ELUCIDATING FUNCTIONAL MECHANISMS IN PROTEINS USING THE DISTANCE CONSTRAINT MODEL

by

Christopher Singer

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2019

Approved by:

_____
Dr. Donald Jacobs

_____
Dr. Irina Nesmelova

_____
Dr. Jun-tao Guo

_____
Dr. Cynthia Gibas

_____
Dr. Jerry Troutman

ABSTRACT

CHRISTOPHER SINGER. Elucidating functional mechanisms in proteins using the Distance Constraint Model. (Under the direction of DR. DONALD JACOBS and DR. IRINA NESMELOVA)

Protein thermodynamics has been shown to be insightful and illuminating to the functional mechanisms behind biological function. By characterizing proteins and protein families according to their enthalpy and entropy we can make inferences to the relationship between specific residues or regions and the activity. Here I will leverage several existing and new computational methodologies on various families of biological importance.

By combining molecular dynamics calculations with thermodynamics I will characterize the importance of disulfide bonds in CXCL7 proteins, which are a trademark characteristic of the Chemokine family. In addition I will use these calculations to approximate the dimerization energy. RiVax being a larger molecule is too computationally expensive to run MD simulations on the scale desired therefore I will create a new higher throughput methodology to better represent the conformational diversity of RiVax. Through better conformational sampling and quantitative stability/flexibility relationships I will elucidate why certain mutations enhance the stability of RiVax making it a more viable drug. Sleeping Beauty presents another set of challenges, as experimental data is scarce for this import link in the CRISPR technology chain. Therefore I will develop a methodology to predict thermodynamic properties of SB without the need for heat capacity data. I will accomplish this by comparing SB to its hyperactive mutation SB100 that is structurally very similar yet has difference in activity

equal to 100 fold.

These new methodologies will enhance the capacity of current methods and build the foundation for future students to develop further, and in time tackle problems outside the scope of current ability.

ACKNOWLEDGEMENTS

As a preface to the technical sections of this dissertation I would like to take this opportunity to thank several amazing people and organizations that offered unwavering support and encouragement throughout this process. Without these people this dissertation would not have come to fruition.

First and foremost my dissertation advisor and committee chair Dr. Donald Jacobs. You were the driving force behind igniting my scientific curiosity as an undergraduate, and became a mentor that reached far past the walls of the classroom. Dr. Irina Nesmelova, your pragmatic approach to science kept me grounded to interesting yet solvable problems rather than purely theoretical curiosities. Without continuous intervention from the two of you it is clear to me that this dissertation would not and could not have happened. I owe a great deal of my successes, abilities, and personality to both of you, and these gifts shall carry me through decades to come. To you I am eternally grateful.

Dr. Cynthia Gibas, even before I became one of your student you were in my corner. I have you to thank for my acceptance into the GAANN Fellowship and I am grateful for your mentorship as a teaching fellow. It is because of this experience that I was able to shift into my career as a Data Scientist.

In reality I have hundreds of people who helped me reach this peak none of which played an negligible role, however here are just a few of them. Dr. Jun-tao Guo and Dr. Jerry Trouman for siting on my committee. Dr. Jenny Farmer and the rest of the BMPG present and former for countless hours of technical help. The Department of Physics and

Optical Science as well as the Department of Bioinformatics and Genomics, and the Graduate School GAANN Fellowship for funding throughout this process as a teaching assistant.

DEDICATION

This dissertation is dedicated to my loving wife Samantha. Thank you for all of your love, support, and motivation over this long journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NMR	Nuclear Magnetic Resonance
mDCM	minimum Distance Constraint Model
MD	Molecular Dynamics
CD	Circular Dichroism
DOF	Degrees of Freedom
FI	Flexibility Index
CC	Correlated Cooperativity
SBNR	Signal Beyond Noise Ratio
sDCM	solvation Distance Constraint Model

Chapter 1 Introduction

1.1 Proteins: The Workforce of the Cell

The directions encoded in our DNA dictate biological function, however proteins are the foot soldiers that carry out the orders sent down from DNA. Molecular Biologists refer to this dissemination of orders from DNA transcription to mRNA, and mRNA translation into proteins as the Central Dogma of Molecular Biology and it represents the multi tiered processes of which life works. Proteins take many forms inside the cell covering structural complexes, which builds our bones and muscles, transport proteins like Hemoglobin which carries oxygen throughout our bodies, signaling proteins that start the chain reaction of cellular processes to maintain homeostasis.

Proteins vary greatly in their structure and sequence. Across species it can be the case that proteins share similar function yet their sequences can differ by more than 50%. On the other hand, a single point mutation, having little effect on sequence similarity compared to the wild type, can have dramatic consequences to protein function, which can lead to diseases. Sickle cell anemia is an unfortunate example of this, where it is known that a mutation from GLU to VAL changes the protein-protein interaction in the red blood cells causing hemoglobin to stick together forming extended patterns. This is extremely painful and completely alters the natural function of an entire cell type (1). Cystic Fibrosis, Neurofiboritois, and even some forms of cancer are linked to a single change to the amino acid structure of the acting protein (2).

The birth of structural biology, followed by structural bioinformatics, began with the first three dimensional structure published in nature in 1958(3). This first structure was

myoglobin obtained by x-ray crystallography, which in addition to Nuclear Magnetic Resonance (NMR) are the two gold standards for experimental protein structure determination even today. These structures give researchers a snapshot of the folded conformation of now thousands of proteins that represent the widest array of functional characteristics. Researchers now have a sequence/structure/function paradigm to help prevent and/or cure major diseases causes by mutations along the amino acid sequence. However, for such a molecular level perspective to be useful, the first steps are to characterize proteins experimentally and build models to glean insight in how each residue interacts with and without each other. From a bioinformatics perspective this means how to collect, and process data and develop tools and pipelines to understand the effects of mutations on protein function.

1.2 Protein Thermodynamics

The laws of thermodynamics govern every process in our universe, and the biological world is no different. Free energy (G) of a system is related to enthalpy (H) and entropy (S) according to (Eq 1).

$$G = H - TS \qquad (1)$$

Enthalpy accounts for all energy contributions from molecular interactions and a work contribution that is related to the system changing its volume. However, near atmospheric pressure this work term from changes in volume is a negligible contribution to enthalpy, meaning we can consider the enthalpy of a system as energy. A simple expression for the energy of a molecular system can be written as one and two body terms as shown in (Eq 2).

$$H = \sum_{i=1}^{n} E_{interal} + \sum_{j=i+1}^{n} Eij \qquad (2)$$

Entropy on the other hand represents how energy is distributed throughout the system among all the possible interactions within the system. Systems tend to increase in entropy by statistical reasons alone, driving systems to become more disordered. While these two properties themselves may be difficult to measure directly we can accurately measure and model differences in G, H and S between two systems. This allows the selection of an arbitrary reference (G=0). Any comparison where the free energy of one system or state of a system is negative with respect to a reference state ($\Delta G < 0$) indicates a more stable thermodynamic state. Conversely, systems with $\Delta G > 0$ are less stable than the reference. Breaking down the components of ΔG also leads to approximating ΔH and ΔS in the same manner.

$$\Delta G = \Delta H - T\Delta S \qquad (3)$$

It is well established that proteins are quite dynamic in nature and take on a multitude of difference conformations. Considering two conformations, two snapshots of its existence, the ΔG can be estimated computationally by approximating the ΔH and ΔS between the two conformations. Since these two conformations are the same protein they have the same atoms, and the covalent bond network is unchanged from conformation to conformation. Two things are changing however: the non-covalent interactions, and the shape of the whole protein. Individual forces between atom pairs are well understood so they may be estimated computationally, however the conformational entropy is much more difficult to accurately approximate. Knowing how to calculate these two quantities its possible to calculate ΔG between two states. Calculating ΔG depending on conformation is a difficult task because proteins are highly dynamic, which requires a

large number of different microstates to be considered. The distance constraint model developed by Dr. Jacobs in the early 2000s allows one to rapidly calculate thermodynamic properties based on an ensemble of constraint topologies, and it also provides insight into mechanistic properties of a protein.

1.3 minimum Distance Constraint Model (mDCM)

The mDCM maps the three-dimensional all-atom structure of the native state of a protein onto a graph, where vertices represent atoms and edges represent different interactions, such as covalent bonding, hydrogen bonds (H-bonds), salt bridges and atomic packing through torsion interactions. A salt bridge is treated as a special type of H-bond. A free energy decomposition scheme is employed, where an energy and entropy contribution is assigned to each interaction, and moreover, each interaction is modeled as one or more distance constraint. Note that covalent bonds are always present in the graph because they do not fluctuate, whereas torsion distance constraints and H-bond distance constraints fluctuate, which generates a large ensemble of accessible constraint networks. The total energy and conformational entropy must be calculated for each network. Although correlations between degrees of freedom (DOF) invalidate an additive reconstitution of free energy components defined by a free energy decomposition(4,5) a non-additive reconstitution is possible(6) by applying the pebble game algorithm(7) to a constraint network to identify independent distance constraints. A lowest upper bound estimate for the total conformational entropy is obtained by preferentially adding the most restrictive independent interactions present in the network, while energy contributions are added over all interactions present. This approach proves to be

sufficient to provide high accuracy within the context of a phenomenological model comprising three free parameters.

The H-bonds identified from the native structure are allowed to break and reform, but no non-native H-bonds enter into the calculation. Torsion constraints fluctuate between native-like and disordered states to model good or poor atomic packing. Good atomic packing lowers energy and conformational entropy, while poor atomic packing increases energy and conformational entropy. To generate an ensemble of constraint networks that characterize conformational fluctuations about the native structure, and to calculate the free energy of the protein, the number of H-bonds, N_{hb}, and the number of good packing constraints, N_{gp}, are order parameters used to specify a macrostate of the protein. Given the maximum number of H-bonds, $N_{hb}^{\max}$, and the maximum number of good packing torsion constraints, $N_{gp}^{\max}$, as determined from an input structure, the number of possible microstates, Ω_m, for a given macrostate (N_{hb}, N_{gp}) is given as $\left(N_{hb}^{\max}!N_{gp}^{\max}!\right)\Big/\left(N_{hb}!\left(N_{hb}^{\max}-N_{hb}\right)!N_{gp}!\left(N_{gp}^{\max}-N_{gp}\right)!\right)$. The mixing entropy S_{mix} is then given by $S_{mix}=R\ln\left(\Omega_m\right)$, where R is the ideal gas constant.

By employing Monte Carlo sampling, the conformational entropy for a macrostate is estimated as:

$$S_c\left(N_{hb},N_{gp}\mid\delta_{gp}\right)=R\left[\delta_{gp}Q_{gp}+\delta_{pp}Q_{pp}+\delta_{hb}^{max}\sum_{k=1}^{N_{hb}^{max}}\left(1+\frac{1}{8}E_k\right)\left\langle q_k n_k\right\rangle\right] \qquad (4)$$

where $R\delta_{gp}$ is the entropy of an independent good packing constraint, $R\delta_{pp}$ is the entropy of an independent poor packing constraint and δ_{hb}^{max} is the entropy of the weakest

possible independent H-bond constraint. The dimensionless parameters δ_{pp} and δ_{hb}^{max} have been determined previously to be equal to 2.53 and 1.89, respectively(8), and they are transferable across globular proteins. Q_{gp} and Q_{pp} represent the average number of independent good packing or poor packing constraints found in the sub-ensemble of constraint networks specified by (N_{hb}, N_{gp}). The variable n_k equals 0 when the k-th native H-bond is broken or 1 when it is present. The quantity q_k counts the number of independent distance constraints associated with the *k*-th H-bond when present. The energy, E_k, for the k-th H-bond (or salt bridge) is limited to the range from 0 to -8 kcal/mol(9), with corresponding entropy being a linear function of its energy. The Q_{gp}, Q_{pp}, and q_i have unique values because of the preferential ordering imposed on Eq. (4) where the lowest entropy constraints are placed in the network before other constraints with higher entropy.

The free energy landscape is calculated by combining total energy and entropy contributions for each macrostate using a free energy functional. Accordingly, the free energy of a macrostate of the protein is constructed as:

$$G(N_{hb}, N_{gp}) = U(N_{hb}) - uN_{hb} + \nu N_{gp} - T\left[S_m(N_{hb}, N_{gp}) + S_c(N_{hb}, N_{gp} \mid \delta_{gp})\right] \quad (5)$$

where U is the average intramolecular H-bond energy given as $U = \sum_{i=1}^{N_{hb}} E_i n_i$. In Eq. (5) the three parameters, (u, v, δ_{gp}) are adjusted for thermodynamic predictions to best agree with experimental measurements. The u parameter represents a favorable (i.e. a negative value) effective protein-solvent H-bond energy, which competes against the formation of intramolecular H-bonds. The v parameter is a favorable energy that occurs when good atomic packing forms, and Rδ_{gp} was described above as the entropy of an independent

good packing constraint. For an exhaustive calculation on an average sized protein (~130 residues), about 2^{720} constraint networks are required, ignoring non-native interactions. Only a few hundred random samples of constraint networks *per macrostate* are sufficient to obtain accurate estimates of the free energy function given in Eq. (5). In most cases the phenomenological parameters *u, v,* and δ_{gp}, were usually determined by fitting the predicted heat capacity curve based on one input structure to the target heat capacity curve from differential scanning calorimetry (DSC). In other cases when the Cp is not available or difficult to measure experimentally GRIDsearch methodology is implemented. GRIDsearch calculates the thermodynamic qualities for a grid of discrete guesses of u,v, and dnat within the physical range from which heat capacity reconstitution is our guide toward physical solutions.

1.4 Applications of the mDCM

Leveraging the speed of thermodynamic and mechanical calculations from the mDCM, I show here how to improve and combine mDCM pipelines with Molecular Dynamics simulations to enhance its computational range. For small proteins like CXCL7 the addition of highly accurate, however expensive MD calculations can better explore the conformational diversity of these signaling proteins. While the community has identified over thirty Chemokines found in humans, very little literature has been written on the topic due to the high cost of experimental samples. For this reason a computational description of the prime characteristics could greatly move the needle in such an underdeveloped field.

RiVax on the other hand is of a different value. It is already established as a viable vaccine for Ricin and has a lot of eyes on the topic. For this reason we have collaborated

with experimentalists who have the ability to clinically test the activity of mutation studies. While this is a larger system and MD is too expensive for such a system it is feasible to use a methodology to sample a more diverse topology of the system in place of the conformation. Through this method I have approximated the melting temperature of mutations and predicted thermostability of various mutations. Mutations with similar function but greater stability could be a better option for drug design.

Finally Sleeping Beauty is a long flexible protein that is a piece of the CRISPR pipeline. Therefore greater understanding of this piece is pivotal to further development in the field. However, no experimental heat capacity data is known for SB due to the difficulty of purification and crystallization. I have developed a new GRIDsearch method to bypass the need for experimental curves through calculating many solutions both realistic and not. Through prior knowledge and internal consistency we have approximated the appropriate properties for SB and its mutations. This new GRIDsearch builds a foundation for many more advancements in the DCM by expanding the scope of potential protein systems previously impossible to calculate.

1.5 Limitations of the mDCM

While the mDCM has been effective in its predictions over the past decade we do not pretend that it is without weakness. The speed of the DCM is its greatest strength yet it is a double-edged sword. Every calculation is based off of a single conformation defined by an experimental structure that has potentials errors in itself. X-ray structures only represent a snapshot of a highly dynamic system and due to the crystallization process, possibly an unnatural conformation. The mDCM widely explores the potential of that single conformation, however it cannot make up for a poor starting structure. The

mDCM has shown to be robust despite the concern of a single reference and family wide studies have been implemented on Lysozyme(10), β-Lactamase(11), and other families with high confidence. With a greater sample of conformations we can be more statistically certain about our analysis of protein mechanics. In addition the mDCM loosely integrates solvent interaction with solution by the fitting parameter u. However, this term can be improved upon to better model the effect of protein surface mutations, which primarily alters solvation interactions. This dissertation in part works around the weakness of the mDCM in finding potential solutions for the purposes to increase accuracy by including additional physical effects along with improved work-flow implementation advancements.

Several different protocols have been employed to circumvent potential artifacts resulting from the single structure bias introduced by static structures. These studies while novel and robust, showed a need for a more sophisticated version of the mDCM in which complex pipelines of multi-tiered software are not necessary for optimal results. In the sections that follow we briefly touch on the project and implementation used to enhance mDCM output. These methods include conformational sampling with Molecular Dynamics (MD) software, large scale cross parameterization, and homology modeling.

Chapter 2 CXCL7 and Molecular Dynamics

2.1 Chemokine CXCL7 – Background

Chemokines are important regulatory proteins associated virtually with all physiologic or pathologic processes that involve immune system cell trafficking(12). The interest to understand how chemokines function increased greatly after they were identified as key players in inflammation-related and infectious disease processes, including autoimmune disease, HIV/AIDS, and cancer(13,14). While the assessment of chemokines in a variety of biological assays provided a wealth of information on their functional activity and new activities are continually being discovered, the structural biology approach allows for the mechanistic understanding of chemokine functioning.

There are 47 members in the family of human chemokines. Chemokine act on receptors that belong to the family of G-protein-coupled receptor (GPCR) superfamily(15). Multiple chemokines bind a single receptor and a single chemokine binds multiple receptors, resulting in a broad range of both unique and shared signaling events. Significant efforts are being made towards understanding the receptor binding affinity and specificity in the chemokine system in order to(16,17) facilitate the development of targeted therapeutic agents(18-20). It is expected that all chemokines have a similarly shaped receptor-binding site, because chemokines are small molecules (8-10 kDa) with essentially the same three-dimensional fold(21). Chemokines have two disulfide bonds joining four conserved cysteine residues per monomer (with few exceptions). The current model of receptor activation by a chemokine ligand, which is based on structure-function studies of several chemokines, proposes two consecutive

interaction sites: the N-loop of chemokine ligand interacting with N-terminal residues of the receptor (Site-I) and the N-terminal residues of chemokine ligand interacting with extracellular loop/transmembrane residues of the receptor (Site-II) (20,22) Accordingly, differences in amino acid composition of the N-terminal and N-loop regions contribute to the affinity and specificity of the receptor binding by a chemokine ligand. Additionally, experimental evidence showing that both the Site-I and Site-II of chemokine/receptor interaction comprise an extended and comparatively flexible region within the chemokine molecule, suggests a mechanistic role of protein dynamics in receptor binding(16,17,23-31) .

An interesting question is whether the differences in protein dynamics or flexibility in the apo structures through intermolecular inter-residue couplings are important in the selection of chemokine-receptor interactions. Indeed, the importance of coupling between Site-I and Site-II in the chemokine ligand for receptor activation has been demonstrated for the CXCL8 chemokine(32,33), suggesting similar effects are present in other chemokines due to the similarity of their structures and disulfide bond locations(33).

Here, we shall employ a combination of computational methods to investigate the dynamics, the inter-residue couplings, and the role of disulfide bonds on chemokine stability and flexibility in the monomer and dimer forms of CXCL7 chemokine. CXCL7 is a strong chemoattractant of neutrophils, and thus plays an important role in inflammation, blood clotting, and wound healing(34). It activates neutrophils via the interactions with cell surface receptors CXCR1 and CXCR2, but has much higher affinity to CXCR2(35-37). Similar to other chemokine ligands of CXCR2 receptor (e.g., CXCL1-3 and CXCL5-8), CXCL7 has a characteristic three N-terminal amino acid residue motif,

ELR, involved in receptor binding and cell activation(38) that was shown to be highly dynamic(17). By analogy to other CXCR2 chemokines, the N-loop (Ile8-His15) and the 30s loop (Gly26-Val34) are expected to be involved in receptor binding as they define relatively dynamic regions of the protein(16,17). Besides the chemotactic activity regulated through the receptor, it was found that the CXCL7 variant is missing just the two C-terminus amino acids, thrombocidin-1 (TC-1) possesses strong antimicrobial activity(39). While a positive patch on protein surface formed by several lysine and arginine residues (Lys17, Lys41, Arg54, Lys56, Lys57, Lys 60, and Lys61) was found to be essential for the antimicrobial activity of folded TC-1(40,41), functional differences (CXCL7 being inactive vs. TC-1 being active) of the two proteins with essentially the same monomer structure were explained by a higher and less restricted mobility of C-terminal residues in TC-1, leading to the increased possibility of interactions with the negatively charged bacterial membranes(16). Thus, protein dynamics plays an important role in both the receptor activation and antimicrobial activity of CXCL7.

The dynamics and stability of CXCL7 is investigated in detail by a combination of methods, where mDCM is combined with all-atom molecular dynamics (MD) simulation in explicit solvent following recent work on antibody fragments(42). For the first time, we extended the mDCM approach to account for solvation effects semi-empirically by extracting free energy differences upon dimerization from experimental circular dichroism (CD) experiments. This latter connection provides a route for quantitatively comparing backbone flexibility and residue pair couplings between the monomer and dimer forms to elucidate the effect of the protein-protein interface. Taken together, we

should arrive at a consistent picture of CXCL7 dynamics and stability in regards to the effect of dimerization and role played by the highly conserved disulfide bonds.

2.2 Molecular Dynamics and mDCM Methodology

To study the effects of the disulfide bonds, each monomer (A and B) was prepared in four cases: both Cys5-Cys31 and Cys7-Cys47 or none of the disulfide bonds form, or only one of the two disulfide bonds form. Residue pair couplings are sensitive to the H-bond network within a protein structure. The asymmetric x-ray crystal structure of a CXCL7 tetramer (pdb code 1NAP) showed differences in H-bonding details among the four monomers. Therefore, MD simulation was used to generate multiple input structures for the mDCM(42). The MD simulation was performed using Gromacs 4.5.5(43,44) in the NVT ensemble with the AMBER99SB-ILDN force field(45). The protein systems were solvated by adding 10.0 Å of TIP3P water in a periodic cubic box, with counter ions added to neutralize the net charge. Before production, the starting structure was obtained by minimizing the potential energy of the system, followed by 1 ns of NPT and 1 ns of NVT equilibration. Pressure (1 atm) was regulated using the extended ensemble Parrinello-Rahman approach(46) and temperature (300 K) was controlled by a Nose-Hoover temperature coupling(47). A cutoff distance of 10.0 Å was used for van der Waals interactions, and the Particle-Mesh-Ewald (48) method was employed to account for the long-range electrostatic interactions. All bonds to hydrogen atoms in proteins were constrained using LINCS(49), and bonds and angles of water molecules were constrained by SETTLE, allowing for a time step of 0.002 ps. A 100 ns trajectory was collected for each structure. A total of 2,000 evenly spaced frames from each trajectory

were clustered using the KCLUST module(50) from the MMTSB tool set(51) based on the RMSD of all heavy atoms. The cluster radius was adjusted so that the ten largest clusters represented 85% or more of the 2000 conformations.

The total number of input structures to mDCM consisted of 40 dimers and 80 monomers in total by considering each of the disulfide bond states for the dimer and monomer (A and B chains) separately. A simple grid search over the three dimensional parameter-space was performed to obtain different sets of parameters *u, v,* and δ_{gp} that would yield heat capacity curves with a peak at the experimentally determined T_m, while consistent with experimental ΔH of unfolding of a monomer. Initially, the grid searches were performed for 20 monomer structures and 10 dimer structures to identify a consensus target heat capacity curve for the monomer and dimer forms. Each input structure (with all disulfide bonds present) produced a unique set of parameters *u, v,* and δ_{gp}, needed to fit the target heat capacity in order to account for conformational differences. The parameters *u, v,* and δ_{gp}, were determined for structures with one or both disulfide bonds removed by computationally reforming the missing disulfide bonds so that the same target heat capacity for when all disulfide bonds are present can be used again. Following prior works(52,53) these new parameters are kept fixed when the disulfide bonds that were just added are subsequently removed because the global features of the protein conformation are unaltered.

2.3 Dimerization Energies

The mDCM phenomenological parameters, which in part reflect solvation contributions to the free energy, are robust across all structures from a MD trajectory because quantities such as heat capacity, average number of H-bonds and mechanical

properties, all being a function of temperature, are not affected by arbitrary constant shifts in the energy or entropy parameters. However, because differences in free energy, energy and entropy are sought between the monomer/dimer states and disulfide bonding states, relative parameter shifts between representative structures from the MD simulation must be accounted for. Importantly, changes in solvent exposed surface area due to displacement of solvent at the dimer interface must be considered. To circumvent modifying the mean-field treatment of solvation effects in the mDCM, a model independent approach was employed that uses the empirical parameters $\{\Delta H_a, \Delta S_a, T_a\}$ for dimer association based on fitting to the CD measurements assuming dimer association is a two state process with $\Delta H_a = T_a \Delta S_a$.

The free energy of a given representative structure relative to an arbitrary reference state is given by $G_k(T) = H_k(T) - TS_k(T) + (h_k - Ts_k)$ where $\{h_k, s_k\}$ are constant shifts in (energy, entropy) that reflect small variations in solvation free energy between MD frames. Since we have 80 different free energy curves for monomers with all disulfide bonds present and 40 such curves for dimers, we average over the respective numbers (80 and 40) and obtain two average free energy curves that serve as reference targets for the monomer and dimer structures. Note that deviations in free energy for a specific structure relative to the reference free energy curve at any given temperature is partly due to the $\{h_k, s_k\}$ shifts poorly modeled by mDCM in addition to structural differences captured well. Therefore, uncertainty in mDCM predictions are obtained by adjusting the $\{h_k, s_k\}$ parameters so that each free energy curve derived from a MD frame matches the reference curve the best it can through a least squares error fitting. Any deviation remaining reflects the intrinsic uncertainty in the mDCM prediction combined with the

conformations explored by the MD simulation. This congruency process yields two free energy curves (for monomer and dimer) resulting in mDCM predictions with appropriate error bars. Although the change in free energy upon formation of the dimer interface is not accounted for (yet), we know $\Delta H_a = H^{(D)} - 2H^{(M)}$ and $\Delta S_a = S^{(D)} - 2S^{(M)}$ where (D) and (M) respectively represent the dimer and monomer forms. At T_a let $\Delta H_a^{(0)} = H_{mDCM}^{(D)}(T_a) - 2H_{mDCM}^{(M)}(T_a)$ and $\Delta S_a^{(0)} = S_{mDCM}^{(D)}(T_a) - 2S_{mDCM}^{(M)}(T_a)$, and using the freedom to add a constant energy and entropy shift between the monomer and dimer, it follows $H_{pred}^{(D)} = H_{mDCM}^{(D)} - \Delta H^{(0)} + \Delta H_a$ and $S_{pred}^{(D)} = S_{mDCM}^{(D)} - \Delta S^{(0)} + \Delta S_a$. Then the predicted free energy curve for the dimer is given as $G_{pred}^{(D)} = H_{pred}^{(D)} - T S_{pred}^{(D)}$. Although this procedure fixes the difference in free energy curves for the dimer minus two monomers, it is still possible to shift the overall energy and entropy of the monomer because this arbitrariness does not modify any response property, nor any measurable differences in energy and entropy of interest.

2.4 Results for CXCL7

The target heat capacity curves and spread of the various heat capacity fits to the targets for the monomer and dimer are shown in Figures 1a and 1b. The mDCM parameters are u = -1.72±0.26 kcal/mol, v = -0.47±0.19 kcal/mol and δ_{gp} = 1.20±0.04 for the monomer, and u = -1.44±0.27 kcal/mol, v = -0.11±0.20 kcal/mol and δ_{gp} = 1.43±0.23 for the dimer. These parameters indicate that the native state of the dimer has greater conformational entropy per residue than that of the monomer (as seen before packing is not as strong(54)), while the destabilizing effect of H-bonding from protein to solvent plays a lesser role. The maximum heat capacity, Cp_{max}, of CXCL7 dimer is more than

twice that of the monomers (Fig. 1b), suggesting that the dimer is more cooperative. Typical free energy landscapes are shown in Figures 1c and 1d for the monomer and dimer. The landscapes with barriers are indicative of a first order transition, however a small fraction of the landscapes do not show a barrier that separates the native and unfolded basins. Moreover, when present, the barriers are low. While the model parameters can be adjusted to produce a clear two-state behavior in all representative structures, it would require a larger change in enthalpy of folding than estimated from CD experiments, hence the unfolding transition is not very cooperative.

The backbone flexibility, quantified by the flexibility index (FI) at temperatures 300K and 350K, is shown in Figure 2a for the monomer and in Figure 2b for the dimer. The temperature 350K is chosen to identify mechanically less stable regions in CXCL7 at elevated temperatures, because at this temperature the protein remains folded with a weakened H-bond network. In general, the backbone flexibility gradually increases in a nearly uniform fashion throughout the protein as temperature increases. In the monomer, the N-terminus and the 50s loop are flexible, while the three beta-strands, C-terminal α-helix, 3_{10} alpha-helical turn and the 40s loop have a negative FI indicating that they are over-constrained. To a lesser degree, the 30s loop is over-constrained at biological temperatures and it becomes flexible just before the protein unfolds. This suggests that the two disulfide bonds mechanically stabilize the 30s loop region. The specific role of disulfide bonds in stabilizing the protein will be addressed below.

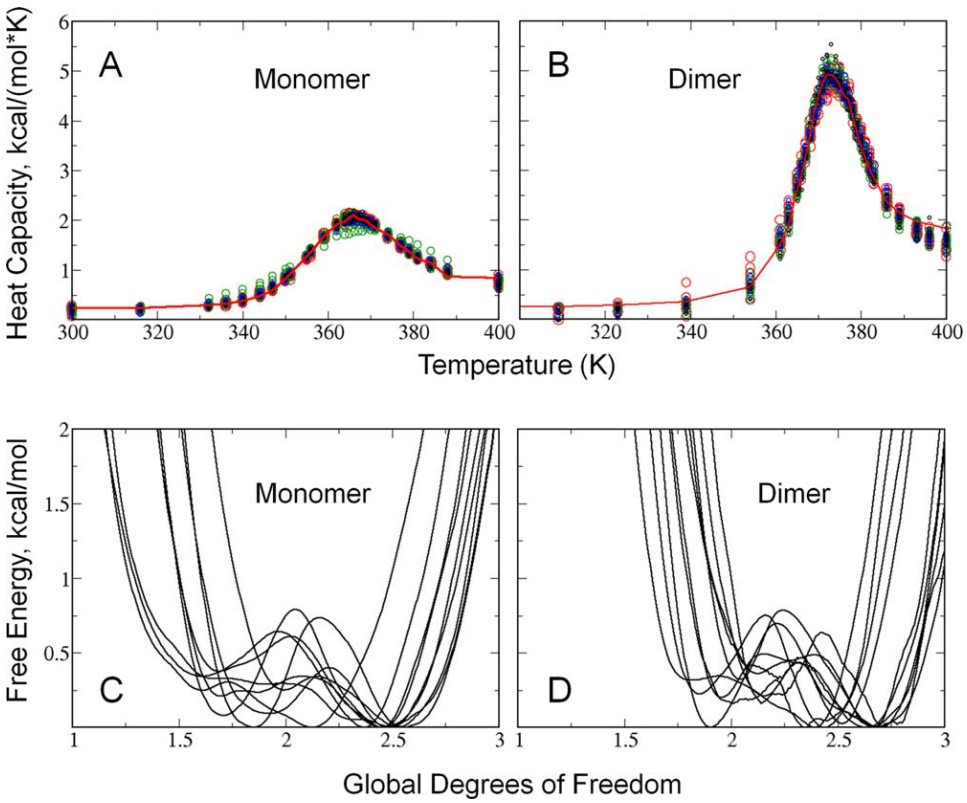

Figure 1: **Heat capacity and free energy of CXCL7**. Heat capacity curves for the monomer (A) and dimer (B). The solid red line is the target heat capacity curve that was initially generated for the case of all disulfide bonds present. The data points are the fits to this target curve for 80 representative structures for monomers and 40 representative structures for the dimer. The corresponding free energy landscapes that result from good fits to the heat capacity are shown for the top 10 representative structures for monomer (C) and dimer (D). The lowest point in each of the free energy landscapes have been set to be zero so that all the curves can be easily compared.

Dimerization enhances mechanical stability in an interesting way. Most H-bonds formed at the interface connect preexisting rigid substructures within the monomers, which lowers the FI in regions already over-constrained in the monomer. A histogram for where H-bonds form across the interface in relation to the flexibility index is shown in Figure 3 for the case that both disulfide bonds are present. On average, about 13 H-bonds are formed at the interface. Despite these additional interfacial H-bonds, backbone flexibility increases in the 30s loop and for several residues at the N-terminus, including

Cys5. Backbone flexibility along the C-terminal α-helix is not significantly affected by dimerization, even though the two α-helices align in the CXCL7 dimer and new side-chain to side-chain contacts form between the two α-helices across the dimer interface.

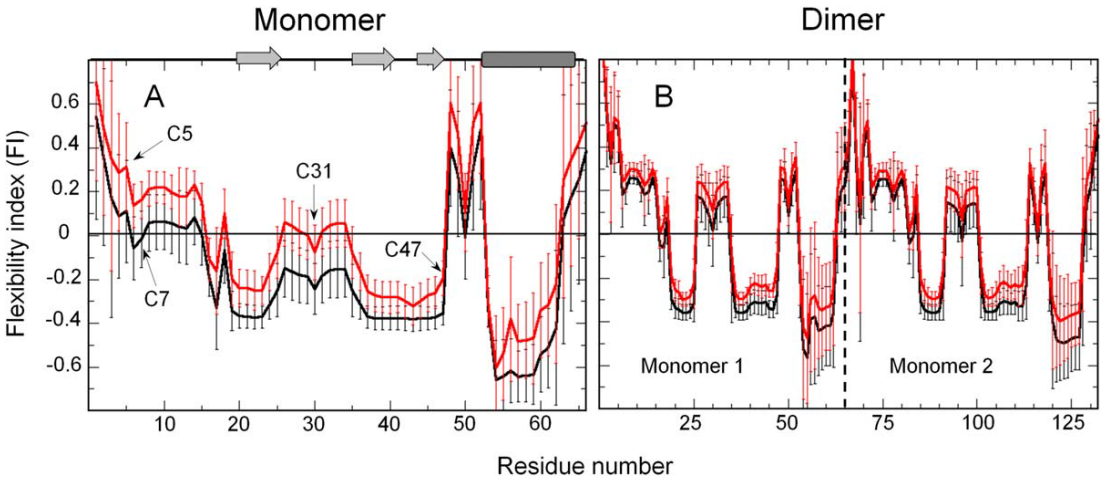

Figure 2: **Flexibility index of CXCL7**. The backbone flexibility is plotted using the flexibility index at T5300K and 350K for the monomer (A) and dimer (B). The secondary structure elements are shown above the plots.

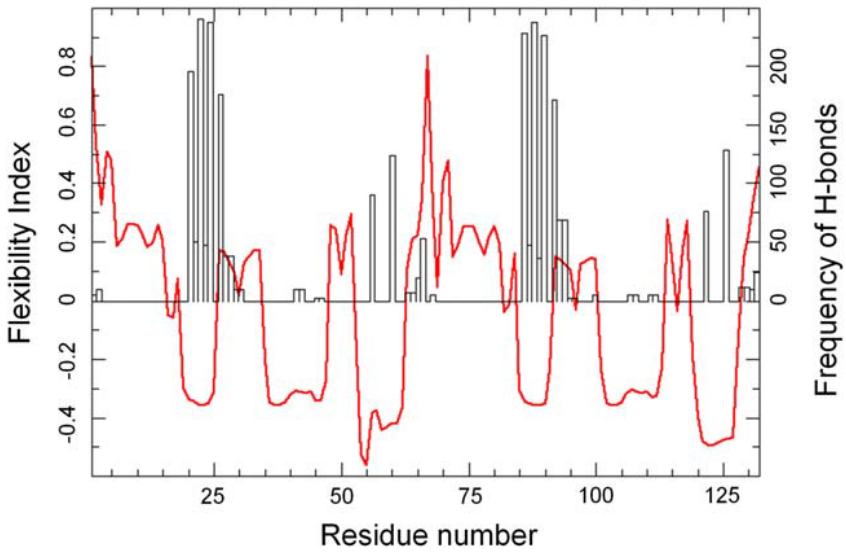

Figure 3: **Location of interfacial H-bonds in CXCL7**. A histogram showing where H-bonds form across the interface in relation to the flexibility index is shown for the case where both disulfide bonds are present.

Residue-to-residue mechanical couplings (the correlations in DOF of atomic motions) within CXCL7 are shown at T=300K and 350K in Figures 4a and 4b for the monomer, and in Figures 4c and 4d for the dimer. At T=300K (Figure 4a), the most rigidly correlated residues are located in the beta-sheet and α-helix secondary structures, but little rigidity propagates between these two secondary structures. Although some correlated flexibility is observed in the N-terminal region between residues Cys5 and His15, overall the monomer forms a fairly rigid molecule at 300K. At 350K the H-bond network weakens, and there is an increase in flexibly correlated motion in the 50s loop hinge region where it couples to the Cys5 to His15 N-terminus section, the 30s loop, and a small section at the end of the C-terminus (Figure 4b). Surprisingly, the C-terminal α-helix is not strongly coupled to other parts of the molecule, including the beta-sheet on top of which it folds. As such, only at low temperature (300K) does the helix have a weak propensity to intermittently stick to the beta-sheet, but otherwise, it moves independently from the rest of the molecule. We observed a similar behavior in a closely related chemokine CXCL4 before(55).

Figures 4c and 4d show that dimerization of CXCL7 greatly increases flexibility correlations. The 30s loop is flexibly correlated with the 50s loop, a large section in the N-terminus, and a small end section of the C-terminus. The rigidity correlation is nearly the same within the beta-sheets and alpha helices as in the monomer, but extends across the monomers, forming a larger rigid core. The dimerization has also enabled new flexibly correlated motions, i.e., the 30s and 50s loops in each monomer, as well as the flexible N-terminus and C-terminus regions are flexibly correlated in different monomers

in the dimer. At 350K, the correlations pattern is similar with weaker levels of rigidity correlation morphing into flexible correlations. This results from loop regions having lower density of H-bonds weakening more readily than in the monomer as temperature increases.

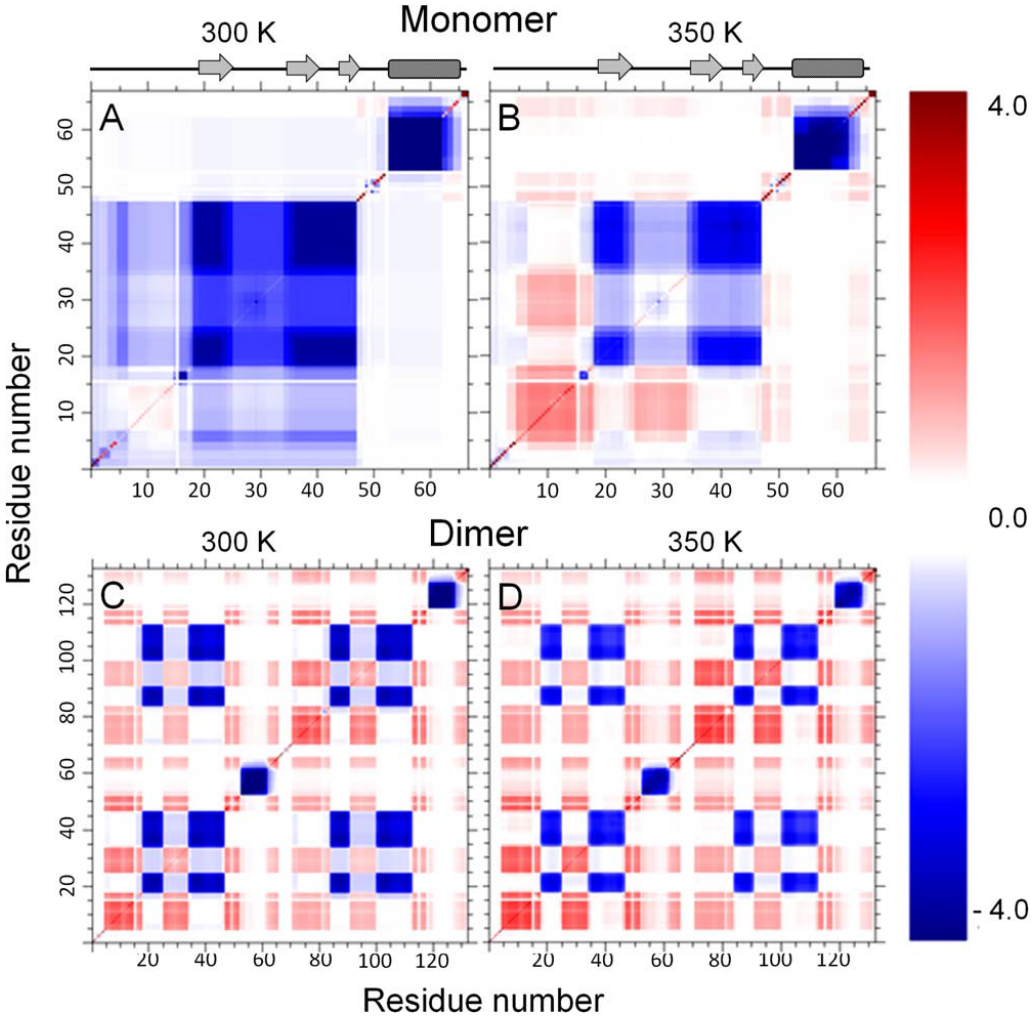

Figure 4: **Cooperativity correlation plots for CXCL7 monomer and dimer.** The residue-residue mechanical couplings are plotted for the monomer at T5300 K (A) and at T5350 K (B). This quantity is also plotted for the dimer at T5300K (C) and at T5350K (D). The coloring scheme is defined by the color bar and is the same for all panels, where white represents no correlation, blue indicates a degree of rigidity correlation, and red indicates a degree of flexibility correlation. The data is shown for the case when all disulfide bonds are present.

The characteristic feature of all chemokines is the presence of two disulfide bonds between highly conserved cysteine residues. To investigate the structural effects of each bond on the dynamics and mechanical couplings in CXCL7 monomer and dimer, we carried out the MD/mDCM analysis for protein states where one or both bonds were removed. The melting temperature only slightly drops with the removal of one or two disulfide bonds, and destabilization is less in the dimer compared to the monomer, suggesting that disulfide bonds do not play a critical role in maintaining structural integrity in either the monomer or dimer. This result agrees with reported experimental data on a closely related chemokine, CXCL4(56).

To quantify changes in thermodynamic stability for each disulfide bond configuration, the change in Gibbs free energy, enthalpy, and entropy upon dimerization is compared at 300K, as well as total energy of the H-bond network and the number of interfacial H-bonds formed. Table 1 shows that dimerization of CXCL7 is energetically favorable, and enthalpically driven for all disulfide bond configurations. Because the overall shape of the protein does not change significantly when the disulfide bond is removed, the quantity $-T\Delta S$ listed in Table 1 is dominated by differences in conformational entropy due to different disulfide bond configurations.

Table 1: **Dimerization Energy of CXCL7:** Enthalpy vs Entropy tradeoff upon dimerization of CXCL7

Table I
Free Energy, Enthalpy, Entropy, and H-Bond Energy of CXCL7 Dimerization Calculated Using the mDCM Approach

Disulfide bonds present	$\Delta G_{D\text{-}2*M}$ ($\Delta H - T\Delta S$)	$\Delta H_{D\text{-}2M}$	$-T\Delta S_{D\text{-}2M}$	$\Delta HB_{D\text{-}2M}$	Mean number of interfacial H-bonds
Cys5-Cys31 Cys7-Cys47	-1.1 ± 0.0 (+ +)	-27.9 ± 0.0 (+ +)	26.8 ± 0.0 (− −)	-1.8 ± 0.2 (+)	13.4
Cys7-Cys47	-2.9 ± 0.4 (−)	-35.9 ± 2.3 (−)	33.0 ± 2.3 (+)	-1.4 ± 0.4 (+)	14.0
Cys5-Cys31	-3.4 ± 0.3 (− −)	-33.6 ± 1.6 (+)	30.2 ± 1.6 (−)	-3.0 ± 0.3 (−)	14.6
None	-2.1 ± 0.4 (+)	-36.8 ± 2.9 (− −)	34.7 ± 2.9 (+ +)	-4.1 ± 0.4 (− −)	16.0

All energies are expressed in kcal/mol. The last column gives the mean numbers of interfacial hydrogen bonds calculated over the MD trajectory. The data is shown for the cases when two disulfide bonds present, either Cys5-Cys31 or Cys7-Cys47 bond removed, or both disulfide bonds removed. (++) and (−−) indicates the most and the least favorable case, respectively.

Although the conformational entropy per residue is higher in the dimer than in the monomer, the total entropy is reduced in the dimer for all disulfide bond configurations due to solvation effects. Interestingly, differences in -TΔS for different disulfide bond configurations span about 7 kcal/mol, which can easily be accounted for by subtle differences in the H-bond network, especially at the inter-monomer interface. However, it is also clear from Table 1 that the free energy differences do not track differences in the total energy of the H-bond network. The correlation coefficient of 0.11 between the change in free energy and change in total H-bond energy indicates that the changes in conformational entropy (which depends on the location of H-bonds and their microenvironments, i.e. the entire H-bond network), play a critical role in stabilizing the dimer

The end result is just what one would expect: the dimerization is favorable at low temperature because a small number of interfacial H-bonds lower the energy enough to overcome the conformational entropy reduction. At elevated temperature, the entropic component to the free energy becomes more important, and overcomes the lower energy,

and hence the dimer dissociates. Because different disulfide bond configurations affect molecular packing and H-bond arrangements, the effect of removing a disulfide bond is non-additive and context dependent. For example, the dissociation temperature is predicted to be lowest when two disulfide bonds are present, next lowest when both are removed and highest when only Cys7-Cys47 is removed.

Removing one or both disulfide bonds leads to interesting and often non-obvious effects, which are revealed by the change in FI. In the monomer, removing either one or both disulfide bonds will generally increase backbone flexibility throughout the protein. In all cases, the main increase in flexibility appears in the C-terminal α-helix. Overall, no substantial FI differences are noticed when either the first or second disulfide bond is removed, although they do affect different regions differently. In all cases, the 50s loop becomes less flexible. The overall characteristics in the differences are insensitive to temperature. In the case of the dimer, removing the disulfide bonds has an opposite effect (decrease in flexibility) in many regions throughout the protein with increases in a few localized regions and in α-helices. Amongst the three possible combinations, removing the Cys7-Cys47 disulfide bond produces the greatest increases in backbone flexibility throughout the protein.

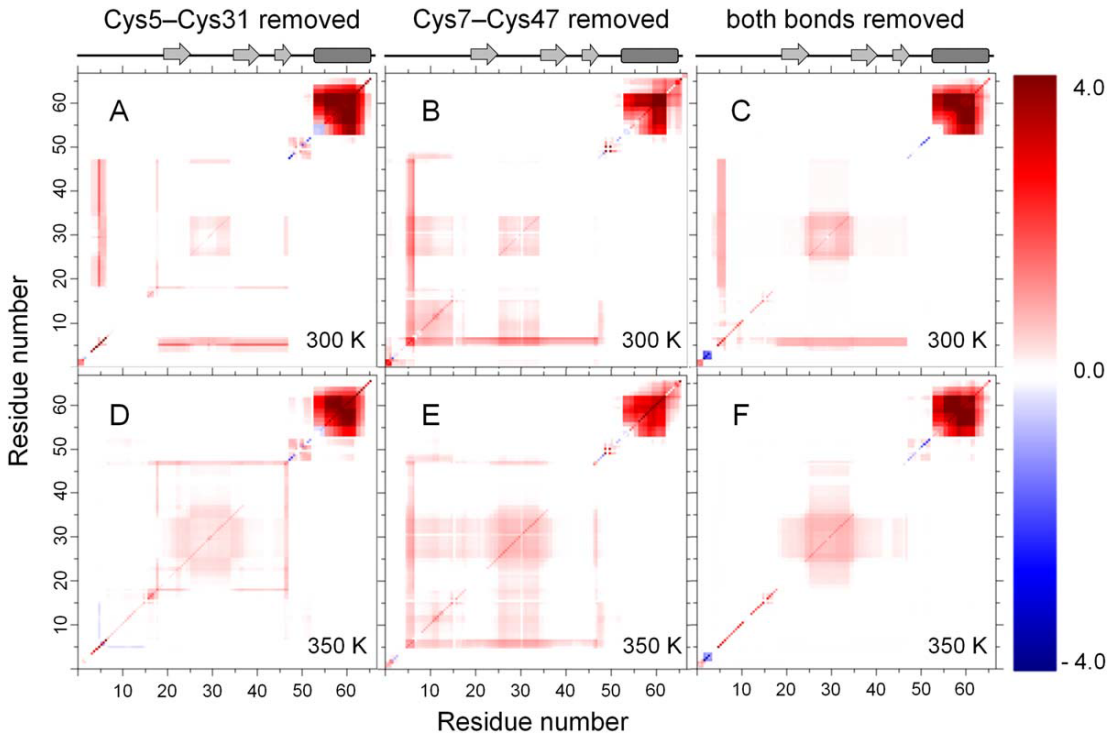

Figure 5: **Cooperativity correlation difference plots for CXCL7 monomer**. The difference in residue-residue mechanical couplings is shown for the monomer at T5300 K in panels (A–C) and at T5350 K in panels (D–F). To construct these plots, the residue-residue mechanical couplings for each case of disulfide bond configurations were subtracted from the residue-residue mechanical couplings when both disulfide bonds are present: the Cys5-Cys31 disulfide bond removed (A,D); the Cys7-Cys47 disulfide bond removed (B,E); and both disulfide bonds are removed (C,F). Shown data is filtered based on the signal beyond noise ratio (SBNR) as explained in the text.

These effects are further analyzed in Figures 5 and 6 that show differences in residue-residue couplings for cases where one or both disulfide bonds are removed relative to when both disulfide bonds are present. Figures 5 and 6 show the data points outside one standard deviation (e.g., SBNR), hence present only statistically significant changes due to the removal of disulfide bonds. For the monomer at T=300K, Figure 5 shows a dramatic increase in flexibly correlation in the C-terminus α-helix, and regardless of which disulfide bond is removed residues Cys5 through Cys7 are more flexibly correlated

with residues from Asn18 through Cys47. However, when only the Cys7-Cys47 disulfide bond is removed, flexibility correlations that couple to residues Cys5 through Cys7 further extend to residues in the N-terminal loop and in the 30s loop. At elevated temperature, little statistically significant correlations remain, and no new patterns of flexibility correlations emerge. The overall increase of flexibility correlations seen in Figure 6 is consistent with an overall increase in conformational entropy within a monomer. In the dimer (Figure 6) removal of one or both disulfide bonds generally increases rigidity correlations. However, as mentioned above, removing the Cys7-Cys47 disulfide bond increases flexibility and increased flexibility correlations as shown in Figure 6b for T=300K, which further increases at T=350K. Flexibility correlation between the 30s and 50s loops is much less in the monomer. As seen in Figure 6 there is a net structural stabilization by removing disulfide bonds within the dimer, which is consistent with the thermodynamic analysis (Table 1) showing conformational entropy decreases upon removal of disulfide bonds. Interestingly, the 30s and 50s loops are less flexibly correlated to one another when both disulfide bonds are missing than compared to when both disulfide bonds are present at T=300K, and this difference gradually decreases as temperature increases.

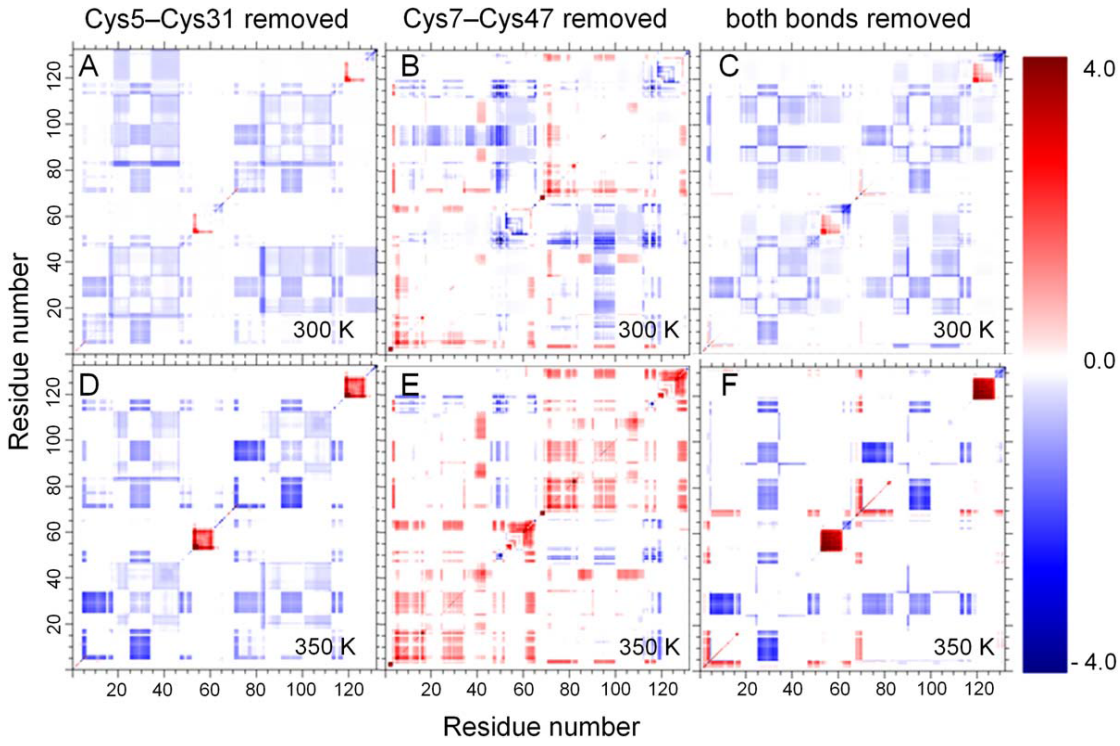

Figure 6: **Cooperativity correlation difference plots for CXCL7 dimer**. The difference in residue-residue mechanical couplings is shown for the monomer at T5300 K in panels (A–C) and at T5350K in panels (D–F). These plots were constructed similarly to the plots shown for CXCL7 monomer and demonstrate the SBNR of the difference in residue-residue mechanical couplings for the case of Cys5-Cys31 disulfide bond removed (A,D), the Cys7-Cys47 disulfide bond removed (B,E), and both disulfide bonds removed (C,F).

2.5 Discussion

In the absence of experimental structure of chemokine-receptor complex, mutagenesis studies have shown that the N-terminal residues preceding the first cysteine residue, the N-loop preceding the first beta-strand (residues 9-19), and the 30s loop connecting the first and second beta-strands are implicated in receptor binding and activation for most chemokines studied to data(20,22,57). The two-step model of chemokine-receptor binding proposes two sites of interactions between the chemokine ligand and cognate receptor. At Site I, the N-loop of the chemokine interacts with N-terminal residues of the

receptor. At Site II, the N-terminal residues preceding the first cysteine and the 30s loop of the chemokine interact with extracellular loop/transmembrane residues of the receptor(20,57). In this regard, CXCL7 is no exception as the triad of N-terminal residues preceding Cys5 has been shown to be critical for neutrophil activation(58). Furthermore, these residues form an ELR motif that is critical for the function of chemokine ligands binding the same receptor as CXCL7(20,22,38).

It is more interesting, however, to analyze the dynamic properties of the 30s loop and the N-loop residues. Although the role of the 30s loop in CXCL7 has not been investigated previously, its importance has been shown experimentally for other CXCR2 chemokine ligands(33). The mDCM analysis shows that at 300K the 30s loop is prone to being flexible in the monomer, and very flexible in the dimer. While the mDCM indicates the N-loop residues are flexible and highly correlated, this does not imply high mobility. The mDCM shows the whole N-terminus of CXCL7 is flexibly correlated to the 30s loop, suggesting that a perturbation at site I (e.g., receptor binding) propagates to the site II through conformational rearrangement. It also suggests that residue substitutions introduced in these parts of the molecule may perturb the dynamics by altering existing mechanical couplings, ultimately changing the receptor-binding properties. Note that the recent study of another CXCR2/CXCR1 ligand, CXCL8, revealed that perturbing the GP motif in the 30s loop causes changes in dynamics and conformational rearrangements altering receptor binding by perturbing the equilibrium between binding competent and incompetent conformations of CXCL8 within the dynamic ensemble(33).

The coupling between N-terminus residues and the 30s loop can be intuitively understood due to the disulfide bond between Cys5 and Cys31 linking these two regions.

In support, our and literature data(16,17) show large R_{ex} contribution for Cys31, indicative of significant slow motion that involves this residue. In other chemokines, similar slow motions in the 30s loop have also been detected for the corresponding cysteine or for the residues in its close proximity(28,31,59). In application to the receptor binding, this would mean that the N-terminus of CXCL7 samples a significantly larger ensemble of conformations and takes longer to find the correct binding conformation. This result is in agreement with observations made in other chemokines, where perturbations of the disulfide bond linking the N-terminus and 30s loop had deleterious effect on receptor binding and function(60,61).

Importantly, the reduction of one or both disulfide bonds was found not to affect the antimicrobial activity of Thrombocidin-1 (TC-1), which differs from CXCL7 only by a two-amino acid C-terminal deletion(40). This result is consistent with the MD/mDCM predictions that disulfide bonds are not critical for maintaining protein stability in CXCL7. The MD/mDCM results also show that the C-terminal α-helix is weakly coupled to the 50s loop, and is otherwise independent of the protein regardless of the disulfide bond states. Moreover, the set of residues Lys17, Lys41, Arg54, Lys56, Lys57, Lys61, Lys62 that were identified experimentally as being critical to function(16,40,41) fall in the correlated rigid regions of the monomer and dimer. In contrast, the helix has relatively large contributions from slow motions in agreement with previous studies(16,17). This result indicates the C-terminal helix is able to move relative to the rest of the protein, which was observed in CXCL4 where the helix exhibited large-amplitude motions relative to the beta-sheet in a MD simulation(55). These results suggest it may be possible to truncate the C-terminal helix and retain receptor-binding

capability. Conversely, residue substitution/deletion into highly correlated parts of a protein often dramatically alters stability and would likely affect receptor binding by shifting the equilibrium conformational ensemble(32,33,62-64).

The MD/mDCM results clearly show that dimerization of CXCL7 induces correlations in flexibility between the 30s and 50s loop. At the same time, there is an increase in thermal and structural stability due to interfacial H-bonding. Recently, the Le Châtelier's principle was invoked(42) to explain in general terms why flexibility tends to redistribute throughout the protein upon a local perturbation to oppose a net shift in rigidity or flexibility. Specifically, to regain a new equilibrium conformational ensemble, a protein attempts to restore a balance in DOF, such that due to the action of rigidifying one region, this drives another region to become more flexible. From this viewpoint, dimerization perturbs two monomers. Each monomer becomes rigidified within beta strand 1 of the beta-sheet, and other areas of the beta-sheet due to interfacial H-bonds. Both monomer structures snap together and extend rigidity across the beta-sheets of each monomer. As this increase in rigidity takes place, a large increase in flexibility in loop regions is induced that become flexibly correlated.

While the CXCL7 fold prevents the removal of disulfide bonds from being detrimental to structural stability, the Le Châtelier tendency to restore equilibrium drives the H-bond network to relax in a way that shifts the CXCL7 monomer-dimer equilibrium toward dimerization. Moreover, a redistribution of flexibility occurs to maintain the number of DOF about the same. In CXCL7, the structure changes in such a way that more H-bonds form at the dimer interface when disulfide bonds are removed. For example, about 3 more interfacial H-bonds form when no disulfide bonds are present

compared to when both disulfide bonds are present. Based on these features, a “snap on” mechanism seems to occur during the process of dimerization. That is, monomers are predisposed to propagate rigidity and flexibility through distinct channels. Upon dimerization, the rigidity channels within the monomers snap together as new interfacial H-bonds form, leading to an extended rigid region (extended beta-sheet). At the same time, the DOF are released in the flexibility channels connecting the 30s and 50s loops in both monomers. Interestingly, the presence of both disulfide bonds creates maximal correlated flexibility between these loops, and at the same time the dimer is least thermally stable. Because the disulfide bonds are highly conserved, this suggests that chemokines function best when there is marginal stability between monomer and dimer forms. In addition to securing marginal stability, the presence of both disulfide bonds provides maximum contrast in flexible correlations between the 30s and 50s loops.

Among the three computational methods, the combined MD/mDCM approach provides detailed information about the native state ensembles, the role of the disulfide bonds in the monomer and dimer and the role of the interface. This work establishes the MD/mDCM approach as a viable means to study the thermodynamic and mechanical properties across the family of chemokine proteins within humans.

Chapter 3 RiVax

3.1 Background

The advent of structural vaccinology coupled with the capacity to synthesize large amounts of recombinant protein antigens in Escherichia coli and other organisms has made subunit vaccines increasingly attractive in the ongoing war against emerging infectious diseases and biothreat agents, including toxins such as ricin for which effective vaccines have proven elusive(65-67). Ricin is a type II ribosome-inactivating protein derived from the seeds of the castor bean plant (Ricinus communis)(68). In its mature form, ricin is a 64 kDa globular glycoprotein composed of a 34 kDa enzymatic subunit (ricin toxin A chain [RTA]) joined by a single disulfide bond to a 32 kDa binding subunit (ricin toxin B chain [RTB])(69,70). RTB, a galactoseand N-acetylgalactosamine (Gal/GalNac) lectin, promotes the attachment and entry of ricin into mammalian cells(71,72). After endocytosis, RTB mediates the retrograde trafficking of ricin from the plasma membrane to the trans Golgi network and the endoplasmic reticulum (ER). Once in the ER, RTA is liberated from RTB and is dislocated across the ER membrane into the cytoplasm(73). RTA is an RNA N-glycosidase that selectively depurates a highly conserved adenosine residue within the sarcin-ricin loop of eukaryotic 28s ribosomal RNA(74). Hydrolysis of the sarcin-ricin loop by RTA results in the cessation of cellular protein synthesis, activation of the ribotoxic stress response, and cell death via apoptosis(75). Ricin is a potential biothreat agent and remains a concern for military and public health officials in the United States, as evidenced by ricin being classified by the Centers for Disease Control and Prevention within the Select Agents and Toxins(67,76). Efforts to develop a ricin toxin vaccine for use by military personnel and certain civilian

populations (e.g., emergency first responders and laboratory staff) initially focused on a formalin-inactivated toxoid. Although the toxoid vaccine proved to be highly efficacious in rodents and nonhuman primates, its development for use in humans was abandoned because of manufacturing and safety concerns(77,78). For more than a decade, efforts have been aimed at the development of a recombinant subunit vaccine, with particular emphasis on attenuated derivatives of the toxin's 267 amino acid enzymatic subunit, RTA(79,80). One of the most advanced candidate subunit antigens is RiVax™, an E coli-based recombinant form of RTA that contains 2 point mutations: V76M and Y80A(81,82). The Y80A mutation attenuates RTA's RNA N-glycosidase activity, whereas the V76M mutation eliminates RTA's capacity to elicit vascular leak syndrome(81,82). X-ray crystallography indicates that the V76M and Y80A substitutions do not alter the tertiary structure of RTA(83). In mice, RiVax immunization by the intramuscular, subcutaneous (s.c.), or intradermal routes elicits toxin-specific serum IgG antibodies that are sufficient to confer protection against a lethal dose of ricin administered by systemic (intraperitoneal) or mucosal (aerosol) routes(79,82,84-87). Moreover, phase I clinical trials have demonstrated that RiVax is safe in healthy human volunteers(88,89). The phase Ib clinical trials did, however, reveal a notable shortcoming associated with RiVax; even after 3 intramuscular immunizations with 100-mg RiVax adsorbed to Alhydrogel, toxin-neutralizing serum antibody titers remained relatively modest(88,89). This finding was not unexpected, considering that inert, none self-assembled subunit vaccines are notorious for being poorly immunogenic. Nonetheless, RiVax may be unusual in this respect as evidenced by studies in our laboratory that compared the onset of toxin neutralizing antibodies (TNAs) in mice after 2 parenteral

immunizations with RiVax or a recombinant protective antigen from anthrax, each adsorbed to Alhydrogel. Whereas anthrax TNAs were achieved within days after the booster immunization, ricin-specific neutralizing antibodies were not detectable for weeks(90). One obvious strategy to augment the overall immunogenicity of RiVax is through the use of next generation adjuvants(65). It was recently reported that a novel type II heat labile enterotoxin when co-administered with RiVax (without Alhydrogel) enhanced the onset of TNAs and protective immunity(87). However, adjuvants themselves may not be sufficient to achieve maximal immunogenicity of RiVax. There is evidence to suggest that the failure of RTA-based antigens, in general, to elicit high titer TNAs may be due to factors intrinsic to RTA, such as its propensity to partially unfold in the absence of its partner, RTB(91). If that is the case, enhancing the immunogenicity of RiVax may require a structure-based redesign of the antigen itself. Reengineering RTA is not a new concept, as the US Army has produced truncated (e.g., removing residues 199-267) and disulfide bond stabilized derivatives of RTA that have proven effective at eliciting protective immunity to ricin in mouse models(80,92-94). With the goal of preserving the RTA's native structure as much as possible, we have chosen a site-directed approach to the redesign of RiVax. In a recent study, orthogonal and complementary computational protein design approaches where used to generate a total of 11 single point mutations in RiVax that were characterized using an array of biophysical and biochemical techniques to assess secondary and tertiary structures, as well as thermostability(95). One approach applied the Rosetta protein modeling suite(96,97) to the RiVax crystal structure (PDB ID: 3BJG) to generate an ensemble of 50 near-native conformations(98,99). Each conformation was analyzed to identify under packed regions

in the protein core and then subjected to RosettaDesign to identify small-to-large or isosteric mutations at these sites that were predicted to stabilize the protein by improving packing(100). This strategy identified several mutations, including V81I, C171L, and V204I. A second design approach utilized a multi-layered filtering scheme to identify mutations predicted to have enhanced stability(101,102). Predicted stabilizing mutations were created and visually inspected in the Molecular Operating Environment (MOE) which lead to the identification of mutations V18P, C171V, and S228K. These single point mutations were introduced onto the RiVax template and subsequently evaluated in a mouse model for the ability to elicit ricin-specific neutralizing antibodies and protective immunity(95). In this present study, we have now created additional derivatives of the RiVax antigen in which we have combined single point mutations to generate double and even triple mutants. The resulting combinations of cavity-filling mutations were tested in a mouse model and led to the identification of 2 derivatives of RiVax, referred to as RiVax-RB and RiVax-RC that are several times more efficient than RiVax at eliciting TNAs.

3.2 Results

To glean insight into the structural basis for the increased immunogenicity conferred by the RB and RC constructs, and increased stability in RC, a QSFR analysis was performed on RiVax (PDB ID: 3SRP) and its mutants (RA, RB, RC, SA, and SB) in the native. Toxin-neutralizing titers in sera of mice immunized with RiVax and RiVax derivatives. Groups of mice (n . 8 mice/group) were immunized with 20 mg of RiVax or the indicated RiVax derivatives, each adsorbed to Alhydrogel.

It was found that RiVax and all 5 mutants have virtually the same backbone

flexibility in the native state, with only 2 notable differences. First, all mutants are slightly less flexible in the region at residues 20-28, and mutants RA, RB, and RC became more flexible at residue 180. In the transition state, RB has the greatest backbone flexibility within the beta-sheet region at residues 55-95. RC is less flexible than RiVax in a few regions, most notably within the beta-hairpin turn at residues 225-245. Greater differences occur in residue pair couplings quantified by the CC-plots. The CC-plot for mutant RB in the native state (Fig. 7a) shows that RB is more rigidly correlated than RiVax in regions that span residues from the N-terminus to residue 45, residues 98-103, and from residue 200 to the C-terminus. As such, RB restricts the N- and C-termini motions compared to RiVax in the native state. Moreover, these 2 swaths of restricted motions couple to residues 98-103, which lie within one of the 2 B-cell epitopes at Asn97ePhe108(103).

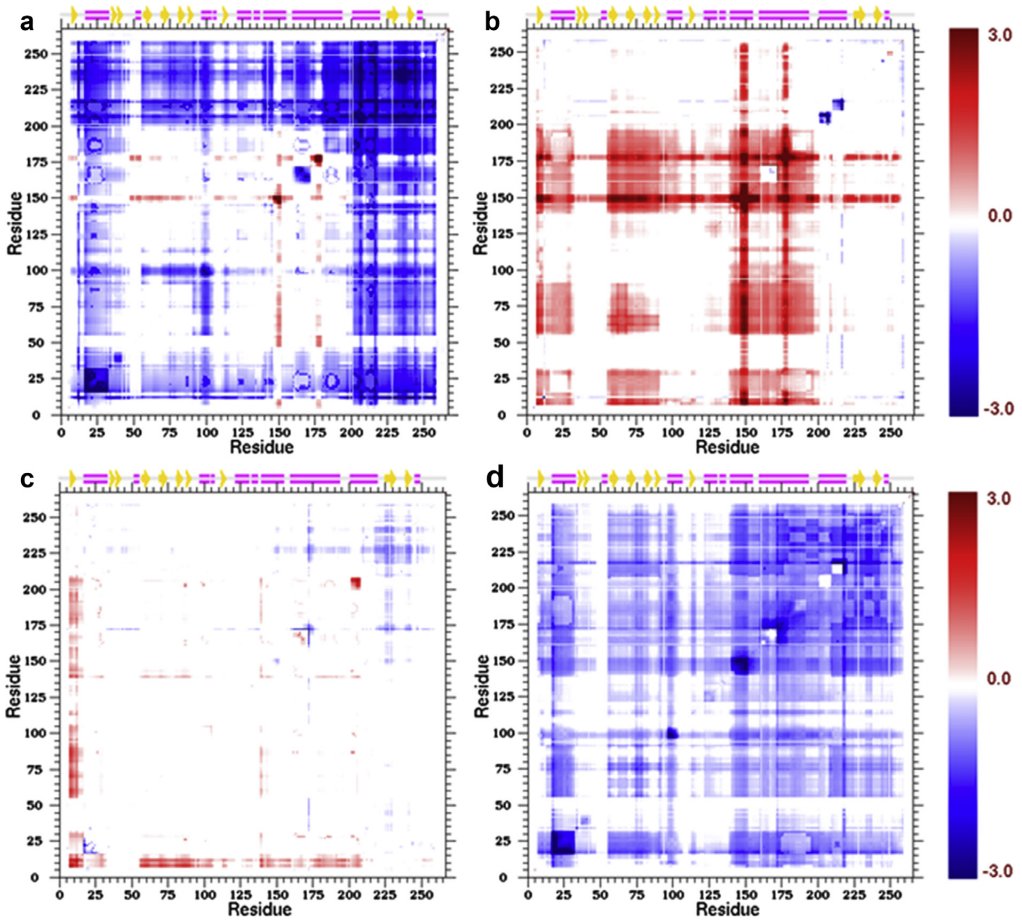

Figure 7: **Cooperativity correlation difference plots of RiVax**. Correlation in flexibility and rigidity between residue pairs is described by CC-plots. Differences in CC-plots are shown for mutant RB relative to RiVax in the native state (a) and the transition state (b). Difference CC-plots are shown for mutant RC relative to mutant RB in the native state (c) and the transition state (d). The coloring scale is the same for all panels, representing the SBNR as defined in the text. White regions correspond to places that are uncorrelated or are filtered out because of weak correlations with magnitude less than baseline noise. Red and blue regions correspond to correlated flexibility and correlated rigidity, respectively. Note that the number 3 on the color scale corresponds to a signal strength that is 3 standard deviations beyond noise.

Despite this broad range of rigidification, 2 localized regions around residue 150 and 177 become more flexibly correlated to one another compared to RiVax. It is worth noting that in all 5 mutants, the regions 18-32, 98-103, and 210-218 become more rigidly correlated. In contrast, RB has much greater flexibly correlated motions relative to RiVax

in the transition state within regions that span residues 7-30, 55-90, and 137-200 (Fig. 7b), in addition to the 2 localized regions around residue 150 and 177. Because the difference CC-plot between RC and RB in the native state (Fig. 7c) shows up as mostly white, this indicates that RB and RC share similar characteristics in the native state. However, the region that spans residues 7 through 17 and the region that spans residues 202-205 is more flexibly correlated and is coupled to one another. Within the region that spans residues 210-250, RC mechanically stabilizes the native state the most, although RB is a close second. The difference CC-plot between RC with respect to RB in the transition state (Fig. 7b) clearly shows that RC also mechanically stabilizes the transition state more than RB. Interestingly, RC is most similar to RB in the native state and most similar to RA in the transition state. The special property about RC is that it appears to be most effective in mechanically stabilizing both the native state and transition state among all 5 mutants analyzed. An effective way to visualize the greatest changes in residue pair correlations is to cluster the most significant changes in correlations within a difference CC-plot (i.e., elements with SBNR>1) and map these changes onto the protein structure (Fig. 8). Juxtaposition is made between mutants RB and RC to visualize on the structure how they differ from RiVax in both the native and transition states. Although RC is slightly more effective than RB in increasing mechanical stability relative to RiVax in the native state, the effectiveness of RC in increasing mechanical stability is much greater than RB in the transition state. Therefore, mutant RC is found to be the most effective alternative mutant for mechanically stabilizing both the native and transition state, while providing the greatest thermodynamic gain in stability together with a comparatively large kinetic free-energy barrier.

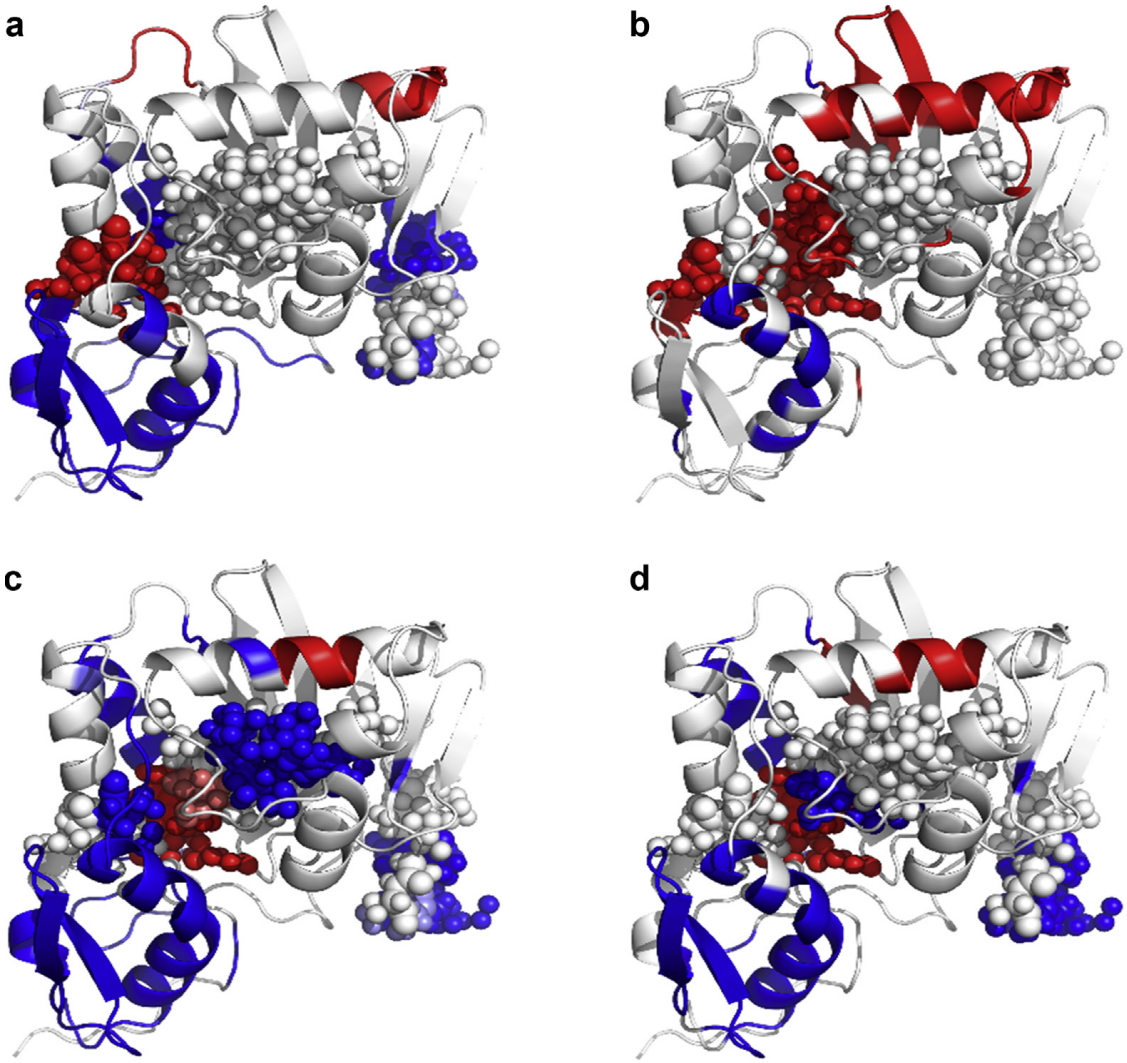

Figure 8: **Extreme backbone cooperativity correlation**. The most significant correlations obtained from difference CC-plots are rendered onto protein models (mutated from PDB: 3SRP) for RB relative to RiVax in the native state (a) and the transition state (b), and for RC relative to RiVax in the native state (c) and the transition state (d). The color scale represents the SBNR and inherits the same color scale used in the difference plots. The small spheres show backbone atoms that fall in immunologically important regions.

3.3 Discussion

The poor immunogenicity of recombinant nonoligomeric protein antigens constitutes a significant impediment to the development of a number of subunit vaccines for biodefense, including ricin toxin(67). We used a computational strategy as a means to

produce derivatives of a leading ricin toxin subunit vaccine, RiVax, with an enhanced capacity to elicit TNAs in mice following a prime-boost regimen. Of the 5 RiVax derivatives that evaluated, the most promising was RB, which carries the 3 point mutations, V81L, C171L and V204I, in addition to RiVax's original V76M and Y80A mutations. The improvement of RB over RiVax was most evident in the low dose immunization studies in which 6 of 8 of the RB-immunized mice had detectable TNA, whereas only 1 of 8 of the RiVax-immunized mice had detectable TNA. After the second boost, 8 of 8 mice immunized with RB had detectable serum TNA as compared to only 2 of 8 in the RiVax-immunized mice. The closely related RC mutant, which differs from RB only by the substitution of Ile at 81 that renders the protein slightly more thermostable was also significantly more immunogenic than RiVax, although no better than RB when tested side-by-side in mice.

To understand why RB and RC were more immunogenic than RiVax, a comparative QSFR analysis was performed. We posit that the flexibility of domains might influence the immunogenicity of specific regions, permitting access to epitopes that are closer in structure to native RTA. Previously, immune-dominant regions on RTA had been identified in the presence of mouse and rabbit sera(103). The QSFR analysis showed that overall backbone flexibility was not altered significantly by any of the mutations, except that RA, RB, and RC increased flexibility at residue 180. However, for residue-pair couplings, it was found that RB and RC mechanically stabilizes both the N- and C-termini regions through correlated rigidity in the native state more than all other mutants, and this rigidification extends to residues 98-103 that overlapped with immune-dominant regions II and V. In addition, RB and RC increased correlated flexibility

between residues 150 and 177 in the native state related to immune-dominant regions III and IV, respectively(103). Even small differences in the flexibility or accessibility of these domains could have a profound effect on the "quality" of antibodies elicited on immunization, considering that we estimated that roughly 90% of RTA-specific B-cell epitopes are conformational (discontinuous) in nature(103,104). Alternatively, the packing mutations could influence antigen stability in situ (e.g., limit protease sensitivity) or even after adsorption onto aluminum salts, thereby prolonging the duration of antibody release and/or deposition in tissue(105). Further examination is required to detail the residues that were mutated in RB and RC and how they might affect overall immunogenicity and/or stability of RiVax. Residues 13-25, including V18 have been tentatively identified as being the target of several so called Cluster 2 toxin-neutralizing monoclonal antibodies such as SyH7(106). Residue V81 is situated within b-strand F and next to a key residue associated with the active site, Y80. We have recently identified a monoclonal antibody that associates with T80 and possibly V81, thereby evoking a possible role of this residue in antibody recognition(107). Residue C171 is located within a-helix E, part of which is the target of several known neutralizing mAbs, including GD12. However, it should be noted that C171 is buried and therefore not surface accessible so it is unclear whether mutations at this site act locally or distally. Residue V204 is situated within a-helix G. There are no known B-cell epitopes within this region, probably because the region is buried and not surface accessible. Finally, residue S228 is situated within the C-terminus of RTA that normally interfaces with RTB. This region is devoid of secondary structure, but it has been suggested that the C-terminus of RiVax (in the absence of RTB) is readily unfolded and promotes instability of the protein. It is

possible that mutation S228 augments immunity by stabilizing the labile nature of the C-terminus. Regardless of the actual mechanism by which mutations within RB and RC potentiate the antibody response to ricin, further studies on these antigens are warranted in rabbits and nonhuman primates. These studies will determine whether the novel RiVax derivatives actually confer a benefit worth the resources required for advanced development.

Chapter 4 Sleeping Beauty Transposase

4.1 Introduction of Transposition and Sleeping Beauty

Traditional drug discovery pipelines are being replaced with new cutting edge methods that reduce administer complexity, complications, and cost. A bright star in this new revolution is gene therapy methods in which the genetic cause for the phenotypic symptoms is replaced all together. The Sleeping Beauty (SB) Transposon/Transposase system is at the forefront of these methods and has already shown to be effective in broad application methods. SB is reconstructed from an ancient Tc1/mariner that is active in vertebrates including humans(108,109). So far the SB System is being considered to develop treatments for Alzheimer's disease, Lymphoma, age-related macular degeneration (AMD) due to the success in pre-trails(110-114). In addition, SB is advancing promising areas of research like reprograming somatic cells into induced pluripotent stem cells (iPSCs)(115-118). This gives promise to a whole new generation of molecular medicine and may unlock the next tier in human health. SB and its nearest homolog Mos1 fall under DD[E/D]-transposases and more specifically fall into the RnaseH-like fold regime.

DD[E/D]-transposases mediate cut-and-paste DNA transposition, where they cleave DNA to excise and reinsert the transposon DNA into the host genome. Due to the presence of three catalytic residues (D, D, and E or D) and the RNase-H like fold that brings the catalytic residues into close proximity for catalysis(119-123). DD[E/D]-transposases belong to a widespread and diverse RNase-H like (RNHL) superfamily that includes enzymes involved in replication, recombination, DNA repair, splicing, (retro)transposition of TEs, RNA interference (RNAi) and CRISPR-Cas immunity

(124,125). Transposases are the most abundant RNHL proteins (124). While the amino acid sequence and structural divergence of RNHL proteins are substantial, there is significant similarity in the architecture of the catalytic core and the catalytic mechanism(126,127).

The RNase-H fold core motif contains a central β-sheet and three α-helices. The β-sheet consists of five strands ordered 32145 with the second strand in an antiparallel arrangement to the other four strands. The α-helices flank the β-sheet. Two α-helices are located on one side of the β-sheet and are inserted between the β-strands in the amino acid sequence and the third α-helix is located on an opposite side of the β-sheet and either immediately follows β-strand 5 or is separated from it by an amino acid insertion of variable size. Other secondary structure elements, in particular α-helices, can be present on both sides of the central β-sheet, however, they are not conserved throughout the RNHL superfamily and can have different positions and orientations with respect to the core motif (124).

Generally protein flexibility is critical to the mechanistic pathways of proteins, and in the cases of Sleeping Beauty and other RNHL proteins this is especially true. Between the RED, PAI, and Catalytic subdomains of SB long super flexible loop regions connect the more structurally significant subdomains like beads on a string. This high intrinsic flexibility allows SB to wrap around the DNA to bind the PAI subdomain to the IR regions of the target genome, while simultaneously positioning the RNHL subdomain into position for action. Within the Catalytic domain itself sits the RNHL fold that cleaves and reconnects the complementary DNA strand. This transposition is a two step hydrolysis of which conformation rearrangement is thought to occur (128). Upon binding

with DNA in Mos1 the highly flexible clamp loop interacts distal with respect to the catalytic site and rigidifies after interaction. Initial flexibility of the loop is crucial to Mos1's ability to interact in dimeric form(128), which is the biologically relevant subunit for activity. This importance has been demonstrated by mutation W119P which disabled the second cleavage event of transposition (128). This produces a shorter than expected 70mer which has been suggested to be a result of misdirected active site catalysis (128). Across the RNHL structures a loop region sits just prior to the third catalytic residue and crystallography shows this loop in Mos1 and other DNA transposases to prefer a small number of conformations of which is moves between (129).

In contrast to this flexible yet limited conformational space retroviral structures that share the RNHL fold have a much higher degree of disorder in this loop region (130). The reason is due to lack of interactions between the loop region and the N and C termini, which stabilizes the transposase loop region. This was confirmed by NMR relaxation measurements (131). Therefore transposases are believed to have a well defined catalytic pocket without the inclusion of DNA while the Integrase family has more disorder opening the door as a target for inhibitors and other drug design methods. For this reason Inhibitors including Ralegravir(132) and Elvitegravir(133) have been developed to target this disorder catalytic pocket prior to DNA interaction. Furthermore the flexibility of the catalytic residues plays a role in the orientation of the two metal ions positioned in the catalytic site, which are necessary for function. While the specific metal ion may be interchanged between Mn^{2+}, Mg^{2+}, and Co^{2+} in some instances(134) as it effects activity, the ions have been shown to significantly effect the stability of the protein through calorimetry experiments (135). However previous computational studies

have shown little difference between systems with and without ions present (136). For this reason we have chosen to exclude metal ions since they are not present in the x-ray structure of 5cr4 which is currently the only xray structure of Sleeping Beauty or any of its direct mutants. On a larger scale overall flexibility and the ability of the catalytic residues to move are directly connected to the function of various examples within the RNHL fold family including HIV-1 Integrase and Ribonuclease H (137-140). Previous studies from this group have already shown the link between correlated flexible regions of RNase H proteins and evolutionary stability indicating possible functional importance (141).

The SB System is broken into two pieces: first of which being the transposon, which carries genetic cargo of interest into the target genome. The second is the transposases, the protein responsible for cleaving and inserting the genetic cargo into the target DNA. SB Transposase itself is defined in three regions: the PAI domain which is responsible for DNA recognition and binding, the RED domain which also acts in DNA interaction, and finally the Catalytic domain which is responsible for the breaking of the DNA complex for insertion. Due to the highly flexible nature of SB and its mutants a full-length structure has not been crystallized, however NMR has now solved both PAI and RED independently(142-144) after efforts made here.

4.2 NMR

To date no complete experimental structure exists for SB or any of its mutants. This partially due to its highly flexible nature and the division of the three subdomains PAI, RED, and Catalytic mentioned previously. However pieces of the structure have

now been pieced together using various methods after publication of our NMR structure. Previously collaborators had solved the NMR structure for the PAI subdomain using CD spectroscopy. As a follow up we employed various methods to enhance the native stability of the RED subdomain in order to solve the NMR solution for the second subdomain. The instability of the three helical subdomains was attempted to be stabilized by altering pH conditions, and adding ionic and non-ionic crowders.

Previously collaborators have shown that the DNA-binding domain of SB transposase does not form a stable structure at physiologic conditions(142). However, the reaction of transposition occurs in cell nucleus, in extremely crowded environment. Therefore, we first investigated whether the crowding induces folding of the RED subdomain. Two agents have been used to mimic the effect of crowding, polyethylene glycol (PEG 6000), and Ficoll-70. These crowding agents are expected to increase the compactness of a protein due to excluded volume effects(145,146). Figure 10(A,B) shows the far UV circular dichroism (CD) spectra of the RED subdomain in the presence of increasing concentrations of PEG 6000 (panel A) or Ficoll-70 (panel B). CD spectra demonstrate that the RED subdomain alone does not have significant secondary structure. Furthermore, the presence of PEG 6000 or Ficoll-70 does not noticeably increase the secondary structure in the RED subdomain at the concentrations used in this study. Figure 10(C,D) shows 2D [1H,15N]-HSQC spectra of the RED subdomain. The narrow distribution of cross-peaks in the proton dimension indicates that the RED subdomain alone is disordered (panel C). The peaks are broadened, which could be caused by the conformational exchange between different disordered states or by association. In the presence of Ficoll-70, peaks in the [1H,15N]-HSQC spectrum of the RED subdomain

become significantly sharper. Such spectral change suggests that the conformational space sampled by the RED subdomain (between different disordered or oligomeric states) is reduced in the presence of crowding. However, the majority of cross-peaks remain in their original positions and the chemical shift distribution remains narrow. Together, the CD and NMR data indicate that the RED subdomain remains disordered at crowded conditions.

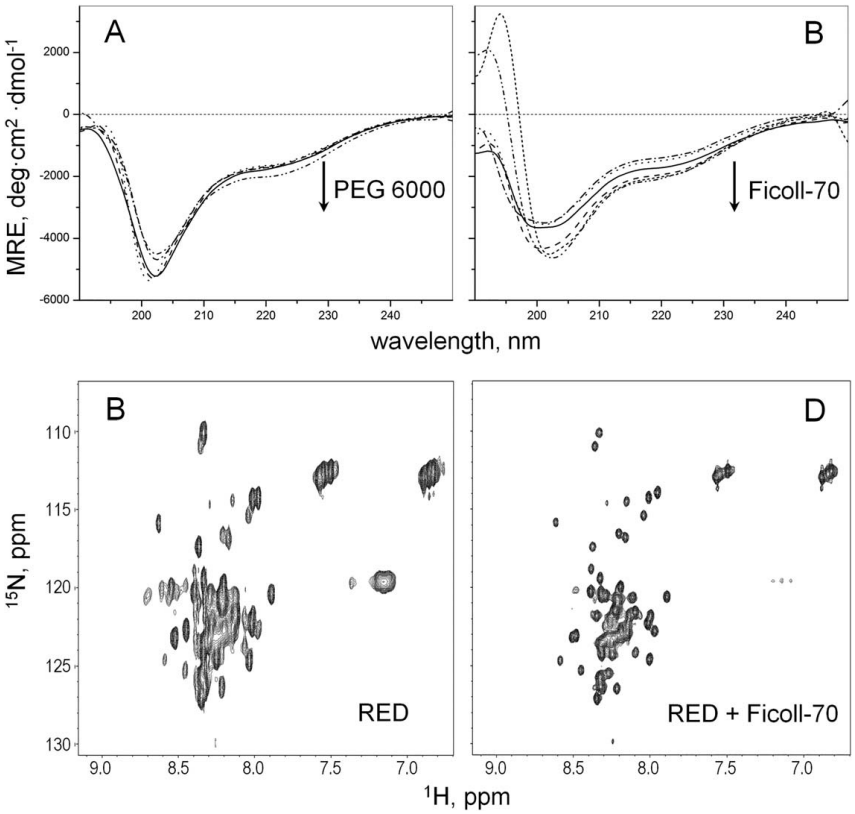

Figure 9: **Folding of the RED subdomain in the presence of crowding**. Far UV CD and [1H,15N]-HSQC NMR spectra of the RED subdomain in the presence of PEG 6000 (A,B) and Ficoll-70 (C,D) were collected in 20 mM aqueous MES buffer at pH 5.0. Increasing the concentration of crowders does not induce a significant amount of alpha-helical structure. Arrows indicate increasing concentrations of crowders.

Next, the effect of salts on the folding of the RED subdomain has been investigated. It is known that kosmotropic ions tend to precipitate proteins and stabilize

their structure, whereas chaotropes tend to enhance protein solubility and favor denaturation(147). This behavior is more pronounced for anions than cations, and the typical order for the anion Hofmeister series is CO22 3 >SO4 2–>H2 4 and for the cation series is K1>Na1>Mg21>Ca21 (kosmotropes on the left), where chloride is usually considered the dividing line between these two types of behavior. Accordingly, we tested the effect of Na2SO4, KCl, NaCl, and NaClO4 on the folding of the RED subdomain. Figure 10 shows the results of titration experiments with these salts. Far UV CD spectra of the RED subdomain show that the characteristic features of alpha-helical secondary structure emerge upon the addition of all salts, i.e., two negative bands at 208 nm and 222 nm and a positive band at 193 nm. Spectral changes are consistent with the observation that positively charged proteins follow the direct Hofmeister series at high salt concentrations (above 200– 300 mM monovalent salt)(148), that is, the addition of NaClO4 has the least effect on the structure of the RED subdomain. The addition of Na2SO4 leads to the largest amount of alpha-helical secondary structure as judged by the magnitude of the ellipticity at 222 nm (Fig. 10). Furthermore, in contrast to Na2SO4, some reduction in intensity of CD spectrum is observed after the addition of KCl, NaCl, and NaClO4, which can be taken as the indirect evidence of self-association of the RED subdomain. Therefore, Na2SO4 was selected for NMR experiments, which led to the structure 5UNK (144).

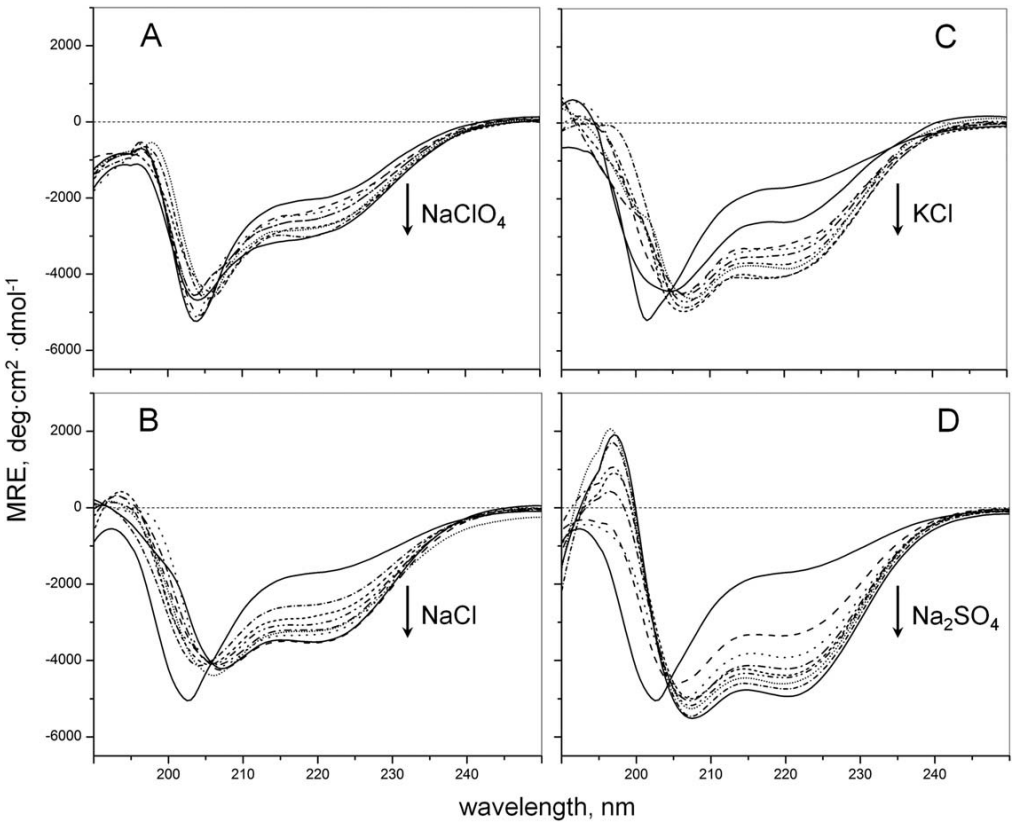

Figure 10: **Folding of the RED subdomain in the presence of different salts**. Far UV CD spectra of the RED subdomain collected in 20 mM aqueous MES buffer at pH 5.0 in the presence of up to 800 mM of NaClO4 (A), NaCl (B), KCl (C), and Na2SO4 (D). Arrows indicate increasing concentrations of salts. The largest content of alpha-helical is observed in the presence of Na2SO4.

4.3 Structure of SB100X

With each of the three subdomains independently solved computational methods where used to pieces together full-length representations of SB. In addition a crystal structure for the catalytic domain of a hyperactive mutation of SB (SB100X) has recently been solved. This mutation shows 100-fold hyperactivity with respect to SB wild type(149). Within the catalytic domain of SB100X it differs from SB wild type by six mutations. Four of which are sequential located on an exterior alpha helix RKEN214-217DAVQ, one sitting directly next to central catalytic residue 244, M243H residing on a

core beta sheet, lastly T314N sitting on the surface of SB100X. Each of these mutation sites has previously shown hyperactivity by themselves(150). In addition to these, three more mutations that lay outside the catalytic domain together that separates the full length SB wild type from SB100X.

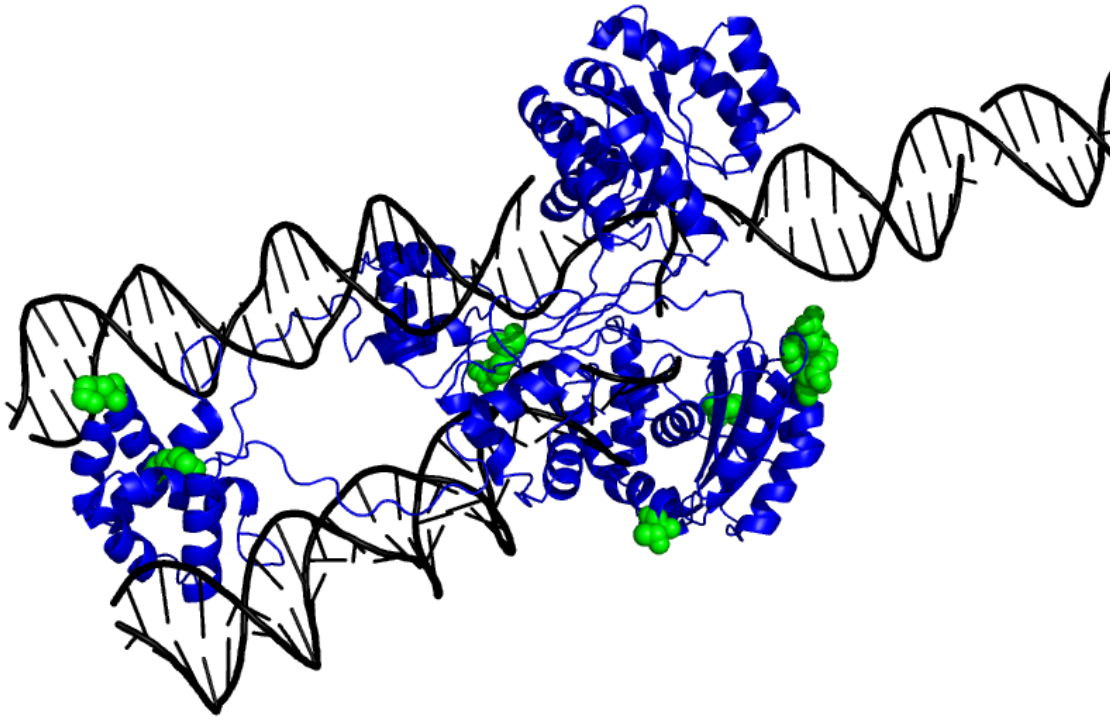

Figure 11: **Full length dimer structure modeled into DNA**. Green regions indicate point mutations that separate SB100x from SB.

As effective as SB has been in trials thus far one of the biggest trade offs for using non viral vectors is potency. Compared to viral vectors SB has lower risk of activating oncogenes near insertion sites, and non viral methods are cheaper and safer to produce. However due to the transient nature and slower transposition rate of non viral vectors the dosage of transposon DNA is much higher(151). SB100X can bridge this gap and reduce the dosage load for treatment and has even shown to rival viral output(150). By better

characterizing the mechanical properties of SB100X the next iteration of gene therapy methods can be even more powerful through more intelligent transposase design.

4.4 Computational Methodology

We used the x-ray crystallographic structure of the catalytic domains of SB100X transposase(149) (pdb code 5CR4) and of the Mos1 transposase(152) (pdb code 2F7T). For comparative analysis of Sleeping Beauty wild type with its hyperactive counter part SB100X a set of seven structures was created. Each these structures represents a single or combination of the six point mutations that separate SB wild type for SB100X. Figure#12 depicts the mutations for each of these structures. Each structure was developed by mutagenesis in PyMol (153) and subsequently minimized in GROMACS under an AMBER99SB-LN force field and TIP3 water model(154). Structure for Mos1 was prepared in identical manner. Several of these intermediate mutations between the wild type and SB100X were functionally screened at the advent of SB100X(150) and their experimental functions are denoted on Figure#12.

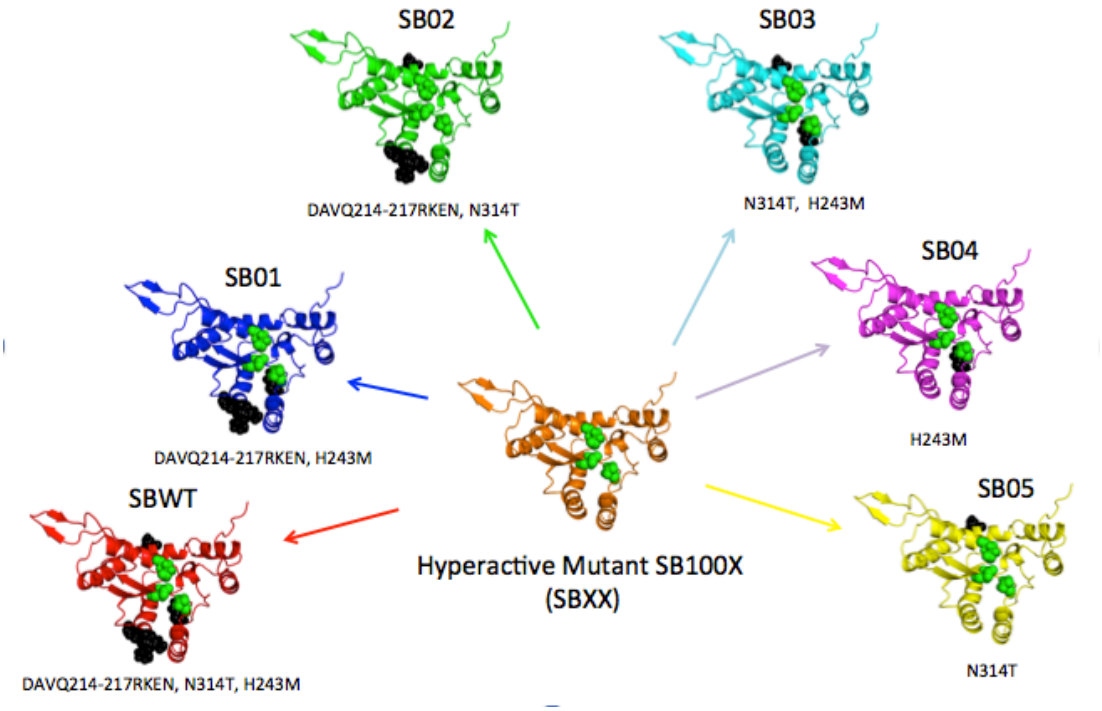

Figure 12: **Mutation steps from SB100X to SB** achieved through PyMol mutagenesis. Each of these structures were the starting point for thermodynamic and mechanical calculations by the mDCM. Green residues show catalytic pocket.

Furthermore the same crystal structures for SB wild type and SB100X were used to construct ten additional mutation structures based off of functional assay data from experiment(155). Ten point mutations were selected from the original eighty based on their activity modulation, location with respect to the active site and the presumed reason for activity alteration. These structures are summarized in TABLE 2 along with their experimental assay values.

Table 2: **Mutations selected for hyperactivity** study with corresponding functions measured experimentally at 300K.

Mutation	Functionality
F198A	0
C197A	5
E216K	5
N217K	10
A205K	60
A205P	100
T203V	400
N217H	450
D210E	700
H207V	700

Each of the structures defined above represent only the catalytic domain of each respective mutant as the PAI, RED, and Catalytic domains area all separated by large flexible loop regions. For the purposes of rigidity calculations we consider each domain independent from each other. Additionally we focus on the monomer despite transposases functionally dimeric nature due to the separation of the dimer interface of the catalytic domains(155).

4.5 Analysis of RNase-H like Fold

Here, we investigate the flexibility and rigidity of two representative DD[D/E]-transposases with the greatest structural similarity, Mos1 and SB. We examine the RNase-H like fold as it is a common motif of DD[D/E]-transposases and RNHL proteins, and also compare protein regions outside the RNase-H like fold. Although, these regions show a great structural diversity between the members of the RNHL family, they are similar in Mos1 and SB transposases.

The RNase H-like fold consists of a central β-sheet comprised of five β-strands (ordered 32145 with the second β-strand B2 antiparallel to other strands) and three surrounding α-helices (Figure 2 A). Two of the three α-helices (H1 and H2) are inserted in the amino acid sequence between β-strands B3 and B4 and B4 and B5, respectively, and are located on one side of the β-sheet, whereas the third α-helix (H3) is located on the opposite side of the β-sheet and either immediately follows β-strand B5 immediately follows β-strand B5.

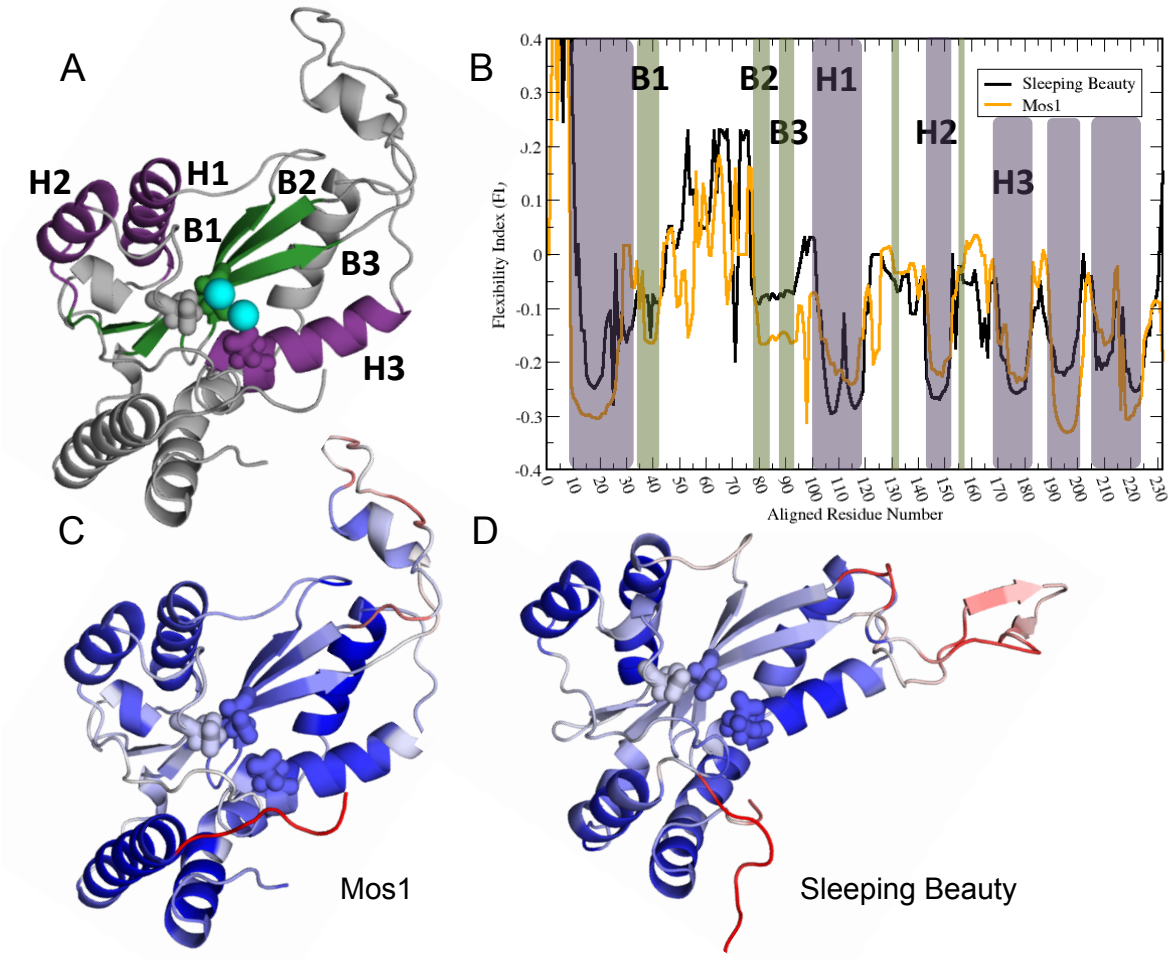

Figure 13: **Mechanical Characteristics of RNase H-like fold**. A) Conserved RNase H-like regions are highlighted in purple (helices) and green (sheets). In addition most transposases work in conjunction with metal ions present in catalytic pocket (cyan). B) Backbone flexibility index (FI) of SB and Mos1. C-D) Backbone FI mapped onto three-dimensional structure. Blue regions indicate rigidity and red flexibility.

Figures 13 B-D show the backbone flexibility index (FI) of Mos1 and SB catalytic domains, as a function of the residue number (B), as well as mapped onto their three-dimensional structures (C-D). Blue color corresponds to negative FI values equivalent to the presence of redundant constraints for a given residue that is determined as more rigid as compared to more flexible or under-constraint residues shown by red color and

positive FI values. As expected, more negative FI values correspond to more rigid secondary structure elements, within or outside the RNase H-like fold, as compared to loops connecting them. The FI profile is similar in Mos1 and SB transposase, indicating that the flexibility and stability trends are common between them. It is interesting that both transposases are generally rigid as revealed by mostly negative FI values and blue color. There are a few exceptions, in particular the long clamp loop that extends out from the catalytic domain and is disordered in the x-ray structure of Mos1or SB catalytic domains solved without DNA (149,152). It is striking that the central β-sheet of the RNase-H like fold is less rigid, more so in SB transposase, in comparison to the surrounding α-helices that are either part or outside the RNase H-like fold, where the shadows of blue correspond to a different degree of rigidity/flexibility with darker color corresponding to a greater rigidity. This unanticipated, yet logical result shows that the flexible core of the protein containing the catalytic site is enclosed by more rigid surroundings forming a cage-like structure, which is likely needed to allow the rearrangements in the catalytic site and its vicinity to accommodate the DNA molecule. The catalytic residues belong to regions with different flexibility and display a similar trend in both transposases. The first catalytic residue, D156 in Mos1 or D153 in SB transposase, located at the end of strand B3, is rigid. In comparison, the second catalytic residue, D249 in Mos1 and D244 in SB transposase, located on a loop joining strand B4 to helix H2, is considerably more flexible. The second catalytic residue is the most flexible of the three catalytic residues. The third catalytic residue, D284 in Mos1 and E279 in SB transposase, located at the edge of helix H3 of the RNase-H-like fold, is the most rigid of the three. Interestingly, namely this residue alternates within the DD[E/D]

motif of DD[E/D]-transposases. These results support our idea that transposases have to be sufficiently flexible to carry out the multi-step reaction of the transposition.

While FI profile is similar in Mos1 and SB transposases, the SB transposase is a more flexible molecule overall. In particular, the beta-sheet and helix H3 of the RNase-H like fold are more flexible in SB transposase. The trend also extends to the second and third catalytic residues, indicating that the constraint network of Mos1 is more densely packed in the core of the protein, while SB has a more distributed network. One of the reasons for higher rigidity of the third catalytic residue of Mos1 versus SB (D284 and E279) could be that a shorter side chain of aspartic acid promotes better molecular packing.

4.6 Effect of DNA binding on the flexibility and rigidity of Mos1 and SB transposases

Upon DNA binding we see the net increase in backbone rigidity. This increase in rigidity is due to the increase in number of constraints on residues, especially in the vicinity of the DNA binding site. While the global shift towards rigidity is expected, it is not a trivial result. The DNA contacts outlined in literature (128,149) contain only a few contacts on a single side of the protein, whereas we see the change in backbone rigidity across the entire protein. The comparison of DNA-bound and DNA-unbound states of Mos1 and SB transposases shows that the increase in rigidity is not uniform (Figure 14). Comparatively, some protein regions gain more rigidity and some gain flexibility, indicative of a global constraint network rearrangement to compensate for the addition of new constraints. The dramatic loss of flexibility is observed in the clamp loop that forms multiple protein-protein and protein-DNA contacts (128). Additionally, flexible loops

joining secondary structure elements of the RNase H-like fold, B3 and H1 or B4 and H2 or B5 and the following helix (not part of the RNase H-like fold), gain rigidity. The central β-sheet of the RNase H-like fold motif also becomes more rigid. In contrast, the helices of the RNase H-like fold and other α-helices flanking the β-sheet become more flexible with the exception of N-terminal helix of the catalytic domain leading to the RNase H-like fold, either the entire helix or its part. At the catalytic site, the most dramatic change (significantly increased rigidity) is observed for the second catalytic residue D249 or D244 in Mos1 or SB transposase, respectively, indicating that the rearrangement of the constraint network caused by DNA binding results in its fixed conformation. No change is observed for the first catalytic residue D156 or D153, originally already rigid. The third catalytic residue shows a different trend, e.g., it becomes more flexible in Mos1 and more rigid in SB transposase, although the change is marginal.

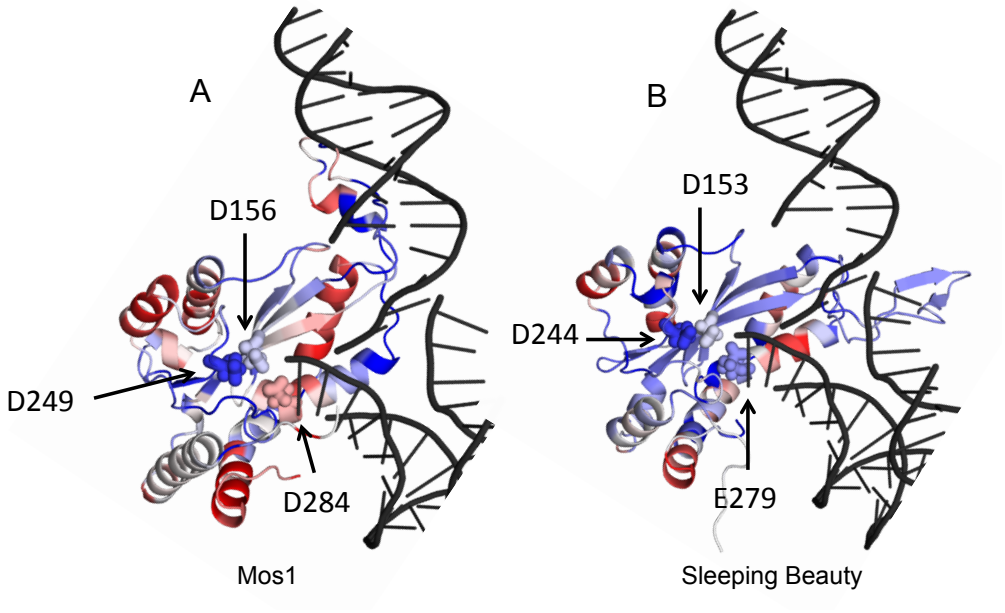

Figure 14: **Difference in Backbone Flexibility Index** between DNA bound and unbound states for Mos1(A) and SB(B) respectively. Blue regions indicate increase in rigidity upon binding, conversely red regions indicate loss of rigidity upon binding.

While the observed trend is true for both Mos1 and SB transposases, the magnitude of the rigidity fluctuations differs. Upon the interaction with DNA, the beta sheet rigidifies more in SB than in Mos1 transposase, and the flexibility of helices H1 and H2 increases more in Mos1 than in SB transposase. We note that SB starts off as a more flexible molecule at the core, and this could be one of the reasons for the different degree of flexibility/rigidity changes observed after DNA binding.

4.7 SB100X Hyperactivity and Rigidity Mechanics

Next, we are interested to determine whether there is a link between the flexibility/rigidity of the transposases and their activity. In this regard, the SB transposase is an ideal model system to study, because the large set of mutagenesis and functional data is available (150,155). The optimization for enhanced transposition activity using

mutagenesis and molecular evolution approaches yielded a hundred times more active version of SB transposase (150), SB100X. SB100X differs from the original SB transposase only by nine mutations. These mutations are distributed throughout the molecule, and, therefore, only minor structural differences are expected (149). However, the role of changes in dynamics and mechanical properties of the protein that could also have a significant effect on protein activity has not been investigated. Within the catalytic domain, SB100X differs from the original SB transposase by six mutations: RKEN214-217DAVQ sequential mutations, located at the beginning of helix H1 of the RNase H-like fold, M243H located on strand B3 next to the catalytic residue D244, and T314N located on the surface of SB100X outside the RNase-H-like fold (Figure 15A). Outside the catalytic domain, SB100X has three additional mutations (K14R, K33A, and R115). In this study, to investigate the relation between the flexibility/rigidity and function of the SB transposase, we selected six hyperactive mutations (RKEN214-217DAVQ, M243H, T314N) located in the SB catalytic domain. We focus on the catalytic domain, because the DNA-binding domain is joined to the catalytic domain by a long and flexible linker, and the two domains remain distantly separated in the protein-DNA complex, limiting the flexibility/rigidity signal propagation from one to another.

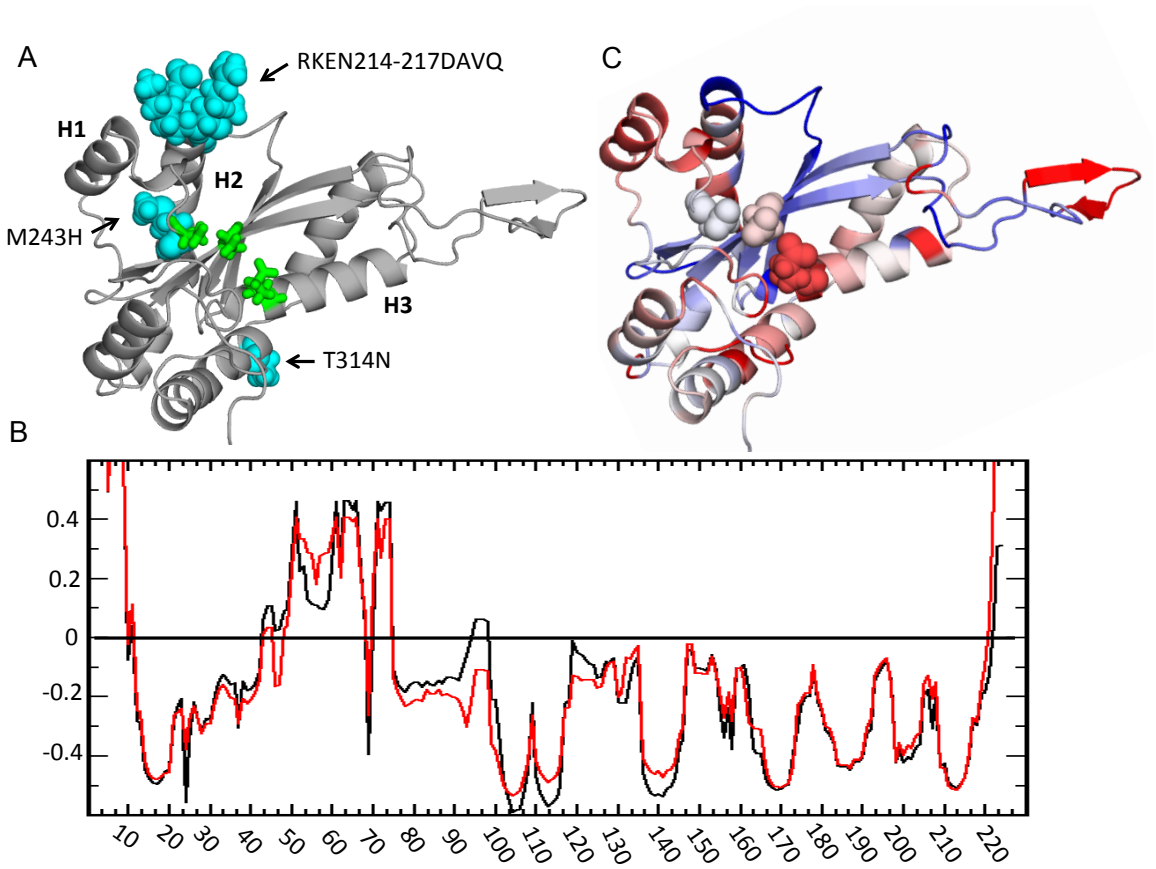

Figure 15: **Hyperactive SB100X**. A) Regions highlighted (cyan) indicate location of 6 point mutations in catalytic domain. B) Backbone FI of SB (black) and SB100x (red). C) Difference of backbone FI between SB100x and SB mapped onto ribbon diagram. Blue regions indicate SB100x is less flexible than SB, conversely red regions indicate more flexibility than SB.

Figure 15B shows the comparison of SB100X to the original transposase. Qualitatively, SB and SB100X transposases share the same FI trends with peaks and valleys in the same locations, however, SB100X exhibits less variation in these values. Quantitative difference of FI values mapped onto the structure of SB100X transposase (Figure 15C), reveals a few important changes in the flexibility/rigidity network of SB transposase introduced by hyperactive mutations. Upon mutation, the central β-sheet becomes more rigid in general. In contrast, the surrounding α-helices become more flexible with the

exception of the end of helix H3. This is the location of RKEN214-217DAVQ substitution. Upon mutations, the additional hydrogen bonding in this region increases mechanical rigidity, and the loop connecting strand B1 to helix H3 becomes significantly more rigid. One of the bonds, the H-bond between E216 and D210 formed after RKEN-DAVQ mutation, has been already revealed by the crystallographic structure of SB100X (149). The mDCM identified this additional bond to be present with the energy of -5.16 kcal/mol, suggesting that this bond is likely to be always present in the topology. In addition, the mDCM identified several weaker hydrogen bonds formed between D216–N217 (-3.22 kcal/mol), R214–E216, R214–K215, and R214 – I212, each of which are between 0-1 kcal/mol. These additional bonds are weak, but one may expect that at least one of these bond can be present at a time, providing additional constraints to increase the rigidity in this protein region. Interestingly, the most flexible of the three catalytic residues, D244, is not affected by hyperactive mutations, whereas the other two catalytic residues, D153 and E279, show increased flexibility. In particular, the catalytic residue E279 along with helix H2 and C terminal helix show the strongest increase of FI.

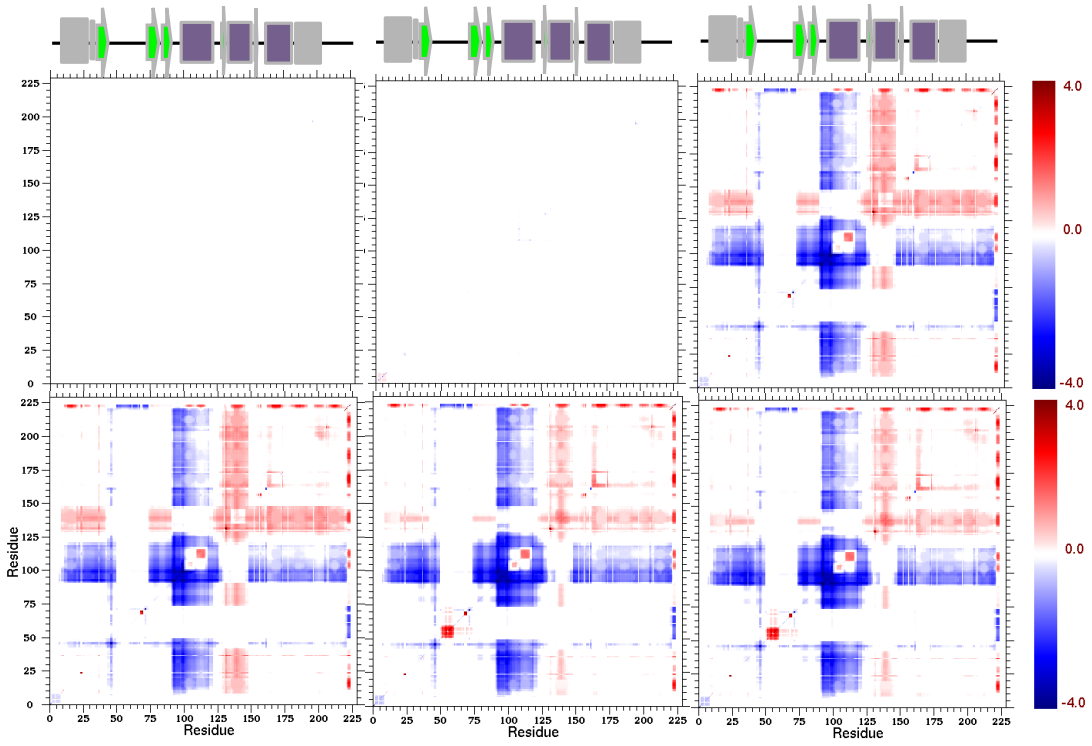

Figure 16: **SBNR Difference in Correlated Cooperativity** of each stepwise mutation from SB to SB100x from SB01-SBXX outlines in figure 12, (A-F respectively).

Differences in CC plots shown in Figure 5 reveal protein regions, in which mutations increase or decrease correlated flexibility (shades of red) or rigidity (shades of blue) between residues. These plots show changes in the constraint network, through which information can propagate within the protein. Figure 5A shows no significant (SBNR) difference in CC properties between the original SB and the T314N mutant. Previously, it was suggested that this mutation increases activity through solubility effects (149). Our results do not contradict this conclusion, and we can further suggest that solubility (protein-protein interactions) is the sole driver for hyperactivity in T314N and does not play a role in long range flexibility/rigidity network effects. Figure 5B represents M243H mutation, which is adjacent to the DDE catalytic triad of residues. Despite the seemingly

important position of M243H mutation and the potential formation of pi bonding between histidine residues as a result of this mutation (149), the changes in the constraint network are less than one standard deviation, which we rank as insignificant. This may indicate that the formation of an additional hydrogen bond in the DDE domain is redundant and potentially not as important as suggested by x-ray structure (149). Figures 5C-F show RKEN214-217DAVQ substitution alone (C) or in combination with T314N and M243H mutations. Qualitatively similar characteristics of these plots suggest that the RKEN214-217DAVQ substitution has the greatest influence on a mechanical framework of the protein that may alter its function. Even though DAVQ set of residues is neither spatially or sequentially near the active site or the DNA-binding domain, it introduces the largest change in protein mechanical properties. The ribbon above CC plots represents secondary structure elements colored in accordance with Figure 13A. The elements of the RNaseH-like fold are also labeled. The RKEN214-217DAVQ substitution leads to a much greater correlated rigidity of the loop joining beta strand B3 to helix H1, increases the rigidity of helix H1 (although to a lesser degree) and of a few residues following beta-strand B1, and increases the correlated flexibility of helix H2 positioned right next to helix H1. Helix H1 contains the DAVQ residues, the substitution of which causes an additional hydrogen bond to form between the helix and the adjacent loop (res 209-215) connecting residues several positions apart and rigidifying the entire area. This change results in existing bonds rearrangement in the helix to better connect other regions of the helix. Figures 5E-F also reveal that M243H substitution in combination with RKEN214-217DAVQ leads to a minor rigidification of helix H2 as demonstrated by less intense red color and to the increased correlated flexibility of the loop connecting helix H3 of the RNase H-like fold

to beta-strand B5. The latter loop, also termed the catalytic loop, was previously shown to be dynamic in other RNHL family members(129,130,156-158), and the loss of its flexibility was shown to be detrimental for function (159). Overall, our data identify the mechanical properties of the loop between beta strand B3, helices H1 and H2, and the loop between helix H3 and beta strand B5, all from the RNase-H like fold, as functionally important.

4.8 Additional Mutations

In addition to hyperactive mutations of SB100X, more than 80 mutations along with their effect on function, albeit much smaller, have been reported (155). Out of these reported mutations, we selected an additional set of ten single mutations (C197A, F198A, T203V, A205P, A205K, H207V, D210E, E216K, N217H, N217K) for this study based on the following criteria: (1) location in the vicinity of helix H1 so that the potential effect on function could be related to the perturbation of the residue constraint network in its vicinity; (2) opposite effects (increase or decrease of the transposition activity); (3) different substitutions of the same residue show adverse functional effects (residue is in orange) (155). These mutations are shown in Figure 17. The activity of these mutations is also indicated as the percent of the original SB transposase activity, where 100% corresponds to no change of activity (light cyan), whereas values smaller and greater than 100% correspond to lesser (blue) and higher activity (red), respectively.

As with SB100X transposase, the primary region of the global correlated flexibility/rigidity network of the molecule affected by mutations contains helices H1 and H2, and the loop connecting beta-strand B3 to helix H1. We note that despite the location of the majority of mutations on beta-strand B1 facing helix H1, a greater change for most

mutations is experienced by helix H2 highlighting a long-range propagation of the signal. Interestingly, as with SB100X transposase, helix H2 becomes more flexibly correlated to the rest of the molecule for activity increasing mutations.

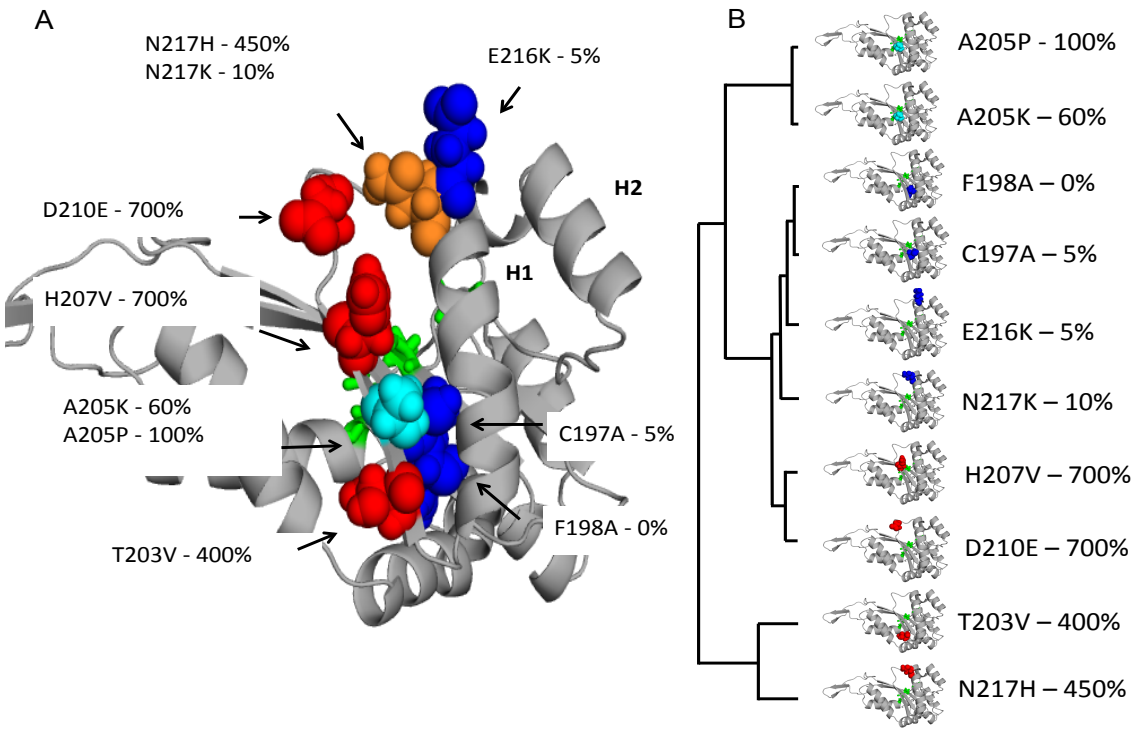

Figure 17: **Point Mutation Analysis**: A) 10 point mutations sampled from literature. Red residues indicate mutations with an increase in function, blue decreased in function, cyan are near wild type functionality, and orange belongs to both hyperactive and low activity depending on mutation. B) Clustering of difference in Flexibility Index of Beta Strands 1,2,3 and Helices 1 and 2.

Unlike hyperactive mutations of the SB100X transposase, none of the 10 mutations that we selected for additional analysis show the trend of increased correlated rigidity in helix H1 and the loop connecting beta-strand B3 to helix H1 to the rest of the molecule, and the activity-enhancing mutations (T203V, H207V, D210E, and N217H) moderately increase the correlated flexibility of helix H1. We identified the hydrogen bond between

two histidine residues H207 and H225 to be responsible for this effect. In the original SB transposase, the hydrogen bond between H207 and H225 exists with the energy of -2.14 kcal/mol. Seven of the mutations disrupt the local environment enough to reduce the bond energy to 0-1 kcal, and in the case of H207V mutation the bond does not form. Upon disruption, the backbone flexibility of H225 dramatically increases, leading to local rearrangement of the intrahelical hydrogen bonds of helix H1. In SB100X the HIS-HIS bond is reduced to -1.43, which makes its probability of existing higher than that of the 7 mutations. In addition no interahelical hydrogen bonds are added in SB100X like are added in other mutations. However 2 strong Hbonds connect the helix to the loop (previously mentioned) and a third connect the TYR218-HIS243 with energy -4.05 kcal/mol. M243H is one of the mutations that separates SB from SB100X and H243 was suggested by crystallographers to potentially participate in pi bonding with H249. However we detect a H-bond with high probability connecting H243 to Y218 instead. This connects the DAVQ loop area to a central beta sheet, and considering the orientation of all three strong hydrogen bonds, acts like tethers holding down H2. This may explain the large correlated flexibility exhibited by SB100X that is not found in other hyperactive variations. Interestingly, there is a correlation between the effect of the mutation on the correlated flexibility/rigidity of the protein and the transposition activity of SB transposase, which is best revealed by using high-dimensional clustering approach adapted by the software Clustergram (160-163). To be able to use the Clustergram, we first converted complex flexibility/rigidity information contained in FI index and CC-plots into single values. Therefore, we selected five segments of SB transposase showing changes upon mutation (beta-strands B1, B2, B3 and alpha-helices H1 and H2) and

averaged the difference between the mutant and the original transposase across all residues in each of these segments. The results represent a tree based on many solution trees spanning the native basin and demonstrated robustness across the native basin.

4.9 Discussion

We performed the analysis of flexibility/rigidity of two representative transposases from the DD[E/D]-transposase family, Mos1 and Sleeping Beauty, containing an RNase H-like fold that occurs in a diverse number of enzymes that belong to the RNase H-like superfamily(124). Our data show both transposases as generally rigid molecules possessing a highly constrained topology with globally conserved flexibility/rigidity profiles.

Within a generally highly constrained topology, the central β-sheet of the RNase-H like fold is less rigid than the surrounding α-helices forming a comparatively more flexible core enclosed in a cage-like structure. The trend reverses upon DNA binding where the central beta-sheet becomes comparatively more rigid than the surrounding helices. The second catalytic residue (D249 in Mos1 and D244 in SB transposase) is considerably more flexible than the other two catalytic residues in the absence of DNA. While it rigidifies upon DNA binding, the other two catalytic residues gain flexibility. These results support our statement that the RNase H-like fold of DD[E/D]-transposases must possess the flexibility sufficient enough to carry out the multi-step reaction of the transposition and agree with experimental data showing that the active site in other RNHL members is conformationally promiscuous (156,164,165) suggesting a conserved trend within the family.

Within the RNase H-like fold, we identified a smaller structural motif, which flexibility/rigidity properties respond to variations in the transposition activity. This motif comprises helices H1 and H2, located on the same side of the central beta-sheet, and the loop leading from beta-strand B3 to helix H1. Accordingly, our results suggest that modifications within this structural motif may be further explored for the optimization of SB transposition activity from the standpoint of flexibility/rigidity.

While the flexibility/rigidity of this structural motif correlates to the transposition activity, the trend is different in SB100X (10,000% of the original transposase activity) and in ten single point mutants that were studied here (0-700% of the original transposase activity). These results demonstrate several important points. First, the changes in the protein due a single residue substitution can propagate to distant sites and alter a global network of correlated motions. Second, while a single mutation can have a significant effect on the transposition activity, a combination of mutations can modify the global network of correlated motions differently and in a more optimal way for function. Finally, the mutations that do exhibit similar flexibility/rigidity profiles may be candidates for modulating functional activity from a dynamics standpoint; however, they may have other properties that affect function.

Chapter 5 Upgrading the DCM

5.1 Complete GRIDsearch

GRIDsearch has already been mentioned several times throughout this dissertation primarily because it has been the go to tool for applying the mDCM to proteins for which no heat capacity curve exists. This problem is only growing with the rapid increase of crystal structures being entered into databases with thermodynamic characterization falling behind. Originally this methodology was developed by Deeptak Verma, a graduate from the Jacobs-Livesay Lab, and has been a staple for the last five years. Traditionally fitting to a heat capacity yields the parameters u, v, and dnat, which were defined in the introduction to the DCM.

GRIDsearch guesses these three parameters in a 3D grid where each node represents a unique parameter set. While each calculation is relatively fast the number of calculations is equal to N_u x N_v x $N_{\delta nat}$ in the net cast by GRIDsearch. In addition we have found that we still have to create a finer mesh grid after the initial calculations to find the optimal solutions. In reality there is a lot of redundant calculation in this process. For each *u, v, δnat* parameter calculation actually calculates solutions in close proximity to the node. For simplicity imagine a 2D grid where each node has a cloud of solutions surrounding it. As the distance between the nodes decreases the overlap of the solution clouds increases. This overlap represents our redundant calculations. Therefore we have reconstructed the GRIDsearch algorithm to take advantage of these extra calculations. We now sweep across every possible *u, v* combination for a given *δnat*. Since nearby calculations are highly similar, the incremental shift made by each step of the algorithm

is nearly instantaneous. This creates a solution continuum and contains the results of every possible *u, v* combination with precision now limited by the rigidity calculations themselves instead of computational cost. This new methodology as tentatively been named completeGRID and was implemented in the analysis of Sleeping Beauty and its mutants in chapter 4 of this dissertation. It has produced new metrics to determine probable solutions to pass onto the mechanical calculations of the DCM.

One of the new features used in SB was the superposition of heat capacity peaks shown in figure 18. In absence of heat capacity completeGRID will generate potential Cp curves, however it is still valuable to have some experimental information about the subject. Thankfully collaborators have determined the melting temperature (Tm) of SB. In addition we have a typical range of Cp max values for proteins of this size. This information is used to discern probable solutions out of completeGRID from improbable ones. Figure18 below represents the Cpmax for each *u, v* node within our net. Looking

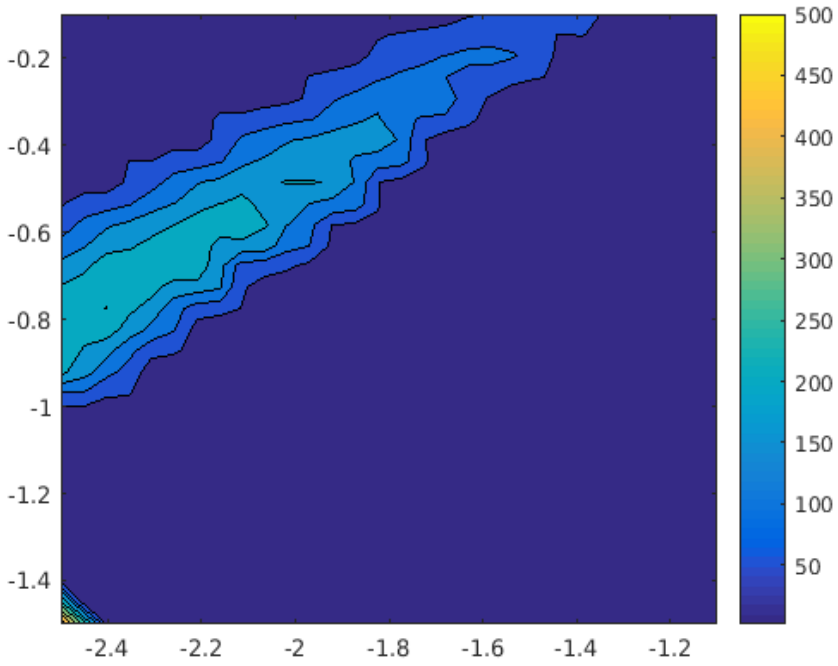

Figure 18: **Heat Capacity Screen**: Heat capacity maximum calculated by Cp reconstitution based on *u, v, δnat* parameters.

at this figure we can immediate narrow down the potential solution range. Everything in dark blue was filtered out as a poor solution and the Cpmax was found to be zero. However notice the line of light blue across the top. These peaks are closer to 100 Kcal/mol, where we expect the peak of a protein this size to be in. Since we don't know exactly what the Cpmax value should be we take all solutions that are within our accepted interval to pass onto the mechanical calculation steps and average the results for greater statistical robustness in our predictions.

Unfortunately the *δnat* term is not so trivially exploited, and it end the clearly the slow variable in this algorithm. Therefore we propose to employ an interpolation methodology to minimize the number of *δnat* sheets necessary to interpolate G across

δnat slices. This methodology will allow us to generate free energy landscapes, and then rigidity metrics for every possible *u, v, δnat* parameter with infinitesimal precision at very low computational costs. From an application standpoint this will enable greater sensitivity to small point mutation studies, or other family wide studies in which the signal may be smaller then the effect generated from GRIDsearch's discrete Δ*u, Δv, Δδnat*.

5.2 Future Application of the DCM

Throughout this dissertation I have outlined several major difficulties in applying the mDCM to modern large scale systems. We have spent considerable time circumventing these issues through our publication history. With small Chemokines we applied MD sampling to increase the conformational diversity that a specific *u, v, δnat* parameter set caused. With RiVax we found that even fitting to experimental heat capacities could result in highly precise parameters yet very unstable in their ΔG values. Both of these issues are automatically resolved with the emergence of completeGRID. With greater parameter sampling we are confident that the results are a better indication of the dynamic nature of proteins as well as not falling into a energy well that may not be physically probable.

completeGRID offers new found computation efficiency that allows that next generation of the BMPG to refocus efforts on increasing the scientific accuracy of the overall model, or use the current state to analyze large scale family studies. This will allow the DCM to remain a potent tool with distinct advantages over the status quo of MD and other computationally expensive methods.

5.3 Publication History

Table 3: **Publication History** Record of all previous and future publications.

Description	Journal	Date Published
Dynamics and Thermodynamic properties of CXCL7 Chemokine	**Proteins**	**August 2015**
Novel Ricin Subunit Antigens With Enhanced Capacity to Elicit Toxin-Neutralizing Antibody Responses in Mice	**Journal of Pharmaceutical Science**	**March 2016**
NMR solution structure of the RED subdomain of the Sleeping Beauty transposase	**Protein Science**	**March 2017**
Rigidity and Hyperactivity of Sleeping Beauty Transposase	**Proteins**	**July 2018**
Large-scale comparative quantitative stability/flexibility relationships	**Protein Dynamics: From fluctuations to function**	**June 2019**

References:

1. Zou, J., Mali, P., Huang, X., Dowey, S.N. and Cheng, L. (2011) Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. *Blood*, **118**, 4599-4608.
2. Ciampi, R., Romei, C., Pieruzzi, L., Tacito, A., Molinaro, E., Agate, L., Bottici, V., Casella, F., Ugolini, C., Materazzi, G. *et al.* (2017) Classical point mutations of RET, BRAF and RAS oncogenes are not shared in papillary and medullary thyroid cancer occurring simultaneously in the same gland. *J Endocrinol Invest*, **40**, 55-62.
3. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662-666.
4. Mark, A.E. and van Gunsteren, W.F. (1994) Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J Mol Biol*, **240**, 167-176.
5. Dill, K.A. (1997) Additivity principles in biochemistry. *J Biol Chem*, **272**, 701-704.
6. Jacobs, D.J., Dallakyan, S., Wood, G.G. and Heckathorne, A. (2003) Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems. *Phys Rev E Stat Nonlin Soft Matter Phys*, **68**, 061109.
7. Jacobs, D.J., Rader, A.J., Kuhn, L.A. and Thorpe, M.F. (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150-165.
8. Jacobs, D.J. and Dallakyan, S. (2005) Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J*, **88**, 903-915.
9. Dahiyat, B.I., Gordon, D.B. and Mayo, S.L. (1997) Automated design of the surface positions of protein helices. *Protein Sci*, **6**, 1333-1337.
10. Verma, D., Jacobs, D.J. and Livesay, D.R. (2012) Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged. *PLoS Comput Biol*, **8**, e1002409.
11. Verma, D., Jacobs, D.J. and Livesay, D.R. (2013) Variations within class-A beta-lactamase physiochemical properties reflect evolutionary and environmental patterns, but not antibiotic specificity. *PLoS Comput Biol*, **9**, e1003155.
12. Baggiolini, M. (1998) Chemokines and leukocyte traffic. *Nature*, **392**, 565-568.
13. Gerard, C. and Rollins, B.J. (2001) Chemokines and disease. *Nat Immunol*, **2**, 108-115.
14. Raman, D., Sobolik-Delmaire, T. and Richmond, A. (2011) Chemokines in health and disease. *Exp Cell Res*, **317**, 575-589.
15. Bonecchi, R., Galliera, E., Borroni, E.M., Corsi, M.M., Locati, M. and Mantovani, A. (2009) Chemokines and chemokine receptors: an overview. *Front Biosci (Landmark Ed)*, **14**, 540-551.
16. Nguyen, L.T., Kwakman, P.H., Chan, D.I., Liu, Z., de Boer, L., Zaat, S.A. and Vogel, H.J. (2011) Exploring platelet chemokine antimicrobial activity: nuclear magnetic resonance backbone dynamics of NAP-2 and TC-1. *Antimicrob Agents Chemother*, **55**, 2074-2083.

17. Young, H., Roongta, V., Daly, T.J. and Mayo, K.H. (1999) NMR structure and dynamics of monomeric neutrophil-activating peptide 2. *Biochem J*, **338 (Pt 3)**, 591-598.
18. Qin, L., Kufareva, I., Holden, L.G., Wang, C., Zheng, Y., Zhao, C., Fenalti, G., Wu, H., Han, G.W., Cherezov, V. *et al.* (2015) Structural biology. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine. *Science*, **347**, 1117-1122.
19. Kufareva, I., Salanga, C.L. and Handel, T.M. (2015) Chemokine and chemokine receptor structure and interactions: implications for therapeutic strategies. *Immunol Cell Biol*, **93**, 372-383.
20. Rajagopalan, L. and Rajarathnam, K. (2006) Structural basis of chemokine receptor function--a model for binding affinity and ligand selectivity. *Biosci Rep*, **26**, 325-339.
21. Clark-Lewis, I., Kim, K.S., Rajarathnam, K., Gong, J.H., Dewald, B., Moser, B., Baggiolini, M. and Sykes, B.D. (1995) Structure-activity relationships of chemokines. *J Leukoc Biol*, **57**, 703-711.
22. Allen, S.J., Crown, S.E. and Handel, T.M. (2007) Chemokine: receptor structure, interactions, and antagonism. *Annu Rev Immunol*, **25**, 787-820.
23. Baryshnikova, O.K. and Sykes, B.D. (2006) Backbone dynamics of SDF-1alpha determined by NMR: interpretation in the presence of monomer-dimer equilibrium. *Protein Sci*, **15**, 2568-2578.
24. Grasberger, B.L., Gronenborn, A.M. and Clore, G.M. (1993) Analysis of the backbone dynamics of interleukin-8 by 15N relaxation measurements. *Journal of Molecular Biology*, **230**, 364-372.
25. Jansma, A.L., Kirkpatrick, J.P., Hsu, A.R., Handel, T.M. and Nietlispach, D. (2010) NMR analysis of the structure, dynamics, and unique oligomerization properties of the chemokine CCL27. *J Biol Chem*, **285**, 14424-14437.
26. Kim, K.S., Rajarathnam, K., Clark-Lewis, I. and Sykes, B.D. (1996) Structural characterization of a monomeric chemokine: monocyte chemoattractant protein-3. *FEBS Lett*, **395**, 277-282.
27. Liou, J.W., Chang, F.T., Chung, Y., Chen, W.Y., Fischer, W.B. and Hsu, H.J. (2014) In silico analysis reveals sequential interactions and protein conformational changes during the binding of chemokine CXCL-8 to its receptor CXCR1. *PLoS One*, **9**, e94178.
28. Mayer, K.L. and Stone, M.J. (2003) Backbone dynamics of the CC-chemokine eotaxin-2 and comparison among the eotaxin group chemokines. *Proteins*, **50**, 184-191.
29. Rajarathnam, K., Li, Y., Rohrer, T. and Gentz, R. (2001) Solution structure and dynamics of myeloid progenitor inhibitory factor-1 (MPIF-1), a novel monomeric CC chemokine. *J Biol Chem*, **276**, 4909-4916.
30. Ye, J., Mayer, K.L., Mayer, M.R. and Stone, M.J. (2001) NMR solution structure and backbone dynamics of the CC chemokine eotaxin-3. *Biochemistry*, **40**, 7820-7831.
31. Ye, J., Mayer, K.L. and Stone, M.J. (1999) Backbone dynamics of the human CC-chemokine eotaxin. *J Biomol NMR*, **15**, 115-124.

32. Joseph, P.R., Sarmiento, J.M., Mishra, A.K., Das, S.T., Garofalo, R.P., Navarro, J. and Rajarathnam, K. (2010) Probing the role of CXC motif in chemokine CXCL8 for high affinity binding and activation of CXCR1 and CXCR2 receptors. *J Biol Chem*, **285**, 29262-29269.
33. Joseph, P.R., Sawant, K.V., Isley, A., Pedroza, M., Garofalo, R.P., Richardson, R.M. and Rajarathnam, K. (2013) Dynamic conformational switching in the chemokine ligand is essential for G Protein coupled-receptor activation. *Biochem J.*
34. von Hundelshausen, P., Petersen, F. and Brandt, E. (2007) Platelet-derived chemokines in vascular biology. *Thromb Haemost*, **97**, 704-713.
35. Gear, A.R. and Camerini, D. (2003) Platelet chemokines and chemokine receptors: linking hemostasis, inflammation, and host defense. *Microcirculation*, **10**, 335-350.
36. Ludwig, A., Petersen, F., Zahn, S., Gotze, O., Schroder, J.M., Flad, H.D. and Brandt, E. (1997) The CXC-chemokine neutrophil-activating peptide-2 induces two distinct optima of neutrophil chemotaxis by differential interaction with interleukin-8 receptors CXCR-1 and CXCR-2. *Blood*, **90**, 4588-4597.
37. Moser, B., Schumacher, C., von Tscharner, V., Clark-Lewis, I. and Baggiolini, M. (1991) Neutrophil-activating peptide 2 and gro/melanoma growth-stimulatory activity interact with neutrophil-activating peptide 1/interleukin 8 receptors on human neutrophils. *J Biol Chem*, **266**, 10666-10671.
38. Veenstra, M. and Ransohoff, R.M. (2012) Chemokine receptor CXCR2: physiology regulator and neuroinflammation controller? *J Neuroimmunol*, **246**, 1-9.
39. Krijgsveld, J., Zaat, S.A., Meeldijk, J., van Veelen, P.A., Fang, G., Poolman, B., Brandt, E., Ehlert, J.E., Kuijpers, A.J., Engbers, G.H. *et al.* (2000) Thrombocidins, microbicidal proteins from human blood platelets, are C-terminal deletion products of CXC chemokines. *J Biol Chem*, **275**, 20374-20381.
40. Kwakman, P.H., Krijgsveld, J., de Boer, L., Nguyen, L.T., Boszhard, L., Vreede, J., Dekker, H.L., Speijer, D., Drijfhout, J.W., te Velde, A.A. *et al.* (2011) Native thrombocidin-1 and unfolded thrombocidin-1 exert antimicrobial activity via distinct structural elements. *J Biol Chem*, **286**, 43506-43514.
41. Nguyen, L.T., Chan, D.I., Boszhard, L., Zaat, S.A. and Vogel, H.J. (2010) Structure-function studies of chemokine-derived carboxy-terminal antimicrobial peptides. *Biochim Biophys Acta*, **1798**, 1062-1072.
42. Li, T., Tracka, M.B., Uddin, S., Casas-Finet, J., Jacobs, D.J. and Livesay, D.R. (2014) Redistribution of flexibility in stabilizing antibody fragment mutants follows Le Chatelier's principle. *PLoS One*, **9**, e92870.
43. Hess, B., Kutzner, C., van der Spoel, D. and Lindahl, E. (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput*, **4**, 435-447.
44. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J. (2005) GROMACS: fast, flexible, and free. *J Comput Chem*, **26**, 1701-1718.

45. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O. and Shaw, D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**, 1950-1958.
46. Martonak, R., Laio, A. and Parrinello, M. (2003) Predicting crystal structures: the Parrinello-Rahman method revisited. *Phys Rev Lett*, **90**, 075503.
47. Hoover, W.G. (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A Gen Phys*, **31**, 1695-1697.
48. Norberto de Souza, O. and Ornstein, R.L. (1999) Molecular dynamics simulations of a protein-protein dimer: particle-mesh Ewald electrostatic model yields far superior results to standard cutoff model. *J Biomol Struct Dyn*, **16**, 1205-1218.
49. Hess, B. (2008) P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput*, **4**, 116-122.
50. Karpen, M.E., Tobias, D.J. and Brooks, C.L., 3rd. (1993) Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry*, **32**, 412-420.
51. Feig, M., Karanicolas, J. and Brooks, C.L., 3rd. (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model*, **22**, 377-395.
52. Jacobs, D.J. and Dallakyan, S. (2005) Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J*, **88**, 903-915.
53. Mottonen, J.M., Jacobs, D.J. and Livesay, D.R. (2010) Allosteric response is both conserved and variable across three CheY orthologs. *Biophys J*, **99**, 2245-2254.
54. Li, T., Verma, D., Tracka, M.B., Casas-Finet, J., Livesay, D.R. and Jacobs, D.J. (2014) Thermodynamic stability and flexibility characteristics of antibody fragment complexes. *Protein Pept Lett*, **21**, 752-765.
55. Carlson, J., Baxter, S.A., Dreau, D. and Nesmelova, I.V. (2013) The heterodimerization of platelet-derived chemokines. *Biochim Biophys Acta*, **1834**, 158-168.
56. Mayo, K.H., Barker, S., Kuranda, M.J., Hunt, A.J., Myers, J.A. and Maione, T.E. (1992) Molten globule monomer to condensed dimer: role of disulfide bonds in platelet factor-4 folding and subunit association. *Biochemistry*, **31**, 12255-12265.
57. Fernandez, E.J. and Lolis, E. (2002) Structure, function, and inhibition of chemokines. *Annu Rev Pharmacol Toxicol*, **42**, 469-499.
58. Yan, Z., Zhang, J., Holt, J.C., Stewart, G.J., Niewiarowski, S. and Poncz, M. (1994) Structural requirements of platelet chemokines for neutrophil activation. *Blood*, **84**, 2329-2339.
59. Crump, M.P., Spyracopoulos, L., Lavigne, P., Kim, K.S., Clark-lewis, I. and Sykes, B.D. (1999) Backbone dynamics of the human CC chemokine eotaxin: fast motions, slow motions, and implications for receptor binding. *Protein Sci*, **8**, 2041-2054.
60. Eigenbrot, C., Lowman, H.B., Chee, L. and Artis, D.R. (1997) Structural change and receptor binding in a chemokine mutant with a rearranged disulfide: X-ray structure of E38C/C50AIL-8 at 2 A resolution. *Proteins*, **27**, 556-566.
61. Rajarathnam, K., Sykes, B.D., Dewald, B., Baggiolini, M. and Clark-Lewis, I. (1999) Disulfide bridges in interleukin-8 probed using non-natural disulfide

analogues: dissociation of roles in structure from function. *Biochemistry*, **38**, 7653-7658.
62. Prado, G.N., Suetomi, K., Shumate, D., Maxwell, C., Ravindran, A., Rajarathnam, K. and Navarro, J. (2007) Chemokine signaling specificity: essential role for the N-terminal domain of chemokine receptors. *Biochemistry*, **46**, 8961-8968.
63. Rajagopalan, L. and Rajarathnam, K. (2004) Ligand selectivity and affinity of chemokine receptor CXCR1. Role of N-terminal domain. *J Biol Chem*, **279**, 30000-30008.
64. Rajagopalan, L., Chin, C.C. and Rajarathnam, K. (2007) Role of intramolecular disulfides in stability and structure of a noncovalent homodimer. *Biophys J*, **93**, 2129-2134.
65. Alving, C.R., Peachman, K.K., Rao, M. and Reed, S.G. (2012) Adjuvants for human vaccines. *Curr Opin Immunol*, **24**, 310-315.
66. Dormitzer, P.R., Grandi, G. and Rappuoli, R. (2012) Structural vaccinology starts to deliver. *Nat Rev Microbiol*, **10**, 807-813.
67. Wolfe, D.N., Florence, W. and Bryant, P. (2013) Current biodefense vaccine programs and challenges. *Hum Vaccin Immunother*, **9**, 1591-1597.
68. Stirpe, F. and Battelli, M.G. (2006) Ribosome-inactivating proteins: progress and problems. *Cell Mol Life Sci*, **63**, 1850-1866.
69. Sandvig, K., Torgersen, M.L., Engedal, N., Skotland, T. and Iversen, T.G. (2010) Protein toxins from plants and bacteria: probes for intracellular transport and tools in medicine. *FEBS Lett*, **584**, 2626-2634.
70. O'Hara, J.M., Yermakova, A. and Mantis, N.J. (2012) Immunity to ricin: fundamental insights into toxin-antibody interactions. *Curr Top Microbiol Immunol*, **357**, 209-241.
71. Sandvig, K., Olsnes, S. and Pihl, A. (1976) Kinetics of binding of the toxic lectins abrin and ricin to surface receptors of human cells. *J Biol Chem*, **251**, 3977-3984.
72. Rutenber, E., Ready, M. and Robertus, J.D. (1987) Structure and evolution of ricin B chain. *Nature*, **326**, 624-626.
73. Spooner, R.A. and Lord, J.M. (2012) How ricin and Shiga toxin reach the cytosol of target cells: retrotranslocation from the endoplasmic reticulum. *Curr Top Microbiol Immunol*, **357**, 19-40.
74. Endo, Y. and Tsurugi, K. (1986) Mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. *Nucleic Acids Symp Ser*, 187-190.
75. Jandhyala, D.M., Thorpe, C.M. and Magun, B. (2012) Ricin and Shiga toxins: effects on host cell signal transduction. *Curr Top Microbiol Immunol*, **357**, 41-65.
76. Radosavljevic, V. and Belojevic, G. (2009) A new model of bioterrorism risk assessment. *Biosecur Bioterror*, **7**, 443-451.
77. Reisler, R.B. and Smith, L.A. (2012) The need for continued development of ricin countermeasures. *Adv Prev Med*, **2012**, 149737.
78. Griffiths, G.D. (2011) Understanding ricin from a defensive viewpoint. *Toxins (Basel)*, **3**, 1373-1392.
79. Smallshaw, J.E. and Vitetta, E.S. (2012) Ricin vaccine development. *Curr Top Microbiol Immunol*, **357**, 259-272.

80. O'Hara, J.M., Brey, R.N., 3rd and Mantis, N.J. (2013) Comparative efficacy of two leading candidate ricin toxin a subunit vaccines in mice. *Clin Vaccine Immunol*, **20**, 789-794.
81. Smallshaw, J.E., Ghetie, V., Rizo, J., Fulmer, J.R., Trahan, L.L., Ghetie, M.A. and Vitetta, E.S. (2003) Genetic engineering of an immunotoxin to eliminate pulmonary vascular leak in mice. *Nat Biotechnol*, **21**, 387-391.
82. Smallshaw, J.E., Firan, A., Fulmer, J.R., Ruback, S.L., Ghetie, V. and Vitetta, E.S. (2002) A novel recombinant vaccine which protects mice against ricin intoxication. *Vaccine*, **20**, 3422-3427.
83. Legler, P.M., Brey, R.N., Smallshaw, J.E., Vitetta, E.S. and Millard, C.B. (2011) Structure of RiVax: a recombinant ricin vaccine. *Acta Crystallogr D Biol Crystallogr*, **67**, 826-830.
84. Smallshaw, J.E., Richardson, J.A. and Vitetta, E.S. (2007) RiVax, a recombinant ricin subunit vaccine, protects mice against ricin delivered by gavage or aerosol. *Vaccine*, **25**, 7459-7469.
85. Neal, L.M., McCarthy, E.A., Morris, C.R. and Mantis, N.J. (2011) Vaccine-induced intestinal immunity to ricin toxin in the absence of secretory IgA. *Vaccine*, **29**, 681-689.
86. Marconescu, P.S., Smallshaw, J.E., Pop, L.M., Ruback, S.L. and Vitetta, E.S. (2010) Intradermal administration of RiVax protects mice from mucosal and systemic ricin intoxication. *Vaccine*, **28**, 5315-5322.
87. Greene, C.J., Chadwick, C.M., Mandell, L.M., Hu, J.C., O'Hara, J.M., Brey, R.N., 3rd, Mantis, N.J. and Connell, T.D. (2013) LT-IIb(T13I), a non-toxic type II heat-labile enterotoxin, augments the capacity of a ricin toxin subunit vaccine to evoke neutralizing antibodies and protective immunity. *PLoS One*, **8**, e69678.
88. Vitetta, E.S., Smallshaw, J.E. and Schindler, J. (2012) Pilot phase IB clinical trial of an alhydrogel-adsorbed recombinant ricin vaccine. *Clin Vaccine Immunol*, **19**, 1697-1699.
89. Vitetta, E.S., Smallshaw, J.E., Coleman, E., Jafri, H., Foster, C., Munford, R. and Schindler, J. (2006) A pilot clinical trial of a recombinant ricin vaccine in normal humans. *Proc Natl Acad Sci U S A*, **103**, 2268-2273.
90. Vance, D.J., Rong, Y., Brey, R.N., 3rd and Mantis, N.J. (2015) Combination of two candidate subunit vaccine antigens elicits protective immunity to ricin and anthrax toxin in mice. *Vaccine*, **33**, 417-421.
91. Argent, R.H., Parrott, A.M., Day, P.J., Roberts, L.M., Stockley, P.G., Lord, J.M. and Radford, S.E. (2000) Ribosome-mediated folding of partially unfolded ricin A-chain. *J Biol Chem*, **275**, 9263-9269.
92. Olson, M.A., Carra, J.H., Roxas-Duncan, V., Wannemacher, R.W., Smith, L.A. and Millard, C.B. (2004) Finding a new vaccine in the ricin protein fold. *Protein Eng Des Sel*, **17**, 391-397.
93. McHugh, C.A., Tammariello, R.F., Millard, C.B. and Carra, J.H. (2004) Improved stability of a protein vaccine through elimination of a partially unfolded state. *Protein Sci*, **13**, 2736-2743.
94. Compton, J.R., Legler, P.M., Clingan, B.V., Olson, M.A. and Millard, C.B. (2011) Introduction of a disulfide bond leads to stabilization and crystallization of a ricin immunogen. *Proteins*, **79**, 1048-1060.

95. Thomas, J.C., O'Hara, J.M., Hu, L., Gao, F.P., Joshi, S.B., Volkin, D.B., Brey, R.N., Fang, J., Karanicolas, J., Mantis, N.J. *et al.* (2013) Effect of single-point mutations on the stability and immunogenicity of a recombinant ricin A chain subunit vaccine antigen. *Hum Vaccin Immunother*, **9**, 744-752.
96. Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, **487**, 545-574.
97. Das, R. and Baker, D. (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem*, **77**, 363-382.
98. Smith, C.A. and Kortemme, T. (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol*, **380**, 742-756.
99. Friedland, G.D., Linares, A.J., Smith, C.A. and Kortemme, T. (2008) A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J Mol Biol*, **380**, 757-774.
100. Sheffler, W. and Baker, D. (2009) RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci*, **18**, 229-239.
101. Li, Y., Zhang, J., Tai, D., Middaugh, C.R., Zhang, Y. and Fang, J. (2012) PROTS: a fragment based protein thermo-stability potential. *Proteins*, **80**, 81-92.
102. Li, Y., Middaugh, C.R. and Fang, J. (2010) A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. *BMC Bioinformatics*, **11**, 62.
103. O'Hara, J.M., Neal, L.M., McCarthy, E.A., Kasten-Jolly, J.A., Brey, R.N., 3rd and Mantis, N.J. (2010) Folding domains within the ricin toxin A subunit as targets of protective antibodies. *Vaccine*, **28**, 7035-7046.
104. O'Hara, J.M., Kasten-Jolly, J.C., Reynolds, C.E. and Mantis, N.J. (2014) Localization of non-linear neutralizing B cell epitopes on ricin toxin's enzymatic subunit (RTA). *Immunol Lett*, **158**, 7-13.
105. Hem, S.L. and Hogenesch, H. (2007) Relationship between physical and chemical properties of aluminum-containing adjuvants and immunopotentiation. *Expert Rev Vaccines*, **6**, 685-698.
106. Toth, R.T.t., Angalakurthi, S.K., Van Slyke, G., Vance, D.J., Hickey, J.M., Joshi, S.B., Middaugh, C.R., Volkin, D.B., Weis, D.D. and Mantis, N.J. (2017) High-Definition Mapping of Four Spatially Distinct Neutralizing Epitope Clusters on RiVax, a Candidate Ricin Toxin Subunit Vaccine. *Clin Vaccine Immunol*, **24**.
107. Rudolph, M.J., Vance, D.J., Cassidy, M.S., Rong, Y., Shoemaker, C.B. and Mantis, N.J. (2016) Structural analysis of nested neutralizing and non-neutralizing B cell epitopes on ricin toxin's enzymatic subunit. *Proteins*, **84**, 1162-1172.
108. Ivics, Z., Hackett, P.B., Plasterk, R.H. and Izsvak, Z. (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**, 501-510.

109. Izsvak, Z., Ivics, Z. and Plasterk, R.H. (2000) Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol*, **302**, 93-102.
110. Boehme, P., Doerner, J., Solanki, M., Jing, L., Zhang, W. and Ehrhardt, A. (2015) The sleeping beauty transposon vector system for treatment of rare genetic diseases: an unrealized hope? *Curr Gene Ther*, **15**, 255-265.
111. Hackett, P.B., Largaespada, D.A., Switzer, K.C. and Cooper, L.J. (2013) Evaluating risks of insertional mutagenesis by DNA transposons in gene therapy. *Transl Res*, **161**, 265-283.
112. Swierczek, M., Izsvak, Z. and Ivics, Z. (2012) The Sleeping Beauty transposon system for clinical applications. *Expert Opin Biol Ther*, **12**, 139-153.
113. Izsvak, Z., Hackett, P.B., Cooper, L.J. and Ivics, Z. (2010) Translating Sleeping Beauty transposition into cellular therapies: victories and challenges. *Bioessays*, **32**, 756-767.
114. Ivics, Z. and Izsvak, Z. (2010) The expanding universe of transposon technologies for gene and cell engineering. *Mob DNA*, **1**, 25.
115. Kaji, K., Norrby, K., Paca, A., Mileikovsky, M., Mohseni, P. and Woltjen, K. (2009) Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*, **458**, 771-775.
116. Davis, R.P., Nemes, C., Varga, E., Freund, C., Kosmidis, G., Gkatzis, K., de Jong, D., Szuhai, K., Dinnyes, A. and Mummery, C.L. (2013) Generation of induced pluripotent stem cells from human foetal fibroblasts using the Sleeping Beauty transposon gene delivery system. *Differentiation*, **86**, 30-37.
117. Grabundzija, I., Wang, J., Sebe, A., Erdei, Z., Kajdi, R., Devaraj, A., Steinemann, D., Szuhai, K., Stein, U., Cantz, T. *et al.* (2013) Sleeping Beauty transposon-based system for cellular reprogramming and targeted gene insertion in induced pluripotent stem cells. *Nucleic Acids Res*, **41**, 1829-1847.
118. Fatima, A., Ivanyuk, D., Herms, S., Heilmann-Heimbach, S., O'Shea, O., Chapman, C., Izsvak, Z., Farr, M., Hescheler, J. and Saric, T. (2016) Generation of human induced pluripotent stem cell line from a patient with a long QT syndrome type 2. *Stem Cell Res*, **16**, 304-307.
119. Jaskolski, M., Alexandratos, J.N., Bujacz, G. and Wlodawer, A. (2009) Piecing together the structure of retroviral integrase, an important target in AIDS therapy. *FEBS J*, **276**, 2926-2946.
120. Nowotny, M. (2009) Retroviral integrase superfamily: the structural perspective. *EMBO Rep*, **10**, 144-151.
121. Montano, S.P. and Rice, P.A. (2011) Moving DNA around: DNA transposition and retroviral integration. *Curr Opin Struct Biol*, **21**, 370-378.
122. Dyda, F., Chandler, M. and Hickman, A.B. (2012) The emerging diversity of transpososome architectures. *Q Rev Biophys*, **45**, 493-521.
123. Nesmelova, I.V. and Hackett, P.B. (2010) DDE transposases: Structural similarity and diversity. *Adv Drug Deliv Rev*, **62**, 1187-1195.
124. Majorek, K.A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K. and Bujnicki, J.M. (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res*, **42**, 4160-4179.

125. Moelling, K., Broecker, F., Russo, G. and Sunagawa, S. (2017) RNase H As Gene Modifier, Driver of Evolution and Antiviral Defense. *Frontiers in Microbiology*, **8**, 1745.
126. Mizuuchi, K. (1992) Transpositional recombination: mechanistic insights from studies of mu and other elements. *Annu Rev Biochem*, **61**, 1011-1051.
127. Craig, N.L. (1995) Unity in transposition reactions. *Science*, **270**, 253-254.
128. Richardson, J.M., Colloms, S.D., Finnegan, D.J. and Walkinshaw, M.D. (2009) Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell*, **138**, 1096-1108.
129. Fitzkee, N.C., Masse, J.E., Shen, Y., Davies, D.R. and Bax, A. (2010) Solution conformation and dynamics of the HIV-1 integrase core domain. *J Biol Chem*, **285**, 18072-18084.
130. Goldgur, Y., Dyda, F., Hickman, A.B., Jenkins, T.M., Craigie, R. and Davies, D.R. (1998) Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc Natl Acad Sci U S A*, **95**, 9150-9154.
131. Wolkowicz, U.M., Morris, E.R., Robson, M., Trubitsyna, M. and Richardson, J.M. (2014) Structural basis of Mos1 transposase inhibition by the anti-retroviral drug Raltegravir. *ACS Chem Biol*, **9**, 743-751.
132. Nguyen, B.Y., Isaacs, R.D., Teppler, H., Leavitt, R.Y., Sklar, P., Iwamoto, M., Wenning, L.A., Miller, M.D., Chen, J., Kemp, R. *et al.* (2011) Raltegravir: the first HIV-1 integrase strand transfer inhibitor in the HIV armamentarium. *Ann N Y Acad Sci*, **1222**, 83-89.
133. Wills, T. and Vega, V. (2012) Elvitegravir: a once-daily inhibitor of HIV-1 integrase. *Expert Opin Investig Drugs*, **21**, 395-401.
134. Klumpp, K., Hang, J.Q., Rajendran, S., Yang, Y., Derosier, A., Wong Kai In, P., Overton, H., Parkes, K.E., Cammack, N. and Martin, J.A. (2003) Two-metal ion mechanism of RNA cleavage by HIV RNase H and mechanism-based design of selective HIV RNase H inhibitors. *Nucleic Acids Res*, **31**, 6852-6859.
135. Cowan, J.A., Ohyama, T., Howard, K., Rausch, J.W., Cowan, S.M. and Le Grice, S.F. (2000) Metal-ion stoichiometry of the HIV-1 RT ribonuclease H domain: evidence for two mutually exclusive sites leads to new mechanistic insights on metal-mediated hydrolysis in nucleic acid biochemistry. *J Biol Inorg Chem*, **5**, 67-74.
136. Stafford, K.A. and Palmer Iii, A.G. (2014) Evidence from molecular dynamics simulations of conformational preorganization in the ribonuclease H active site. *F1000Res*, **3**, 67.
137. Bahar, I., Erman, B., Jernigan, R.L., Atilgan, A.R. and Covell, D.G. (1999) Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J Mol Biol*, **285**, 1023-1037.
138. Das, K., Martinez, S.E., Bandwar, R.P. and Arnold, E. (2014) Structures of HIV-1 RT-RNA/DNA ternary complexes with dATP and nevirapine reveal conformational flexibility of RNA/DNA: insights into requirements for RNase H cleavage. *Nucleic Acids Res*, **42**, 8125-8137.
139. Stafford, K.A., Trbovic, N., Butterwick, J.A., Abel, R., Friesner, R.A. and Palmer, A.G., 3rd. (2015) Conformational preferences underlying reduced activity of a thermophilic ribonuclease H. *J Mol Biol*, **427**, 853-866.

140. Billamboz, M., Bailly, F., Lion, C., Calmels, C., Andreola, M.L., Witvrouw, M., Christ, F., Debyser, Z., De Luca, L., Chimirri, A. *et al.* (2011) 2-hydroxyisoquinoline-1,3(2H,4H)-diones as inhibitors of HIV-1 integrase and reverse transcriptase RNase H domain: influence of the alkylation of position 4. *Eur J Med Chem*, **46**, 535-546.
141. Livesay, D.R. and Jacobs, D.J. (2006) Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins*, **62**, 130-143.
142. Carpentier, C.E., Schreifels, J.M., Aronovich, E.L., Carlson, D.F., Hackett, P.B. and Nesmelova, I.V. (2014) NMR structural analysis of Sleeping Beauty transposase binding to DNA. *Protein Sci*, **23**, 23-33.
143. Leighton, G.O., Konnova, T.A., Idiyatullin, B., Hurr, S.H., Zuev, Y.F. and Nesmelova, I.V. (2014) The folding of the specific DNA recognition subdomain of the sleeping beauty transposase is temperature-dependent and is required for its binding to the transposon DNA. *PLoS One*, **9**, e112114.
144. Konnova, T.A., Singer, C.M. and Nesmelova, I.V. (2017) NMR solution structure of the RED subdomain of the Sleeping Beauty transposase. *Protein Sci*, **26**, 1171-1181.
145. Tokuriki, N., Kinjo, M., Negi, S., Hoshino, M., Goto, Y., Urabe, I. and Yomo, T. (2004) Protein folding by the effects of macromolecular crowding. *Protein Sci*, **13**, 125-133.
146. Minton, A.P. (2005) Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. *Biophys J*, **88**, 971-985.
147. Zhang, Y. and Cremer, P.S. (2006) Interactions between macromolecules and ions: The Hofmeister series. *Curr Opin Chem Biol*, **10**, 658-663.
148. Zhang, Y. and Cremer, P.S. (2010) Chemistry of Hofmeister anions and osmolytes. *Annu Rev Phys Chem*, **61**, 63-83.
149. Voigt, F., Wiedemann, L., Zuliani, C., Querques, I., Sebe, A., Mates, L., Izsvak, Z., Ivics, Z. and Barabas, O. (2016) Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. *Nat Commun*, **7**, 11126.
150. Mates, L., Chuah, M.K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D.P., Schmitt, A., Becker, K., Matrai, J. *et al.* (2009) Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet*, **41**, 753-761.
151. Thomas, C.E., Ehrhardt, A. and Kay, M.A. (2003) Progress and problems with the use of viral vectors for gene therapy. *Nat Rev Genet*, **4**, 346-358.
152. Richardson, J.M., Dawson, A., O'Hagan, N., Taylor, P., Finnegan, D.J. and Walkinshaw, M.D. (2006) Mechanism of Mos1 transposition: insights from structural analysis. *EMBO J*, **25**, 1324-1334.
153. Schrodinger, LLC. (2015).
154. Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D. *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845-854.

155. Abrusan, G., Yant, S.R., Szilagyi, A., Marsh, J.A., Mates, L., Izsvak, Z., Barabas, O. and Ivics, Z. (2016) Structural Determinants of Sleeping Beauty Transposase Activity. *Mol Ther*, **24**, 1369-1377.
156. Bujacz, G., Alexandratos, J., Qing, Z.L., Clement-Mella, C. and Wlodawer, A. (1996) The catalytic domain of human immunodeficiency virus integrase: ordered active site in the F185H mutant. *FEBS Lett*, **398**, 175-178.
157. Dyda, F., Hickman, A.B., Jenkins, T.M., Engelman, A., Craigie, R. and Davies, D.R. (1994) Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science*, **266**, 1981-1986.
158. Lee, M.C., Deng, J., Briggs, J.M. and Duan, Y. (2005) Large-scale conformational dynamics of the HIV-1 integrase core domain and its catalytic loop mutants. *Biophys J*, **88**, 3133-3146.
159. Greenwald, J., Le, V., Butler, S.L., Bushman, F.D. and Choe, S. (1999) The mobility of an HIV-1 integrase active site loop is correlated with catalytic activity. *Biochemistry-Us*, **38**, 8892-8898.
160. Bar-Joseph, Z., Gifford, D.K. and Jaakkola, T.S. (2001) Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17 Suppl 1**, S22-29.
161. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**, 14863-14868.
162. DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.
163. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.
164. Davies, D.R., Mahnke Braam, L., Reznikoff, W.S. and Rayment, I. (1999) The three-dimensional structure of a Tn5 transposase-related protein determined to 2.9-A resolution. *J Biol Chem*, **274**, 11904-11913.
165. Rice, P. and Mizuuchi, K. (1995) Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration. *Cell*, **82**, 209-220.