RICE UNIVERSITY

Design and Structural Characterization of Self-Assembling Triple Helical Heterotrimers

by

Jorge A. Fallas

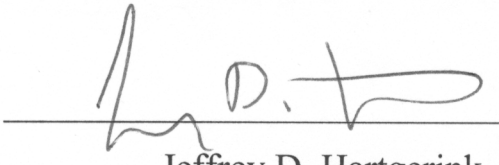
A THESIS SUBMITTED

IN PARTIAL FULFILLMENT OF THE

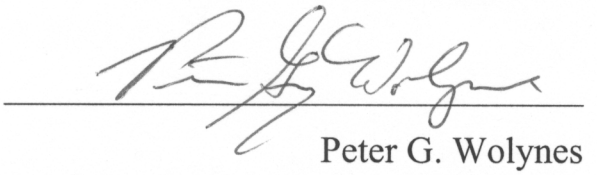
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

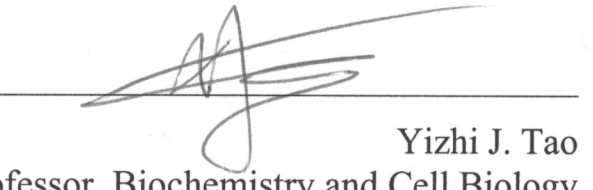
APPROVED, THESIS COMMITTEE:



Jeffrey D. Hartgerink, Chair
Associate Professor, Chemistry



Peter G. Wolynes
Bullard-Welch Professor of Science, Chemistry



Yizhi J. Tao
Associate Professor, Biochemistry and Cell Biology

HOUSTON, TEXAS

MAY 2012

ABSTRACT

Design and Structural Characterization of Self-Assembling Triple Helical Heterotrimers

by

Jorge A. Fallas

The design of self-assembling ABC-type collagen mimetic heterotrimers is challenging due to the number of distinct species that can form in a ternary mixture of peptides with a high propensity to fold into triple helices. Given the required one amino acid stagger between adjacent peptide strands in this fold, a ternary mixture of peptides can form 27 triple helices with unique composition or register. Using a combination of X-ray crystallography and NMR spectroscopy, we carry out a detailed study on pair-wise interactions between positively and negatively charged amino acids in triple helices. We find important differences in the side chain conformation of amino acids in the crystalline and solution state. Two types of contacts with distinct sequence requirements depending on the relative stagger of the interacting chains are observed: axial and lateral. We demonstrate that axial interactions can be used to bias the population towards a desired target through an in-depth structural characterization of a previously designed ABC heterotrimer from the Hartgerink laboratory. Despite the formation of the desired targer the NMR analysis shows that homotrimeric helices are also present in the peptide mixture. We modified the previous design protocol by including residues at positions that allow for axial contacts between the first and third chain of the desired target state while destabilizing the competing homotrimeric state. These changes lead to a system in which only ABC heterotrimers fold but the presence of two registers of the desired

composition is confirmed by NMR experiments. Finally, we use a computational strategy to accomplish the formation of a single-register ABC heterotrimer. By maximizing the energy gap between the desired target state and the next most stable competing state while minimizing the stability of the target state using a simple sequence-based stability function, we are able to select sequences that selectively fold into one of the possible 27 states. This approach successfully yields sequences that fold into single-register triple helices and they are characterized using CD and NMR spectroscopy.

Acknowledgments

I would like to thank my advisor, Dr. Jeffrey Hartgerink, for his guidance during the course of my time in his laboratory. He took a chance on me as I was trying to find a new home at Rice and without his support I would not have been able to make it this far.

I would also like to thank my committee, Dr. Peter Wolynes and Dr. Yizhi Tao, for their time and input on my work. I feel really honored to have such brilliant scientists review my research. Additionally, I would like to thank Dr. Tao for allowing me to collaborate with her laboratory and learn from her expertise.

A special acknowledgement is due to my undergraduate advisor, Dr. Leticia Gonzalez, for her mentorship and her help in finding a graduate program. I would also like to acknowledge the help of all the SEA members, whose support was invaluable during my time at Rice. In particular, Dr. Sean Moran for all of his help during the early stages of my research.

I would also like to thank my extended Houston family: the Valhalla crowd, the Rice Chemistry department and everyone else that had a positive impact on my experience in the past five years.

A very special thanks to all of the past and present members of the Hartgerink laboratory; you contributed towards a great work environment and for that I am grateful. Specifically, I would like thank the people that went far beyond my academic experience and became some of my most trusted and beloved friends: Lesley, Erica and Marci. Thank you for everything, you really helped to put smile on my face when nothing else seemed to be going right and made this a wonderful experience.

To my Houston friends: Soph, Al, Eoghan, Amanda, Denise, Jason, Brad, Paul and Alan. I could not have asked for a better group of people to share this experience with. Your support both personally and academically during this difficult endeavor was invaluable. I will always treasure your friendship.

To my friends from Costa Rica: Glenn, Wins, Gary, Steph, Ana, Dani, Pava and Alfredo and to my german friends: Maciek, Franzi, Magda, Henrike, Jan and Basti. You guys have known me the longest and believed in me from day one. Thank you for all support and friendship.

Finally, I would like dedicate this thesis to my family: Mami, Papi, David, Giny and Cristian. You mean the world to me. Your love has helped make it through this difficult time and I hope that I have made you proud. I would like to specially thank my parents for their example in how to conduct myself as scientist, as a friend and as a partner. You believed in my dreams and allowed me to follow them, even if that meant for us to be apart. You also gave me a strong set of values to follow and for that I will always be grateful. Last, I would like to thank Cristian for being here every step of the way. I know it was not easy but you deserve much credit because without your love and care none of this would have been possible. Thank you Love.

Table of Contents

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 5

Chapter 6

LIST OF TABLES

ABBREVIATIONS

CD	circular dichroism spectroscopy
CMP	collagen mimetic peptide
ECM	extracellular matrix
ESI-TOF	electrospray ionization time of flight mass spectrometry
FACIT	fibril associated containing interrupted triple helices
Fmoc	9-fluoronylmethoxycarbonyl
GA	genetic algorithm
HBTU	O-benzotriazole N,N,N',N'-tetramethyluronium hexafluorophosphate
HATU	O-(7-Azabenzotriazol-1-yl)-N,N,N',N'-tetramethyluronium hexafluorophosphate
HNHA	3D-correlation NMR experiment: $\{^{1}H_{amide}\}$-$\{^{15}N_{amide}\}$-$\{^{1}H_{\alpha}\}$
HNHB	3D-correlation NMR experiment: $\{^{1}H_{amide}\}$-$\{^{15}N_{amide}\}$-$\{^{1}H_{\beta}\}$
HPLC	high pressure liquid chromatography
HSQC	heteronuclear single quantum coherence
MALDI-TOF	matrix desorption assisted time of flight mass spectrometry
MMP	matrix-metalloprotease
MRE	molar residual ellipticity
NMR	nuclear magnetic resonance spectroscopy
NOESY	nuclear Overhauser effect spectroscopy
TFA	trifulouroacetic acid
SPPS	solid phase peptide synthesis
TOCSY	total correlated spectroscopy

Preface

Supramolecular chemistry is a new discipline that aims to utilize non-covalent, cooperative interactions to build well-defined complexes using molecular building blocks. One of the most important ideas behind the advances in this field is the encoding and propagation of information through complementary chemical interactions. Concepts like preorganiztion, where synthetic restrictions are placed on the designed molecules to adopt conformations that facilitate complex formation and self-assembly, where molecular recognitions motifs are utilized to induce the formation of particular macromolecular structures from smaller, non-covalently bound protomers are two of the most important examples of information transfer through rationally designed chemical interaction potentials. The building blocks in supramolecular chemistry have diversified from synthetic macrocycles to biologically inspired molecules such as peptides and lipids which has led to a tremendous growth in the field and the preparation of novel self-assembling structures and materials.

Despite of all the advances in the synthetic realm of supramolecular chemistry our level of control over self-assembling systems is crude compared to that attained by the molecular machinery found in cells. Living organisms also utilize non-covalent interactions to encode and propagate information but through millions of years of evolution they have achieved an unparalleled level of complexity and precision. Proteins are exceedingly interesting from a chemical perspective because naturally occurring amino acids contain different moieties that can engage in a variety of non-covalent interactions. Thus, a protein's amino acid sequence, which is naturally selected due to evolutionary pressures, contains information in the form of complementary, non-covalent

interactions that allows it to reliably adopt intricate three-dimensional structures or selectively bind particular partners. The underlying principles that direct the behavior observed in proteins have been thoroughly studied from a theoretical physics perspective. The advances in this field have lead to important hypothesis about the inherent shape of the process' energy landscape as well as the emergence of practical concepts like the stability gap requited between a protein's native three-dimensional structure and other plausible conformations, which can utilized in the laboratory to design new amino acid sequences and proteins with novel folds or functions.

During the course of my doctoral research I have been exposed to these novel concepts working in a laboratory that utilizes short peptidic sequences and ideas from supramolecular chemistry to build functional biomaterials. Particularly, I have been working with the most abundant protein folds in the human body: the collagen triple helix. The main goal of my work in the Hartgerink laboratory has been to design, synthesize and characterize triple helical peptides that are able to self-assemble into well defined heterotrimeric states while avoiding the formation of undesired competitors and this thesis describes the different steps that were required to achieve this goal.

Chapter 1: Introduction*

1.1 The Collagen Protein Family and Collagen Mimetic Peptides

The collagen protein family encompasses some of the most abundant proteins in the human body.[1] They are large multi-domain proteins found in the extracellular matrix (ECM) and are known for providing structural stability to a variety of tissues. There are 28 known isoforms of collagen in the human species,[2] found in a wide variety of tissues such as cartilage, bones and skin[3] and are arranged in carefully crafted supramolecular structures that range from planar sheet-like networks to fibrils and fibers in order to fulfill specialized functions.[4] Collagens are usually classified into five groups, according to their supramolecular morphology, as 1) fibril forming, 2) fibril associated containing interrupted triple helices (FACIT), 3) beaded filament, 4) anchoring fibril, and 5) network forming and transmembrane collagens.[5] Despite the vastly different architectures that they adopt, all proteins in this family share a common structural domain known as the collagen triple helix. The triple helix is formed by polypeptide strands, known as alpha chains, that adopt a left-handed type II polyproline helix conformation and wind around each other to form a tightly packed right-handed super-helix. Collagens can be either homotrimers, with identical sequences for all alpha chains (AAA) or heterotrimers that can include either two distinct sequences (AAB) or three distinct sequences (ABC) in their triple helices. From a biochemical perspective, collagens are interesting because they participate in cell-ECM interactions through specialized receptors with a high

---

* This chapter is largely based on the following publication:
Fallas, J. A.; O'Leary, L. E.; Hartgerink, J. D. *Chem. Soc. Rev.* **2010**, *39*, 3510-3527.

affinity to proteins in a triple helical conformation.[6,7] Furthermore, collagen turnover, a process required for ECM remodeling (and associated with both homeostatic processes such as angiogenesis[8] and pathologic conditions like arthritis, periodontal disease and cancer metastasis)[9] is mediated by a specialized set of matrix-metalloprotease (MMPs) that require the triple-helical conformation to be effective.[10] From a supramolecular chemistry perspective, the collagen protein family is an interesting subject because the proteins in it are able to adopt a wide variety of complex and multi-hierarchical supramolecular structures but the mechanism by which these are achieved is not well understood.

Despite their importance it is very difficult to study the structure, assembly and biochemistry of natural collagens because of the inherent heterogeneity of their native environment[11] and the size of collagenous proteins, which ranges from a few hundred to over one thousand amino acids per alpha chain. To overcome these difficulties, smaller synthetic peptides that adopt a triple helical fold have been used for decades to study the molecular structure,[12] stability[13] and biochemistry[14] of the collagen triple helix as well its further self-assembly into higher-order structures.[15] These smaller systems, usually referred to as Collagen Mimetic Peptides (CMPs), are better suited for high-resolution structural techniques such as X-ray crystallography[16-19] and nuclear magnetic resonance (NMR) spectroscopy[20-23] and have led to major advances such as the first high-resolution triple helical structure[24] and the interaction mechanism of collagenous proteins with cell surface receptors such as the I-domain of the α2β1 integrin.[25] Despite having gained considerable knowledge on the structure and stabilization of the triple helical folding motif and how it interacts with cell-surface receptors using short model peptides there is

still much left unknown. Significantly, much published work uses homotrimeric collagen mimics, which are good models for some collagen types like fibril forming collagen type II and III found in cartilage and skin respectively. However, very little work is available on heterotrimeric triple helices primarily because, until recently, there was no straightforward method for their assembly. Some of the most abundant proteins within the collagen family are heterotrimers, like type I found in bones and teeth (AAB heterotrimer) and type IV, the major component of basement membranes (both AAB and ABC varieties exist), making heterotrimer research very promising to gain a deeper understanding of these important proteins.

1.1a Synthesis and Characterization of Collagen Mimetic Peptides

The design of CMPs is based on the sequence of natural collagens, following its three amino acid repeat of the form X-Y-Gly, where the X position is predominantly occupied by proline and the Y position is most commonly 4R-hydroxyproline, a post-translationally modified amino acid with a hydroxyl group on the γ-carbon of the proline side chain (single letter code O and the three letter code Hyp). Because 4R-hydroxyproline is not encoded in the standard genome and the systems available for protein biosynthesis, until recently, lacked prolyl hydroxylase, the enzyme responsible for the transformation, CMPs have been mostly produced by chemical synthesis. The first CMPs were synthesized with the advent of the solid phase peptide synthesis methodology[26] (SPPS) in the 1960's. Initial efforts showed that synthetic peptides with the sequence $(PPG)_{10}$ behaved like natural collagen in terms of the temperature dependence of their optical rotation properties (ORD).[27,28] Unlike natural collagens, the

synthetic peptides were able to produce single crystals[29] that yielded diffraction patterns[30] consistent with the fiber diffraction patterns obtained for natural collagens,[31,32] corroborating their similarities at the molecular level and validating their use as mimics of collagenous proteins.

CMPs are usually synthesized using standard *N*-(9-fluorenyl)methoxycarbonyl- (Fmoc) based SPPS, including the use of benzotriazole coupling reagents and piperidine for Fmoc deprotection. Some modifications are required to maximize the yield of difficult couplings, particularly the sequential coupling of imino acids. This can be addressed by coupling triplets instead of amino acids[33] or increasing the coupling time, utilizing a mixture of diaza(1,3)bicyclo[5.4.0]undecane and piperidine during the deprotection.[34] An alternative approach is the implementation of double couplings for imino acids in the X position when the Y position is occupied by proline or a proline derivative. For particularly difficult or expensive sequences, like those including isotopically enriched amino acids, manual couplings using low loading resins and more reactive coupling reagents like HATU tend to improve the yield and facilitate purification. Recent advances in microwave assisted SPPS have also been implemented for CMP synthesis.[35]

After cleavage, peptides can be purified by reverse phase high-pressure liquid chromatography (HPLC) using C18 or biphenyl columns with a water-acetonitrile gradient.[36] The pure peptides are usually characterized by matrix adsorption assisted– or electrospray ionization- time of flight (MALDI-TOF, ESI-TOF) mass spectrometry. An appropriate matrix for MALDI-TOF measurements in the mass range of most CMPs is α–cyano-4-hydroxycinnamic.

The association of the peptides is usually characterized by circular dichroism (CD) spectroscopy. The signature spectrum for triple helical assemblies includes a maximum around 225 nm and a minimum around 190 nm. The exact position of the extrema is sequence dependent and may vary by a few nanometers when comparing different assemblies. Other common methods used to study the supramolecular assembly of CMPs are solution NMR and X-ray crystallography. CMPs up to ten triplets can be readily characterized using homonuclear experiments but it is difficult to obtain correlation spectra for longer peptides due to their anisotropic tumbling in solution and thus longer peptides require the use of ^{15}N- and/or ^{13}C- labeled amino acids. For X-ray crystallography studies CMPs have been notoriously difficult to crystallize but recent developments including the use of polyethylene glycol as a co-precipitant have been successful in producing well-ordered crystals.

1.1b Triple Helical Structure

Early attempts to determine the molecular structure of collagen were based on fiber diffraction data of stretched kangaroo[37] and rat-tail[38] tendon collagen fibers. Ramachandran used the diffraction patterns to estimate a fiber period of 28.6 Å[39] and used it to put forth the first triple helical model of collagen.[40] The proposed model had the correct basic features, involving three left-handed chains super-coiled around each other but it overestimated the number of hydrogen bonds present per tripeptide unit. This structure was later revised by Rich and Crick[41] on the basis of their findings for the poly-glycine helix[42] to the structure that is currently accepted.[43] Simultaneously, Cowan et. al.[38] proposed a very similar model in the context of their work on the poly-proline type

II helix.[44] The structure proposed by Rich and Crick and Cowan et al. was later found to be very similar to the structure of the synthetic $(PPG)_{10}$ peptide, which was refined to a low resolution electron density map based on single crystal diffraction data.[24] The first high-resolution structure (1.9 Å) of a collagen triple helix was not available until the early 1990's when Brodsky and co-workers[45] solved the structure of a triple helical peptide, validating the Rich and Crick model (Figure 1.1a).

The primary structure of the triple helix contains a three amino acid repeat (X-Y-Gly), requiring glycine as every third amino acid. Each of the component chains forms a polyproline type II left handed helix that associates with two other chains to form a tightly packed right handed superhelix stabilized through a hydrogen-bonding network and van der Waals interactions. The hydrogen-bonding network goes along the backbone of the polypeptide chains, perpendicular to the helical axis, between the carbonyl group of the amino acid in the X position in one chain and the amide nitrogen of glycine in a different chain (Figure 1b). To allow for tight packing of the chains and to maximize the contact area for van der Waals interactions, glycine, the third amino acid in every triplet, is always facing towards the core of the helix (Figure 1c). This forces the chains to assemble in a staggered manner, offset by a single amino acid, so that there is always a glycine residue in every cross-section of the helix taken perpendicular to the helical axis.

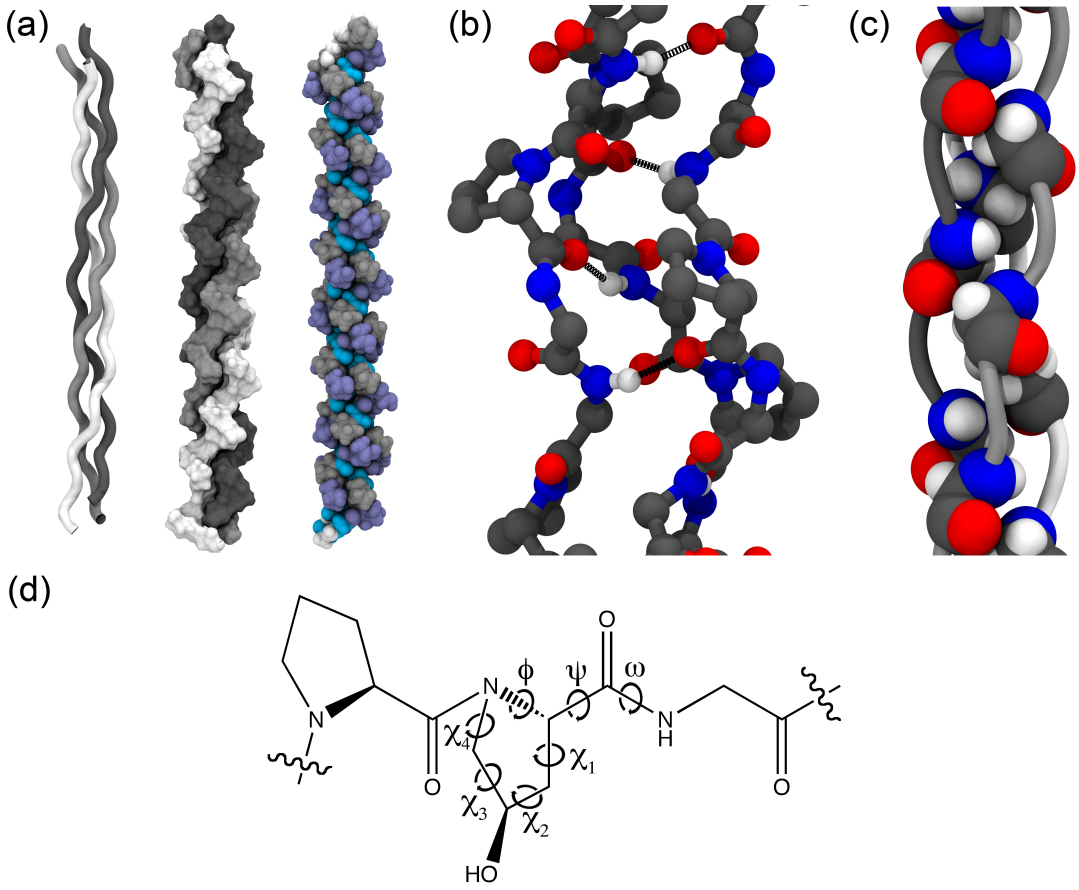


Figure 1.1. Triple helical structure (a) Cartoon, surface and space filling models of the $(POG)_{10}$ structure (gray = P, purple = O, cyan = G) (b) Hydrogen bonding network, highlighted by the dotted lines (c) Glycine interactions at the core of the triple helix. The chains are colored as follows in (a) and (c): leading strand-dark gray, middle strand -light gray, trailing strand-white. (d) Schematic of the prototypical POG sequence highlighting dihedral angles. All structural models are based on the $(PPG)_{10}$ crystal structure (PDB 1K6F[46]) , modified to include hydroxyproline following the protocol described in Fallas et al.[47]

The predominance of imino acids in the X and Y positions of collagenous sequences can also be understood in terms of its structural consequences. The pyrrolidine ring in proline and its derivatives can adopt two different conformations (endo or exo) that fix the χ1 and φ dihedrals. The χ1 dihedral is the torsion angle that describes the rotation along the C_α-C_β bond in the amino acid side chains while the φ dihedral describes the rotation around the N-C_α bond in the peptide backbone. For the endo

conformer the dihedral values are 19° and -75°, for χ1 and φ respectively and the exo conformer has a χ1 value of -6° and a φ value of -68°.[20] The φ value of the endo conformer coincides with the value observed for residues in the X position of the triple helix and conversely, the φ value of the endo conformer agrees with the φ dihedral for the Y position in the triple helix. It is not surprising, then, to find that the pyrolidine ring in proline, the most common amino acid in the X position of collagens, has a slight preference to adopt the endo conformation or that 4R-hydroxyproline, the most common amino acid found in the Y-position in collagens, adopts preferentially the exo conformation.

Despite sharing the same structural elements, there is a significant historical difference between the triple helical models derived from native collagen and CMPs: the helical symmetry. The Rich and Crick model proposed on the basis of fiber diffraction data is a 10/3 helix, a helix with 10 scattering units (30 amino acids) and 3 helical turns in each axial repeat,[32] with an axial repeat (or fiber period) of 28.6 Å. The single-crystal diffraction structures derived from CMPs have an axial repeat of 20 Å with 7/2 helical symmetry, containing 7 scattering units (21 amino acids) and 2 helical turns in each axial repeat.[32] Both helices have the same unit height, or translation along the helical axis per tripeptide unit, at 2.86 Å,[48] but have different unit twists with the 7/2 model being more tightly wound than its 10/3 counterpart (Figure 1.2). One difficulty with this difference is that the fiber diffraction data does not contain enough reflections to determine the axial repeat unambiguously and can be explained with either model,[49] making both structures suitable candidates for the correct structure of native collagen even if they have not always been considered on equal footing in the literature.

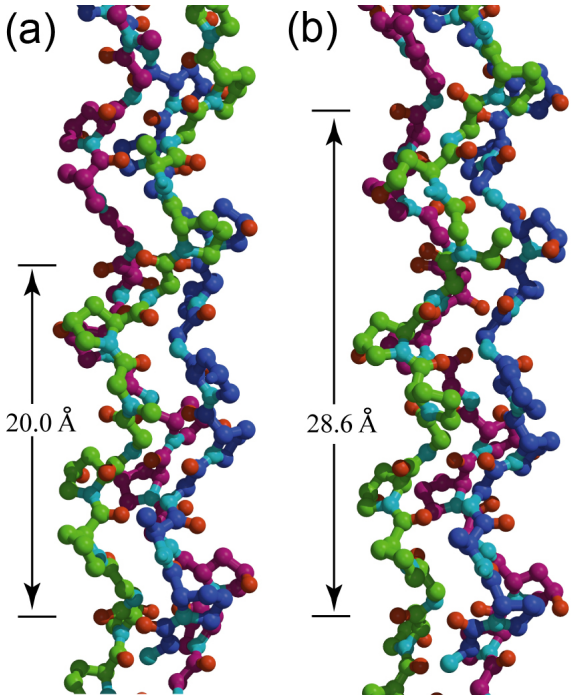


Figure 1.2. Triple helical symmetry. (a) 7/2 helix (b) 10/3 helix. Figure reproduced from Okuyama *et. al.*[49]

Some light was shed on this question with the solution of high-resolution structures of CMPs that included sequences taken from natural collagens. Such peptides require N- and C-terminal flanking regions to drive the formation of stable triple helical assemblies and may include any amino sequence taken from a natural collagen in the middle. An example is the T3-785 peptide[50] that contains amino acids 785-796 from human collagen type III, with the sequence ITGARGLAG and is flanked by three repeats of the POG sequence at the N- and C-termini. An analysis of the helical symmetry of this peptide demonstrated that the guest region, which has a very low content of imino acids, agrees better with the looser 10/3 model than the 7/2 model.[51] Conversely, the imino-acid rich flanking regions show a 7/2 helical symmetry, similar to the imino acid rich peptides that were used for previous crystallographic studies.[16,17,46] Since then, other crystal structures that include guest sequences from natural collagens and show a similar behavior have been solved.[52,53] Thus, the experimental evidence points towards both structures being

correct and the helical symmetry in the collagen triple helix varying along the main axis depending on the imino acid content of a particular stretch

The conformation of CMPs in solution has been studied by multi-dimensional NMR experiments. Using the prototypical POG sequence, Brodsky et al.[20] found two sets of spin systems for each residue type with 80% of the triplets being in a similar chemical environment and thus having identical chemical shifts. Furthermore, using nuclear Overhauser effect spectroscopy (NOESY) they were able to analyze the topology of the molecule. Comparing the NOE crosspeaks measured for these triplets with the ones expected based on the fiber diffraction model the authors concluded that the solution conformation of the peptide is similar to that present in the fibers. The remaining 20% of the triplets were assigned to a less ordered form of the helix present on the termini of the molecule.

1.1c Triple Helical Stability

In the same way that CMPs have been integral in uncovering key features of the triple helical structure, they have also helped reveal non-covalent interactions that stabilize this structure. The central feature of CMPs used to compare their stabilities is the unfolding temperature, measured through CD, ORD melting studies or calorimetry experiments. As previously mentioned, proline and hydroxyproline are the two most abundant amino acids in the X- and Y-positions of collagenous proteins, respectively. The importance of hydroxylation, specifically hydroxyproline as opposed to hydroxylysine, to triple helical stability was initially reported over 25 years ago in thermal melting studies performed on digested collagen type I.[54] The necessity for O to be in the Y-position within the collagen triplet as opposed to the X position was proven

by the inability of the CMP $(OPG)_{10}$ to fold despite the fact that both $(PPG)_{10}$ and $(POG)_{10}$ form stable trimers.[55] There are two leading arguments to explain the increased thermal stability provided by the presence of hydroxyproline in the Y-position of the collagen triplet: stereoelectronic effects and water mediated hydrogen bonding. Schematic representations of the two theories are shown in Figure 1.3.

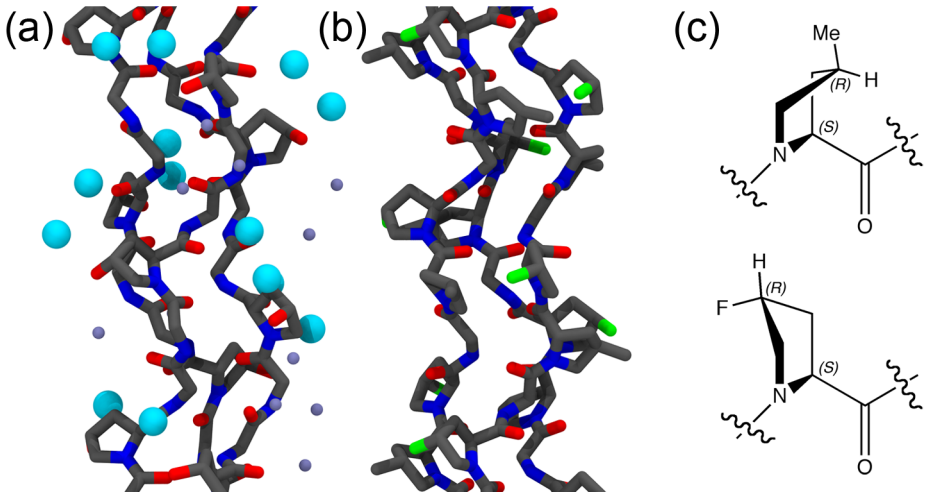


Figure 1.3. Triple helical stabilization mechanisms. (a) Hydration network of a triple helical peptide containing hydroxyproline (pdb 1CGD). Water molecules involved in inter-strand bridges are represented as large cyan spheres and water molecules involved in intra-strand water bridges are represented by small purple spheres. (b) Structure of a triple helical peptide containing 4R-methylproline and 4R-fluoroproline (pdb 3IPN). (c) The endo ring pucker can be observed in 4R-methylproline residues and the exo ring pucker in 4R-fluoroproline residues (fluorine is depicted green).

In 1994, Brodsky reported the first high resolution x-ray crystal structure of a derivative of $(POG)_{10}$ containing a single glycine to alanine mutation.[45] Along with several key observations on the puckering of the proline and hydroxyproline residues, endo and exo respectively, the most significant observation from the high resolution structure was a well defined hydration shell in which hydroxyproline bonds with water instead of engaging in direct contacts with other amino acids. Based on this observation it was suggested that the hydroxyl group present in hydroxyproline allows for water

mediated hydrogen bonds thus stabilizing the CMP triple helix. In 2001, the x-ray crystal structure of a CMP containing a sequence from collagen type III showed the presence of water-mediated hydrogen bonds that connect the N-H of an amino acid in position X to the carbonyl oxygen of the glycine residue.[50] Thus through a series of X-ray crystal structures on CMPs, the hydration shell surrounding this protein fold was elucidated and more importantly, the fact that the structure is enhanced by water-mediated interactions and the ability of hydroxyproline to form bonds with water, not just other amino acids.[56]

Raines has argued that the stabilizing effect of hydroxyproline is largely a stereoelectronic effect.[57-60] This argument is based on a series of papers studying proline derivatives with electronegative substituents in the γ-carbon such as 4R-hydroxyproline and 4R-fluoroproline (Flp) that compare the thermal stability of different triple helices: $(PPG)_{10}$, $(POG)_{10}$ and $(Pro\text{-}Flp\text{-}Gly)_{10}$. The stability of the CMPs was highest in $(Pro\text{-}Flp\text{-}Gly)_{10}$ with a melting temperature of 91 °C and lowest in $(PPG)_{10}$ which melted at 41 °C, with $(POG)_{10}$ in-between the two with a melting temperature of 69 °C.[59,60] As mentioned in section 1.2b the pucker of the pyrolidine ring in imino acids controls the φ backbone dihedral and can adopt two different conformations, one that suits the requirements of the X position and one that is more apt for the Y position of the triple helix. Proline shows a slight tendency to adopt the endo conformer with a φ dihedral corresponding to the X position and this tendency can be reinforced by either adding a bulky substituent, such as a methyl group, to the γ-carbon in the R configuration or an electronegative group, like fluorine or a hydroxyl group, in the S configuration (Figure 1.3c). On the other hand electronwithdrawing groups, like the hydroxyl and fluoro groups in 4R-hydroxyproline and 4R-fluoroproline with R configuration favor an exo ring pucker (Figure 1.3c) that

preorganizes the φ value towards the dihedral observed for the Y position of the triple helix. Overall, by choosing the right stereochemistry in the proline derivatives to drive the endo and exo conformations in the X and Y positions respectively, one can pre-organize the backbone dihedrals and reduce the entropic penalty for triple helical formation leading to a stabilization of the folded state. While the merits of hydrogen bonding versus stereoelectronics have been argued as competing theories for the stabilization of the collagen triple helix, both likely play a role with stereoelectronics as the main component contributing to the higher stability seen in hydroxyproline containing systems.

While examining the hydrogen bonding network surrounding hydroxyproline in CMPs, Brodsky et. al. explored how single amino acid mutations affect the stability of triple helical assemblies. The melting temperature of a library of peptides containing all possible single point mutations in the 4th triplet of a $(POG)_8$ template were recorded.[61,62] The most important finding from this work is that mutations from P or O to any other naturally occurring amino acid in a triple helical sequence will cause a decrease in the melting temperature of the resulting assembly. Other observations from this work include the correlation between stabilities observed and the occurrence in fibril-forming collagens of the residue in the specific position of the collagen triplet and the predominance for negatively and positively charged amino acids to have higher melting temperatures in the X and Y positions respectively. The X position is known to be more solvent exposed[63,64] and thermodynamic calculations performed on the 1996 melting data of hydrophobic residues showed that the preference for the X position is entropically driven.[62] However, the driving force for the prevalence of negatively charged residues in this position was

initially unclear. In order to explore this question, CMPs containing double mutations in a $(POG)_9$ template were synthesized using charged and hydrophobic amino acids.[65] The experimental thermal stabilities collected from CD melting studies were compared to the predicted unfolding temperature based on the contributions from individual triplets. The most surprising result from this study was the high stability seen for the CMPs containing GPKGEO or GPKGDO. In both peptides, lysine is the positively charged amino acid in the Y position and a negatively charged residue, glutamic acid or aspartic acid, is in the X position. Computer modeling on the GPKGEO peptide, shown in Figure 1.4, revealed the close proximity of the side chains of the charged amino acids allowing for cross-chain hydrogen bonding. This work laid the framework for a large portion of the current research on synthetic triple helical peptides and will be expanded upon in the coming Chapter 2.

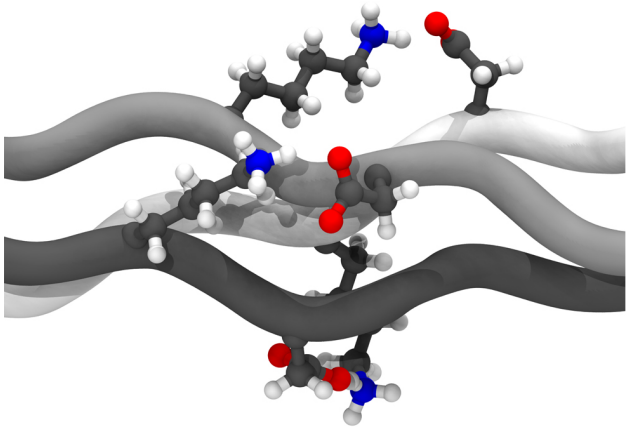


Figure 1.4. Computer models of pairwise interactions between lysine and glutamic acid made to explain the increased thermal stability of GPKGEG-containing homotrimeric CMPs.[65] The model is based on the $(PPG)_{10}$ crystal structure (PDB 1K6[46]) , modified to include charged residues and minimized following the protocol described in Fallas et al.[47]

A different approach that has been widely utilized to study sequences with low triple helical propensities is the templated assembly of triple helices. This method utilizes small organic molecules attached to the peptides to drive the trimerization of the helix (Figure 1.5). Molecules with three reactive moeities such as cis,cis-1,3,5-trimethylcyclohexane-1,3 acid (KTA)[66,67], tris(2-aminoethyl)amine (TREN)[68], tris(hydroxymethyl)aminomethyl (TRIS)[69,70] and triserine lactone (TSL)[71] have been used to maintain the peptide chains close to one another and promote their assembly into triple helices. An interesting consequence of the use of such scaffolds, particularly KTA, has been the successful incorporation of peptoid N-isobutylglycine, in the X and Y position of a triple helix[72-76]. Other related approaches include the use of disulfide bridges[77,78], lysine-lysine cross-linking[79], metal ion coordination sites[80,81] and alkyl chains[82,83] to promote triple helical nucleation. A recent review by Greg Fields goes into the details of the synthetic protocols utilized for these constructs and readers interested in this aspect are referred to this document as this beyond the scope of this thesis.[34]

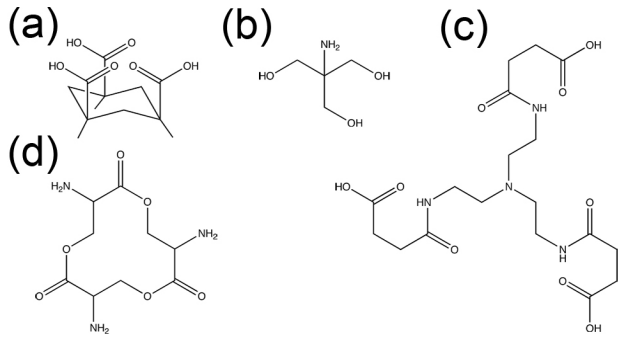


Figure 1.5. Chemical Structures of small molecules commonly used for the templated assembly of triple helical peptides. (a) KTA (b) TRIS (c) TREN (d) Triserine.

1.2 Heterotrimeric Collagen Mimics

Despite some of the most abundant collagens in nature being heterotrimeric the first heterotrimeric CMPs were not reported until the mid 90's using Lys-Lys covalent tethering,[84] about thirty years later than their homotrimeric counterparts. In the last ten years different strategies have been used to synthesize heterotrimeric CMPs to study collagen degradation, integrin binding and connective tissue diseases. The major approaches to heterotrimer synthesis and findings from these studies will be discussed in the next two sections.

1.2a Covalently-linked Heterotrimers

The most common approach for the covalent synthesis of heterotrimeric CMPs has been regioselective disulfide bridging.[85] The synthetic procedure has been described in detail elsewhere[86] and is depicted in Figure 1.6. For this approach a single cysteine residue is required in the leading and in the lagging peptide strands, as well as two cysteines in the middle strand. The general procedure involves the protection of the two cysteines in the middle chain using the tertbutylthio (StBu) and acetamidomethyl (Acm) groups, meanwhile the cysteines in the leading and lagging strands are protected with the StBu functionality. The StBu derivatives can be readily deprotected by reducing the disulfide bond using tributylphosphine (PBu_3) to yield free thiols. The free thiol in the middle strand is then activated using 2,2'-dithio-di(5-nitro)pyridine ($Npys_2$) to yield the 5-nitropyridil-2sulefenyl (Npys) cysteine derivative which can readily react with a free thiol from another strand leading to the covalent crossliking of two of the peptides. The

Acm group in the second position is subsequently deprotected using Npys-Cl to yield the Npys derivative in the second position which in turn readily reacts with the free thiol in the third strand completing the heterotrimer. This approach has been successful in synthesizing heterotrimers with POG triplets at the N-terminus as triple helical nucleating sequences followed by guest sequences from heterotrimeric collagens with the register fixed by the C-terminal cysteine knot.

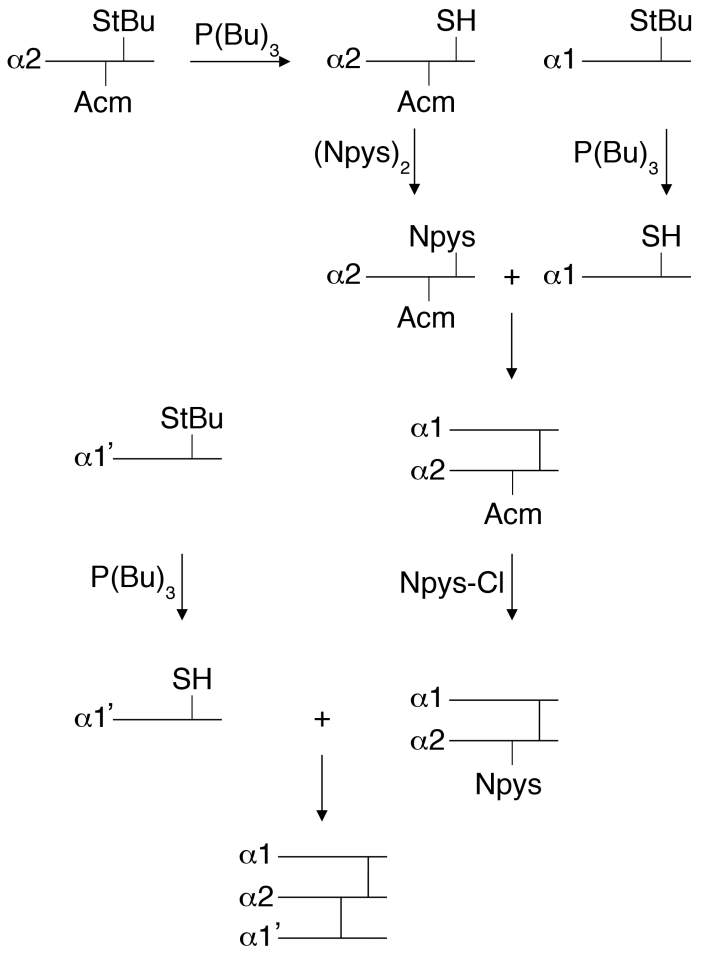


Figure 1.6. Regioselective cysteine chemistry for heterotrimeric CMP synthesis, full names of protecting groups and reagents are available in section 3.1. Adapted from Renner et al.[87]

One of the sequences studied using this approach corresponds to the matrix metalloprotein-1 (MMP-1) and MMP-8 cleavable site from type I collagen, which includes residues 772-784. Collagenases perform a highly selective and conformationally dependent scission of the peptide bonds between glycine and leucine in the α1 chain and glycine and isoleucine in the α2 chain (amino acids 775 and 776). Peptides were designed to include the type I sequence, an N-terminal nucleation domain $(POG)_n$, with values for n of 3 and 5. The peptides were fixed in the α1α2α1 register using the cysteine knot strategy described above.[88,89] All the heterotrimers fold into triple helices but their thermal stability is dependent on the length of the nucleation sequence, with melting temperatures ranging from 9 °C to 41 °C respectively, with the C-terminal domain of the longer peptide only forming a partially ordered structure as shown by NMR experiments.[90] Enzyme digestion assays performed at room temperature on the peptides showed that MMP-8 proteolysis of the partially-folded CMP is efficient but the completely unfolded trimer shows a very slow process. Conversely, MMP-2, also known as gelatinase A, cleaves the unfolded peptide preferentially.[91,92]

The integrin binding sequence of the most abundant form of collagen type IV, an AAB heterotrimer, is located between residues 457-468 and amino acids in different chains are thought to be important for binding. Peptides containing the type IV sequences flanked by three POG repeats at the N-terminus and two at the C-terminus were synthesized to study the binding of this sequence to the $\alpha_2\beta_1$ integrin.[93] Using the cysteine knot approach two different heterotrimer registers were synthesized and tested for their integrin binding affinity. The experiments showed that the CMP with the α2α1α1 register shows a slightly higher binding affinity than the α1α2α1 register.[94] Interestingly, the

α2α1α1 register has been proposed on the basis of fluorescence resonance energy transfer measurements of cyanogen-bromide derived fragments extracted from type IV collagen.[95] The thermal stability and folding rate was also found to be register dependent with the α1α2α1 register showing a higher melting temperature and higher rate constant.[96] The change in melting temperature can be partially explained by the work on self-assembled heterotrimers driven by electrostatic interactions that will be described in the next section. The sequences used to study $\alpha_2\beta_1$ integrin binding contain several lysine and aspartic acid residues, but only in the α1α2α1 do they have the correct three-dimensional arrangement to form an inter-chain salt-bridge similar to the ones that drive the formation of the heterotrimers discussed below.

1.2b Self-assembled Heterotrimers

Self-assembly of heterotrimers have two challenges that are not faced by comparable homotrimeric assemblies: control over composition and register. Heterotrimers can be either AAB or ABC depending on the number of unique peptide chains used. In both cases control over composition is required in which the desired composition is favored while the undesired composition is disfavored. For example, peptides A and B may associate to form either A_2B or AB_2 heterotrimers or either of the two homotrimers, A_3 and B_3. A mixture of A, B and C peptides is even more complicated, potentially leading to any of ten different triple helical compositions including three homotrimeric helices (A_3, B_3, and C_3), any of six possible two component heterotrimers (A_2B, AB_2, A_2C, AC_2, B_2C, and C_2B) or the desired ABC heterotrimer. But the problem is made even more complicated by the issue of register. The three peptides of the collagen triple helix are offset from one another by a single amino acid creating

leading, middle and lagging strands. In a homotrimer, it makes no difference which peptide is in which position as they all have identical sequences. For heterotrimers, however, each composition can be formed in several different registers. For example a triple helix with the composition AAB could be any of three different registers (AAB, ABA or BAA) while a triple helix with the composition ABC could be any of six different registers (ABC, ACB, BAC, BCA, CAB, or CBA). Figure 1.7 shows a scheme of all triple helical registers and compositions possible in a ternary mixture of peptides.

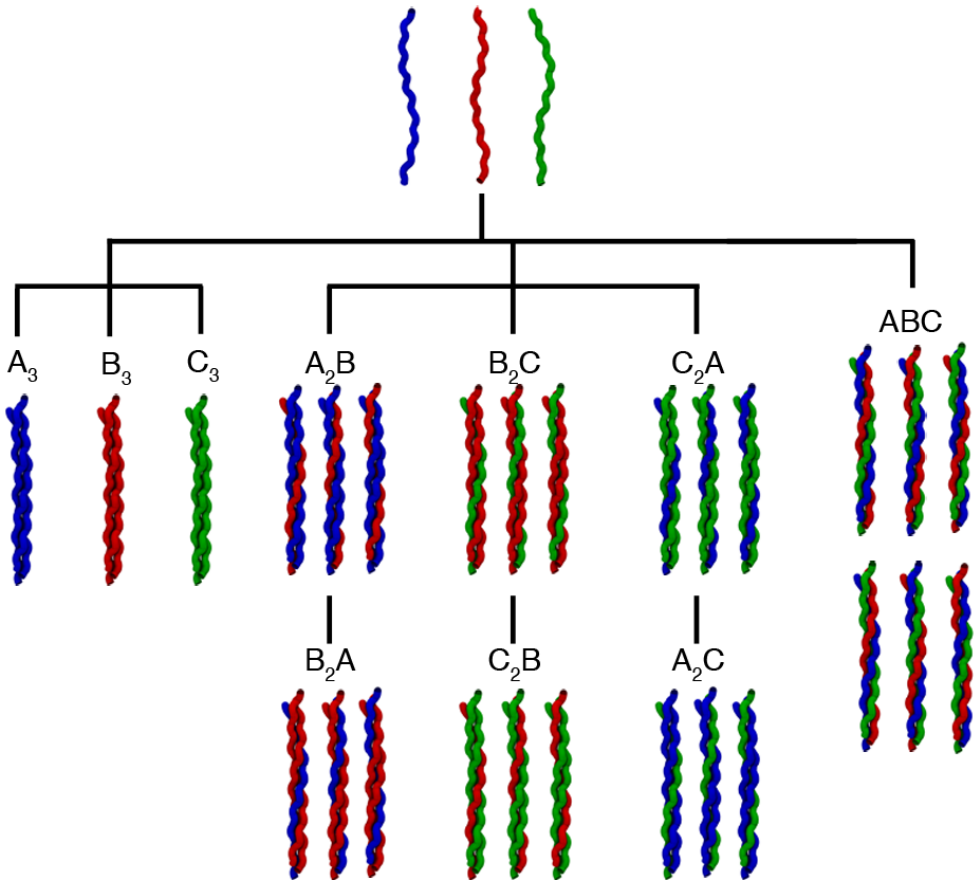


Figure 1.7. Schematic representation of the 27 different triple helices that can form in a mixture of 3 peptides with unique sequences.

Early studies performed on heterotrimeric self-assembled CMPs had some major shortcomings such as low thermal stability[97] or lack of specificity with respect to composition and register. An annealed mixture of $(POG)_{10}$ and $(PPG)_{10}$ peptides shows the formation of heterotrimers but with no control over the composition of the assemblies.[98] Brodsky synthesized a peptide system with POG flanking regions and sequences from the α1 and α2 chains of collagen type IV in the interior that forms a AAB type heterotrimer visible by CD, DSC and NMR[97]. Although this was one of the first examples of a self-assembled AAB type collagen triple helix, a significant difficulty with this design was the low thermal stability of the system, 14.5 °C. Longer peptides containing type IV guest sequences and C-terminal cysteine knots to stabilize the self-assembled heterotrimer through disulfide bridges showed a mixture of several registers when analyzed by solution NMR.[99] Raines and co-workers were able to form a self-assembled heterotrimer from $(PPG)_7$ and a peptide containing fluoroproline derivatives with the appropriate stereochemistry for the X and Y positions. Despite having a large degree of pre-organization, the fluoroproline containing peptide failed to fold into a stable triple helix due to unfavorable steric interactions arising from the fluorine atoms. The inclusion of $(PPG)_7$ is able to mitigate this effect by spacing the fluorine atoms and the two peptides form a triple helix in a 2:1 ratio.[100] While having control over the composition, nothing is known about the register of triple helices self-assembled with this method.

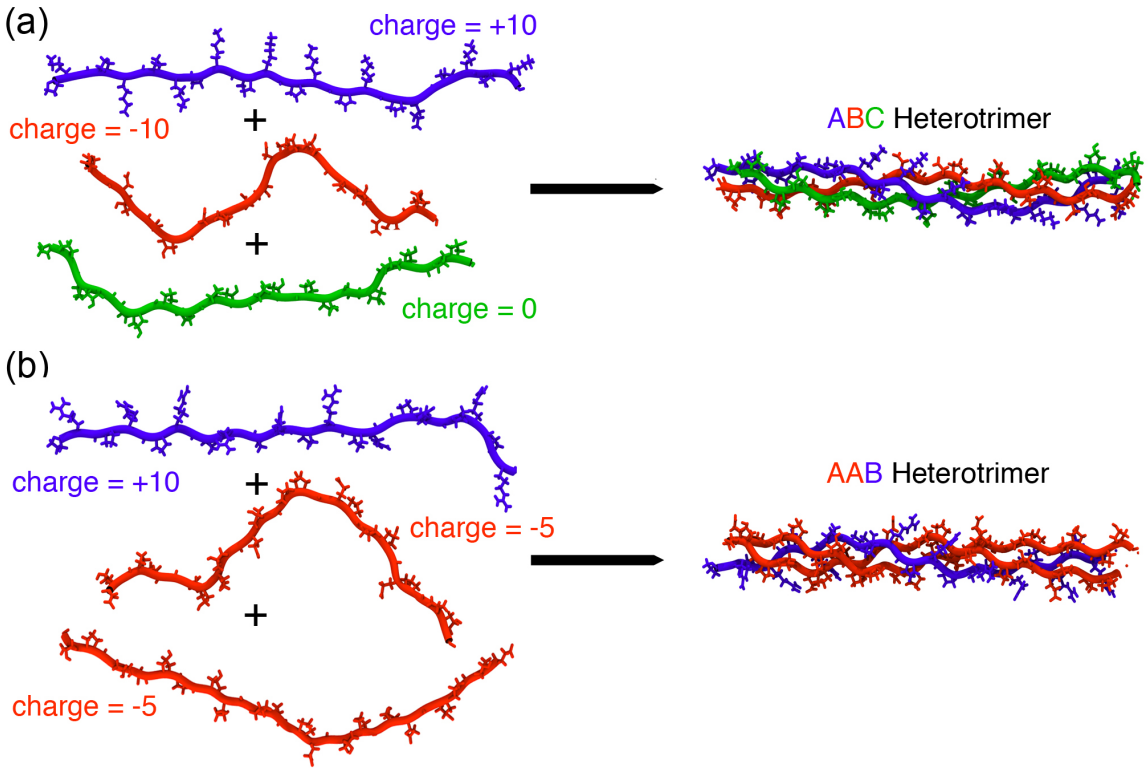


Figure 1.8. Strategy for the self-assembly of heterotrimeric CMPs with (a)ABC and (b) AAB composition.

A different approach to the problem using oppositely charged amino acids to drive the self-assembly of a stable heterotrimeric CMPs has been successful in our group to minimize many of these problems. Experiments mixing three peptides: one with charge $+\eta$, one with charge $-\eta$ and a neutral peptide, show that a zwitterionic ABC triple helix can be formed (Figure 1.8a). The effects of different values for η were inspected for CMPs with 10 triplets using arginine and glutamic acid as the charged moieties and the prototypical POG sequence as the neutral species.[101] The peptides were designed to include the positive amino acid in the Y position of the $(X\text{-}Y\text{-}Gly)_n$ triplet and the negative amino acid in the X position because this arrangement is most commonly found in natural collagens. From these experiments, it was found that the peptides with $\eta = 10$ including 20 mutations from the prototypical $(POG)_{10}$ sequence formed the most stable ABC heterotrimers. In order to optimize the heterotrimeric assembly, other plausible

charge pairs were studied following the same design pattern and it was found that lysine–aspartate interactions provide the most stable heterotrimer, with a melting temperature only a few degrees below that of a $(POG)_{10}$ homotrimer.[101,102] The structural characterization of this system and shortcomings of the design protocol will be the main focus of Chapter 3, while improvements towards its selectivity for a specific ABC-type triple helix will be discussed in Chapters 4 and 5.

This approach has also been used as the basis to design a new host-guest system to study the effects of *OI* mutations on the thermal stability and folding rate of heterotrimeric triple helices, which mimic collagen type I.[103] In the new host-guest system the flanking regions are composed of five triplets of the designed sequences, which direct triple helical self-assembly though electrostatic interactions and the guest region of nine amino acids from type I collagen. This new model system allows for both α_1 and α_2 sequences to be included and thus the effects of one or two mutations in the triple helix to be assessed (in contrast to homotrimeric models which must have either zero or three mutations) making this model a more accurate representation of the disease.[104] The results showed that the first mutation causes a drastic drop in thermal stability but subsequent mutations, although still lowering the thermal stability, have a less pronounced effect.

The approach of driving heterotrimeric assemblies through electrostatic interactions was also successful in producing an AAB heterotrimers. Mixing two peptides with a -5 charge, $(EOGPOG)_5$, and one with a +10 charge, $(PRG)_{10}$, results in the formation of a zwitterionic AAB triple helix (Figure 1.8b).[105] An interesting feature of this system is that the heterotrimer is the most stable assembly in the system, which is not the case in for the

ABC heterotrimers studied. Furthermore, this peptide assembly shows a greatly improved thermal stability when compared to other AAB self-assembled peptides available in the literature.[97,100]

1.3 References

(1) Di Lullo, G. A.; Sweeney, S. M.; Korkko, J.; Ala-Kokko, L.; San Antonio, J. D. *J. Biol. Chem.* **2002**, *277*, 4223-4231.

(2) Heino, J. *BioEssays* **2007**, *29*, 1001-1010.

(3) Kadler, K. E.; Baldock, C.; Bella, J.; Boot-Handford, R. P. *J. Cell Sci.* **2007**, *120*, 1955-1958.

(4) van der Rest, M.; Garrone, R. *FASEB J.* **1991**, *5*, 2814-2823.

(5) Khoshnoodl, J.; Cartailler, J.-P.; Alvares, K.; Cels, A.; Hudson, B. G. *J. Biol. Chem.* **2006**, *281*, 38117-38121.

(6) Tuckwell, D. S.; Ayad, S.; Grant, M. E.; Takigawa, M.; Humphries, M. J. *J. Cell Sci.* **1994**, *107*, 993-1005.

(7) Vandenberg, P.; Kern, A.; Ries, A.; Luckenbill-Edds, L.; Mann, K.; Kuhn, K. *J. Cell Biol.* **1991**, *113*, 1475-1483.

(8) Kalluri, R. *Nat Rev Cancer* **2003**, *3*, 422-433.

(9) Lauer-Fields, J. L.; Tuzinski, K. A.; Shimokawa, K.; Nagase, H.; Fields, G. B. *J. Biol. Chem.* **2000**, *275*, 13282-13290.

(10) Lauer-Fields, J. L.; Juska, D.; Fields, G. B. *Biopolymers* **2002**, *66*, 19-32.

(11) Orgel, J. P.; Irving, T. C.; Miller, A.; Wess, T. J. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9001-9005.

(12) Okuyama, K.; Wu, G.; Jiravanichanun, N.; Hongo, C.; Noguchi, K. *Biopolymers* **2006**, *84*, 421-432.

(13) Shoulders, M. D.; Raines, R. T. *Annu. Rev. Biochem.* **2009**, *78*, 929-958.

(14) Farndale, R. W.; Lisman, T.; Bihan, D.; Hamaia, S.; Smerling, C. S.; Pugh, N.; Konitsiotis, A.; Leitinger, B.; de, G., PG; Jarvis, G. E.; Raynal, N. *Biochem Soc T* **2008**, *36*, 241-250.

(15) Brodsky, B.; Thiagarajan, G.; Madhan, B.; Kar, K. *Biopolymers* **2008**, *80*, 345-353.

(16) Nagarajan, V.; Kamitori, S.; Okuyama, K. *J Biochem* **1998**, *124*, 1117-1123.

(17) Nagarajan, V.; Kamitori, S.; Okuyama, K. *J Biochem* **1999**, *125*, 310-318.

(18) Mohs, A.; Silva, T.; Yoshida, T.; Amin, R.; Lukomski, S.; Inouye, M.; Brodsky, B. *J. Biol. Chem.* **2007**, *282*, 29757-29765.

(19) Kramer, R. Z.; Venugopal, M. G.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *J. Mol. Biol.* **2000**, *301*, 1191-1205.

(20) Li, M.-H.; Fan, P.; Brodsky, B.; Baum, J. *Biochemistry* **1993**, *32*, 7377-7387.

(21) Mohs, A.; Popiel, M.; Li, Y.; Baum, J.; Brodsky, B. *J. Biol. Chem.* **2006**, *281*, 17197-17202.

(22) Li, Y.; Brodsky, B.; Baum, J. *J. Biol. Chem.* **2008**, *282*, 22699-22706.

(23) Boudko, S. P.; Engel, J.; Okuyama, K.; Mizuno, K.; Bachinger, H. P.; Schumacher, M. A. *J. Biol. Chem.* **2008**, *283*, 32580-32589.

(24) Okuyama, K.; Arnott, S.; Takajagani, M.; Kakudo, M. *J. Mol. Biol.* **1981**, *152*, 427-443.

(25) Emsley, J.; Knight, C. G.; Farndale, R. W.; Barnes, M. J.; Liddington, R. C. *Cell* **2000**, *101*, 47-56.

(26) Merrifield, R. B. *J. Am. Chem. Soc.* **1963**, *85*, 2149-&.

(27) Sakakibara, S.; Kishida, Y.; Kikuchi, Y.; Sakai, R.; Kakiuchi, K. *B Chem Soc Jpn* **1968**, *41*, 1273.

(28) Berg, R. A.; Olsen, B. R.; Prockop, D. J. *J. Biol. Chem.* **1970**, *245*, 5759-5763.

(29) Sakakibara, S.; Tanaka, N.; Kakudo, M.; Okuyama, K.; Ashida, T.; Kishida, Y. *J. Mol. Biol.* **1972**, *65*, 371.

(30) Okuyama, K.; Tanaka, N.; Ashida, T.; Kakudo, M.; Sakakibara, S.; Kishida, Y. *J. Mol. Biol.* **1972**, *72*, 571-&.

(31) Yonath, A.; Traub, W. *J. Mol. Biol.* **1969**, *43*, 461.

(32) Okuyama, K.; Takayanagi, M.; Ashida, T.; Kakudo, M. *Polym J* **1977**, *9*, 341-343.

(33) Fields, G. B.; Prockop, D. J. *Biopolymers* **1996**, *40*, 345-357.

(34) Fields, G. B. *Org Biomol Chem* **2010**, *8*, 1237-1258.

(35) Banerjee, J.; Hanson, A. J.; Muhonen, W. W.; Shabb, J. B.; Mallik, S. *Nat Protoc* **2010**, *5*, 39-50.

(36) Fields, C. G.; Grab, B.; JL, L.; Fields, G. B. *Anal. Biochem.* **1995**, *231*, 57-64.

(37) Cohen, C.; Bear, R. S. *J. Am. Chem. Soc.* **1953**, *75*, 2783-2784.

(38) Cowan, P. M.; McGavin, S.; North, A. C. *Nature* **1955**, *176*, 1062-1604.

(39) Ramachandran, G. N.; Ambady, G. K. *Curr. Sci.* **1954**, *23*, 349-350.

(40) Ramachandran, G. N.; Kartha, G. *Nature* **1955**, *176*, 593-595.

(41) Rich, A.; Crick, F. H. *Nature* **1955**, *176*, 915-916.

(42) Crick, F. H. C.; Rrich, A. *Nature* **1955**, *176*, 780-781.

(43) Rich, A.; Crick, F. H. C. *J. Mol. Biol.* **1961**, *3*, 483.

(44) Cowan, P. M.; McGavin, S. *Nature* **1955**, *176*, 501-503.

(45) Bella, J.; Eaton, M.; Brodsky, B.; Berman, H. M. *Science* **1994**, *266*, 75-81.

(46) Berisio, R.; Vitagliano, L.; Mazzarella, L.; Zagari, A. *Protein Sci.* **2002**, *11*, 262-270.

(47) Fallas, J. A.; Gauba, V.; Hartgerink, J. D. *J. Biol. Chem.* **2009**, *284*, 26851-26859.

(48) Okuyama, K. *Connect Tissue Res* **2008**, *49*, 299-310-PII 905233052.

(49) Okuyama, K.; Xu, X. Z.; Iguchi, M.; Noguchi, K. *Biopolymers* **2006**, *84*, 181-191.

(50) Kramer, R. Z.; Bella, J.; Brodsky, B.; Berman, H. M. *J. Mol. Biol.* **2001**, *311*, 131-147.

(51) Kramer, R. Z.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *Nat Struct Biol* **1999**, *6*, 454-457.

(52) Emsley, J.; Knight, C. G.; Farndale, R. W.; Barnes, M. J. *J. Mol. Biol.* **2004**, *335*, 1019-1028.

(53) Schumacher, M. A.; Mizuno, K.; Bachinger, H. P. *J. Biol. Chem.* **2006**, *281*, 27566-27574.

(54) Berg, R. A.; Prockop, D. J. *Biochem. Biophys. Res. Commun.* **1973**, *52*, 115-120.

(55) Inouye, K.; Kobayashi, Y.; Kyogoku, Y.; Kishida, Y.; Sakakibara, S.; Prockop, D. J. *Arch. Biochem. Biophys.* **1982**, *219*, 198-203.

(56) Xu, Y.; Hyde, T.; Wang, X.; Bhate, M.; Brodsky, B.; Baum, J. *Biochemistry* **2003**, *42*, 8696-8703.

(57) Bretscher, L. E.; Jenkins, C. L.; Taylor, K. M.; DeRider, M. L.; Raines, R. T. *J. Am. Chem. Soc.* **2001**, *123*, 777-778.

(58) Eberhardt, E. S.; Panasik, N.; Raines, R. T. *J. Am. Chem. Soc.* **1996**, *118*, 12261-12266.

(59) Holmgren, S. K.; Bretscher, L. E.; Taylor, K. M.; Raines, R. T. *Chem. Biol.* **1999**, *6*, 63-70.

(60) Holmgren, S. K.; Taylor, K. M.; Bretscher, L. E.; Raines, R. T. *Nature* **1998**, *392(6677)*, 666-667.

(61) Persikov, A. V.; Ramshaw, J. A.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2000**, *39*, 14960-14967.

(62) Shah, N. K.; Ramshaw, J. A.; Kirkpatrick, A.; Shah, C.; Brodsky, B. *Biochemistry* **1996**, *35*, 10262-10268.

(63) Fraser, R. D.; MacRae, T. P.; Suzuki, E. *J. Mol. Biol.* **1979**, *129*, 463-481.

(64) Jones, E. Y.; Miller, A. *J. Mol. Biol.* **1991**, *218*, 209-219.

(65) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2005**, *44*, 1414-1422.

(66) Melacini, G.; Feng, Y. B.; Goodman, M. *J. Am. Chem. Soc.* **1996**, *118*, 10359-10364.

(67) Feng, Y. B.; Melacini, G.; Taulane, J. P.; Goodman, M. *J. Am. Chem. Soc.* **1996**, *118*, 10351-10358.

(68) Kwak, J.; De Capua, A.; Locardi, E.; Goodman, M. *J. Am. Chem. Soc.* **2002**, *124*, 14085-14091.

(69) Kinberger, G. A.; Cai, W.; Goodman, M. *J. Am. Chem. Soc.* **2002**, *124*, 15162-15163.

(70) Cai, W.; Wong, D.; Kinberger, G. A.; Kwok, S. W.; Taulane, J. P.; Goodman, M. *Bioorg. Chem.* **2007**, *35*, 327-337.

(71) Cai, W. B.; Taulane, J. P.; Sorto, N. A.; Oganesyan, A.; Gutierrez, C. G.; Goodman, M. *Lett Org Chem* **2007**, *4*, 96-101.

(72) Goodman, M.; Melacini, G.; Feng, Y. B. *J. Am. Chem. Soc.* **1996**, *118*, 10928-10929.

(73) Melacini, G.; Feng, Y. B.; Goodman, M. *J. Am. Chem. Soc.* **1996**, *118*, 10725-10732.

(74) Feng, Y.; Melacini, G.; Taulane, J. P.; Goodman, M. *Biopolymers* **1996**, *39*, 859-872.

(75) Feng, Y.; Melacini, G.; Goodman, M. *Biochemistry* **1997**, *36*, 8716-8724.

(76) Melacini, G.; Feng, Y.; Goodman, M. *Biochemistry* **1997**, *36*, 8725-8732.

(77) Barth, D.; Musiol, H.; Schutt, M.; Fiori, S.; Milbradt, A. G.; Renner, C.; Moroder, L. *Chem-eur J* **2003**, *9*, 3692-3702.

(78) Barth, D.; Kyrieleis, O.; Frank, S.; Renner, C.; Moroder, L. *Chem-eur J* **2003**, *9*, 3703-3714.

(79) Henkel, W.; Vogl, T.; Echner, H.; Voelter, W.; Urbanke, C.; Schleuder, D.; Rauterberg, J. *Biochemistry* **1999**, *38*, 13610-13622.

(80) Cai, W.; Kwok, S. W.; Taulane, J. P.; Goodman, M. *J. Am. Chem. Soc.* **2004**, *126*, 15030-15031.

(81) Kinberger, G. A.; Taulane, J. P.; Goodman, M. *Inorg. Chem.* **2006**, *45*, 961-963.

(82) Yu, Y. C.; Berndt, P.; Tirrell, M.; Fields, G. B. *J. Am. Chem. Soc.* **1996**, *118*, 12515-12520.

(83) Yu, Y. C.; Tirrell, M.; Fields, G. B. *J. Am. Chem. Soc.* **1998**, *120*, 9979-9987.

(84) Fields, C. G.; Grab, B.; Lauer, J. L.; Miles, A. J.; Yu, Y. C.; Fields, G. B. *Lett Pept Sci* **1996**, *3*, 3-16.

(85) Ottl, J.; Battistuta, R.; Pieper, M.; Tschesche, H.; Bode, W.; Kuhn, K.; Moroder, L. *FEBS Lett.* **1996**, *398*, 31-36.

(86) Ottl, J.; Moroder, L. *Tetrahedron Lett.* **1999**, *40*, 1487-1490.

(87) Renner, C.; Sacca, B.; Moroder, L. *Biopolymers* **2004**, *76*, 34-47.

(88) Ottl, J.; Moroder, L. *J. Am. Chem. Soc.* **1999**, *121*, 653-661.

(89) Ottl, J.; Musiol, H. J.; Moroder, L. *J Pept Sci* **1999**, *5*, 103-110.

(90) Fiori, S.; Sacca, B.; Moroder, L. *J. Mol. Biol.* **2002**, *319*, 1235-1242.

(91) Ottl, J.; Gabriel, D.; Murphy, G.; Knauper, V.; Tominaga, Y.; Nagase, H.; Kroger, M.; Tschesche, H.; Bode, W.; Moroder, L. *Chem Biol* **2000**, *7*, 119-132.

(92) Muller, J. C. D.; Ottl, J.; Moroder, L. *Biochemistry* **2000**, *39*, 5111-5116.

(93) Sacca, B.; Moroder, L. *J Pept Sci* **2002**, *8*, 192-204.

(94) Sacca, B.; Sinner, E. K.; Kaiser, J.; Lubken, C.; Eble, J. A.; Moroder, L. *Chembiochem* **2002**, *3*, 904-907.

(95) Golbik, R.; Eble, J. A.; Ries, A.; Kuhn, K. *J. Mol. Biol.* **2000**, *297*, 501-509.

(96) Sacca, B.; Renner, C.; Moroder, L. *J. Mol. Biol.* **2002**, *324*, 309-318.

(97) Madhan, B.; Xiao, J.; Thiagarajan, G.; Baum, J.; Brodsky, B. *J. Am. Chem. Soc.* **2008**, *130*, 13520-13521.

(98) Slatter, D. A.; Miles, C. A.; Bailey, A. J. *J. Mol. Biol.* **2003**, *329*, 175-183.

(99) Slatter, D. A.; Foley, L. A.; Peachey, A. R.; Nietlispach, D.; Farndale, R. W. *J. Mol. Biol.* **2006**, *359*, 289-298.

(100) Hodges, J. A.; Raines, R. T. *J. Am. Chem. Soc.* **2005**, *127*, 15923-15932.

(101) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 2683-2690.

(102) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 15304-15041.

(103) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7509-7515.

(104) Brodsky, B.; Baum, J. *Nature* **2008**, *453*, 998-999.

(105) Russell, L. E.; Fallas, J. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2010**, *132*, 3242-3243.

(109) Renner, C.; Sacca, B.; Moroder, L. *Biopolymers* **2004**, 76, 34.

Chapter 2: Stabilizing Pair-wise Interactions in Triple Helical Proteins[1]

Triple helical proteins have a high content of imino acids, glycine and charged amino acids[1]. The latter are important because they participate in molecular recognition events[2], stabilizing pair-wise interactions[3,4], and the packing of triple helices into staggered arrays[5,6]. From a perspective of molecular design, understanding the mechanism by which ionizable residues stabilize this protein fold and participate in packing interactions would serve as valuable tool to rationally bias the self-assembly of designed CMPs to a particular heterotrimeric or fibrillar target states. Host-guest CMPS containing an imino acid-rich region at the termini and a guest region which follows the PKGXOG motif , where X is either glutamic or aspartic acid are ideally suited for this purpose. In this chapter we use a combination of X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy to study the conformation of the side chains both in solution, where inter-strand interactions are important in determining the thermal stability of the helix and in a crowded macromolecular state, where inter-triple helical contacts play a large role.

2.1 Circular Dichroism Melting Studies

As mentioned in the previous section, when considering only naturally occurring amino acids, sequences of the form $(POG)_n$ self-assemble into the most stable triple helices in an aqueous environment. It is known that any point mutation in the X or Y position from this template will destabilize the resulting helix[7]. Particularly mutations

---

[1] This chapter is based on the following publication:

Fallas, J. A.; Dong, J.; Tao, Y. J.; Hartgerink, J. D. *J. Biol. Chem.* **2012**, *287*, 8039-8047.

from O to K destabilize the helix by 10 °C, while mutations from P to D and P to E by 7 °C and 4 °C, respectively. Despite this fact, previous studies have shown that double mutations will have different effects on the stability of the triple helix. Studying peptides of the form $(POG)_3XYGX'Y'G(POG)_3$ Brodsky *et al.* found that some sequences containing double mutations involving oppositely charged amino acids approximately behave as expected from an addition of two point mutations and some that exhibit a higher thermal stability. In particular, the sequences PKGEOG and PKGDOG were found to be highly stabilizing[3], while the sequence EKGPOG and DKGPOG were not[8]. The difference in thermal stability is hard to explain based on the pairing of oppositely charged residues because molecular models showed that in both cases it is geometrically possible for the charged moieties to come in contact. Because peptides containing the XKG amino acid sequence do not provide a particularly high thermal stability we will focus on the KGX case. The melting curves for each of the two peptides containing the KGX guest sequence were repeated (Figure 2.1) and they agree were with the results available in the literature. The peptide sequences and melting temperatures, as determined by the minimum in the first derivative of the melting profile with respect to temperature, are available in table 2.1.

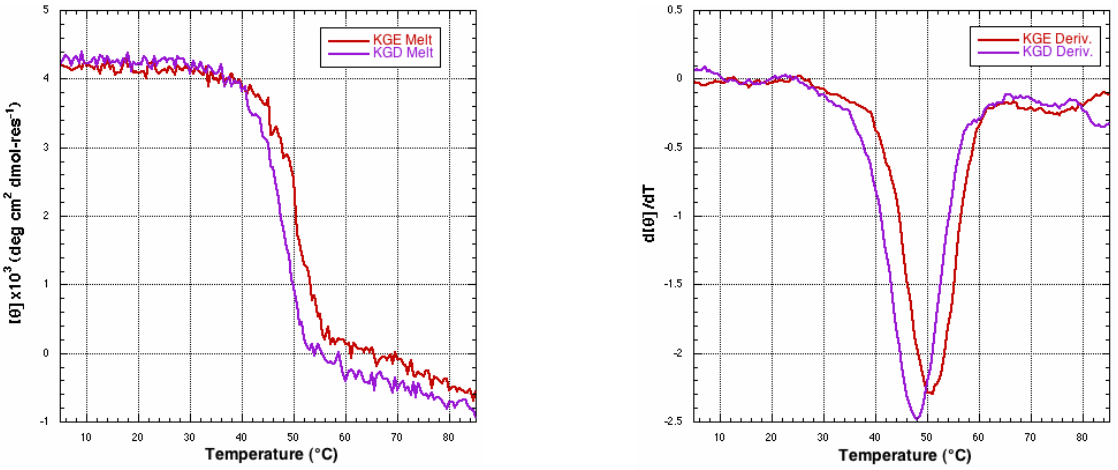


Figure 2.1. CD spectroscopy. CD melting curves and derivatives with respect to temperature for the a) **KGE** peptide and b) **KGD** peptide.

	Sequence[a]	T_m (°C)[b]
KGE	$(POG)_3PK\boldsymbol{G}EOG(POG)_3$	51
KGD	$(POG)_3PK\boldsymbol{G}DOG(POG)_3$	48

[a]Highlighted amino acids are ^{15}N-labelled. [b]The melting temperature is defined here as the minimum in the derivative of the circular dichroism melting curve.

Table 2.1. Peptide Sequences, abbreviations and melting temperatures.

2.2 Crystallographic Studies

To understand the differences between both sequence motifs (KGX vrs XKG) we focus on the structural characterization of host-guest triple helical peptides containing lysine-glutamate (**KGE** peptide) and lysine-aspartate (**KGD** peptide) salt bridges. By means of X-ray crystallography and NMR spectroscopy we are able to study the side-chain interactions in two different stages of the life of a collagen molecule and find important differences in their conformation that give insight into the different functions that the same residues can perform during the life-span of a single triple helix. The X-ray studies serve as an example of the triple helix in a crowded molecular environment, akin to the one found in the supramolecular architectures made by tightly packed triple helices in collagen fibrils.

2.2a KGE Crystal Structure

KGE readily crystallizes around neutral pH in a tacsimate buffer. The structure of the peptide was solved by molecular replacement and contains two anti-parallel triple helices in the asymmetric unit packed in a quasi-hexagonal lattice. As is commonly observed in triple helical peptides the structure shows some disorder at the termini[9], evidenced by the B-factors obtained for the terminal triplets (Figure 2.2a and c). This is particularly pronounced at the C-terminus, where poor density prohibited the modeling of the Gly24(C). Table 2.2 summarizes the data acquisition and refining parameters.

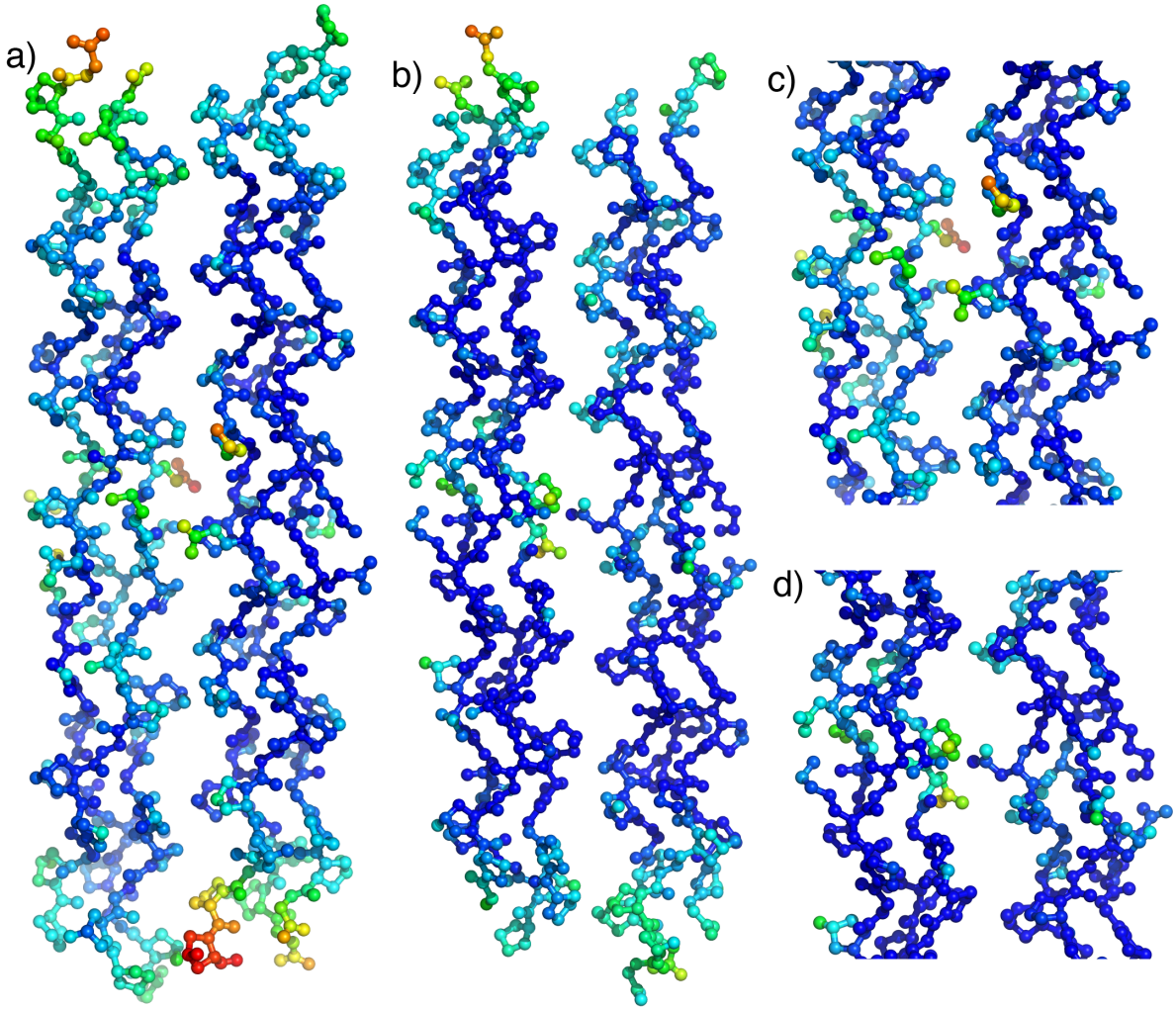


Figure 2.2. Overall structure of a) **KGE** and b) **KGD**. The guest regions of each triple

helix are highlighted in panels c) and d) respectively. Atoms are colored by B-factors. Hotter colors signify higher B-factors. All of the side chain atoms of the charged residues in the guest region were included in the final model but some of them show a higher degree of flexibility, as evidenced by their high B-factors. Image generated using pymol[10].

	3T4F (**KGE**)	3U29 (**KGD**)
Data Collection		
Space Group	P1	P1
Unit Cell Dimensions		
a, b, c (Å)	14.1, 23.8, 67.7	14.2, 23.2, 67.5
α, β, γ (°)	95.0, 94.7, 94.9	94.3, 94.7, 93.3
Resolution (Å)	1.68 (1.74 – 1.68)	2.00 (2.05 – 2.00)
Completness	80.1 %	76.9 %
Redundancy	1.1	1.1
R_{merge}	5.3 (12.0)	9.8(17.1)
I/σI	19.4	21.0
Refinement		
Resolution (Å)	1.68 (1.74-1.68)	2.00 (2.05-2.00)
Total reflections	7957	4220
R_{work}/R_{free}¶	19.5/20.8	23.9/25.0
No. atoms		
Protein	935	915
Water	218	180
r.m.s deviation		
Bond lengths (Å)	0.07	0.08
Bond angles (°)	1.4	1.5

* Data in parenthesis corresponds to the highest resolution shell.
¶ 5% of reflections were included in the test set

Table 2.2. Data Collection and Refinement Statistics.

The final model contains 935 peptide atoms and 219 water molecules and was refined to an R_{work}/R_{free} value of 18.9 / 20.7. Figure 2.3a) shows the contents of the asymmetric unit, highlighting the charged residues in the guest region by coloring lysine residues cyan, glutamic acid residues red and water molecules light blue. The $2F_o$-F_c map contoured at 1.2σ is also shown as a transparent surface to illustrate the accuracy of the phases.

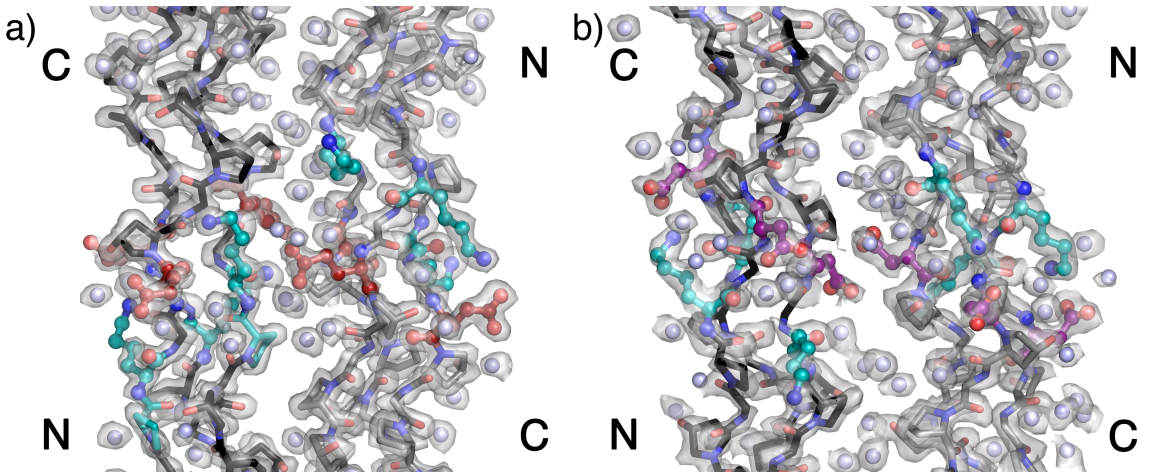


Figure 2.3. Atomic structure of a) **KGE** and b) **KGD.** Contents of the asymmetric unit of the crystals highlighting the two anti-parallel triple helices. The position of the N- and C- of each helix is noted. The 1.68 Å $2F_o$-F_c map in a) and 2.00 Å $2F_o$-F_c map in b) is contoured at 1.2σ and depicted as transparent surfaces. Image generated using pymol[10].

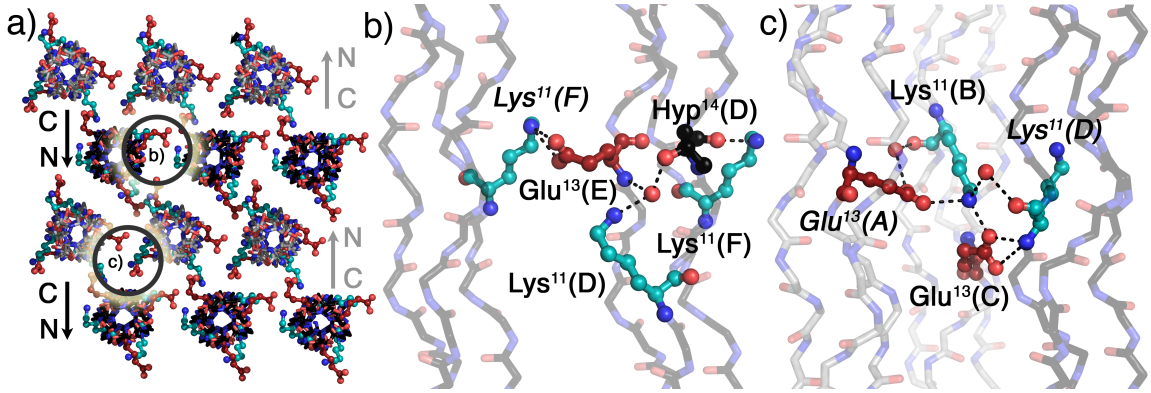


Figure 2.4. Crystal packing and Molecular interactions of **KGE**. a) Crystal packing of the **KGE** peptide highlighting the positions of the lysine(cyan) and glutamate (red) side-chains.). Side view of the areas highlighted by circles are depicted in the following panels. Triple helices oriented N- to C- terminus are shown in gray (A-leading chain, B-

middle chain, C-lagging chain) and triple helices oriented C- to N- terminus in black (D-leading chain, E-middle chain, F-lagging chain). b), c) Inter- and intra-helical hydrogen bonding networks involving the charged side-chains at the interface of b) two parallel tripe helices and c) three anti-parallel triple helices. Amino acids are labeled using their three-letter code, sequence position and chain. Image generated using pymol[10].

In general three types of interactions are possible for each of the charged moieties: i) a salt-bridge or ionic hydrogen bond, where there is a direct contact between two oppositely charged residues ii) a hydrogen bond, where the charged side chain group shares a hydrogen with a neutral peptidic polar atom and iii) a water mediated contact, where a water molecule forms a bridge between the charged side chain moiety and another polar atom. The extensive network of hydrogen bonds observed in the quasi-hexagonal crystal packing of our structure shows all three of these cases and selected examples are available in Figure 2.4.

The contacts observed in the structure can be either intra-strand, inter-strand or inter-helical depending on the relative positions of the interacting amino acids. Figure 2.5 shows different inter-strand and inter-helical contacts involving the amino acids in the guest region of the **KGE** crystal structure with symmetry equivalent positions denoted by color: triple helices depicted gray are looking down from the N-terminus in figure 2.3(chains A-leading, B-middle and C-trailing) and triple helices depicted black are looking down from the C-terminus in figure 2.3(chains D-leading, E-middle and F-trailing). Panel a) shows an ionic hydrogen bond between Lys^{11}(B) and Glu^{13}(C). Panel b) shows the equivalent interaction in the second helix of the asymmetric unit. In this case Lys^{11}(E) forms a hydrogen bond with the backbone carbonyl of Lys^{11}(F) instead of a direct salt-bridge with Glu^{13}(F). In panel c) the Lys^{11}(F) side-chain also prefers a backbone hydrogen bond, in this case to the Hyp^{14}(D) carbonyl, instead of the direct

hydrogen bond to the negatively charged Glu[13](A) carboxylate. Panel d) shows a top view of the packing interactions involving the residues in panel b) to highlight the fact that although the Glu[13](F) side-chain does not interact directly with Lys[11](E) there is a water mediated hydrogen-bond between both residues. Further water mediated-contacts are observed between Glu[13](F) and Lys[11](A) and Glu[13](A), with the latter two amino acids coming from distinct symmetry related helices.

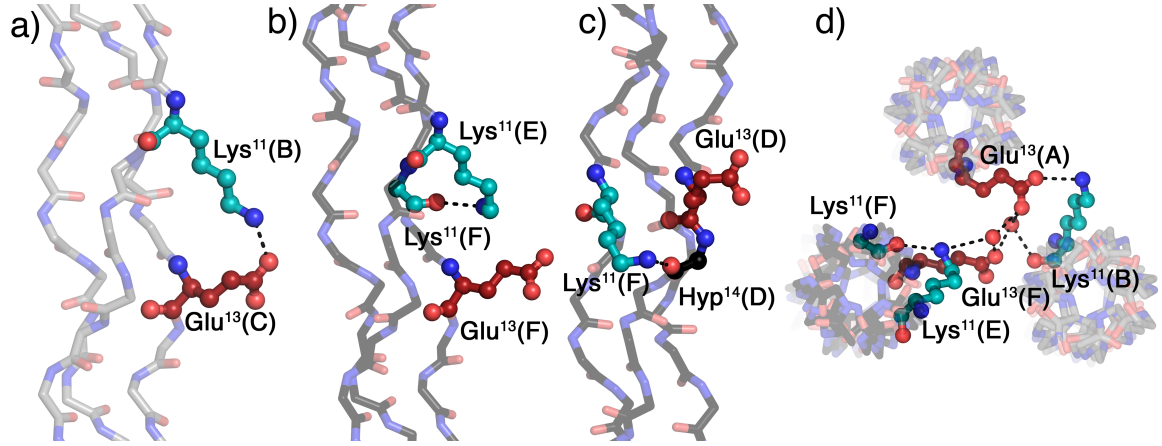


Figure 2.5 Inter-strand, intra-strand and inter-helical interactions in **KGE**. a),b) Inter-strand hydrogen bonds involving the charged side-chains. c) Intra-strand hydrogen bond d) Top view of the packing interactions involving the residues depicted in b) and c). Triple helices oriented N- to C- terminus in figure 1 are shown in gray (A-leading chain, B-middle chain, C-lagging chain) and triple helices oriented C- to N- terminus in figure 1 are depicted in black (D-leading chain, E-middle chain, F-lagging chain). Lysines are shown in cyan and glutamates in purple. Amino acids are labeled using their three-letter code, sequence position and chain. Images generated using pymol[10].

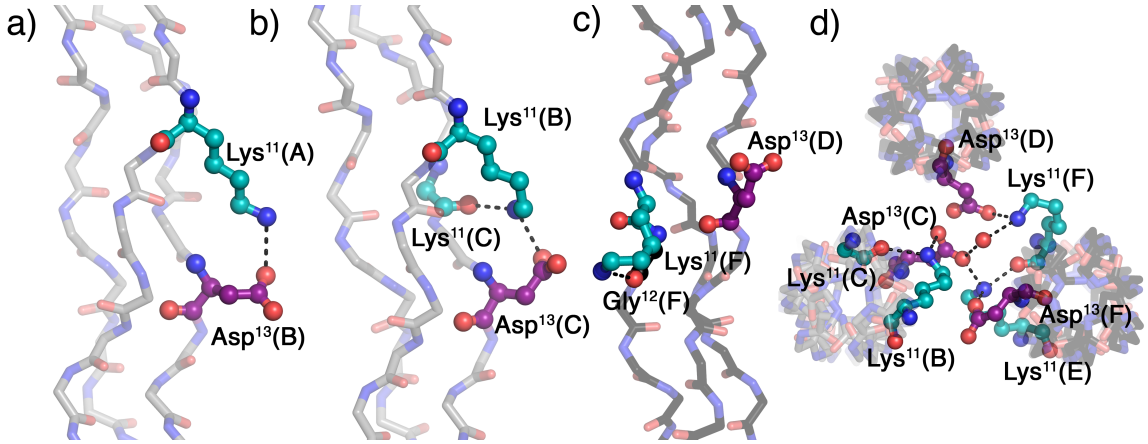


Figure 2.6 Inter-strand, intra-strand and inter-helical interactions in **KGD**. a),b) Inter-strand hydrogen bonds involving the charged side-chains. c) Intra-strand hydrogen bond d) Top view of the packing interactions involving the residues depicted in b) and c). Triple helices oriented N- to C- terminus in figure 1 are shown in gray (A-leading chain, B-middle chain, C-lagging chain) and triple helices oriented C- to N- terminus in figure 1 are depicted in black (D-leading chain, E-middle chain, F-lagging chain). Lysines are shown in cyan and aspartates in purple. Amino acids are labeled using their three-letter code, sequence position and chain. Images generated using pymol[10].

2.2b KGD Crystal Structure

KGD readily crystallizes around neutral pH in a tacsimate buffer. The structure of the peptide was solved by molecular replacement and it is similar to the **KGE** structure, with two anti-parallel triple helices in the asymmetric unit packed in a quasi-hexagonal lattice and higher B-factors at the termini (Figure 2.2b and d). The final model contains 915 peptide atoms and 180 water molecules and was refined to an R_{work}/R_{free} value of 23.9/25.0. Figure 2.3b) shows the contents of the asymmetric unit, highlighting the charged residues in the guest region by coloring lysines cyan and aspartates purple with waters are depicted in light blue. The 2.00 Å $2F_o$-F_c map contoured at 1.2σ is depicted as a transparent surface. Table 2.2 summarizes the refinement statistics.

Figure 2.6 shows different inter-strand, intra-strand and inter-helical contacts involving the amino acids in the guest region of the **KGD** crystal structure with symmetry equivalent positions denoted by color as previously described. Panel a) shows a salt-bridge between Lys^{11}(A) and Glu^{13}(B). Panel b) shows a similar interaction between Lys^{11}(B) and Glu^{13}(C), however, because of differences in the lysine side-chain conformation a second hydrogen bond between the amino group and the Lys^{11}(C) backbone carbonyl is also possible. Panel c) depicts the Lys^{11}(F) side-chain engaging in an intra-strand backbone hydrogen bond to the Gly^{12}(F) carbonyl instead of a direct interaction with its nearest neighbor carboxylate. Panel d) shows a top view of the packing interactions involving the residues from panels b) and c) to highlight that even though Glu^{13}(D) does not participate in inter-strand interactions it engages in several inter-helical contacts, including an ionic hydrogen bond and a water mediated contact to Lys^{11}(F) in an adjacent, symmetry related helix. Furthermore, it shows Glu^{13}(C) participating both in inter-strand hydrogen bonds, as described above, and in inter-helical contacts with an ionic hydrogen bond to Lys^{11}(E) and a water-mediated hydrogen bond to Lys^{11}(F).

2.3 Solution NMR Studies

The NMR experiments serve as a model for individual collagens in solution. Using this technique it is possible to observe preferential pair-wise interactions between the different peptide strands within a single triple helix that serve to stabilize the native state.

2.2a KGE Solution Conformation

The solution conformation of the **KGE** peptide was also investigated by means of multi-dimensional NMR experiments. In order to facilitate the analysis a ^{15}N-labelled glycine residue was included in position 12 (table 2.1). The ^{1}H,^{15}N-HSQC spectrum of the system (Figure 2.7a) shows three distinct peaks for the triple helical assembly and also a monomer peak.

Despite having a homotrimeric composition the one-residue stagger characteristic of this protein fold lifts the degeneracy associated with symmetry equivalent positions in the guest region of our triple helix, which is observed for the host region (Figure 2.8a). Because of the degeneracy observed in this region the NMR analysis will focus on amino acids Lys^{11}-Glu^{13}. In the following text when referring to a particular proton the superscript next to the three-letter code will denote its chain and the atom name will be given in parenthesis.

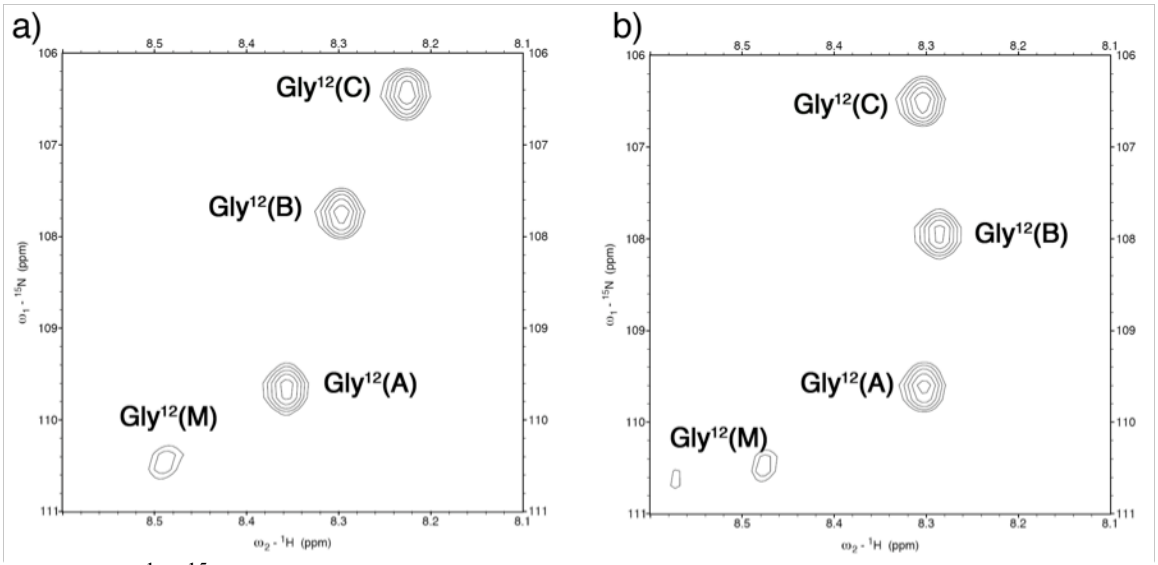


Figure 2.7 ^{1}H,^{15}N-HSQC spectra. a) **KGE** spectrum and b) **KGD** spectrum. Amino acids are labeled using their three-letter code, sequence position and chain (A-leading chain, B-middle chain, C-lagging chain and M- monomer).

The side chain conformation of the charged residues can be studied using the NOESY spectrum of **KGE (**figure 2.9). The glutamic acid residues in both chain B and C show a rather rigid conformation with distinct chemical shifts for each of the four side chain methylene protons (figure 2.9a). Besides the intra-residue NOEs each of these Glu(NH) protons shows cross peaks to the Lys(H_ε) methylene in the preceding chain, which shows distinct resonances for each of its diasteriotopic protons (figure 2.9 c). It is possible to unambiguously assign the observed resonances to be $Lys^A(H_{\varepsilon 2})$-Glu^B(NH), $Lys^A(H_{\varepsilon 3})$-Glu^B(NH) and $Lys^B(H_{\varepsilon 2})$-Glu^C(NH), $Lys^B(H_{\varepsilon 2})$-Glu^C(NH), the latter pair showing stronger peaks. On the other hand the glutamate in the leading chain shows distinct chemical shifts only for its β-protons while the γ-methylene presents a single chemical shift for both hydrogens indicating a more dynamic conformation. This residue lacks a resonance to the lysine ε-methylene of the lagging strand, which shows degenerate shifts for both ε-protons. In order to illustrate the side chain conformation observed in solution for the guest region of the **KGE** peptide a family of structures was generated to approximate the native ensemble of the triple helical assembly. Molecular dynamics simulations were carried out starting from the crystal structure to sample alternate conformations, which were then subjected to a constrained minimization using distance restraints extracted from the NOE cross peak intensities. This methodology allows for efficient sampling of the relevant conformational space by biasing the geometry of the structures visited during the MD simulations towards the native free energy basin using experimental constraints derived from the NOESY spectra. It has been shown that molecular dynamic simulations can generate ensembles that show significant overlap with those obtained by a traditional NMR structure determination process[11]. We

do not attempt to treat these structures as a quantitative thermodynamic ensemble due to the difficulties associated with accurate computation of electrostatics to protein stability[12].

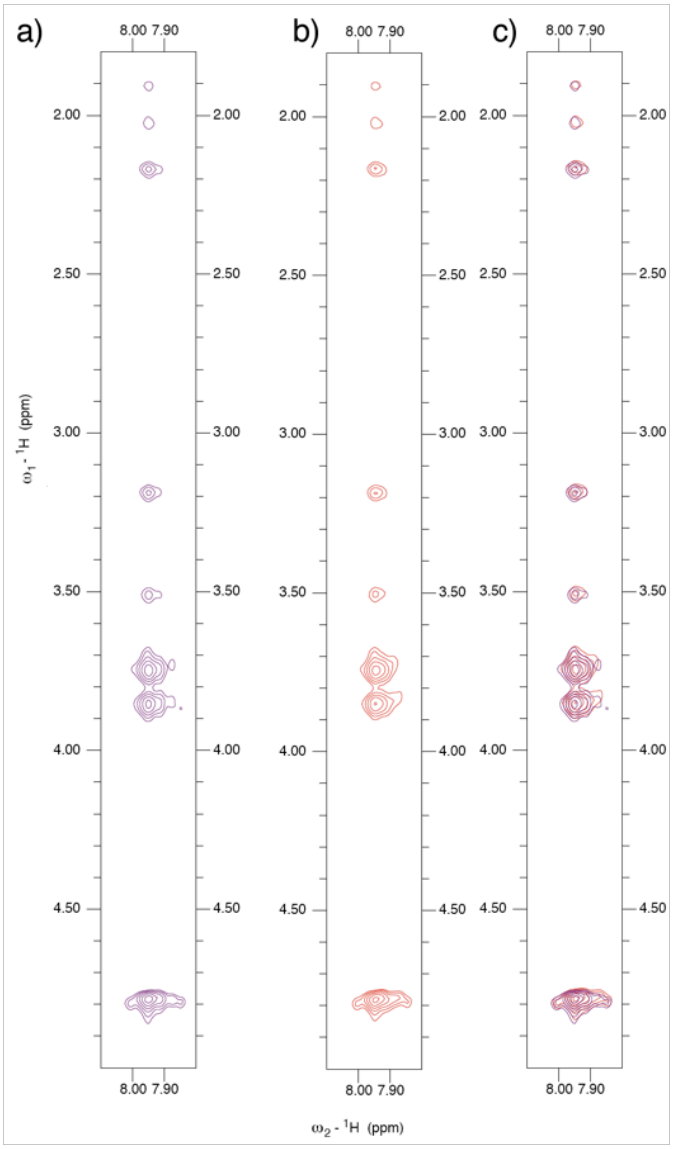


Figure 2.8 ^{1}H,^{1}H-NOESY Spectra. Strip of the spectrum showing the chemical shift of glycine in the host region for a) **KGE** peptide, b) **KGD** peptide and c) overlay showing the similarity observed between the host regions of both peptides.

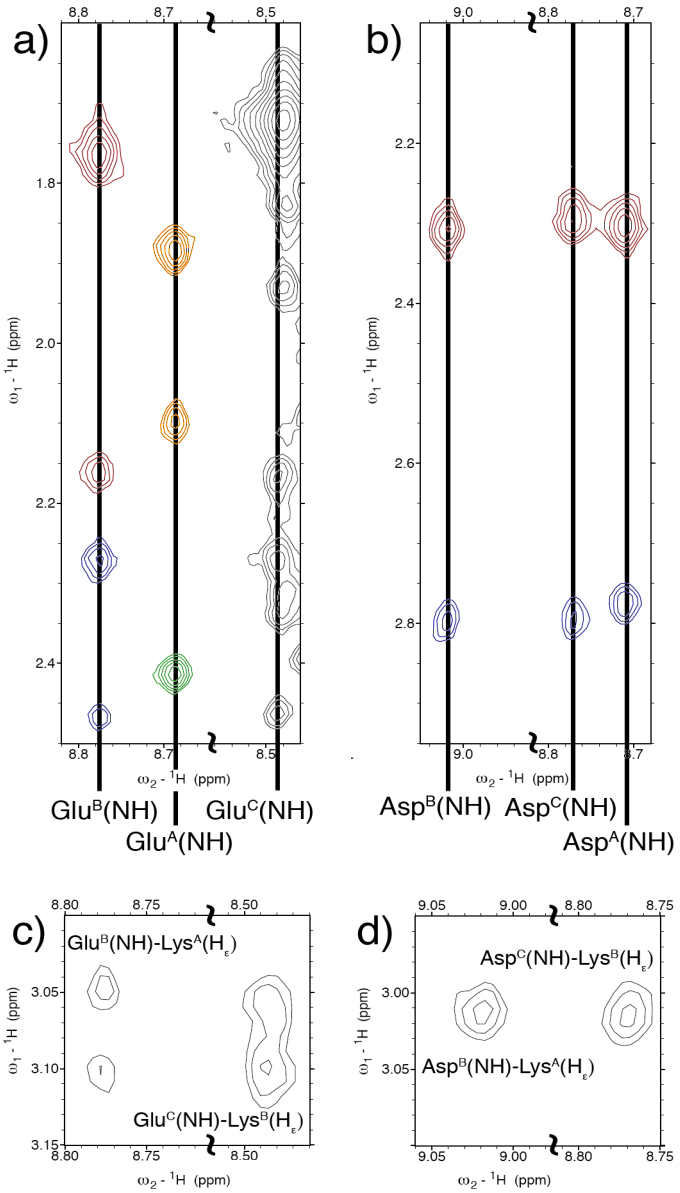


Figure 2.9 $^{1}H,^{1}H$-NOESY spectra. Strips of the **KGE** NOESY spectrum (panels a & c) and strips of the **KGD** NOESY spectrum (panels c & d). *Panel a* shows intra-residue Glu(NH)-Glu($H_{\beta,\gamma}$) resonances in **KGE**. Three columns can be observed corresponding to Glu(NH) in chain B, A and C. Glu^{B}(NH) shows cross peaks with $Glu^{B}(H_{\beta})$ protons (red) and $Glu^{B}(H_{\gamma})$ protons (blue). Glu^{A}(NH) shows cross peaks with $Glu^{A}(H_{\beta})$ protons (orange) and $Glu^{A}(H_{\gamma})$ protons (green). A similar pattern is observed for Glu^{C}(NH) column, although the peak corresponding to $Glu^{C}(H_{\beta 3})$ overlaps with the intense $Lys^{A,B}$(NH)-$Lys^{A,B}(H_{\gamma 3})$ resonance. *Panel c* shows intra-residue Asp(NH)-Asp(H_{β}) resonances in **KGD**. Three columns can be observed corresponding to Asp(NH) in chains B, C and A. The peaks near 2.78 (blue) and 2.30 ppm (red) that correspond to the correlation with Asp($H_{\beta 3}$) and Asp($H_{\beta 2}$) protons respectively. *Panel c* shows inter-chain Glu(NH)-Lys(H_{ε}) resonances in **KGE**. *Panel d* shows inter-chain Asp(NH)-Lys(H_{ε}) resonances in **KGD**.

The 100 lowest energy structures were selected for the final ensemble with an overall backbone rmsd of 0.83 Å. Figure 2.10 a-c shows the ensemble highlighting interactions between chains A-B (a), B-C (b) and C-A (c). It is possible to divide the observed contacts into two subsets based on their geometry. The first one includes contacts between chains A-B and chains B-C, specifically Lys^{11}(A)-Glu^{13}(B) and Lys^{11}(B)-Glu^{11}(C) and will be referred to as an axial interaction since the interacting side chains are arranged along the helical axis (Figure 2.11a), leading to a geometry that facilitates the formation of ionic hydrogen bonds between the charged side chains. Furthermore, there is a slight difference in the interaction between chains A-B and B-C, which can be observed in the NMR ensemble of the assembly as depicted in Figures 3a and 3b, with the B-C interaction presenting a higher degree of conformational flexibility. In terms of sequence this interaction occurs between a lysine residue in position *n* and a glutamic acid in position *n+2* in subsequent chain of a triple helix, provided that the lysine is either in chain A or B. The axial interaction between chains C-A, although possible, requires a different amino acid sequence, namely for glutamic acid to be in position *n+5* in chain A if lysine occupies position *n* in chain C. Instead, we observe a lateral interaction between the two remaining chains (Figure 2.11b), which is characterized by a larger degree of conformational flexibility. In this case there is a competing interaction between the Lys^{11}(C)-Glu^{13}(A) salt bridges and a Lys^{11}(C)-Hyp^{14}(A) hydrogen bond involving the lysine H_ζ protons and the hydroxyproline backbone carbonyl. While it is possible to satisfy both contacts simultaneously, some conformers sampled interacted with neither. The corresponding lateral interaction

between chains A-B or B-C is also possible but would require for glutamic acid to be inposition *n-1* (in chain B or C) if lysine is in position *n* (in chain A or B respectively).

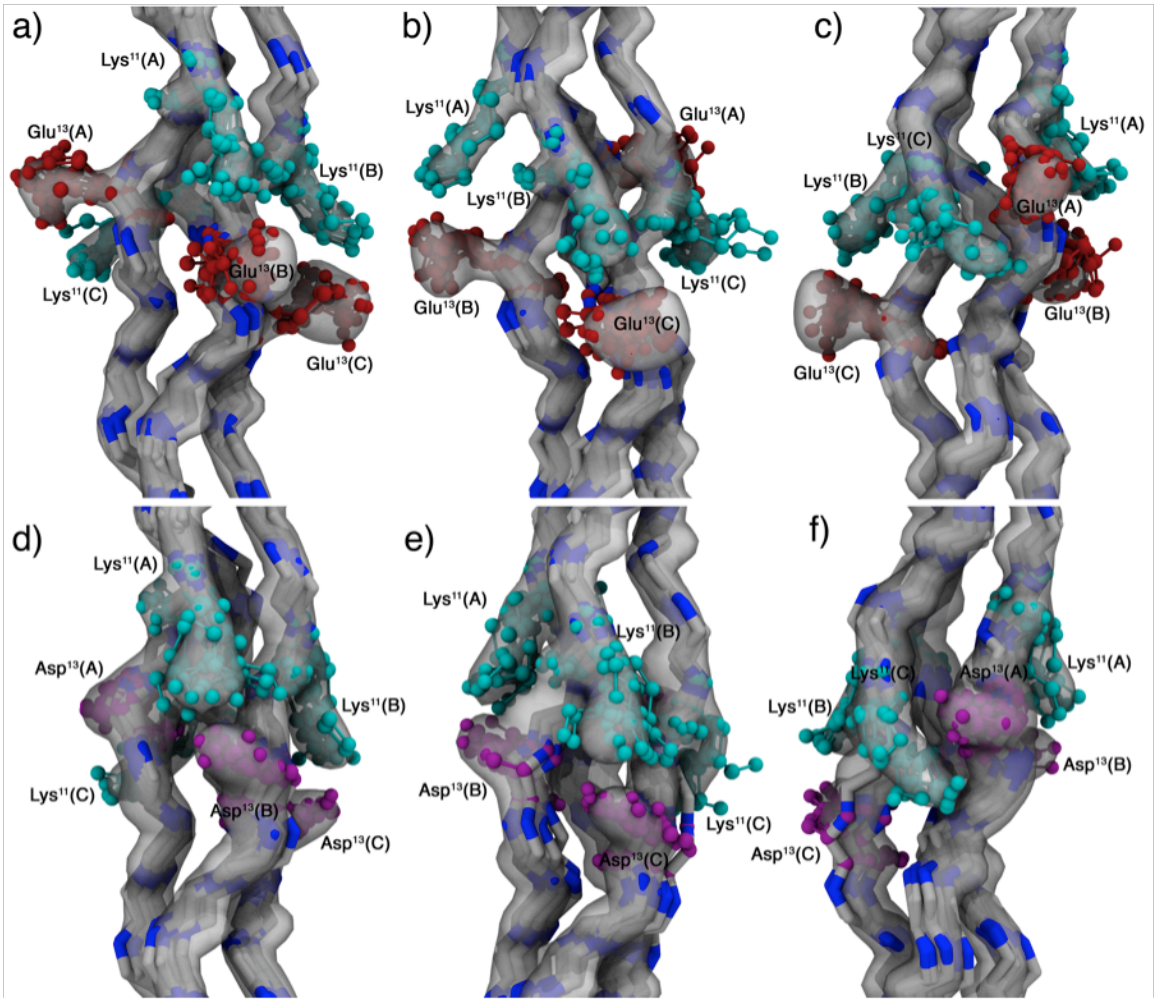


Figure 2.10 NMR Ensembles. a-c) **KGE** ensemble and d-f) **KGD** ensemble. Interactions between the charged residues in chains A-B (a,d), B-C (b,e) and C-A (c,f) are highlighted. The reweighted atomic density for each ensemble[13] is depicted as a semi-transparent surface at a value of 40% and 10 representative structures are shown as a backbone trace with the lysine (cyan), glutamic acid (red) and aspartic acid (purple) residues depicted in a cpk model. Amino acids are labeled using their three-letter code, sequence position and chain (A-leading chain, B-middle chain and C-lagging chain). Images generated using vmd-xplor [14].

2.2b KGD Solution Conformation

The solution structure of the **KGD** peptide was also studied by multi-dimensional NMR experiments. The $^{1}H,^{15}N$-HSQC spectrum of the system (Figure 2.7b) also shows three distinct peaks for the triple helical assembly and two monomer peaks, which we ascribe to cis-trans isomerization of the prolyl amide bonds surrounding the guest region[15]. As for **KGE**, the host region (Figure 2.8b) shows mostly degenerate peaks and thus further analysis will focus in the guest region (Lys^{11}-Asp^{13}).

Again, the side chain conformation of the charged residues is studied using the NOESY spectrum (Figure 2.9b and d). In this case all three aspartic acid residues show similar conformations with distinct chemical shifts for the $Asp(H_{\beta 2})$ and $Asp(H_{\beta 3})$ protons (Figure 2.9b). Inter-chain NOEs are also observed between Asp^{B}(NH) and Asp^{C}(NH) and the $Lys^{A,B}(H_{\varepsilon})$ protons, which show a single chemical shift for both methylene protons. Based on structural constraints it is possible to assign the resonances to be $Lys^{A}(H_{\varepsilon})$-Asp^{B}(NH) and $Lys^{B}(H_{\varepsilon})$-Asp^{C}(NH). There is no comparable resonance between chains C and A (Figure 2.9d).

An NMR ensemble was also generated for this peptide sequence using the same methodology described earlier with modified coordinates to include aspartic acid. Figure 6 d-f shows the 100 lowest energy structures, which have an overall backbone rmsd of 0.91 Å. The conformations observed mirror those of the **KGE** ensemble, presenting both axial interactions between chains A-B and B-C and lateral interactions between C-A(Figure 7). The **KGD** ensemble shows no significant difference between the axial interactions. Furthermore, there are few geometries with salt bridges between chains C-A,

instead the H_ζ atoms of lysine interact preferentially with the backbone carbonyl of Hyp^{14}(A) and Asp^{13}(A).

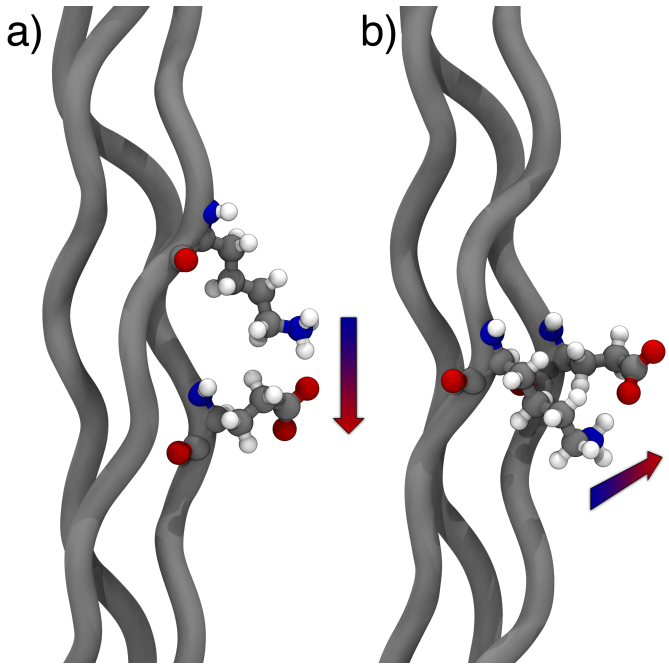


Figure 2.11 Molecular models highlighting the different contact geometries observed in triple helical proteins. a) Axial interaction between lysine and glutamate b) Lateral interaction between lysine and glutamate.

2.3 Biological Relevance

Most atomic-resolution data concerning charge pair interactions in collagens comes from the crystal structures of homotrimeric triple helical peptides[5,6]. However, due to their large surface area (2/3 of all amino acids are surface exposed), which can lead to extensive inter-protein contacts, crystallography may not the best analytical tool to study events that concern individual triple helices in solution. This fact is illustrated in our studies by the different rotamers observed for the same charged amino acid in different triple helices of the **KGE** and **KGD** asymmetric units. The NMR spectra on the other hand suggest a more homogeneous ensemble for both peptides. After analyzing the

rotamers observed in the crystal structure it seems clear that they adopt conformations dictated by the crystal packing. Furthermore, even the residues that are involved in contacts with other amino acids in the same triple helix present interactions with other helices either directly or through water mediated hydrogen bonds. It is interesting to note that the packing interactions observed in our quasi-hexagonal lattice are able to accommodate different interaction networks for the same charged amino acids. This finding supports the idea that lysine and glutamic acid rearrange between different sets of possible interactions in collagen type I fibers[16]. Because we have two anti-parallel triple helices in the unit cell, we observe interfaces between anti-parallel triple helices, which are relevant as chemical models for some heterotypical collagen assemblies. For instance type II collagen fibers, which pack in a hexagonal array [17], are known to be decorated in the surface by type IX molecules in an anti-parallel fashion[18], a collagen type that has a particularly high content of glutamic acid residues[19].

When trying to understand the molecular basis of stabilizing interactions in triple helical proteins it is necessary to study the amino acid interactions in the appropriate environment. As mentioned previously, the charged amino acids in the guest region engage in extensive networks of inter-triple helical interactions in the crystal structure. On the other hand, in solution such interactions are minimal and therefore do not contribute to the side-chain conformation leading to a different arrangement of the charged amino acids. Despite these fundamental differences the crystallographic study is vital because it allows to obtain information on the overall backbone structure of the assemblies and shows some examples of conformations that allow for intra- and inter-strand hydrogen bonds. Nonetheless, we feel that solution NMR experiments are the most

accurate way to study the amino acid interactions that lead towards an increase in the thermal stability of triple helical proteins.

Our structural study shows that it possible to divide the observed interactions in solution into two categories: the first one comprises interactions between the leading chain and middle chain as well as interactions between the middle chain and lagging chain and we have labeled them as axial contacts, the second one comprises interactions between the lagging chain and leading chain which we labeled a lateral contact. Despite having identical sequences these differences arise from the relative stagger of the peptide strands within the triple helix and lead to different conformations of the interacting amino acids.

In the axial contact the lysine side-chain extends down the helical axis reaching towards the acidic residue in the neighboring strand. In terms of amino acid composition, the lysine-glutamate pair shows a more rigid conformation for the lysine side-chain than the lysine-aspartate one. This can be rationalized by noting that lysine and glutatmate are only able to make a direct ionic hydrogen bond, which we deem the most stable interaction, if all the lysine c dihedrals adopt a trans conformation. On the other hand the shorter aspartate side-chain allows for more flexibility forming salt-bridges with different rotamers, including the all trans conformation and a related rotamer were the χ_4 dihedral adopts a gauche conformation.

The lateral interaction is characterized by an overall increase in side-chain flexibility. An analysis of the solution conformation of this interaction shows that there are several hydrogen bonds and salt-bridges that are possible for this contact geometry leading to less rigid conformation for both interacting amino acids.

The stabilizing effect of the sequences present in this study was first noted by Brodsky *et al.*[3]. Using molecular modeling the authors hypothesized the presence of contacts similar to the ones described in the previous sections. The authors also carried out a statistical analysis of human collagen type I, II, III, V and type X that showed a higher occurrence of KGE/D sequences than that expected based on the number of occurrences for each of the individual amino acids, indicating that this motif may have been selected for evolutionarily as an alternate stabilization mechanism for triple helical proteins. However, no difference was made between axial and lateral interactions in the analysis.

A computational study of the **KGE** and **KGD** peptides carried out by Stultz *et. al.* [20] shows asymmetry in the salt bridges between the different peptide chains, which agrees with our interpretation of the structural data. Their methodology allowed for an estimation of the free energy contribution of each pair using an explicitly solvated model and assigns low free energy contributions to the lateral pairs of both peptides. This study allows us to make a direct link between the observed structural differences in the contact geometry and their effect on the stability of the triple helix. In this context, it is known that the EKG sequence does not provide any thermal stabilization to triple helical peptides[8]. In the EKG arrangement where K is in position *n* and E is position *n-1*, the sequence alignment resulting from the one residue stagger characteristic of collagenous domains allows for only lateral interactions from leading to middle and middle to lagging chains and does not allow for any efficient interactions between the lagging and leading chains because the residues are too far apart[3]. It should be noted that the structure of the peptide with the sequence EKG in its guest region [5], which packs into a staggered parallel

array, does not exhibit any salt bridges. Additionally, it shows conformers for the lysine residues in its leading and middle chains similar to the one observed in lagging chain F of our crystal structure. More generally it agrees with the conformations observed for the lagging chain of our NMR ensemble, which participates in a lateral interaction. Given that no increased thermal stability is observed for the sequence containing only lateral interactions we propose that they only contribute marginally to the stability of triple helical proteins. By this logic, the increased thermal stability observed in the KGE/D sequences comes from the two axial interactions that we observe in the NMR ensemble of the peptides.

Assuming that there is an energetic difference between the lateral and axial interactions one can speculate about the assembly and energy landscape of type I collagen, an AAB heterotrimer, particularly the segments that have the sequences presented in this study. Further analysis of the collagen type I sequence shows that the KGD sequence occurs only in the α1 chain while the KGE sequence appears in both the α1 and α2 chains. However, all but one of the KGD occurrences in the α1 chain correspond to KGE triplets in the α2 chain. Given the asymmetry observed in the different charge pair interactions this sequence arrangement suggests that type I collagen may utilize this as a form of register specific stabilization.

2.4 Conclusions

The structural study presented here follows the conformation of charged amino acids in different environments that are relevant for the function of collagenous proteins. We are able to identify how the charge pairs stabilize this fold and, particularly, we

identify what amino acid interactions are energetically favorable. Using these ideas we are able to speculate on the registration process of natural collagens an important question that remains unanswered. Furthermore, we are characterize the differences between the solution conformation of ionizable residues and their conformation in a crowded macromolecular state in which inter-helical interactions are important. In this state we find hydrogen-bonding networks that can direct the packing of anti-parallel triple helices into quasi-hexagonal arrays. Similar packing of triple helices is found in collagen type II heterotypical fibers decorated with FACIT type IX collagen.

The fact that Lys-Glu and Lys-Asp salt bridges show distinct interactions depending on the relative stagger of the interacting peptide chains also has important implications for the design of single-register and single-composition heterotrimeric triple helices. In the coming sections the use of oppositely charged amino acids to drive heterotrimeric triple helical assemblies will be discussed.

2.5 Experimental

Peptide Synthesis and Purification - The peptides were synthesized in house with an Advanced Chemtech Apex 396 solid phase peptide synthesizer using standard Fmoc chemistry and a Rink MBH amide resin. During the automated procedure a manual addition of 2 equivalents ^{15}N-labelled glycine, purchased form Cambridge Isotope Laboratories, was carried out in position 12. The final products are N-terminally acetylated and C-terminally amidated to provide increased thermal stability. Purification was performed on a Varian PrepStar220 HPLC using a preparative reverse phase C-18

column with a linear gradient of water and acetonitrile each containing 0.5 % TFA and analyzed by ESI-TOF mass spectrometry on a Bruker microTOF instrument

Crystallization and Data Collection – The pure and lyophilized peptide powder was dissolved at concentrations of 8, 10, 12, 15 mg/ml in water and pH adjusted to a value of 7.0 using 0.1 M sodium hydroxide. The peptides were crystallized using the hanging drop vapor diffusion method by mixing 1 µl of the peptide solution with 1 µl of 50% tacsimate solution, purchased from Hampton Research. Crystals grew as thin plates at pH values ranging from 5.9 – 7.1 in approximately 3 days. Crystals at pH 6.4 and peptide concentration of 12 mg/ml were chosen for data acquisition. The sample were flash-cooled in a N_2 cryostream at 100K using 7.5 % glycerol in the mother liquor as cryoprotectant. Data was collected at 1.54 Å using a Rigaku RUH3R rotating anode X-ray generator with a Rigaku R-axis IV++ detector in 0.5° wedges. The **KGE** crystals diffracted to 1.68 Å and were indexed to a triclinic unit cell, space group P1, with dimensions: a = 14.0 Å, b = 23.8 Å, c = 67.7 Å, α = 95.0° , β = 94.7° and γ = 94.9° using the hkl2000 software[21]. The **KGD** crystals diffracted 2.00 Å and were indexed to a triclinic unit cell, space group P1, with dimensions: a = 14.2 Å, b = 23.2 Å, c = 67.5 Å, α = 94.3° , β = 94.7° and γ = 93.3° using the hkl2000 software[21].

Structure Determination and Refinement - The structures were solved by molecular replacement using the epmr software[22]. Several search models were tried for the **KGE** crystal but a modified version of the structure 1QSU[5] containing alanine mutations at positions 13, 14 and 16 and reduced from ten to eight triplets in length yielded the highest correlation coefficient (CC = 0.722, R-factor=0.45) for the with two anti-parallel triple helices in the asymmetric unit. THE **KGD** structure was solved using a modified version

of the **KGE** structure containing alanine mutations at position 13 of each chain and was also found to contain two anti-parallel triple helices in the asymmetric unit (CC = 0.75, R-factor=0.42). The initial phases were improved by rigid body refinement followed by rounds of simulated annealing and anisotropic B-factor refinement starting at 3.0 Å resolution and gradually increasing using the CNS suite[23]. The models were rebuilt in coot[24] when the composite omit map showed clear density for the missing side chains. -CNS since it is known that triple helical peptides are often associated with an extensive water network that contributes significantly to the total scattering of the asymmetric unit[525]. After each round of automated water picking further rounds of atomic position, temperature factor refinement and model rebulding were carried out with increasing resolution until the limiting value of 1.68 Å was reached for the **KGE** structure and 2.01 Å for the **KGD** structure. The final **KGE** model contains 935 peptide atoms and 219 water molecules. The C-terminal glycine of the B chain was not modeled due to poor density in that region. The final **KGD** model contains 915 peptide atoms and 180 water molecules. The N-terminal proline and hydroxyproline of the D chain were not modeled due to poor density in that region. The final CNS models were subjected to TLS refinement in refmac[26], where each of the six peptide chains in the asymmetric unit was treated as a rigid body in the procedure. The final **KGE** structure has an R_{work} / R_{free} value of 18.9 / 20.7 % and the final **KGD** structure has an R_{work} / R_{free} value of 23.9 / 25.0 % .

NMR Spectroscopy- All NMR experiments were recorded in an 800 MHz Varian spectrometer equipped with a triple resonance probe at 5° C. The spectra were processed using NMRpipe[27] and analyzed using ccpnmr[28]. Square Cosine bell window functions were used as apodization functions and the data was zero-filled to the next power of two

in both dimensions. Drift corrections were applied when necessary. Samples of each peptide were prepared with a total peptide concentration of 3 mM, determined by mass, in a 9:1 ratio of H_2O to D_2O and 10 mM phosphate buffer at neutral pH. Each sample was characterized using 2D total correlated spectroscopy (TOCSY), nuclear Overhauser effect spectroscopy (NOESY), $^{1}H,^{15}N$-heteronuclear single quantum coherence (HSQC) and 3D NOESY-^{15}N-HSQC experiments. The sequential assignment procedure was carried out using a combination of $^{1}H,^{1}H$-TOCSY and $^{1}H,^{1}H$-NOESY experiments. All sequential NOEs from the NH of residue *i* to the $C_{\alpha}H$ of residue *i-1* are observed in the guest region. The chain register assignment was determined using the resonance between $Lys^{C}(H_{\alpha})$ and Glu^{A}(NH) or Asp^{A}(NH), which, due to structural constraints, only occurs between the lagging and leading strands. The side chain resonances were assigned using a combination of $^{1}H,^{15}N$,^{1}H-NOESY-HSQC experiments. TOCSY spectra with a 30 ms spinlock duration at 8 kHz were acquired with a total of 1360 complex points recorded in 8 scans for the directly acquired dimension while 480 increments were used in the indirect dimension. NOESY spectra with a 75 ms mixing time were acquired with a total of 1360 complex points recorded in 8 scans for the directly acquired dimension while 480 increments were used in the indirect dimension. A square spectral window of 8000 Hz was used for all spectra. A total of 1192 complex points in 32 scans for the direct dimension and 50 increments in the indirect dimension were acquired for the $^{1}H,^{15}N$-HSQC experiments using a spectral window of 8000 Hz in the hydrogen dimension and 1620 Hz in the nitrogen dimension. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in both dimensions. For the 3D NOESY-^{15}N-HSQC spectra a mixing time of 100 ms was used and a total of 1360

complex points in 8 scans for the direct dimension, 120 increments for the first indirect dimension and 12 increments for the second indirect dimensions were acquired using a spectral window of 8000 Hz for direct dimension, 1376 for the hydrogen indirect dimension and 809 Hz for the nitrogen indirect dimension. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in all dimensions. Forward backwards linear prediction was used.

Stereospecific Assigments- The stereo-specific assignment of methylene protons in the guest region was done qualitatively using the cross peak intensity of the protons in question and other anchor atoms in the peptide with unambiguous chemical shift assignments. In the following section we will describe the process for several particular cases and mention what other protons were assigned using similar distance constraints. In general, several assumptions were made: i) diasterotopic protons with unique chemical shifts for each their methylene protons adopt a rigid conformation and preferentially populate a particular side-chain rotamer ii) only rotamers that avoid steric clashes with the triple helical backbone are populated and iii) only trans and gauche conformations are allowed for each c dihedral. Most NOEs used could be resolved in the 2D ^{1}H,^{1}H-NOESY spectrum, however in some cases the 3D ^{1}H,^{15}N,^{1}H-NOESY-NHSQC spectrum was used to avoid ambiguity.

The easiest diasterotopic protons to assign are the glycine α-protons. Because of the backbone dihedrals observed in the triple helix Hα3 is closer to the glycine amide proton and thus produces a stronger NOE. This fact can be used to assign all the backbone diasterotopic atoms in the guest region.

The assignment of diasterotopic β-protons was done following the procedure described by Clore *et al.* [29] using the intra-residue NOEs between the β-protons and the backbone amide and α-proton expected for the allowed values of the χ_1 dihedral. Table 2.3 shows the relative intensity expected for the different cross peaks based on the distance measured in the model for E37. If the distance between the anchor atom (either Ha or NH) and each of the b-methylene protons is similar an "~" is presented in the table for both pairs, indicating that NOEs with comparable intensities are expected. On the other hand, if one of the β-protons is closer to the anchor atom then an "**+**" is presented, signifying that a stronger NOE is expected for that pair, and "**–**" in the complementary case. The qualitative assignment of the observed NOEs is presented in the last column for comparison. Using this information is possible to assign both the conformer of the χ_1 dihedral and the identity of each of the β-protons. An identical analysis is used to assign residues E13 and E61 in the **KGE** spectra as well as D13, D37 and D61 in the **KGD** spectra.

	Gauche (+)	Trans	Gauche (-)	Observed
$\alpha - \beta_2$	**~**	**+**	**-**	**-**
$\alpha - \beta_3$	**~**	**-**	**+**	**+**
$H - \beta_2$	**-**	**~**	**+**	**+**
$H - \beta_3$	**+**	**~**	**-**	**-**

Table 2.4. Stereospecific assignments for the E37 β–protons using NOE cross peak intensities

The γ-protons in for the glutamates were assigned using a similar procedure with one aditional constraint, the χ_1 dihedral was assumed fixed at the value previously determined. In this case the anchor atoms used for the assignment are $H\beta_2$ and NH. The same conventions as above were utilized and the results for E37 are summarized in table 2.4. The same analysis for E61 leads to the same conclusions.

	Gauche (+)	Trans	Gauche (-)	Observed
$\beta_2 - \gamma_2$	-	+	~	~
$\beta_2 - \gamma_3$	+	-	~	~
$H - \gamma_2$	~	+	-	-
$H - \gamma_3$	~	-	+	+

Table 2.5. Stereospecific assignments for the E37 γ–protons using NOE cross peak intensities

The γ-protons in for the lysine residues in both peptides are more challenging to assign given the lack of unambiguous information for the χ_1 dihedral in that residue from NOE data. In order to overcome this problem we utilized geometrical constraints derived from the observed NOEs. For instance, only 4 out of the 81 rotamaters available to the lysine side-chain allow for the observed NOE Glu37(NH)-Lys11(H_ε) while simultaneously avoiding clashes with the peptide backbone in triple helical conformation. All of these four conformers have identical χ_1 and χ_2 dihedrals, in the trans conformation. Using this assumption and the relative intensity of the inter-residue Lys11($H_{g2,3}$)-Gly12(NH) NOEs the stereospecific assignment is possible. A comparable analysis using the pair of inter-strand Lys11($H_{\varepsilon1,2}$)-Glu37(NH) resonances and the intra-residue Lys11(H_{g2})- Lys11($H_{\varepsilon1,2}$) resonances can be used to assign the ε-protons.

NMR Ensemble Calculation- For the **KGE** peptide the ensemble was generated starting from its crystal structure and for the **KGD** peptide the glutamic acid residues were mutated to aspartic acid using using PyMOL[10]. Beacause not enough experimental constraints are available for a traditional NMR structure determination conformational sampling was achieved by running langevin dynamic simulations in implicit solvent for 2.5 ns at 248.15, 298.15, 348.15 and 398.15 K using the AMBER99[30] force field. Weak harmonic constraints were placed at the terminal residues[20] and parameters to bias the

hydroxyproline towards the observed ring pucker were used[31] to provide a more efficient sampling of relevant triple helical conformations. Snapshots were taken every 2.5 picoseconds along the trajectory and sorted according to their energy. The 125 lowest energy conformations from each temperature were then subjected to a minimization procedure including distance constraints derived from NOE data. The distance constraints were generated from the acquired NOESY spectra, assuming a r^{-6} proportionality between intensity and distance and using the intensity of the resonance between the acidic amide proton and the glycine HB2 proton in the middle chain together with its distance from the crystal structure as a reference. Only constraints involving the charged residues were used during the minimization procedure. The 100 lowest energy structures after the minimization step were selected for the final ensemble.

Circular Dichroism- All CD experiments were performed with a Jasco J-810 spectropolarimeter equipped with a Peltier temperature control system. 300 μM samples were prepared in 10 mM phosphate buffer at pH 7 and incubated overnight at room temperature. Spectra were acquired between 215-250 nm and the maximum around 222 nm, was monitored during unfolding curves. Melting experiments were performed from 5 to 85 °C with a heating rate of 10 °C/hr. The first derivative of the melting curve was taken in order to determine the melting temperature (T_m) of the sample, which we define as the minimum in the derivative graph. The molar residual ellipticity (MRE) is calculated from the measured ellipticity using the equation:

$$[\theta]= \frac{\theta \times m}{c \times l \times n_r}$$

where θ is the ellipticity in mdeg, m is the molecular weight in g/mol, c is the concentration in mg/mL, l is the pathlength of the cuvette in cm, and n_r is the number of amino acids in the peptide.

(1) Emsley, J.; Knight, C. G.; Farndale, R. W.; Barnes, M. J. *J. Mol. Biol.* **2004**, *335*, 1019-1028.

(2) Emsley, J.; Knight, C. G.; Farndale, R. W.; Barnes, M. J.; Liddington, R. C. *Cell* **2000**, *101*, 47-56.

(3) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2005**, *44*, 1414-1422.

(4) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 15304-15041.; Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 2683-2690.; Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7509-7515.; Russell, L. E.; Fallas, J. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2010**, *132*, 3242-3243.; O'Leary, L. E.; Fallas, J. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2011**, *133*, 5432-5443.

(5) Kramer, R. Z.; Venugopal, M. G.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *J. Mol. Biol.* **2000**, *301*, 1191-1205.

(6) Boudko, S. P.; Engel, J.; Okuyama, K.; Mizuno, K.; Bachinger, H. P.; Schumacher, M. A. *J. Biol. Chem.* **2008**, *283*, 32580-32589.

(7) Persikov, A. V.; Ramshaw, J. A.; Brodsky, B. *J. Biol. Chem.* **2005**, *280*, 19343-19349.

(8) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *J. Mol. Biol.* **2002**, *316*, 385-394.

(9) Schumacher, M. A.; Mizuno, K.; Bachinger, H. P. *J. Biol. Chem.* **2006**, *281*, 27566-27574.

(10) Delano, W. L. *The PyMOL Molecular Graphics System, Delano Scientific, San Carlos, CA, USA* **2002**,

(11) Abseher, R.; Horstink, L.; Hilbers, C. W.; Nilges, M. *Proteins* **1998**, *31*, 370-382.

(12) Morozov, A. V.; Kortemme, T.; Baker, D. *J Phys Chem B* **2003**, *107*, 2075-2090.

(13) Schwieters, C. D.; Clore, G. M. *J. Biomol. NMR* **2002**, *23*, 221-225.

(14) Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *149*, 239-244.

(15) Buevich, A. V.; Dai, Q. H.; Liu, X. Y.; Brodsky, B.; Baum, J. *Biochemistry* **2000**, *39*, 4299-4308.

(16) Jelinski, L. W.; Torchia, D. A. *J. Mol. Biol.* **1980**, *138*, 255-272.

(17) Antipova, O.; Orgel, J. P. R. O. *J. Biol. Chem.* **2010**, *285*, 7087-7096.

(18) Wu, J. J.; Woods, P. E.; Eyre, D. R. *J. Biol. Chem.* **1992**, *267*, 23007-23014.

(19) Pihlajamaa, T.; Perala, M.; Vuoristo, M. M.; Nokelainen, M.; Bodo, M.; Schulthess, T.; Vuorio, E.; Timpl, R.; Engel, J.; Ala-Kokko, L. *J. Biol. Chem.* **1999**, *274*, 22464-22468.

(20) Gurry, T.; Nerenberg, P. S.; Stultz, C. M. *Biophys. J.* **2010**, *98*, 2634-2643.

(21) Otwinowski, Z.; Minor, W. *Method Enzymol* **1997**, *276*, 307-326.

(22) Kissinger, C. R.; Gehlhaar, D. K.; Fogel, D. B. *Acta Crystallogr D* **1999**, *55*, 484-491.

(23) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr D* **1998**, *54*, 905-921.

(24) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. *Acta Crystallogr D* **2010**, *66*, 486-501.

(25) Kramer, R. Z.; Bella, J.; Brodsky, B.; Berman, H. M. *J. Mol. Biol.* **2001**, *311*, 131-147.

(26) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallogr D* **1997**, *53*, 240-255.

(27) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277-293.

(28) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, P.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. *Proteins* **2005**, *59*, 687-696.

(29) Powers, R.; Garrett, D. S.; March, C. J.; Frieden, E. A.; Gronenborn, A. M.; Clore, G. M. *Biochemistry* **1993**, *32*, 6744-6762.

(30) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

(31) Park, S.; Radmer, R. J.; Klein, T. E.; Pande, V. S. *J. Comput. Chem.* **2005**, *26*, 1612-1616.

Chapter 3: Solution NMR studies on a Designed of ABC Collagen Heterotrimer*

Most of the studies performed on collagen mimetic peptides utilize triple helices with three identical chains. Such systems are good models for some types of collagen, for example type II collagen found in cartilage.[1] However, many of the most abundant types of collagen such as type I, IV and IX are heterotrimeric species containing two (AAB) or three (ABC) different chains.[2] The Hartgerink laboratory recently introduced a robust method to prepare heterotrimeric collagen like triple helices via non-covalent interactions[3,45] (Figure 3.1). Initially, these systems were characterized through circular dichroism (CD) spectroscopy, which is a good indicator for the fold and stability of the peptides but lacks the ability to give detailed structural information.

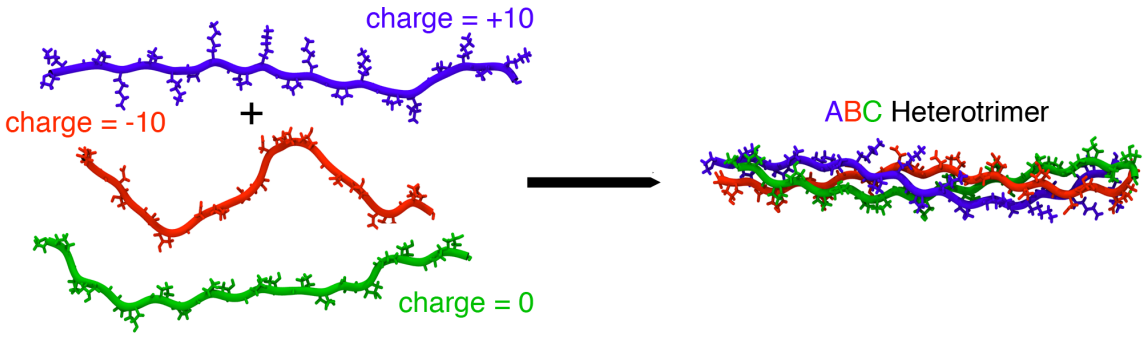


Figure 3.1. Strategy for the self-assembly of heterotrimeric CMPs.

A system of particular interest is composed of three peptides, (Pro -Lys-Gly)$_{10}$, (Asp-Hyp-Gly)$_{10}$, and (Pro-Hyp-Gly)$_{10}$, which will be abbreviated **K**, **D** and **O** respectively (Table 3.1). Upon mixture and annealing of the peptides, CD studies indicate that an ABC triple helix with the highest thermal stability reported for synthetic

---

* This chapter is largely based on the following publication:
Fallas, J. A.; Gauba, V.; Hartgerink, J. D. *J. Biol. Chem.* **2009**, *284*, 26851-26859.
The text was modified in pertinent sections to fit in the current format and highlight our improved understanding of the subject matter since its publication.

heterotrimeric collagen mimics is formed,[4] which was analyzed using solution NMR spectroscopy.

Abbreviation	Sequence
K	$(PKG)_{10}$
D	$(DOG)_{10}$
O	$(POG)_{10}$
K*	$(PKG)_4P\underline{\mathbf{KGP}}KG(PKG)_4$
D*	$(DOG)_4DO\underline{\mathbf{GD}}OG(DOG)_4$
O*	$(POG)_4PO\underline{\mathbf{GP}}OG(POG)_4$

Table 3.1. Abbreviation and chemical sequence of the peptides discussed in this chapter. Highlighted amino acids amino acids are uniformly ^{15}N,^{13}C-labeled.

3.1 Spin System Identification

The number of species present in the sample was determined from a nitrogen ^{1}H,^{15}N-HSQC experiment using the peptides **K***, **D*** and **O*** with uniformly ^{15}N, ^{13}C labeled amino acids (chemical sequences shown in Table 3.1 and spectrum in Figure 3.2). Some of the peaks can easily be identified as the monomeric forms of the highly charged **D** and **K** peptides using the information from TOCSY spectra of samples composed of each peptide separately. These are labeled M in the spectrum. The **O** peptide readily forms **O•O•O** homotrimers in solution and the presence of this species in the mixture was identified using homonuclear spectra containing exclusively this peptide.[6] That peak is labeled as $G_{\mathbf{H}}$ in the spectrum. The remaining 5 resonances correspond to the novel assembly formed when the three peptides come together and represent our subset of interest. To simplify further, the following notation will be used when referring to a particular atom: $Z_{\mathbf{N}}(B)$, where Z is the amino acid single letter code, **N** is the peptide chain (either **D**, **O**, **K** or **H** for the **O•O•O** homotrimer) and B specifies which particular atom in the amino acid is being discussed. In cases where the peptide chain is ambiguous

or amino acids from different chains are being discussed the index specifying the register is omitted. For example P(C_αH) refers to all proline alpha protons.

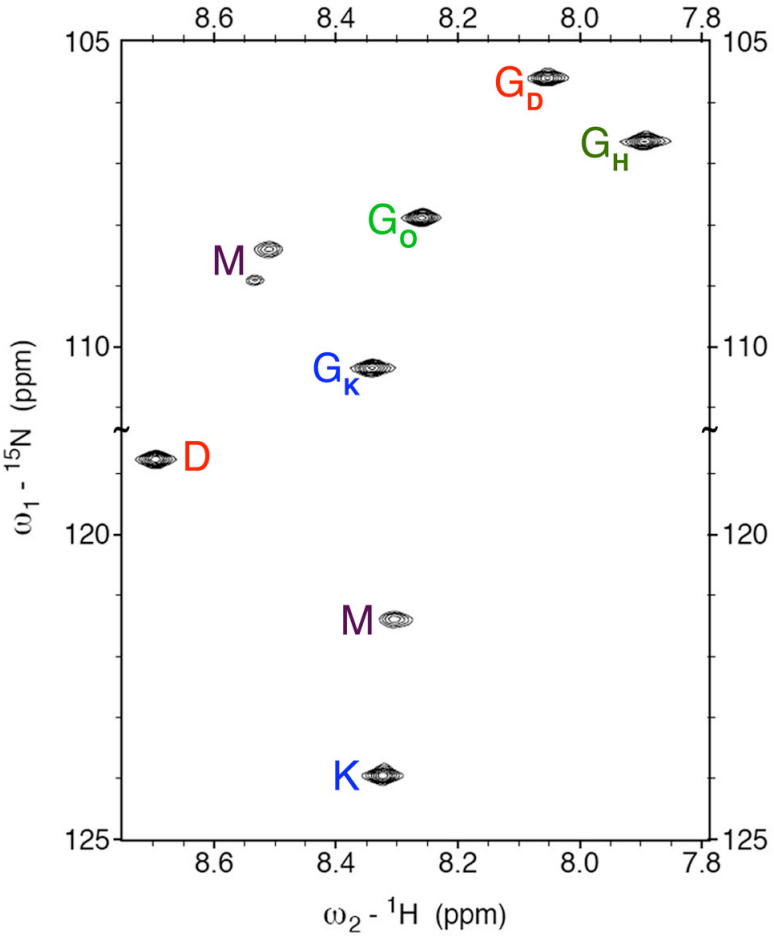


Figure 3.2. 2D ^{1}H,^{15}N-HSQC spectrum of the **K*•D*•O***. Triple helical resonances are labeled using single letter codes. For glycine, the chain it belongs to is specified as a subscript (H stands for homotrimer). Monomeric resonances are labeled M.

The spin systems belonging to each chain were determined by homonuclear sequential assignment using TOCSY and NOESY spectra at 15 and 25 °C. Intra-residue connectivity can be readily identified in the TOCSY spectra and all possible inter-residue NOEs from the NH of residue *i* to the C_αH of residue *i-1* are present. Even though they lack amide protons all P(C_αH) and O(C_αH) resonance frequencies were identified through

the NOEs with the NH of the next amino acid except for the proline on the **O** chain. In that case the sequential following of two imino acids makes it impossible to determine the $P_{\mathbf{O}}(C_{\alpha}H)$ chemical shift this way. Thus this residue was assigned based solely on the resonances present in the aliphatic region of the spectra.

Most methylene groups presented unique chemical shifts for both their protons with the exception of the γ-protons of proline and the δ- and ε-protons of the lysine. Stereospecific assignments for the methylene groups with non-degenerate chemical shifts for the proline and hydroxyproline residues were carried out using the NOE intensities of the crosspeaks between the β-, δ- and α-protons and the β-, δ- and γ-protons, respectively. Due to conformational restrictions placed on the methylene groups by the proline rings, these assignments are straightforward. In the case of the α-protons of the glycine residues, a combination of NOE data and the crosspeak intensity in the HNHA spectrum was used. A similar approach was taken for the β-protons of lysine and aspartic residues but using the information from HNHB spectrum instead. The γ-protons of the lysine residues were assigned exclusively based on NOE crosspeaks intensity.

3.2 Assessment of Triple Helical Topology

The amide region of the NOESY spectrum (Figure 3.3) shows a set of resonances at the chemical shifts corresponding to the position of the $G_{\mathbf{K}}(NH)$, $G_{\mathbf{D}}(NH)$, $G_{\mathbf{O}}(NH)$, $G_{\mathbf{H}}(NH)$, D(NH) and K(NH) peaks in the ^{1}H-dimension of the $^{1}H,^{15}N$-HSQC spectrum, indicating that most amino acids in each peptide show an ordered structure that is very similar to that of the central triplets, thus having identical chemical shifts. This phenomenon is characteristic of triple helical peptides, where the majority of the triplets

show an identical chemical environment due to the symmetry of the helix. However, in order to assure that are our peptides are indeed folded in a triple helical conformation, we need to compare the NOEs observed to those expected from a triple helix. To this effect, we compared all unique inter-chain NOEs expected for the **O•O•O** homotrimer[6] and our heterotrimer (Table 3.2). We were able to find analogous peaks and although not all of them can be resolved unequivocally due to the overlap of the $P_{\mathbf{O}}(C_{\gamma}H)$, $O_{\mathbf{O}}(C_{\beta}H_1)$ and $O_{\mathbf{D}}(C_{\beta}H_1)$ resonances, this comparison gives us confidence that the observed NOEs are indeed consistent with a triple helical fold of the peptides.

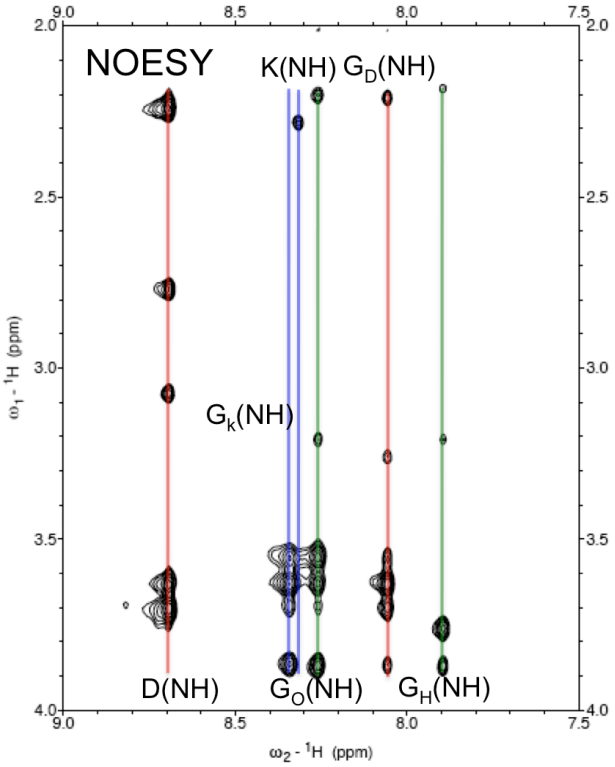


Figure 3.3. Amide region from the 2D NOESY spectrum of the mixture K/D/O. The vertical lines denote the chemical shifts of the different amide resonances using the nomenclature described in the text.

Due to the symmetry breaking induced by the heterogeneity in the chemical composition of our assembly, some of the resonances that are degenerate in the case of a

homotrimer can be easily resolved in our system. Such resonances are of interest because they are not only characteristic of a collagen triple helix but also demonstrate that the three chains are in close proximity, as expected for an ABC heterotrimer.

Expected Homotrimer NOE	Observed Heterotrimer NOE
G(NH)-P($C_\delta H_1$)	$G_{\mathbf{D}}$(NH)-$P_{\mathbf{O}}(C_\delta H_1)$#, $G_{\mathbf{O}}$(NH)-$P_{\mathbf{K}}(C_\delta H_1)$#
O($C_\delta H_2$)-P($C_\delta H_1$)	$O_{\mathbf{O}}(C_\delta H_1)$-$P_{\mathbf{K}}(C_\delta H_1)$, $O_{\mathbf{D}}(C_\delta H_1)$-$P_{\mathbf{O}}(C_\delta H_1)$
O($C_\alpha H$)-P($C_\delta H_1$)	$O_{\mathbf{O}}(C_\alpha H_1)$-$P_{\mathbf{K}}(C_\delta H_1)$, $O_{\mathbf{D}}(C_\alpha H_1)$-$P_{\mathbf{O}}(C_\delta H_1)$
O($C_\alpha H$)-P($C_\delta H_2$)	$O_{\mathbf{O}}(C_\alpha H_1)$-$P_{\mathbf{K}}(C_\delta H_2)$, $O_{\mathbf{D}}(C_\alpha H_1)$- $P_{\mathbf{O}}(C_\delta H_2)$
G($C_\alpha H_1$)- P($C_\alpha H$)	$G_{\mathbf{O}}$ ($C_\alpha H_1$)- $P_{\mathbf{K}}(C_\alpha H)$
P($C_\beta H_1$)-O($C_\beta H_2$)	$P_{\mathbf{K}}(C_\beta H_1)$-$O_{\mathbf{O}}(C_\beta H_2)$*,$P_{\mathbf{O}}(C_\beta H_1)$-$O_{\mathbf{D}}(C_\beta H_2)$*
P($C_\gamma H_1$)-O($C_\alpha H$)	$P_{\mathbf{K}}(C_\gamma H)$-$O_{\mathbf{O}}(C_\alpha H)$
P($C_\gamma H_1$)-O($C_\beta H_1$)	$P_{\mathbf{O}}(C_\gamma H)$-$O_{\mathbf{D}}(C_\beta H_1)^+$
P($C_\gamma H_1$)-O($C_\beta H_2$)	$P_{\mathbf{O}}(C_\gamma H)$-$O_{\mathbf{D}}(C_\beta H_2)^+$
P($C_\gamma H_2$)-O($C_\beta H_2$)	$P_{\mathbf{O}}(C_\gamma H)$-$O_{\mathbf{D}}(C_\gamma H)^+$
P($C_\gamma H_1$)-O($C_\gamma H$)	$P_{\mathbf{O}}(C_\gamma H)$-$O_{\mathbf{D}}(C_\gamma H)^+$
P($C_\delta H_1$)-O($C_\beta H_2$)	$P_{\mathbf{K}}(C_\delta H_1)$-$O_{\mathbf{O}}(C_\beta H_2)$, $P_{\mathbf{O}}(C_\delta H_1)$-$O_{\mathbf{D}}(C_\beta H_2)$
P($C_\delta H_2$)-O($C_\beta H_2$)	$P_{\mathbf{K}}(C_\delta H_2)$-$O_{\mathbf{O}}(C_\beta H_2)$, $P_{\mathbf{O}}(C_\delta H_2)$-$O_{\mathbf{D}}(C_\beta H_2)$

#Highlighted in Figure 2; *Overlapping peaks; $^+P_{\mathbf{O}}(C_\gamma H)$ overlaps with $O_{\mathbf{O}}(C_\beta H_1)$ and $O_{\mathbf{D}}(C_\beta H_1)$

Table 3.2 Expected interchain NOEs based on a POG homotrimer model and observed interchain NOEs for our heterotrimer.

A set of cross peaks that is very illustrative of these two facts is the one arising from the G(NH) to the G($C_\alpha H$) protons. Besides the intra-chain NOEs identified in the TOCSY spectrum, a set of inter-strand NOEs is present (Figure 3.4a). Although all possible G(NH)-G($C_\alpha H$) inter-chain correlations are present, some of them overlap and only the ones that can be unambiguously assigned are highlighted. These peaks confirm the spatial proximity of the α-protons and the amide protons on all the chains in the core of the helix, as expected from the collagen model (Figure 3.4b). The same set of resonances

cannot be observed in the homotrimer spectrum as the G(NH) and G(C_αH) chemical shifts are indistinguishable between the different chains.

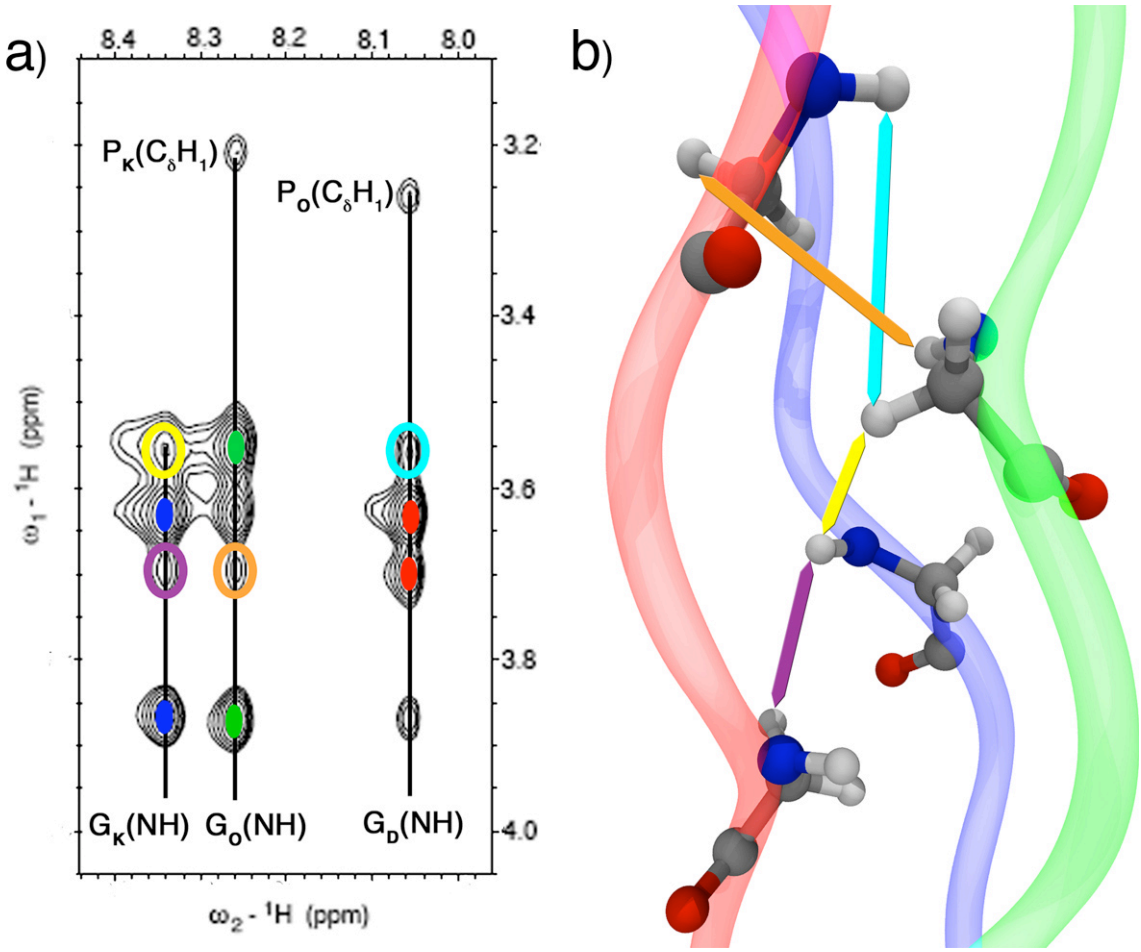


Figure 3.4 2D NOESY spectrum and molecular model. a) The NOESY spectrum shows the resonances from the glycine amide hydrogen to the α-protons. The vertical lines mark the NH chemical shifts, the solid colored ellipses the position of the intra-strand cross peaks (red stands for **D**, green for **O**, blue for **K**) and the hollow ellipses unambiguous inter-strand interactions. The P($C_\delta H_1$)-G(NH) cross peaks are also shown. b) Model highlighting the atoms that that give rise to the inter-stand cross peaks in the spectrum using colored arrows.

Another way to probe the conformation of peptides in solution is to measure the $^3J_{HNH\alpha}$ coupling constant since it can be directly linked to the φ backbone dihedral angle via the Karplus relation. We measured the coupling constant for our heterotrimeric helix using the HNHA experiment and for the residual **O•O•O** homotrimer in our system. The

values obtained range from 4 to 7 Hz and agree with previously measured values for homotrimeric triple helices.[7] Table 3.3 shows a comparison between the φ dihedral angles computed from these values and those of a high resolution **O•O•O** crystal structure[8] and 2 model helices, one with 7/2 symmetry[9] and one with 10/3 symmetry.[10] The angles obtained for aspartic acid and lysine agree with those expected from amino acids in the X and Y position of a collagen triple helix. The values obtained for the glycines of all three chains also agree with those determined for the homotrimer using X-ray crystallography and our NMR measurements.

	Heterotrimer			Homotrimer		
	PKG	DOG	POG	POG[a]	7/2[b]	10/3[c]
X	---	-72±6	---	-73±4	-76	-72
Y	-63±10	---	---	-58±5	-63	-75
G	-80±10	-80±20	-79±10	-75±6	-70	-67

[a]Nagarayan et al. (1999) crystal srtructure at 1.9 Å resolution;
[b] Okuyama etal. (1980) model; [c]Fraser et al. (1979) model.

Table 3.3. Dihedral angles calculated from the $^{3}J_{HNH\alpha}$ coupling constants using Karplus equation. Values from a high resolution crystal structure of the **O•O•O** homotrimer and two models with different helical symmetry are included for comparison.

The ratio between the homotrimeric and heterotrimeric species was determined using the peak intensity observed for each triple helix in the $^{1}H,^{15}N$-HSQC experiment. For the **K*/D*/O*** sample, which contains a 1:1:1 mixture of the peptides, the ratio of heterotrimer to homotrimer is approximately 3 to 1. Changing the relative amount of one of the peptide strands in the mixture can shift the equilibrium towards the formation of the heterotrimer. This was observed in the **K/D/O*** sample, which contains an excess of the K peptide (1.5:1:1), and has a ratio of the heterotrimer to homotrimer of approximately 11 to 1. Because all peptide concentrations were determined by mass, the

actual ratio of peptides in the mixture is most likely skewed thus the reported values are only intended for a qualitative comparison.

3.3 Chain Registration

Because we observe only one set of resonances for each amino acid type in each peptide chain of the heterotrimeric triple helix, we hypothesize that that the peptides preferentially assemble in only one of the six possible registers (Figure 3.5). The following analysis is strongly dependent on this hypothesis. In general, this is a robust statement and a more heterogeneous spectrum would normally be expected for a sample containing several distinct supramolecular assemblies.[11] Nonetheless, the high symmetry of the heterotrimers in this system, the rigidity of the triple helical structure and the repetitive nature of the sequences may cause three of the six registers to have identical resonant frequencies. However, we cannot prove that this is the case with our current experimental setup and thus the initial hypothesis will be assumed to be valid.

Under this assumption and in order to determine which register is formed, we built a homology model for each possible assembly and compared the NOEs observed with those expected from each model. A set of resonances that is very useful when analyzing this problem arises between the K($C_\alpha H_1$)-D(NH), K($C_\varepsilon H$)-D(NH) and Go($C_\alpha H_1$)-D(NH) protons. These are depicted in the strip corresponding to the NOESY spectrum shown in Figure 3.6a (labeled N). Figure 3.6b summarizes the expected results of this experiment for each register. When an inter-proton distance less than 5 Å is observed for any of the aforementioned pairs in the model, an ✖ is placed in the column corresponding to the NOESY spectrum (N) of that register; otherwise an O is places in that spot. The result of

the actual NMR experiment is summarized in the last row of the table, where an ✖ has been placed for each of the observed resonances. Any inconsistencies between the spectra and the models are highlighted in red. Using this comparison, three registers (**D•K•O**, **O•D•K** and **K•O•D**) can be discarded. However, due to the periodic nature of our sequences, unambiguously determining the register using only 2D NOESY experiments is not possible.

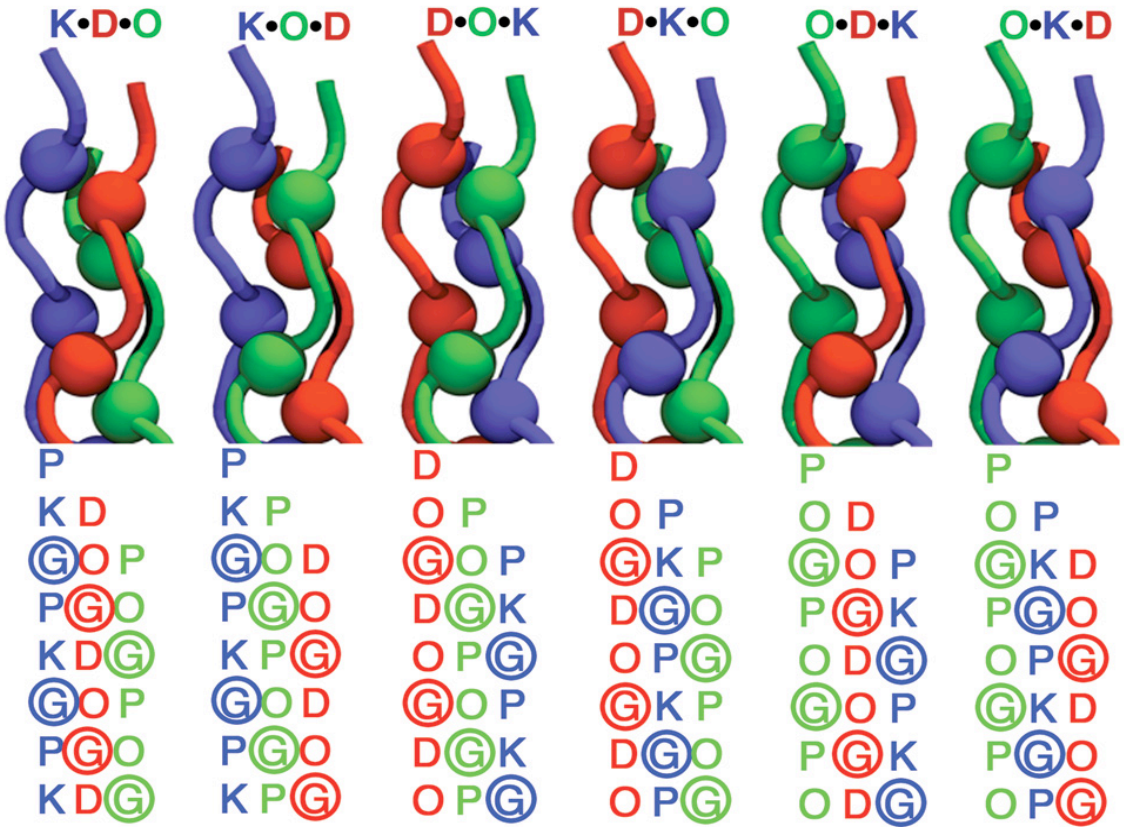
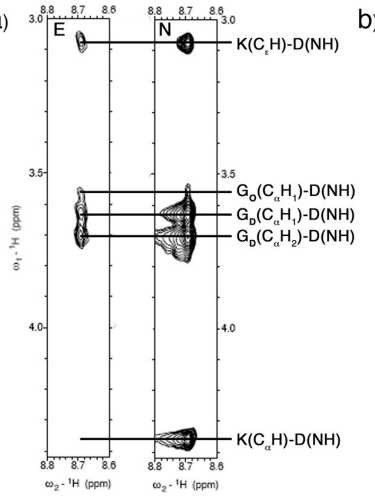


Figure 3.5. Schematic N-terminal representation of the six possible registers for the heterotrimeric triple helix. The difference in the sequence is highlighted under each representation, where the position of glycine residues is marked by colored spheres.

In order to distinguish between the three remaining registers, knowledge about which triplet along the sequence of the amino acids gives rise to these resonances is required. That is, we need to know if the ε-protons of lysine in triplet *n* are close to the

amide proton of aspartic acid in triplet *n, n+1* or *n+2*. To obtain this information, we used a 2D $^{13}C,^{15}N$-edited NOESY spectrum (refer to the Experimental section for details), where the observed resonances occur only between labeled amino acids (K14, D16, G**O**15). A strip of the spectrum corresponding to the D(NH) chemical shift is shown in Figure 3.6a (labeled E). The main difference between the edited and regular spectrum is the absence of the K($C_\alpha H_1$)-D(NH) peak. This means that the α-proton of K14 (fifth triplet, **K** chain) is not close to amide proton of D16 (sixth triplet, **D** chain). Meanwhile, the presence of the K($C_\varepsilon H$)-D(NH) resonance indicates that the ε-protons of K14 are near the amide proton of D16 and the G**O**($C_\alpha H_1$)-D(NH) peak that D16 is within 5 Å of G**O**15 (fifth triplet, **O** chain). Using the same convention as before, the results of the experiment expected for each register are summarized in Figure 3.6b under the column corresponding to the edited NOESY spectrum (E). Due to the arrangement of the chains, the K($C_\alpha H_1$)-D(NH) resonance would be expected instead of K($C_\varepsilon H$)-D(NH) for the **D•O•K** register and the G**O**($C_\alpha H_1$)-D(NH) peak should be absent for **O•K•D** according to the model. This comparison yields, in agreement with our hypothesis based on the number of peaks seen in the spectra, only one register: **K•D•O**.

a)



b)

	$K(C_\alpha H_1)$-D(NH)		$K(C_\varepsilon H)$-D(NH)		$G_O(C_\alpha H_1)$-D(NH)	
	E	N	E	N	E	N
D•K O	✖	✖	✖	✖	O	O
O•D•K	O	O	O	O	O	O
K•O•D	O	O	O	O	O	O
D•O•K	✖	✖	✖	✖	O	✖
O•K•D	O	✖	O	✖	✖	✖
K•D•O	O	✖	✖	✖	✖	✖
Observed	O	✖	✖	✖	✖	✖

Figure 3.6 2D Edited NOESY (E) and NOESY (N) spectra of **K•D•O**. a) Strips from both experiments corresponding to the D(NH) chemical shift. b) Table showing the expected outcome of both spectra for the six possible registers of the assembly according to our models. The ✖ indicates that a peak should be observable in the spectrum and an **O** that no peak is expected. The last row summarizes the results of the strips on the right. Inconsistencies between the models and the spectra are highlighted in red.

3.4 Solution Structure

With knowledge about the register, the NOEs observed can be assigned to proton pairs (or groups in the case of overlapping methylene resonances) along the chemical sequence of the peptides and, together with the constraints obtained from the HNHA and HNHB experiments, used to calculate an a ensemble of structures that are representative of the solution conformation of the triple helix. A summary of the constraints and structural statistics is provided in Table 3.4 and details about the protocol used for structural determination are given in the methods section.

NMR distance and dihedral constraints	
Distance constraints	
Total NOE	771
Intra-residue	253
Inter-residue	518
Sequential ($\|i - j\| = 1$)	180
Interchain	338
Total dihedral angle restraints	120
ϕ	72
$\chi 1$	48
Structure statistics	
Violations (mean and s.d.)	
Distance constraints (Å)	0.07 ± 0.05
Dihedral angle constraints (º)	1.86 ± 1.65
Max. dihedral angle violation (º)	4.97
Max. distance constraint violation (Å)	0.244
Deviations from idealized geometry	
Bond lengths (Å)	0.0097 ±0.0001
Bond angles (º)	2.34 ± 0.04
Average pairwise r.m.s. deviation, 15 structures (Å)	
Heavy	0.68
Backbone	0.53

Table 3.4. NMR and refinement statistics.

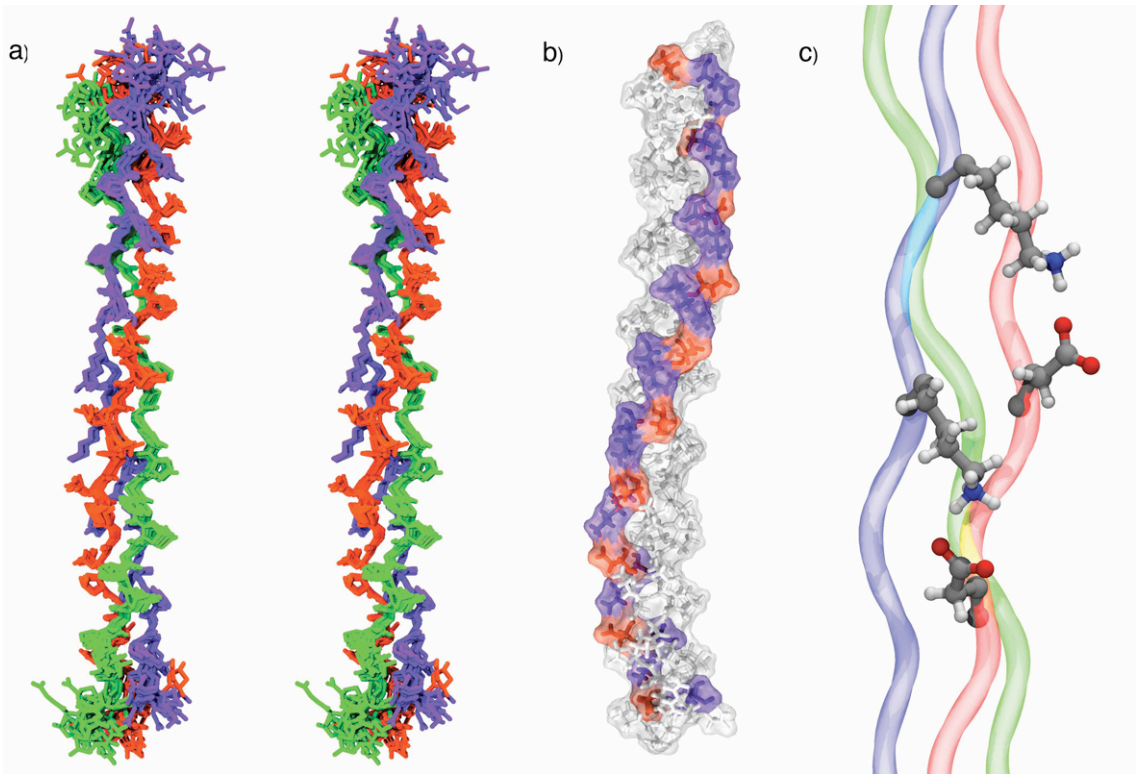


Figure 3.7. NMR structure of **K•D•O**. a) Stereo view of the superposition of the 10 lowest energy conformers. The **K** peptide is blue, **D** is red and **O** is green. b) Surface and CPK representation of the lowest energy conformer highlighting the position of charged amino acids, lysine in red and aspartic acid in blue, along the triple helix. c) Expanded view of two of the salt bridges observed in b) with distinct conformations.

The backbone of our refined NMR structure behaves in a similar way to the homotrimeric system. Most of the points in the Ramachadran plot for the ensemble (Figure 3.7) are grouped in a narrow region corresponding to the poly-proline type II helix and only the unconstrained residues populate different regions of the (φ,ψ) space. The hydrogen-bonding network along the backbone of the peptides, which goes from the carbonyl of the amino acid in position X to the amide proton of glycine in a neighboring chain, is also conserved, although no explicit hydrogen bonding-type restraints were used during the refinement procedure. The helical pitch can very hard to determine by NMR because of its long-range character, but the coupling constants measured indicate that our helix is probably closer to a 7/2 helix than to a 10/3 helix, like the **O•O•O** homotrimer.

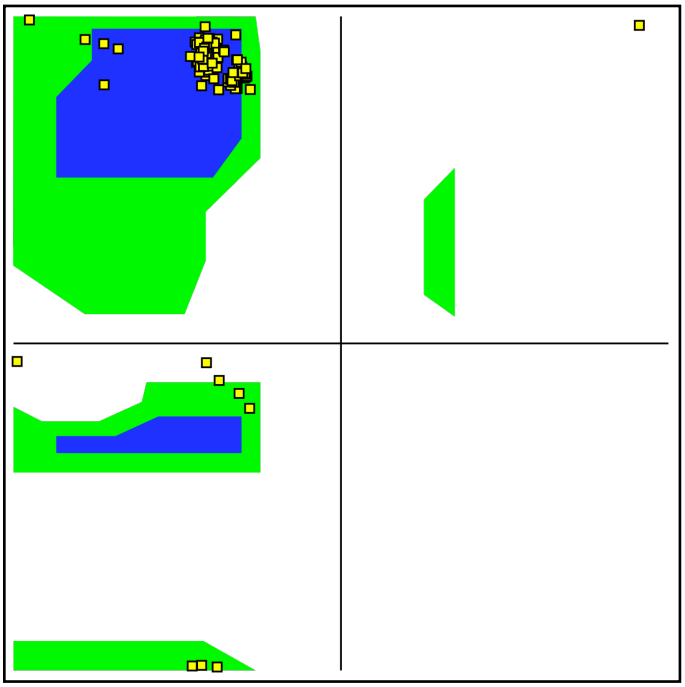


Figure 3.8 Representative Ramachandran plot of the conformers from our NMR ensemble.

The most interesting feature of the structure is the side-chain conformation of the charged aspartic acid and lysine residues, which form a network of ionic hydrogen bonds spiraling along the helical axis (Figure 3.8b), following the helicity of a single peptide strand within the triple helix with an axial repeat of approximately 60 Å.[12] These ionic hydrogen bonds are formed exclusively between lysine in triplet *n* of one peptide chain and aspartate in triplet *n+1* of the adjacent peptide, which corresponds to the axial salt bridge geometry identified in the previous chapter. The salt bridges formed are highly dynamic, which can be seen in the structure of the lysine side chain resonances. The β- and γ-protons show two distinct chemical shifts, in contrast to the γ- and δ-methylene groups, which present only a single chemical shift. This indicates that the χ1 and χ2 dihedral angles have well defined values, but the χ3 and χ4 dihedrals sample a wider

variety of conformations, similar to what we found for the Asp-Lys pairs in the previous chapter. Due to the length of the lysine side chain, the different conformers are still able to interact effectively with aspartic acid, which is primarily locked in a single conformation (Figure 3.7c).

3.5 Conclusions

This study represents one of the first attempts to elucidate the molecular mechanisms of triple helical recognition in heteortrimeric systems. We show that specific amino acid interactions, rather than overall electrostatic complementarity, are the main reason behind the surprising stability of this designed heterotrimeric system. Furthermore, we are able to see the formation of axial salt bridges, which we described in the previous chapter, using a solution NMR approach. This is relevant because the symmetric nature of the designed primary structure allows for the formation of both axial and lateral contacts.[13] The lateral contacts are plausible between amino acids in identical triplets and their structural proximity is validated in solution by the presence of backbone NOEs between the charged residues, which are not possible in an axial arrangement. Although it is possible that the availability of both states introduces some of the structural heterogeneity observed in the side chain geometry of the interacting amino acids, their preference for an axial geometry can be observed in the solution structure.[14]

Despite the formation of a high stability ABC heterotrimer, this system contains some residual homotrmeric $(POG)_{10}$ helices. Although this result is not surprising given the higher thermal stability of the homotrimeric helix, it makes the system less attractive from the perspective of future applications, particularly in the realm of structural biology.

Nonetheless, the system is able to avoid the formation of several other competing species, including some lower stability AAB heterotrimers that fold in the absence of the last complementary chain. This result is encouraging and the following chapters will build upon this initial specificity towards ABC heterotrimers in order to improve it and gain a better understanding of the rules that govern the self-assembly of heretrimeric collagens.

3.6 Experimental

Peptide Synthesis and Purification of the labeled peptides - The **D***, **O*** and **K*** peptides were synthesized with an Advanced Chemtech Apex 396 solid phase peptide synthesizer using standard Fmoc chemistry and a Rink MBH amide resin on a 0.05 mM scale. The unlabeled amino acids were added in a 4:1 molar ratio to the peptide chain with HBTU/HoBT as activating agents and a coupling time of 45 minutes. The uniformly labeled amino acids were purchased form Cambridge Isotope Laboratories and added in a 1.5:1 molar ratio to the peptide chain. For these steps, HATU was used as the activating agent and the coupling reaction was carried out for 4 hours. The peptides were protected at the N-terminus using acetic anhydride and cleaved from the resin with a 38:1:1 mixture of TFA, triisopropylsilane and water yielding an amide C-terminus. Purification was done on a Varian PrepStar220 HPLC using a preparative reverse phase C-18 column with a linear gradient of water and acetonitrile gradient each containing 0.5% TFA. HPLC fractions were analyzed by MALD/TOF mass spectrometry on a Bruker autoflex II using a pre-spotted anchor chip with a-cyano-4-hydroxycinnamic acid as the matrix.

All NMR experiments were performed on an 800 MHz Varian spectrometer equipped with a cryogenic probe.

NMR characterization of the D and K peptides – 2.4 mg of **D** and 2.4 mg of **K** were each dissolved separately in 560 μL of H_2O and mixed with 70 μL 100 mM phosphate buffer and 70 μL D_2O to afford a 1.2 mM sample. TSP was used as an internal proton standard. TOCSY spectra with a 75 ms spinlock were acquired for each sample on an 800 MHz Varian spectrometer equipped with a cryogenic probe at 25 °C. A total of 1918 complex points were recorded in 16 scans for the directly acquired dimension and 320 increments were recorded in the sates mode for the indirect dimension. A spectral width of 9600 Hz was used in both dimensions. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in both dimensions.

NMR characterization of the K/D/O mixture – 2.3 mg of **O** and 2.4 mg of **K** were dissolved together in 200 μL H_2O and 2.4 mg of **D** was dissolved in 70 μL 100 mM phosphate buffer and 130 μL H_2O affording a 4.2 mM aqueous solution for each peptide. After mixing them, heating to 85 °C for 15 minutes and incubating at room temperature for at least 18 hours, 70 μL D_2O, 30 μL H_2O and a proton standard (TSP) were added. The final concentration for each strand was 1.2 mM and the total peptide concentration in the sample was 3.6 mM. TOCSY spectra with a 75 ms and a 10 ms spinlock duration were recorded at 15 °C and 25 °C. A total of 1918 complex points were recorded in 16 scans for the directly acquired dimension for all spectra while 560 increments were used in the indirect dimension for the 25 °C data and 320 increments were used for the 15 °C

data. NOESY spectra with a 75 ms mixing time were recorded at 15 °C and 25 °C. A total of 3271 complex points were recorded in 32 scans for the directly acquired dimension and 480 increments in the indirect dimension. A square spectral window of 9600 Hz was used for all spectra.

NMR characterization of the K/D*/O* mixture* – A sample was prepared using a similar methodology as described above. $^{1}H,^{13}C$- and $^{1}H,^{15}N$-HSQC, HNHA, HNHB and a 2D ^{13}C, ^{15}N-edited NOESY experiments were performed on an 800 MHz Varian spectrometer at 25 °C. A total of 630 complex points in 16 scans for the direct dimension and 400 increments in the indirect dimension were acquired for the $^{1}H,^{15}N$-HSQC using a spectral window of 8000 Hz in the hydrogen dimension and 4000 Hz in the nitrogen dimension. A total of 961 complex points in 24 scans for the direct dimension and 450 increments in the indirect dimension were acquired for the $^{1}H,^{13}C$- HSQC using a spectral window of 9600 Hz in the hydrogen dimension and 18100 Hz in the carbon dimension. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in both dimensions. Forward backwards linear prediction was used to improve the resolution in the heteroatom dimension of both spectra. For the 3D HNHA spectrum a total of 682 complex points in 16 scans for the direct dimension, 120 increments for the first indirect dimension and 18 increments for the second indirect dimensions were acquired using a spectral window of 8000 Hz for direct dimension, 48000 for the hydrogen indirect dimension and 2432 Hz for the nitrogen indirect dimension. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in all dimensions. Forward

backwards linear prediction was used to improve the resolution in the indirect hydrogen dimension. For the 3D HNHB spectrum a total of 1024 complex points in 4 scans for the direct dimension. 100 increments for the first indirect dimension and 30 increments for the second indirect dimensions were acquired using a spectral window of 12000 Hz for direct dimension, 8000 for the hydrogen indirect dimension and 3243 Hz for the nitrogen indirect dimension. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in all dimensions. A total of 1024 complex points in 96 scans for the direct dimension and 200 increments in the indirect dimension were acquired for the 2D ^{13}C, ^{15}N-edited NOESY using a spectral window of 1150 Hz and 8000 Hz.

NMR characterization of the K/D/O mixture* – A sample was prepared using a similar methodology as described above but using a 1.5:1:1 ratio of the peptides (**K:D**:**O***). In this case the D and O concentration was 1.2 mM and the K concentration was 1.8 mM making the total peptide concentration 4.2 mM. A ^{15}N-HSQC experiment was performed on an 800 MHz Varian spectrometer at 25 °C. A total of 1024 complex points in 8 scans for the direct dimension and 180 increments in the indirect dimension were acquired using for a spectral window of 13000 Hz in the hydrogen dimension and 2432 Hz in the nitrogen dimension. The data was processed by zero filling to the next power of two and cosine bell apodization functions were applied in both dimensions. Linear prediction was used to improve the resolution in the indirect dimension of both spectra.

Details about NMR experiments – 2D ^{13}C, ^{15}N-edited NOESY. It corresponds to a 2D version of a 4D ^{1}H,^{13}C-HMQC-NOESY-^{1}H,^{15}N-HSQC experiment was recorded at 25 °C;[15] to ease further discussion of this experiment it will be referred to as a 2D ^{13}C, ^{15}N-edited NOESY. The initial ^{1}H,^{13}C-heteronuclear multiple quantum coherence (HMQC) pulse train selectively creates magnetization on protons attached to labeled carbon atoms, at the end of which a mixing time is allowed for NOE buildup; the transferred magnetization is then read through an ^{1}H,^{15}N-HSQC pulse sequence. Our experiment was carried out keeping both heteroatom evolution times (t1 and t3) constant, yielding a 2D NOESY spectrum where cross peaks between hydrogens that are directly to bonded to ^{13}C nuclei and hydrogens that are directly bonded to ^{15}N nuclei are observed. 3D HNHA. For this experiment, the ratio between the intensity of the cross and diagonal peaks can be used to compute the $^{3}J_{HNH\alpha}$ value. A systematic error is incurred in the measurement because the anti-phase magnetization that gives rise to the cross peaks relaxes at a faster rate than the in-phase component which gives rise to the diagonal peak, attenuating the ratio and underestimating the coupling constant. A correction factor of 1.16 has been previously determined for a collagen mimetic peptide to account for this phenomenon.[7] Because this factor is a function of the rotational correlation time, τ_c,[16] we have applied the same correction to the values reported here. The experimental error in the measurement comes from the integration of the peaks, which was done by fitting the peak profile to a Gaussian shape using a least squares procedure. The error in the integration was estimated by the residual of the fit. The residual was calculated by the sum of the squares of the difference of the value of the peak and the value of the best fit at each point. Using this approach the minimum and maximum possible intensity ratios

were computed and used to calculate the error bars on the ^{3}J values. The errors where then propagated when solving the Karplus equation[16] as parameterized in reference 16 to obtain the φ backbone dihedral angles.

Molecular Modeling – Homology models were built starting from the crystal structure of a triple helical peptide (pdb identifier: 1k6f).[17] The necessary sequence changes were made using PyMOL[18] to generate a preliminary structure for each of the six possible registers. Each structure was then minimized using the AMBER99 force field[19] with implicit water (generalized Born approximation). Additional force field parameters to account for the stereo electronic effects of the hydroxyl group on the proline side chain conformation were included.[20] Short constant temperature Langevin dynamics runs at 300K, 200K and 100K were used within the minimization algorithm in order to equilibrate the structures and obtain low energy conformers.

Conformational Restraints and Structure Calculation – Distance restraints were generated from the 2D NOESY experiments. The peaks were mapped onto the shortest stretch of the chemical sequence that could unambiguously accommodate all inter- and intra-strand resonances (PKGPKG for **K**, OGDO for **D** and POG for **O**). A qualitative approach was taken and the peaks were divided into four categories (very strong, strong, medium and weak) according to their intensity. Three types of dihedral restraints were used in the calculations. The φ backbone dihedral angles for K, D and G were constrained using the values obtained from the HNHA experiment, the χ_1 angle of K and D were loosely constrained based on the results of the HNHB experiment and the the χ_1 and φ

dihedral of proline and hydroxyproline were constrained according to the ring puckering of the side chain, as determined by the intensity ratio of the β- and δ-protons.[6] Details on the restraints are available in the supplementary methods. The restraints were propagated along the sequence from triplet 2 to triplet 9, leaving the N- and C-terminal triplets unconstrained since those amino acids have been shown to populate a less ordered conformation in homotrimeric triple helices.[6] Distance restraints. Intensity categories defined by peak volume: very strong (2.0 - 2.5 Å), strong (2.2 -2.8 Å), medium (2.5–3.5 Å) and weak (2.8-5.0 Å). Dihedral restraints. From the HNHA experiments see table 2 in the text; from the HNHB experiment the χ_1 angle of K (180° ± 40°) and D (-60° ± 40° for D); from proline side ring pucking the the χ_1 and φ values of P (χ_1 = 19° ± 30°, φ = -72° ± 30°) and O (χ_1 = -6° ± 30°, φ = -58° ± 30°).

Structure calculations were done using cycles of simulated annealing (SA) followed by a refinement in implicit solvent. In the SA stage 300 trial structures were calculated using a combination of torsional and cartesian dynamics with the standard protocol available in the Crystallography and NMR System (CNS) software.[21] The refinement stage was done in AMBER99, performing a minimization in implicit solvent subjected to the same constraints utilized in the SA stage on the 150 conformers that showed the lowest CNS target function. In the initial cycle, structure calculations were started from extended polypeptide chains and only backbone dihedral constrains were used. The minimum energy conformer was then used to start a new cycle, in which only cartesian dynamics were used in the SA stage, but all the constraints available were included. The 15 conformers with the lowest energy, as calculated by AMBER, were then selected for the final ensemble.

3.7 References

(1) Brodsky, B.; Thiagarajan, G.; Madhan, B.; Kar, K. *Biopolymers* **2008**, *80*, 345-353.

(2) Ricard-Blum, S. *CSH. Perspect. Biol.* **2011**, *3*, ARTN a004978.

(3) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 2683-2690.; Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7509-7515.

(4) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 15304-15041.

(5) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 2683-2690.

(6) Li, M.-H.; Fan, P.; Brodsky, B.; Baum, J. *Biochemistry* **1993**, *32*, 7377-7387.

(7) Li, Y.; Brodsky, B.; Baum, J. *J. Biol. Chem.* **2008**, *282*, 22699-22706.

(8) Nagarajan, V.; Kamitori, S.; Okuyama, K. *J. Biochem.* **1999**, *125*, 310-318.

(9) Okuyama, K.; Arnott, S.; Takajagani, M.; Kakudo, M. *J. Mol. Biol.* **1981**, *152*, 427-443.

(10) Fraser, R. D.; MacRae, T. P.; Suzuki, E. *J. Mol. Biol.* **1979**, *129*, 463-481.

(11) O'Leary, L. E.; Fallas, J. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2011**, *133*, 5432-5443.; Slatter, D. A.; Foley, L. A.; Peachey, A. R.; Nietlispach, D.; Farndale, R. W. *J. Mol. Biol.* **2006**, *359*, 289-298.

(12) Okuyama, K.; Wu, G.; Jiravanichanun, N.; Hongo, C.; Noguchi, K. *Biopolymers* **2006**, *84*, 421-432.

(13) Fallas, J. A.; Dong, J.; Tao, Y. J.; Hartgerink, J. D. *J. Biol. Chem.* **2012**, *287*, 8039-8047.

(14) Fallas, J. A.; Gauba, V.; Hartgerink, J. D. *J. Biol. Chem.* **2009**, *284*, 26851-26859.

(15) Muhandiram, D. R.; Guang, Y. X.; Kay, L. E. *J. Biomol. NMR* **1993**, *463-470*, 463-470.

(16) Vuister, G. W.; Bax, A. *J. Am. Chem. Soc.* **1993**, *115*, 7772-7777.

(17) Berisio, R.; Vitagliano, L.; Mazzarella, L.; Zagari, A. *Protein Sci.* **2002**, *11*, 262-270.

(18) Delano, W. L. *The PyMOL Molecular Graphics System, Delano Scientific, San Carlos, CA, USA* **2002**,

(19) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668-1688.

(20) Park, S.; Radmer, R. J.; Klein, T. E.; Pande, V. S. *J. Comput. Chem.* **2005**, *26*, 1612-1616.

(21) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr. D* **1998**, *54*, 905-921.

Chapter 4: Rational Design of Single Composition ABC Heterotrimers*

In order to improve upon the specificity towards ABC heterotrimers within our zwitter-ionic system described in the previous chapter, the stability of the **O•O•O** homotrimer, the main competing state, needs to be reduced. Utilizing the sequence-structure relationship between the positively charged amino acid lysine and the negatively charged amino acids glutamate and aspartate described in chapter 2, we also rationally engineer aditional salt-bridges to widen the stability gap between the desired target state and the competing homotrimer. This multi-state approach to rational design results in the self-assembly of high-stability single composition ABC-type collagen heterotrimers in which the designed interactions are satisfied.

4.1 Peptide Design

Work by Brodsky et al. shows that the substitution of hydroxyproline by any other amino acid leads to a loss of thermal stability in the resulting triple helix.[1] This element of negative design is particularly attractive for our purposes since homotrimeric assemblies will include three times the number of substitutions relative to their heterotrimeric counterparts. Peptide **C** (Table 4.1) is based on the $(POG)_{10}$ template but includes two substitutions in the Y position of the 2nd and 7th triplets (O7K and O22K). This peptide, as well as all others in this chapter, includes an N-terminal tyrosine to facilitate accurate calculations of concentration.

---

* This chapter is largely based on the following communication:
Fallas, J. A.; Lee, M. A.; Jalan, A. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2012**, *134*, 1430-1433.
The text was modified in pertinent sections to fit in the current format and highlight our improved understanding of the subject matter since its publication.

To complement the negative design element, we included a positive design component to widen the stability gap between the desired state and competing states. This was achieved by pairing the lysine residues in chain **C** with aspartate residues in an adjacent strand. Since the **A•B•C** register of the heterotrimer is the target state, two aspartate residues were placed in the X position of the 4th and 9th triplets of the $(PKG)_{10}$ template sequence (P12D and P27D) to make peptide **A**. This sequence arrangement (K at position *n* in the lagging strand and D at position *n*+5 in the leading strand) was chosen based on modeling and previous structural studies,[2] which indicate that this arrangement places the charged side-chain moieties in an ideal position to engage in inter-strand ionic hydrogen bonds. To utilize this relationship, we complete our heterotrimeric system with peptide **B**, which follows the $(DOG)_{10}$ template. The effect of replacing aspartate with glutamate in the X position of the B chain on the specificity of the system was also explored (peptide **B1**, Table 4.1). Previously, we have observed that replacing aspartate with glutamate in heterotrimeric triple helical systems decreases the melting temperature of the heterotrimers,[3] which could lead towards improved specificity in the redesigned system. In order to facilitate the analysis, the peptides will be divided into two groups: system I composed of peptides **A**, **B** and **C** and system II composed of peptides **A**, **B1** and **C**.

Abbreviation	Sequence
A	$YG(PKGPKGPKGDKGPKG)_2$
B	$YG(DOG)_{10}$
B1	$YG(EOG)_{10}$
C	$YG(POGPKGPOGPOGPOG)_2$
O	$(POG)_{10}$

Table 4.1. Abbreviation and chemical sequence of the peptides discussed in the paper. All peptides include a ^{15}N-labelled glycine at position 17 and are free amines at the N-terminus and amides at the C-terminus.

4.2 Circular Dichroism Melting Studies

4.2a System I: A/B/C

Peptide **C** forms a homotrimeric triple helix, **C•C•C**, as evidenced by the sigmoidal transition observed in circular dichroism (CD) thermal unfolding experiments. Figure 4.1B shows the first derivative of the CD melting curve of **C•C•C**, with a transition temperature of 51 °C which is 14 °C lower than that of the parent sequence. Neither of the other peptides this group **(A** and **B)** forms a homotrimer under the examined conditions as evidenced by the linear profile observed in the melting experiments (Figure 4.1a and b).

The binary mixtures involving peptides with opposite charges (**A/B** and **B/C**) form AAB-type heterotrimers, which can be differentiated from the **C•C•C** homotrimer by a lowering of the melting temperature in the CD thermal unfolding curve (Figure 4.2c and d). The **A/C** mixture melts at the same temperature as the sample containing only peptide **C**, which indicates that there is no significant interaction between the peptides. This is not surprising as both peptides have an overall positive charge and the most stable species in the mixture is expected to be the **C•C•C** homotrimer.

A 1:1:1 mixture of peptides **A**, **B** and **C** forms a highly stable ABC-type heterotrimer as evidenced by CD thermal unfolding studies. The first derivative of the CD melting curve of an annealed mixture of the three peptides is shown in Figure 4.1c. The melting temperature of the ternary mixture, at 60° C, is approximately 8 °C higher than that of the **C•C•C** homotrimer, as well as that of any of the binary mixtures, making the ABC heterotrimer the most stable species in the redesigned system (Figure 4.1e and f).

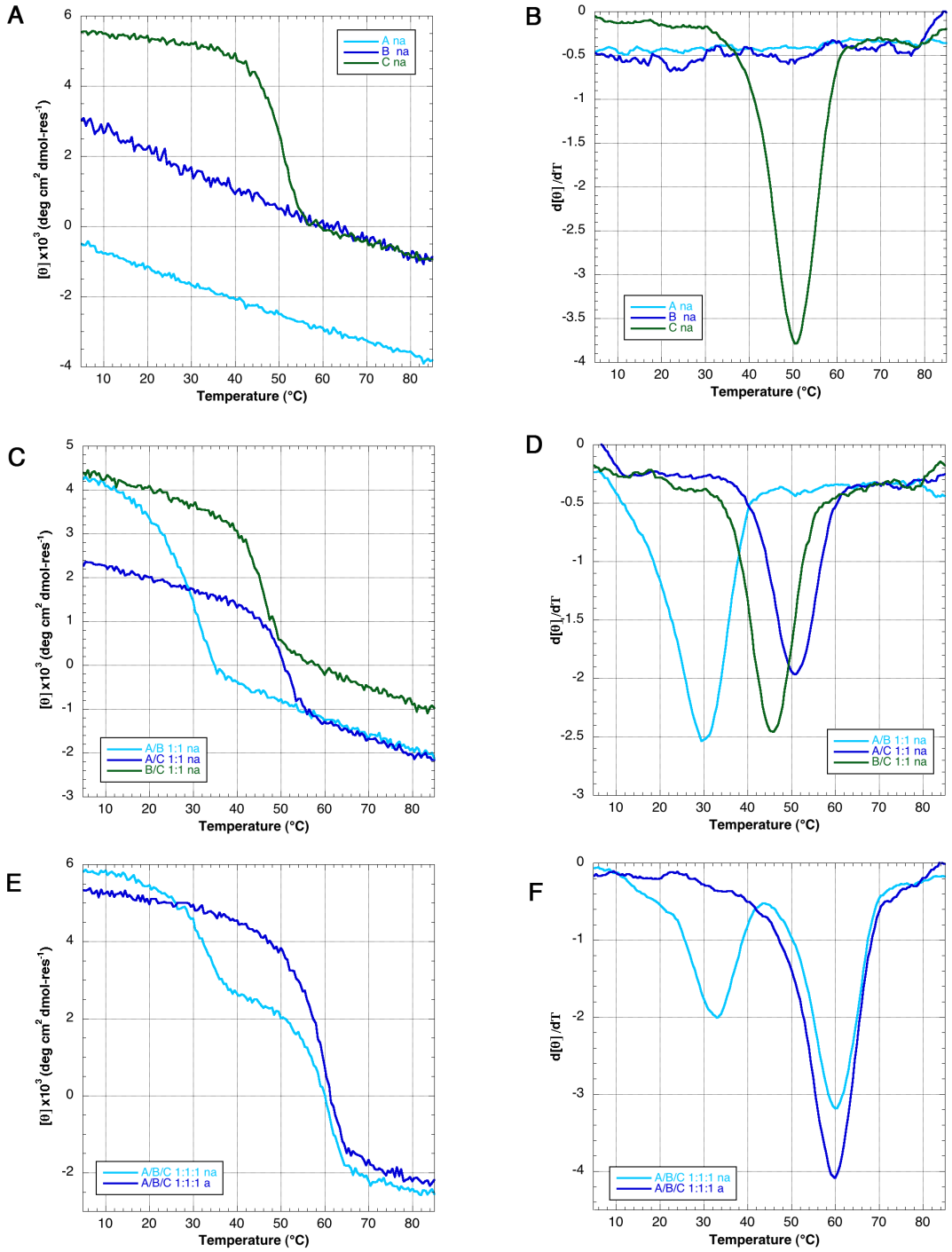


Figure 4.1 CD thermal unfolding curves (right) and first derivatives (left) for different mixture of peptides in system I. (a) and (b) correspond to individual solutions of the three peptides, **A**, **B** and **C**. (c) and (d) correspond to the binary mixtures of the peptides: **A**/**B**, **B**/**C**, **C**/**A**. (e) and (f) correspond the mixture containing all peptides before (na) and after annealing (a).

4.2b System II: A/B1/C

The melting profiles of peptides **A** and **C** and their mixture **A/C** were described in the previous section and peptide **B1** shows a linear melting profile indicating the absence of a homotrimer (Figure 4.2a and b). The binary mixtures (Figure 4.2c and d) behave similar to system I with oppositely charged peptide mixtures (**A**/**B1** and **B1**/**C**) forming AAB-type heterotrimers.

A 1:1:1 mixture of peptides **A**, **B1** and **C** forms an ABC-type heterotrimer as evidenced by CD thermal unfolding studies. The melting temperature for this trimer is lower that of the **A/B**/**C** mixture at 52° C and overlaps with that of the **C•C•C** homotrimer. Despite this overlap, it is possible to differentiate both species by the disappearance of the peak corresponding to the **A**/**B1** trimer upon annealing of the sample (Figure 4.2c and f). Solution NMR experiments will be presented in the next section to complement the CD analysis.

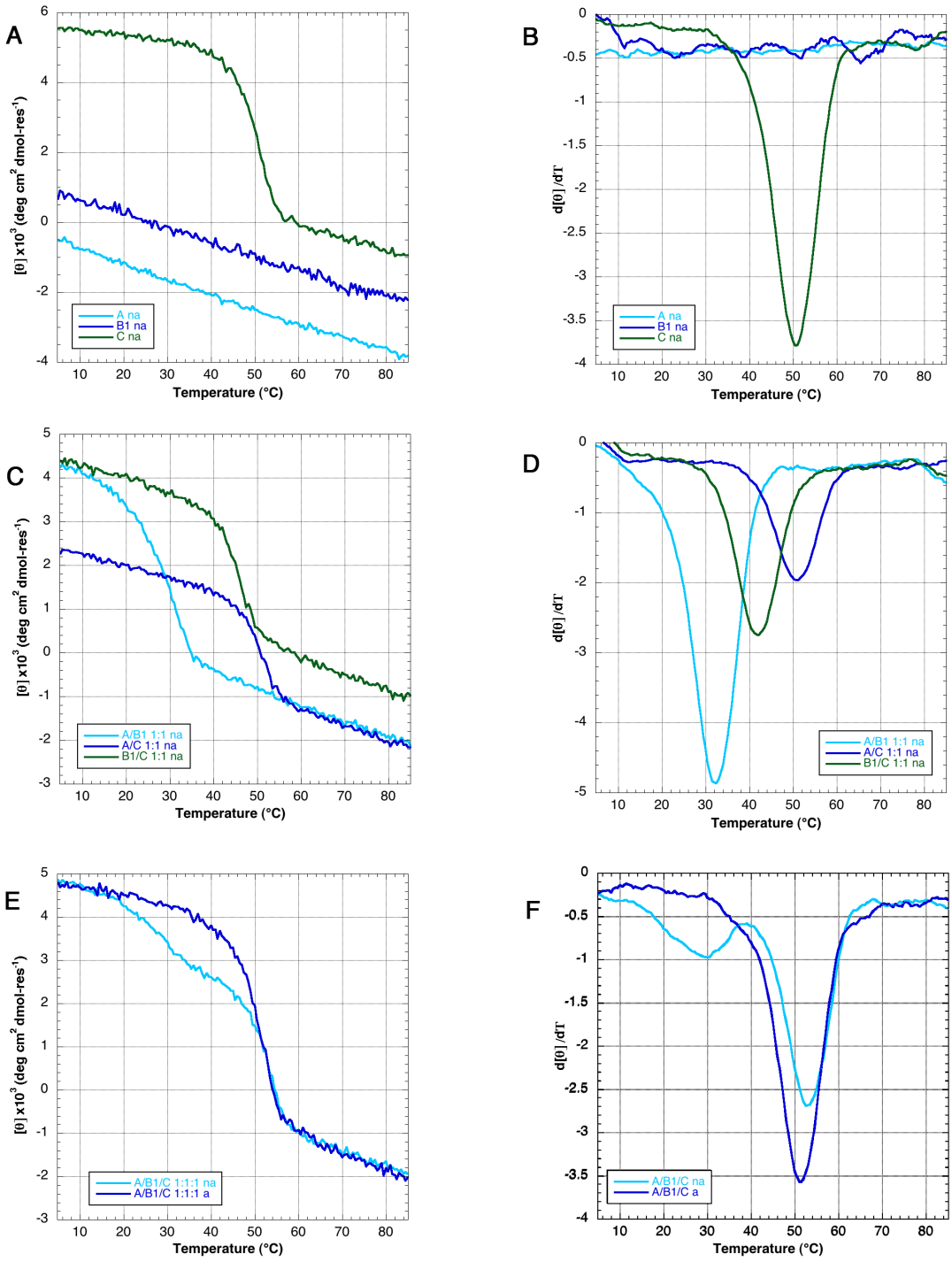


Figure 4.2 CD thermal unfolding curves (right) and first derivatives (left) for different mixture of peptides in system I. (a) and (b) correspond to individual solutions of the three peptides, **A**, **B1** and **C**. (c) and (d) correspond to the binary mixtures of the peptides: **A**/**B1**, **B1**/**C**, **C**/**A**. (e) and (f) correspond the mixture containing all peptides both before (na) and after (a) annealing.

4.3 Solution NMR studies

4.3a System I: A/B/C

To aid the study of the molecular conformation of the peptides in solution, a ^{15}N-labelled glycine was included in the 5th triplet of each peptide (G17). The ^{1}H,^{15}N-Heteronuclear Single Quantum Coherence (HSQC) spectrum of the peptide **C** is shown in Figure 4.3a. The peak corresponding to the **C•C•C** homotrimer has the expected chemical shift, similar to that of **O•O•O**. Two additional peaks are observed and we have assigned them to monomeric peptide with different cis-trans isomerization states of the prolyl-peptide bonds surrounding the labeled glycine.[4]

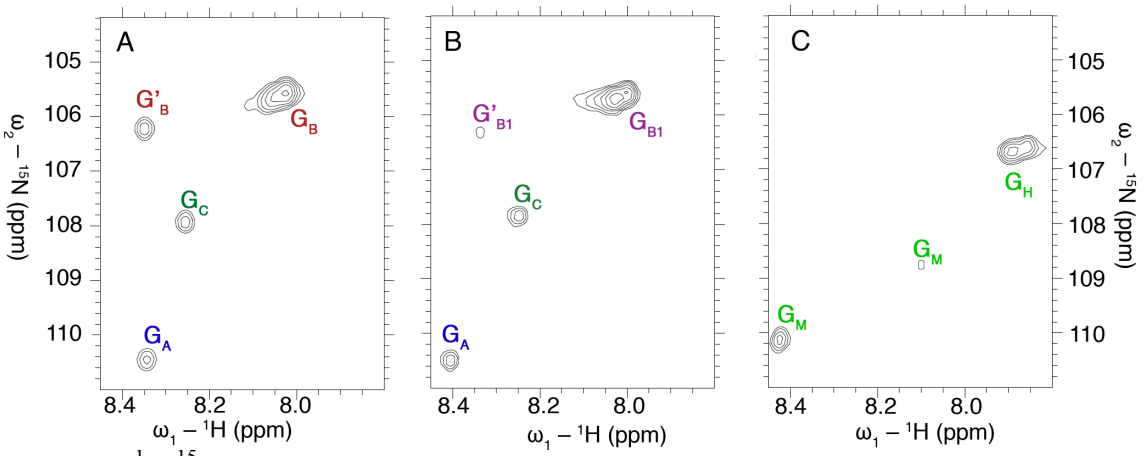


Figure 4.3 ^{1}H,^{15}N-HSQC spectra of the (a) **A•B•C**, (b) **A•B1•C** and (c**) C•C•C** trimers at 25 °C. Each peptide strand contains a ^{15}N-labelled glycine in the 5th triplet, denoted by the subscript next to each peak (H stands for homotrimer and M).

Despite the apparent success observed by CD, we are interested in studying the molecular conformation of the mixture. Particularly, we would like to verify that the difference in stability between the ABC heterotrimer and the homotrimer is sufficient to preclude self-assembly of the latter. The ^{1}H,^{15}N-HSQC spectrum of the mixture (Figure 4.3b) shows the three peaks expected from the heterotrimer, with chemical shifts

comparable to those of the template sequences discussed in the previous chapter.[2] Importantly, it also lacks the peak corresponding the **C•C•C** homotrimer. Thus, within the experimental limits of CD and NMR we only observe the single composition ABC heterotrimic helix, validating our design protocol. Besides the peaks corresponding to the main register of the ABC system, an additional crosspeak is observed in the spectrum (highlighted by a prime symbol), which we assigned to a competing register of the ABC helix. This assignment is based on a combination of Nuclear Overhauser Effect Spectroscopy (NOESY) and $^{1}H,^{15}N$-HSQC experiments at 45 °C to avoid ambiguity about composition since the highest stability AAB heterotrimer within this system has a melting temperature of 45 °C. The analysis focuses on NOEs between the aspartate amide proton and the glycine amide protons in the three different chains. Figure 4.4 highlights the resonances utilized in the assignment of the main register and also shows a molecular model where the spatial arrangement of the atoms utilized in the analysis is illustrated. Figure 4.5 highlights the resonances utilized in the assignment of the secondary register and also shows a molecular model where the spatial arrangement of the atoms utilized in the analysis is illustrated. The peaks corresponding to the other two chains of the competing ABC register were not identified in the $^{1}H,^{15}N$-HSQC due to chemical shift overlap with the peaks arising from the main register.

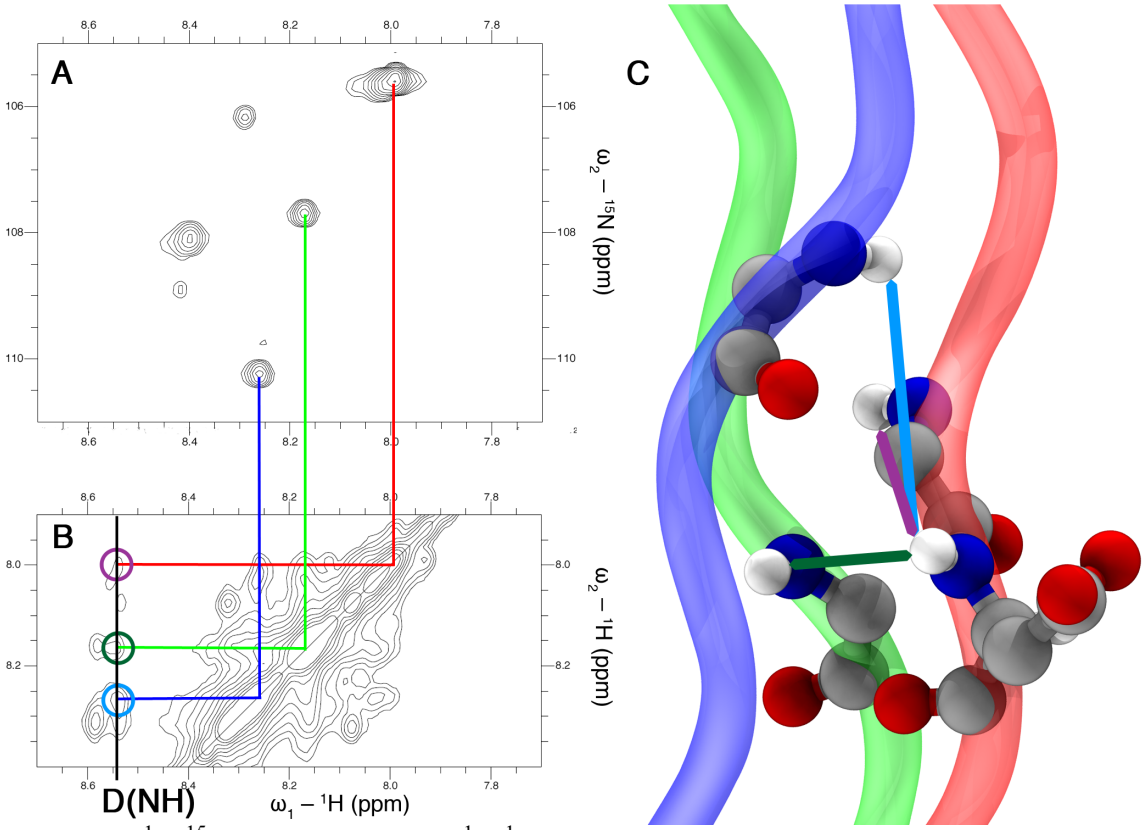


Figure 4.4 $^{1}H,^{15}N$-HSQC spectrum $^{1}H,^{1}H$-NOESY spectrum and model highlighting the resonances arising from the **A•B•C** register of system I. (a) $^{1}H,^{15}N$-HSQC spectrum. (b) NOESY spectrum highlighting amide-amide cross peaks between aspartic acid and three glycine residues in different chains (b) Molecular model showing the protons that give rise to the resonance in (a); Chain **A** is blue, **B** is red and **C** green. Circles in (b) correspond to lines of the same color in (c). Lines connecting resonances in (a) and (b) correspond to the chain color in (c).

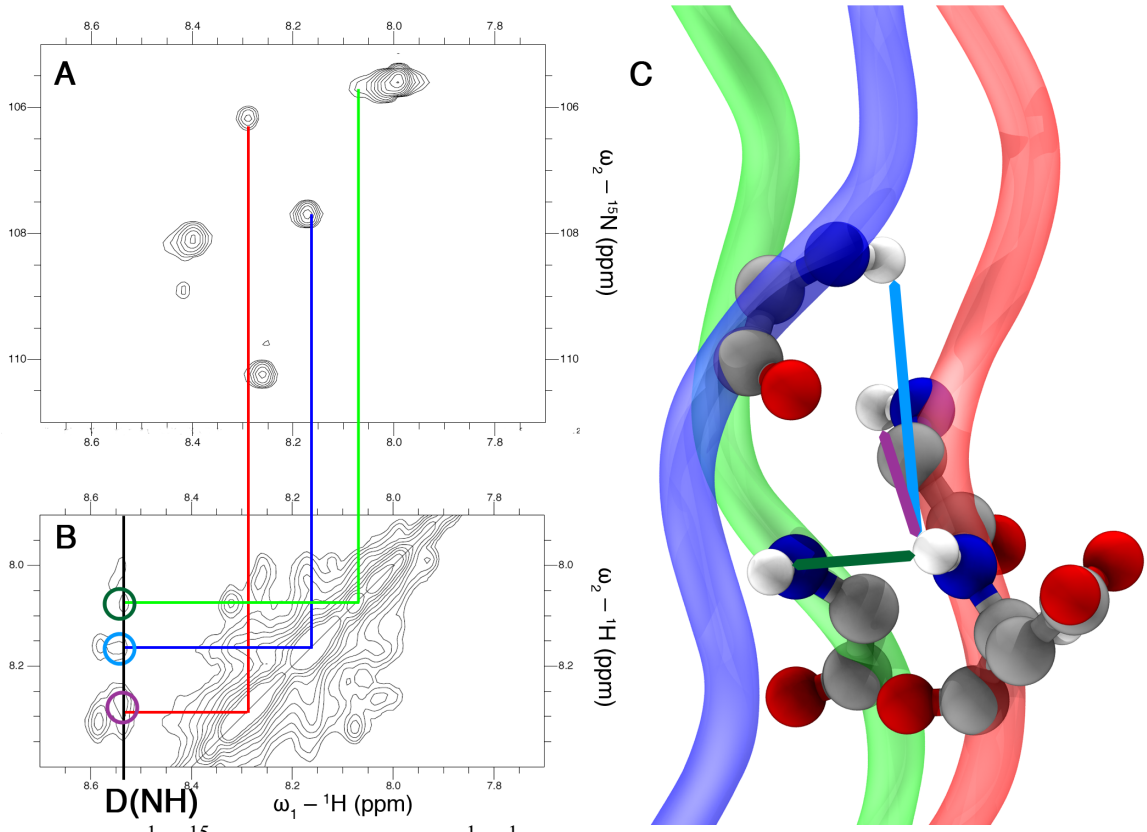


Figure 4.5 ^{1}H,^{15}N-HSQC spectrum ^{1}H,^{1}H-NOESY spectrum and model highlighting the resonances arising from the **C•A•B** register of system I. (a) ^{1}H,^{15}N-HSQC spectrum. (b) NOESY spectrum highlighting amide-amide cross peaks between glutamic acid and three glycine resiudes in different chains (b) Molecular model showing the protons that give rise to the resonance in (a); Chain **A** is blue, **B** is red and **C** green. Circles in (b) correspond to lines of the same color in (c). Lines connecting resonances in (a) and (b) correspond to the chain color in (c).

The CD melting studies corroborate the success of our negative design component, but in order to confirm the success of our multi-state strategy, the structure based positive design element needs to be validated. We do this by studying the molecular conformation of the redesigned amino acids using solution NMR experiments. Because of the symmetry of the triple helix, only a single set of chemical shifts is expected for each of the newly incorporated charged residues,[5] facilitating the analysis using this analytical technique. Figure 4.6a shows two sections of the NOESY spectrum of a 3 mM 1:1:1 mixture of peptides **A**, **B** and **C** at 45 °C. The cross peak at 8.58 and

3.04 ppm arises from the asparate amide proton in chain **A** and the lysine ε-methylene in chain **C**, which shows a single chemical shifts for both diasteriotopic protons. This crosspeak is characteristic of salt bridges in triple helical peptides,[6] and therefore validates our positive design strategy. Furthermore, another peak at 3.64 and 3.04 ppm, arising from lysine ε-protons in chain **C** and the glycine α-protons preceding the aspartate in chain **A**, is also observed. The corresponding resonances between chains **A** and **B** are also highlighted in the figure. Figure 4.6b shows a triple helical structure generated using the flexible backbone modeling capabilities[7] of the Rosetta macromolecular modeling suite[8] and highlights the atoms that give rise to the resonances mentioned above.

These peaks are important, not only because they validate our positive design strategy, but also because they serve to unambiguously determine the register of the triple helix as **A•B•C**. Given the relative position of the charged amino acids in chains **A** and **C,** the only register in which those residues can come close enough to one another to generate NOEs is the target state, the **A•B•C** register of the heterotrimeric helix (Figure 4.7).

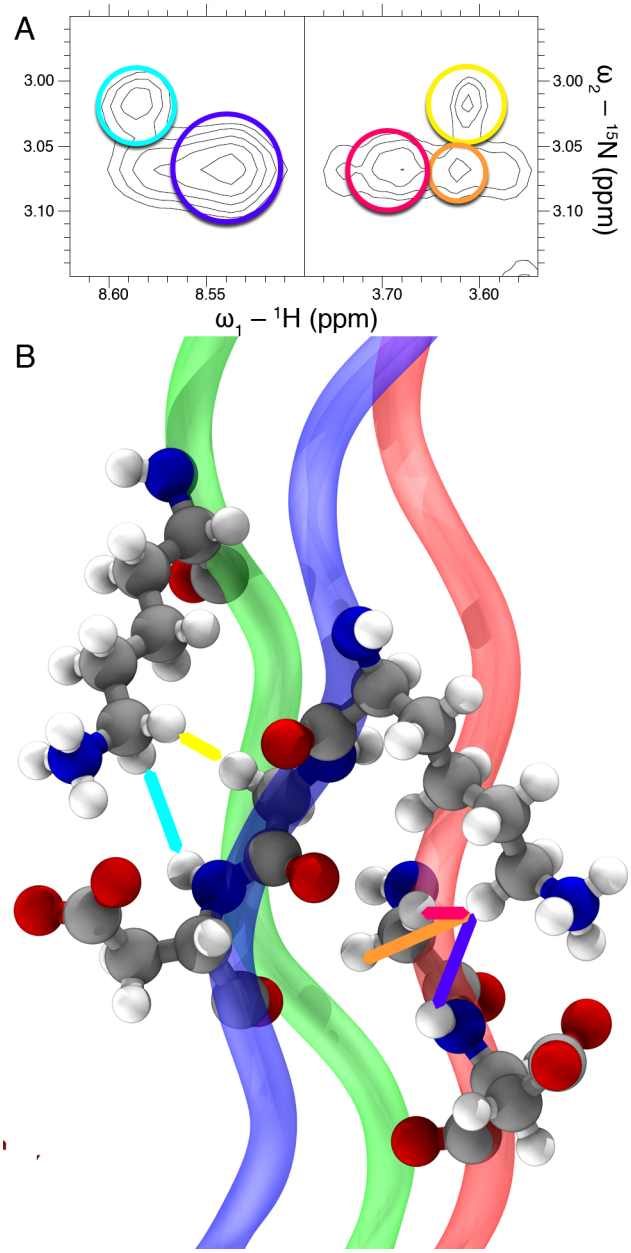


Figure 4.6 ^{1}H,^{1}H-NOESY spectum and model. (a) NOESY spectrum of the **A•B•C** heterotrimer highlighting inter-strand interactions involving the lysine side-chains and aspartate/glutamate and glycine backbone atoms. (b) Molecular model showing the protons that give rise to the resonance in (a); Chain **A** is blue, **B** is red and **C** green. Lines and circles of the same color in (a) and (b) correspond to one another.

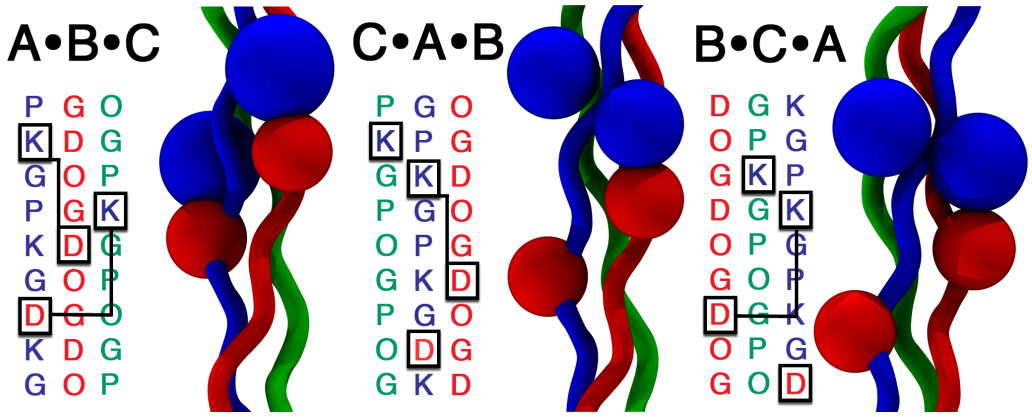


Figure 4.7 Registers of an ABC heterotrimer. Amino acids highlighted by boxes in the aligned sequence correspond to beads in the structure. Pairs of amino acids participating in inter-strand salt bridges are connected by lines.

4.3b System II: A/B/C

The $^{1}H,^{15}N$-HSQC spectrum of system II (Figure 4.3b) lacks the peak corresponding to the **C•C•C** homotrimer, confirming that this system is also composed solely of ABC heterotrimers despite the similar melting temperature of the **C•C•C** homotrimer. This can be rationalized utilizing the chemical potential of the solution and the relative stability of the different available states. The mixture will seek to minimize the chemical potential to reach equilibrium and this achieved by populating only heterotrimeric helices, as the self-assembly of **C•C•C** would force the two remaining peptide chains to fold into relatively unstable AAB heterotrimers. The pattern observed in this spectrum is very similar to that of the mixture containing **B** (Figure 4.3a), and contains both a main register (Figure 4.8) and a secondary register (Figure 4.9). The main difference between system I and system II is that the peak corresponding to the secondary register is significantly weaker, which we interpret as a reduction in the relative population of this state. Thus, by including both K/E and K/D charge pairs, we have improved upon the specificity towards a particular heterotrimeric register with the caveat

that the overall stability of the system is decreased. This trade-off between the overall stability in a designed protein system and the specificity towards a particular state is not unusual,[9,10] but has only been recently been explored for triple helical systems.[11]

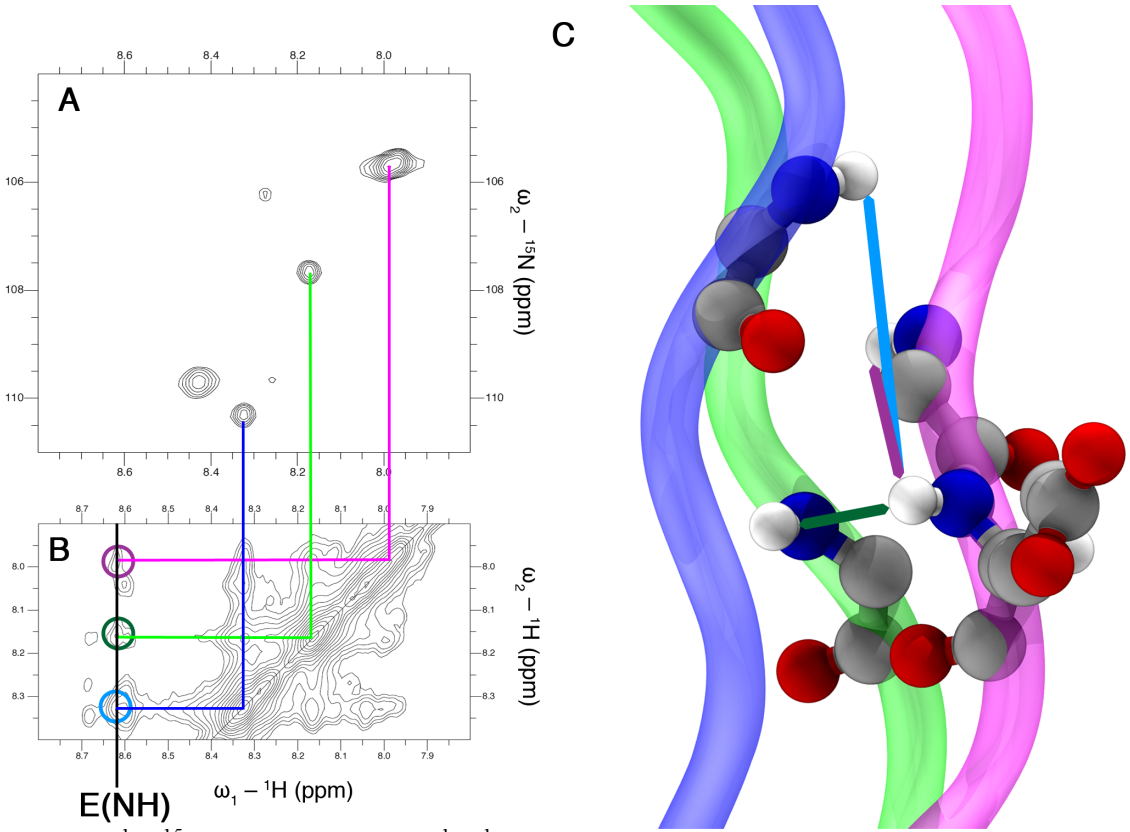


Figure 4.8 $^{1}H,^{15}N$-HSQC spectrum $^{1}H,^{1}H$-NOESY spectrum and model highlighting the resonances arising from the **A•B1•C** register of system II. (a) $^{1}H,^{15}N$-HSQC spectrum. (b) NOESY spectrum highlighting amide-amide cross peaks between glutamic acid and three glycine resiudes in different chains (b) Molecular model showing the protons that give rise to the resonance in (a); Chain **A** is blue, **B1** is magenta and **C** green. Circles in (b) correspond to lines of the same color in (c). Lines connecting resonances in (a) and (b) correspond to the chain color in (c).

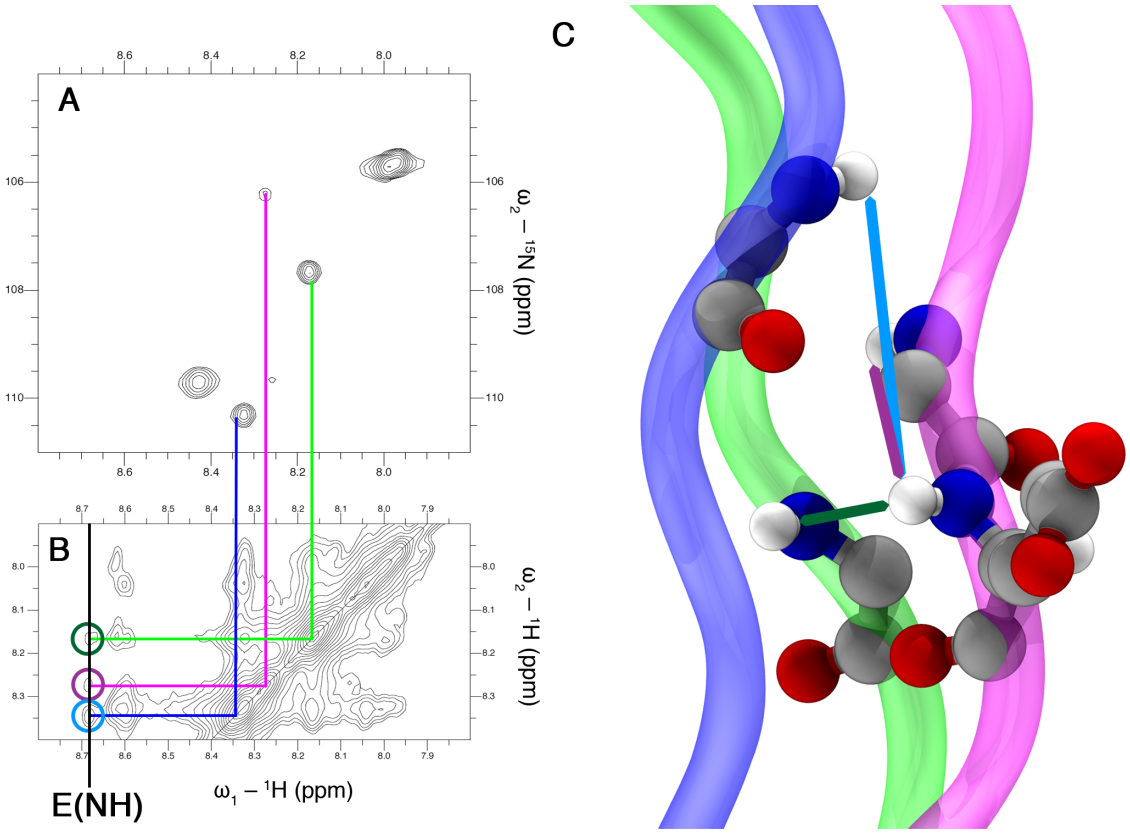


Figure 4.9 $^{1}H,^{15}N$-HSQC spectrum $^{1}H,^{1}H$-NOESY spectrum and model highlighting the resonances arising from the **C•A•B1** register of system II. (a) $^{1}H,^{15}N$-HSQC spectrum. (b) NOESY spectrum highlighting amide-amide cross peaks between glutamic acid and three glycine resiudes in different chains (b) Molecular model showing the protons that give rise to the resonance in (a); Chain **A** is blue, **B1** is magenta and **C** green. Circles in (b) correspond to lines of the same color in (c). Lines connecting resonances in (a) and (b) correspond to the chain color in (c).

The register of the ABC-type heterotrimer of system II was also studied. Using a similar methodology than for system I, we studied the molecular conformation of the redesigned amino acids using solution NMR experiments. Figure 4.9a shows two sections of the NOESY spectrum of a 3 mM 1:1:1 mixture of peptides **A**, **B1** and **C** at 45 °C. The cross peak at 8.61 and 3.02 ppm arises from the aspartate amide proton in chain **A** and the lysine ε-methylene in chain **C**, which shows a single chemical shifts for both diasteriotopic protons. This crosspeak is characteristic of salt bridges in triple helical peptides[6] and once again validates our positive design strategy. Another peak at 3.68 and 3.02 ppm, arising from lysine ε-protons in chain **C** and the glycine α-protons preceding the aspartate in chain **A**, is also observed. The corresponding resonances between chains **A** and **B1** are also highlighted in the figure. These peaks are important, not only because they determine the register of the triple helix as **A•B1•C**. Given the relative position of the charged amino acids in chains **A** and **C,** the only register in which those residues can come close enough to one another to generate NOEs is the target state, the **A•B1•C** register of the heterotrimeric helix. Figure 4.10b shows a triple helical structure generated using the flexible backbone modeling capabilities[7] of the Rosetta macromolecular modeling suite[8] and highlights the atoms that give rise to the resonances mentioned above.

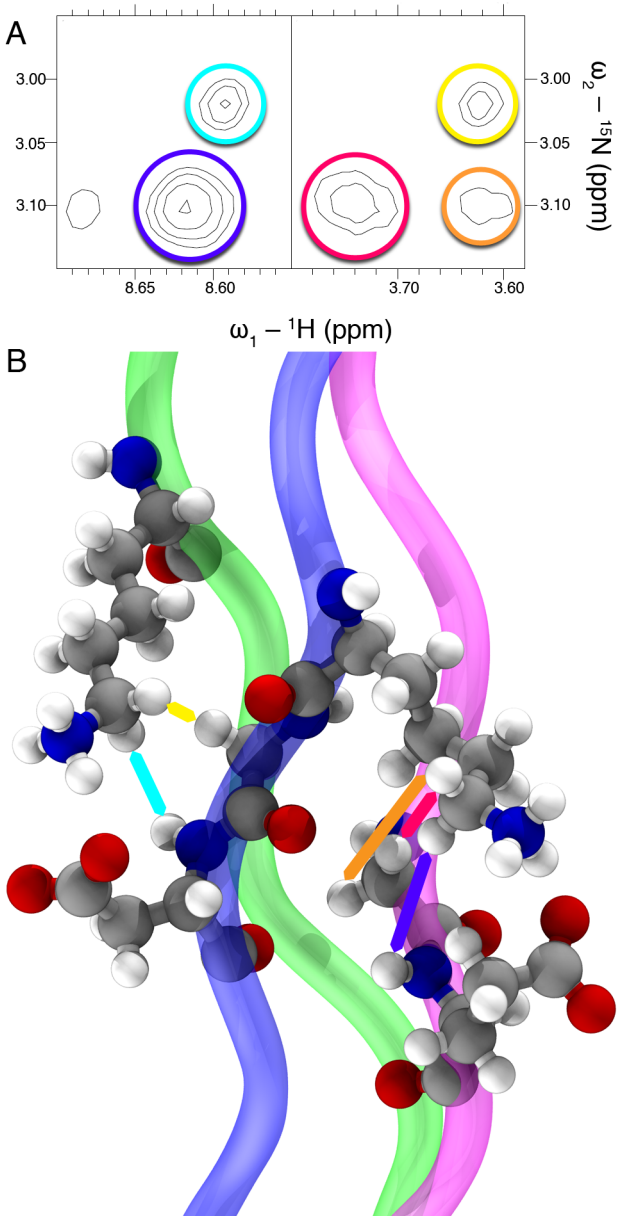


Figure 4.10 ^{1}H,^{1}H-NOESY spectum and model. (a) NOESY spectrum of the **A•B1•C** heterotrimer highlighting inter-strand interactions involving the lysine side-chains and aspartate/glutamate and glycine backbone atoms. (b) Molecular model showing the protons that give rise to the resonance in (a); Chain **A** is blue, **B** is red and **C** green. Lines and circles of the same color in (a) and (b) correspond to one another.

4.4 Conclusions

The synthesis of a single register self-assembling ABC triple helical heterotrimer of high stability is a major challenge for *de novo* protein design. Such a system is desirable because it can be used as a scaffold in host-guest peptides to study the structure, biochemistry, stability and multi-state self-assembly of heterotrimeric collagenous proteins, mirroring what has been done with homotrimeric triple helices, which has been pivotal in our understanding of the most abundant protein family in the human body.[12-18] This study successfully improves upon previous efforts by generating a high-stability single composition ABC heterotrimeric system utilizing a rational multi-state design strategy that exploits novel sequence-structure relationships in triple helical proteins. Despite populating more than one register of the desired heterotrimer, NMR studies on the system are able to confirm that the main component in the mixture corresponds to the target state and that the stabilizing pair-wise interactions that were included based upon structural modeling are satisfied. Furthermore, we are able to reduce the relative population of alternative registers by modifying the amino acid composition of one the peptide chains within the system. In order for future systems to improve upon selectivity towards a particular register, the stability of all competing states needs to be included in the design protocol. This idea will be expanded upon in the next chapter.

4.5 Experimental

*Peptide Synthesis and Purificatio***n –** All peptides were synthesized with an Advanced Chemtech Apex 396 solid phase peptide synthesizer using Fmoc chemistry and a Rink MBH amide resin. During the automated procedure, a manual addition of 2 equivalents

^{15}N-labelled glycine, purchased form Cambridge Isotope Laboratories, was carried out in position 17. The final sequences include a tyrosine (for concentration determination) and a glycine spacer at the N-terminus. All peptides are C-terminally amidated. The peptides were purified on a Varian PrepStar220 HPLC with a preparative reverse phase C-18 column using a linear water/acetonitrile gradient each containing 0.5% TFA and analyzed by ESI-TOF mass spectrometry on a Bruker microTOF instrument.

Concentration Determination – Concentration of stock solutions was determined by UV/Vis absorption at 275 nm using a molar extinction coefficient of 1400 cm^{-1}/M.

Circular Dichroism – All CD experiments were performed with a Jasco J-810 spectropolarimeter equipped with a Peltier temperature control system. Non-annealed samples were mixed and incubated at room temperature overnight before measurements were preformed. Annealed samples were heated above 80 °C for 30 minutes in the spectrometer and incubated for at least 12 hours at room temperature before measurements were performed. Samples were prepared to a total concentration of 300 μM in 10 mM phosphate buffer at pH 7 by mixing the desired peptides in the appropriate ratio. Samples containing binary mixtures were prepared in a 1:1 ratio and samples containing ternary mixtures were prepared in a 1:1:1 ratio. Spectra were acquired between 215-250 nm to locate the maximum near 222 nm, which was monitored during unfolding experiments. Melting curves were performed from 5 to 85 °C with a heating rate of 10 °C/hr. The first derivative of the melting curve was taken in order to determine the melting temperature (T_m) of the sample, which we define as the minimum in the

derivative graph. The molar residual ellipticity (MRE) is calculated from the measured ellipticity using the equation:

$$[\theta] = \frac{\theta \times m}{c \times l \times n_r}$$

where θ is the ellipticity in mdeg, m is the molecular weight in g/mol, c is the concentration in mg/mL, l is the pathlength of the cuvette in cm, and n_r is the number of amino acids in the peptide. All annealed samples were repeated after a 2-week incubation period to ensure that systems under thermodynamic equilibrium were measured. For ease of discussion peptides were grouped into two systems: system I refers to peptides **A**, **B** and **C** and system II refers to peptides **A**, **B1** and **C**.

Molecular Modeling – Homology models for the **A•B•C**, **C•A•B** and **B•C•A** registers of system I and the **A•B1•C**, **C•A•B1** and **B1•C•A** registers of system II were prepared using the Rosetta software suite[8] using the crystal structure of a triple helical peptide (pdb id: 1K6F) as a template.[19] After mutating the residues to their corresponding identity in each of the register using the fixed backbone design application rounds of flexible backbone modeling using the backrub application (100000 monte-carlo steps with an internal temperature value of 0.3) and side chain relaxation were carried out. Because this particular macromolecular software suite lacks explicit electrostatic scoring terms but includes directional hydrogen-bonding potentials, distance constraints were placed upon the charged residues to bias them towards the observed experimental conformations, as evidenced by NOE cross peak patterns and previous structural studies, during both procedures. The backrub/relax protocol was repeated until convergence was reached.

NMR Spectroscopy – All NMR experiments were recorded in an 800 MHz Varian spectrometer equipped with a triple resonance probe. Samples were prepared at two different total peptide concentrations (1 mM and 3 mM) in a10 mM phosphate buffer at pH 7 and a 9:1 ratio of H_2O to D_2O. Heterotrimer samples were mixed in a 1:1:1 ratio, heated to 85 °C and incubated overnight at room temperature before starting the experiments. Homotrimer samples were only studied at a 1 mM concentration and were not annealed. The spectra were processed using NMRpipe[20] and analyzed using ccpnmr.[21] Square Cosine bell window functions were used as apodization functions and the data was zero-filled to the next power of two in both dimensions. Drift and baseline corrections were applied when necessary. Each sample was characterized using 2D total correlated spectroscopy (TOCSY), nuclear Overhauser effect spectroscopy (NOESY), $^{1}H,^{15}N$-heteronuclear single quantum coherence (HSQC) and 2D NOESY-^{15}N-HSQC experiments at 45 °C (3 mM samples for heterotrimers, 1 mM sample for the homotrimer) and $^{1}H,^{15}N$-HSQC and $^{1}H,^{1}H$-NOESY experiments at 25 °C (1 mM samples). TOCSY spectra with a 50 ms spinlock duration at 8 kHz were acquired with a total of 1700 complex points recorded in 8 scans for the directly acquired dimension while 500 increments were used in the indirect dimension. NOESY spectra with a 100 ms mixing time were acquired with a total of 1700 complex points recorded in 8 scans for the directly acquired dimension while 500 increments were used in the indirect dimension. A square spectral window of 1000 Hz was used for all homonuclear spectra. A total of 1208 complex points in 32 scans for the direct dimension and 50 increments in the indirect dimension were acquired for the $^{1}H,^{15}N$-HSQC experiments using a spectral

window of 10000 Hz in the hydrogen dimension and 1200 Hz in the nitrogen dimension. For the 2D NOESY-^{15}N-HSQC spectra a mixing time of 100 ms was used and a total of 1600 complex points in 128 scans for the direct dimension and 200 increments for the indirect dimension were acquired using a spectral window of 8000 Hz for the direct dimension and 7200 for the indirect dimension.

Sequential Assignment and Species Identification – Because of the symmetry of the triple helix, most of the amino acids in repetitive sequence show identical chemical shifts. The substitutions included in the present study partially break this symmetry but there is still a set of residues that are in a symmetric chemical environment and give rise to a set of stronger resonances. The chemical sequence of those residues in the main register was determined using a combination of ^{1}H,^{15}N-HSQC, ^{1}H,^{1}H-NOESY, ^{1}H,^{1}H-TOCSY experiments and the data from the template sequences as a reference.[2] The sequential assignment procedure for the redesigned sequences (GDK in chain **A** and OGD in chain **C**) was carried out using a more traditional sequential assignment procedure using ^{1}H,^{1}H-TOCSY and ^{1}H,^{1}H-NOESY experiments at 45 °C and ^{1}H,^{1}H-NOESY experiments at a temperature of 25 °C. The peaks for the secondary register were assigned using a combination of ^{1}H,^{1}H-NOESY and ^{1}H,^{15}N-HSQC experiments at 45 °C.

4.6 References

(1) Persikov, A. V.; Ramshaw, J. A.; Brodsky, B. *J. Biol. Chem.* **2005**, *280*, 19343-19349.

(2) Fallas, J. A.; Gauba, V.; Hartgerink, J. D. *J. Biol. Chem.* **2009**, *284*, 26851-26859.

(3) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 15034-15041.

(4) Buevich, A. V.; Dai, Q. H.; Liu, X. Y.; Brodsky, B.; Baum, J. *Biochemistry* **2000**, *39*, 4299-4308.

(5) Li, M.-H.; Fan, P.; Brodsky, B.; Baum, J. *Biochemistry* **1993**, *32*, 7377-7387.

(6) O'Leary, L. E.; Fallas, J. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2011**, *133*, 5432-5443.

(7) Humphris, E. L.; Kortemme, T. *Structure* **2008**, *16*, 1777-1788.

(8) Leaver-Fay, A. et al. *Methods Enzymol.* **2011**, *487*, 545-574.

(9) Bolon, D. N.; Grant, R. A.; Baker, T. A.; Sauer, R. T. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 12724-12729.

(10) Grigoryan, G.; Reinke, A. W.; Keating, A. E. *Nature* **2009**, *458*, 859-864.

(11) Xu, F.; Zahid, S.; Silva, T.; Nanda, V. *J. Am. Chem. Soc.* **2011**,

(12) Kramer, R. Z.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *Nat. Struct. Biol.* **1999**, *6*, 454-457.

(13) Persikov, A. V.; Ramshaw, J. A.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2000**, *39*, 14960-14967.

(14) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *J. Mol. Biol.* **2002**, *316*, 385-394.

(15) Emsley, J.; Knight, C. G.; Farndale, R. W.; Barnes, M. J.; Liddington, R. C. *Cell* **2000**, *101*, 47-56.

(16) Farndale, R. W.; Lisman, T.; Bihan, D.; Hamaia, S.; Smerling, C. S.; Pugh, N.; Konitsiotis, A.; Leitinger, B.; de, G., PG; Jarvis, G. E.; Raynal, N. *Biochem. Soc. T.* **2008**, *36*, 241-250.

(17) Carafoli, F.; Bihan, D.; Stathopoulos, S.; Konitsiotis, A. D.; Kvansakul, M.; Farndale, R. W.; Leitinger, B.; Hohenester, E. *Structure* **2009**, *17*, 1573-1581.

(18) Kar, K.; Ibrar, S.; Nanda, V.; Getz, T. M.; Kunapuli, S. P.; Brodsky, B. *Biochemistry* **2009**, *48*, 7959-7968.

(19) Berisio, R.; Vitagliano, L.; Mazzarella, L.; Zagari, A. *Protein Sci.* **2002**, *11*, 262-270.

(20) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277-293.

(21) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, P.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. *Proteins* **2005**, *59*, 687-696.

Chapter 5: Computational Design of Register-Specific ABC Collagen Heterotrimers*

Proteins have mastered the cooperative use of non-covalent interactions to self-assemble into complex three-dimensional architectures. A rather stringent test of our understanding of the principles that determine a protein's structure from the physicochemical information encoded in its amino acid sequence lies in the design of synthetic polypeptide chains that are able to replicate this feat; that is, to accurately fold into a particular conformation while avoiding the population of closely related states. Computational design protocols have been successful at this task, particularly when dealing with globular proteins[1-3] and α-helical coiled coils.[4-6] These structural motifs benefit from the presence of a hydrophobic core that is buried upon exposure to an aqueous environment and acts as a major driving force in the folding and association of the peptide chains.[7] A structural motif that, despite its predominance in higher organisms, has seen rather limited success in this field is the collagen triple helix.[8,9] The large number of competing states that need to be explicitly modeled and the fact that only solvent-exposed amino acids can be used to bias the chain association in this fold, make it a challenging system for *de novo* design.

In this chapter, we develop a multi-state computational design protocol using a sequence-based scoring function that exploits the sequence-structure relationships derived in Chapter 2.[10] This approach allows us to explicitly calculate all the possible triple helical states within a peptide mixture and optimize the stability of the desired

---

* This chapter is based upon a manuscript that has been submitted for publication in a peer-reviewed scientific journal.

target state while maximizing the energy gap between the target and the most stable decoy. As a proof of principle, we use this methodology to design three peptides that fold into an ABC heterotrimer with complete control over the helix composition and register.

5.1 Computational Design Methodology

We developed a sequence-based scoring function for triple helical proteins based on our understanding of the non-covalent interactions that stabilize this protein fold. We set the prototypical homotrimeric sequence, $(POG)_{10}$, as the reference state and give it a numerical value of 0 in our relative scale. Single point mutations with respect to this scaffold, which are known to be destabilizing,[11] are given a positive numerical value. Pairs of amino acids that are known to interact favorably and stabilize the fold[12] are given negative numerical value. In principle any single and double substitutions can be allowed but we have restricted ourselves to oppositely charged amino acids, particularly, lysine and aspartic acid as they have shown to engage in the most stabilizing inter-chain ionic hydrogen bonds in the context of rationally designed collagen heterotrimers[13], as exemplified by the results presented in Chapter 4. Furthermore, we restrict the amino acid identity of the X position to either P or D and that of the Y position to either O or K following the pattern observed in naturally occurring collagens where negatively charged amino acids have higher propensity for the X position and positively charged amino acids have a higher propensity for the Y position.[14] Even in this reduced space, two distinct contact geometries between the oppositely charged amino acids are possible, which we refer to as lateral and axial interactions in Chapter 2.[10] As discussed in that chapter lateral contacts are only marginally stabilizing in triple helices[15] while axial contacts have been

shown to bias self-assembling peptides towards a specific heterotrimeric target state as discussed in Chapters 3 and 4,[16] thus only the latter will be considered in our current approach.

With these considerations in mind, the energy score (E) of a particular sequence is given by

$$E = \varepsilon_1 M - \varepsilon_2 N \qquad (1)$$

where M is the number of ionizable residues, N is the number of axial salt bridges and ε_1 and ε_2 their respective contributions. Figure 5.1a shows the relative position of interacting amino acids in axial salt bridges in terms of aligned triple helical sequences and Figure 5.1b is a molecular representation of the interacting side-chains (pdb ID 3U29). We hypothesize that this function, despite its simplistic form and the numerous approximations used in its formulation, captures the dominant contributions to the free energy difference between triple helical states in the sequence space of interest by i) penalizing point mutations from the POG scaffold and ii) rewarding double mutations that lead to the formation of ionic hydrogen bonds between adjacent strands. Furthermore, although we arrive at our expression using intuitive supramolecular considerations, it can be independently derived using a rigorous theoretical approach. It can be shown that equation (1) corresponds to a truncated, simplified version of the cluster expansion, recently applied by Keating et al.[17] to evaluate protein energies from their amino acid sequences.

The second component of the design protocol is a search algorithm that is able to explore the space of interest and select sequences that satisfy a given set of constraints. We use a genetic algorithm (GA) for this purpose since it has been successful in multi-

state protein design problems.[18,19] For this approach, a fitness function needs to be defined and optimized. We define our fitness function, χ, as

$$\chi = E_T - \lambda\Delta E, \quad \Delta E = E_T - \min[E_i]_{i \neq T} \qquad (2)$$

where E_T represents the stability of the target state, λ is a proportionality constant and ΔE is the difference in stability between the target state and the most stable member of the competing state ensemble, which is a measure of the specificity of the system towards the target state. The first term biases the search to sequences that have low energy scores and thus a large proportion of paired charged amino acids or a high content of P and O residues. The second term biases the search towards sequences where there are more unpaired basic and acidic residues in the most stable decoy than in the target structure.

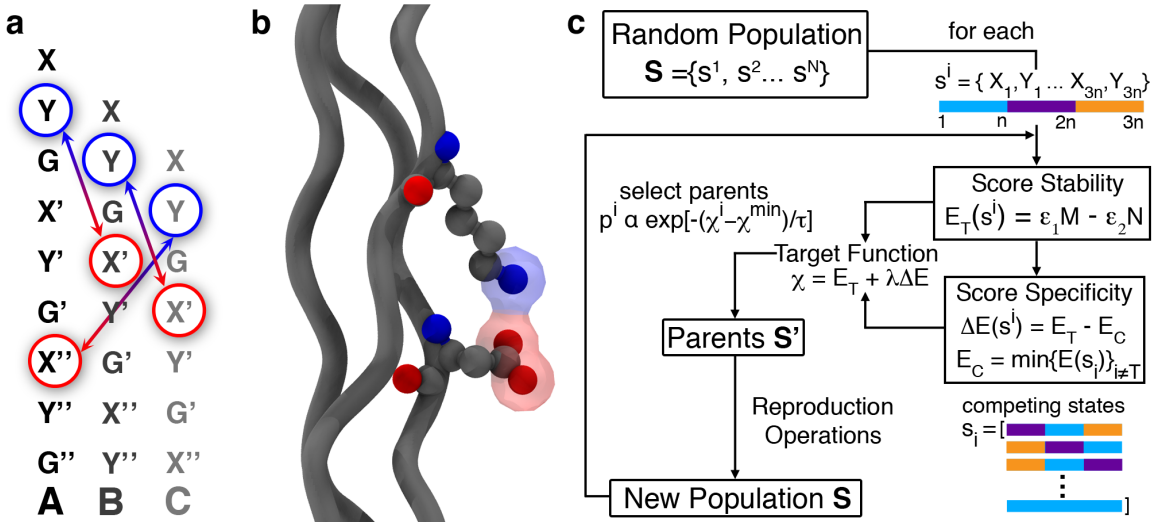


Figure 5.1 Inter-chain interactions and computational design protocol. **a**, Position of the interacting amino acids in axial salt bridges in terms of aligned triple helical sequences. **b**, Molecular representation of the contacts shown in **a** (pdb id 3U29). **c**, Schematic of our sequence selection genetic algorithm.

In our GA (Figure 5.1c) we start with a random population of sequences that are scored according to their fitness. A second subset is generated that is augmented with some of the fittest members of the initial population which are then subjected to

reproduction operations to generate an offspring generation. This process is repeated until a target fitness is met or a preset number of generations is produced. Details on the GA are available in the Materials and Methods section.

The best fitness score found for ABC-type sequences was -12 a.u.; this means that there are 12 more unpaired ionizable residues in the most stable competing state than in the desired triple helix, where all oppositely charged residues are paired. This solution is not unique and although we cannot prove that it corresponds to the global minimum of the fitness function, we show experimentally that it is sufficient to preclude the self-assembly of any alternative states when all three sequences are present in solution.

5.2 Experimental Characterization

Table 5.1 shows the three sequences that were selected for experimental characterization, which will be referred to as α, β and γ respectively. These peptides have a smaller net charge (-2, +2 and 0, respectively) than the rationally designed triple helical heterotimers studied in the previous chapters despite having a higher content of charged residues. There are 14 possible axial contacts, which are satisfied in the desired register $\alpha\bullet\beta\bullet\gamma$. The next most stable configuration corresponds to 8 paired salt bridges with 12 unpaired ionizable residues and there are several triple helices with that arrangement: 2 alternative ABC registers ($\beta\bullet\gamma\bullet\alpha$ and $\gamma\bullet\alpha\bullet\beta$) and 10 AAB-type helices ($\alpha\bullet\alpha\bullet\beta$, $\alpha\bullet\beta\bullet\alpha$, $\alpha\bullet\beta\bullet\beta$, $\beta\bullet\alpha\bullet\beta$, $\alpha\bullet\alpha\bullet\gamma$, $\alpha\bullet\gamma\bullet\gamma$, $\beta\bullet\gamma\bullet\gamma$, $\gamma\bullet\beta\bullet\gamma$, $\beta\bullet\gamma\bullet\beta$ and $\beta\bullet\beta\bullet\gamma$).

Abbreviation	Sequence
α	PKGPKGDOGPOGDK**G**DKGPKGPOGDKGPOGGY
β	POGDOGDKGPOGPO**G**DKGDOGDKGPKGDOGGY
γ	PKGPOGPKGDKGPO**G**POGDKGPOGDOGDOGGY

Table 5.1 Peptide sequences and abbreviations. All peptides include a ^{15}N-labelled glycine at position 15 and are acetyl-esters at the N-terminus and amides at the C-terminus.

5.2a Circular Dichroism Melting Studies

In order to assess the performance of our GA algorithm, samples were prepared for CD melting studies with a total peptide concentration of 0.3 mM in 10 mM phosphate buffer at pH 7. We utilize the minimum in the first derivative of the unfolding curve to define the melting temperature in our analysis. Each sequence was examined individually, in 1:1 binary mixtures and in a 1:1:1 ternary mixture (Figure 5.2). Only peptide γ shows the formation of a homotrimeric helix under the examined conditions, as evidenced by the weak cooperative transition observed in the unfolding experiment (Figure 5.2 a and b). All binary mixtures show cooperative transitions with the 1:1 α/β mixture having the lowest molar residual ellipticity (MRE) and melting temperature (T_m). The 1:1 α/γ and β/γ mixtures both show transitions with the same T_m (43 °C, figure 5.2 c and d) and comparable MRE. The ternary mixture shows the highest T_m of the system with an unfolding transition at 58 °C, 15 °C higher than the most stable competing AAB heterotrimers (Figure 5.2 e and f). We attribute this difference in thermal stability to the difference in the number of charge pairs between the desired register and the AAB competing states.

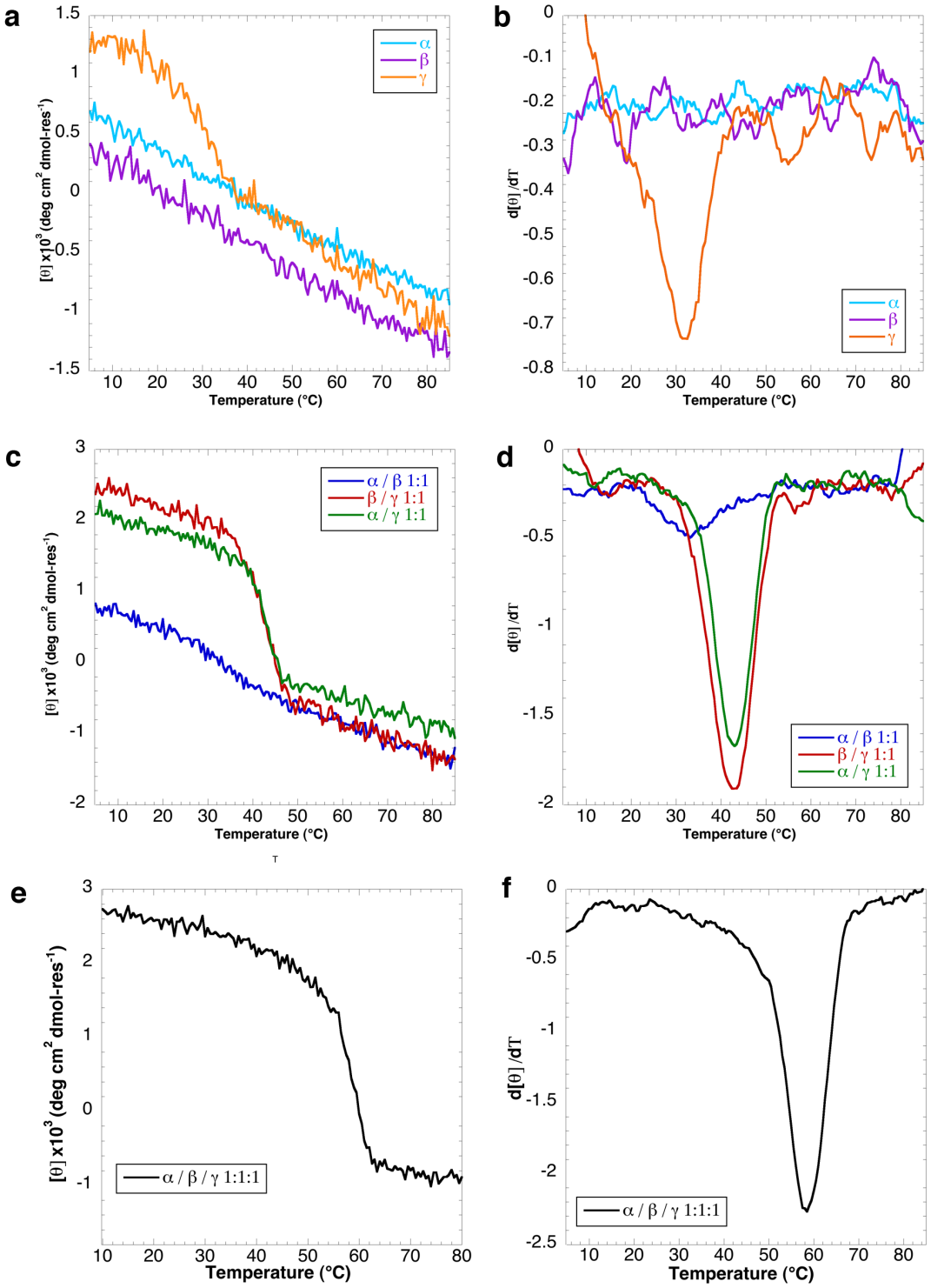


Figure 5.2 CD thermal unfolding curves (right) and first derivatives (left) for different different peptide mixtures; **a** and **b** correspond to individual solutions of the three peptides: α, β and γ; **c** and **d** correspond to the 1:1 binary mixtures: α/β, β/γ and α/γ; **e** and **f** corresponds to the 1:1:1 α/β/γ mixture.

Although this result is encouraging, the presence of competing states can be easily masked in CD melting studies. Furthermore, this technique cannot differentiate between different registers of a given helix to show that the cooperative transition observed in the ternary mixture indeed corresponds to the designed register. For this reason solution NMR studies were carried out to corroborate that the ternary mixture, within the detection limits of this analytical technique, is indeed composed solely of the desired $\alpha\bullet\beta\bullet\gamma$ heterotrimer.

5.2b Structural Characterization

Samples for NMR were prepared in 10 mM phosphate buffer at pH 7 with 10% D_2O. Once again, each sequence was examined individually, in 1:1 binary mixtures and in a 1:1:1 ternary mixture. Figure 5.3 shows the $^{1}H,^{15}N$-Heteronuclear Single Quantum Coherence (HSQC) spectra of the different samples at 37 °C. Each of the peptide sequences contains a ^{15}N-labeled glycine at position 15 to facilitate the analysis. A single peak is expected from every unique chemical environment that each of the peptides encounter. No homotrimeric triple helices are present at this temperature, as expected from the CD melting studies and evidenced by the absence of trimeric peaks originating from the samples containing a single sequence. The overlaid spectra (Figure 5.3a) show only the presence of broad monomeric peaks. Figure 5.3b showcases the overlaid spectra of the binary mixtures. The blue peaks correspond to the α/β mixture, which are identical to the peaks observed for the individual sequences, thus indicating the absence of $\alpha_2\beta/\alpha\beta_2$ trimers at this temperature. On the other hand, both the α/γ and β/γ mixtures show distinct trimeric peaks, green and red respectively. These peaks correspond to the molecular fingerprint of the competing states of alternative composition

and can be used to investigate their presence or absence from the ternary mixture. The annealed **α/β/γ** mixture shows only 3 distinct heterotrimeric cross peak of equal intensity as well as residual monomeric peaks. The three peaks in this spectrum (Figure 5.3 c) can be unambiguously assigned to the **α**, **β** and **γ** chains.

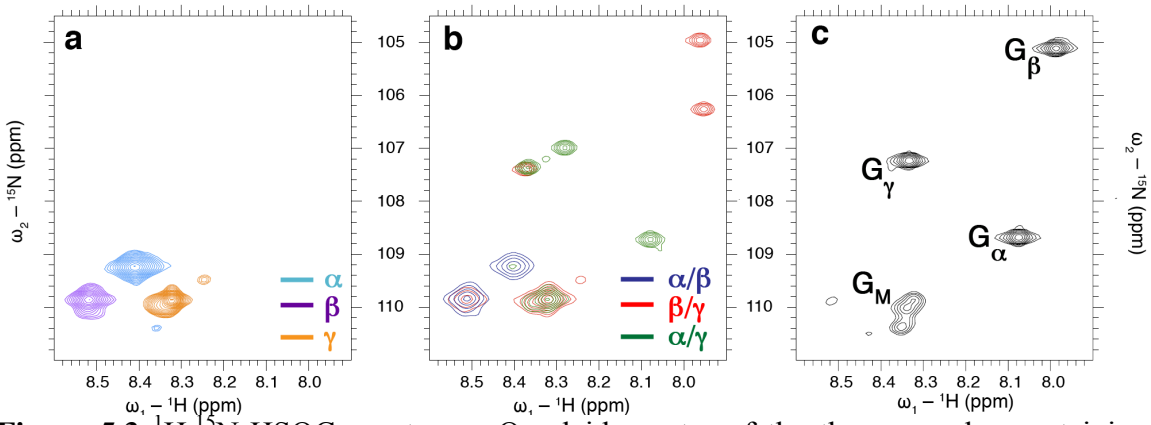


Figure 5.3 $^{1}H,^{15}N$-HSQC spectra. **a**, Overlaid spectra of the three samples containing individual sequences. **b**, Overlaid spectra of the three samples containing binary mixtures. **c**, Spectrum of the annealed ternary mixture. All experiments performed at 37 °C.

The chemical shift of the labeled glycines (position 15 in each chain) was determined using a combination of $^{1}H,^{15}N$-HSQC , $^{1}H,^{1}H$-NOESY, $^{1}H,^{1}H$-TOCSY and 2D $^{1}H,^{1}H$-NOESY-^{15}N-HSQC spectra at 37 °C. In the case of peptide α (figure 5.4) the chemical shift of K14(Hα) proton, K14(Hγ1) and K14(Hγ2) can be identified using the sequential NOE to the labeled G15(NH) in the $^{1}H,^{1}H$-NOESY-^{15}N-HSQC spectrum as well as the intra-residue NOEs and TOCSY cross peaks arising from the unlabeled lysine residue. Although the intra-residue peaks K14(Hγ1)-K14(Hγ2) and K14(Hγ1)-K14(Hα) in Figure 5.4a cannot be unambiguously assigned because most of the lysine side-chains present similar shifts for the γ-methylene protons, their unique aliphatic chemical environment gives rise to a characteristic chemical shift that can be used to unequivocally identify the labeled glycine corresponding to the **α** chain since none of the remaining

sequences have lysine residues preceding the labeled position. Similarly, in the case of peptide **β** (Figure 5.5) the chemical shift of O14(Hα), O14(Hβ1), O14(Hβ2) can be identified using the sequential NOE to the labeled G15(NH) in the ^{1}H,^{1}H-NOESY-^{15}N-HSQC spectrum as well as the intra-residue NOEs and TOCSY cross peaks arising from the unlabeled hydroxyproline residue, O14(Hβ1)-O14(Hα) and O14(Hβ2)-O14(Hα). The chemical shift of D16(NH) can be identified from the sequential D16(NH)-G15(Hα1) and D16(NH)-G15(Hα2) NOEs in the ^{1}H,^{1}H-NOESY spectrum, these are necessary to differentiate sequences β (O^{14}G^{15}D^{16}) and γ (O^{14}G^{15}P^{16}). Finally, in the case of peptide **γ** (Figure 5.6) the chemical shift of O14(Hα) proton, O14(Hβ1), O14(Hβ2) can be identified using the sequential NOE to the labeled G15(NH) in the ^{1}H,^{1}H-NOESY-^{15}N-HSQC spectrum as well as the intra-residue NOEs and TOCSY cross peaks arising from the unlabeled hydroxyproline residue, O14(Hβ1)-O14(Hα) and O14(Hβ2)-O14(Hα).

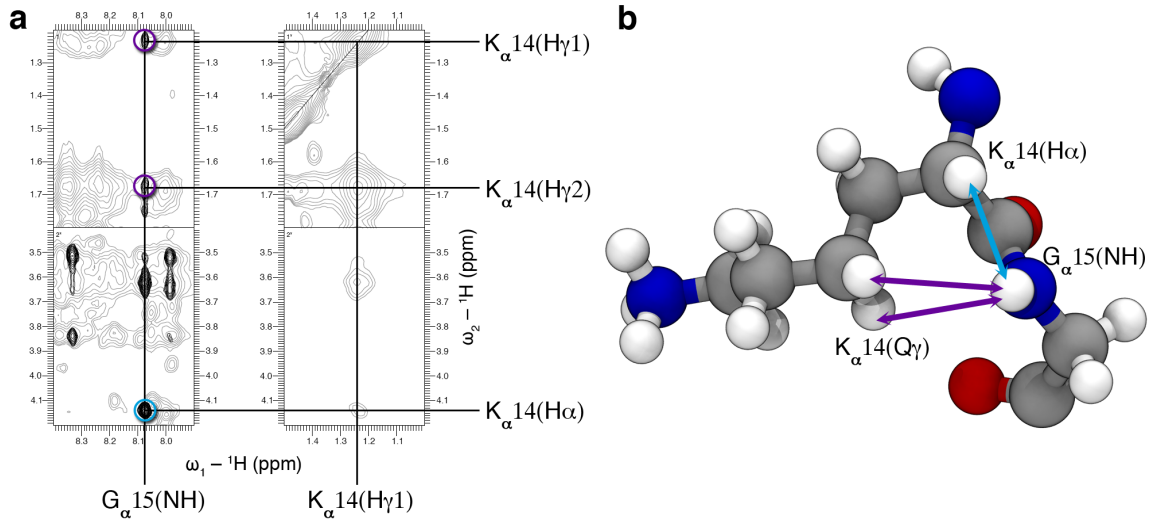


Figure 5.4. Sequential assignment of the α peptide. **a**, overlaid ^{1}H,^{1}H-NOESY(gray) and 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC (black) spectra of the α/β/γ mixture with sequential NOEs highlighted by colored circles; the aliphatic region of the spectrum is also depicted to show intra-residue NOEs in the lysine side-chain. **b**, Atomic model showing the KG dipeptide that gives rise to the sequential NOEs in **a**. Colored circles in **a** correspond to colored arrows in **b**.

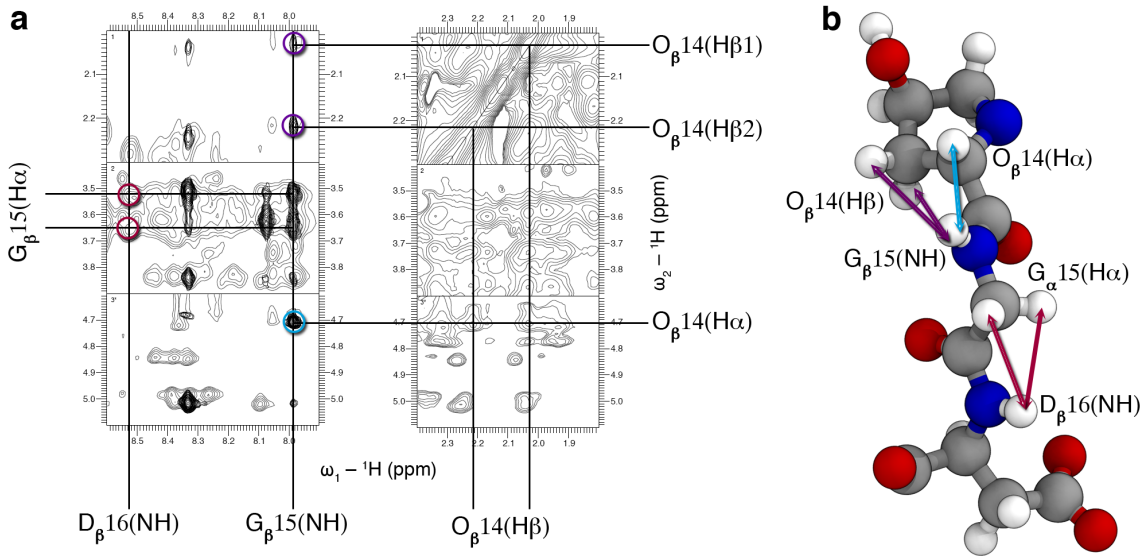


Figure 5.5 Sequential assignment of the β peptide. **a**, overlaid ^{1}H,^{1}H-NOESY(gray) and 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC (black) spectra of the α/β/γ mixture with sequential NOEs highlighted by colored circles; the aliphatic region of the spectrum is also depicted to show intra-residue NOEs in the hydroxyproline ring. **b**, Atomic model showing the OGD tripeptide that gives rise to the sequential NOEs in **a**. Colored circles in **a** correspond to colored arrows in **b**.

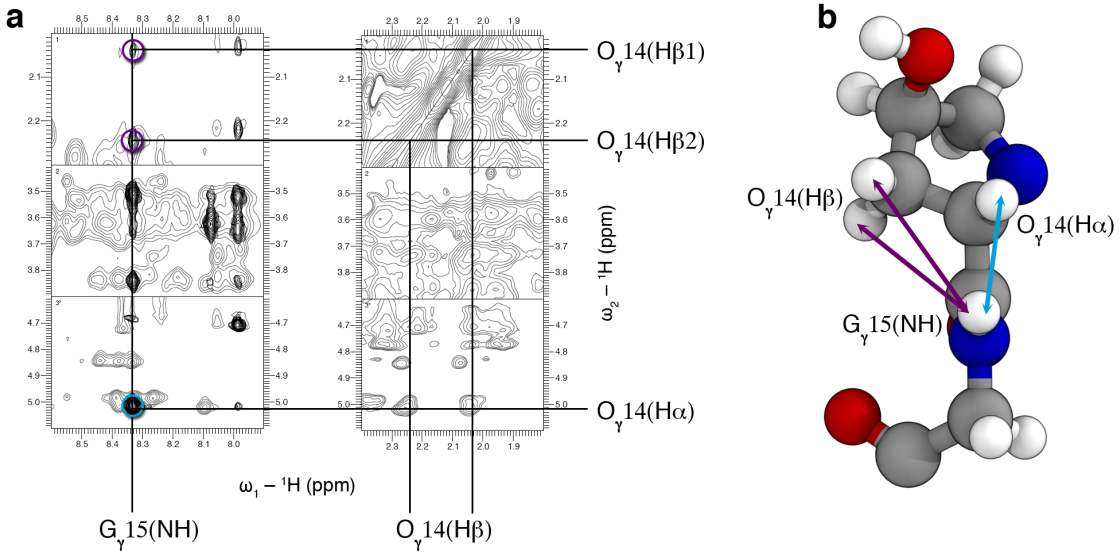


Figure 5.6. Sequential assignment of the γ peptide. **a**, overlaid ^{1}H,^{1}H-NOESY(gray) and 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC (black) spectra of the α/β/γ mixture with sequential NOEs highlighted by colored circles; the aliphatic region of the spectrum is also depicted to show intra-residue NOEs in the hydroxyproline ring. **b**, Atomic model showing the OG dipeptide that gives rise to the sequential NOEs in **a**. Colored circles in **a** correspond to colored arrows in **b**.

The last step required to validate our design protocol is to experimentally characterize the chain stagger or register of the three peptide strands. For this purpose we utilize the ^{1}H,^{1}H-NOESY-^{15}N-HSQC spectrum (Figure 5.7a). The spectrum shows symmetric $G_{\beta}15(NH)$-$G_{\alpha}15(NH)$ and $G_{\beta}15(NH)$-$G_{\gamma}15(NH)$ cross peaks that arise from the glycine packing in the core of the helix, and position peptide **β** as the middle chain. The stagger of peptide **γ** with respect to **β** is evidenced by the $O_{\gamma}14(Ha)$-$G_{\beta}15(NH)$ NOE, which puts the **γ** chain in the lagging position with a one amino acid stagger with respect to **β**. Together these peaks show that the register of the self-assembling ABC heterotrimer is **α•β•γ**, in agreement with our design protocol. An *in silico* model in Figure 5.7b shows the spatial arrangement of the amino acids utilized for register determination.

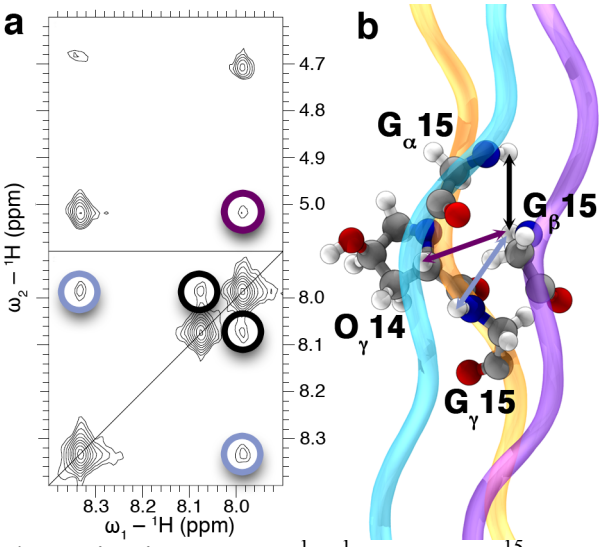


Figure 5.7. Register determination. **a**, 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC spectrum of the annealed ternary mixture at 37 °C. **b**, *In silico* model showing the backbone NOEs highlighted in **a** with the peptide a colored cyan, b strand purple and g orange. Colored circles in **a** correspond to the colored arrows in **b**

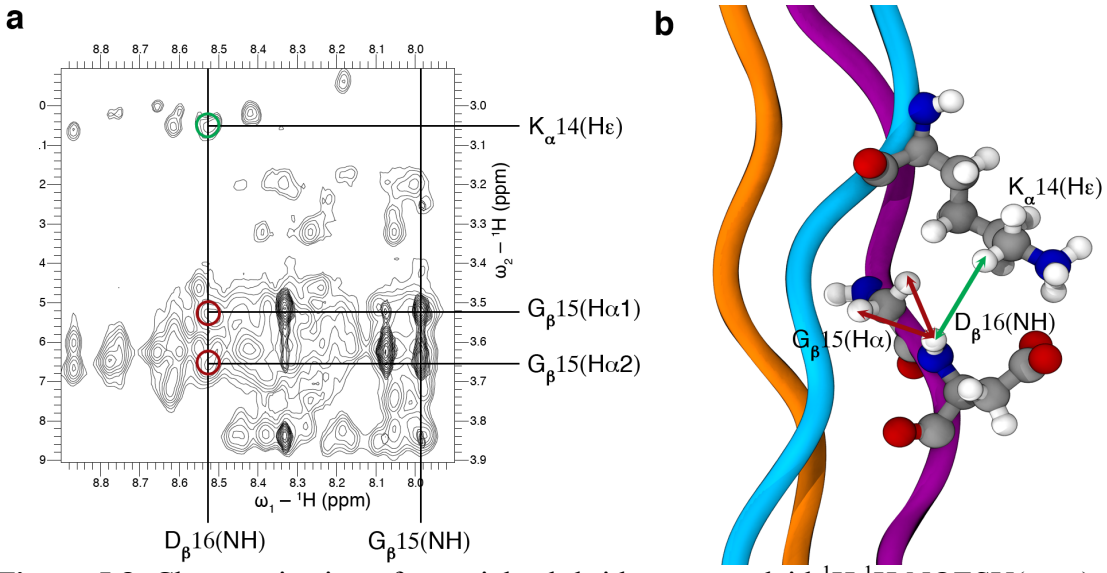


Figure 5.8. Characterization of an axial salt bridge. **a**, overlaid ^{1}H,^{1}H-NOESY(gray) and 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC (black) spectra of the a/b/g mixture with sequential and inter-chain NOEs highlighted by colored circles. **b**, Homology model showing the spatial arrangement of the atoms that gives rise to the NOEs in **a** with the peptide a depicted cyan, b purple and g orange. Colored circles in **a** correspond to colored arrows in **b**.

Although the chemical shift of most charged amino acids cannot be unambiguously determined a combination of ^{1}H,^{1}H-NOESY and 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC spectra at 37 °C can be used to assign one of the axial salt bridges that stabilize our designed triple helix. Figure 5.8 shows the resonances used in the characterization of this inter-strand interaction between K_α14 and D_β16. The chemical shift D_β16(NH) can be identified using the sequential NOE to the labeled G_β15(Hα) in the NOESY spectrum. There is also a clear resonance between D_β16(NH) and a lysine ε-methylene. Most ε protons have comparable chemical shifts and thus the assignment can only be made considering the sequence, but this resonance is characteristic of K-D axial salt bridges and validates our design hypothesis by showing that axial salt bridges are indeed present in our system.

Our experimental characterization of the peptide sequences generated by the GA agree with the initial hypothesis that our minimalistic energy function captures the

dominant contributions to the chemical potential of triple helical peptide mixtures within the set sequence constraints. Although other factors besides the formation of axial salt bridges, such as electrostatic repulsion and contributions of different single and double substitutions, could be incorporated to improve the accuracy of the model, their relative strength needs to be carefully weighted for triple helical systems. Nanda et al.[9] recently utilized a comparable sequence-based scoring function adapted from coiled-coil design and a simulated annealing Monte Carlo search algorithm to tackle the problem of compositional control in ABC-type heterotrimers. Their study generated sequences with significantly lower thermal stability, approximately 30 °C, and does not differentiate based on register. Additionally, that study explored a larger sequence space by allowing lysine residues in the X position as well as aspartic acid residues in the Y position, relied heavily on repulsion between amino acids of identical charge and weighted equally all pair-wise configurations that allowed for geometrical contacts between oppositely charged residues. We believe that the main reason for the difference in melting temperature between the two designed peptide systems lies in the fact that axial salt-bridges dominate the energy landscape. If other interactions are to be included within the model, their relative contributions need to be weighted more effectively. Establishing proper weighting for additional pairwise interactions with collagen triple helices is an important goal for full understanding of the structure and self-assembly of collagen helices.

5.3 Conclusions

This chapter presents a minimalistic approach to the design of heterotrimeric triple helical peptides. By constraining the sequence space and understanding what amino

acid configurations are stabilizing and destabilizing for triple helices within those constraints, we are able to generate sequences that form ABC-type triple helices with a high thermal stability and control over the relative stagger of the peptide chains within the helix. Our automated sequence selection algorithm is successful because of the balance struck in our scoring function between the destabilization induced on triple helical assemblies by changing conformationally restricted imino acids to ionizable residues and the stabilization conferred upon the formation of axial inter-strand ionic hydrogen bonds.

Currently, the registration process in heterotrimeric members of the collagen family, such as types I, IV and IX, is poorly understood. It is thought that globular domains capable of setting the composition play a dominant role in this process, but our synthetic analog shows that it is indeed possible to control the register of a triple helical system using information encoded solely in the collagenous domain. Our simple scoring function can be expanded to account for other amino acids, and their respective interactions, to study the stability and specificity profiles of natural heterotrimeric collagens and shed light on their registration mechanism and the role that triple helical domains have in that process.

Finally, this methodology can be used to generate flanking regions for heterotrimeric host-guest peptide studies. The designed N- and C-terminal domains can be used to set the composition and chain register as well as drive triple helix formation, similar to POG triplets in homotrimers, and the guest domain can be used to include wild type sequences or mutants opening a whole new chapter in the study of the biochemistry and biophysics of this important protein family.

5.4 Experimental

Scoring Function – Each triple helical sequence composed, of 30 amino acids per chain is encoded as a 60-bit string, odd bits represents the X positions and even the Y positions, glycines are excluded as they are not designable amino acids in this context. Bits 1-20 represent chain A, 21-40 chain B and 41-60 chain C. Each sequence is scored according to equation (1) by counting the number of charged residues and axial salt bridges. The e_1/e_2 ratio in (1) can be utilized to explore different regions in sequence space however we utilize a value of 1 for e_1 and 2 for e_2, with the rationale that a paired salt bridge approximately cancels out the destabilization caused by the point mutations.[12]

Genetic Algorithm – We start with a population of 80 random 60-bit strings. The fitness, c, of each member of the population is calculated using the energy score of the sequence, the energy score of the most stable member of the competing state ensemble and a value of 1 for the proportionality constant. The competing state ensemble is generated from the 26 remaining combinations of the three segments corresponding to chains A, B and C. A second population is generated by picking members of the initial random population using a metropolis-type criterion with a probability, p, proportional to $\exp[-(c-c^{min})/t]$, with t=1. All members of this set are paired and a new generation is produced using variable, randomly-selected single crossover combinations of the parent sequences. A mutation rate of 0.5% is used to keep genetic variability and it is increased ten-fold if both parent sequences happen to be identical. This algorithm is repeated until a target fitness is met or a set number of generations produced.

Peptide Synthesis – Peptides were synthesized with an Advanced Chemtech Apex 396 synthesizer using Fmoc solid phase peptide chemistry and a Rink MBH amide resin. During the automated procedure, a manual addition of ^{15}N-labelled glycine, purchased form Cambridge Isotope Laboratories, was carried out in position 15. All peptides include a tyrosine (for concentration determination) and a glycine spacer at the C-terminus and are C-terminally amidated and N-terminal acetylated to eliminate any competing electrostatic interaction at the termini. The peptides were purified on a Varian PrepStar220 HPLC with a preparative reverse phase C-18 column using a linear water/acetonitrile gradient each containing 0.05% TFA and analyzed by ESI-TOF mass spectrometry on a Bruker microTOF instrument (available in the supplementary information).

Sample Preparation – Concentration of stock solutions was determined by UV/Vis absorption at 275 nm using a molar extinction coefficient of 1400 cm^{-1}/M. All peptide mixtures were prepared, annealed at 85 °C and incubated for a week at room temperature before experimental measurements were performed.

Circular Dichroism – CD experiments were performed with a Jasco J-810 spectropolarimeter equipped with a Peltier temperature control system. Samples were prepared to a total concentration of 300 μM in 10 mM phosphate buffer at pH 7 by mixing the desired peptides in the appropriate ratio (1:1 for binary samples and 1:1:1 for the ternary sample). Spectra were acquired between 215-250 nm to locate the maximum near 222 nm, which was monitored during unfolding experiments. Melting curves were

performed from 5 to 85 °C with a heating rate of 10 °C/hr. The first derivative of the melting curve was taken in order to determine the melting temperature (T_m) of the sample, which we define as the minimum in the derivative graph. The molar residual ellipticity (MRE) is calculated from the measured ellipticity using the equation:

$$[\theta] = \frac{\theta \times m}{c \times l \times n_r}$$

where θ is the ellipticity in mdeg, m is the molecular weight in g/mol, c is the concentration in mg/mL, l is the pathlength of the cuvette in cm, and n_r is the number of amino acids in the peptide.

Nuclear Magnetic Resonance – NMR experiments were recorded in an 800 MHz Varian at 37 °C spectrometer equipped with a triple resonance probe. Samples were prepared at two different total peptide concentrations (1 mM for samples containing a single peptide and 3 mM for peptide mixtures) in a10 mM phosphate buffer at pH 7 and a 9:1 ratio of H_2O to D_2O. The spectra were processed using NMRpipe[20] and analyzed using ccpnmr[21] . A list of all experiments performed as well as acquisition and processing parameters are available in the supplementary information. Each sample containing a mixture of peptides was characterized using 2D total correlated spectroscopy (TOCSY), nuclear Overhauser effect spectroscopy (NOESY), $^1H,^{15}N$-heteronuclear single quantum coherence (HSQC) and 2D $^1H,^1H$-NOESY-^{15}N-HSQC experiments while samples containing single sequences were characterized using $^1H,^{15}N$-HSQC spectra at 37 °C. Additional $^1H,^{15}N$-HSQC spectra for the ternary mixture were acquired at 5 °C, 25 °C and 45 °C. TOCSY spectra with a 50 ms spinlock duration at 8 kHz were acquired with a total of 1700

complex points recorded in 8 scans for the directly acquired dimension while 500 increments were used in the indirect dimension. NOESY spectra with a 100 ms mixing time were acquired with a total of 1700 complex points recorded in 8 scans for the directly acquired dimension while 500 increments were used in the indirect dimension. A square spectral window of 1000 Hz was used for all homonuclear spectra. For the 2D ^{1}H,^{1}H-NOESY-^{15}N-HSQC spectra a mixing time of 100 ms was used and a total of 1600 complex points in 32 scans for the direct dimension and 400 increments for the indirect dimension were acquired using a spectral window of 8000 Hz for the direct dimension and 7200 for the indirect dimension. A total of 1208 complex points in 32 scans for the direct dimension and 100 increments in the indirect dimension were acquired for the ^{1}H,^{15}N-HSQC experiments using a spectral window of 10000 Hz in the hydrogen dimension and 1200 Hz in the nitrogen dimension. Square Cosine bell window functions were used as apodization functions and the data was zero-filled to the next power of two in both dimensions. Drift and baseline corrections were applied when necessary.

Homology Modeling – A model of the a•b•g register was prepared using the Rosetta software suite[22] using the crystal structure of a tiple helical peptide (pdb id: 1K6F) as a template.[23] After mutating the residues using the fixed backbone design application rounds of flexible backbone modeling using the backrub and side chain relaxation were carried out. Because this particular macromolecular software suite lacks explicit electrostatic scoring terms but includes directional hydrogen-bonding potentials, distance constraints were placed upon the charged residues to bias them toward the axial salt

bridge conformation since this is expected based on the D(NH)-K(He) resonances observed in the $^{1}H,^{1}H$-NOESY spectrum.

5.5 References

(1) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. *Science* **2003**, *302*, 1364-1368.

(2) Fleishman, S. J.; Whitehead, T. A.; Ekiert, D. C.; Dreyfus, C.; Corn, J. E.; Strauch, E. M.; Wilson, I. A.; Baker, D. *Science* **2011**, *332*, 816-821.

(3) Stranges, P. B.; Machius, M.; Miley, M. J.; Tripathy, A.; Kuhlman, B. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 20562-20567.

(4) Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S. *Science* **1998**, *282*, 1462-1467.

(5) Grigoryan, G.; Reinke, A. W.; Keating, A. E. *Nature* **2009**, *458*, 859-864.

(6) Zaccai, N. R.; Chi, B.; Thomson, A. R.; Boyle, A. L.; Bartlett, G. J.; Bruning, M.; Linden, N.; Sessions, R. B.; Booth, P. J.; Brady, R. L.; Woolfson, D. N. *Nat Chem Biol* **2011**, *7*, 935-941.

(7) Das, R.; Baker, D. *Annu. Rev. Biochem.* **2008**, *77*, 363-382.

(8) Xu, F.; Zhang, L.; Koder, R. L.; Nanda, V. *Biochemistry* **2010**, *49*, 2307-2316.

(9) Xu, F.; Zahid, S.; Silva, T.; Nanda, V. *J. Am. Chem. Soc.* **2011**,

(10) Fallas, J. A.; Dong, J.; Tao, Y. J.; Hartgerink, J. D. *J. Biol. Chem.* **2012**, *287*, 8039-8047.

(11) Persikov, A. V.; Ramshaw, J. A.; Brodsky, B. *J. Biol. Chem.* **2005**, *280*, 19343-19349.

(12) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2005**, *44*, 1414-1422.

(13) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 15034-15041.

(14) Persikov, A. V.; Ramshaw, J. A.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2000**, *39*, 14960-14967.

(15) Gurry, T.; Nerenberg, P. S.; Stultz, C. M. *Biophys. J.* **2010**, *98*, 2634-2643.

(16) Fallas, J. A.; Lee, M. A.; Jalan, A. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2012**, *134*, 1430-1433.

(17) Grigoryan, G.; Zhou, F.; Lustig, S. R.; Ceder, G.; Morgan, D.; Keating, A. E. *PLoS Comput Biol* **2006**, *2*, e63.

(18) Havranek, J. J.; Harbury, P. B. *Nat Struct Biol* **2003**, *10*, 45-52.

(19) Leaver-Fay, A.; Jacak, R.; Stranges, P. B.; Kuhlman, B. *PLoS One* **2011**, *6*, e20937.

(20) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277-293.

(21) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, P.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. *Proteins* **2005**, *59*, 687-696.

(22) Leaver-Fay, A. et al. *Methods Enzymol.* **2011**, *487*, 545-574.

(23) Berisio, R.; Vitagliano, L.; Mazzarella, L.; Zagari, A. *Protein Sci.* **2002**, *11*, 262-270.

Chapter 6: Conclusions

The main focus of the work described in this thesis has been the design and structural characterization of self-assembling collagen triple helices with control over both helical composition and register. This is a challenging task mainly due to the number of possible states that can be populated in a mixture of three peptides with a high propensity to fold into collagen-like triple helices and the fact that our understanding of the stabilization of homotrimeric helices is based on single amino acid contributions and thus cannot be used to bias the self-assembly towards a particular heterotrimer rather than a mixture of different triple helices.

In order to achieve our goal we studied sequence-structure relationships within this protein fold, specifically how different arrangements of oppositely charged amino acids impact the thermal stability of a collagen triple helix. We were characterized two distinct interaction geometries for ionizable residues in triple helices. We named these interactions axial and lateral and noted that they have different sequence requirements depending on the relative stagger or register of the interacting chains within the helix (Figure 6.1). Furthermore, we are able to asses their effect on the thermal stability of self-assembled CMPs. Despite the fact that in both contact geometries the oppositely charged moieties can form ionic hydrogen bonds, the axial interaction provides a large increase in stability while the lateral interaction behaves approximately as one would expect from the addition of the thermal destabilization caused by each point mutation. Although we only explored a limited sequence space in this thesis, it should be possible to expand the amino acid identities and include different residues in the X, Y, X', Y', positions of a host-guest peptide library to study the impact of different axial and lateral pairs on the thermal

stability of triple helical proteins. An exhaustive experimental survey of all possible canonical amino acid pairings would be impractical but a simple computational screening algorithm can be implemented to find promising candidates. Combining homology models with either a rotamer repacking algorithm or a short molecular dynamics simulation followed by gradient minimization, favorable atomic contacts between different residue identities can be identified. Although this does not guarantee that there will be an effect on the thermal stability of triple helical proteins, as observed for the axial salt bridges, it can be used as a starting point to select sequences that have a potential to exhibit an interesting behavior and characterize them experimentally.

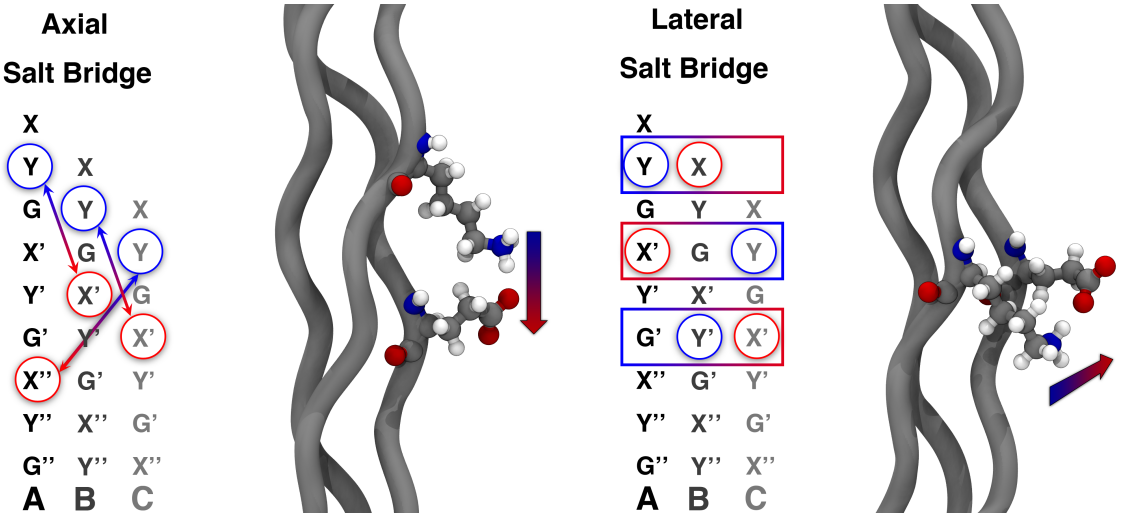


Figure 6.1. Schematic representation of axial and lateral salt bridges in triple helical proteins including the relative position of interacting amino acids in aligned triple helical sequences.

Another important accomplishment presented in this work is the in-depth structural characterization of heterotrimeric collagen triple helices. We modified the experimental methodology available for homotrimeric helices and expanded it to address problems unique to their heterotrimeric counterparts. Using this approach, we were able to show that the formation of interstrand axial salt-bridges can be used to drive the

formation of triple helical heterotrimers. Furthermore, the developed experimental methodology is able to show the presence of competing states that were missed utilizing biophysical techniques, indicating that the rational supramolecular design strategy developed initially developed in the laboratory, although able to direct the self-assembly of the desired heterotrimer, was unsuccessful in encoding sufficient information in the amino acid sequences to select a single state, leading to a system composed of a mixture of helices.

In order to improve upon the selectivity towards heterotrimers observed in this system, a rational approach to modify the sequences was developed. Utilizing the fact that axial salt-bridges have different register-specific sequence requirements additional charged residues were included in positions that allowed for the formation of axial contacts between the lagging and leading chain in the desired target state but not in any of the other possible triple helices while simultaneously destabilizing the main competing state, a homotrimeric helix (Figure 6.2). This change led to a system in which only ABC heterotrimers folded but the presence of at least two registers of the desired composition was confirmed.

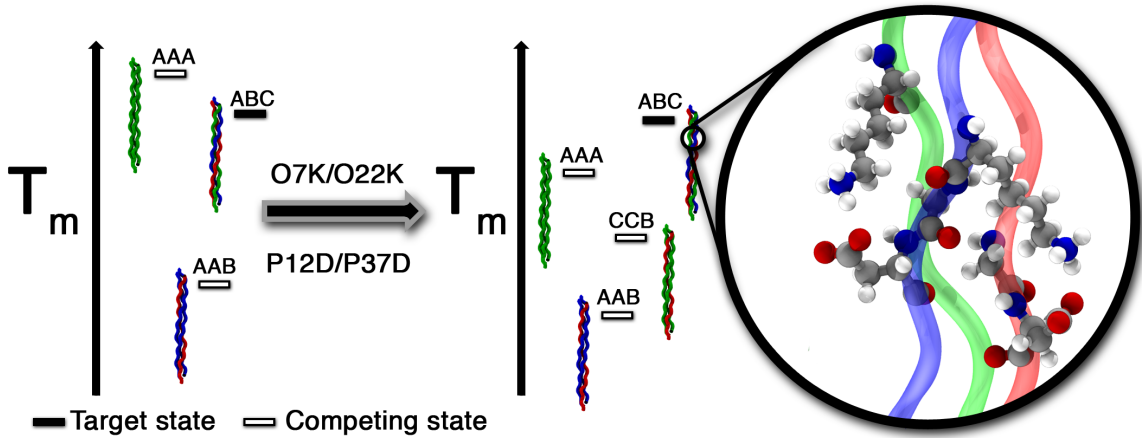


Figure 6.2. Rational design strategy for the design of single-composition ABC collagen heterotrimers.

In order to realize the final goal, the formation of a single-register ABC heterotrimer, a computational design strategy was utilized. We developed a simple, sequence based stability scoring function and used it to maximize the energy gap between the desired target and the next most stable competing state while minimizing the stability of the target state (Figure 6.3). This approach succesfully yielded sequences that fold into highly stable triple helices that are able to select one of the 27 states available in the ternary mixture. The strategy is mainly successful because, at least to a first-order approximation, the axial salt bridges govern the association landscape of triple helical peptides in the sequence space explored in this thesis. Thus, by combining knowledge gained using structural biology techniques with rational supramolecular considerations and basic assumptions about the energy landscape of triple helical peptides we were able to generate sequences using an automated procedure that contain enough information to fold into register-specific collagen heterotrimers.

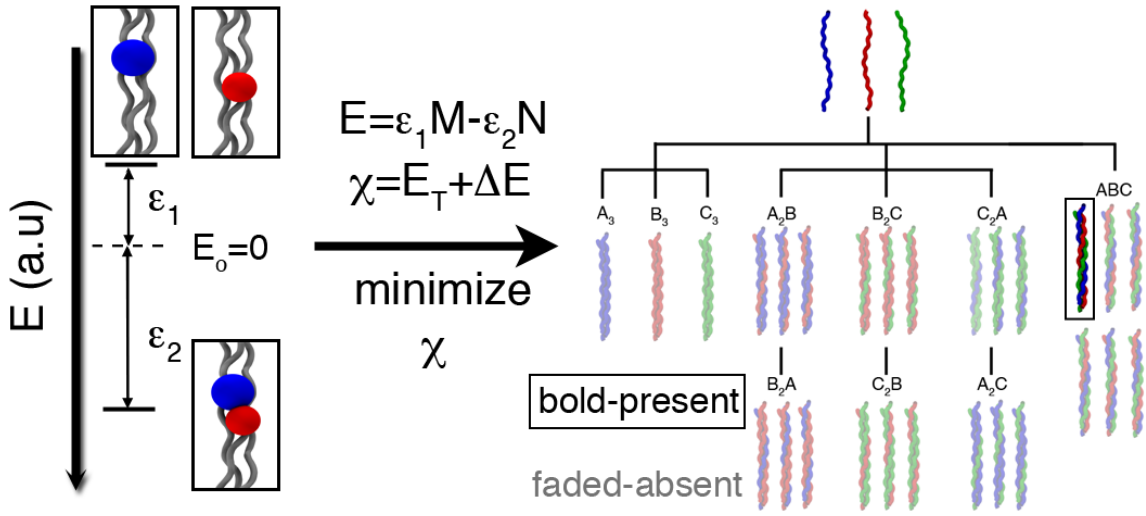


Figure 6.3. Computational design strategy for the design of single-register ABC collagen heterotrimers.

The sequence-based scoring function described here can be expanded together with the work outlined earlier on pair-wise interactions in triple helical peptides to include different amino acid identities. Such a function can be used to calculate stability profiles, expanding on work done for homotrimeric collagens, and specificity profiles of the heterotrimeric members of the collagen protein family. The idea of calculating specificity profiles, the difference between the most stable and 2nd most stable registers of a particular sequence stretch of a heterotrimeric member of the collagen family, is interesting as it could be used to locate natural sequences with enough information to direct the self-assembly of triple helices. The role of collagenous domains in the folding process of natural collagens is poorly understood. The initial association happens trough trimeric globular domains and they play a dominant role in choosing the composition of the helices but how or if they have a role in choosing a particular register in the case of heterotrimers is still an open question. Specificity profiles, coupled with experimental characterization of promising sequences, could be used to start unraveling the folding mechanism of the heterotrimeric members of this important protein family.

Appendix 1: Publication List

The work discussed in this thesis has been highlighted in several publications. They will be listed in order of the chapters that are based on them:

1. **Fallas, J. A.**; O'Leary, L. E.; Hartgerink, J. D. *Chem. Soc. Rev.* **2010**, *39*, 3510-3527.
2. **Fallas, J. A.**; Dong, J.; Tao, Y. J.; Hartgerink, J. D. *J. Biol. Chem.* **2012**, *287*, 8039-8047.
3. **Fallas, J. A.**; Gauba, V.; Hartgerink, J. D. *J. Biol. Chem.* **2009**, *284*, 26851-26859.
4. **Fallas, J. A.**; Lee, M. A.; Jalan, A. A.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2012**, *134*, 1430-1433.
5. **Fallas, J.A.**; Hartgerink J.D. *submitted* **2012**.

The research discussed in this thesis corresponds to the projects were I was the primary investigator. I also worked in collaboration with other members of the Hartgerink laboratory in their projects. My contributions related to my expertise in structural biology techniques (both NMR spectrocospy and X-ray diffraction) and computational algorithms to generate triple helical homology models. That work was also highlighted in several publications, listed chronologically:

1. Russell, L. E.; **Fallas, J. A.**; Hartgerink, J. D. *J. Am. Chem. Soc.* **2010**, *132*, 3242-3243.
2. O'Leary, L. E.; **Fallas, J. A.**; Hartgerink, J. D. *J. Am. Chem. Soc.* **2011**, *133*, 5432-5443.
3. O'Leary, L. E.; **Fallas, J. A.**; Bakota, E. L.; Kang, M. K.; Hartgerink, J. D. *Nat. Chem.* **2011**, *3*, 821-828.
4. Wei, F.; **Fallas, J.A.**; Hartgerink, J.D. *submitted* **2012**

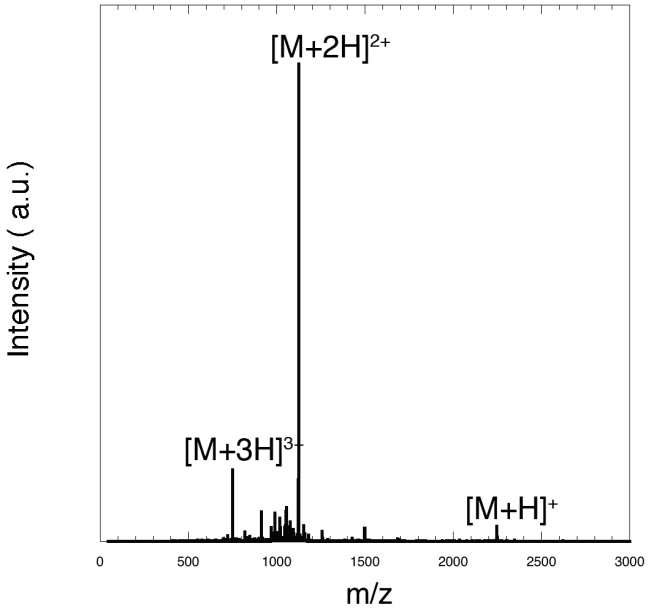
Appendix 2: Peptide Library

Sequence*	Abbreviation
$(POG)_3PK\mathbf{G}EOG(POG)_3$	**KGE**
$(POG)_3PK\mathbf{G}DOG(POG)_3$	**KGD**
$(PKG)_{10}$	**K**
$(DOG)_{10}$	**D**
$(POG)_{10}$	**O**
$(PKG)_4P\mathbf{\underline{KGP}}KG(PKG)_4$	**K***
$(DOG)_4DO\mathbf{\underline{GD}}OG(DOG)_4$	**D***
$(POG)_4PO\mathbf{\underline{GP}}OG(POG)_4$	**O***
YGPKGPKGPKGDKGPK**G**PKGPKGPKGDKGPKG	**A**
$YG(DOG)_4DO\mathbf{G}(DOG)_5$	**B**
$YG(EOG)_4EO\mathbf{G}(EOG)_5$	**B1**
YGPOGPKGPOGPOGPO**G**POGPKGPOGPOGPOG	**C**
PKGPKGDOGPOGDK**G**DKGPKGPOGDKGPOGGY	**α**
POGDOGDKGPOGPO**G**DKGDOGDKGPKGDOGGY	**β**
PKGPOGPKGDKGPO**G**POGDKGPOGDOGDOGGY	**γ**

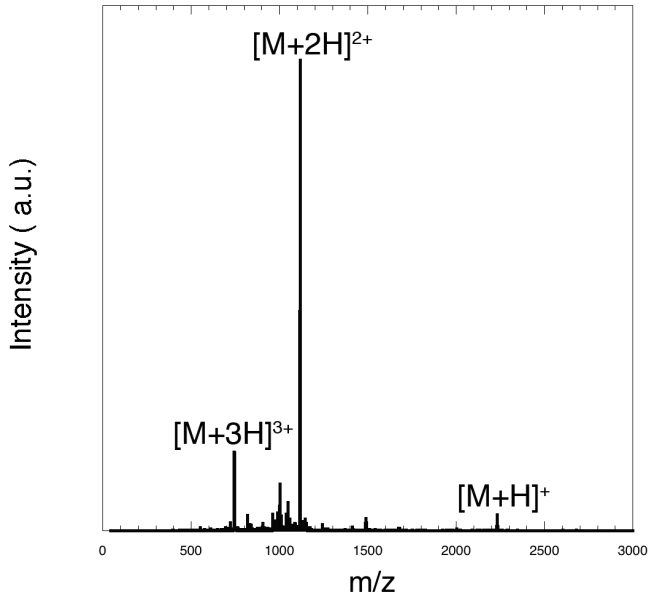
* Bold and underlined amino acids are uniformly 15N- and 13C-labeled; bold amino acids are uniformly 15N-labeled.

Appendix 3: Mass Spectrometry of the Peptides Synthesized

ESI-TOF MS data for the peptides synthesized for Chapter 2

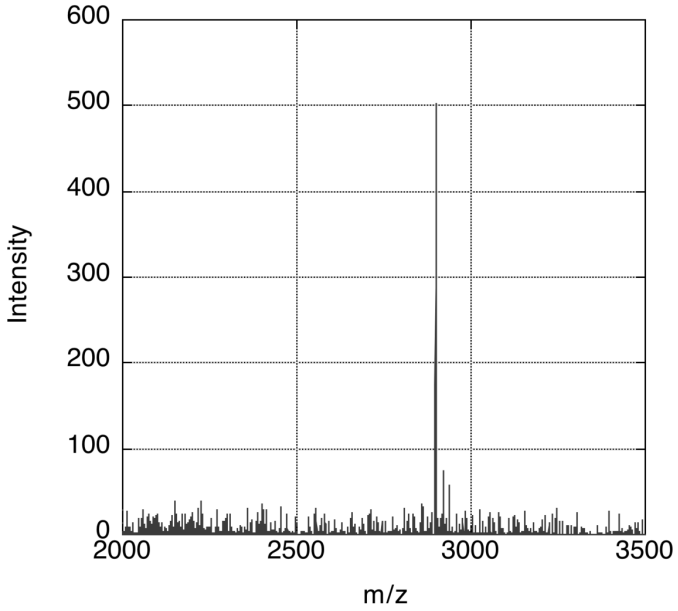


Peptide **KGE** expected: 1123.0$[M+2H]^{2+}$, Observed = 1122.9$[M+2H]^{2+}$

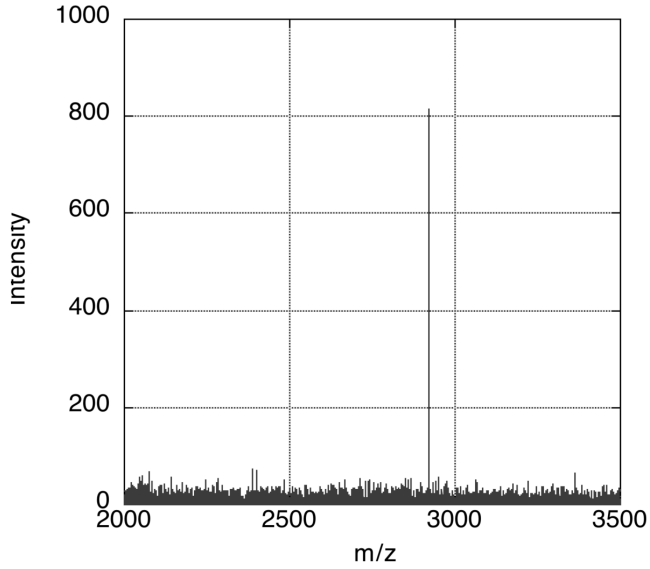


Peptide **KGD**expected: 1116.0$[M+2H]^{2+}$, Observed = 1115.6 $[M+2H]^{2+}$

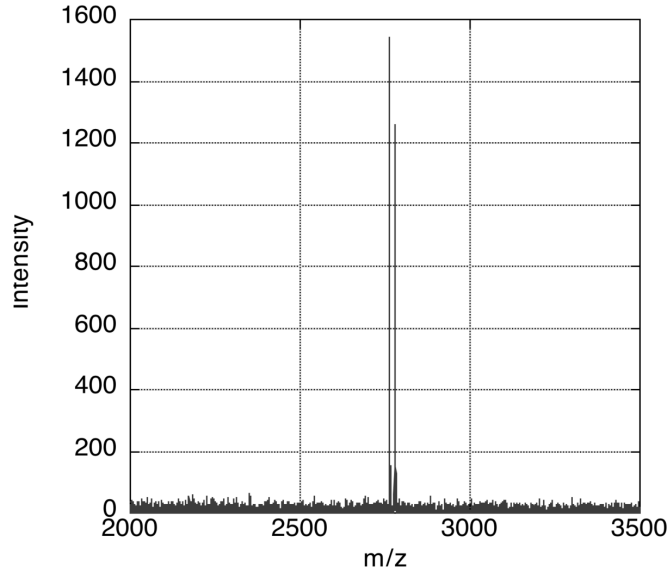
MALDI-TOF MS data for the peptides synthesized for Chapter 3



Peptide **K*** Expected: 2897.7[H+], Observed: 2898.1[H+]

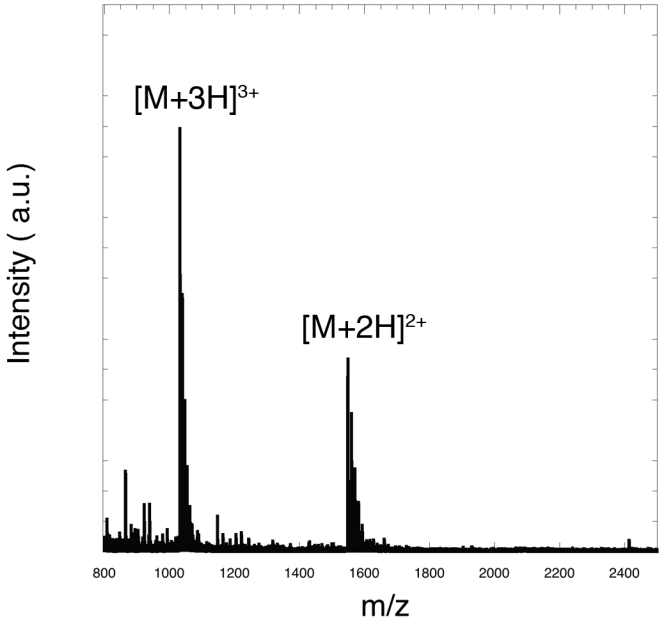


Peptide **O*** Expected: 2762.2[Na+], Observed: 2762.4[Na+]

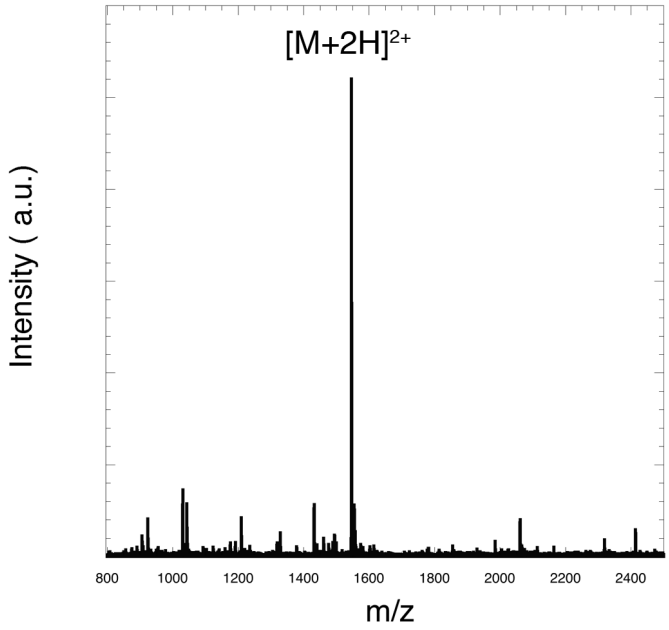


Peptide **D*** Expected: 2919.6[H+], Observed: 2920.1[H+].

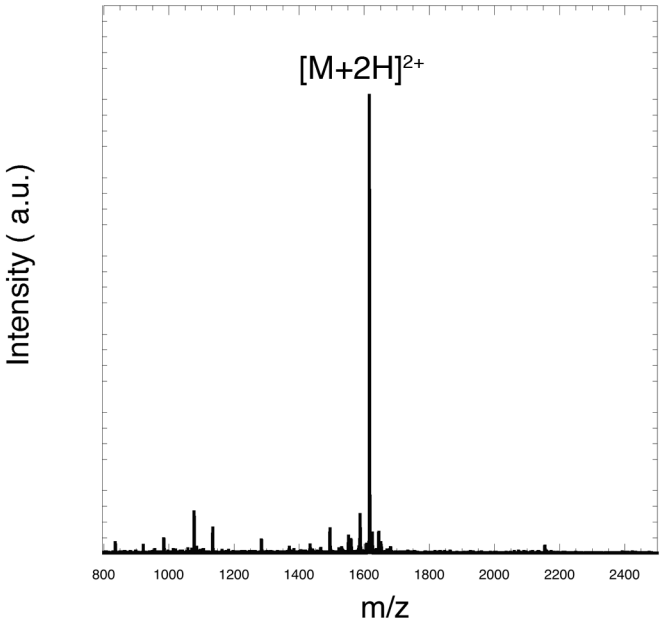
ESI-TOF MS data for the peptides synthesized for Chapter 4



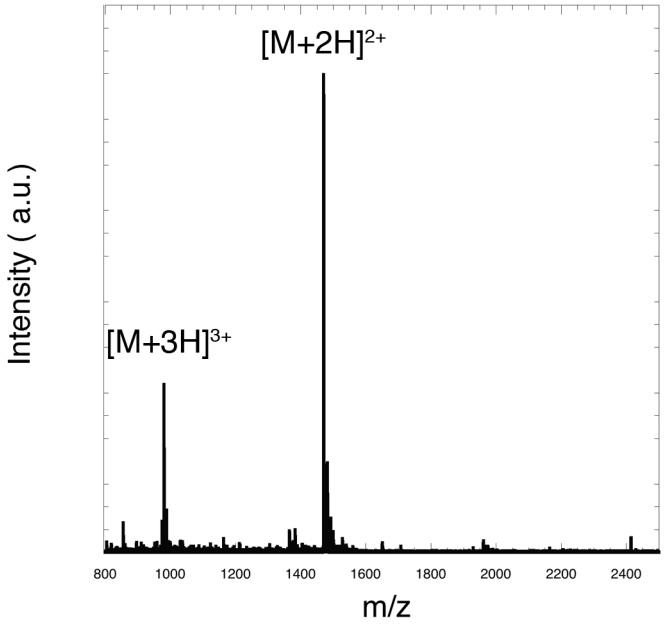
Peptide **A** Expected: 1032.9 $[M+3H]^{3+}$, Observed = 1032.8 $[M+3H]^{3+}$



Peptide **B** Expected: 1546.0 $[M+2H]^{2+}$, Observed = 1545.8 $[M+2H]^{2+}$

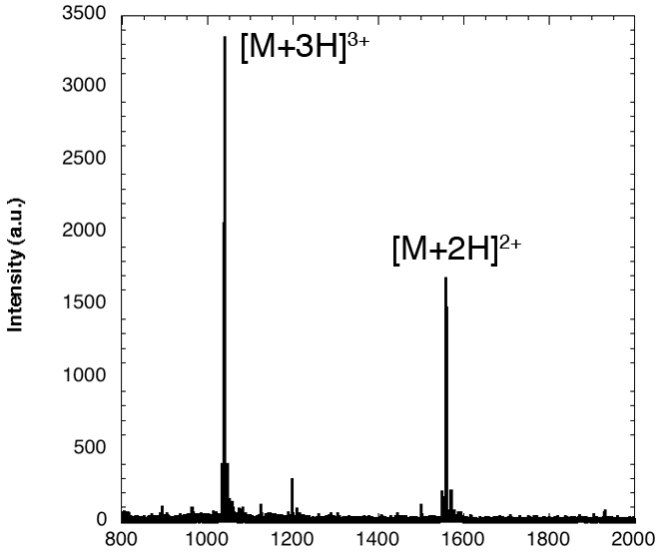


Peptide **B1** Expected: 1616.1 $[M+2H]^{2+}$, Observed = 1615.8 $[M+2H]^{2+}$

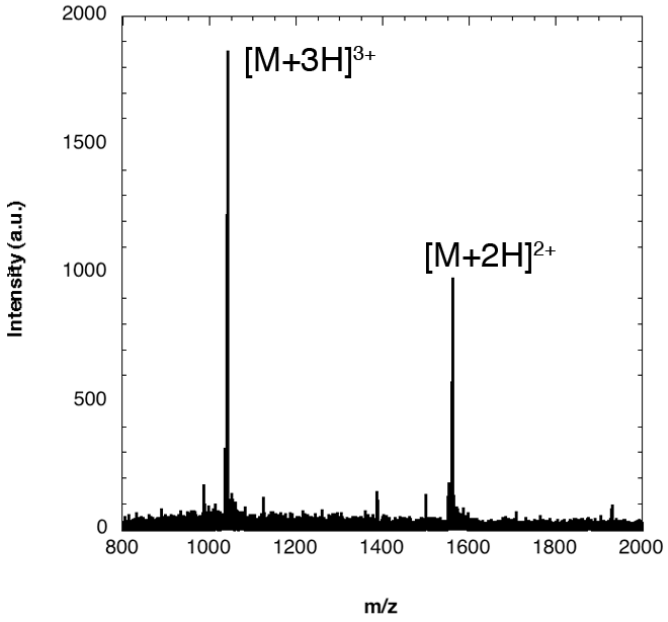


Peptide **C** Expected: 1471.2 $[M+2H]^{2+}$, Observed = 1470.9 $[M+2H]^{2+}$

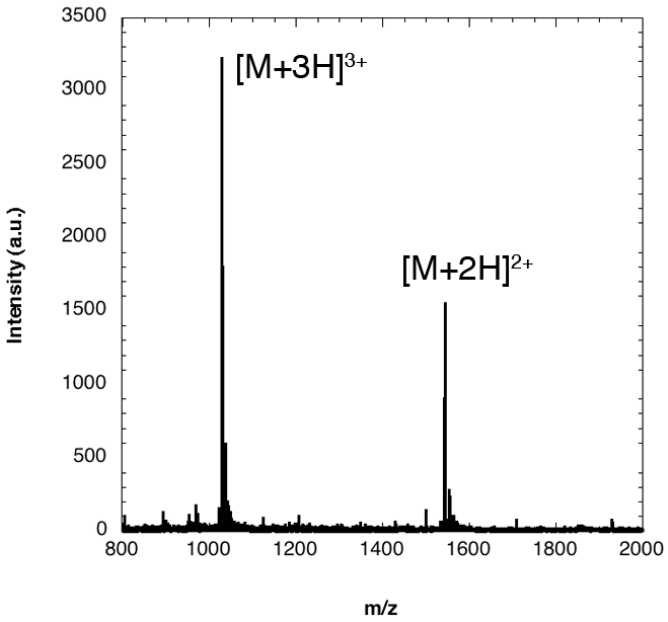
ESI-TOF MS data for the peptides synthesized for Chapter 5



Peptide **α** Expected: 1557.8 $[M+2H]^{2+}$, Observed = 1558.0 $[M+2H]^{2+}$



Peptide β Expected: 1560.7 $[M+2H]^{2+}$, Observed = 1561.0 $[M+2H]^{2+}$



Peptide γ Expected: 1542.7 $[M+2H]^{2+}$, Observed = 1543.0 $[M+2H]^{2+}$

Appendix 4: Melting Temperatures of the Studied Peptides and Peptide Mixtures[*]

Ref.	Abbr.	Sequence[^]	T_m (°C)[≠]
(1)	**K**[⌘]	$(PKG)_{10}$	-
(1)	**D**[⌘]	$(DOG)_{10}$	-
(1)	**O**[⌘]	$(POG)_{10}$	68
(1)	**K/D**[⌘]		41
(1)	**K/D/O**[⌘]		64
p. 32	**KGE**[⌘]	$(POG)_3PK\mathbf{G}EOG(POG)_3$	51
p. 32	**KGD**[⌘]	$(POG)_3PK\mathbf{G}DOG(POG)_3$	48
p. 91	**A**[§]	YGPKGPKGPKGDKGPK**G**PKGPKGPKGDKGPKG	-
p. 91	**B**[§]	$YG(DOG)_4DO\mathbf{G}(DOG)_5$	-
p. 91	**C**[§]	YGPOGPKGPOGPOGPO**G**POGPKGPOGPOGPOG	51
p. 91	**A**/**B**[§]		30
p. 91	**B**/**C**[§]		46
p. 91	**A**/**C**[§]		51
p. 91	**A**/**B**/**C**[§]		60
p. 93	**B1**[§]	$YG(EOG)_4EO\mathbf{G}(EOG)_5$	-
p. 93	**A**/**B1**[§]		32
p. 93	**B1**/**C**[§]		42
p. 93	**A**/**B1**/**C**[§]		52
p. 118	**α**[⌘]	PKGPKGDOGPOGDK**G**DKGPKGPOGDKGPOGGY	-
p. 118	**β**[⌘]	POGDOGDKGPOGPO**G**DKGDOGDKGPKGDOGGY	-
p. 118	**γ**[⌘]	PKGPOGPKGDKGPO**G**POGDKGPOGDOGDOGGY	32
p. 118	**α**/**β**[⌘]		34
p. 118	**β**/**γ**[⌘]		43
p. 118	**α**/**γ**[⌘]		43
p. 118	**α**/**β**/**γ**[⌘]		58

[≠] Melting temperature defined as the minimum in the derivative of the unfolding curve with respect to temperature. Samples showing linear transitions are denoted by "-".

[^] Bold and underlined amino acids are uniformly 15N- and 13C-labeled; bold amino acids are uniformly 15N-labeled.

[⌘] Peptides are acetylated at the N-terminus and amides at the C-terminus

[§] Peptides are free amines at the N-Terminus and amides at the C-terminus.

Reference

(1) Gauba, V.; Hartgerink, J. D. *J. Am. Chem. Soc.* **2007**, *129*, 15304-15041.

[*]All samples tabulated under annealed conditions and equimolar ratios for mixtures, for details refer to the experimental section of each chapter.