution of the emsemble that currently has the distribution $\boldsymbol{\pi}_n$ to the new distribution $\boldsymbol{\pi}_{n+1}$. Thus the transition matrix is the operator that moves a probability distribution of the Markov process forward in time. The immediate question is then, does a limit exist and is it unique? That is, does there exist $\boldsymbol{\pi}^*$ such that

$$\boldsymbol{\pi}^* = \lim_{n\to\infty} \boldsymbol{\pi}_0 \mathbf{P}^n$$

If so, how does it depend on $\boldsymbol{\pi}_0$? For the moment, posit that such a $\boldsymbol{\pi}^*$ does exist and is unique. Then we have that

$$\boldsymbol{\pi}^* = \lim_{n\to\infty} \boldsymbol{\pi}_0 \mathbf{P}^{n-1}\mathbf{P} = \boldsymbol{\pi}^*\mathbf{P}$$

And thus, if it exists, $\boldsymbol{\pi}^*$ is an eigenvector of $\mathbf{P}$ with eigenvalue 1. Such a probability vector is a **stationary distribution** of $\mathbf{P}$. If the system described by $\mathbf{P}$ is a physical thermodynamic system, then $\boldsymbol{\pi}^*$ corresponds to the equilibrium distribution of thermodynamic states in the system.

Under the right conditions, the stationary distribution exists and is unique, and thus independent of the initial probability distribution of the system. In the next section, we discuss the Perron-Frobenius theorem from spectral theory, which when applied in the context of a stochastic transition matrix yields the appropriate stationary distribution. Before moving on, however, let us consider some failure modes where this may not hold. Given the above two-state system, let the conditional probability distribution be given by

$$p_{AA} = p_{BB} = 1$$

The transition matrix is then the identity matrix, and we have that every probability vector is an eigenvector with eigenvalue 1, and more specifically, is stationary. While a trivial example, it is clear that non-negative entries and row sums of 1 are not sufficient to ensure the uniqueness of stationary distributions. More generally, we have that given a transition matrix $\mathbf{P}$, if there exists a permutation matrix $\mathbf{A}$ such that $\mathbf{AP}$ is a block diagonal matrix, then there is not a unique stationary distribution. An intutive way to see this is to consider the graph of the Markov model that a block diagonal transition matrix describes. Such a graph would be disconnected, and though each connected subgraph may describe a Markov model that has a proper stationary distribution, there can be no such distribution for the total Markov model because the system is disconnected. In particular, any convex combination of stationary distributions of the subgraphs is a stationary distribution of the overall system. Thus, we are interested in Markov models described by strongly connected graphs, or alternatively **irreducible** matrices, those which are not similar via a permutation to a block upper triangular matrice.

In practice, we can deal with disconnected systems quite easily, as the disconnectedness implies that the connected subgraphs of the system are independent of each other, and can be analyzed individually. More problematic in the case of studying a protein dynamics system is that disconnected graphs can result from insufficient sampling rather than actual theoretical independance. This is usually addressed either by dropping disconnected subgraphs, or, when possible, running more and/or longer simulations to increase the available data.

More troubling than disconnected systems are weakly connected systems, which can and do arise in empirical studies, where the graph is connected, but there exist pairs of states i, j such that $\forall n \in \mathbb{N}, \mathbf{P}_{ij}^n = 0$. In other words, state j is unreachable from i regardless of the number of steps into the future the system progresses. This cannot happen in an irreducible matrix, and as a stochastic matrix is non-negative, an irreducible stochastic matrix can be characterized by the existence of an $n \in \mathbb{N}$ such that $\mathbf{P}^n$ has all strictly positive values.

For another failure mode, consider the transition matrix defined by $p_{AB} = p_{BA} = 1$, that is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The process defined by this transition matrix is again Markovian, the matrix is stochastic and a stationary distribution exists and is unique. It is clear that the only stationary distribution is the vector $\boldsymbol{\pi}^* = [0.50.5]$, as the transition matrix simply swaps the two entries of the probability vector at each time step. However, while the stationary distribution exists and is unique, we see that $\forall \boldsymbol{\pi} \neq \boldsymbol{\pi}^*$ such that $\boldsymbol{\pi}$ is a probability vector, and $\forall n \in \mathbb{N}$,

$$||\boldsymbol{\pi}\mathbf{P}^{n+1} - \boldsymbol{\pi}\mathbf{P}^{n}|| > 0$$

Hence the sequence $\boldsymbol{\pi}, \boldsymbol{\pi}\mathbf{P}, \boldsymbol{\pi}\mathbf{P}, ...$ is not Cauchy convergent. Thus although the stationary distribution exists, it is not the limit point of any orbit of probability vectors other than itself, and in fact, any distribution other than the stationary distribution is part of a two-element closed orbit set. Thus, non-stationary distributions do not relax toward the stationary distribution, which in thermodynamic terms means that the ensemble does not equilibrate. Intuitively, such a transition matrix cannot model a thermodynamic ensemble as it would not relax toward the maximum entropy distribution. More generally, this is the requirement that the Markov chain be **aperiodic**. Aperiodicity of a matrix can be directly characterized number theoretically, but for the case of irreducible stochastic matrices, it suffices that the trace be non-zero, and so in practice, we will not encounter aperiodic Markov systems calculated from empirical molecular dynamics data.

3.2.1 Perron-Frobenius

The Perron-Frobenius theorem is the basis of, essentially, a branch of spectral theory. In the simplest form, the theorem dates back to work by Oskar Perron in 1907 on square matrices with positive values, later extended by Georg Frobenius in 1912 to a subset of non-negative matrices. This section will state the theorem and some consequences of it for stochastic matrices, but a proof is omitted; for a modern proof using spectral theory, see the article by Smyth[6]. The theorem is spectral in the sense that it describes properties of eigenvalues and eigenvectors of real positive/non-negative matrices, and in this concrete

[6] M.R.F. Smyth: A Spectral Theoretic Proof of Perron-Frobenius, in: Mathematical Proceedings of the Royal Irish Academy 2002.

form finds many applications in probability theory (of interest to the present work), as well as dynamical systems theory and numerous applications, perhaps most notably as an aspect of the PageRank algorithm, which is based on Markov chain theory.

The Krein-Rutman theorem generalizes the Perron-Frobenius theorem to infinite dimensional Banach spaces. Although not the subject of the present work, the value here is that this extends the Perron-Frobenius theorem to the general theory of transfer operators. The thermodynamics of a protein system can be described in full generality by the transfer operator that moves an initial ensemble of conformations towards equilibrium, and an excellent analysis of this approach is described by Prinz et al[7]. Though we will not delve further into this subject at present, it is worth noting that the Markov state model formalism is essentially a numerical approximation to calculating the spectrum of the transfer operator of the protein dynamical system that moves probability mass toward the equilibrium distribution. The transfer operator formalism can thus be used to describe how effective an approximation is, and by similarly casting the tICA decomposition into the same setting, it turns out that both tICA and MSMs are numerical approximations to the transfer operator spectrum, yielding a theoretical reason for the effectiveness of tICA in preprocessing data for analysis with the MSM approach.

Returning to the subject at hand, the Perron-Frobenius theorem asserts several properties about the eigenvectors and eigenvalues of a square, positive matrix. Without loss of generality, assume that the transfer matrix of our Markov model has all positive values. Since we will work, theoretically, only with aperiodic, irreducible transfer matrices we have that if the transfer matrix $\mathbf{T}$ is not positive, then there exists an $n \in \mathbb{N}$ such that $\mathbf{T}^n$ has all positive values.

Let $\mathbf{T}$ be a square matrix with positive real entries over $\mathbb{C}$. Then there exists a positive real number r such that r is a simple eigenvalue of $\mathbf{T}$, and for all $\lambda \neq r$ that are eigenvalues

[7] J.-H. Prinz et al.: Markov models of molecular kinetics: Generation and validation, in: J. Chem. Phys. 134 (2011).

of $\mathbf{T}$, $|\lambda| < r$. Furthermore, the eigenvector v corresponding to r has all positive real values, and for any other eigenvalue w of $\mathbf{T}$ such that w is not a positive multiple of v, then the entries of w include at least one negative or complex value.

Unpacking this a bit, theorem tells us that there exists a unique maximal eigenvalue and that the corresponding eigenvector is strictly positive, and furthermore, no other eigenvector has all positive real entries. We will ultimately be concerned with symmetric transition matrices, for which the eigenvector basis can always consist of only real valued vectors. The interpretation of this, then, is that the eigenvector v corresponding to the eigenvalue r is, under normalization to length 1, the stationary distribution of the Markov process, and the other eigenvectors describe the degrees of freedom along which the system relaxes toward equilibrium, with the corresponding eigenvectors describing the timescale of the relaxation modes. Eigenvectors other than the stationary distribution have both negative and positive values because they describe flows of probability mass through the system, while a strictly positive or strictly negative valued eigenvalue would correspond to a source or sink of probability mass over time, which should not happen. The stationary distribution is all positive as it describes the distribution at equilibrium, and furthermore, if it is in fact the stationary distribution, then $r = 1$, and as we will see later, the corresponding timescale is infinitely long, as we would expect of a thermodynamic system.

It remains to show that $r = 1$. However, since $\mathbf{T}$ is a stochastic matrix, it maps probability vectors to probability vectors. Let $\boldsymbol{\pi}^* = \frac{\mathbf{v}}{||\mathbf{v}||_1}$. Then

$$||\boldsymbol{\pi}^*\mathbf{T}||_1 = ||r\boldsymbol{\pi}^*||_1 = ||\boldsymbol{\pi}^*||_1$$

It follows that $r = 1$.

Frobenius generalized these results to the case case of irreducible non-negative matrices, and in the general study of Markov models this is a highly useful tool. However, we are currently only interested in the stationary distribution, for which we have that some finite power of $\mathbf{T}$ has all positive values, and furthermore we have that the stationary distribution $\boldsymbol{\pi}$ obeys

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{T}$$

and so also obeys

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{T}^n$$

So $\boldsymbol{\pi}$ is an eigenvector with eigenvalue 1 for both $\mathbf{T}$ and $\mathbf{T}^n$. The utility of this is that the general case of the Perron-Frobenius theorem only guarentees that $|\lambda| \leq r$, and there exist up to h eigenvalues, with $h-1$ taking negative or complex values, of maximal absolute value, where h is the period of the matrix. From this, it follows that if we require that our transition matrix be aperiodic, then $h = 1$, and there is a single eigenvalue of maximal absolute value among the spectrum of the transition matrix. This justifies the restriction that a Markov model of a thermodynamic system be aperiodic.

3.2.2 Detailed Balance

The principle of detailed balance is a fundamental principle of chemical kinetics and thermodynamics that states the at equilibrium, each elementary process of a chemical system is at equilibrium with its reverse process. In the case of a Markov chain, this requires that the Markov process be reversible, which can be expressed as

$$\boldsymbol{\pi}_i^*\mathbf{T}_{ij} = \boldsymbol{\pi}_j^*\mathbf{T}_{ji}$$

Note that this does not imply that given two states $A, B \in \Omega$ that $p_{AB} = p_{BA}$, which would say that the forward and reverse probabilities of a particular state change are the same, rather, the ensemble does not experience a net probability flow one state to another at equilibrium. In kinetic terms, this is the idea that at equilibrium, the net rates of reaction are balanced, where the net rate is the rate of a reaction times the concentration. We can write this as a first order rate reaction,

$$k_{A \to B} [A] = k_{B \to A} [B]$$

which is the standard statement of equilibrium for a reversible first order chemical reaction.

Kolmogorov's criterion is necessary and sufficient for a transition matrix to obey detailed balance[8]. Kolmogorov's criterion states that every finite closed cycle of states has the same product probability as the reverse cycle. Formally, for any finite sequence of states $\{s_i\}_{i=1}^{n}$,

$$p_{s_1 s_2} p_{s_2 s_3} \cdots p_{s_{n-1} s_n} = p_{s_n s_{n-1}} p_{s_{n-1} s_{n-2}} \cdots p_{s_2 s_1}$$

Clearly, if a transition matrix models a chemical process, it must obey detailed balance.

3.2.3 Maximum Likelihood Estimation

So far we have studied the properties of a Markov model in relationship to the transition matrix that defines the model. This is vital to understanding the nature of the Markov model and to analyzing and intepreting a model once written down. This is a fine state of affairs to stop at if we are only analyzing models written down *ab initio*, but we are interested in empirical models built from data.

Let Ω be the discrete, finite state space of the system of interest. Our data set is a finite sequence of draws from this state space, $\{X_t : X_t \in \Omega\}_{t=1}^{N}$. The problem is to determine the transition matrix $\mathbf{T}$ that maximizes the probability of observing the $\{X_t\}$ sequence.

This is an instance of the general problem of statistical inference, given some data set $\{x_t \in \Omega\}$ that is drawn from $p(x|\theta)$, the probability density function p parameterized by θ. Then the maximum-likelihood approach to estimating θ is to solve the optimization problem

[8] F.P. Kelly: Reversibility and Stochastic Networks, 1979.

given by

$$\max_{\theta} \mathcal{L}(\theta; x_1, x_2, ..., x_n) = \max_{\theta} p(x_1, x_2, ..., x_n|\theta)$$

where $f(x_1, x_2, ..., x_n|\theta)$ is the joint probability distribution of the data set given the parameter set θ. In the simple case where the data are drawn independent and identically distributed, the joint probability distribution factors as

$$p(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta),$$

and in this case the maximum likelihood estimation of θ is

$$\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta)$$

This is a straightforward optimization problem, and for some families of distributions p, there exists a closed-form solution. Much more discussion of the method of MLE can be found in numerous textbooks on statistical inference, and much ink has been spilled on the topic since it was popularized in 1912 by Fisher. It is worthwhile to note that MLE is not the only method of statistical inference, however a deeper consideration of the topic is outside the current scope, as we will assume that MLE is sufficient and appropriate to the task at hand.

Our particular problem is not quite so simple, as the draws are not independent and identically distributed since at each step the next draw from the Markov process, if it is of any interest as a Markov process, depends on the previous result. So the probability distribution does not immediately factor as a product.

3.3 Markov State Models

The Markov state model formalism is an approach to modeling the dynamics of a protein using Markov models and Markov chain theory[9]. MSMs are constructed from molecular dynamics data, making them a form of ad-hoc coarse grained models built empirically from directly sampling the system of interest, rather than attempting to fit parameters of an analytic coarse-grained model.

The advantage is that the Markov model is more likely to capture the actual dynamics of the system, and can be tuned to be coarser or finer in a systematic manner so that finer models can be used to predict experimental results while coarser models lend themselves to better human intuition and understanding of the major aspects of dynamics, while maintaining a link to the finer models for potential verification. The models themselves can be simulated exceedingly efficiently, however, as the models are, by construction, irreducible aperiodic transition matrices, we may employ the full theory of Markov chains to study them analytically. The Perron-Frobenius theorem immediately yields the equilibrium distribution of the model, and further eigenvalues and eigenvectors show the relaxation degrees of freedom, and the timescales of these computed quantities can be several orders of magnitude longer than the simulated data used to generate the model, effectively stretching the data.

The downside of the MSM approach is that a large amount of data is nonetheless required to estimate transition matrices, and this can be quite costly compared to simulating a coarse-grained system derived analytically with fitted parameters.

3.3.1 Microstate Model Construction

Markov state models are built in two stages: first a ‘microstate’ model is built, and then a ’macrostate’ model is constructed from an analysis of the microstate model. The microstate model is build by directly clustering the data frames into small clusters, which does not

[9] F. Noé/S. Fischer: Transition networks for modeling the kinetics of conformational change in macromolecules, in: Current Opinion in Structural Biology 2008.

achieve the goal of kinetically clustering the data, but rather allows the kinetically-clustered macrostate model to be bootstrapped from the microstate model.

A major motivation in the construction of Markov state models is clustering data frames of a simulation. In Chapter 2, we studied two common clustering algorithms, k-means and k-medoids, which cluster together data points on the basis of a similarity metric, often (only, in the case of k-means) the Euclidean distance between points. Of non-hierarchical clustering methods, the more sophisticated methods generally take the form of a transformation of the space followed by k-means of k-medoids, for example spectral clustering which uses the spectrum of the similar matrix of the data to perform dimensionality reduction before clustering, similar to the process described here where tICA is used to perform dimensionality reduction before clustering with k-medoids. The clustering problem with respect to molecular dynamics simulations is that the similarity metric must respect the kinetics of the system, otherwise if we cluster two data points point (simulation frames) together that are kinetically separated, we have accidentally removed an energy barrier from the output model. The fundamental flaw of clustering using standard geometric criteria, the RMSD between conformations, is that RMSD can easily hide kinetic barriers, such as sterically hindered ϕ/ψ angle movements that separate two local energy wells. Small ϕ/ψ angle changes may separate two local energy wells of the conformational space, but clustering on the RMSD of the frames that live near the transition may show small RMSD changes as the angles may not strongly alter the α carbon positions, but still represent a large kinetic barrier. If the clustering accidentally links these frames together, the kinetic barrier is lost and the model underestimates transitions between states, or else combines disparate states together.

We avoid this problem in two ways. The more recent approach takes a page from the more sophisticated clustering methods and pre-processes the data with a transform and projection of the data space[10]; here that method is tICA, which is itself a linear approximation to the

[10] C.R. Schwantes/V.S. Pande: Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9, in: J. Chem. Theory Comput. 9 (2013), pp. 2000–2009; R.T. McGibbon/V.S. Pande: Learning Kinetic Distance Metrics for Markov State Models of Protein Conforma-

spectrum of the transfer operator[11], making even the microstate model a clustering using kinetic rather than geometric data. The second method, used in concert with tICA or other data pre-processing, is to build a model of many small clusters, with the goal that each cluster is sufficiently small that the kinetic similarity of the frames inside the cluster is very high. Using a large number of clusters is often quite important when clustering on solely geometric criteria, i.e. RMSD distance between frames, but less so when using an effective pre-processing method. The value here is that with fewer clusters, we have more samples per cluster, which will improve the statistics of model estimation. Make no mistake, though the Markov state model method can extract significantly information from the data, it is still an immensely data hungry analysis.

With all of the tools built up in Chapter 2 and the last few sections, constructing a microstate model turns out to be quite simple. For clarity, we describe the process only for the case of a single long trajectory, but construction of a model from multiple trajectories is not much more involved than the single trajectory case and is well described in the literature.

Let the intial data set be a sequence of N frames $\{f_t\}_{t=1}^{N}$ generated by a molecular dynamics simulation. Each frame is taken to be a vector describing the molecular structure of interest; this may be the direct Cartesian positions of the atoms in XYZ space, a vector of ϕ/ψ angles (suitably transformed by sin and cos to account for periodicty) or another representation. Cartesian coordinate and other representations that have an external frame of reference must be aligned to remove irrelevant center of mass drift and molecular tumbling. This is the stage at which data pre-processing is used, so that we run the data representation through tICA or another pre-processing method as desired by the investigator. The data, transformed or in its original form, is clustered using k-means or k-medoids. The clustering metric is a free parameter of the analysis, though is usually dependent on the data representation. Using the XYZ coordinates would naturally lead to using the pairwise RMSD or

tional Dynamics, in: J. Chem. Theory Comput. 2013.

[11] M. Sarich/J-H Prinz/C. Schütte: Markov Model Theory, in: (An Introduction to Markov State Models and Their Applications to Long Timescale Molecular Simulation), 2014.

mass-weighted RMSD as similarity metrics, while the standard Euclidean metric generally fits the sin/cos representation of the dihedral angles. More exotic choices are possible, the Hamming distance is fitting for a binary contact map representation of the protein, which can be effective in the context of protein folding. In the case of pre-processed data, the Euclidean metric is standard when the output space is $\mathbb{R}^n$, as we are effectively clustering using the induced metric of the pre-processing technique.

The clusters generated by the clustering algorithm become the state space of the microstate model, and the data is transformed to a sequence of states; each data frame in the trajectory is mapped to one of the clusters, and our transformed data is the sequence of cluster indices. At this point, the data is a sequence of draws from a finite state space, so we can estimate the transition matrix using MLE as described in the previous section.

3.3.2 Time Lag

There is a major flaw in the MSM construction process laid out in the previous section– the transition matrix estimated from the procedure as written assumes that the data sequence is drawn from an ergodic system, and that assignment of a frame to a particular cluster-state implies that the system is in a local equilibrium for that cluster-state. Essentially, there is some error introduced by discretizing the system into clusters, and the discretized system may not be Markovian on the discretized time-scale, that is, looking at the system one frame at a time may violate the Markov property. The solution is to build the model at a longer timescale.

The process of building the models at longer time lags is straightforward. To build an MSM with timelag τ instead of counting transitions of the sequence $s_1, s_2, s_3, ...$, we instead count transitions from the sequence $s_1, s_{\tau+1}, s_{2\tau+1}, ...$ So we run the same MLE analysis on the sequence generated by subsampling the original data sequence at a rate τ. When dealing with MD data, τ is usually expressed in terms of simulation time rather than frame number.

This results in an important trade-off – the longer the time-lag, the less descriptive the model becomes, but the more accurately it models a process. Practically, longer time lags also require more data, build a model at timelag τ, expressed in frames, from N frames of data, we only have $\left\lfloor \frac{N}{\tau} \right\rfloor$ entries in the trajectory sequence, reducing the data by a factor of τ. This can be ameliorated by using a sliding window rather than subsampling, reducing the data by an additive factor of τ rather than multiplicative, but then transition counts are no longer independent, introducing an error into the estimated transition matrix.

This implies a need for a test to determine the appropriate timelag to build a model at. Several tests have been explored in the literature[12]. The most common and the one used in Chapter 4 is to look for convergence of implied timescales.

3.3.3 Implied Timescales

The implied timescales of an MSM correspond to the relaxation rates of the degrees of freedom of the MSM. It is important to note that the degrees of freedom of the MSM do not correspond to the physical system's degrees of freedom, such as those found by tICA. The degrees of freedom of the MSM are probability fluxes between collections of states that system undergoes as it relaxes toward equilibrium, that is, the eigenvectors of the transition matrix. It is unsurprising then, that the implied timescales are proportional to the eigenvalues. However, rather than directly studying the eigenvalues of the transition matrix, we calculate the implied timescales to link the relaxation modes of the system to real time in terms of either simulated lab time or frames. The i-th implied timescale is given by

$$t_i = -\frac{\tau}{\log(\lambda_i)}$$

where τ is the timelag of the MSM and λ_i is the i-th eigenvalue of the transition matrix. Note that $\lambda_0 = 1$ by the Perron-Frobenius theorem, so $t_0 = -\frac{\tau}{\log(1)} = \infty$, corresponding to equilibrium occuring 'at infinity'.

[12] S. Park/V.S. Pande: Validation of Markov state models using Shannon's entropy, in: J. Chem. Phys. 2006; Prinz et al.: Markov models of molecular kinetics: Generation and validation (see n. 7).

The simplest method for determining an MSM time-lag is building MSMs of the same system at a sequence of time-lag values and checking for convergence of the slowest time scales, corresponding to the first few non-trivial eigenvalues. We only need the first few timescales to converge because from a theoretical standpoint, we are largely interested in the slowest motions of the system, and from a practical consideration, when we build the macrostate system with n states, it will only maintain the information from the slowest $n-1$ degrees of freedom.

3.3.4 Macrostate Model Construction

The final macrostate MSM is useful for intuitive understanding of the system dynamics and more coarse-grained analysis, such as we are interested in when analyzing the TCR dynamics for alternative conformational states. Microstate models, or macrostate models with more states are effective for predictive calculations, i.e. NMR or EPR parameters.

The passage from microstate model to macrostate model is essentially another clustering process; this time the data frames are not clustered directly but rather the microstate clusters are clustered together to generate clusters-of-clusters that show maximal meta-stability, that is, we want the probability of leaving the macrostate clusters to be low. Ultimately, this corresponds to maximizing the trace of the transition matrix of the clustered graph, carried out using the PCCA+ algorithm, which groups nodes of the microstate graph together into larger stable nodes, using the eigenstructure of the transition matrix[13]. Once new clusters are assigned, the final macrostate transition matrix is estimated directly from the data, using the same lag-time analysis and maximum likelihood estimation of parameters used to estimate the microstate model from data.

The major analysis choices fall to the choice of lag-time, which follows the same convergence procedure as microstate model building, and determining the number of clusters

[13] P. Deuflhard/M. Weber: Robust Perron cluster analysis in conformational dynamics, in: Linear Algebra and its Applications 2005.

to build the macrostate model with. There is no obvious choice for the number of clusters, and it is partially a trade-off between interpretability and detail. However, a reasonable approach and one taken in the analysis of the TCR data is to choose one more cluster than there are slow timescales in the microstate model that separate out from the other implied timescales at convergence. A markov model with n states has $n-1$ degrees of freedom, and implied timescales correspond to the slow degrees of freedom. Hence, if we are interested in the slowest three degrees of freedom, it is sensible to model the macrostate model with four states to incorporate these degrees of freedom.

CHAPTER 4
RESULTS ON THE T CELL RECEPTOR

Having completed our tour of the mathematical constructs underpinning the analysis, we come to the results on the T cell receptor. We aim to shed light on the nature of the CDR loop flexibility and movements. We have two specific aims in this regard. First, we want to determine how structured the motions of the loops are, can they be well described by a low-dimensional system? And are these motions structured in some fashion, displaying clustering and pathways? Second, do there exist metastable states of the system? This second question ties into the binding hypotheses spectrum described in Chapter 1; are there metastable states in solution that could be states for an equilibrium selection mechanism, or seeds of states that support the conformational melding hypothesis?

With these aims in mind, we have simulated 10 independent MD trajectories of the 2C TCR for 300 ns each, for a total of 3 μs of data. For a comparison system, we have also simulated 10 independent 100 ns trajectories of the NKT15 system, a Class I NK T cell that recognizes α-GalCer, for a total of 1 μs of data. Technical details of the simulations and analysis can be found in Appendix C.

4.1 tICA Analysis shows distinct conformations and low dimensional motion

We studied the conformational changes of the CDR3α and CDR3β loops individually by analyzing their backbone dihedral angles under the tICA decomposition. As described in Chapter 2, the tICA algorithm can be intuitively understood as taking a dataset and a timescale parameter chosen by the investigator as inputs, and returning a set of combinations of the input degrees of freedom, here sets of dihedral angles, that are independent of one another and display long-lived behavior.

We applied the tICA decomposition to the phi and psi angles of CDR3α and CDR3β independently, each with an autocorrelation time of 5ns. Considering only the first two degrees of freedom resulting from this analysis, we see a local maxima of the probability distribution for the CDR3α loop (figure 4.1A), while there are four regions with local maxima for the CDR3β loop (figure 4.1B). These islands of locally high probability are long-lived regions of conformaitonal space that are frequently visited by the simulation, suggesting that these conformations are relatively stable, and indicating the existence of stable conformational states.

An outstanding question drawn from crystallographic studies asks how free are the motions of the CDR3 loops? Are they weakly structured with a large number of degrees of freedom to move in, or are they tightly choreographed, moving in distinct conformational states? To address these questions, and confirm the value of our two dimensional distributions, we consider the probability distributions of the first eight tICA degrees of freedom. CDR3α shows an assymetric distribution in the first and third tICA degrees of freedom, and a highly peaked distribution in the second tICA degree of freedom centered away from zero (figure 4.1C). The remaining tICA degrees of freedom are more Gaussian with means near zero, suggesting that CDR3α has some mild internal structure to its motions, with at most only the first three tICA degrees of freedom capturing interesting behavior. The system appears to be well described by two degrees of freedom. CDR3β shows significantly more interesting behavior in it's first two tICA degrees of freedom, both of which show multiple peaks, while the remaining degrees of freedom show much more Gaussian-like appearences (figure 4.1D). This strongly suggests CDR3β's motions are primarily captured by the first few, and in particular the first two, tICA degrees of freedom, indicating highly structured motions and a largely two dimensional phase space of non-thermal noise motions.

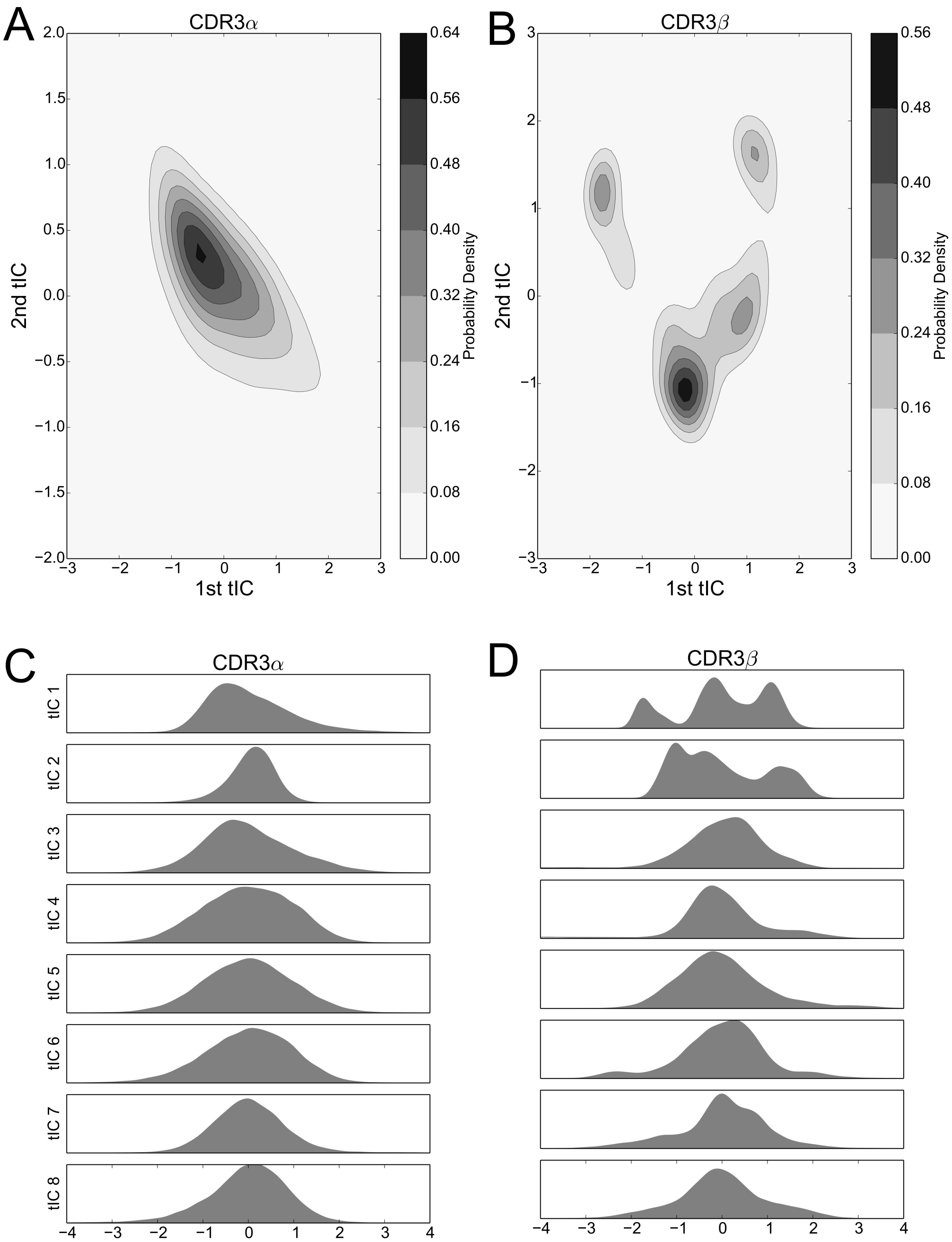

Figure 4.1: Kernel density estimates of tICA projections. (A,B) 2-D Kernel density estimate of the simulation data projected onto the first two degrees of freedom discovered by tICA for the CDR3α and CDR3β loops, respectively, using a 2-D Gaussian kernel. The KDEs estimate the probability density function for finding a randomly selected frame in a region of conformational space described by the tICA degrees of freedom. (C,D) 1-D probability density graphs of the first eight tICA degrees of freedom for CDR3α and CDR3β, respectively, using a Gaussian kernel.

4.2 Markov state model of CDR3β shows discrete metastable states

Next, we clustered the frames in the tICA projection via k-medoids into a microstate model and estimate MSMs as described in chapter three. The microstate models of CDR3α do not converge over the timescales analyzed, and the trajectory data is insufficient to go to longer timescales for MSM construction. This implies that CDR3α has a very slow degree of freedom that is not sufficiently explored in the simulation. We address this further in a later section with a reverse simulation, and note that very slow timescales of CDR3α compared to CDR3β has been described in simulations of A6, where a single trajectory of several simulated trajectories showed a major conformational change of the CDR3α loop. As discussed later, this is indicative of a large kinetic barrier between a bound-like conformation and the current unbound-like conformation of CDR3α.

On the other hand, the three slowest timescales of the CDR3β models separate out from the faster timescales when the implied timescales converge (figure 4.2B). The CDR3β data samples the phase space of CDR3β sufficently to build a qualitative MSM of CDR3β dynamics.

The separation of three slow timescales of the CDR3β loop implies a four state macro model of CDR3Eβ dynamics, and agrees with the four high-probability islands observed in the 2D projection of the data under the tICA analysis (figure 4.1B). We construct a four state model of CDR3β and extract the centroids of the macrostate clusters; these are the orientations that are in centers of the clusters drawn from data frames, so the centroids are conformations observed in the simulations, not mathematical averages. The centroids are shown in figure 4.3A. All four states are well populated at equilibrium (figure 4.3B), as predicted from the macrostate MSM, with the fourth state showing the highest equilibrium population. Interestingly, state four is poorly populated in the empirically observed data. The discrepency is due to the equilibrium distribution being calculated from eigenvalue anal-

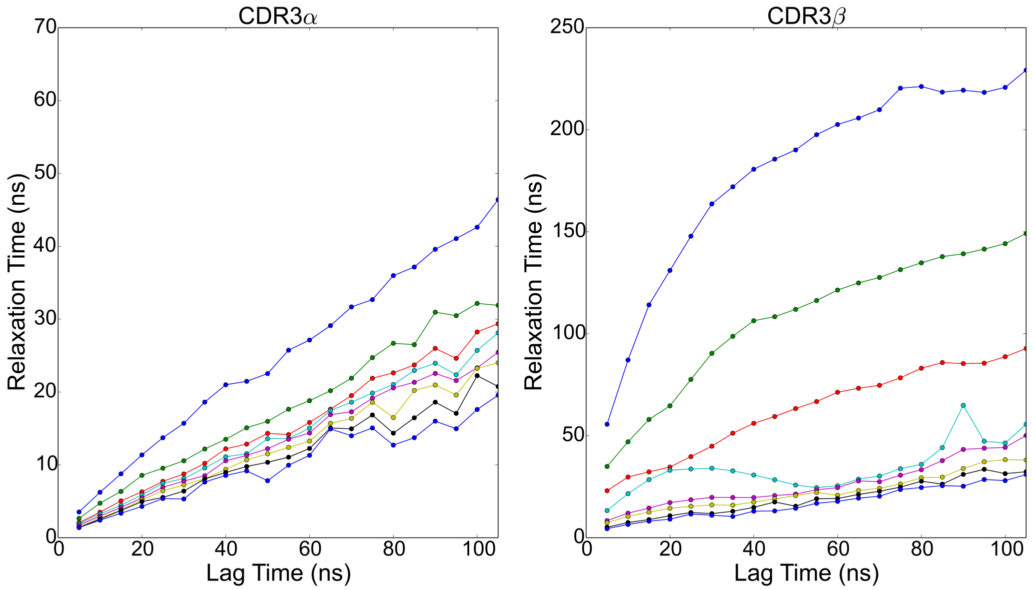

Figure 4.2: Implied timescales/relaxation timescales derived from eigenvalue analysis of microstate MSMs for CDR3α (A) and CDR3β (B) loops.

ysis of the MSM, rather than from the directly observed data, meaning that the equilibrium is determined by analysis of the kinetic model, not direct sampling. Importantly, despite a low empirically observed frequency of state four, state four appears to involve hydrogen bonding interactions with the CDR3α loop, and demontrates immense stability due to a mixture of hydrogen bonds and hydrophobicity. This phenomena is described in more detail in a later section, and demonstrates the MSM picking up and emphasizing kinetic details over the directly sampled data in an undersampled data regime.

The backbone ϕ/ψ angles of the eight central residues of the CDR3β loop are shown in figure 4.3D. G205 and L210 show minimal variation between the centroids, suggesting that flexibility at these positions is not required to generate the observed collection of metastable states. Diversity is seen in both of the angles of G207, while S204 separates out the state 1 and 2 orientations along the others along the ϕ and ψ angles respectively. G206, G208, and Y211 primarily separate a single centroid orientation from the other three along a single ϕ or ψ angle, while showing minimal varation in the non-separating angle. T209 appears to separate centroids along the ϕ angle, however variation is seen under re-clustering of the original microstate clusters, while the behavior of the other angles is stable, implying that T209 is flexible but does not meaningfully describe the different states.

The macrostate Markov model of CDR3β shows distinct pathways between the different metastable clusters and differing levels of metastability in the states, with states 3 and 4 showing strong metastable behavior, while states 1 and 2 are only weakly stable (figure 4.4). Despite the relative instability of state 2, it acts in a hub-like fashion, with the largest rates into states 1, 3, and 4 all coming from state 2. Rates into state 2 are also highly relative to all other state transitions, with the exception of the state 1 to state 3 transition which shows similar magnitude to the state 1 to state 2 transition. The other transitions show much lower flux rates, so that state 3, although only weakly metastable, acts as a central metastable intermediate. This high flux into state 2 accounts for the high population observed in the equilibrium distribution of the state despite the weak stability. State 1 is also weakly metastable, but does not have a counter-balancing inward flux, leaving it as a simpler weakly metastable state, which accounts for its low equilibrium population. State 1 has a large outward flux to both state 2 and state 3, with the most significant in-flow coming from the hub-like state 2, positioning state 1 as an alternate pathway to access the much more stable state 3. State 4 only shows significant exchange with the hub-like state 2, and shows strong stability and high equilibrium population similar to state 3.

4.3 CDR3α and CDR3β loops of Type I NKT TCRs have metastable states

Unlike CD4$^+$ and CD8$^+$ $\alpha\beta$ TCRs, type I NKT $\alpha\beta$ TCRs recognize lipids presented by CD1d, a monomorphic MHC-like protein. NKT TCRs do not show significant variation in their bound state footprint, and crystal structures show comparatively little movement between free and bound conformaitons, despite variation in the chemical structures of the presented lipids. Type I NKT TCRs show significantly higher binding affinities than CD4$^+$/8$^+$

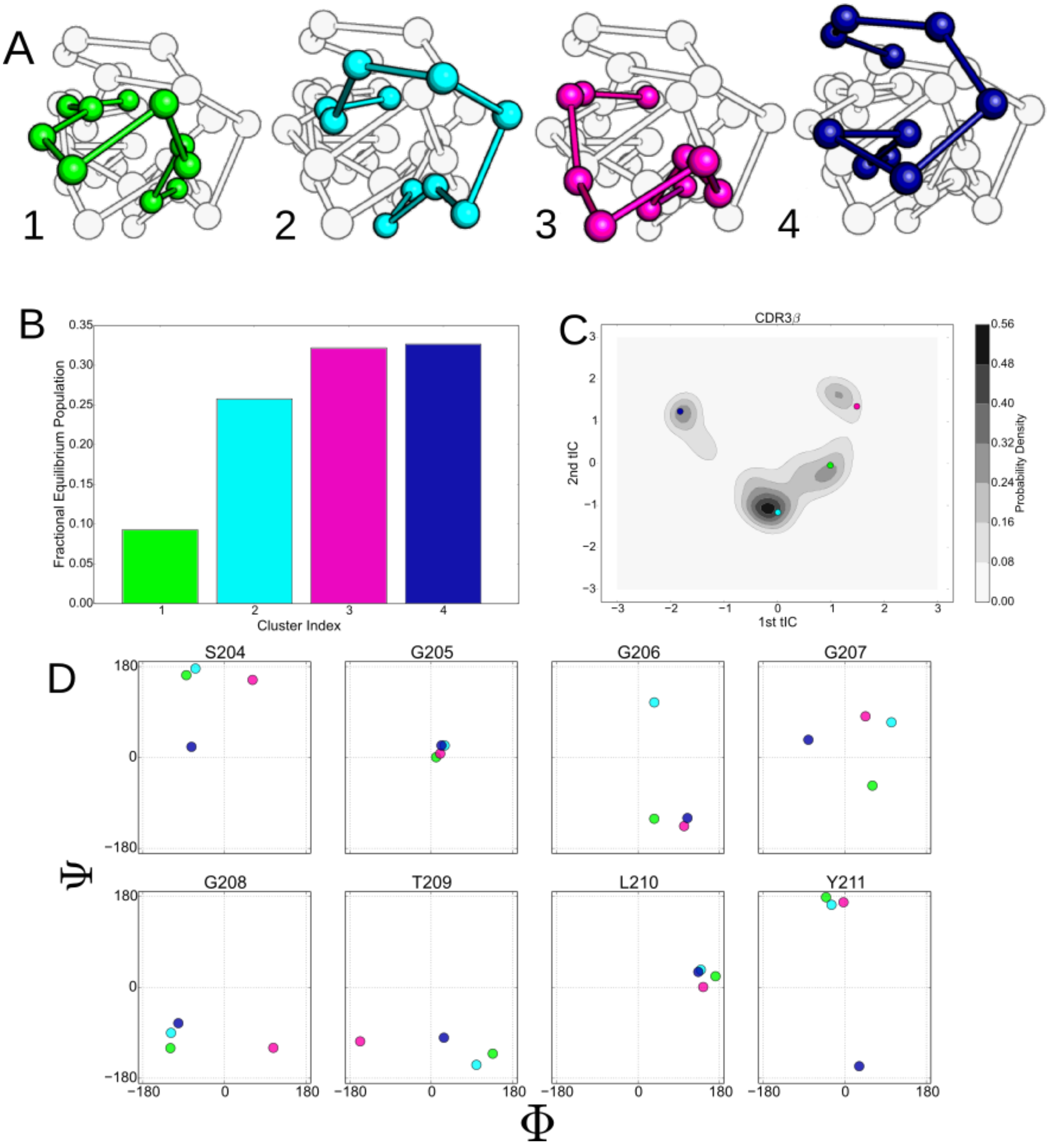

Figure 4.3: (A) Ball and stick model of the CDR3β loop showing the centroids of the four macrostate clusters determined by the MSM. Centroids were determined by finding the orientation that minimizes the distance to all other members of the cluster under the tICA projection distance. (B) Equilibrium populations of the four clusters, determined by eigenvalue analysis of the macrostate MSM. (C) Projection of the centroids onto the first two tICA dimensions overlaid on the kernel density estimate of the projected data. (D) ϕ/ψ backbone angles of eight residues along the CDR3β loop. Colors are consistent throughout for states 1 (green), 2 (light blue), 3 (purple), and 4 (dark blue).

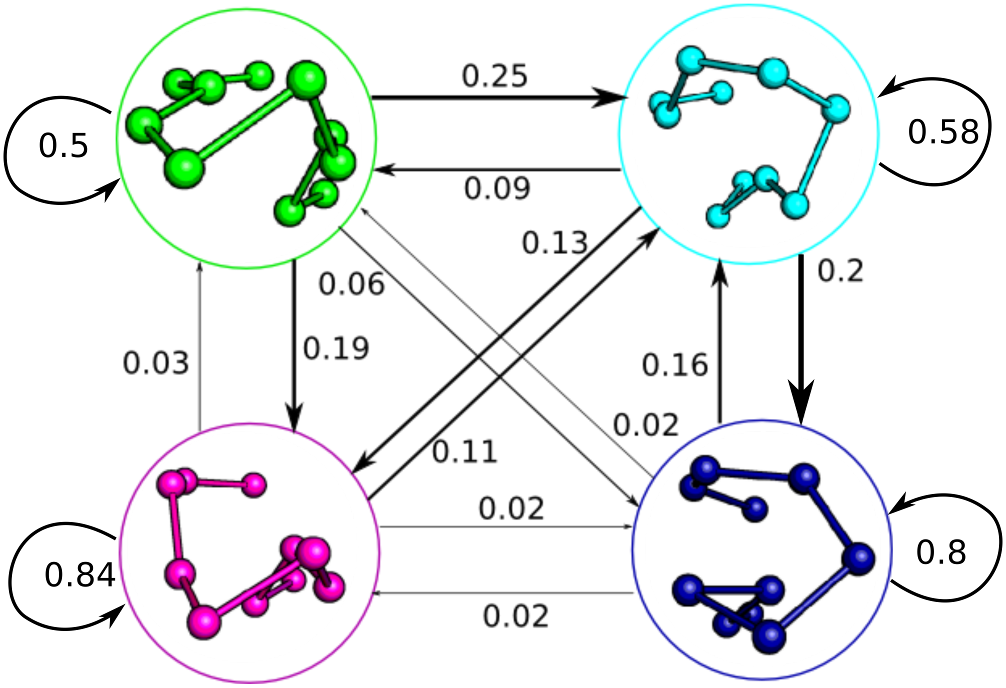

Figure 4.4: Macrostate Markov State Model of 2C CDR3β built with a time-lag of 115 nanoseconds. State clusters are represented by their centroids as initially described in figure 4.3A, and jump probabilities are described by arrows labeled by the probability of that state transition occuring in a 115 nanosecond time step. Arrow size is proportional to jump probability.

TCRs, and have binding kinetics that suggest an innate-like response. As they use the same immunoglobulin architecture as standard CD4$^+$/8$^+$ TCRs, we investigated the unbound dynamics of the NKT TCR as a comparison to the dynamics of the 2C system. We ran 10 independent trajectories of the NKT15 TCR for 100ns, collectively totaling 1μs of data.

We applied the tICA decomposition to the backbone dihedral angles of the CDR3α and CDR3β loops of NKT15 with a timelag of 5ns, just as with the 2C system. Similar to the 2C TCR, the tICA decomposition is indicative of low-dimensional, structured motions. Most of the tICA degrees of freedom consist of Gaussian motions around a mean of zero, thus consisting of thermal motion, with only one degree of freedom for each loop showing multiple peaks that suggest metastable conformational regions (figure 4.5A). Plotting the density estimates of the first two degrees of freedom for each loop, we find that both loops show two distinct high probability regions separated by lower probability transition regions (figure 4.5B, C). In both CDR3α and CDR3β, the two local probability maxima are separated along a single axis, so only a single degree of freedom is responsible for the transition ebtween these high-probability regions. Furthermore, in both systems, one of the high-probability regions shows a much higher probability relative to the other, suggesting the existence of a single major local energy minima, and a kinetically nearby metastable state with higher energy. In contrast to 2C, we observe distinct metastable regions in both systems, although CDR3β is much simpler in NKT15 than in 2C, with only two metastable states separated along a single degree of freedom, implying that NKT15's motions are more restrained than 2C.

4.4 CDR3α and CDR3β loops interact in 2C through hydrogen bonds

Previous work has shown that there is weak, if any, coupling between the overall loop dynamics of CDR3α and CDR3β loops in the A6 TCR. However, we do observe direct hydrogen

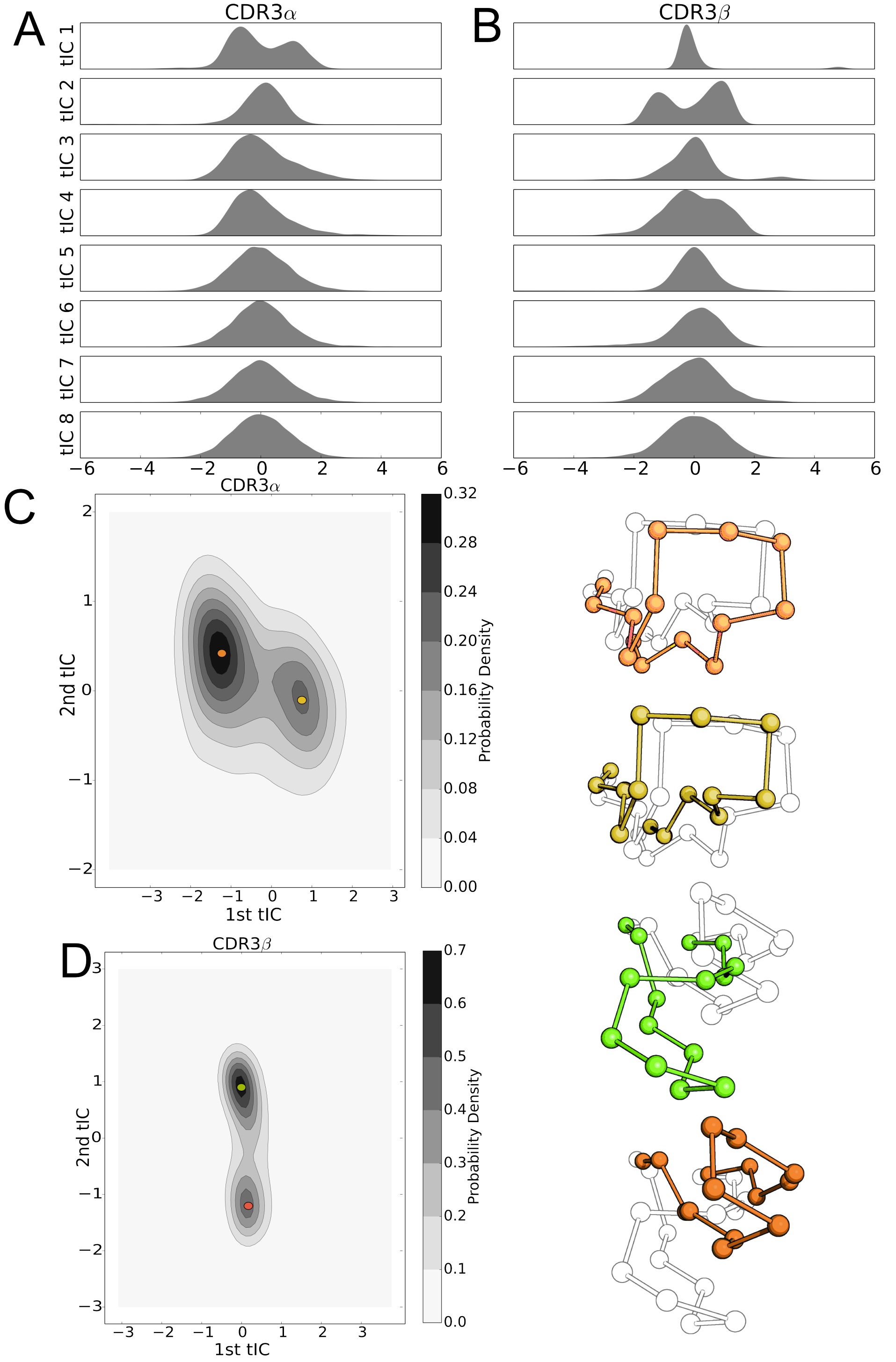

Figure 4.5: (A) Probability distributions of the NKT15 CDR3α and CDR3β conformations projected onto each of the first eight tICA degrees of freedom, computed by a 1-D kernel density estimate with a Gaussian kernel. (B) 2-D probability distribution of NKT15 CDR3α projected onto the first two tICA degrees of freedom; selected conformations from the simulation are shown in orange and gold and overlaid on the probability distribution plot. (C) As in panel (B) for the CDR3β loop with selected conformations shown in light green and ochre. Probability distributions were computed by a 2-D kernel density estimate with a Gaussian kernel over all collected trajectory data.

bond interactions between the CDR3α and CDR3β loops of 2C when CDR3β adopts the metastable state 4. In two of the ten trajectories of 2C, the CDR3β loop adopts a conformation that permits a hydrogen bond between the sidechain of Ser93 on CDR3α and backbone of Gly207 on CDR3β (figure 4.6A). The CDR3α loop's conformation that permits this bond is near the high-probability region observed in the tICA projection, and may account for some of the long-tail spread observed in the first tICA degree of freedom for CDR3α (figure 4.6B). The CDR3β loop of 2C appears to be able form this bond only in state 4 where the CDR3β loop is oriented to make the Gly207 backbone contact with the CDR3α Ser97. The hydrogen bond demonstrates significant stability, appearing in 25% of frames assigned to state 4. The persistence of this interaction and the specificity of the orientation required to allow it accounts for the high equilibrium population of state 4 in the Markov state model. As the model relies on kinetic information to determine equilibrium populations, rather than directly observed conformations, the model indicates that this hydrogen will tend be a high-population, dominant state over a long time scale relative to the observed sample from the simulation.

This conformation is further stabilized by multiple intra-loop polar contacts and a hydrophobic 'shell' that protects the hydrogen bonds from solvent interactions. In addition to the inter-loop contact between Ser93 and Gly207, in the sample frame we observe a CDR3α-CDR3α hydrogen bond between Ser93 and the backbone of Gly206, as well as interactions between Thr209 and Gly206 (figure 4.6A). Surrounding these hydrogen bonds are numerous hydrophobic residues that can shield the hydrogen bonds from solvent, as depicted by the pink residues in figure 4.6A. There are nine hydrophobic residues within 6 angstroms of either Ser 93 or Gly207, creating a hydrophobic shell around the inter-loop hydrogen bond and shielding some of the intra-loop interactions as well. Surrounding hydrogen bonds with hydrophobic residues has been shown to enhance stability[1], suggesting this hydrophobic shell

[1] Christopher M. Fraser/Ariel Fernandez/L. Ridgway Scott: Wrappa: A screening tool for candidate dehydron identification, tech. rep. TR-2011-5, University of Chicago, 2011; idem: Dehydron analysis: quantifying the effect of hydrophobic groups on the strength and stability of hydrogen bonds, in: (Advances in

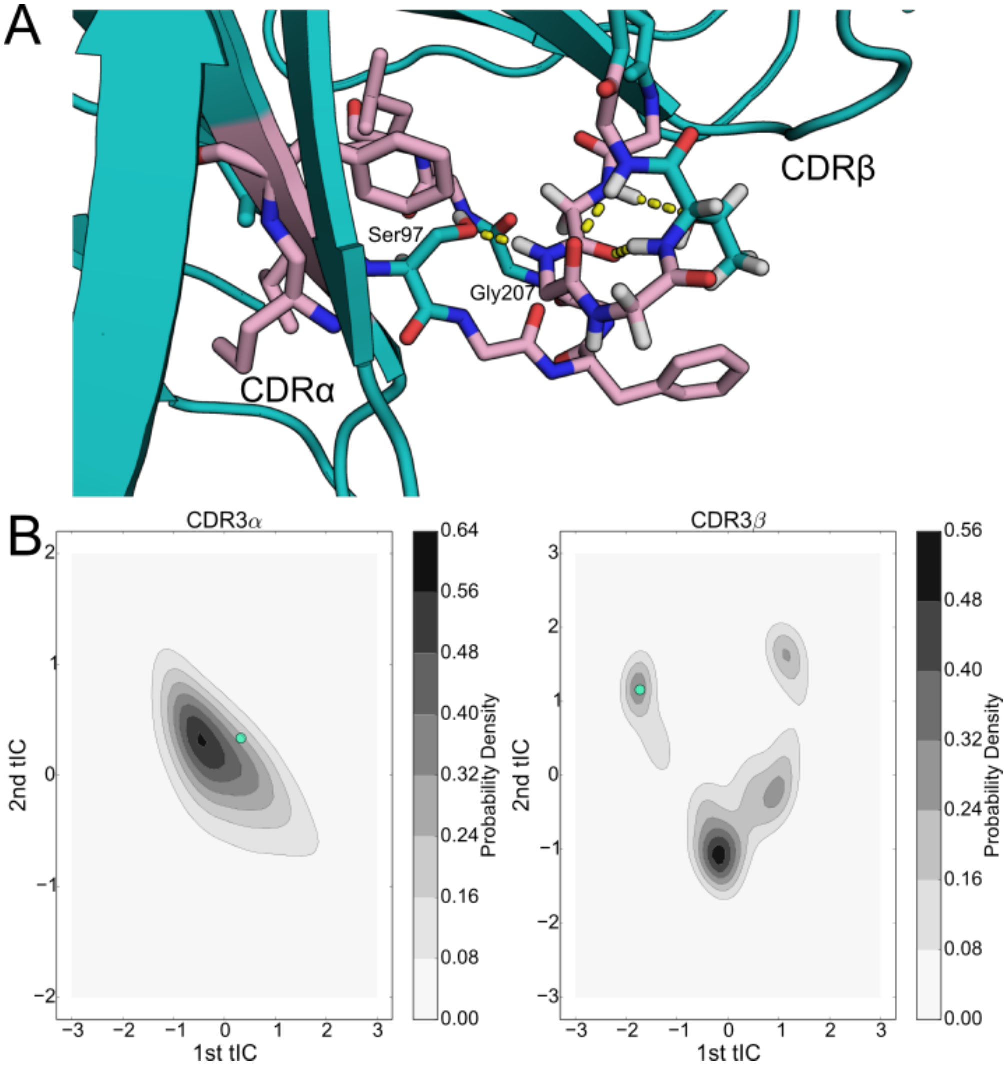

Figure 4.6: (A) Structural rendering of the 2C hydrogen bond interaction between Ser93 and Gly207 with the Vα domain shown on the left and the Vβ domain shown on the right; hydrophobic residues surrounding the hydrogen bonds are shown in pink. (B) Projection of the data frame onto the tICA projections of the CDR3α (left) and CDR3β (right) loops overlaid in cyan.

is responsible for the significant stability of the hydrogen bond and the conformation. The stability is notable in the simulations, as there are no transitions out of this state observed in the simulation trajectories where it occurs. This 'hydrophobic collapse' conformational state is both structurally and kinetically separated from the other conformations, notably looking unlike known bound states of 2C, and potentially acting as a hydrophobically driven 'off' state that reduces the overall affinity of the TCR by stabilizing a binding-incabable state.

At the same time, the hydrophobic sidechains that contribute to the stability of state 4 may explain the instability of states 1 and 2 in which the CDR3β loop is more extended and thus more solvent accessible. The increased solvent exposure of the hydrophobic sidechains will create unstable conformations, leading the CDR3β loop to 'search' for a conformation that once again buries the hydrophobic residues, leading to the transition-state behavior of states 1 and 2 where the CDR3β loop is frequently sampling, possibly unsuccessfully, transitions out of the conformational state.

4.5 Simulations reproduce CDR3β bound crystal structure orientations

We are able to compare our results with experimentally determined crystal structures in two ways. First, as the tICA projection matrix can project previously unobserved data, we projected the CDR3β loop conformations of three bound structures of 2C in complex with H-2k^b/SIYR (PDB 1G6R), H-2K^b/dEV8 (PDB 2CKB), and H-2L^d/QL9 (PDB 2OI9) onto the two dimensional space of the first two tICA degrees of freedom (figure 4.7A). We omit CDR3α projections because no bound states of the CDR3α loop are found in the simulation trajectories, implying either a much slower transition time as observed for A6[2], or that the conformation of the CDR3α's bound state is unfavorable without the environment of the

Computational Biology), 2010, pp. 473–479.

[2] Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism (see n. 13).

peptide-MHC.

Projecting the bound conformations onto the first two tICA degrees of freedom, we find that H-2K^b/SIYR and H-2K^b/dEV8 both appear near the most frequently observed region of the tICA conformational space, but are themselves in low probability regions that appear to be transition regions between two metastable states. This indicates that although the bound conformation for these antigens are closely sampled in solution, they are unlikely to directly be the result of selection from a pre-existing equilibrium. However, they are kinetically close to two well-populated metastable states, making it plausible that if a binding event is initiated from either of these two metastable regions, then CDR3β will be able to rapidly find the correct orientation observed in the bound state. In contrast, the bound conformation for the alloreactive H-2L^d/QL9 falls into the region corresponding to state 2 of the MSM, which is the lowest equilibrium population state of the model. Intriguingly, both antigens that use the H-2K^b MHC fall into the transition-like region, but nearer to the hub-like state 2, while H-2L^d/QL9 falls into a distinct region in the projection, and biologically presents in a different context than H-2K^b.

4.6 Reverse simulations indicate slow CDR3α dynamics

In our main dataset, CDR3α did not transition to a bound-like conformation in any of the ten trajectories. This strongly suggests that the bound conformation lives in a stable, local energy minima with slow kinetics between the bound and unbound-like regions of phase space. To test the stability of the bound state, we ran an additional ten trajectories of 2C, initialized with the coordinates of the bound state for 2C bound to H-2K^b/SIYR. Trajectories were run for 100 ns each, collecting an aggregate of 1μs. CDR3α remained near the bound conformation for the entirety of all ten trajectories (figure 4.8), in line with the hypothesis that the bound state is a stable local well. Because no transitions are observed in any trajectory, we are unable to construct a Markov state model of the CDR3α, however the data indicate that CDR3α is stable in bound conformation independent of the environment

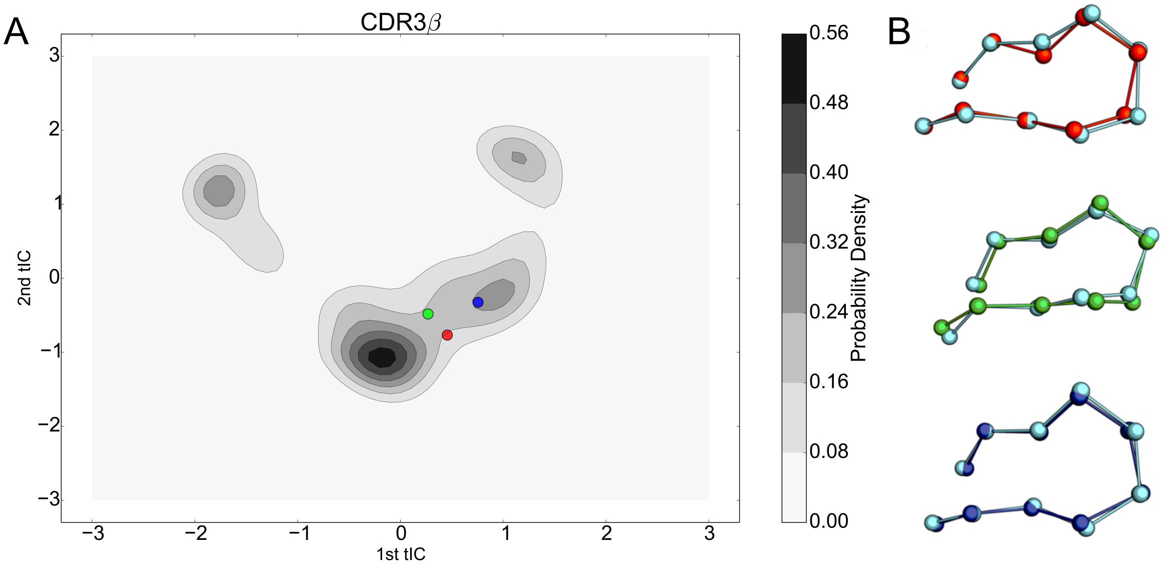

Figure 4.7: (A) tICA projections of the bound 2C CDR3β loop conformations for 2C bound to H-2L^b/SIYR (red), H-2K^b/dEV8 (green), and H-2L^d/QL9 (blue) overlaid on the 2-D probability density. (B) Ball and stick render of the CDR3β bound crystal structures overlaid with the nearest simulation frame by RMSD of the Cαs after aligning the β variable domains. Simulation data is shown in cyan.

of the peptide-MHC, and the kinetics of transitions between these states are very slow. This is in line with observations of A6, where simulations yielded only a single transition of CDR3α in an aggregate data set of 460 ns, suggesting that slow CDR3α dynamics may be a general feature of CD4$^+$/CD8$^+$ TCRs.

4.7 Analysis of 2C and NKT15 loop fluctuations

One major difficulty of tICA and similar projection techniques is detailed comparison between data sets of different molecules. The projection is parameterized to the data set that generated it, and in general, scaling between different data sets analyzed with tICA is unrelated. To address the issue of loop flexibility in both 2C and NKT15 directly, we look at the fluctuations of the Cα of the tip region of the loop over the course of each trajectory. This was determined by taking the Cα of the loop residues that show the largest average displacement in each trajectory and plotting the frame-by-frame displacement from the initial position of the trajectory.

The CDRα loop of 2C shows significant flexibility and variance in most trajectories (Figure 4.8 A). The hydrophobic collapse state that promotes the hydrogen bonding interaction between CDR3α and CDR3β is visible in the CDR3α traces as large displacements from the original orientation but sharp drop-offs in variance as the loop is constrained to that conformation. This aligns well with the idea that 2C's CDR3α has a broad energy well in the unbound-like state with relatively little structure, with the exception of the tightly controlled collapse state. The CDR3α loop of NKT15 (Figure 4.8B) showed lower overall displacement from the original position and lower variance, though some regions show higher local variability, usually coupled with larger displacement from the starting conformation; these two regions may separate the two regions observed in the tICA 2D projection. The CDR3α loop of NKT15 does appear less flexible than the CDR3α loop of 2C over the timescales observed.

The CDRβ loops of 2C and NKT15 show similar levels of displacement and variance, demonstrating similar levels of flexibility, and the maximum displace of NKT15's CDR3β was larger than that observed for 2C. Given this data, CDR3β's flexibility does not appear to be related to germline selected V[β] segments. However, CDR3β is significantly less important to NKT recognition than CDR2β and CDR3α.

The overall flexibility of NKT15 compared to 2C is unexpectedly high given the minimal variation in conformation observed in crystallographic data of bound NKT structures, and the rapid binding kinetics. It is possible that the flexibility is necessary to ensure the lipid adopts the proper conformation observed in the bound state, but that the system essentially falls down an energy well toward the bound state upon interacting in a binding-capable state, leading to the fast binding kinetics. However, it is notable that the CDR3α loop, which is more important to recognition than CDR3β in NKT15, shows lesser flexibility than in 2C, and restricting this flexibility should be more important to the system than restricting the CDR3β loop. The restricted Vα segment repertoire may be in part to restrict the flexibility of the CDR3α loop, and simulating a larger variety of $\alpha\beta$ TCRs with different Vα segments

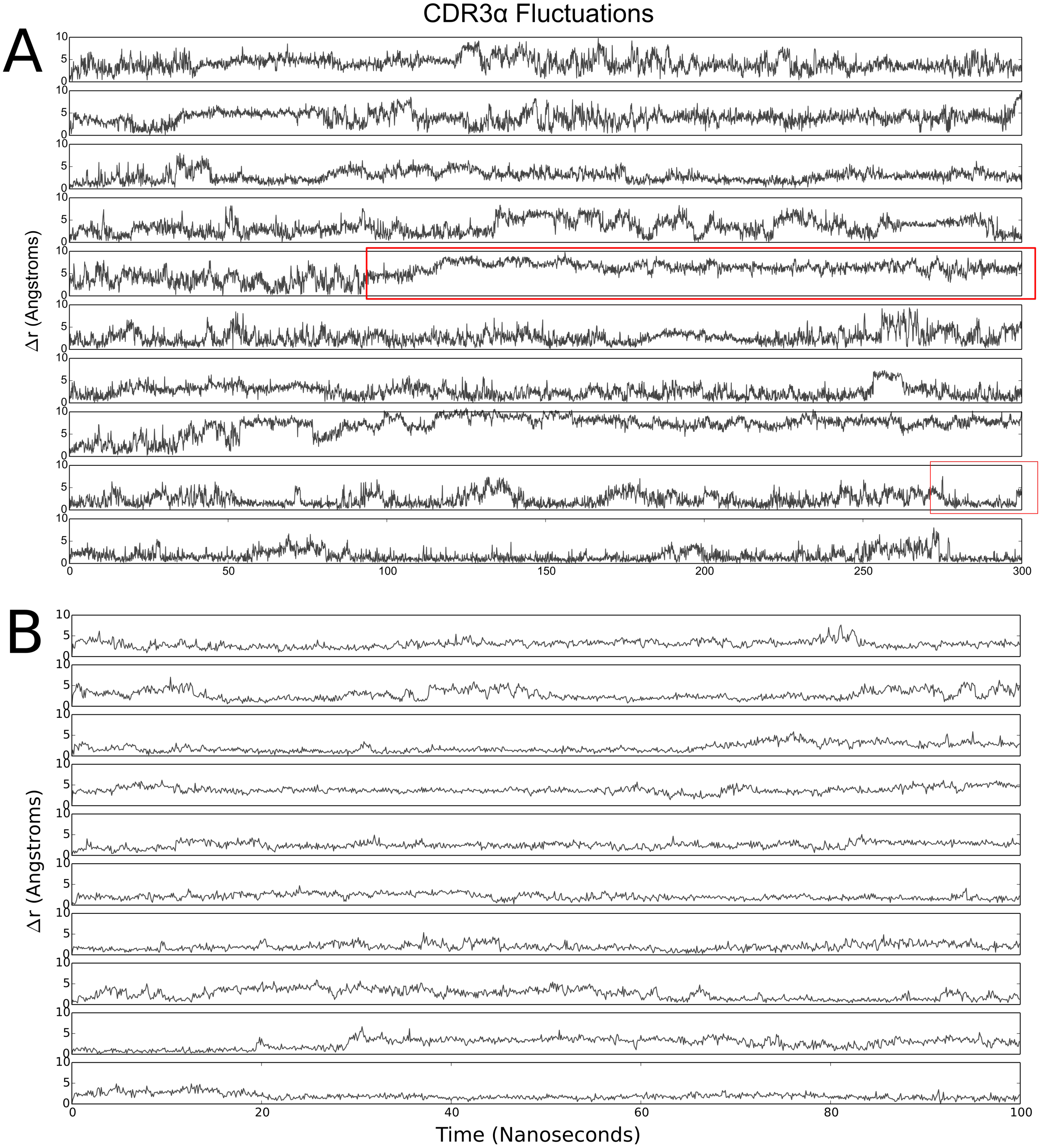

Figure 4.8: Trajectories of CDR3α loop tip motions. Traces for each trajectory of the Cα displacement magnitude of the Cα with the largest average displacement over the trajectory for 2C (A) and NKT15 (B). Red box highlights regions of 2C trajectories where the 'hydrophobic collapse' and accompanying hydrogen bonds occur.

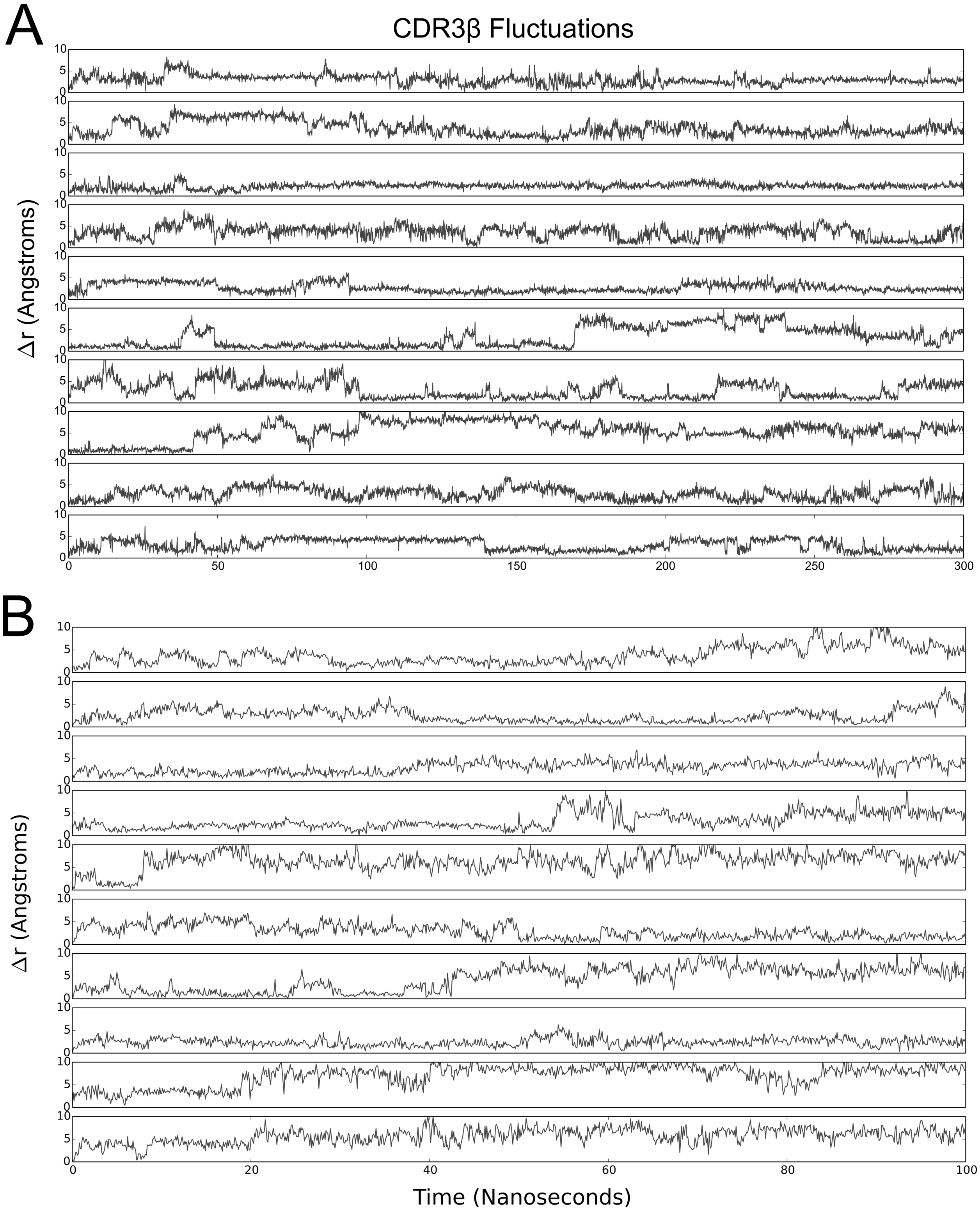

Figure 4.9: Trajectories of CDR3β loop tip motions. Traces for each trajectory of the Cα displacement magnitude of the Cα with the largest average displacement over the trajectory for 2C (A) and NKT15 (B).

would shed more light on the issue.

CHAPTER 5
CONCLUSIONS

5.1 On the T cell receptor

The flexibility and dynamics of the CDR loops of T cell receptors have long been a topic of speculation and interest. Crystallographic work has demonstrated the existence of multiple loop conformations in the bound state of the CDR loops and indicated that loop flexibility must necessarily play a role in cross-reactivity. Here, we have used the Markov state model framework to show that in 2C's CDR3β loop, there exist clusters of conformations that are distinct and exist independent of the environment of the final binding state, and that these conformations are much broader even than those variations observed in the known crystal structures of 2C, our model system. We have shown that these individual states, made of many kinetically related conformations, are inherently stable in a fashion that makes them fitting of the term 'state', and there exists a distinctive structure in the movements of these loops between these states. Previous pioneering work by Scott et al. demonstrated the existence of distinct clusters of conformations in the unbound A6 TCR[1], and provided evidence for a slow mode of motion in the CDR3α loop, and faster, more diverse motion in the CDR3β loop. Our results find good agreement with this work, suggesting a common behavior, that of slower, simpler dynamics in the CDR3α loop and faster, more complex dynamics in the CDR3β loop, for $\alpha\beta$ CD4$^+$/CD8$^+$ TCRs. We have furthermore demonstrated the stability of these clusters, showing them to be true local minima, providing distinct conformational groups that can potentially act as a source of initial conformations from which

[1] Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism (see n. 13); Scott et al.: Limitations of time-resolved fluorescense suggested by molecular simulations: assessing the dynamics of T cell receptor binding loops (see n. 31).

selections can be made, in either a conformational selection or conformational melding model of TCR-pMHC recognition.

The tICA decomposition is a powerful too for understanding the complexity of the motions we observe. Previous work, both computational and crystallographic, has firmly established the flexibility of CDR3 loops in CD4^{+}/CD8^{+} $\alpha\beta$ TCRs, but it has been difficult to understand how well structured those flexible motions are, that is, are the motions precise and organized through specific degrees of freedom, or are the loops more like a rope, able to flex anywhere along its length? With the tICA decomposition into linear, orthogonal degrees of freedom, we can characterize these motions by the number of orthogonal degrees of freedom that meaningfully contribute to the state transformations, in the case of 2C CDR3β, we observe two orthogonal degrees of freedom that captured by the tICA decomposition that reveal evidence of substates and probability densities that are distinctly non-Gaussian. Thus, 2C's CDR3β loop moves, with respect to its internal motions, through a two-dimensional space and has a restricted flexibility. Even more strikingly, we see that with a tICA decomposition of the available data for NKT15, both CDR3α and CDR3β are described by a single tICA degree of freedom. The Cα at the top of the CDR3β loop NKT15 shows a larger variation in its location in real space than the corresponding measurement of 2C's CDR3β loop, however NKT15 is less flexible in that is has fewer degrees of freedom, forcing it to adopt simpler motions than those available to 2C.

This difference in the dimensionality of flexible motion of the TCRs is a qualitative demarcation between 2C and A6 on one hand, and NKT15 on the other. While we have not examined A6 using the tICA decomposition, the similarity in conformational state diversity observed in previous work to the diversity we observe in 2C makes the extrapolation a reasonable hypothesis, and the simplicity observed in NKT15's dynamics yields a possible explanation for the different kinetics observed. In the present work, only one of 2C's bound states falls into the locally most probable region (QL9-Ld antigen), while the other two bound states appear in a lower probability transition region between two wells. This supports the conformational melding hypothesis; there are clear clusters of conformations that would be capable of more quickly finding the bound state, but the actual bound states are not so likely that the binding mechanism is well described as conformational selection. On the other hand, we achieve an incredibly close match between the CDR3β loop's bound states and simulation frames, and find the bound states project well onto the observed data under the tICA decomposition's projection, demonstrating that the environment of the peptide-MHC is not required for CDR3β to find a bound-like conformation, as would be expected from a pure fold-upon-binding mechanism.

We can roughly partition agonist $\alpha\beta$ TCR kinetics into two classes, those which have slow off rates, and those with on rates fast enough to rebind before diffusing away where analysis of re-binding events have been shown to effectively predict signaling[2]. In the more classical, slow off rate case, 'local search' could explain the slow observed binding kinetics, as put forward in the conformation search and conformational melding hypotheses[3]. Conformational melding effectively argues that the search is local, and thus must be seeded by a conformation that is initially selected from a set of equilibrium conformations; the observed state clusters in our Markov model provide distinct initial states for such a seeding in accordance with the melding hypothesis.

On the other hand, the innate-like kinetics of type I NKTs would suggest simpler motions[4], which are apparent in the tICA decomposition of the NKT15 simulation data. The crystallography of NKTs demonstrates little variation in binding orientation, unlike 2C, the footprints of type I NKT TCRs are nearly identical across different antigens, which is fitting with the faster kinetics. The need for faster kinetics and thus simpler loop dynamics can potentially explain the reduced selection of variable domains: the reduced selection has been evolutionarily selected specifically for the tendency to create simpler loop dynamics,

[2] Govern et al.: Fast on-rates allow short dwell time ligands to activate T cells (see n. 42).

[3] Gagnon et al.: T cell receptor recognition via cooperative conformational plasticity (see n. 27).

[4] Rossjohn et al.: Recognition of CD1d-restricted antigens by natural killer T cells (see n. 5).

while still using recombination to make minor adjustments to the enthalpic contacts and potentially alter the equilibrium distribution of states. As we observe two states for each of the CDR3α and CDR3β loops in 2C, it is reasonable to postulate that these alternative states may act as simple 'off' states that do not permit binding, thus acting modulate overall affinity.

A major outcome of 2C's flexibility is the creation of the hydrogen bond interaction between CDR3α and CDR3β and the hydrophobic region that stabilizes the interaction. It is likely this state is binding-incapable, as the cluster conformations differ sharply from the bound conformations present in crystal structures, which suggest a dual-role for the CDR3β in both MHC recognition and overall affinity adjustment. The hydrogen bonded state 4, and the less well characterized, but similarly stable state 3 in our MSM of CDR3β appear to be 'off' states, whose equilibrium populations would control affinity by altering the probability that the TCR is binding competent or binding incompetent. A similar role has been suggested for CDR3α in the context of A6 due to its slow motions. If these states are also reachable in the bound system, they may also adjust the off-rates depending on how accessible they are. On the other hand, states 1 and 2 divide the bound conformations by MHC, suggesting that CDR3β conformations contribute to MHC recognition as well as peptide specificity. Despite the length of our simulations, transitions out of the hydrogen bonded state are not observed, which limites our understanding of the state dynamics and limits the quantitative value of the CDR3β MSM. Nonetheless, the qualitative results, the existence of four distinct conformational clusters, is clear.

Finally, we note that the existence of these slow dynamics and long-lived metastable states indicates a need for significantly longer trajectories and larger data sets. We have contributed a large data set for a single TCR, which we believe to be the largest set of trajectories for a free TCR that deals with only a single system and thus is comparable across trajectories, as well as allowing for independent trajectories to evolve. Much work has largely used 100 ns or shorter trajectories[5] , often with fewer trajectories, with trade-offs between deeper sampling of a particular phenomna or broadly sampling many comparable systems forced by technological and resource constraints. Using MSMs to knit together multiple trajectories into a larger picture and taking advantage of GPU-enhanced calculations to greatly extend the size and scope of simulations offers a much more comprehensive picture for single systems.

5.2 Conformational Dynamics and Models of TCR Binding

The data presented most strongly support the conformational melding model, rejecting strong forms of both the induced fit and conformational selections models.

Induced Fit

The data does not support the strongest form of the induced fit model on the simple basis that distinct states exist in solution, as demonstrated by the kinetic clustering of the MSM, and the CDR3 loop motions are constrained to few degrees of freedom. Comparing to the tICA projection and the known bound states, we might expect the bound states to occur far from solution-state conformations if induced fit is accurate, shown as the orange region in figure 5.1. We do not observe the bound states here, instead finding them in the transition region between states 1 and 2. This does fit with weaker forms of the induced fit model, where the induced conformations are more the result of freezing out alternative conformations, but this arguably the conformational melding model aside from focusing solely on the TCR's behavior.

[5] Scott et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism (see n. 13); Knapp/Dunbar/C.M.: Large Scale Characterization of the LC13 TCR and HLA-B8 Structural Landscape in Reaction to 172 Altered Peptide Ligands: A Molecular Dynamics Simulation Study (see n. 48).

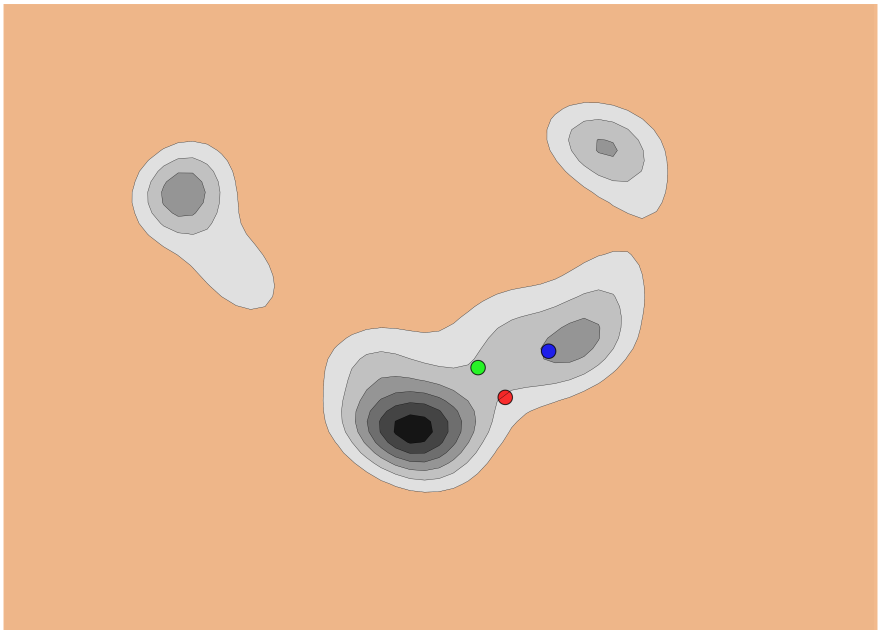

Figure 5.1: The tICA projection of the primary simulation dataset with the bound states' projections shown in red, blue, and green. The orange region shows the expected location of bound states under the induced fit model.

Conformational Selection

The conformational selection model appears at first to be well supported by the data; the CDR3 loops have restricted degrees of freedom and CDR3β displays distinct conformational states that the loop transitions between and holds. However, if the selection model were true, we would expect that the bound structures would be located in the regions of high local probability, as shown in orange in figure 5.2. State 4 is excluded as an expected bound state in the figure, as the hydrophobic collapse is likely to be binding incapable and is furthest from the known bound structures. Only one bound state is near such a region, and it is a particularly unstable region, while the other bound states are located between two states in a less well-populated region.

However, we find that the bound states are located between states 1 and 2, with only one bound structure, curiously a antagonist ligand[6], near a high probability region. This

[6] Stone/Chervin/Kranz: T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity (see n. 23).

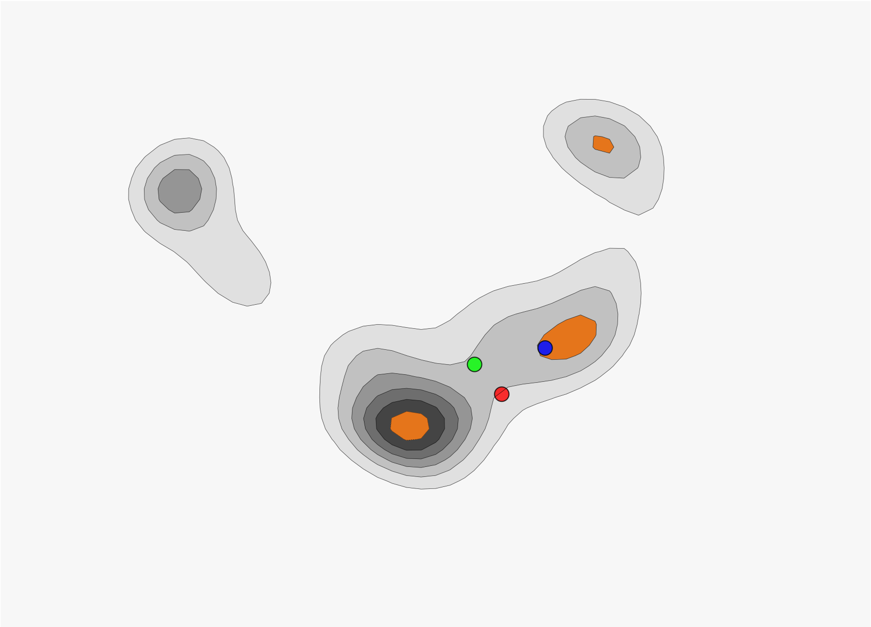

Figure 5.2: The tICA projection of the primary simulation dataset with the bound states' projections shown in red, blue, and green. The orange regions show expected locations of bound states under the conformational selection model. The region corresponding to state 4 is omitted because of the likelihood it would be binding incapable.

strongly suggests that stable states are not selected as bound conformations, rejecting the central idea of the conformational selection model.

Conformational Melding

Conformational melding fits the data best. From the perspective of only the TCR's dynamics, the conformational melding model is largely a golden mean; the loops are not unstructured and may have distinct states, but those states are not directly the bound conformations. Rather, the states may seed the search process. In particular, conformational melding implies the need for weakly stable states, as the system needs to be able to perform a local search in order to find the correct bound conformation. This is reflected in the orange region of figure 5.3, where we might expect that the bound states would be in marginally populated regions near the less stable states, particularly near the hub-like state 2. This is precisely where we find the bound conformations.

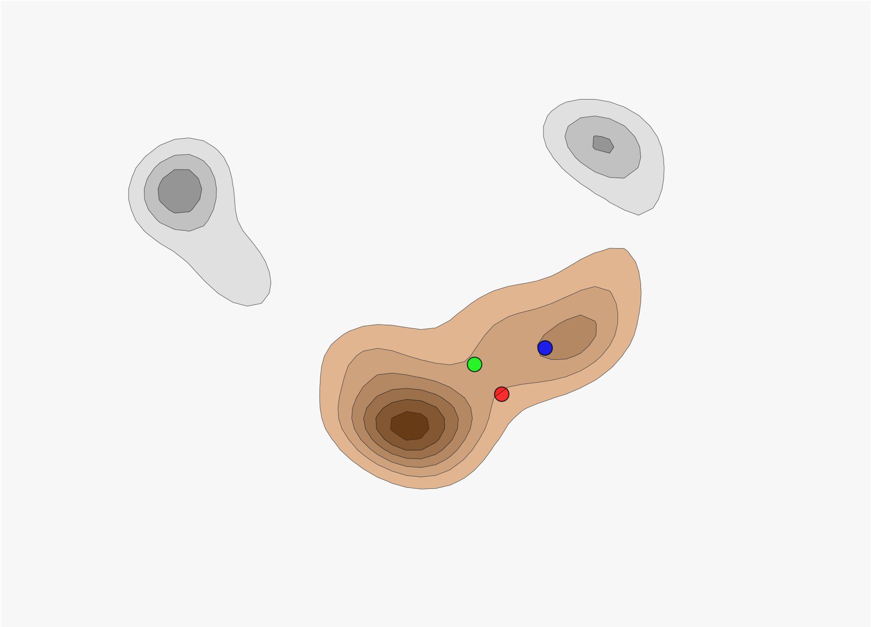

Figure 5.3: The tICA projection of the primary simulation dataset with the bound states' projections shown in red, blue, and green. The orange region shows the expected location of bound states under the conformational melding model.

5.3 Future Directions

This work presents one of, if not the, largest simulation datasets of a single T cell receptor to date, and provides a novel analysis of the system using kinetic clustering methods and machine learning techniques to discover previously unobserved conformational states and behaviors. However, we have found more questions than answers in doing so. This section describes several ways to extend the present work, and some experimental apporoaches that could yield more insight.

5.3.1 Experimental Probes

A major motivation for exploring the CDR3 loop dynamics via simulation is that there are not currently many effective experimental techniques for getting at the level of detail. Major methods with the potential for sufficient spatial and temporal resolution are single-molecule approaches using florescent probes and NMR. The single-molecule approach has been tried and shown to be unable to resolve the loop dynamics, though they did broadly validate simulation results[7]. NMR has sufficient resolution to determine larger conformational shifts, and has been shown to resolve multiple peaks in an NMR experiment examining the 2C TCR clone bound to a peptide-MHC ligand; this experiment was capable of resolving muliple conformations of the CDR3β loop, and implies that the CDR3β loop of 2C has multiple conformations in the bound state[8]. So NMR is possible and effective for T cell receptors, however the experiment is quite difficult; the form of 2C crystallized in the 1TCR pdb structure has 439 amino acids, making it roughly a 48 kDa protein, much larger than NMR can easily handle without significant resource investment. Cutting down to just the variable domain, there are still roughly 220 amino acids, which is around a 24 kDa protein before including the linker used in variable-domain only experimental set ups for, e.g. SPR. This is a managable size for NMR, demonstrated by NMR analysis of 2C[9], but still quite large, and thus exceptionally expensive to work with, particularly for systems that have to be expressed in more exotic cell cultures.

All of this is to say that NMR is likely to be highly impractical. Nonetheless, if it is feasible, then the Markov State Model can be used to directly calculate NMR parameters, and then compared to experimental values, which has shown excellent results in other contexts[10]. This is the cleanest way to directly assess the MSM results.

Another approach, though painful, is to make key mutations to the 2C sequence. The goal here is not to assess the MSM directly, but rather to use it to explore the consequences of the conformational dynamics. In particular, the serine at position 93 in the CDR3α

[7] Scott et al.: Limitations of time-resolved fluorescense suggested by molecular simulations: assessing the dynamics of T cell receptor binding loops (see n. 31).

[8] Hawse et al.: TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility (see n. 27).

[9] Ibid.

[10] D. Sezer/B. Roux: Markov State and Diffusive Stochastic Models in Electron Spin Resonance, in: (An Introduction to Markov State Models and Their Applications to Long Timescale Molecular Simulation), 2014.

loop makes the hydrogen bond that appears central to the 'hydrophobic collapse' state that appears as state 4 in the MSM of CDR3β in the present work. Mutating this position to eliminate the sidechain-mainchain hydrogen bond should destabilize this state. If it is indeed an 'off' state, this should cause a general increase in the overall affinity of any ligand, on the order of a factor of 1.5 increase in affinity if the state is completely eliminated, given that it has an occupancy of approximately 30% according to the MSM. However, this effect is tempered and possibly reversed if the binding ligand makes enthalpic contacts with the serine in the bound state, making this a tricky proposition. In the crystal structure 2CKB, of 2C bound to QL9/K^b, Ser93 makes a backbone hydrogen bond with a lysine on the peptide, as well as a backbone hydrogen bond to another loop via the backbone nitrogen, and a sidechain hydrogen bond to alanine at position 103 in the CDR3α loop. This suggests that inserting a hydrophobic residue like alanine, which would be the standard approach, at the Ser93 position could lose some stability due to the loss the sidechain hydrogen bond and introduction of a methyl group. A glycine should eliminate the sidechain-mainchain hydrogen bond observed in state 4 of the MSM without causing significant disruption of the TCR-pMHC interface in this particular system; extrapolating from the crystal structure we would expect only the loss of the intra-chain CDR3α hydrogen bond, and no disruption of the binding interface with the antigen. Other systems with bound crystal structures of 2C show that Ser93 forms hydrogen bonds to water molecules that in turn interact with the antigen, e.g. in the PDB file 2OI9.

Having identified a possible mutation that should have primarily conformational implications and not effect the interface contacts, the next step is perform the mutation *in silico* and perform the simulation of the mutated 2C TCR, repeating the MSM procedure. This will yield predicted states, which should align well with the states found in the current work due to the very small difference, though state 4 should be suppressed or eliminated. An increase in affinity, measured by SPR, should correlate with the decrease in occupancy of state 4. This would directly demonstrate both that state 4 is indeed an 'off' state and that the conformational dynamics play a direct role in binding behavior.

5.3.2 Further in silico Methods

There are several *in silico* extensions to the present work. The simplest but of clear value is to run longer and additional simulations. The presented data set does not show transitions between bound-like and unbound-like states of the CDR3α loop. This indicates long timescales involved in the transition, but does not reveal any informational about the transition pathway itself, and provides only a lower bound on the kinetics of the transition. As the CDR3α loop interactions are critical to binding, understanding the relationship between on-rates and conformational dynamics is quantitatively impossible without more accurate assessment of the CDR3α state transitions.

Path Sampling CDR3β

Similarly, we have described the CDR3β state transitions phenomenologically, and a deeper analysis of the transitions is warrented, as this would allow direct, specific inspection of the transitions modes between states that is only globally described by the tICA projection. CDR3β transitions can be assessed by either longer simulations or additional trajectories, ideally starting from the states extracted by the MSM 'kinetic clustering' analysis presented here. This would yield an unbiased assessment of the transitions between states, and in particular would allow exploration of transitions paths that are less likely. Furthermore, using the MSM states to seed standard MD simulations would make it possible to find further unexplored states, or provide evidence that such states do not exist. On the other hand, for studying the major transitions directly, biased path sampling methods such as the string method could be used to assess specific state transitions of interest. One major caveat of using biased methods is that the most interesting state presented here, state 4, relies on expulsion of solvent from the inter-loop region by stochastic fluctuations; in general, current biased methods do not reliable handle these contexts, so this transition is difficult to study via biased path-sampling approaches. One acceleration method that might work in metadynamics. The tICA decomposition has demonstrated that most of the behavior of the CDR3β loop is contained in the first two degrees of freedom spanned by the tICA projection. Since tICA is a linear projection operator, these degrees of freedom can be algebraically described in terms of the dihedral angles of the system, making them amenable reaction coordinates for use with metadynamics. This would allow for significantly more rapid sampling of the states accessible by variation in these two degrees of freedom, and demonstrate convergence in population frequency much more quickly than can be accomplished by direct simulation.

Simulating the Bound State

The single biggest target of value is simulating the bound state. Recent NMR work has shown that the CDR3β loop of 2C is mobile in the bound state with the SIYR peptide system[11]. The peptide is also mobile. This has led to speculation that the motions of the peptide and CDR3β loops are correlated. Simulation of the system would yield direct information about the motions of the bound state, and correlations can be examined in two methods. First, correlations can be directly inspected by calculating the mutual information between the two systems, either directly, which is likely to be difficult, or after reduction via a method such as tICA. Furthemore, if the CDR3β loop and peptide exhibit state-like behavior, which is suggested by the distinct NMR peaks observed, then a Markov State Model can be built to describe the system, and trajectories can be assigned to state for each of the peptide and the CDR3β loop, and correlations between state occupancy can be directly calculated. If the motions occur on a similar timescale, then both the peptide and the CDR3β loop can be described by a single Markov State Model, which would directly describe how well the system moves together. This can potentially be pushed further, using *in silico* mutation of the peptide to less favorable ligands, where if conformational selection is correct, we should

[11] Hawse et al.: TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility (see n. 27).

expect to see decreasing correlation in the movements of the CDR3β loop and the peptide. These experiments are straight-forward in light of the present work, but require a large investment of computational resources.

The bound state simulations and increasing the accuracy and depth of the current models will likely yield the best results at present. However, as described earlier, there are more classes of T cell receptors, both in terms of conventional $\alpha\beta$ TCRs with different binding kinetics and non-conventional TCRs such as the invariant TCRs associated with Natural Killer T cells or $\gamma\delta$ T cells, that we expect to show significantly different dynamics. Extending this analysis to more systems would make possible comparison between different systems that serve different immunological purposes but share, to various degrees, a fundamental architecture.

APPENDIX A
SIMULATING STOCHASTIC DIFFERENTIAL EQUATIONS AND THE TOY SYSTEM

If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.

– John von Neumann

This appendix covers the simulation of stochastic differential equations, and in particular the toy system in greater detail, including the code used to simulate the system. We begin with a 1-D Langevin system and then move to the toy system presented in Chapter 2.

All code has been tested with Python 3.4.

A.1 Langevin in One Dimension

The simplest SDEs are Brownian motion and Langevin dynamics in one dimension. In one dimension, Brownian motion is described by

$$\frac{d}{dt}x(t) = \sigma\eta(t)$$

Where $\eta(t)$ is a Gaussian process with zero mean, variance σ^2, and independent increments, i.e.

- $\langle\eta(t)\rangle = 0$
- $\langle\eta(t)\eta(t')\rangle = \delta(t - t')$

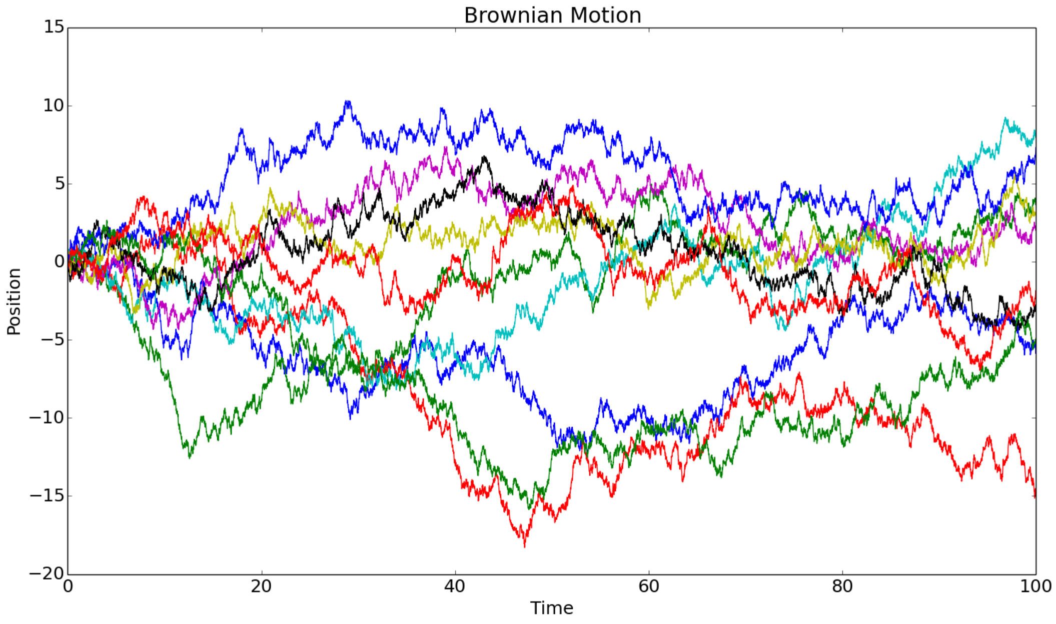

Figure A.1: Ten Brownian motion/Wiener process simulations over 100 time units.

With δ the Dirac delta function. Integrating yields the standard Wiener process, but our immediate interest is in simulating paths. We can do so using the Euler-Maruyama method, which yields

$$Y_{n+1} = Y_n + \sqrt{\delta t}\sigma \Delta W_n$$

Where the ΔW_n are independent and identically distributed normal random variables with expected value zero and unit variance. We can simulate several paths with

```
## Brownian Motion Simulation
## y is the simulated trajectory
import numpy as np

num_sims = 10
t0, t1, dt = 0, 100, 0.001
t = np.arange(t0, t1, dt)
sqrtdt = np.sqrt(dt)
y = np.zeros((t.size, num_sims))

for i in range(1, t.size):
    y[i] = y[i-1] + sqrtdt * np.random.normal(loc=0.0, scale=1.0, size=num_sims)
```

Langevin dynamics in the diffusive limit are only slightly more complicated, we simply add a deterministic term representing the potential energy field to the equation of motion for Brownian dynamics. Let $U(x)$ represent the one-dimensional potential. Then the equation of motion is

$$\frac{d}{dt}x(t) = -\frac{d}{dx}U(x) + \sigma\eta(t)$$

We can use Euler-Maruyama again to get the discretized form

$$Y_{n+1} = Y_n - \frac{d}{dY}U(Y_n) + \sqrt{\delta t}\sigma\Delta W_n$$

This is the same as the discretized form of the equation of motion for Brownian motion, with the addition of the deterministic spatial dependance on Y_n due to the potential energy field. We can simulate paths using

```python
## Builds a 1D SDE system
import numpy as np

def build_SDE(dV, sigma):
    # Builds a 1D Langevin IVP
    # Assumes a Wiener measure (isotropic white noise)
    # for the random component
    # dV: derivative of the potential field,
    # should take and return a float/double
    # sigma: variance of the process (a free parameter)
    def sde_model(t0, t1, dt, IC = None):
        t = np.arange(t0, t1, dt)
        rdt = np.sqrt(dt)
        y = np.zeros(t.size)
        if IC:
            y[0] = IC
        else:
            y[0] = 0

        for i in range(1, t.size):
            y[i] = y[i-1] - dt * dV(y[i-1]) + \
                    sigma * rdt * np.random.normal(loc=0.0, scale=1.0)
        return (t, y)
    return sde_model
```

Now, consider a potential of the form

$$V(x) = -\frac{1}{2}e^{-\beta(x-x_0)^2}$$

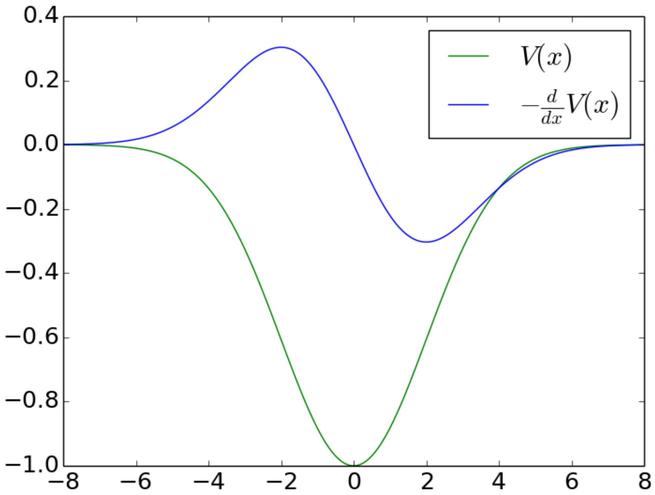

Figure A.2: Graph of the potential energy function $V(x) = -\frac{1}{2}e^{-\beta(x-x_0)^2}$ and the negative derivative of $V(x)$. While the system is within the energy well, movement should tend toward 0. The potential rapidly tends toward 0 outside of $[-4, 4]$. If the system escapes this interval, dynamics will return to standard Brownian motion until the system falls back into the potential well.

Where $\beta \in \mathbb{R}$ is an arbitrarily chosen constant. The potential is a Gaussian potential well, corresponding to the physical idea of a local energy well. The potential has negative derivative (and thus exerts a force proportional to)

$$-\frac{d}{dx}V(x) = -\beta(x - x_0)e^{-\beta(x-x_0)^2}$$

A simulated path is shown in Figure 3, where the trajectory acts like a random walk as observed with Brownian motion, but has a much stronger tendency to fall toward the zero position due to the influence of the deterministic potential energy function. However, when the random fluctuations drive the system away from zero toward positions where the potential is weak, we observe that Brownian motion dominates. This is notably true around the 600 time point where the system has drifted very far away, eventually returning to the potential well by random walk. Additionally, as we expect from statistical mechanics, the histogram of observed positions shown in Figure 3 approximates the shape of the inverse of

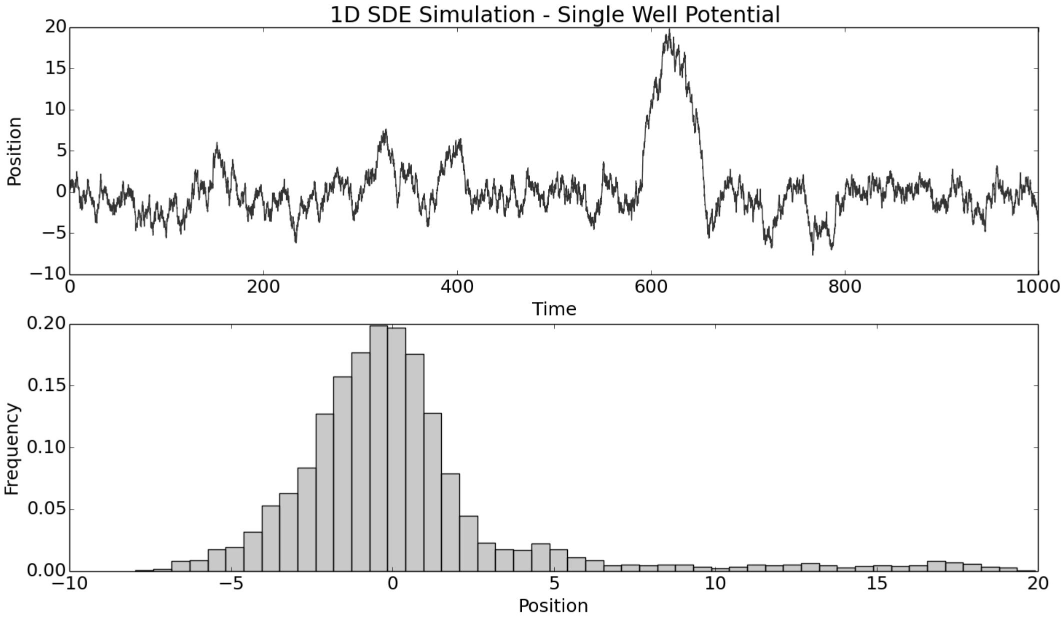

Figure A.3: Simulation of the Langevin system with potential $V(x) = -\frac{1}{2}e^{-\beta(x-x_0)^2}$ showing the trajectory over units of time on the top and the histogram of observed positions on the bottom. The histogram reproduces the negative of the potential energy function near the center of the well at $x = 0$ but the system performs a random walk in areas with little influence from the potential.

the potential energy function.

Next let's consider a more interesting potential function - two wells. Although simple, a two-well system has two identifiable states and is capable of undergoing transitions between the two states. This is the essential behavior we are interested in at the level of protein dynamics.

Let our new potential take the form

$$V(x) = -\frac{1}{2}e^{-\beta(x+3)^2} - \frac{1}{2}e^{-\beta(x-3)^2} + 4x^2$$

This is a simple two well potential with an additional harmonic restraint around zero to prevent the system from wandering away from the basic two-state set-up. The harmonic restraint will essentially act as a 'differentiable box', but is otherwise uninteresting for our purposes. The potential has negative derivative

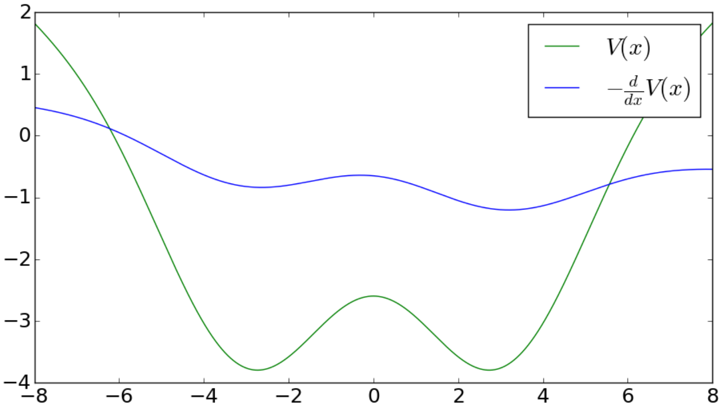

Figure A.4: Graph of the potential energy function $V(x)$ and $-\frac{d}{dx}V(x)$ for the double well potential in a harmonic box. Shown with parameters $\alpha = 8$, $\beta = 1/8$ and $\gamma = 1/32$.

$$-\frac{d}{dx}V(x) = -\beta(x+3)e^{-\beta(x+3)^2} - \beta(x-3)e^{-\beta(x-3)^2} - 8x$$

and the graphs of both potential and negative derivative are shown in Figure 4.

Figure 5 shows a realization of the system over the course of 1000 units of time. In general, the system moves randomly in a narrow region about the potential local minima at $x_0 = -3$ and $x_1 = 3$ with a tendancy to move toward the minima, as we would expect. Occasionally the system jumps between the two wells driven by large random fluctuations. As we observed for the single well system, the histogram of positions shown in Figure 5 is proportional to the inverse of the potential energy function. Unlike the single-well system, we don't observe a random walk region, as the harmonic energy term prevents escape.

A.2 Toy System

In Chapter 2 we described a toy system: a massless particle undergoing 2-D Brownian motion in a potential field. The toy system is Langevin dynamics in 2D, described by

$$\frac{dx}{dt} = -\nabla U(x) + \sqrt{2}\sigma\eta(t)$$

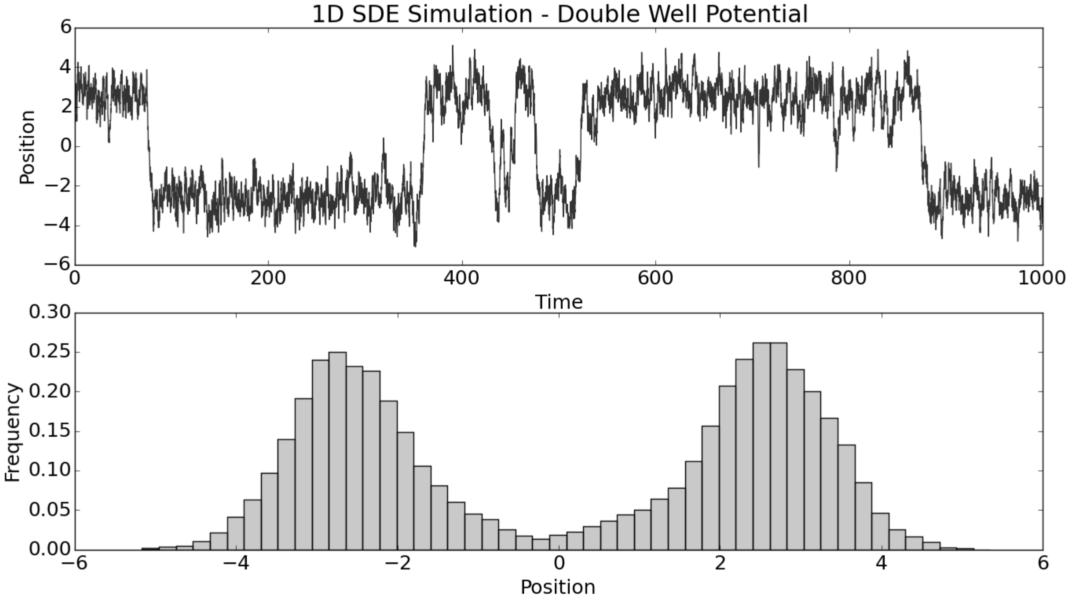

Figure A.5: Single trajectory of double-well SDE with harmonic constraint (top) and histogram of positions visited (bottom). The trajectory moves randomly within the narrow region of each harmonic well, with occasional large random fluctuations causing the system to jump between the wells.

Where $x \in \mathbb{R}^2$, ∇ is the gradient operator, $U(x)$ is the potential energy field, and $\eta(t)$ is a Gaussian process with the conditions that

- $\langle \eta(t) \rangle = 0$
- $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$

Where δ is the Dirac delta function.

The toy system is a first-order stochastic differential equation. The constraints on the Gaussian process imply that the stochastic part of the SDE is pure diffusion with no drift (first constraint) and has independent increments (second constraint). Furthermore, the noise is isotropic, so spatial dependence is constrained to the potential energy term. We can solve the SDE by integrating, and get

$$x(t) - x(0) = \int_0^t -\nabla U(x) ds + \sqrt{2}\sigma \int_0^t dW_s,$$

where $dW_s = \eta(t)ds$ is a standard Wiener measure. Using the Euler-Maruyama approximation, we have the discretized form

$$X_{n+1} - X_n = -\nabla U(X_n)\Delta t + \sqrt{2}\sigma \Delta W_n,$$

where W_n is a standard random normal variable with mean 0 and variance Δt. Since the noise is isotropic, we can draw random 2-vectors by simply drawing two random normal variables. We simulate trajectories in Python with

```python
## Builds an SDE system that steps through a trajectory
import numpy as np

def build_SDE(dims, gradV, sigma):
    # Builds a simple Langevin IVP
    # Assumes a Wiener measure (isotropic white noise)
    # for the random component
    # dims: number of dimensions in the system
    # gradV: function that takes and returns an array of size dims
    # sigma: variance of the noise in the model (a free parameter)
    def sde_model(t0, t1, dt, IC = None):
        t = np.arange(t0, t1, dt)
        rdt = np.sqrt(dt)
        y = np.zeros((t.size, dims))
        if IC:
            y[0] = IC
        else:
            y[0] = np.zeros(dims)

        for i in range(1, t.size): # use xrange in Python 2.7
            y[i] = y[i-1] - dt * gradV(y[i-1] + \
                    sigma * rdt * np.random.normal(loc=0.0, scale=1.0, size=dims)

        return (t, y)
    return sde_model
```

Chapter two described two potential functions, given by different parameterizations of the potential given by $U(x, y)$, defined

$$U(x,y) = \frac{1}{2}\exp\left(-\beta(\kappa(x-x_0)^2 + y^2)\right) - \frac{1}{2}\exp\left(-\beta(\kappa(x-x_1)^2 + y^2)\right) + \alpha(x^2 + y^2),$$

where α, β, and κ are parameters that control the shape of the potential, and x_0 and x_1 control the separation of the two wells on the x-axis. The last term is a harmonic potential for

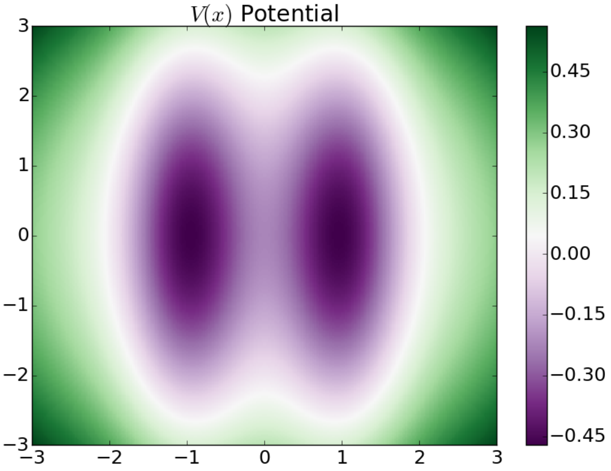

Figure A.6: Graph of the $V(x, y)$ potential.

the purpose of constraining simulations to the region of interest. To perform the simulations, we will also need the gradient of the potential, given by

$$\nabla U(x,y) = \begin{bmatrix} \beta\kappa(x-x_0)e^{-\beta(\kappa(x-x_0)^2+y^2)} + \beta\kappa(x-x_1)e^{-\beta(\kappa(x-x_1)^2+y^2)} + 2\alpha x \\ \beta y e^{-\beta(\kappa(x-x_0)^2+y^2)} + \beta y e^{-\beta(\kappa(x-x_1)^2+y^2)} + 2\alpha y \end{bmatrix}$$

We refer to the $V(x, y)$ potential as the U potential with the parameterization $\alpha = \frac{1}{8}$, $\beta = \frac{1}{2}$, $\kappa = 16$, $x_0 = -1$, and $x_1 = 1$. The V potential is shown in Figure 2.1, as two elliptical potential wells with major axis along the y-axis and separated by a barrier along the x-axis.

Figure 6 shows a colormap of the potential energy surface with the wells centered on the x axis at $x = -1$ and $x = 1$. The wells are anisotropic, with elongations along the y axis and a transition boundary along $x = 0$.

We can inject the gradient into the solver and simulate via

```python
## 2-D SDE Simulation - Anisotropic Double Well and Harmonic Box (V Potential)
x0 = -1
y0 = 0
x1 = 1
y1 = 0

alpha = 1./8
beta = 1./2
kappa = 16

def V(x,y):
    return (-1./2) * np.exp(-1 * beta * (kappa*(x-x0)**2 + (y-y0)**2)) + \
        (-1./2) * np.exp(-1 * beta * (kappa*(x-x1)**2 + (y-y0)**2)) + \
        alpha * ((x)**2 + (y)**2)

def gradV(v):
    x = v[0]
    y = v[1]
    return np.array([
                beta * kappa * (x-x0) * \
                np.exp(-1 * beta * (kappa*(x-x0)**2 + y**2)) + \
                beta * kappa * (x-x1) * \
                np.exp(-1 * beta * (kappa*(x-x1)**2 + y**2)) + \
                2 * alpha * x,

                beta * y * np.exp(-1 * beta * \
                        (kappa*(x-x0)**2 + y**2)) + \
                beta * y * np.exp(-1 * beta * \
                    (kappa*(x-x1)**2 + y**2)) +\
                2 * alpha * y
                ])


sigma = np.sqrt(2)/4
sde = build_SDE(2, gradV, sigma)
t0 = 0
t1 = 200
dt = 0.001

t, results = sde(t0, t1, dt)
```

Figure 7 shows a trajectory of the 2D Langevin system subject to the $V(x, y)$. The trajectory primarily moves within the two wells, as expected, with movement between the wells provided by random fluctuations just as with the 1D system. Figure 8 shows the probability density of the system's positions as estimated by a Gaussian kernel density estimator. Unlike the close mirroring of the energy wells seen in the 1D systems, the kernel density estimation allocates a lot more probability to the left energy well. This is due to

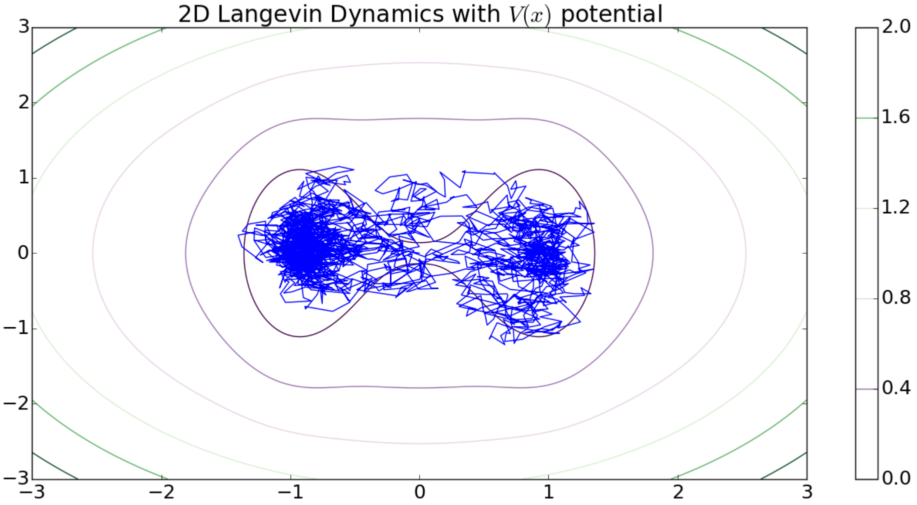

Figure A.7: Trajectory of a single 200 time unit simulation of the toy system with the $V(x, y)$ potential with $\sigma = 1/4$. Trajectory is overlaid on contours of the potential energy surface.

insufficient sampling of the system, as the wells are perfectly symmetric and should have equal occupancy in the limit of infinite data. As discussed in Chapter 2, this is one form of the Curse of Dimensionality, occuring here in the increased sampling needed to explore both wells in a 2D setting, as well as the increased computational cost of simulation. Nevertheless, the kernel density estimator accurately captures the topology of the wells and their transition region, which is ultimately the main information we want to extract in the full setting of protein dynamics.

We refer to the second toy system potential as the $W(x, y)$ potential energy function. The $W(x, y)$ potential is defined as the U potential field with parameterization $\alpha = \frac{1}{8}$, $\beta = \frac{1}{2}$, $\kappa = 48$, $x_0 = -0.25$, and $x_1 = 0.25$. This potential is very nearly identical to V, with the potential wells elongated along the y-axis and brought closer together, yet still separated by a barrier along the x-axis.

Simulation of the $W(x, y)$ system is handled identically to the $V(x, y)$ system, swapping out the gradient function for the gradient of $W(x, y)$, which is just a change of the constant parameters in the python code. A trajectory of the $W(x, y)$ system is shown in Figure A.10 which shows similar behavior as the $V(x, y)$ system.

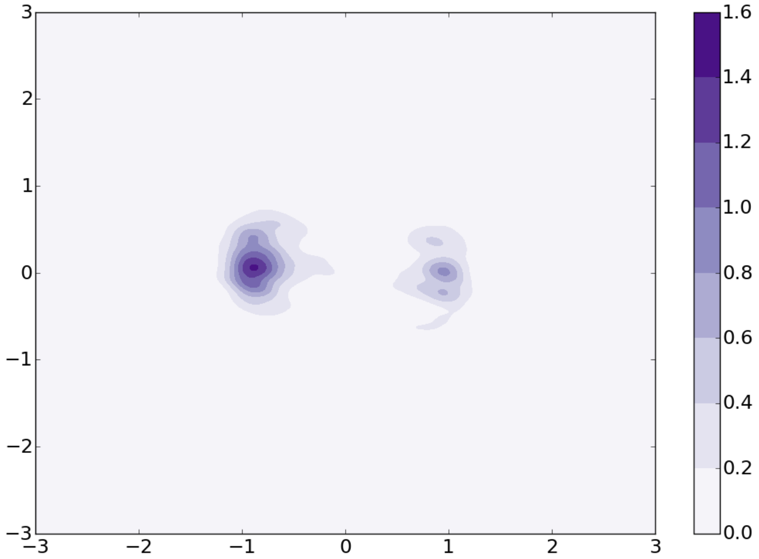

Figure A.8: 2D probability density of the positions occupied by the trajectory of the simulation with the $V(x, y)$ potential and $\sigma = 1/4$. The density was generated by kernel density estimation using a Gaussian kernel and Scott's rule. The density resembles the potential energy surface, but finite sampling error has caused the right well at $x = 1$ to be undersampled compared to the left well at $x = -1$.

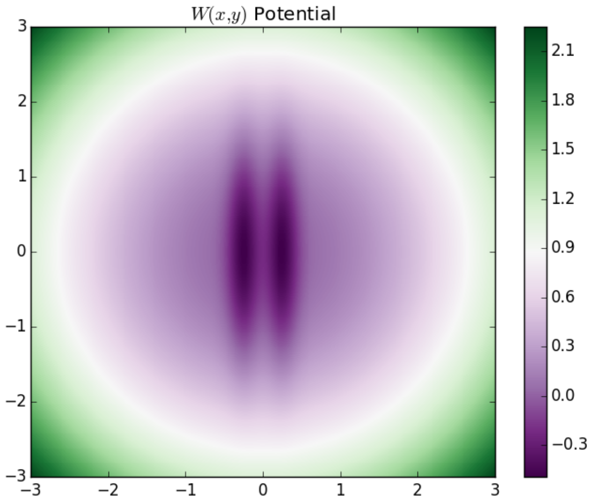

Figure A.9: Graph of the $W(x, y)$ potential energy function.

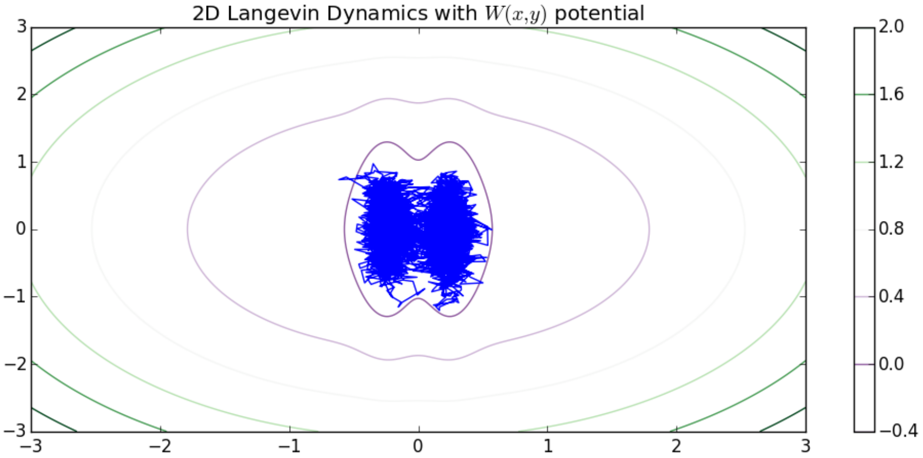

Figure A.10: Trajectory of a single 200 time unit simulation of the toy system with the $V(x, y)$ potential with $\sigma = 1/4$. The trajectory is overlaid on a contour plot of the potential energy surface.

APPENDIX B

NONLINEAR DIMENSIONALITY REDUCTION AND MOLECULAR DYNAMICS

Chapter 2 presented two dimensionality reduction techniques, Principal Component Analysis and time-structured Independent Component Analysis. Both of these methods fall into the class of linear dimensionality reduction techniques, methods that transform a data set to a lower dimension, typically taking data with values in $\mathbb{R}^n$ to values in $\mathbb{R}^m$ with $m < n$. The fundamental goal of dimensionality reduction techniques is reduce the dimension of the space while losing as little information as possible, ideally the dimensions that are discarded are only noise. This is useful in both the context of exploratory data analysis, where an investigator is working with the data to determine patterns or using automated data mining tools, but is also useful when applying other machine learning techniques. Many machine learning methods are brittle to noise and overfitting, and similarity and distance metrics tend to behave poorly in high dimensions, as described at the end of Chapter 2 and in Appendix B. Reducing the dimensionality to 2 or 3 can be particularly valuable in exploratory analysis, where visualization is possible – though subject to bias, the human visual system is extraordinarily powerful and direct visualization of the data can be cruicial to investigator understanding of the data and interpretation. Of purely practical value, dimensionality reduction also reduces the workload of future processing by reducing the size of that data that needs to be processed. Dimensionality reduction is thus vital to working with high dimensional data.

Linear techniques reduce the dimension by projecting into a linear subspace of the original data space. The reduced space has a direct and clear connection to the original space, often improving the intuitive meaning of results, and making it simple to project new data points

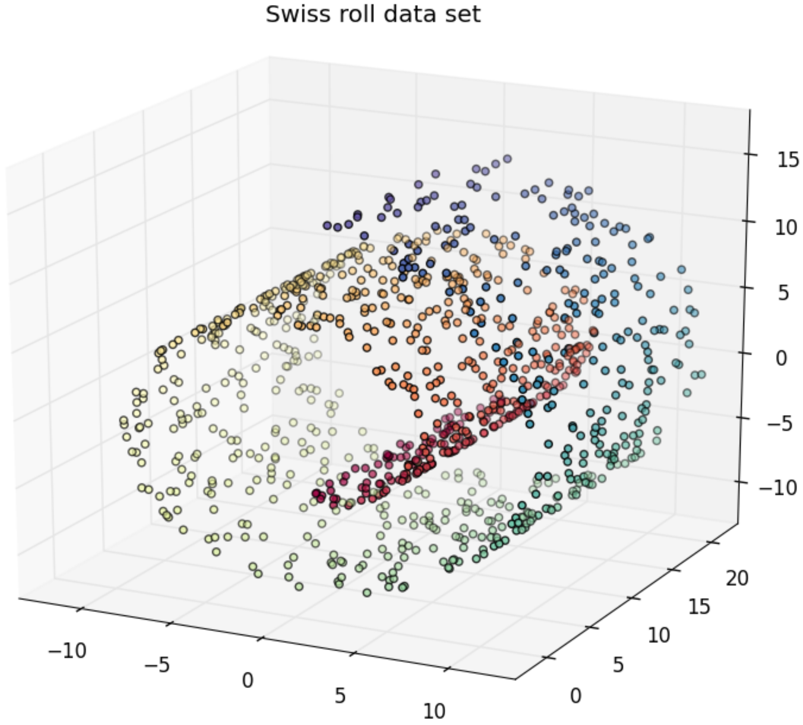

Figure B.1: The Swiss Roll, a canonical example of a 2D data set non-linearly embedded in 3D Euclidean space. Data generated using scikit-learn datasets generator.

onto the reduced space by simply applying the linear projection. When appropriate, linear techniques are preferred; they are robust and generally very fast.

Unfortunately, data is often not so accomodating. Often it is the case that data is generated by a process that lives on a manifold, or something even more complex, and does not neatly decompose into linear combinations of the variables studied. As a simple example, consider the classic Swiss Roll data set, which is the 2D plane rolled up, and which is presented in standard Euclidean space; an example is shown in Figure 5.1. This is clearly a 2D data set, however because it embedded in 3D space in a manner that cannot be linearly projected, linear dimensionality reduction techniques fail to capture the shape of the data.

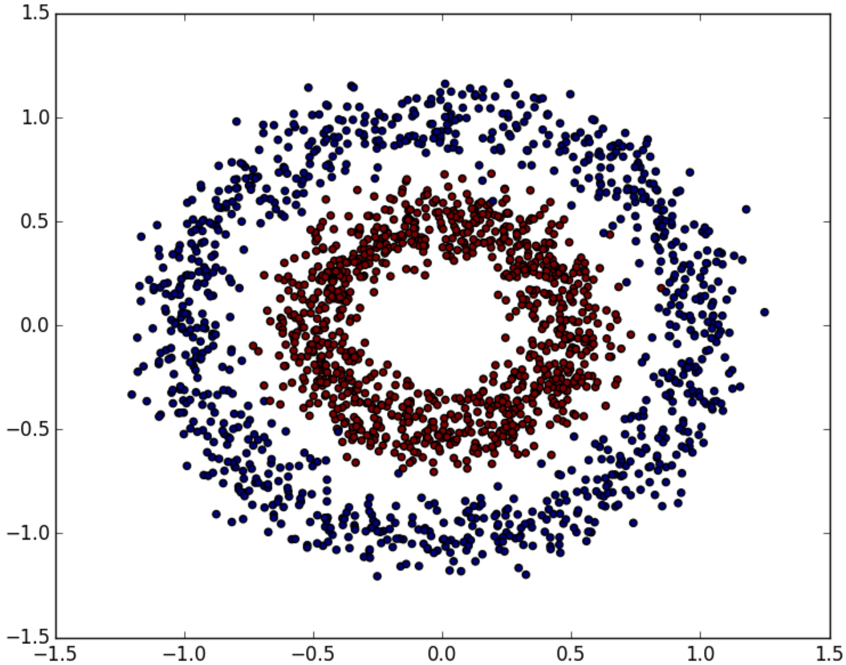

Figure B.2: Data drawn from two clearly distinguishable circular distributions that are not linearly separable in the plane, and hence cannot be clustered by a linear clustering method such as k-means. Data generated using scikit-learn datasets generator.

The mathematical details of manifolds are deep and outside the scope of this text, but for understanding the algorithms laid out here, it suffices to understand a manifold to be a space where for any given point, we can find a small region around that point where everything behaves like Euclidean space.

A similar problem occurs in clustering, as mentioned in Chapter 2, where clusters are grouped in a manner that is not linearly seperable, such as the data set consisting of two rings shown in Figure 5.2. Non-linear dimensionality reduction techniques are applicable here as well, as the methods search for non-linear transforms of the data – simply setting the target output dimension equal to the input dimension can find a non-linear transform of the data the matches the 'shape' of the data, so that the output becomes linearly separable.

Given the complexity of molecular dynamics data, it comes as no surprise that non-linear dimensionality reduction appears a useful tool, though the literature on the topic is somewhat sparse, perhaps indicative of the somewhat esoteric combination of disciplines involved. In practice, we find that the promises of non-linear dimensionality reduction do not hold particularly well. This chapter details thus details some non-linear dimensionality reduction methods, and the misadventures experienced applying them to MD data on the 2C T cell receptor, culminating in a theoretical proposal for future work that might do better. Previous work has been done before using Isomap, Locally Linear Embeddings, and Autoencoders to study the reconstruction error of an 8-member ring, finding Autoencoders to be the most effective overall[1].

B.1 Isomap

Isomap is an extension of Multidimensional scaling (MDS), the simplest form of which is PCA as described in chapter 2. Isomap augments MDS by changing the distance metric of MDS, rather than compute a direct pair-wise distance matrix, Isomap computes an approximation to the geodesic distance between points as determined by the underlying manifold that generates the data. Since the data is presumed to lie on a manifold structure, the geodesic distance between nearby points is approximately the standard Euclidean metric. A graph is constructed taking the data points as vertices, and edges are drawn between each vertex and it's k nearest neighbors, where k is an input parameter of the algorithm. The edge weights are the Euclidean distance between the data points. From this graph, the geodesic distance between two points is approximated by the shortest path distance between the two vertices representing those data points on the graph. This construction empirically approximates the geodesic distance as the path passes through the different local maps making up the manifold atlas.

The main weakness of Isomap is 'short-circuiting', in which two points are joined by the nearest neighbor search that should be separated due to noise or too large of a selection of

[1] M.W. Brown et al.: Algorithmic dimensionality reduction for molecular structure analysis, in: J. Chem. Phys. 2008.

k relative to the data density. This creates an incorrectly short path between two distinct regions of the manifold, and warps the geodesic distance measurements, potentially of much of the data. This has two major downsides, one is straightforward susceptibility to noise in the data, the other is more subtle: if the data is not evenly sampled at each region, then the choice of k has to follow the most poorly sampled region. Furthermore, short-circuiting can result from a single noisy datapoint, making it particularly weak to the presence of outliers. Somewhat unsurprisingly, Isomap performs rather poorly for MD data; much of the machinery of the MSM approach is specifically dedicated to avoiding linking kinetically unrelated regions, while small perturbations can cause Isomap to ignore kinetic barriers.

B.2 Locally Linear Embedding

Locally linear embedding is conceptually like applying PCA to small patches of the data, finding a linear projection of each path and then gluing these locally linear patches together to transform the whole data set. The idea arises from the fact that PCA has historically proven to be an effective tool for linear dimensionality reduction, and manifolds behave locally like Euclidean space, so for a small enough region around any given data point, PCA finds a linear projection that well approximates the manifold structure. LLE is surprisingly effective for MD data; the major problem encountered in applications to TCR data was brittleness with respect to input parameter perturbation. The output results can change, sometimes wildly, with only small changes in input parameters; future use demands validation and parameter selection methods to make analysis practical and assure the investigator against spurrious results.

An intriguing idea is to consider LLE as a kernel method, in which it may be possible to design a variant of LLE that performs a local tICA calculation rather than PCA. Doing so may keep the valuable aspects of tICA for molecular dynamics analysis, while allowing for non-linear transformations.

B.3 Diffusion Maps

Diffusion maps bears similarity to a combination of Isomap and a simplified version of the MSM machinery; diffusion maps approximates a geodesic distance between data points by constructing a graph, as Isomap does, but computes distances as the result of a random walk on the data. This renders diffusion maps more robust to noise and immune to the outlier effect that plagues Isomap – a short-circuiting outlier only creates a single low-probability pathway, and so doesn't contribute much weight to distance measurement, which is a weighted average over the paths between two points.

Diffusion maps struggles with MD data because MD data has locally varying spatial structure, which diffusion maps doesn't capture as it uses a fixed size kernel across all data. The Markov State Model method finds multi-scale models by using the microstate clusters as a fine discretization of the phase space, followed by adjustable coarse-graining of that data, as well as allowing for varying time-lags. This flaw of diffusion maps for molecular dynamics was addressed by adaptively varying the kernel size based on the data, however, the resulting method is computationally expensive, indeed early implementations by the thesis author showed several orders of magnitude more computation time on small test sets than other methods. Combined with super-linear scaling, the method, although promising, is too expensive for practical use, particularly as the computation resources required are of a magnitude that they would be more likely to be better spent on producing more data and using a simpler method. In the limit of sufficient data, nearly any method will eventually suffice, and certainly the combination of tICA and MSMs are highly effective in the data-rich regime, while consisting of an efficient processing pipeline.

B.4 Autoencoders

Autoencoders are a form of Artificial Neural Network, where the network attempts to learn the identity function to transform the data – the network is trained to output its input.

The network learns important features of the data through one or both of two techniques, reduced hidden layer size or regularization, both of which have the effect of forcing the network to learn an encoding and decoding of the data. After training, the decoding layers are stripped from the network, and the encoding layer can be used as a feature extraction or dimensionality reduction tool, and have shown effective results in other fields such as document analysis[2]. When applied to MD data, autoencoders showed the best reconstruction results compared to Isomap, LLE, and PCA[3].

B.4.1 Connections to tICA

Second-order independent component analysis, that is, tICA, in k dimensions can be represented by a recurrent neural network with k neurons[4]. Recurrent neural networks are of course significantly more powerful than the feed-forward network model of an Autoencoder, however, denoising autoencoders are known to be capable of manifold learning[5], and feed-forward architectures have been used to model time-delay learning using input layers divided into ’present’ and ’future’ inputs with excellent results in video using convolutional networks[6].

[2] G.E. Hinton/R.R. Salakhutdinov: Reducing the Dimensionality of Data with Neural Networks, in: Science 313 (2006).

[3] Brown et al.: Algorithmic dimensionality reduction for molecular structure analysis (see n. 1).

[4] L. Molgedey/H.G. Schuster: Separation of a Mixture of Independent Signals Using Time Delayed Correlations, in: Physical Review Letters 72.23 (1994).

[5] P. Vincent et al.: Extracting and Composing Robust Features with Denoising Autoencoders, tech. rep. 1316, Université de Montrèal.

[6] A. Karpathy et al.: Large-Scale Video Classification with Convolutional Neural Networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

APPENDIX C
SIMULATION AND ANALYSIS DETAILS

This appendix briefly covers the specifics of the MD simulations used to produce the data set in this work, and the analysis parameters.

C.1 Molecular dynamics simulations

All simulations were carried out using the Amber14 package. Input coordinates were prepared from PDB files 1TCR (2C) and 2EYS (NKT15), truncated to the variable domains and prepared using pdb4amber processing scripts. These structures were solvated with TIP3P waters in an octohedral unit cell at 12 angstroms, neutralized with NaCl at 150mM concentrations, and parameterized using the AMBER99SB forcefield and Joung/Cheatham ion parameters using xleap. Two rounds of 2000 steps of minimization were carried out, first with restraints on the protein, and then secondoly without restraints. These minimized states were the initial seeds for each of the ten trajectories run our for each of 2C and NKT15. Each trajectory was set to 300K through initial velocity randomization, and allowed to equilibrate in NPT for 10ns using a Langevin thermostat ($\gamma = 1$) and the Amber Monte Carlo barostat at 1 atm, allowing the trajectories to diverge independently. All data presented in the analysis was collected following the 10ns equilibration stage, with each trajectory run for an additional 300ns (2C) or 100ns (NKT15) using SHAKE to allow 2fs timesteps. Calculations were performed using the CUDA-enhanced pmemd Amber module on the University of Chicago's Midway cluster, utilizing either K20 or K40 Tesla GPUs.

C.2 Data processing and dimensionality reduction

Raw simulation data was processed using cpptraj to re-image the system and extract protein data. Structure alignments and RMSD calculations were carried out using VMD. Hydrogen bonds were determined with a 3.2 angstrom distance cutoff and 20 degree angle cutoff in VMD. Further data processing used custom Python scripts with trajectory featurization and data handling provided by the MDTraj library. We used the MSMBuilder3 library to perform tICA analysis, clustering, and Markov state model generation as described in their sections. Kernel density estimates were calculated using the gaussian_kde module from the Scipy library; kernel bandwidth was selected automatically using Scott's Rule.

C.3 Markov state model construction

Our analysis follows the procedure outlined in Chapter 3 for building MSMs from MD data. Our initial data was taken from the MD simulation by extracting the CDR3α and CDR3β loops as independent datasets. The datasets were featurized as dihedral angles, so that the actual analyzed data is the set time-series of ϕ and ψ angles for each of CDR3α and CDR3β. Taking the dihedral angles eliminates variation due to whole protein motion and minimizes the effect of individual domain drift. The tICA decomposition and projection were applied to these ϕ/ψ angle time-series. The first 16 degrees of freedom determined by the tICA decomposition were used for clustering; we applied the k-medoids algorithm as described in Chapter 2 using the Euclidean distance metric after the tICA projection. The first 16 degrees of freedom were chosen as they account for $> 90\%$ of the energy of the eigenvalues (eigenvalues are shown in Figure C.1).

After clustering, a microstate Markov models are generated by estimation of a state transition matrix using Maximum Likelihood Estimation (MLE). The estimator used a sliding window to maximize the data available for estimation; the algorithms used in MSMBuilder3 correct for the non-independence of the sliding window. Fictional transition counts were added to smooth numerical issues, on the order of 0.1, yielding minimal perturbation of the model but preventing extensive noise as the timelag varies. Models were constructed over a series of timelags as shown in the implied timescale analysis in Chapter 4 (figure 4.2). The final choice of timelag was determined by finding the smallest timelag where the slowest degrees of freedom showed convergence to a local stable value, graphically where the curves appear to flatten out.

The final macrostate Markov model, which is described in Chapter 4, was extracted from the microstate model by Perron Cluster Clustering Analysis, as detailed in Chapter 3. PCCA serves to construct the macrostate clusters themselves, acting a clustering algorithm for the data that relies on the kinetics of the microstate model. The transition matrix for the macrostate model was estimated directly from the time-series of the macrostate cluster assigned data frames, and the microstate model is discarded after the macrostate cluster assignments are generated.

In the models presented, CDR3α data was clustered into 16 clusters and CDR3β data was clustered into 32 clusters for building the microstate models. The numbers of clusters chosen were those that showed good coverage of the state space spanned by the two-dimensional projection of the data under tICA, and therefore could be inspected visually, and were low enough to provide reasonable statistics and convergence in building the microstate MSMs. Despite this, CDR3α did not show convergence under any selection of clustering parameters; the simplest interpretation is insufficient data to capture the very slow motions of the CDR3α dynamics. Macrostate assignment was via the PCCA+ algorithm, which is a more robust variant of the standard PCCA algorithm. The CDR3β microstate model was built with an 8 nanosecond timelag, and clustered into four macrostates as indicated by inspection of the implied timescales graph (figure 4.2B) and the four local maxima that appear in the kernel density estimate of the data projected under the first two degrees of freedom of tICA (figure 4.1).

BIBLIOGRAPHY

Allen, M. P. and D. J. Tildesley: Computer Simulation of Liquids, 1989.

Aloise, D. et al.: NP-hardness of Euclidean sum-of-squares clustering, in: Machine Learning 2009.

Amdahl, Gene M.: Validity of the single processor approach to achieving large scale computing capabilities, in: AFIPS spring joint computer conference, 1967.

Archbold, J.K. et al.: Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition, in: J. of Exp. Med. 2009.

Armstrong, K.M., F.K. Insaidoo, and B.M. Baker: Thermodynamics of T-cell receptor-peptide/MHC interactions: progress and opportunities, in: Journal of Molecular Recognition 104 (2008).

Baker, Brian M. et al.: Structural and dynamic control of T-cell receptor specificity, cross-reactivity, and binding mechanism, in: Immunological Reviews 250 (2012), URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1600-065X.2012.01165.x/full.

Beyer, Kevin et al.: When Is "Nearest Neighbor" Meaningful?, in: vol. 1540 (Lecture Notes in Computer Science), 1999, pp. 217–235.

Blanchard, Philippe and Erwin Brüning: Mathematical Methods in Physics, vol. 69 (Progress in Mathematical Physics), 2015.

Boniface, J.J. et al.: Thermodynamics of T cell receptor binding to peptide-MHC: evidence for a general mechanism of molecular scanning, in: Proceedings of the National Academy of Sciences of the United States of America 96 (1999).

Borbulevych, O. Y. et al.: T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility, in: Immunity 31 (6 2009), URL: http://www.sciencedirect.com/science/article/pii/S1074761309004981.

Borbulevych, O.L., K.H. Piepenbrink, and B.M. Baker: Conformational melding permits a conserved binding geometry in TCR recognition of foreign and self molecular mimics, in: J. Immunol. 2011.

Brown, M.W. et al.: Algorithmic dimensionality reduction for molecular structure analysis, in: J. Chem. Phys. 2008.

Case, D.A. et al.: Amber14, 2014, URL: ambermd.org.

Colf, L.A. et al.: How a single T cell receptor recognizes both self and foreign MHC, in: Cell 129 (2007).

Dai, S. et al.: Crossreactive T Cells spotlight the germline rules for alphabeta T cell-receptor interactions with MHC molecules, in: Immunity 2008.

Degano, M. et al.: A functional hot spot for antigen recognition in a superagonist TCR/MHC complex, in: Immunity 12 (2000).

Deuflhard, P. and M. Weber: Robust Perron cluster analysis in conformational dynamics, in: Linear Algebra and its Applications 2005.

Fabian, H. et al.: HLA-B27 subtypes differentially associated with disease exhibit conformational differences in solution. In: J. Mol. Biol. 2008.

Francois, P. et al.: Phenotypic model for early T-cell activation displaying sensitivity, specificity, and antagonism, in: PNAS 2013.

Fraser, Christopher M., Ariel Fernandez, and L. Ridgway Scott: Dehydron analysis: quantifying the effect of hydrophobic groups on the strength and stability of hydrogen bonds, in: (Advances in Computational Biology), 2010, pp. 473–479.

Idem: Wrappa: A screening tool for candidate dehydron identification, tech. rep. TR-2011-5, University of Chicago, 2011.

Gagnon, S.J. et al.: T cell receptor recognition via cooperative conformational plasticity, in: Journal of Molecular Biology 363 (2006).

Garcia, K.C. and E.J. Adams: How the T cell receptor see antigen - a structural view, in: Cell 122 (2005).

Garcia, K.C. et al.: An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. In: Science 274 (1996), pp. 209–219.

Idem: Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen, in: Science 1998.

Germain, R.N. and I. Stefanová: The Dynamics of T Cell Receptor Signaling: Complex Orchestration and the Key Roles of Tempo and Cooperation, in: Annu. Rev. Immunol. 1999.

Govern, C. C. et al.: Fast on-rates allow short dwell time ligands to activate T cells, in: Proceedings of the National Academy of Sciences of the United State of America 107 (19 2010), URL: www.pnas.org/cgi/doi/10.1073/pnas.1000966107.

Hawse, W.F. et al.: TCR scanning of peptide/MHC through complementary matching of receptor and ligand molecular flexibility, in: J Immunol 192 (2014), pp. 2885–2891, URL: http://www.jimmunolo.org/content/192/6/2885.

Hinton, G.E. and R.R. Salakhutdinov: Reducing the Dimensionality of Data with Neural Networks, in: Science 313 (2006).

Holler, P.D. and D.M. Kranz: T cell receptors: affinities, cross-reactivities, and a conformer model, in: Molecular Immunology 40 (2004).

Hopfield, J.J.: Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity, in: Proceedings of the National Academy of Sciences of the United States of America 1974.

Huseby, E.S. et al.: How the T cell repertoire becomes peptide and MHC specific, in: Cell 2005.

Janin, J.: The Kinetics of Protein-Protein Recognition, in: Proteins: Structure, Function, and Genetics 1997.

Jolliffe, I.T.: Principal Component Analysis, 2002.

Karpathy, A. et al.: Large-Scale Video Classification with Convolutional Neural Networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

Kaufman, L. and P.J. Rousseeuw: Clustering by means of Medoids, in: Statistical Data Analysis Based on the L_1-Norm and Related Methods (Statistics for Industry and Technology), 1987.

Kelly, F.P.: Reversibility and Stochastic Networks, 1979.

Kjer-Nielsen, L. et al.: A structural basis for selection and cross-species reactivity of the semi-invariant NKT cell receptor in CD1d/glycolipid recognition, in: J. Exp. Med. 2006.

Klein, L. et al.: Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see), in: Nature Reviews Immunology 2014.

Knapp, B., J. Dunbar, and Deane C.M.: Large Scale Characterization of the LC13 TCR and HLA-B8 Structural Landscape in Reaction to 172 Altered Peptide Ligands: A Molecular Dynamics Simulation Study, in: PLoS Computational Biology 10 (8 2014), URL: `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003748`.

Li, Y. et al.: The Vα14 invariant natural killer T cell TCR forces microbial glycolipids and CD1d into a conserved binding mode, in: J. Exp. Med. 2010.

Lloyd, S.P.: Least squares quantization in PCM, in: IEEE Transactions on Information Theory 1982.

Lopez-Sagaseta, J. et al.: The molecular basis for MAIT cell recognition of MR1, in: Proceedings of the National Academy of Sciences of the United States of America 2013.

Mason, Don: A very high level of crossreactivity is an essential feature of the T-cell receptor, in: Immunology Today 19.9 (1998), pp. 395–404, URL: `http://dx.doi.org/10.1016/S0167-5699(98)01299-7`.

McGibbon, R.T. and V.S. Pande: Learning Kinetic Distance Metrics for Markov State Models of Protein Conformational Dynamics, in: J. Chem. Theory Comput. 2013.

McGibbon, R.T. et al.: Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models, in: Proc. 31st Intl. Conf. on Machine Learning, 2014.

McKeithan, T.W.: Kinetic proofreading in T-cell receptor signal transduction, in: Proceedings of the National Academy of Sciences of the United States of America 1995.

Molgedey, L. and H.G. Schuster: Separation of a Mixture of Independent Signals Using Time Delayed Correlations, in: Physical Review Letters 72.23 (1994).

Mombaerts, P. et al.: RAG1-deficient mice have no mature B and T lymphocytes, in: Cell 1992.

Noé, F. and S. Fischer: Transition networks for modeling the kinetics of conformational change in macromolecules, in: Current Opinion in Structural Biology 2008.

Pande, Vijay S., Kyle Beauchamp, and Gregory R. Bowman: Everything you wanted to know about Markov State Models but were afraid to ask, in: Methods 52 (1 2010), pp. 99–105.

Park, S. and V.S. Pande: Validation of Markov state models using Shannon's entropy, in: J. Chem. Phys. 2006.

Pellicci, D.G. et al.: Differential recognition of CD1d-alpha-galactosyl ceramide by the V beta 8.2 and V beta 7 semi-invariant NKT T cell receptors, in: Immunity 2009.

Prinz, J.-H. et al.: Markov models of molecular kinetics: Generation and validation, in: J. Chem. Phys. 134 (2011).

Resier, J.B. et al.: CDR3 loop flexibility contributes to the degeneracy of TCR recognition, in: Nature Immunology 4 (2003).

Robey, E. and B.J. Fowlkes: Selective events in T cell development, in: Annu. Rev. Immunol. 1994.

Roldan, E.Q. et al.: Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells, in: Immunogenetics 1995.

Rossjohn, J. et al.: Recognition of CD1d-restricted antigens by natural killer T cells, in: Nature Reviews Immunology 12 (2012), pp. 845–857.

Rudolph, M.G., R.L. Stanfield, and I.A. Wilson: How TCRs bind MHCs, peptides, and coreceptors, in: Annual Review of Immunology 24 (2006).

Sandstrom, A. et al.: The B30.2 domain of Butyrophilin 3A1 binds phosphoantigens to mediate activation of human Vγ9Vδ2 T cells, in: Immunity 2014.

Sarich, M., J-H Prinz, and C. Schütte: Markov Model Theory, in: (An Introduction to Markov State Models and Their Applications to Long Timescale Molecular Simulation), 2014.

Schwantes, C.R. and V.S. Pande: Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9, in: J. Chem. Theory Comput. 9 (2013), pp. 2000–2009.

Scott, D.R. et al.: Disparate degrees of hypervariable loop flexibility control T-cell receptor cross-reactivity, specificity, and binding mechanism, in: Journal of Molecular Biology 414 (2011).

Idem: Limitations of time-resolved fluorescense suggested by molecular simulations: assessing the dynamics of T cell receptor binding loops, in: Biophysical Journal 103 (2012).

Sewell, A.K.: Why must T cells be cross-reactive?, in: Nature Reviews Immunology 12 (2012), pp. 669–677, URL: http://www.nature.com/nri/journal/v12/n9/abs/nri3279.html.

Sezer, D. and B. Roux: Markov State and Diffusive Stochastic Models in Electron Spin Resonance, in: (An Introduction to Markov State Models and Their Applications to Long Timescale Molecular Simulation), 2014.

Shinkai, Y. et al.: RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement, in: Cell 1992.

Smyth, M.R.F.: A Spectral Theoretic Proof of Perron-Frobenius, in: Mathematical Proceedings of the Royal Irish Academy 2002.

Stone, J.D., A.S. Chervin, and D.M. Kranz: T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity, in: Immunology 2009.

Strook, Daniel W.: An Introduction to Markov Processes, vol. 230 (Graduate Texts in Mathematics), 2005.

Vincent, P. et al.: Extracting and Composing Robust Features with Denoising Autoencoders, tech. rep. 1316, Université de Montrèal.

Wu, L.C. et al.: Two-step binding mechanism for T-cell receptor recognition of peptide-MHC, in: Nature 418 (2002).

Wun, K.S. et al.: Ternary crystal structure of the human NKT TCR-CD1d-C20:2 complex, in: J. Biol. Chem. 2012.