THERMODYNAMIC AND MECHANICAL EFFECTS OF DISULFIDE BONDS IN CXCL7 CHEMOKINE

by

Christopher Singer

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Applied Physics

Charlotte

2015

Approved by:

_____
Dr. Donald Jacobs

_____
Dr. Irina Nesmelova

_____
Dr. Tsing-Hua Her

ABSTRACT

CHRISTOPHER SINGER. Thermodynamic and mechanical effects of disulfide bonds in CXCL7 chemokine. (Under the direction of DR. DONALD JACOBS)

Chemokines are a family of signaling proteins mainly responsible for the chemotaxis of leukocytes, where their biological activity is modulated by their oligomerization state. Here, the dynamics and thermodynamic stability are characterized in monomer and homodimer structures of CXCL7, one of the most abundant platelet chemokines. The effects of dimerization and disulfide bond formation are investigated using computational methods that include molecular dynamics (MD) simulations and the Distance Constraint Model (DCM). A consistent picture emerges for the effect of dimerization and role of the Cys5-Cys31 and Cys7-Cys47 disulfide bonds. Surprisingly, neither disulfide bond is critical for maintaining structural stability in the monomer or dimer, although the monomer is destabilized more than the dimer upon removal of disulfide bonds. Instead, it is found that disulfide bonds influence the native state dynamics as well as modulates the relative stability between monomer and dimer. The combined analysis elucidates how CXCL7 is mechanically stable as a monomer, and how upon dimerization flexibly correlated motions are induced between the 30s and 50s loop within each monomer and across the dimer interface. Interestingly, the greatest gain in flexibility upon dimerization occurs when both disulfide bonds are present in each domain, and the homodimer is least stable relative to its two monomers. These results suggest the highly conserved disulfide bonds in chemokines facilitate a structural mechanism for distinguishing functional characteristics between monomer and dimer.

ACKNOWLEDGMENTS

I would like to take a moment to thank everyone who has been a key player in making this thesis possible. First and foremost, thank you to my advisors Dr. Donald Jacobs, and Dr. Irina Nesmelova for the education, guidance, and support over the past few years. Also thank you to Dr. Tsing-Hua Her for sitting on the masters committee for the defense of this thesis. I would also like to thank Dr. Dennis Livesay as well as the entire BMPG research group for the continual support and direction, especially Mathew Brown and Charles Herring. Without all of your help this would not have been possible. Thank You.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANM	Anisotropic Network Model
CC	correlated cooperativity
FI	flexibility index
CD	circular dichroism
DOF	degree of freedom
DSC	differential ccanning calorimetry
NMR	nuclear magnetic resonance
GPCR	G-Protein-Coupled Receptor
MD	molecular dynamics
mDCM	minimum Distance Constraint Model
MOE	Molecular Operating Environment
QSFR	Quantitative Stability/Flexibility Relationships
SBNR	signal beyond noise ratio

CHAPTER 1: INTRODUCTION

1.1 Biological Basis

The ability to predict protein function and the mechanisms in which proteins carry out this function within living organisms is the Holy Grail of structural biology. Unfortunately this level of predictive power lies in the distant future and we are forced to consider less ambitious questions and then piece their answers together to understand the complex situations biology bestows upon us. It is the responsibility of scientists to create cleverly designed experiments and to build models to chip away at the big picture, while drilling down into detailed explanations of how processes work at the molecular level. This is evident, as numerous experimental methods have emerged to probe different physical characteristics of proteins including: x-ray crystallography, heat capacity, dynamics through Nuclear Magnetic Resonance (NMR), stability through Circular Dichroism (CD) and melting temperatures, binding affinities, and so forth.

The central dogma of molecular biology is the cornerstone of biological research, and provides the basis for understanding of how life works at the molecular level, as we know it on planet Earth. DNA is coded through the translation of DNA into RNA through transcriptase proteins. Then the RNA is transcribed into amino acid polypeptides called proteins, which are the focus of this thesis. Proteins are the workhorses of the seemingly endless number of molecular processes taking place in cellular environments, where they help govern everything from DNA replication to oxygen transport. In the study of proteins, biophysicists tease out mechanistic pathways that carry out metabolic function. This scientific inquiry is a fundamental

component of current pharmaceutical research that is concerned with how to up regulate and down regulate specific metabolic pathways by enhancing or inhibiting protein function. That being said, understanding such complex functions require significant effort using quantified modeling approaches that are germane to physics. Indeed, the molecular world of a living system or otherwise, is governed by the laws of physics, and therefore we must discover the physical properties of these proteins to understand what makes them "tick". In this thesis, certain models will be employed to help elucidate mechanisms as physical and biological principles are combined to arrive at molecular understanding of life.

1.2 Chemokine Family and CXCL7

We will be focusing on a specific signaling protein within the chemokine family known as CXCL7. The chemokine family is important to nearly all homeostatic and regulatory processes[1] within the human system. The primary interest in the field currently is to understand how chemokines play such key roles in signaling inflammation response in infectious diseases, HIV/AID, and caner.[2,3] Despite increases in experimental research on chemokines over the past several years there has been a lack of computational analysis of any of the family members.

There are 47 known members of the human Chemokine family and each of these members have the ability to interact with multiple receptors that belong to the G-protein-coupled receptor (GPCR) superfamily.[4] This diversity in interactions opens the door for an enormous amount of functional variation with relatively few protein structures in comparison to their span of functional space. A reason for the ambiguity of binding events is partially due to a highly conserved fold among large portions of

the family.[5] This high level of conservation is especially true within the binding sites located near the N-terminus and the N-loop. This conservation allows binding to multiple receptors, which then directly affects the function of the protein-ligand complex[6-16]. In addition to the ability to bind to multiple different receptors, monomer chains of chemokines have been seen to dimerize into both homodimer and heterodimers further expanding the functional space of the chemokine family.

Additional characteristics conserved across the chemokine family are a highly mobile N-terminus, followed by a long N-loop, a 3_{10} helical turn, a three-stranded beta-sheet, and finally a C-terminal alpha helix. The N-loop is a critical part of the receptor-binding site and due to it being highly flexible, it causes substantial diversity in its binding events as previously mentioned.[17,18] Another defining characteristic, which will become a point of interest in this study, is the existence of two highly conserved disulfide bonds[19] connecting four cysteine residues in each monomer chain. Indeed, there are only a few exceptions to this rule across the family, and this begs the question what is their importance to the function of chemokines? Or at least, what is their role in altering the stability and dynamics of chemokines?

Here we intend to investigate the dynamics of inter-residue couplings as well as the role of these disulfide bonds on the stability of CXCL7 structure. In addition the biologically relevant form of CXCL7 is in the monomer and homodimer[20] subunits within a tetramer superstructure obtained from x-ray crystallography. Therefore we are also interested in probing the stability differences between the monomer and dimer, which is important in the regulation of function efficiency.

Furthermore, our interest is to determine mechanical couplings that may imply some form of mechanism that underlies biological function.

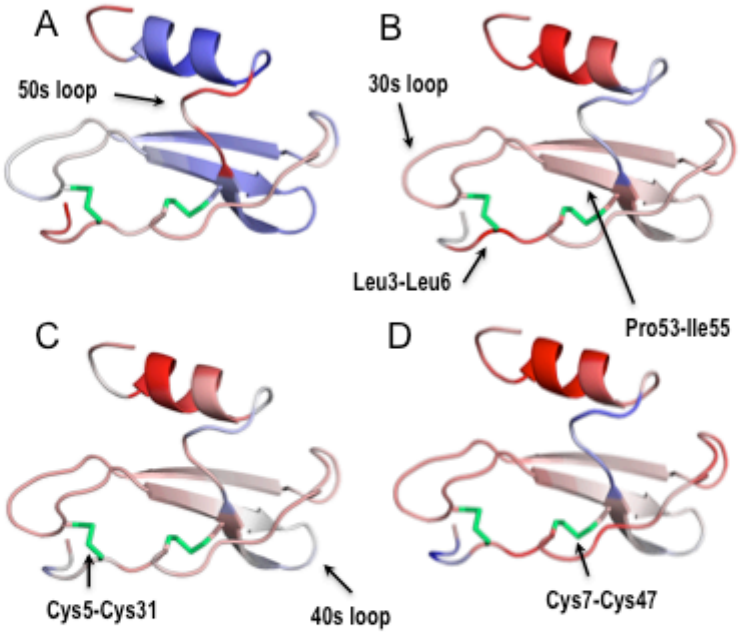

Figure 1: Ribbon diagram representation of CXCL7 monomer. Disulfide bonds are highlighted in green, as well as indicators for other important regions of CXCL7. A) Flexibility Index mapped onto monomer structure with blue meaning rigid and red meaning flexible. B) Difference in FI between A and structure with first disulfide removed. Red indicates increase in flexibility. C) Difference in FI for second DS bonds removed. D) Difference in FI for both bonds removed.

1.3 Molecular Simulations

A common tool in the structural biology field is molecular dynamics simulation. For example, Newton's coupled differential equations of motion is solved numerically as it calculates all intermolecular forces and propagates the trajectory of motion of all atoms within the structure over a femtosecond time scale. Moreover, it is possible to include the details of atoms within the molecular structure of the solvent in the calculation. However, these all-atom molecular dynamics (MD) calculations can be costly, requiring specialized super computers. As such, many different

models[21-23] and a variety of methods at different levels of detail exist for MD; here we will briefly survey a few of the most common.

First a united atom model[21] is primarily used to probe the dynamics of large-scale events. Grouping clusters of atoms together into single objects that roughly reflect the chemical and physical parameters of the cluster of atoms reduces the complexity of the problem. This makes the model suitable for studying intermolecular motions, but is limited in uncovering detailed mechanisms. Often, solvent effects are not carefully accounted for, although implicit solvent models are frequently used in united models. Explicit solvent models can also be used with united atom modes, where a water model is usually treated as a single rigid object with a fixed charged distribution on it.

By accounting for molecular complexity at another level, and therefore driving up computational time, greater accuracy can be generated with all atoms modeled explicitly within the protein. While the resolution of the protein is increased to all atoms, it is also possible to treat the solvent implicitly. In this case, interactions between protein atoms and solvent are handled through surface terms in an energy function plus the electrostatic interactions are handled by solving for electric fields in a continuum dielectric medium[22,23]. The entire solvent is treated as a continuum as opposed to individual molecules of water or other co-solvents. The next level of detail is to replace the implicit solvent with an all atom explicit model, which is another step further in realistic modeling. At this all atom level of detail, accuracy of the chemistry near the surface of proteins is greatly increased, but it also significantly increases the number of particles in which to calculate energies to simulate. Usually the solvent

consumes more than 95% of the computational cost in these all atom explicit models[22].

In addition to these basic models other methods have been introduced in effort to increase both speed and efficiency of MD simulations. For example meta-dynamics is an approach commonly used to steer the simulation away from its starting point in effort to explore greater conformational space. It does this by adding additional potential to conformations in which have already been explored causing a bias to move away from that region of conformational space. Monte Carlo[21,24] is another method that has been employed in many computational fields. Monte Carlo uses statistical mechanics and a Markov Chain procedure to determine the next step in the simulation.

These listed methods are merely a drop in the bucket as far as the number of computational models that have been developed, each with there own unique benefits and cost. This speaks to our ambition to approach the same problem from as many ways as conceivable in effort to paint a coherent story about what nature is doing. In many ways this has become a common practice for computational scientists with the explosion of computing capacity over the past few decades. These models are a brute force approach to generating protein dynamics by simultaneously calculating driving forces for every particle within the simulation space. These methods are computationally expensive and are only as accurate as the force field driving the calculations, however they are typically regarded as the most accurate within the field.

1.4 Protein Thermodynamics

In contrast to MD, which describes dynamics based off of a force field, thermodynamic models aim to describe a protein's free energy based on enthalpy and entropy contributions. In general most proteins can be approximated to exist in one of two thermodynamic states. Typically the biologically functional state is the folded state, which we often refer to as the native state. The native fold's free energy is dominated by favorable non-covalent interactions between residues, which are expressed within the enthalpy term of Eq. (1). These interactions can be hydrogen bonds, ionic interactions, van der walls, London dispersion, or any other weak interaction. Second is the highly disordered or what is usually referred to as the unfolded state. Here there is a dramatic loss of intramolecular non-covalent interactions that increases the enthalpy (non favorable), and associated with a dramatic increase in conformational entropy (favorable). As such, at high temperature the protein will thermodynamically be more stable in the unfolded state. This is because the loss in favorability by reducing inter-residue interactions is outweighed by the increase in entropy multiplied by the temperature in Eq. (1), yielding an overall lower free energy G. The two state approximation is widely accepted and used in the protein community and is assumed in many protein models. Strictly speaking, the two state model deals with a binary designation of folded or unfolded macrostates, meaning in reality there exists two predominate ensembles of conformations. The nature of these conformations and changes between conformations is related to mechanisms. Furthermore, how the protein changes from the folded to unfolded or unfolded to folded thermodynamic state requires following a continuous change in

conformation. This is referred to as protein folding pathways, and again the study of this detail is concerned with mechanisms, for which the 2 state model cannot address.

$$\Delta G = \Delta H - T\Delta S \quad (1)$$

The minimum Distance Constraint Model (mDCM), which will be a large focus of this thesis, takes advantage of the statistical mechanics behind these thermodynamic laws. Using an x-ray crystal structure as an input, the model calculates the energy of all of the hydrogen bonds within the protein structure as well as the entropy associated with that conformation. With this information the mDCM is able to approximate the free energy of that conformation. Additionally the model can flicker these bonds on and off as well as perturb the original fold of the protein and recalculate all of the same quantities. By doing this we built an ensemble of conformations, which allows us to use statistical tools in order to tease out mechanical properties of the native domain. The details of this model will be expressed more rigorously in a later section.

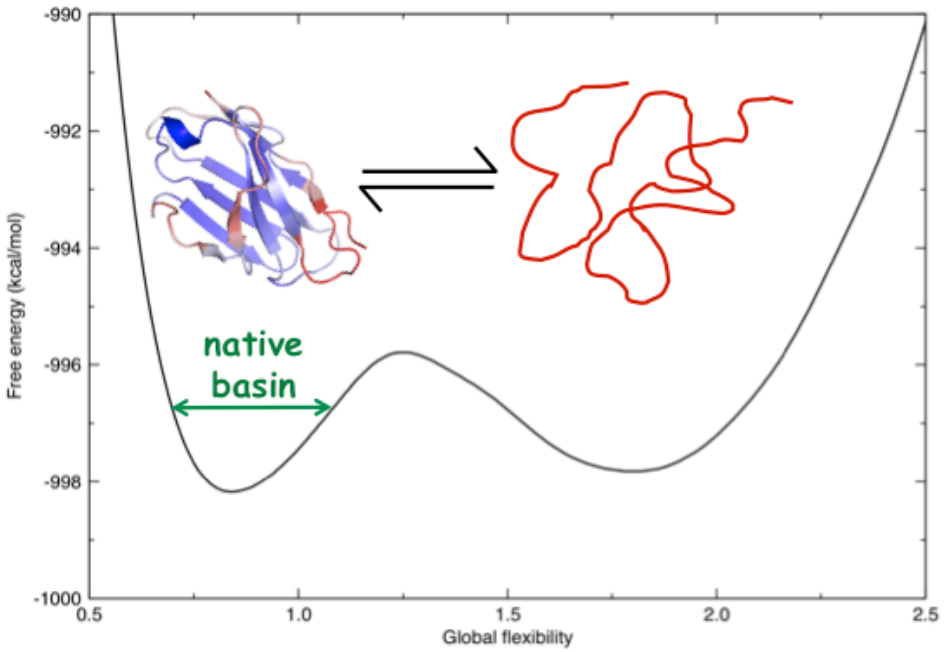

Figure 2: Free energy landscape of a protein. To the left the native basin has a local minimum describing the most probable fold. To the right is the unfolded state with its own energy minimum. This shows the two state model with a single transition from one basin to the other.

The barrier between the domains as seen in figure 2 dictates the transition between the folded and unfolded basins. In general the higher the barrier the more time it takes for the protein to transition from one state to another. Low thermodynamic barriers express a capacity to fluctuate more freely between native and disordered states. This is due to the fact that all protein conformation propensities are based on Boltzmann statistics, meaning that small differences in free energy translate into large differences in probabilities. This is why a two state model is such a good approximation for real systems; the conformation which exists at the minimal energy will exists with far more prevalence than similar structures with higher energy.

In practice this transition is correlated with the height of the heat capacity peak of the protein unfolding. The higher the peak of the heat capacity the more energy need to be added into the system in order to cause slight shifts in temperature, meaning the structure of the protein has a very large transitional barrier. This piece of information is interesting however they are merely correlated values, as this is only a typical case across globular proteins. The thermodynamic model and other models that calculate a free energy landscape with respect to a single order parameter as shown in Figure 2 reflect the typical cases, but more subtle cases that deviate from this paradigm are observed in nature, and, computational models that are based on physical principles should not rely on these characteristics, but rather predict them.

While the entire heat capacity curve is useful for accurately modeling the statistical weights for protein conformations within computational models; in practice obtaining the full curve through experimental methods like Differential Scanning Calorimetry (DSC) is difficult. However it is clear that experimental values are crucial to the validity of statistical methods. Therefore using other more readily available information is ideal, for example, melting temperature T_m. The melting temperature is important because is describes the point where the entropy begins to outweigh the enthalpy in Eq. (1) indicating a swift transition between the two basins. In addition melting temperatures can be experimentally determined using fairly simple procedures, Circular Dichroism (CD), for example. Little information about the thermodynamic properties of any chemokines has been found in the literature. Therefore in order to use any of the statistical modeling approaches mentioned thus far we must employ the use of CD spectroscopy in order to approximate the melting

temperature of CXCL7. Here we will briefly highlight the theory behind the CD spectroscopy approach used as a launch pad for this work.

The protein sample is concentrated in solution and maintained at a specific temperature while polarized incident light to emitted onto the sample. Secondary structure (α-helices and β-sheets) in proteins changes the polarization state of the induced light depending on the specific bonding structure within the secondary structure element as well as the wavelength of the incident light. Therefore by measuring the change in polarization state after light of a specific wavelength we can infer information about the presence of secondary structure. Alpha-helices and beta-sheets presence manifests near 208/222nm and 218nm respectively, meaning as we scan across the wavelength range (190-250nm here), the presence of secondary structure will be identifiable at these regions such as in panel A of Figure 3. By repeating this process over a span of increasing temperature values the change in the CD curve indicates a change in the overall secondary structure of CXCL7. Furthermore, by plotting the values of the CD curve at 220nm as a function of temperature we can see the transition of CXCL7 from the native to the disordered domain. Due to the mixed helix/sheet fold, 220nm was chosen to best capture the unfolding of both secondary structure elements. This is shown in panel B of Figure 3. The sharpness of the transition is an indicator of how accurate the two state approximation is for a particular sample. Here we see a wide transition indicating a slower unfolding pathway and less of a switch which is expressed in other protein families. In addition the inflection point indicates the temperature of which the

transition between native and disordered macrostates is occurring most rapidly, also known as the T_m.

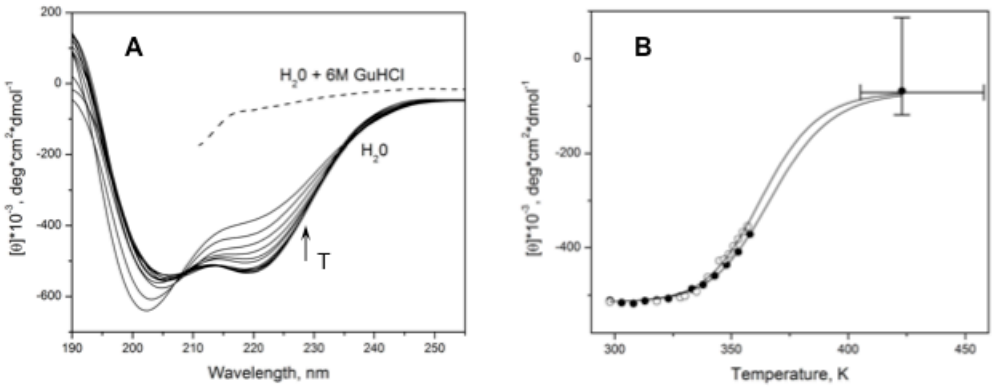

Figure 3: Temperature-induced dissociation and unfolding of CXCL7. A) Far-UV CD spectra of 43 μM CXCL7 dissolved in water (solid lines) at pH 5.0 acquired at different temperatures ranging from 25° C to 85° C. Dashed line represents the CD spectrum recorded on the same sample upon addition of 6 M GuHCl. B) Ellipticity at 222nm as a function of temperature, determined from the CD spectra recorded at the CXCL7 concentration of 43 μM (solid circles) and 12 μM (open circles). Solid lines represent global fits of experimental data, Error bars represent the ranges of the Ellipticity of unfolded protein and of the temperature, at which CXCL7 becomes fully unfolded that were used to determine the error in the determination of T_m.

CHAPTER 2: METHODS

2.1 Overview

It is clear that each of the methods reviewed have some advantageous or disadvantageous with respect to each other. Molecular Dynamic software is highly accurate, however computationally expensive and it cannot produce a sufficient exploration of conformational states to capture thermodynamics. Statistical models use empirical calculations based off experimental measurements, but can miss fine details, which are important to mechanistic detection. In addition these models can be very heavily biased by the reference structure. Experimental procedures are of course physically accurate, however, without a model to interpret the data, do little to provide insight into the mechanisms that involve the conformational space of a dynamic protein. Therefore this thesis outlines a hybrid MD/mDCM approach that captures the best of both worlds. MD is used to explore a large conformational space and used as more physical input structures for the mDCM's statistical toolbox. This procedure is outlines in the flowchart[25] in Figure 4.

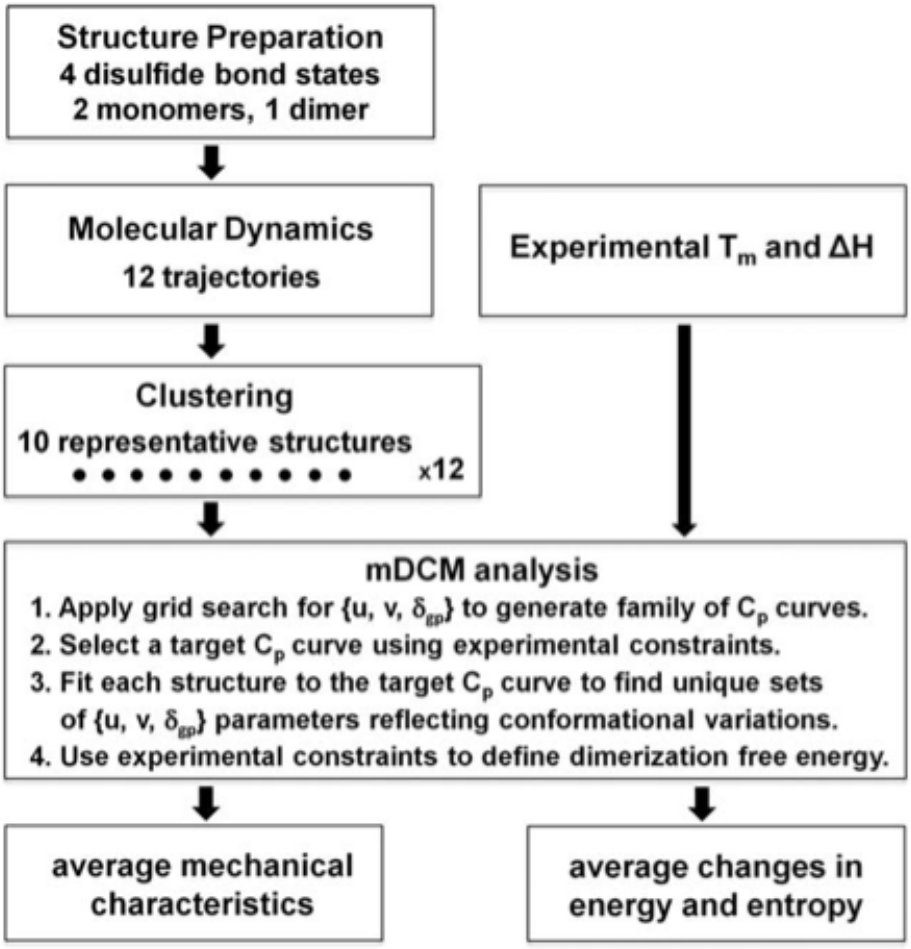

Figure 4: Flowchart of steps of combined CD/MD/mDCM analysis

In essence this workflow is simple despite some of the procedural complexity. Every structure in this study originated from a single x-ray crystallographic PDB. Using this single structure we computationally separated it into two monomer chains and one homodimer each with four disulfide bond permutations. Each of the 12 resulting structures was subjected to a long molecular dynamic simulation in order to change the original conformation to reflect the loss of the disulfide bonds. From each of these 12 trajectories 10 "snapshots" of the structures were taken that best reflect the conformational drift throughout the trajectories and used as starting points for the statistical mDCM model. At which point experimental parameters were used to add

physicality to the mDCM calculations across all 120 structures, which explored the mechanical and thermodynamic characteristics of CXCL7. Averages of mechanical and thermodynamic properties gave us a robust platform on which to analysis differences between disulfide bond states as well as the difference between monomer and dimer characteristics.

2.2 Structure Preparation

The X-ray Crystal structure of Chemokine CXCL7 (PDB ID: 1NAP) was used throughout this study.[26] 1NAP is a tetramer complex in which chains A and B where stripped out to create two separate monomer chains as well as a single homodimer. Aside from minor coordinate arguments the two chains of the dimer are nearly symmetrical and superimposable. In all cases software from the Chemical Computing Group called MOE (Molecular Operating Environment) was used to model all missing residues from crystal structure. Another piece of software was used to protonate based on environmental pH. This was in the form of a web server developed by Virginia Tech called H++. [27] After protonation state calculation MOE was used once again to minimize the free energy of the protein systems in attempt to best mimic natural conformations of CXCL7. After minimization steps in MOE three structures exist: monomer A, monomer B, and dimer AB.

2.3 Molecular Dynamics Simulations

GROMACS molecular dynamics software[28] was used to then remove disulfide bonds in order to create a set of structures for future analysis. The two disulfide bonds that connect Cys5-Cys31 and Cys7-Cys47 in each monomer chain where both individually removed and removed together. Thus each monomer had

four separate bond permutation structures: 1) The original monomer chain 2) Monomer chain with Cys5-Cys31 removed 3) Monomer with Cys7-Cys47 removed and 4) Monomer with both Cys5-Cys31 and Cys7-Cys47 removed. The dimer followed the same rule with each chain maintaining their homodimer status i.e. if Cys5-Cys31 was removed from chain A then it was also removed from chain B. This is a total of 12 PDB structures (4 monomer A, 4 monomer B, 4 dimer AB). Each of the twelve structures was minimized through GROMACS before at 100ns trajectory in the NVT ensemble with an AMBER99SB-ILDN force field[29] solvated by adding 10 angstrom of TIP3P water[30] in a periodic cubic box with counter ions added to neutralize the net charge. The potential energy minimization was followed by 1 ns of NPT and 1ns of NVT equilibration. Pressure was regulated using the extended ensemble Parrinello-Rahman approach[31] and Nose-Hoover temperature coupling controlled temperature. A cutoff of 10 angstroms was used for van der waals interactions, and the Particle-Mesh-Ewald method[32] was employed to account for the long range electrostatic interactions. All bonds to hydrogen atoms in proteins were constrained using LINCS[26], and bonds and angles of water molecules were constrained by SETTLE[27] allowing for a time step of 0.002ps. A total of 2,000 evenly spaced frames were gathered from the 100ns trajectory.

Following completion of the 100ns trajectories, each of the twelve MD runs were clustered by a K means clustering algorithm using KCLUST module[33] from MMSTB tool set[34] in such a way that the top ten representative structures covered at least 85% of the structural space. This was an attempt to maximize conformational diversity across the simulation as to more accurately mimic the dynamic nature of

proteins in biological systems. Of each K means cluster a centroid was chosen at the center of each respective cluster as a representative structure that was used to move forward with thermodynamic analysis of that conformation. At this point each of the twelve trajectories were clustered into ten representations for a total of 120 PDB structures.

These 12 different archetypes each with 10 representatives gives a more reliable significance to our conformational sampling. This is highly important due to the strength of the disulfide bonds in protein structure. These high-energy bonds are thought to be major stabilizers within protein structure, therefor by removing them we had to consider the possibility of the structure unfolding upon removing these bonds. MD simulations were seen to be the best method for modeling the conformation of the protein after the disulfide bonds were removed. Therefore the combined MD/mDCM approach as presented in this work as well as recent other works[35], is a two birds with one stone approach. Not only are we testing the validity of using the original crystal structure for structures in which the bond networks have been perturbed, we also gain much greater statistical power in the event that this is true. Which is the case here.

2.4 minimum Distance Constraint Model (mDCM)

The minimum Distance Constraint Model[36-38] (mDCM) is a statistical mechanics model that probes mechanical and thermodynamic characteristics of protein structures. The mDCM is built on calculating the free energy of protein conformations and developing a free energy landscape to establish the most probable conformations based on Boltzmann statistics. However where the mDCM steps apart

from many other models is the non-additive entropy of redundant constraints while calculating this free energy as described below. In this way it more accurately describes the thermodynamic states of the protein and has shown to reconstruct experimental heat capacities based off of the parameters that we will define here.

The crystal structure used as a reference for the model is used to construct a constraint network of permanent bonds (ie covalent bonds, disulfide bonds, etc) of which these bonds are unbreakable within the model. Using the reference conformation from the crystal structure as a "ground state" the model generates an ensemble of structures by perturbing away from the ground state. It breaks native contacts defined by the crystal structure. It defines two order parameters that are the number of hydrogen bonds N_{hb}, and the number of natively packed torsions N_{gp}. 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 $\left(N_{hb}^{\max}!N_{gp}^{\max}!\right)/\left(N_{hb}!\left(N_{hb}^{\max}-N_{hb}\right)!N_{gp}!\left(N_{gp}^{\max}-N_{gp}\right)!\right)$. The mixing entropy S_{mix} is then given by $S_{mix}=R\ln\left(\Omega_m\right)$, where R is the ideal gas constant.

Furthermore using Monte Carlo sampling the conformational entropy for each macrostate is estimated by:

$$S_c\left(N_{hb},N_{gp}\mid\delta_{gp}\right)=R\left[\delta_{gp}Q_{gp}+\delta_{pp}Q_{pp}+\delta_{hb}^{max}\sum_{k=1}^{N_{hb}^{max}}\left(1+\frac{1}{8}E_k\right)\langle q_k n_k\rangle\right] \quad (2)$$

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.[39] 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,[34] with corresponding entropy being a linear function of its energy.

The free energy landscape of CXCL7 is calculated by combining total energy and entropy contributions for each macrostate using the functional described in equation 3:

$$G(N_{hb}, N_{gp}) = U(N_{hb}) - uN_{hb} + \nu N_{gp} - T\left[S_m(N_{hb}, N_{gp}) + S_c(N_{hb}, N_{gp} \mid \delta_{gp})\right] \quad (3)$$

where U is the average intramolecular H-bond energy given as $U = \sum_{i=1}^{N_{hb}} E_i N_i$. Three parameters (u, v, δ_{gp}) are floating parameters designed to model specifically to each thermodynamic system based on experimental parameters. First the u parameter represents protein solvent hydrogen bonds. The v parameter represents native like packing of atoms within the protein, and $R\delta_{gp}$ as mentioned previously represents entropy of the packing constraint. Typically the three phenomenological parameters u, v, and δ_{gp}, were determined through a fitting mechanism to heat capacity data. However, here the only thermodynamic information known was T_m from CD spectroscopy experiments. Therefore multiple different input structures are considered using another approach called GRIDsearch, which will be reviewed in more detail later.

The free energy functional defined in Eq. (3) is applied to all nodes of an N_{hb} vs. N_{gp} two-dimensional continuum. By flickering each parameter on and off two basins emerge from the free energy landscape associated with this parameter space. The first of which is the native or folded basin which is typically where most of the QSFR calculations discussed later occur, and a second basin for the disordered or unfolded state of the protein. Nodes in which the free energy is minimized are selected as part of the native and disordered basins, but the models effectiveness is in the calculation of this energy. Each hydrogen bond has an associated energy value that gets added to the enthalpy when considering the difference in energy between two structures. This is a simple additive process; therefore the total ΔH between two structures is merely the difference in the total hydrogen bond energies within the conformation. However with the addition of each constraint we also must consider the change in conformational entropy of the structure. If the hydrogen bond connects two highly mobile regions, which are for the most part otherwise unconstrained, then the entropic cost of that bond is very high. However if a bond is introduced into an already constrained area (near secondary structure for example), then the addition of the extra hydrogen bond did little to change the conformational entropy of the structure. Therefore the mDCM considers whether or not each flickering constraint is redundant or not. If so then the bond is associated with a drop in enthalpy without incurring an entropic cost.

The reference enthalpy and entropy are buried into three fitting parameters used (u, v, δ_{gp}) during simulated annealing step. The importance of these experimental values was all outlined previously in this section. Each of the 80

monomer structures and 40 dimers as described in the previous section were fit to the predicted heat capacity curve resulting from GRIDsearch which will be described in a later section. After careful parameterization of the data it was determined that δ_{gp} could become fixed across the dataset with a value of 1.2 and 1.5 for the monomers and dimers respectively. This is a method that is commonly deployed by members of the BMPG group typically in protein family studies as the δ_{gp} parameters reflects structural similarity[40,41] of structures which is obviously the case with this bonding network permutation study. This fixation also shows the degree of "U-V compensation" which speaks to the enthalpic-entropic compensation that happens in protein thermodynamics. Figure 5 shows an example of such U-V compensation in anti-body fragments.[42]

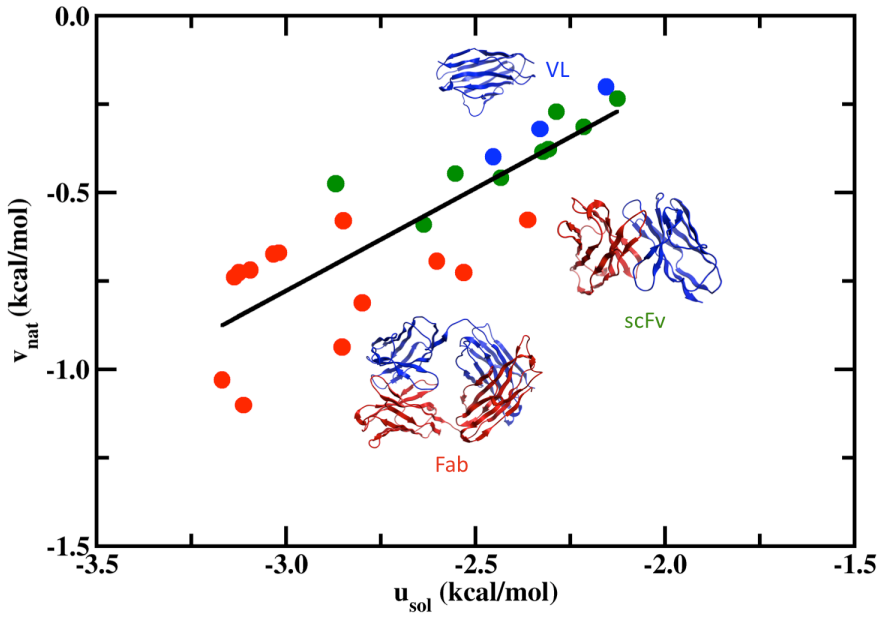

Figure 5: U-V Compensation example. Given a fixed δ_{gp} parameter *u* and *v* have been seen to correlate with each other.

2.5 GRIDsearch

Ideally experimental data would be inserted into the model in the form of heat capacity curves or melting temperature information. In the case of known heat capacity data the mDCM can be fit to these curves using three primary empirical parameters (u, v, δ_{nat}). However when only T_m is known (as in the case here) a procedure known as GRIDsearch is used to screen thermodynamic information in order to predict heat capacity curves. GRIDsearch was run for all 10 representative structures on three different trajectories. Monomer A, Monomer B, and Dimer AB trajectories that maintained their original disulfide bond network were the three trajectories in which this operation was run. By creating a three dimensional grid of our empirical parameters (u, v, δ_{nat}) and reconstructing the heat capacity curve based off of these parameters and constraint network an ensemble of potential curves are generated. This method uses all of the same calculations discussed in the mDCM by calculated the free energy functional based off of the fitting parameters. However the difference here is that the parameters (u, v, δ_{nat}) where calculated in an ensemble of combinations and the solution field is reduced using filtering methods. Collaborators have been able to experimentally determine melting temperature of CXCL7 monomer and dimer with its original disulfide bonds configuration. This information allows us to filter the large number of potential solutions created by GRIDsearch down to solutions within five degrees Celsius of the experimental melting temperature of 365K for monomer and 372K for dimer. Of all of the reasonable solutions across all of the representative structures a single curve was chosen for monomer and another for the dimer that best satisfied the solutions of the entire dataset.

2.6 Quantitative stability/flexibility relationships (QSFR)

Two primary metrics used here are backbone Flexibility Index (FI) and Cooperativity Correlation (CC).[37,38,43] When calculating these metrics we must consider phi and psi backbone torsion angles, as well at chi side chain torsion angles, which are associated with rotatable bonds. Such rotatable bonds are the biological analog to a simple hinge, albeit a hinge which can become rigidified and restricted due to local or global interactions. In flexible regions the number of rotatable bonds is denoted as H, and the number of independent disordered torsions as A. Therefore we can define the density of independent DOF within a region as $h_i = A/H$. Similarly within a rigid region the number of locked torsion angle is L, the number of redundant torsion angles as B, therefore the density of redundant constraints within a region is: $l_i = B/L$. The calculation of FI is defined as $f_i = (h_i - l_i)$ for the i-th rotatable bonds within a given microstate. FI values for a given rotatable bond over 0 denotes an excess number of degrees of freedom, while conversely a FI of less than 0 denotes an over constrained region. While any local region cannot be both rigid and flexible simultaneously, the reported FI is an ensemble of fi calculations across a multitude of possible constraint networks.

Cooperativity Correlations is calculated in a similar manner to that of FI however it reveals slightly subtler coupling between regions of the protein. That is that the CC matrix reveals information about rotatable bonds that flex together, or lock together to form rigid clusters. If two rotatable bonds i and j fall within the same flexible region, their CC matrix values are assigned a value equal to the density of independent DOF as defined in the FI calculation. $CC_{ij} = CC_{ji} = h_i$. Likewise, two

rotatable bonds within the same rigid cluster are given a value equal to the density, which fall in isostatic regions or belong to independent regions, are assigned a value of 0.

As mentioned previously in this section these calculations are based off of an average among some large ensemble of potential constraint networks. In reality each of these microstates is assigned a thermodynamic statistical weight based off of the probability of such a state occurred generated by the free energy landscape constructed by the mDCM. Keep in mind that this is all done for a single input structure and a single run of the mDCM. Here we combine this approach with another level of statistical weighting by considering the coverage of conformational space among the 10 representative structures. Therefore we consider the weighted average of the FI and CC for each MD trajectory. This approach both increases the robustness of the model by reducing the sensitivity to input structure as well as increasing the practicality of analyzing such a large amount of data.

A final filtering step was applied to the differences in the CC matrices in attempt to differentiate between statistically significant couplings and noise within the dataset. This filtering is a signal beyond noise ratio (SBNR) and defined as follows. The standard deviation sigma is calculated for all differences between matrix elements denoted as x. If $x > \sigma$, then $SBNR(x) = (x - \sigma)/\sigma$. If $x < -\sigma$, then $SBNR(x) = (x + \sigma)/\sigma$. Otherwise signal is below noise and assigned a value of 0. The result of this filtering only shows coloration in CC plots for differences which are greater than 1 standard deviation in either direction from the mean.

2.7 Free Energy of Solvation upon Dimerization

Estimating the free energy contribution from solvent exposure is part of all models that deal with implicit solvent. The mDCM uses a couple of parameters to effectively provide the effect of solvent as the protein unfolds, but when multiple structures are being compared at the same time within the same calculation, the solvent free energy contribution must be included. In this work, the monomer and dimer structures are being compared directly, as well as the different representatives that have been generated for each case. The problem we encounter is how to best model this contribution and how to make sure the solvation model is consistent with the rest of the mDCM free energy function. This modeling task is one that is currently being pursued in the lab. However, there is a way to obtain this information within an empirical model in a much easier way using experimental inputs. This is the approach taken in this work.

Given a specific mDCM calculation the free energy associated with an ensemble can be shifted by some arbitrary constant. The parameters of the DCM in some way reflect this effect however not to the degree of a comparison between states. Therefore in place of modifying the mean field approach to the solvation effects in the mDCM, an independent calculation was employed that uses the enthalpy, entropy, and temperature of dimer association based on the mDCM fitting procedure to experimental measurements. The free energy of any given structure compared to the reference state can be expressed as follows:

$G_k(T) = H_k(T) - TS_k(T) + (h_k - Ts_k)$, where h_k and s_k are constant shifts in energy and enthalpy due to small variation in solvation terms between MD frames. The

reference frames were determined from an average over all 80 monomer and 40 dimer structures. These h_k and s_k parameters were used to shift all structures free energy landscape as close to the reference from as possible according to a least squares error fitting. Further variation across the dataset reflects the change in conformation across the sampled space from MD trajectories as well as intrinsic error within the mDCM. Therefore we can approximate the change in free energy upon dimerization by the following logic. $\Delta H_a = H^D - 2H^M$ and $\Delta S_a = S^D - 2S^M$ where D and M represent the dimer and monomer respectively. 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 invoke the right to add any constant energy or entropy shift between the forms, it follows $H_{pred}^D = H_{mDCM}^D - \Delta H^0 + \Delta H_a$ and $S_{pred}^D = S_{mDCM}^D - \Delta S^0 + \Delta S_a$. Therefore the free energy of the dimer can be predicted as $G_{pred}^D = H_{pred}^D - TS_{pred}^D$.

CHAPTER 3: RESULTS

3.1 Stability/Flexibility of CXCL7

In figure 6 below we show the target heat capacity curves along with the associated fits as well as an example free energy landscape of each MD trajectory. 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.[42] As shown in Figure 6 the maximum heat capacity of the CXCL7 dimer is more than double than the monomer. This suggests a cooperativity effect that develops upon dimerization. The free energy landscapes shown do show evidence of a first order transition, however several of these landscapes have barriers that are below thermal fluctuation. Meaning that the two state model of native and disordered with a swift transition is not as accurate in this study as many previous efforts. Due to this all statistical sampling of thermodynamic parameters are considered over the entire conformational ensemble in place of only in the native basin.

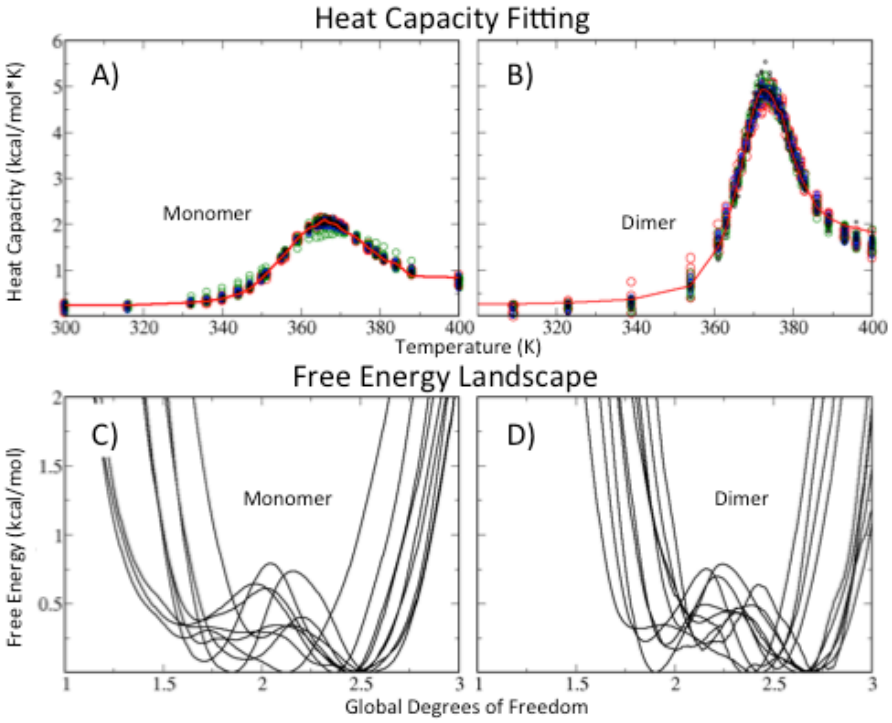

Figure 6: Heat capacity fits, and free energy landscapes. A) mDCM fitting results for monomer target (red line). B) Fitting results for dimer target (red line). C) FEL for 10 example monomer structures. D) FEL for 10 example dimers.

The backbone flexibility was analyzed at two different temperatures for both the monomer and the dimer cases. T=300K and T=350K were chosen specifically to show mechanically less stable regions of CXCL7. Note that this temperature is still below the experimental melting temperature determined by CD experiments, however still exposes weakness in the hydrogen bond network. Typically the backbone flexibility increases uniformly with temperature due to the increase in rotatable bonds with increased thermal energy. In the monomer, the N-terminus and 50's loop are flexible, while the three beta-strands, C-terminal, α-helix, 3_{10} alpha-helical turn are rigid expressing a negative value for FI. 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.

It is assumed that dimerization stabilizes the structure of CXCL7, as the biologically relevant for exists in dimer state. One of the contributing factors to this stabilization effects is the formation of interfacial hydrogen bonds. It is evident that these hydrogen bonds form in regions that connect rigid regions of monomer A to rigid regions of monomer B. The histogram below shows the location of interfacial hydrogen bonds mapped along the flexibility index. On average, about 13 H-bonds are formed at the interface. Backbone flexibility is seen to increase in the 30s loop even though this is a

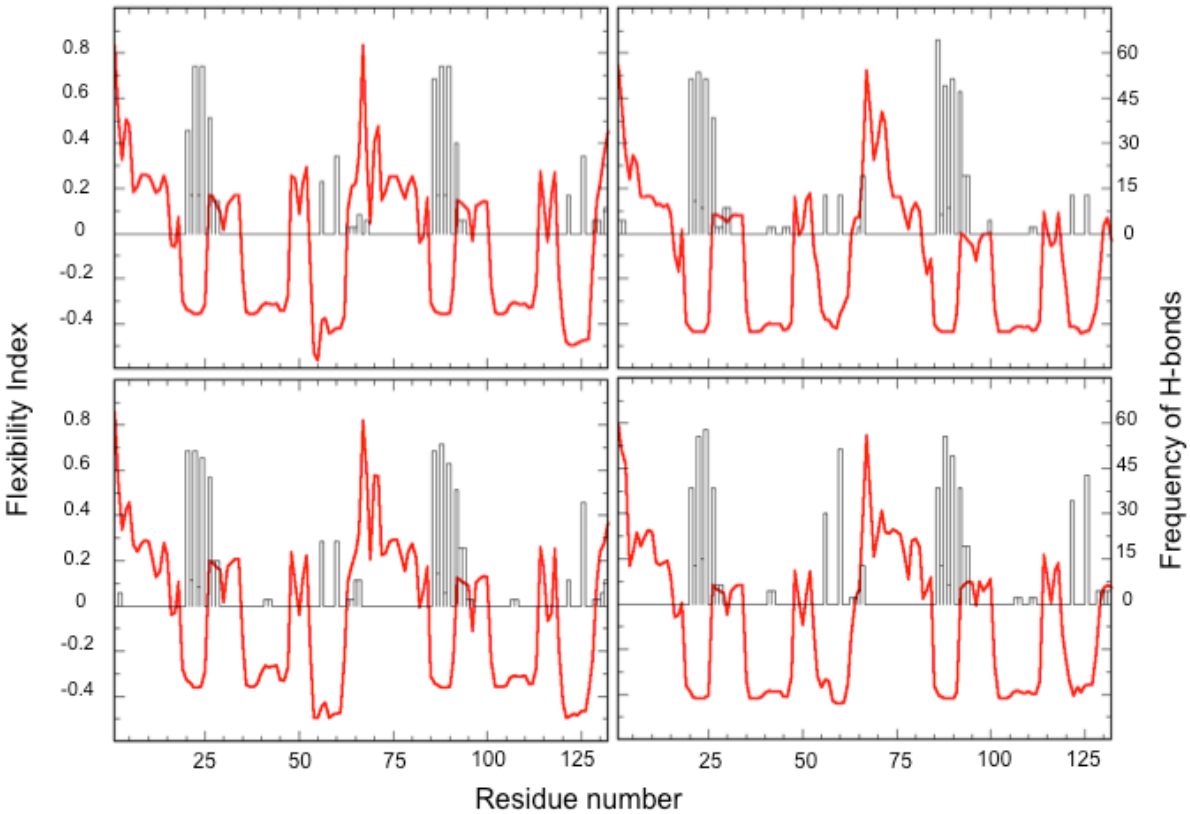

Figure 7: Histogram of Interfacial H-bonds according to residue number. Mapped against backbone flexibility of dimer chain. Both disulfide bonds present (upper left), 1st bond removed (upper right), 2nd removed (bottom left), and both removed (lower right).

region in which these additional hydrogen bonds form. Throughout the C-terminal alpha helices backbone flexibility is not significantly changed due to dimerization, even though the two C-terminal helices line up and form additional side chain hydrogen bonds across the interface.

Figure 8 shows the residue-to-residue mechanical couplings again at T=300K and T=350K for both monomer and dimer averages. At 300K, the secondary structure elements dominate the rigid clusters with connections between these beta sheets and alpha helices. As a whole CXCL7 is a fairly rigid system at 300K, however when the temperature is increased to 350K the fluctuation in the hydrogen bonds network shows as there is an increase in flexibly correlated motion in the 50's loop coupled with the N-terminus. It is also shown that the residues which interconnect the beta sheets (residues 25-35) significantly increase in flexibility with respect to the 300K structure. In the case of the dimer a similar trend is exhibited however with smaller magnitude of differences. Interestingly at 300K the dimer form of CXCL7 tends to have more innate flexibility with respect to the monomer. It is unclear whether or not this is functionally relevant effect or just a byproduct of the crystal structures. Again as with the monomer we see rigid clusters in locations of secondary structure.

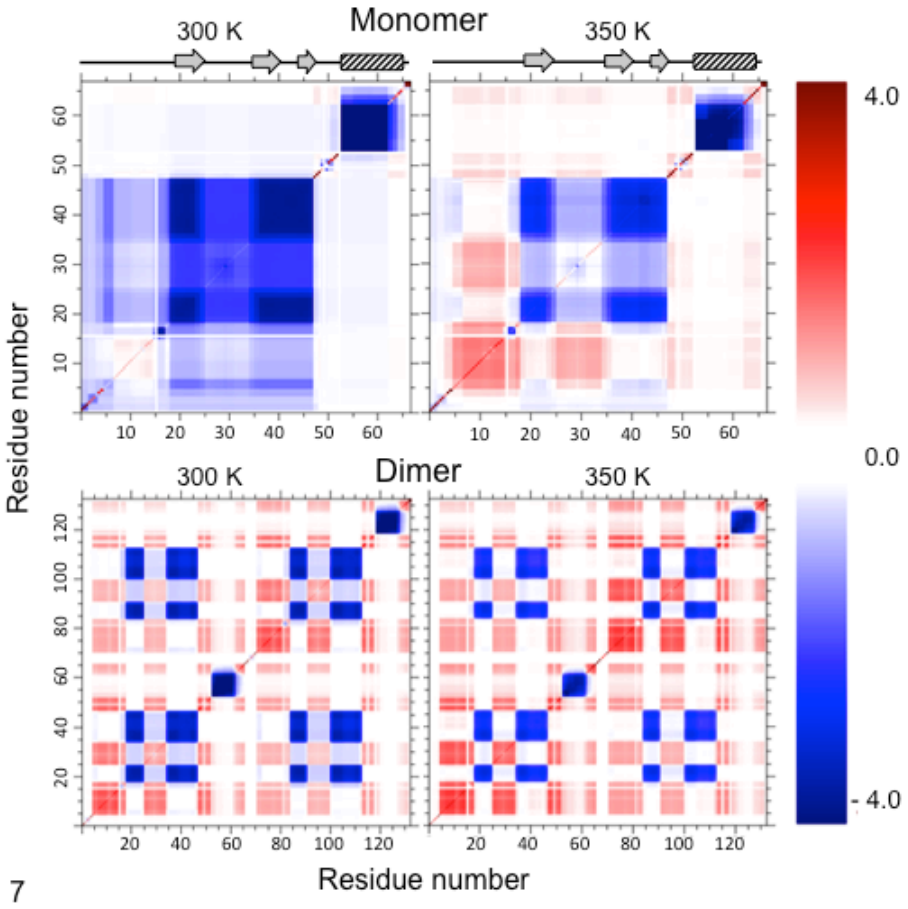

Figure 8: Correlated Cooperativity plots of ensemble averages. Blue represents correlated rigid clusters and red represents correlated flexible regions.

3.2 Effect of Disulfide Bonds

As mentioned previously one of the defining characteristics of the Chemokine family is the presence of two disulfide bonds in each monomer chain connecting Cysteine residues with only a few exceptions to this rule across the family. In effort to investigate the structural significance of these disulfide bonds to the stability of the structure we employed the MD/mDCM hybrid approach in which one or both disulfide bonds were removed from the structure. The melting temperature was computationally determined to drop slightly with the removal of each disulfide bond. This effect appears to be greater in case of the monomer as opposed the dimer cases

suggesting that disulfide bonds do not play as crucial of a role as commonly thought. This result agrees with reported experimental data on a similar chemokine CXCL4.

Table 1: Description of thermodynamic quantities as variation with disulfide bonds. 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 removed. (++) and (--) indicated the most and the least favorable case, respectively.

Disulfide bonds present	$\Delta G_{D\text{-}2*M}$ (ΔH-$T\Delta S$)	$\Delta H_{D\text{-}2M}$	$-T\Delta S_{D\text{-}2M}$	$\Delta HB_{D\text{-}2M}$	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

Table 1 quantifies the thermodynamic stability for each disulfide bond configuration in terms of change in Gibbs free energy enthalpy, entropy, total energy of Hydrogen bond network, and the number of interfacial hydrogen bonds formed during the dimerization process. It is seen that dimerization is energetically favorable, and enthaplically driven for each of the disulfide bonds configurations. Note that the overall shape of CXCL7 does not change significantly across each of these disulfide bonds permutations, meaning that the differences in $-T\Delta S$ are due to small conformational entropy changes. The differences in $-T\Delta S$ varies by about 7 kcal/mol across the disulfide bond permutations, which can be accounted for by minor perturbations to the hydrogen bond network. However it is worth noting that free

energy differences do not mirror differences in total energy of the hydrogen bonds network. The correlation coefficient of 0.11 between the change in free energy and change in total hydrogen bond energy suggests that changes in the conformational entropy play a large role in dimer stabilization.

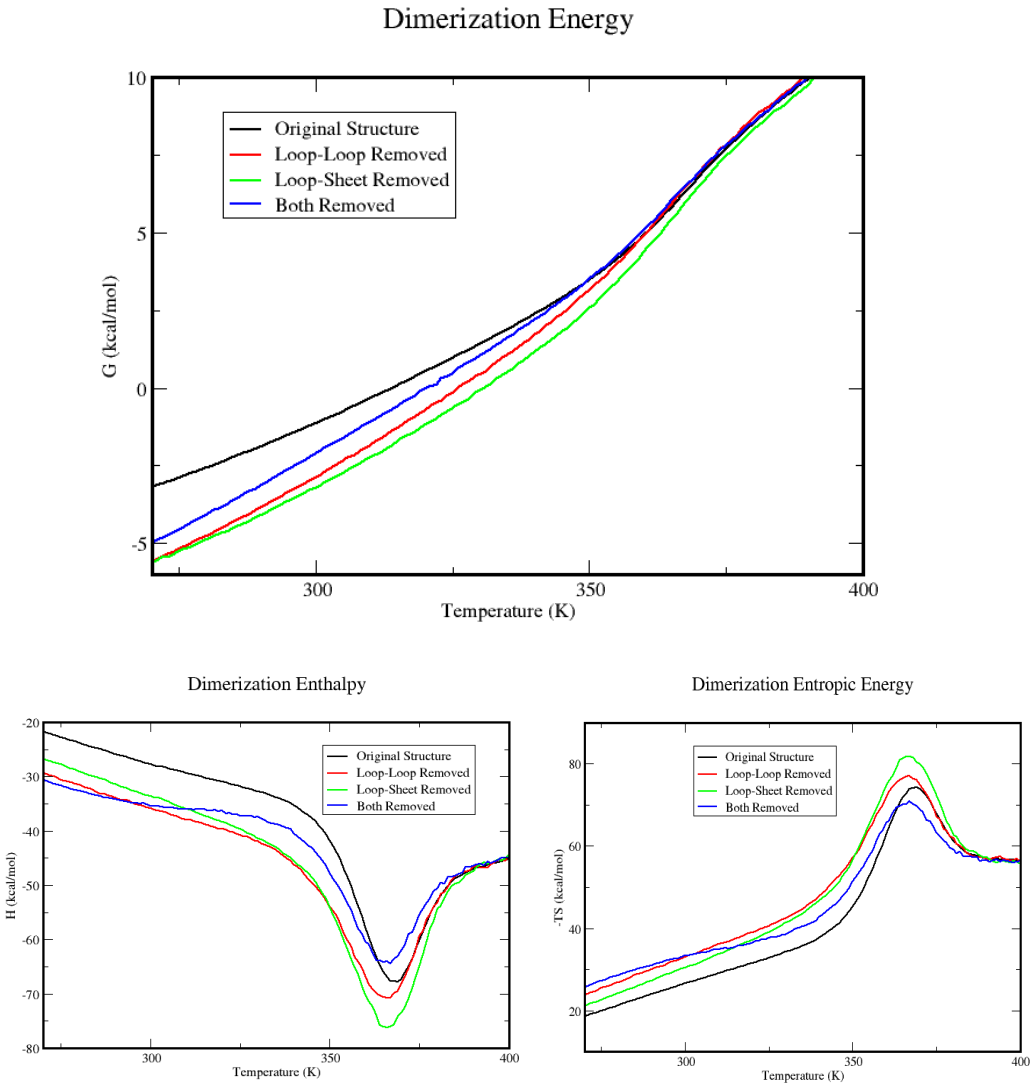

Figure 9: Thermodynamic Decomposition of Dimerization A) Free energy of disulfide bond states. B) Enthalpy of disulfide bond states. C) Entropic energy of disulfide bond states.

Figure 9 above describes the total free energy as well as the enthalpy entropy components of free energy. Interestingly we see that dimerization stabilizes the structure in which the loop-loop disulfide bond is removed. Looking at the enthalpy this is due to a spike in enthalpic stabilization, indicating the formation of more and/or stronger hydrogen bonds during dimerization. However this also comes with the greatest entropic price. In addition we can consider the differences in thermodynamic quantities from disulfide bond states in the monomer and dimer independently. Figure 10 below describes the differences (DS bond case – Original structure) in free energy

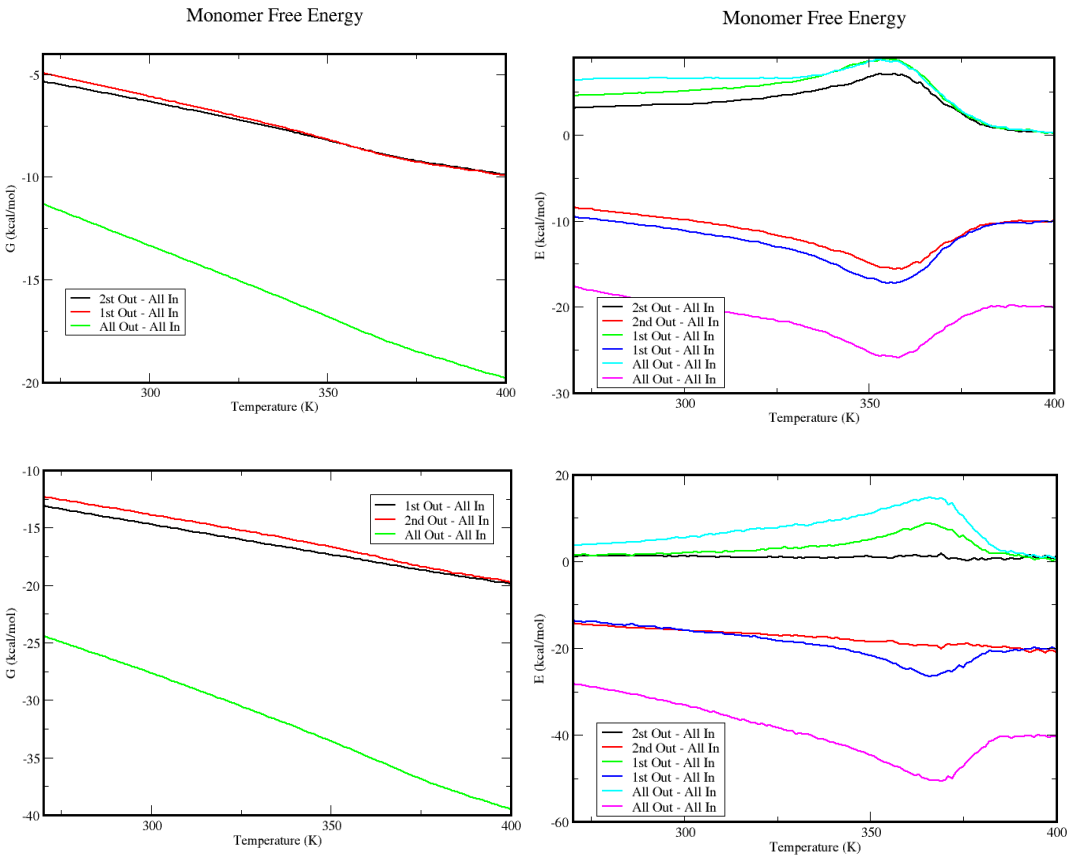

Figure 10: Thermodynamic Decomposition of Disulfide Bonds A) Free energy of disulfide bond differences (monomer). B) Enthalpy and Entropy of disulfide bond differences (monomer). C) Free energy of disulfide bond differences (dimer). D) Enthalpy and Entropy of disulfide bond differences (dimer).

Enthalpy and entropy. Here we have been able to tease out differences between the monomer and dimer states. For both monomer and dimer (panels A and C) the free energy looks rather uninteresting. Both of the single disulfide bond permutations appear to have similar thermodynamic relevance. However looking at the components independently shows a more interesting story. For the monomer (panel B) we see greater enthalpic stabilization with the structure in which both bonds were removed; however the entropic destabilization is similar in all three cases. The dimer (panel D) however, shows that removing the 1st disulfide bond has no enthalpic or entropic effect with respect to the original structure with both bonds present. However we see an increase in both of these values for the second bond. This indicates that removing the loop-loop disulfide bond has a greater thermodynamic impact on the structure than removing the loop-sheet bond.

Dimerization is favorable at low temperatures due to a few key hydrogen bonds at the interface, which lower the enthalpy enough to overcome the conformational entropy reduction. At higher temperatures the entropic piece becomes more important in the free energy balance and causes the dimer to dissociate. Interestingly the predicted temperature of dissociation is highly context dependent due to molecular packing and hydrogen bond arrangements as a result of changes in the bond network. The dissociation temperature is predicted to be lowest when both disulfide bonds are present, next lowest when both are removed, and maximum when only Cys7-Cys47 is removed from each monomer chain. This suggests that the two disulfide bonds do not contribute equally to the stability of CXCL7 as removing one or both of these bonds leads to non-obvious effects. These effects are reflected in the

change in flexibility index. In the case of the monomer removing either disulfide bond increases the global backbone flexibility of the protein. However the greatest increase in flexibility for all of the monomer cases is present in the C-terminus, the location of the alpha helix on each monomer chain. Even though these differences are rather small in magnitude each bond affects different regions of the protein. However in the case of the dimer the removal of disulfide bonds tend to decrease flexibility, which is contrary to the monomer results. Removing Cys7-Cys47 yielded the greatest increases in backbone flexibility, which is believed to be a result of redundancy within the constraint network. Meaning the disulfide bond was removed from a region which connected the N-terminal loop to the edge of central beta sheets which already have rigid structure.

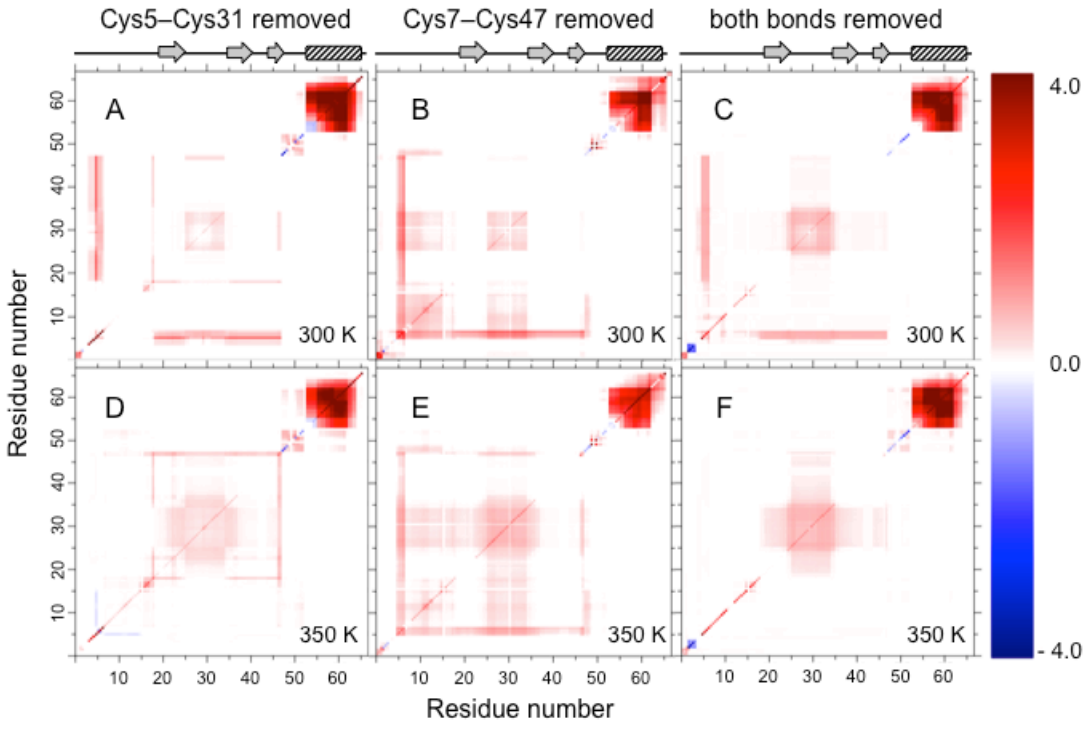

Figure 11: The difference in residue-residue mechanical couplings is plotted for the monomer at T=300K. That is, the residue-residue mechanical couplings for the Cys5-Cys31 disulfide bond removed in each monomer subtracted by the residue-residue mechanical couplings when both disulfide bonds are present in each monomer (A). The same type of plot is made, but for the Cys7-Cys47 disulfide bond removed (B) and then for the case that both disulfide bonds are removed (D). Instead of plotting raw differences, this data is filtered based on the signal beyond noise ratio (SBNR) as explained in the text. The same corresponding difference plots for T=350K are plotted using the SBNR in panels D, E and F.

Similar effects are seen as we move upward into the higher order coupling with differences in CC matrices. Figure 9 shows the residue-residue couplings for cases where one or both of the disulfide bonds have been removed. Note that these differences plots only show statistically significant differences which respect to the both bond present structures based on the signal beyond noise ration (SBNR) discussed previously. The most significant increase in correlated flexibility is seen in the C-terminal alpha helices at which a partial unfolding event is happened when the

constraint network is destabilized. This increase in global flexibility is consistent with the increase in conformational entropy within the monomer. Figure 9 depicts a similar situation with the monomer however with a few more interesting results. Removal of Cys5-Cys31 from each monomer chain in the homodimer shows a slight increase in flexibility throughout the alpha helix at T=300K, however at 350K as far more significant increase in flexibility occurs in the same region. This trend continues across the of bond permutations indicating that there is a stabilizing effects of removing bonds at low temperatures with is consistent with thermodynamic analysis. However once the temperature is increased to 350K the stabilization of the dimer no longer can overcome the fluctuations within the constraint network and the C-terminal alpha helices partially unfold similar to the monomer cases.

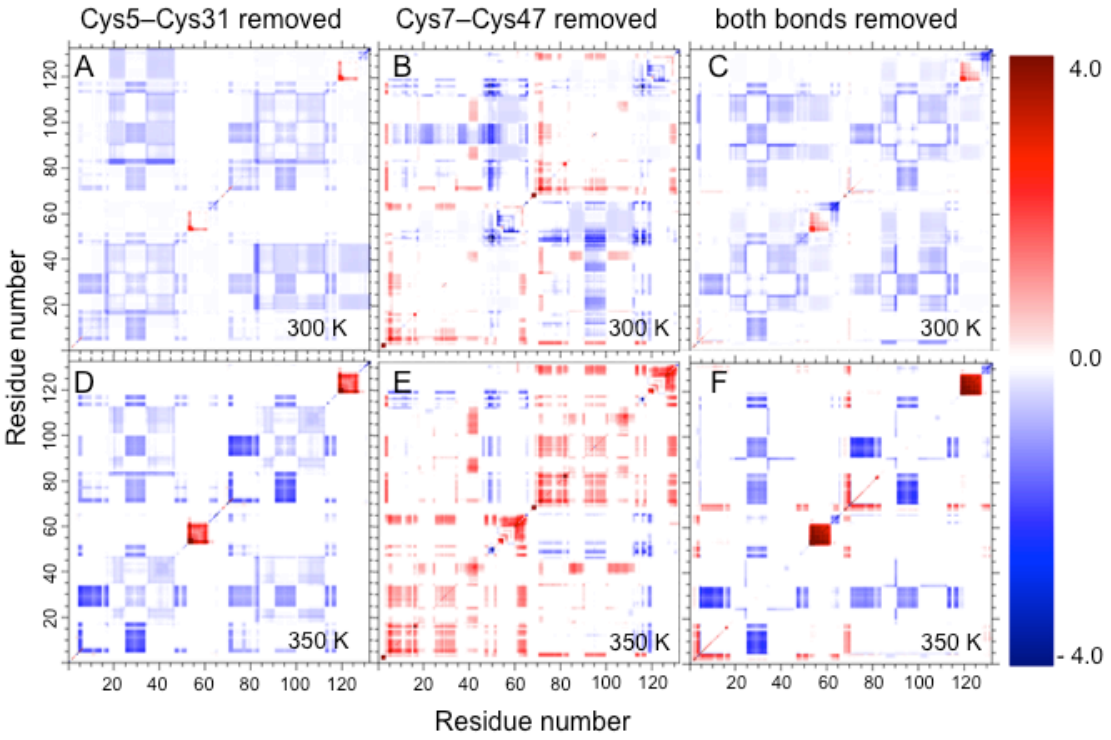

Figure 12: The difference in residue-residue mechanical couplings is plotted for the dimer at T=300K. That is, the residue-residue mechanical couplings for the Cys5-Cys31 disulfide bond removed in each monomer subtracted by the residue-residue mechanical couplings when both disulfide bonds are present in each monomer (A). The same type of plot is made, but for the Cys7-Cys47 disulfide bond removed (B) and then for the case that both disulfide bonds are removed (D). Instead of plotting raw differences, this data is filtered based on the signal beyond noise ratio (SBNR) as explained in the text. The same corresponding difference plots for T=350K are plotted using the SBNR in panels D, E and F.

3.3 Discussion

The results gathered from the described MD/mDCM workflow in addition to knowledge gathered from collaborative efforts in other computational and experimental procedures (ANM, CD, and NMR) lead to a consistent description for the dynamics of CXCL7. Minor differences across the methods pose inconsistencies at the superficial level, however with deeper analysis we see that these are in fact

complementary. For example: ANM analysis show large fluctuations in the 30's and 40's loop, with less so in the 50's loop. NMR relaxation shows that Val34-Val36 go through high frequency motions, but no other residues in the 30-50 loop's undergo such motion. The mDCM predicts high flexibility in the 50's loop with lower flexibility in the 30's and 40's loops. The 50's loop is seen to be less flexible in the dimer than in its monomer counterparts, we believe that mobility in this region is contributed from low frequency modes as determined from ANM analysis. However mDCM indicates with the residue-residue couplings shown in figure that a small degree of rigidity propagates between the C-terminus alpha helix and the beta-sheet dominated core. As temperature increases this relationship changes into a weak flexibility correlation. Note that these figures represent average correlations over statistical ensemble of structures. Therefore this hinge represents an average mechanism across many conformations, which has an appreciable amount of mobility, mainly in higher frequency motions.

Mutagenesis studies have shown N-terminal residues preceding the first cysteine residue, the N-loop preceding the first beta-strand, and the 30's loop connecting the first and second beta-strands all play crucial roles in the receptor binding in most Chemokines.[17,18,44] These receptor-binding sites are broken up into two sites. At Site I, the N-loop interacts with the N-terminus of the receptor chain. At Site II, the N-terminal residues of CXCL7 as well as the 30's loop interact with extracellular loop/transmembrane resides of the receptor.[17,44] In other Chemokines the N-terminal residues before Cys5 has been shown to play a pivotal role in neutrophil activation.[45] In addition the ANM analysis showed that these residues, which are

responsible for binding to several different receptors, have an affinity for being highly mobile.[46]

Now looking more deeply in the dynamics of the 30's loop and the N-loop resides, which have previously not been investigated in depth, other Chemokines show that this range is important for binding CXCR2 ligands.[19] The mDCM shows high flexibility in the 30's loop at 300K in the monomer state and being very flexible in the dimer state. However the ANM analysis did not detect large motion amplitudes in this region indicating that the loop adopts a specific conformation. Even though the mDCM predicts a high level of flexibility in a region this only represents a high number of degrees of freedom and not actual dynamics. Additionally ANM analysis shows that residues in the 30's loop move collectively within the loop as well as isolated regions across the protein, indicating that the motion is not random but high coupled to the global conformation. The mDCM also shows the N-terminus of CXCL7 as flexibly correlated to the 30's loop, again suggesting a coupling between binding Site I and Site II. A perturbation at either of these sites will propagate to the other through conformational rearrangement. It also infers that mutations at these binding sites may perturb the dynamics by altering the constraint network, therefore directly changing the mechanical couplings. Another recent study of a binding event between CXCL8 and CXCR2/CXCR1 ligand shows that altering the GP motif in the 30's loop caused changes in dynamics and conformational rearrangement of the binding event.[19]

This coupling between Site I and Site II is intuitive in the scope of the constraint network due to the presence of the disulfide bond between Cys5 and Cys31

which links these two regions. We have seen large Rex contributions for Cys31, which suggests significant slow motions involving Cys31. Other Chemokines have shown similar slow motions in the 30's loop for residues in close proximity with the relative cysteine.[6,8,47] Meaning that during ligand binding the N-terminus of CXCL7 samples a larger conformational space and takes longer to establish the correct binding conformation. This is in agreement with other chemokine binding studies, in which altering the N-terminal disulfide bond had deleterious effects on the binding affinity of receptors.[48,49]

Surprisingly the mDCM predicted that the disulfide bonds are not critical for maintaining protein stability. This result is consistent with a study carried out on Thrombocidin-1 (TC-1), which contains only two C-terminal deletions with respect to CXCL7.[50] The mDCM also shows that the C-terminal alpha helix is weakly coupled to the 50's loop. Otherwise there is no coupling between the helix and the rest of the protein regardless of the disulfide bond state. Note that a few key residues have been experimentally determined as functionally significant Lys17, Lys41, Arg54, Lys56, Lys57, Lys61, and Lys62.[10,50,51] Each of these residues fall in the correlated rigid regions for both monomer and dimer. The C-terminal helix is seen to be able to move more relative to the rest of the protein, which was also observed in CXCL4 where large scale motions relative to the beta sheet in MD simulation.[52] This suggests that it may be possible to remove the C-terminal alpha helix while retaining receptor binding ability. However residue substitution in highly correlated regions of the protein have been shown to significantly alter stability of the protein.[41] Furthermore,

these mutations would likely affect the receptor binding by shifting the equilibrium conformational ensemble.[19,53-56]

Our approach shows a clear increase in flexible correlation between the 30's and 50's loop upon dimerization. In addition we see an increase in the structural stability upon dimerization due to interfacial hydrogen bonds. We see that upon dimerization each monomer becomes more rigid within the beta sheets due to interfacial hydrogen bonds. Simultaneously we see a large increase in flexibility and correlated flexibility within loop regions.

Removing the disulfide bonds does not cause total unfolding of the CXCL7 structure, however the perturbation to the constraint network of the monomers pushes the equilibrium towards dimerization as a stabilization effect. Meaning that there is an increase in the number of hydrogen bonds formed across the interface when disulfide bonds are removed from monomer structures. This infers a "snap on" mechanism occurs during the dimerization process. Monomers are predisposed to propagate rigidity and flexibility through specific channels. Dimerization allows hydrogen bonds to form between these rigid channels producing an extended rigid cluster. Interestingly we see that the presence of both disulfide bonds maximize the flexibility couplings in the 30's and 50's loop regions, however the dimer with this configuration is the least thermally stable. This information suggests that biologic function is maximal which a marginal stability between the monomer and dimer forms. In addition, this configuration also produces the maximal contrast in couplings between the 30's and 50's loops.

CHAPTER 4: CONCLUSION

This MD/mDCM approach outlined in this work establishes a viable path to study the thermodynamic and mechanical properties of chemokines on a larger scale. The approach outlines detailed information about native state ensembles, as well as the effect of disulfide bonds on structure stability of monomer and dimer states. Combing the accuracy of MD simulations to explore conformational space with the high speed mechanical and thermodynamic ensemble calculations of the mDCM a robust yet sensitive platform has been developed for analyzing inter-residue couplings and interactions. This physical based information provides insight into the mechanisms that are responsible for carrying out certain functions.

A consistent picture emerges for the effect of dimerization and role of the Cys5-Cys31 and Cys7-Cys47 disulfide bonds. The presence of disulfide bonds is not critical for maintaining structural stability in the monomer or dimer. However even though CXCL7 did not denature when the disulfide bonds were removed there were large changes to the hydrogen bond network as a result, possibly altering mechanistic pathways. The monomer is destabilized more than the dimer upon removal of disulfide bonds. This is due to interfacial hydrogen bonds connecting rigid regions of each monomer chain, extending the rigid clusters across the interface. Disulfide bonds play a key role in shaping the characteristics of native state dynamics and modulating relative stability between monomer and dimer. The combined analysis elucidates how CXCL7 is mechanically stable as a monomer, and how upon dimerization flexibly correlated motions are induced between the 30s and 50s loop within each monomer and across the dimer interface. Interestingly, the greatest gain

in flexibility upon dimerization occurs when both disulfide bonds are present in each domain, and the homodimer is least stable relative to its two monomers. This last finding is counter-intuitive to most biochemists, but the explanation is rooted to how rigidity propagates through molecular structure. This simple physical explanation now provides insight into why the disulfide bonds are highly conserved in the chemokines family. Based on the findings in this work, it appears the disulfide bonds facilitate a structural mechanism that is turned to optimally distinguish functional characteristics between monomer and dimer.

CHAPTER 5: FUTURE WORK

This work shows the validity of the mDCM as a thermodynamic and mechanical model for CXCL7. Moving forward we intend to implement similar methods across the entire Chemokine family, which has high conservation of fold class, disulfide bonds location, and overall size. However the diversity of this family is far reaching making it an interesting target for a structural bioinformatics study. The chemokine family of proteins has 47 members, many of which have x-ray crystal structures or at least Nuclear Magnetic Resonance (NMR) structures that can be used as templates for this hybrid approach. This is a small number of targets when considering the diversity of dimerization and receptor binding events that are possible. By probing the thermodynamic and mechanical properties of the entire family we will be able to determine which inter-residue couplings are conserved across the family, possibly indicating a mechanistic reason for the multiple receptors binding affinity or the heterodimer propensity.

Due to the range of bodily functions that elicit chemokine responses a study of this magnitude could be critical to the reaching the next step in therapeutic development of both acute and chronic immune response. The conservation of sequence and tertiary structure across the family could be the key to developing inhibitors or enhancers for a slue of immunological responses.

REFERENCES

1 Baggiolini, M. Chemokines and leukocyte traffic. *Nature* **392**, 565-568, doi:10.1038/33340 (1998).

2 Raman, D., Sobolik-Delmaire, T. & Richmond, A. Chemokines in health and disease. *Exp Cell Res* **317**, 575-589, doi:S0014-4827(11)00007-3 [pii] 10.1016/j.yexcr.2011.01.005 (2011).

3 Gerard, C. & Rollins, B. J. Chemokines and disease. *Nat Immunol* **2**, 108-115, doi:10.1038/84209 (2001).

4 Bonecchi, R. *et al.* Chemokines and chemokine receptors: an overview. *Front Biosci (Landmark Ed)* **14**, 540-551, doi:3261 [pii] (2009).

5 Clark-Lewis, I. *et al.* Structure-activity relationships of chemokines. *J Leukoc Biol* **57**, 703-711 (1995).

6 Mayer, K. L. & Stone, M. J. Backbone dynamics of the CC-chemokine eotaxin-2 and comparison among the eotaxin group chemokines. *Proteins* **50**, 184-191, doi:10.1002/prot.10238 (2003).

7 Ye, J., Mayer, K. L., Mayer, M. R. & Stone, M. J. NMR solution structure and backbone dynamics of the CC chemokine eotaxin-3. *Biochemistry* **40**, 7820-7831, doi:bi010252s [pii] (2001).

8 Ye, J., Mayer, K. L. & Stone, M. J. Backbone dynamics of the human CC-chemokine eotaxin. *J Biomol NMR* **15**, 115-124 (1999).

9 Grasberger, B. L., Gronenborn, A. M. & Clore, G. M. Analysis of the backbone dynamics of interleukin-8 by 15N relaxation measurements. *Journal of Molecular Biology* **230**, 364-372, doi:S0022-2836(83)71152-6 [pii]

10.1006/jmbi.1993.1152 (1993).

10 Nguyen, L. T. *et al.* Exploring platelet chemokine antimicrobial activity: nuclear magnetic resonance backbone dynamics of NAP-2 and TC-1. *Antimicrob Agents Chemother* **55**, 2074-2083, doi:AAC.01351-10 [pii] 10.1128/AAC.01351-10 (2011).

11 Rajarathnam, K., Li, Y., Rohrer, T. & Gentz, R. Solution structure and dynamics of myeloid progenitor inhibitory factor-1 (MPIF-1), a novel monomeric CC chemokine. *J Biol Chem* **276**, 4909-4916, doi:10.1074/jbc.M005085200
M005085200 [pii] (2001).

12 Young, H., Roongta, V., Daly, T. J. & Mayo, K. H. NMR structure and dynamics of monomeric neutrophil-activating peptide 2. *Biochem J* **338 (Pt 3)**, 591-598 (1999).

13 Baryshnikova, O. K. & Sykes, B. D. Backbone dynamics of SDF-1alpha determined by NMR: interpretation in the presence of monomer-dimer equilibrium. *Protein Sci* **15**, 2568-2578, doi:15/11/2568 [pii] 10.1110/ps.062255806 (2006).

14 Jansma, A. L., Kirkpatrick, J. P., Hsu, A. R., Handel, T. M. & Nietlispach, D. NMR analysis of the structure, dynamics, and unique oligomerization properties of the chemokine CCL27. *J Biol Chem* **285**, 14424-14437, doi:M109.091108 [pii]

10.1074/jbc.M109.091108 (2010).

15 Kim, K. S., Rajarathnam, K., Clark-Lewis, I. & Sykes, B. D. Structural characterization of a monomeric chemokine: monocyte chemoattractant protein-3. *FEBS Lett* **395**, 277-282, doi:0014-5793(96)01024-1 [pii] (1996).

16 Liou, J. W. *et al.* 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, doi:10.1371/journal.pone.0094178 PONE-D-13-48443 [pii] (2014).

17 Rajagopalan, L. & Rajarathnam, K. Structural basis of chemokine receptor function--a model for binding affinity and ligand selectivity. *Biosci Rep* **26**, 325-339, doi:10.1007/s10540-006-9025-9 (2006).

18 Allen, S. J., Crown, S. E. & Handel, T. M. Chemokine: receptor structure, interactions, and antagonism. *Annu Rev Immunol* **25**, 787-820, doi:10.1146/annurev.immunol.24.021605.090529 (2007).

19 Joseph, P. R. *et al.* Dynamic conformational switching in the chemokine ligand is essential for G Protein coupled-receptor activation. *Biochem J*, doi:BJ20130148 [pii]10.1042/BJ20130148 (2013).

20 von Hundelshausen, P., Petersen, F. & Brandt, E. Platelet-derived chemokines in vascular biology. *Thromb Haemost* **97**, 704-713, doi:07050704 [pii] (2007).

21 Adcock, S. A. & McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chemical Reviews* **106**, 1589-1615, doi:10.1021/cr040426m (2006).

22 Anandakrishnan, R., Drozdetski, A., Walker, Ross C. & Onufriev, Alexey V. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophys J* **108**, 1153-1164, doi:10.1016/j.bpj.2014.12.047.

23 Karplus, M. Molecular dynamics of biological macromolecules: A brief history and perspective. *Biopolymers* **68**, 350-358, doi:10.1002/bip.10266 (2003).

24 Deng, Y. & Roux, B. Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. *J Chem Phys* **128**, 115103, doi:10.1063/1.2842080 (2008).

25 Herring, C. A. *et al.* Dynamics and thermodynamic properties of CXCL7 chemokine. *Proteins* **83**, 1987-2007, doi:10.1002/prot.24913 (2015).

26 Malkowski, M. G., Wu, J. Y., Lazar, J. B., Johnson, P. H. & Edwards, B. F. The crystal structure of recombinant human neutrophil-activating peptide-2 (M6L) at 1.9-A resolution. *J Biol Chem* **270**, 7077-7087 (1995).

27 Gordon, J. C. *et al.* H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* **33**, W368-371, doi:33/suppl_2/W368 [pii]10.1093/nar/gki464 (2005).

28 Eyal, E., Yang, L. W. & Bahar, I. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **22**, 2619-2627, doi:btl448 [pii]10.1093/bioinformatics/btl448 (2006).

29 Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-293 (1995).

30 Goddard, T. D. & Kneller, D. G. (University of California, San Francisco).

31 Lipari, G. & Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity. *Journal of the American Chemical Society* **104**, 4546-4559, doi:Doi 10.1021/Ja00381a009 (1982).

32 Bieri, M. & Gooley, P. R. Automated NMR relaxation dispersion data analysis using NESSY. *BMC Bioinformatics* **12**, 421, doi:1471-2105-12-421 [pii]10.1186/1471-2105-12-421 (2011).

33 Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **44**, 150-165, doi:10.1002/prot.1081 [pii] (2001).

34 Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Science* **6**, 1333-1337 (1997).

35 Li, T. *et al.* Redistribution of flexibility in stabilizing antibody fragment mutants follows Le Chatelier's principle. *PLoS One* **9**, e92870, doi:10.1371/journal.pone.0092870PONE-D-13-44266 [pii] (2014).

36 Jacobs, D. J. & Dallakyan, S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J* **88**, 903-915, doi:DOI 10.1529/biophysj.104.048496 (2005).

37 Verma, D., Jacobs, D. J. & Livesay, D. R. Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged. *PLoS Comput Biol* **8**, e1002409, doi:10.1371/journal.pcbi.1002409 PCOMPBIOL-D-11-01030 [pii] (2012).

38 Mottonen, J. M., Xu, M., Jacobs, D. J. & Livesay, D. R. Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family. *Proteins* **75**, 610-627, doi:10.1002/prot.22273 (2009).

39 Jacobs, D. J. & Dallakyan, S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys J* **88**, 903-915, doi:S0006-3495(05)73163-5 [pii] 10.1529/biophysj.104.048496 (2005).

40 Mottonen, J. M., Jacobs, D. J. & Livesay, D. R. Allosteric response is both conserved and variable across three CheY orthologs. *Biophys J* **99**, 2245-2254, doi:S0006-3495(10)00915-X [pii] 10.1016/j.bpj.2010.07.043 (2010).

41 Jacobs, D. J., Livesay, D. R., Hules, J. & Tasayco, M. L. Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model. *J Mol Biol* **358**, 882-904, doi:S0022-2836(06)00184-7 [pii] 10.1016/j.jmb.2006.02.015 (2006).

42 Li, T. *et al.* Thermodynamic Stability and Flexibility Characteristics of Antibody Fragment Complexes. *Protein Pept Lett*, doi:PPL-EPUB-53865 [pii] (2013).

43 Livesay, D. R. & Jacobs, D. J. Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* **62**, 130-143, doi:10.1002/prot.20745 (2006).

44 Fernandez, E. J. & Lolis, E. Structure, function, and inhibition of chemokines. *Annu Rev Pharmacol Toxicol* **42**, 469-499, doi:10.1146/annurev.pharmtox.42.091901.115838 42/1/469 [pii] (2002).

45 Yan, Z. *et al.* Structural requirements of platelet chemokines for neutrophil activation. *Blood* **84**, 2329-2339 (1994).

46 Yang, L. W. & Bahar, I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* **13**, 893-904, doi:S0969-2126(05)00167-X [pii] 10.1016/j.str.2005.03.015 (2005).

47 Crump, M. P. *et al.* Backbone dynamics of the human CC chemokine eotaxin: fast motions, slow motions, and implications for receptor binding. *Protein Sci* **8**, 2041-2054, doi:10.1110/ps.8.10.2041 (1999).

48 Eigenbrot, C., Lowman, H. B., Chee, L. & Artis, D. R. 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, doi:10.1002/(SICI)1097-0134(199704)27:4<556::AID-PROT8>3.0.CO;2-8 [pii] (1997).

49 Rajarathnam, K., Sykes, B. D., Dewald, B., Baggiolini, M. & Clark-Lewis, I. Disulfide bridges in interleukin-8 probed using non-natural disulfide analogues: dissociation of roles in structure from function. *Biochemistry* **38**, 7653-7658, doi:10.1021/bi990033v bi990033v [pii] (1999).

50 Kwakman, P. H. *et al.* Native thrombocidin-1 and unfolded thrombocidin-1 exert antimicrobial activity via distinct structural elements. *J Biol Chem* **286**, 43506-43514, doi:M111.248641 [pii] 10.1074/jbc.M111.248641 (2011).

51 Nguyen, L. T., Chan, D. I., Boszhard, L., Zaat, S. A. & Vogel, H. J. Structure-function studies of chemokine-derived carboxy-terminal antimicrobial peptides. *Biochim Biophys Acta* **1798**, 1062-1072, doi:S0005-2736(09)00417-9 [pii] 10.1016/j.bbamem.2009.11.021 (2010).

52 Carlson, J., Baxter, S. A., Dreau, D. & Nesmelova, I. V. The heterodimerization of platelet-derived chemokines. *Biochim Biophys Acta* **1834**, 158-168, doi:S1570-9639(12)00214-2 [pii] 10.1016/j.bbapap.2012.09.010 (2013).

53 Joseph, P. R. *et al.* 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, doi:M110.146555 [pii] 10.1074/jbc.M110.146555 (2010).

54 Prado, G. N. *et al.* Chemokine signaling specificity: essential role for the N-terminal domain of chemokine receptors. *Biochemistry* **46**, 8961-8968, doi:10.1021/bi7004043 (2007).

55 Rajagopalan, L. & Rajarathnam, K. Ligand selectivity and affinity of chemokine receptor CXCR1. Role of N-terminal domain. *J Biol Chem* **279**, 30000-30008, doi:10.1074/jbc.M313883200 M313883200 [pii] (2004).

56 Rajagopalan, L., Chin, C. C. & Rajarathnam, K. Role of intramolecular disulfides in stability and structure of a noncovalent homodimer. *Biophys J* **93**, 2129-2134, doi:S0006-3495(07)71467-4 [pii] 10.1529/biophysj.107.108761 (2007).