Self-assembling nanosystems: Collagen, an escape strategy from plaque formation

Villő Katalin Pálfi

PhD Thesis

Supervisor: Prof. Dr. András Perczel

Program: Synthetic Chemistry, Material Sciences, Biomolecular Chemistry
Head of the Program: Prof. Dr. István Tamás Horváth

Doctorate School of Chemistry
Head of the Doctorate School: Prof. Dr. György Inzelt

Eötvös Loránd University
Institute of Chemistry

2008.

Acknowledgements

There are a number of people that I have to thank a lot, namely:

My family: my parents and my sister, who have helped me whenever they could.

My supervisor: Prof. András Perczel, who was supportive and encouraged my work every time. Peoples of the Laboratory of Structural Chemistry and Biology were also very supportive and provided me with useful advices.

Special thank to Prof. Imre Géza Csizmadia for helping and correcting my works and papers.

Furthermore, I have to acknowledge people and institutes for providing me with computational possibilities: Zoltán Fekete, Ferenc Bartha (HPC Szeged), Tamás Fekete, Levente Tatár (KFKI), Svend Knak Jensen (Aarhus University), Imre Jákli, and Árpád Kiss (ELTE).

Gábor Burján has also helped me to tame the computers and use them more efficiently.

Last but absolutely not least, I have to thank Zoltán Szabadka for all the support and encouragement.

Table of contents

Abbreviations and notations

O, Hyp, hydroxyproline: 4-R-hydroxyl-S-proline

Flp, fluoroproline: 4-R-flour-S-proline

a: D-alanine

Sa, Sar: sarcosine, N-methyl-glycine

SFigure: Supplementary Figure, can be found in the Appendix

STable: Supplementary Table, can be found in the Appendix

$\hat{H}$: Hamiltonian-operator

Ψ: wave-function

χ: molecular orbital

ϕ: spatial molecular orbital

DFT: density functional theory

BSSE: basis set superposition error

LCAO-MO: linear combination of atomic orbitals to molecular orbitals

MP: Møller-Plesset perturbation theory

CC: Coupled Cluster theory

CI: Configuration Interaction theory

PES: potential energy surface

PBC: periodic boundary condition (type of calculation)

HF: Hartree-Fock method

RHF: restricted Hartree-Fock method

UHF: unrestricted Hartree-Fock method

LDA: local density approximation

GGA: generalized gradient approximation

1 Introduction

Protein folding is a relatively fast physical process by which a so called unfolded polypeptide chain folds into its characteristic and functional 3D structure. Hydrophobic, hydrophilic, electrostatic and additional type of interactions are made responsible for the formation of this native structure. The failure to fold into the expected 3D-shape makes globular proteins become inactive. The threat is that these misfolded proteins not only fail to execute their desired task, but that they form aggregates, as aggregation is a spontaneous incident of misfolded or partially folded globular proteins, and was made to be responsible for many neuro-degenerative diseases, such as Alzheimer's.

Protein aggregates can be connected not only to illnesses but they can also be a requirement to maintain a healthy organism. The most prominent case for this is the self assembly of proteins such as the case of collagen formation. Up to our knowledge, the aggregation of amyloid and other proteins involved in lethal diseases cannot be controlled and the aggregates themselves cannot be dismantled by the organism. Collagen formation, however, is thoroughly regulated, and the resulting structure – as can be observed for example in the bone – can also be degraded.

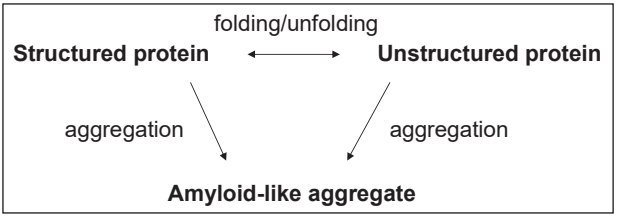


Figure 1. A simplified scheme of an emerging new paradigm of structural biology

In this work collagen-like molecular folds fall in the category of structured proteins (**Figure 1**), while its monomers are regarded as unfolded proteins. Structures that contain multiple strands with extended backbone conformation represent the amyloid-like aggregates.

1.1 Amyloid

The primary structure (amino acid sequence) of a globular protein encodes its three dimensional structure related to its biological function.[1] However, growing evidence supports that an alternative, well organized, but different 3D structure could exist for many proteins.[2] Dozens of ordinary proteins (*e.g.* SH3, β-2 microglobulin, lysozyme, myoglobin)[3,4] tend to aggregate if misfolded under abnormal cellular conditions,[5] producing an architecture similar to the amyloid peptide, which is responsible for the development of Alzheimer's disease.[6] Similar aggregates play a role in the case of the Creutzfeldt-Jacob disease and the Huntington chorea, although the aggregating protein is different, the prion protein (PrP) is responsible for the first, and the poly-glutamine for the second disease.[7] The conversion from the globular form into an amyloid-like aggregate is the transformation of the physiologically "healthy" structure into a non-functional, pathogenic conformer. Thus, understanding the molecular mechanisms of such a transformation and deciphering the underlying thermodynamic cause has a wide range of applications.

These protein aggregates have the same macroscopic structure, so-called amyloid fibrils, and are supposed to have the same or at least similar molecular structure. Several research groups performed pioneering contributions to the structure elucidation of such fibrils.[2,8,9,10,11,12,13] However, the exact atomic build-ups are still uncertain. It is thought that these aggregates are rich in β-sheet structures, where in one peptide chain there are two elongated parts connected by a turn region, and peptide chains are placed parallelly in an endless crystal. When the appropriate conditions are set, practically any protein investigated could be trapped in such a toxic β-layer form.[2,14,15,16,17] Thus, the folding propensities encoded by the protein side-chains has apparently little or no impact on the formation of the aggregate.[3,18,19] So much the more, as a recently proposed backbone-based theory of protein folding emphasizes that the energetics of backbone hydrogen bonds dominate the overall folding process[20] even for "normally" folded proteins. Thus, strong backbone-backbone interactions (primarily interchain hydrogen bonds) that are well pronounced in β-strands are expected to be the driving force of amyloid-like aggregation.

To prevent proteins from aggregating and preserve their native structure is believed to be a major driving force in biological evolution.[17] There are a number of alternative strategies besides molecular chaperones, ubiquitination enzymes and proteasomes to protect proteins from any unwanted type of aggregation.[21,22,23] These are: *i,* insertion of "structural

gatekeepers"[22], meaning that charged side chains prevent aggregation by interrupting contiguous stretches of hydrophobic residues in the primary sequence. *ii,* the use of domains of low (30-40%) sequence homology in multidomain proteins[21] and *iii,* the application of a "negative design"[23] for the protection and cover of all otherwise free edge β-strands, that are highly prone to dimer formation and further aggregation.

One may ask the question of why peptide chains prefer this β-pleated sheet conformation and whether it is the only structure available for aggregate formation.

1.2 Collagen

Collagen is an essential extracellular protein. Its importance is best exemplified by the fact that about one quarter of the mass of all the proteins in a human body is collagen.[24] It is a major structural protein, forming molecular "cables" that strengthen the tendons and sheets that support the skin and internal organs. These cables and sheets are built up of collagen molecules strongly attached to each other.[25] The first level of this attachment in collagen is tropocollagen, which consists of three protein chains that are self-assembled into a triple-helical structure.[25] Furthermore, seven tropocollagen triple helices form microfibrils in a hexagonal arrangement.[25] (see **Figure 2**) It is interesting to note that the individual collagen strands form a left-handed helix, and three of them build a right-handed triple helix. The hexagonal filament built from triple helices is also right-handed.

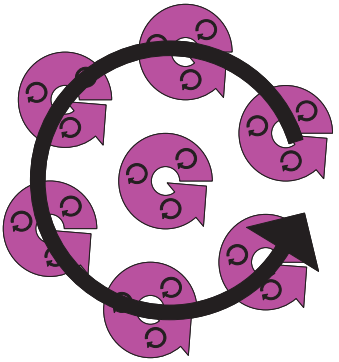


Figure 2. Right-handed collagen filament, formed from right-handed tropocollagen triple helices (pink), that, in turn are built up of left-handed helices of protein chains (small black).

There are about 20 types of collagen, each type consists of different protein chains. The most frequent is type I collagen, where the tropocollagen triple helix is built up of two identical and one slightly different collagen protein.[26]

There are three well-known diseases connected to collagen. The inheritable brittle-bone disease (osteogenesis imperfecta)[27,28], when the triple helices are not properly folded and the bones (where collagen also plays an important role) can be easily broken. Osteoporosis imperfecta[28] is mostly common for women. Here the bones also become weaker, but because the disintegration of collagen is faster than its build-up, not because of inherent weakness. The most well-known disease is scurvy, the vitamin C deficiency[29]. In this case the collagen also weakens, that is caused by the decreasing amount of hydroxyproline amino acid. Hydroxyproline can only be produced in the presence of vitamin C.

Collagen fibril formation starts with procollagen synthesis. This is a protein, that has globular ends on both sides, and the middle (approximately 900) residues are sequentially predetermined to tropocollagen formation (every third amino acid is glycine). The C-terminal end is then attached to 2 other C-terminal parts, and then as the three middle parts get closer to each other, triple helix formation begins. When the whole chain is wound procollagen peptidase cleaves down the globular ends and only a tropocollagen triple helix remains.[24]

1.2.1 Atomic structure and H-bonding characteristics of collagen triple helix and β-pleated sheets

The triple helix of tropocollagen is built up of almost identical conformational elements (homoconformers) often referred to as the polyproline II or shortly PPII conformation[30,31,32], of elongated backbone structures with the following parameters: $\phi \approx -70°$, $\psi \approx +160°$ (ε_L)[33], similar those of β-pleated sheet type aggregates, β_L, ($\phi \approx -150°$, $\psi \approx +150°$)[33] (**Figure 3**).

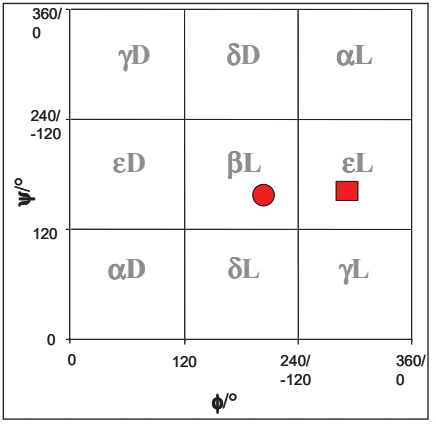


Figure 3. The nine typical conformational backbone building units in peptides and proteins.[34] The square represents building units of collagen, also known as PPII (or ε_L for short), while the circle represents β_L type conformers, typical of β-pleated sheet structures. These are the two most important peptide folds for the present study.

The primary structure of collagen is characterized by the repetitive motif Xxx-Yyy-Gly, where the Xxx and Yyy positions are typically occupied by proline (Pro, P) and hydroxyproline (Hyp or O in short) residues, respectively.[24] The latter repetitive motif, Xxx-Yyy-Gly, is called an amino acid triplet, or triplet for short.

Although β-sheet formation is favored by some and disfavored by other natural amino acid residues[35] no such strict primary sequence motif is required. As β-layers,[36] collagen triple-helices are also self-stabilized by specific H-bonds. In collagen triple helix, the amide hydrogen atom (N-H) of each Gly is attached to the carbonyl oxygen (C=O) of residue Xxx of the adjacent strand *via* strong hydrogen bonding.[24,30,37] In addition, there are two weak (non-

classical or improper) H-bond systems along the collagen microfiber: the H_α of residue Yyy is connected to the carbonyl oxygen atom of residue Xxx of the adjacent strand, and the two H_α(s) of glycines are hooked to the C=O of another glycine in one of the adjacent strands.[37,38,39,40,41,42] (**Figure 4a**) The β-sheet secondary structure incorporates also a fine tuned but rather different interchain H-bond network as described elsewhere.[36](**Figure 4b**)

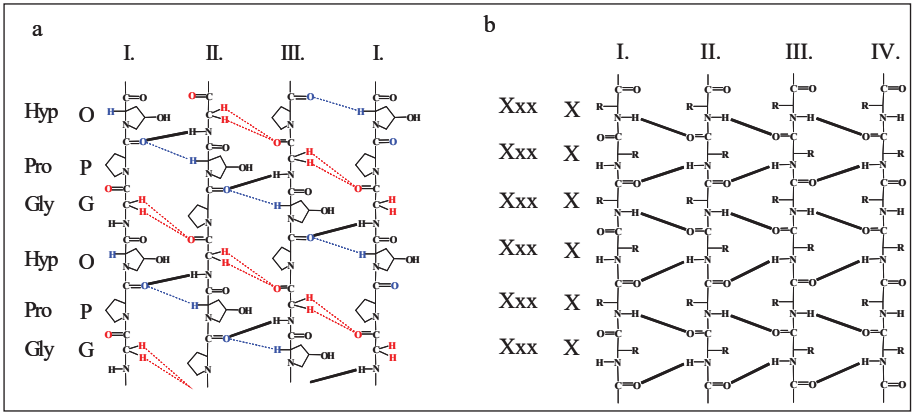


Figure 4. a, Schematic diagram of the H-bond network typical for a collagen triple helix: both the classical (C=O...HN; strong black) and weak (C=O...HC; dashed blue or red) hydrogen bonds along the polypeptide chains are depicted. (The first strand is duplicated to show the complete H-bond network.) **b,** Schematic diagram of the H-bond network of parallel β-pleated sheet structures.

There are several papers in the literature dealing with the stability of collagen.[43,44,45,46,47,48] It is well known that proline and hydroxyproline residues are vital for the stability of the triple helix. It is also known that electron withdrawing substituents at 4(R) position on the proline ring stabilize the helix at the Yyy position, as they even more decrease the conformational freedom of a proline ring and the dihedral angles of the backbone.[44,45,48] However, in these studies the conformational preference of amino acid dipeptides was compared to each other rather than directly the stability of the resulting secondary structures.

An advance from these results is the work of Parthasarathi *et al.*[49], where they compare the collagen forming Gibbs energy of various triplets. However, as they mention, a work using a full collagen triple helix is needed.

Dannenberg and colleagues[50] have recently reported the energetics of a variety of triple helix collagen models of six amino acids in length. One of the important statements in their comprehensive study is that the relative energy (stability) of a triple helix can be calculated with respect to more than one reference state by using either amino acid dipeptides or single stranded polypeptides, leading to different stability results. The destabilizing effect of a Gly → L-Ala mutation (that causes the brittle-bone disease) has been observed both in native collagen and in model systems [39,50,51,52] showing that it induces a "bulge" within the collagen triple helix by distorting the characteristic H-bridges.[53] Another fascinating suggestion arising from Dannenberg's work is the use of a D-amino acid residue instead of glycine.[50] They found that D-Ala or D-Ser at the right sequential position stabilizes a collagen triple helix structure.

However, neither of these studies examined a triple helix built up only from amino (non-imino) acids, or the effect of sarcosine (N-methyl-glycine) on the formation energetics.

In collagen, like in amyloid, protein strands need to be nearly parallel and also need to maintain strong contact with each other, by forming strong interchain hydrogen bonds. As mentioned above, several strategies exist in nature to avoid protein aggregation and collagen formation is usually not regarded as one of them. However, in this thesis we will introduce the following concept: to save collagen proteins from forming amyloid-like plaques nature has designed it to be an "anti-amyloid", that is to maintain an inherently different conformation (poly-glycine II), and, in the meantime, depositing the protein strands parallelly and bound by hydrogen bonds.

Sarcosine (or N-methyl glycine) is an amino acid that is not among the 20 that are encoded in the DNA. However, it is quite common in the human body, as it is a metabolite of choline to glycine, to which it is rapidly degraded. In principle, sarcosine could replace either proline or hydroxyproline in tropocollagen as it is an imino- rather than an amino acid residue (see below).

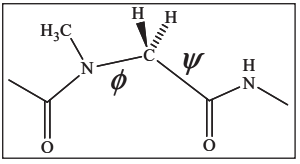


Figure 5. Molecular structure of N-acetyl-sarcosine-methylamide.

1.2.2 The hydration structure of collagen

At least six [Pro-Hyp-Gly] or [Pro-Pro-Gly] triplet units, $[POG]_6$ or $[PPG]_6$ in short, are required in solution to form a stable triple-helical secondary structural element.[54,55,56] From the experimental side it is known that hydroxyproline at the Yyy position makes collagen more stable than if it were only a proline.[48,57] However, the precise explanation of how $^{\gamma}$OH groups stabilize the triple-helical collagen structure is still not yet fully understood. Previous results[38,39,53,58] claim that the part of the overall stability of collagen triple helix is due to its special hydration shell: water molecules become specifically attached to the surface of the protein. Other type of results[48,59,60] emphasize the electron withdrawing effect of the OH group on the pirrolidine ring, thus reducing the conformational freedom of the backbone, so "freezing" the polypeptide in the required secondary structure. A confirmation for this theory is that substituting the hydroxyl group with a fluorine atom at the Yyy position greatly increases the melting point of the triple helix.[42] A third theory[41] (apparently combining the previous two) suggests, that the required backbone conformation of the polypeptide chain is maintained *via* a water molecule that is bound to a carbonyl oxygen atom and to the OH group of hydroxyproline. As (to the best of our knowledge) there is no atomic structure data available on fluoro-proline containing collagen, we cannot quantify the extent of how such a substitution modifies either the structure or the hydration shell of it.

The water content bound to tropocollagen triple helix is said to be also connected to the aging of collagen.[61] Lysines form covalent bonds with each other between tropocollagens. The process starts with an oxidative desamination, when the amino group of one lysine is transformed into a carbonyl group that can subsequently form Schiff-basis with another lysine from another collagen triple helix.[61, 62] The number of the above inter-helix covalent links grows over time, while the amount of structurally bound water shells slowly decreases and the rigidity of collagen increases.[61] This seems to confirm the importance of maintaining the optimal hydration level of skin (and all of our body).

The atomic structure of the water shell around collagen was first described meticulously by Bella et al.[53] and was observed in several other X-ray structures as well.[42,58,63,64,65] They describe that some water molecules are not just absorbed on the surface of tropocollagen but also maintain a regular pattern. Hydrogen-bond mediated water chains connect different parts of tropocollagens.[53] We refer to these water-chains as "water bridges" when the first and the last water molecules of the chain are connected to the same tropocollagen unit. However, in crystal structures triple helices can be as close to each other as 5 Å (in the 1V7H X-ray

structure[63]), which is in the same scale as the distance between the two ends of a water bridge on the same helix (in the 1V7H X-ray structure[63]). This distance can be easily covered by a water chain of 2 or 3 molecules. Therefore, regarding the number of water molecules in one bridge there is no difference between inter- or intra-tropocollagen water chains. Deciphering characteristics of inter-tropocollagen water chains is not the focus here, as in this thesis the internal stability of tropocollagen and its first hydration layer is examined.

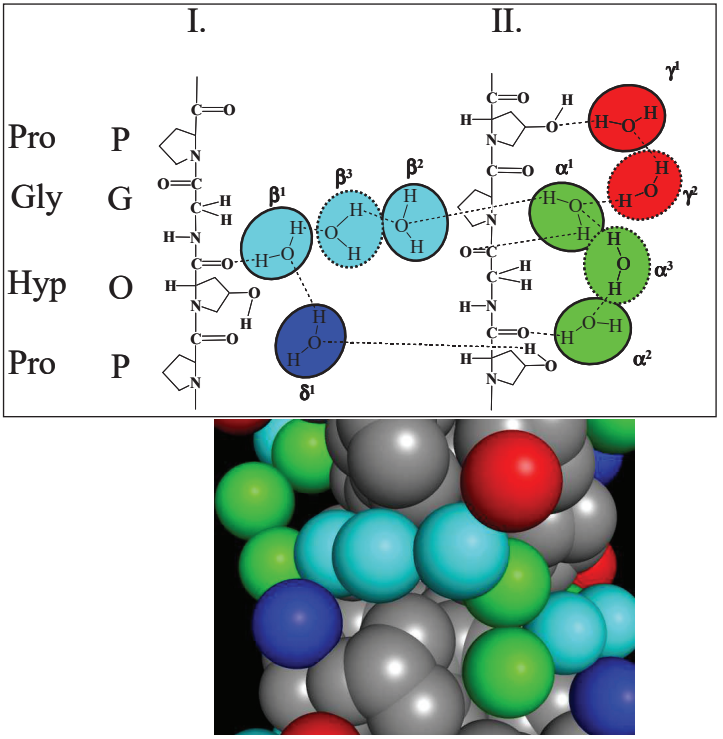


Figure 6. upper, Schematic water connectivity (first layer hydration network) of a triple helical collagen. (α; green (3 H_2O), β; light blue (3 H_2O), γ; red (2 H_2O) and δ; marine (2 H_2O, from which there is only one at the images, as the other one is considered as part of the β-bridge, see text)). The names of the water bridges are taken from Bella et al.[53], while labeling of the individual water molecules were completed accordingly. Those H_2Os circled with a solid line can always be found in the X-ray structure, while those encircled by a dashed line can be missing. **lower,** The water bridges as seen in the 1V7H PDB structure.[63] The same coloring pattern is used as for the scheme above, while all atoms of the polypeptide chains are grey. (No hydrogen atoms are shown.)

Bella et al.[53] described four different types of water bridges, called α, β, γ and δ according to the type of atoms they are attached to on the surface of the collagen triple-helix.[53] **(Figure 6)** In all of the so far reported X-ray structures these four different types of water bridges (**Figure 6)** can be identified. However, as reported in 1CGD by Bella et al.[53] and can be

observed in 1V7H[63], the total number of H_2Os within the same water-bridge may vary. It changes for the α, β and γ-type bridges; namely 2 or 3 for the α-, 3 or 4 for the β-, and 2 or 3 water molecules are there in the γ-bridges. (see **Figure 6** and **Table 1**) As shown in **Figure 6** the individual water molecules were named according to the parent bridge and the type of connection involved. There are two water molecules, α^1 and β^1, that are part of two or three water bridges, however named only after one. The existence of γ and δ water bridges greatly depends on the amino acid type at the Yyy position: namely Yyy should either be hydroxyproline or threonine. All in all, the first hydration shell of the POG-type collagen consists of minimum 6 and maximum 9 water molecules per triplet unit, of which 5 is attached directly to a collagen atom by a hydrogen-bond.

Table 1. The maximum number of water molecules forming α, β, γ and δ water bridge, the first and most specific hydration level of collagen as assigned in the unit cell of 1V7H[63]

Water-bridge binding place in the unit-cell [b]	**Type of bridge**[a]			
	α	**β**	**γ**	**δ**
1st	2[c,d], N.A.[e]	4	N.A.[e]	2
2nd	1, N.A.[e]	4	2	2
3rd	2	3	2	2
4th	3	3	2	2
5th	2	4	3	2
6th	3	3, N.A.[e]	2, N.A.[e]	1, N.A.[e]
7th	2	4	2	2
Summary	2-3	3-4	2-3	2

a, the water bridges were first described by Bella et al.[39]
b, as the repeating unit of 1V7H[63] is seven residue long, there are seven different binding places for water bridges
c, The γ-bridge has one water molecule shared with that of the α- and β-bridge. The δ-bridge has one water molecule shared with that of the β-bridge (**Figure 6**). Thus, in these bridges the total number of water molecules is lesser by 3.
d, The water molecule is considered to be attached to another one or to collagen if the distance between the appropriate oxygen atoms in the X-ray structure is less than 3.25 Å.
e, Not Available: water molecules are not found, or no more is found, although the geometrical arrangement suggests an incomplete water bridge

There is an additional type of structural water associated with collagen as suggested by Ramachandran and Chandrasekharan.[66] It can be assigned only if amino acid Xxx has a

backbone amide hydrogen atom. This water molecule connects to two polypeptide chains of the same tropocollagen unit, thus can be regarded as a water-bridge composed of a single molecule only (**Figure 7** and **Figure 12**). As found in 1BKV structure[58], in which at Xxx non-imino acids can also be found, out of the 9 sequentially allowed positions a total of 8 inserted H_2O(s) are observable.[58,67] For these 8 binding places, a single and somewhat isolated oxygen atom is positioned at an optimum H-bonding distance, with respect to both polypeptide chains. Further on, this interchain inserted single water molecule is called as ζ-type water, first introduced by Bella et al.[53].

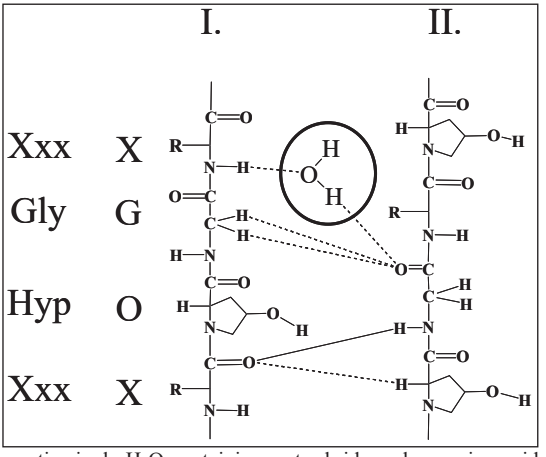


Figure 7. A schematic single H_2O containing water bridge when amino acid Xxx is neither proline nor any other imino acid residue, in which no amide hydrogen atom would be available. These types of collagen subunits are typical of collagen cleaving sites. Here the one and only water molecule is marked by a circle and called as ζ-type water. The direct inter-chain H-bonds (**Figure 4**) are marked with black.

Melacini et al.[68] have measured the residence time of water molecules around the triple-helix. They found that even the directly bound water molecules move in and out on a nano- to sub-nano-second timescale, which means that they possess liquid-like properties.

Fullerton et al.[69,70,71] have recently measured the hydration shells of collagen by NMR spectroscopy. For the most attached water molecule they concluded that in average there is one per 3 amino acid residue. At the next layer approximately 3, while at the third regime some 20 water molecules are incorporated per triplets. Another (calorimetric) measurement by Boryskina[72] showed that the total number of the mostly attached water molecules is around

3 per triplet, but their enthalpy of hydration was found to be 0.71 kcal.mol^{-1} per water molecule. However, these experiments fail to specify the type of atoms of collagen to which these water molecules are attached to.

It is interesting to note that in the 1V7H X-ray structure[63] water molecules are localized and can clearly be identified, while the above measurements indicate liquid-like characteristics for the bound waters. To explain both of these measurements a "hopping" theory[67,68] was suggested: water molecules hop in and out of these specific binding sites, with a short residence time however preserving their well defined atomic position.

Also an interesting result from Boryskina[72] is that the native triple-helical collagen structure is maintained only with a minimal number of 4-5 water molecules per triplet.

1.3 Theoretical introduction

All of the results are obtained using theoretical calculations, therefore a very short introduction is given here to the techniques used. If the reader would like to learn more details about them, the book 73 and article 74 are recommend, and further references therein.

1.3.1 Introduction to the Hartree-Fock and DFT methods

To calculate the energy of a non-relativistic quantum chemical system the Schrödinger-equation is used, whose time-independent form is the well-known:

$$\hat{H}\Psi = E\Psi \,, \qquad \text{eq. 1.}$$

where $\hat{H}$ is the Hamiltonian-operator and Ψ is the wave function of the system. There are infinite number of eigenfunctions, the solution corresponding to the i^{th} lowest energy (E_i) is denoted by Ψ_i.

The form of the Hamiltonian-operator without electric or magnetic field is:

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A=1}^{M}\frac{1}{M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \qquad \text{eq. 2.}$$

where A and B stands for the nuclei from 1 to M, i and j for the electrons from 1 to N. M_A is the mass of nucleus A in atomic units, ∇^2 is the Laplace-operator. Note that *R* stands for the coordinates of the nuclei, and *r* for the coordinates of the electrons. The distances are calculated the following way: $r_{iA}:=|r_i–R_A|$; $r_{ij}:=|r_i-r_j|$; $R_{AB}:=|R_A-R_B|$. In equation 2 the first and the second terms describe the kinetic energy of the electrons and the nuclei, the third the electron-nucleus attraction, the fourth and the fifth the electron-electron and nuclei-nuclei repulsion, respectively. This equation uses *atomic units* that greatly simplify the equations that describe quantum systems. Physical quantities are expressed as multiples (or combination) of fundamental constants, which can be therefore dropped from the equations. There are atomic units for mass, charge, length and energy that are further detailed in **Table 2**.

Table 2. The atomic units of the physical quantities and their units in SI.

Quantity	Atomic unit	Value in SI units	Symbol (name)
Mass	Rest mass of electron	$9{,}109 \cdot 10^{-31}$ kg	m_e
Charge	Elementary charge	$1{,}602 \cdot 10^{-19}$ C	e
Action	Planck's constant/2π	$1{,}055 \cdot 10^{-34}$ Js	$\hbar$
Length	Bohr-radius: $4\pi\varepsilon_0\hbar/m_e e^2$	$5{,}292 \cdot 10^{-11}$ m	a_0 (bohr)
Energy	Hartree-energy: $\hbar^2/m_e a_0^2$	$4{,}360 \cdot 10^{-18}$ J	E_h (hartree)

To use the Schrödinger-equation for eventual calculations lots of simplifications have to be made. The first is the *Born-Oppenheimer* (or clamped nuclei) approximation. This states, that as nuclei are much heavier than electrons (at least 1836 times, for the smallest hydrogen atom), nuclei seem to be fixed from the point of view of the electrons. Therefore the Schrödinger-equation can first be rewritten as if the kinetic energy of the nuclei were zero, e.g. they did not move, and their positions (R) are included only as parameters, not as variables.

$$\hat{H}_{BO}(R)\Psi(x;R) = E_{BO}(R)\Psi(x;R) \qquad \text{eq. 3.}$$

This $E_{BO}(R)$ is also called Potential Energy Surface (PES), and is widely used to describe reactions as well as conformational changes of molecules.

$$\hat{H}_{BO}(R) = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} = \\ = \hat{T} + \hat{V}_{eN}(R) + \hat{V}_{ee} + \hat{V}_{NN}(R) = \hat{H}_e(R) + \hat{V}_{NN}(R) \qquad \text{eq. 4.}$$

The $\hat{V}_{NN}$ part is only a number, as it depends solely on the position of the nuclei that are parameters. The rest can be treated as a Hamiltonian operator of the electrons, and further on we will consider the solving of this eigenvalue-equation.

$$\hat{H}_e(R)\Psi(x;R) = E_e(R)\Psi(x;R) \qquad \text{eq. 5.}$$

From this the potential energy surface can be obtained as

$$E_{BO}(R) = E_e(R) + \hat{V}_{NN}(R) \qquad \text{eq. 6.}$$

The wave-function for a many-electron system can be defined as

$$\Psi(x;R) = \Psi(x_1, x_2, x_3, ..., x_N; R) \qquad \text{eq. 7.}$$

where N is the number of electrons and x_i is the coordinate of the ith electron. It includes the three spatial coordinates of the electrons, and also the spin coordinate. Further on the coordinates of the nuclei (R) will not be marked.

The wave-function itself cannot be measured, only the square of its absolute value ($\left|\Psi(x_1, x_2, x_3, ..., x_N)\right|^2 dx_1 dx_2 dx_3 ... dx_N$), which means the probability of finding one of the electrons in the $[x_1, x_1 + dx_1] \times [x_2, x_2 + dx_2] \times ... \times [x_N, x_N + dx_N]$ part of the space. As the electrons are indistinguishable, the absolute value cannot change if we switch two of them:

$$\left|\Psi(x_1, x_2, ..., x_i, x_j, ..., x_N)\right|^2 = \left|\Psi(x_1, x_2, ..., x_j, x_i, ..., x_N)\right|^2 . \qquad \text{eq. 8.}$$

Thus, the two wave functions can only differ by a unimodular complex multiplier $e^{i\phi}$. It can be shown, that upon the switching of two particles the wave function of the system must remain identical, or, at most, it can change its sign. For *bosons* (particles with integer spins) the wave function remains identical (symmetric wave function), while for *fermions* (particles that have half spins), the sign changes (antisymmetric wave function). Electrons are fermions, therefore the electronic wave function changes its sign upon the change of two electrons. From this follows Pauli's exclusion principle, that is "no two electrons can have the same states" (or else the wave function would have to be zero).

As the probability of finding one of the electrons in the full space is one we get the following equation:

$$\int ... \int \left|\Psi(x_1, x_2, x_3, ..., x_N)\right|^2 dx_1 dx_2 dx_3 ... dx_N = 1 \qquad \text{eq. 9.}$$

Putting it into other words, the wave function is normal.

The wave function of a system "encodes" not only the electron density, but also the expectation value of the result of every measurement made on the system. If the operator corresponding to the measurement is $\hat{O}$, then the expectation value of that measurement can be calculated as follows:

$$\left\langle \Psi \middle| \hat{O} \middle| \Psi \right\rangle = \int ... \int \Psi^*(x_1, x_2, x_3, ..., x_N) \hat{O} \Psi(x_1, x_2, x_3, ..., x_N) dx_1 dx_2 dx_3 ... dx_N \qquad \text{eq. 10.}$$

The operator corresponding to the energy of the system is the Hamiltonian-operator, so

$$E_0 = \left\langle \Psi_0 \middle| \hat{H} \middle| \Psi_0 \right\rangle \qquad \text{eq. 11.}$$

The *variational principle* states that the expectation value of the Hamiltonian-operator with the exact ground electronic state wave function (Ψ_0) is lower than the expectation value with *any* trial wave function (Ψ_{trial}).

$$\langle \Psi_{trial} | \hat{H} | \Psi_{trial} \rangle = E_{trial} \geq E_0 = \langle \Psi_0 | \hat{H} | \Psi_0 \rangle \qquad \text{eq. 12.}$$

The closer the trial function is to the exact ground state wave function, the lower the energy will be. Therefore, if we know the location of the nuclei and the number of electrons, we can construct the Hamiltonian-operator, and from that the $\Im[\Psi] = \langle \Psi | \hat{H} | \Psi \rangle$ functional. Optimizing this functional with a $\langle \Psi | \Psi \rangle = 1$ constraint we may reach, in a limiting sense Ψ_0 and E_0.

The *Hartree-Fock* methods search for the minimal energy by using only a subset of the possible wave functions. It builds the wave function of the whole many-electron system from one-electron wave functions in a determinant form (to maintain the antisymmetric nature of the wave function). This determinant is usually called the Slater-determinant.

$$\Psi_{trial} \approx \Psi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \cdots & \chi_N(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \cdots & \chi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(x_N) & \chi_2(x_N) & \cdots & \chi_N(x_N) \end{vmatrix} \qquad \text{eq. 13.}$$

These one electron functions (χ_i) are spin orbitals, as they contain a spin coordinate as well.

$$\chi_i = \chi_i(r,s), \qquad \text{eq. 14.}$$

where r contains the spatial coordinates, and $s \in \left\{-\frac{1}{2}, \frac{1}{2}\right\}$ is the spin state. These spin-orbitals can be decomposed into two spatial orbitals as follows:

$$\chi_i(x) = \phi_i^{\alpha}(r)\alpha(s) + \phi_i^{\beta}(r)\beta(s), \qquad \text{eq. 15.}$$

where

$$\phi_i^{\alpha}(r) = \chi_i\left(r, \frac{1}{2}\right), \ \phi_i^{\beta}(r) = \chi_i\left(r, -\frac{1}{2}\right) \qquad \text{eq. 16., 17.}$$

$$\alpha\left(\frac{1}{2}\right) = 1, \ \alpha\left(-\frac{1}{2}\right) = 0, \ \beta\left(\frac{1}{2}\right) = 0, \ \beta\left(-\frac{1}{2}\right) = 1 \qquad \text{eq. 18., 19., 20., 21.}$$

In order to guarantee the Slater-determinant to be normal, we have to constrain the spin-orbitals to be normal, and pairwise orthogonal:

$$\langle \chi_i | \chi_j \rangle = \delta_{ij} \qquad \text{eq. 22.}$$

The generalized *Hartree-Fock* method searches for the global minimum of the $\Im[\Psi]=\langle\Psi|\hat{H}|\Psi\rangle$ functional, with the restriction that Ψ is a Slater-determinant (eq. 13.), and that eq. 22 holds. Further restrictions can be made to produce the Unrestricted Hartree-Fock (UHF) method: the χ_i spin orbitals are real valued, and have only α or only β spin component. The resulting Slater-determinant will be automatically the eigenfunction of the spin-operator ($\hat{S}_z$), but not necessarily of the total spin-operator ($\hat{S}^2$). To get to the Restricted Hartree-Fock method (RHF) one further restriction is made: half of the electrons have only α and half only β spin state, and the electrons can be paired in such a way that each pair shares the same spatial orbital.

$$\begin{array}{ll}\chi_1(x)=\phi_1(r)\alpha(s) & \chi_{N/2+1}(x)=\phi_1(r)\beta(s)\\ \vdots & \vdots\\ \chi_{N/2}(x)=\phi_{N/2}(r)\alpha(s) & \chi_N(x)=\phi_{N/2}(r)\beta(s)\end{array} \qquad \text{eq. 23.}$$

The resulting wave function will be the eigenfunction of the total spin-operator as well.

The expectation value of the Hamiltonian-operator with a Slater-determinant in the RHF framework can be expressed as

$$E_{HF}=\langle\Psi_{SD}|\hat{H}|\Psi_{SD}\rangle=2\sum_{i=1}^{N/2}(i|\hat{h}|i)+\sum_{i=1}^{N/2}\sum_{j=1}^{N/2}\left[2(ii|jj)-(ij|ji)\right], \qquad \text{eq. 24.}$$

where the one electron integrals describe the kinetic energy of the electrons and the electron-nuclei attraction:

$$(i|\hat{h}|i)=\int\phi_i(r_1)\left\{-\frac{1}{2}\nabla^2-\sum_{A=1}^{M}\frac{Z_A}{r_{1A}}\right\}\phi_i(r_1)dr_1 \qquad \text{eq. 25.}$$

and the two-electron integrals describe the interaction of two electrons.

$$J_{ij}=(ii|jj)=\iint\phi_i^2(r_1)\frac{1}{r_{12}}\phi_j^2(r_2)dr_1dr_2 \qquad \text{eq. 26.}$$

$$K_{ij}=(ij|ji)=\iint\phi_i(r_1)\phi_j(r_1)\frac{1}{r_{12}}\phi_j(r_2)\phi_i(r_2)dr_1dr_2. \qquad \text{eq. 27.}$$

The first is the Coulomb-integral, the second is the exchange integral.

So the task is to vary the Slater-determinant to get the lowest possible energy, and in the meantime maintain the orthogonality of the molecular orbitals. As this is a constrained

optimization problem, the method of the *Lagrangian multipliers* is used. The Lagrangian-functional of the problem is:

$$\tilde{\mathfrak{I}}\left[\{\phi_i\}_{i=1}^{N/2}\right] = \langle \Psi_{SD} | \hat{H} | \Psi_{SD} \rangle - \sum_{1 \le i,j \le N/2} \lambda_{ij} \left(\langle \phi_i | \phi_j \rangle - \delta_{ij} \right), \qquad \text{eq. 28.}$$

where λ_{ij} are the Lagrangian multipliers. From now on we will denote the dependence of a functional or operator on a system of molecular orbitals by $[\phi]$. We can get the extremum points of the original functional by taking the functional derivatives of the Lagrangian functional with respect to the molecular orbitals, and requiring the results to be zero.

$$\frac{\delta \tilde{\mathfrak{I}}[\phi]}{\delta \phi_k} = \hat{F}^{RHF}[\phi]\phi_k - \sum_{j=1}^{N/2} \lambda_{kj}\phi_j = 0, \qquad \text{eq. 29.}$$

where

$$\hat{F}^{RHF}[\phi] = \hat{h} + \sum_{j=1}^{N/2} \left(2\hat{J}_j[\phi] - \hat{K}_j[\phi] \right) \qquad \text{eq. 30.}$$

is called the Fock-operator. Applying a diagonalization on the Lagrangian multipliers we get the Hartree-Fock equations:

$$\hat{F}^{RHF}[\phi]\phi_j = \varepsilon_j \phi_j, \qquad j=1,\ldots, N/2 \qquad \text{eq. 31.}$$

Solving this integro-differential equation we can obtain (among all the extremum points) the global minimum of the $\mathfrak{I}[\Psi]$ functional with the constraint that Ψ is a Slater-determinant built from the pairwise orthogonal molecular orbitals. (This minimum is the Hartree-Fock approximation of the ground state energy.) To transform it into a well-know matrix eigenvalue problem the ϕ functions are expressed as a linear combination of basis functions (usually called atomic orbitals), (LCAO-MO). This leads to the following equation:

$$\underline{\underline{\hat{F}}}[\varphi]\underline{\varphi_j} = \varepsilon_j \underline{\varphi_j}, \qquad j=1,\ldots, N/2 \qquad \text{eq. 32.}$$

Here the Fock-matrix ($\underline{\underline{\hat{F}}}[\varphi]$) still depends on the system of coefficients $\{\varphi_1,\ldots,\varphi_{N/2}\}$. To overcome this difficulty we start with an initial set of coefficients (ansatz), build the Fock-matrix, solve the eigenvalue-problem (eq. 32.) to obtain a new set of coefficients, then update the Fock-matrix. This is continued until the eigenvalues (ε_i) and the coefficients (φ_i) are sufficiently close to each other in the subsequent steps. This procedure is called the Self-Consistent-Field (SCF) method.

As the Fock-operator is not the exact Hamiltonian, and the Slater-determinant does not represent the best type of trial function, the resulting energy of the molecule is not the lowest possible, i.e. not the exact. The difference between the exact and the Hartree-Fock energy is called the correlation energy.

There are two basic approaches that try to approximate this correlation energy. One of them uses the exact non-relativistic Hamiltonian, and approximates the exact wavefunction by a superposition of determinants. These methods include the Møller-Plesset perturbation theory (MP) series, the Coupled Cluster (CC) series and the Configuration Interaction (CI) series, called altogether post-Hartree-Fock methods. Due to the *ab initio* nature of these methods the results can always be systematically improved to solve the original Schrödinger-equation with the desired accuracy. However, these represent highly demanding computations, and for example the Coupled Cluster calculations are unfeasible for systems larger than 10 atoms.

The other approach that tries to include the correlation energy into calculations is the density functional theory (DFT). With DFT techniques even large peptide molecules of around 20 amino acid residues can be easily calculated, therefore we used this technique to get the most exact results.

Density functional theory is based on the theory that the electron density contains all the relevant information of the system. The electron density is the integral of the wave-function in all the coordinates of the electrons in the system, except for the spatial coordinates of one.

$$\rho(r) = N\int ...\int \left|\Psi(x_1, x_2, x_3, ..., x_N)\right|^2 ds_1 dx_2 dx_3 ... dx_N \qquad \text{eq. 33.}$$

The electron density is a non-negative function that converges to zero as the coordinates (r) converge to infinity. Its integrate with r gives the electron number, N.

$$\int \rho(r) dr = N \qquad \text{eq. 34.}$$

The electron density is an observable (e.g. measurable by X-ray crystallography). Its important property is that it has its maximum in the places of nuclei. There it is continuous but not differentiable, which property is called cusp (Kato cusp theory). That is why the electron density holds all the information: it contains the places and charges of nuclei and the number of electrons. From these the Hamiltonian can be constructed, and every other desired property can be calculated.

The first Hohenberg-Kohn theorem states that the electron density of the ground state uniquely determines the Hamiltonian, and so the ground state wave function.[75] (That is, no two different electron densities can have the same Hamiltonian.) The second Hohenberg-Kohn theorem states that the energy that corresponds to a certain nuclei configuration (with a

given number of electrons) is the lowest if we apply the exact density function, any trial function gives higher energy. Therefore it is the variational principle for electron density. To put it in a formula:

$$\left\langle \tilde{\Psi} \middle| \hat{H} \middle| \tilde{\Psi} \right\rangle = E(\tilde{\rho}) \geq E(\rho) = \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle \qquad \text{eq. 35.}$$

where $\hat{H}$ is the Hamiltonian, Ψ is the exact and $\tilde{\Psi}$ is the trial wave function.

The exact functional that provides the energy directly and solely from the electron density is not known, however, through the wave function we can give a definition. That is, the energy that corresponds to a certain electron density is the minimum of the expectation value of the Hamiltonian-operator with those wave functions that integrate to the given electron density.

$$E(\rho) := \min_{\Psi:\int \Psi = \rho} \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle \qquad \text{eq. 36.}$$

Even if we accept it as a "functional", we have to see that it depends on the system in focus.

Therefore the mathematical protocol to find the energy and the electron density of a system is the following: $E_0 = \min_{\rho:\int \rho = N} \left(\min_{\Psi:\int \Psi = \rho} \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle \right)$. That is, to minimize the energy first so, that we change the wave function but the density remains the same, and afterwards we change the density too, to get the lowest possible energy. This is a mathematically correct formula and the density is unambiguously determined by it, however, unfortunately, in practice it is absolutely useless.

Let us, however, expand it:

$$E_0 = \min_{\rho:\int \rho = N} \left(\min_{\Psi:\int \Psi = \rho} \left\langle \Psi \middle| \hat{H} \middle| \Psi \right\rangle \right) = \min_{\rho:\int \rho = N} \left(\min_{\Psi:\int \Psi = \rho} \left\langle \Psi \middle| \hat{T} + \hat{V}_{ee} + \hat{V}_{Ne} \middle| \Psi \right\rangle \right) \qquad \text{eq. 37.}$$

or,

$E[\rho]=T[\rho]+E_{ee}[\rho]+E_{Ne}[\rho]$, where T (or $\hat{T}$) stands for the kinetic energy of the given electron density, E_{ee} (or $\hat{V}_{ee}$) for the electron-electron interactions, and E_{Ne} (or $\hat{V}_{Ne}$) for the electron-nuclei interaction. The first two terms are called "universal", as they do not depend on the given system. The third term contains the position of the nuclei, therefore it depends on the system. However, after we specify the system (give the positions of the nuclei), the third term will be defined, too. Moreover it can be expressed without the wave-function as:

$$E_{Ne}[\rho] = \int \hat{V}_{Ne} \rho(r) dr \qquad \text{eq. 38.}$$

The minimization of the system dependent energy functional with respect to the density will not only provide the energy, but also the ground state density, and therefore other physical properties, too.

The variational problem of minimizing the energy functional E[ρ] can be solved by introducing a fictitious density functional of a non-interacting system: $E_f[\rho]=T_f[\rho]+V_f[\rho]$, or $E_f = \langle \Psi | \hat{T}_f + \hat{V}_f | \Psi \rangle$. If we chose this fictitious $\hat{V}_f$ functional as $\hat{V}_f = \hat{T} + \hat{V}_{ee} + \hat{V}_{Ne} - \hat{T}_f$, then the density functional of the non-interacting system will be the same as the density functional of the original, interacting system. Since the exact eigenfunctions of a non-interacting system have the form of a Slater-determinant, one can get the Kohn-Sham equations for the non-interacting system:

$$\hat{F}^{KS}\phi_i = \left[-\frac{1}{2}\nabla^2 + \hat{V}_f \right]\phi_i = \varepsilon_i \phi_i , \qquad \text{i=1, … N,} \qquad \text{eq. 39.}$$

where $\hat{F}^{KS}$ is the Kohn-Sham operator. Note that as in the Hartree-Fock method the Kohn-Sham operator depends on the system of orbitals as well. These (integro-differential) equations provide orbitals (ϕ_i, called Kohn-Sham orbitals) that reproduce the original density:

$$\rho_0 = \rho_f = \sum_{i=1}^{N} |\phi_i|^2 . \qquad \text{eq. 40.}$$

Furthermore:

$$\hat{V}_f = \hat{T} + \hat{V}_{ee} + \hat{V}_{Ne} - \hat{T}_f = \hat{V}_{Ne} + (\hat{T} - \hat{T}_f) + \hat{V}_{Coulomb} + (\hat{V}_{ee} - \hat{V}_{Coulomb}) =$$
$$\hat{V}_{Ne} + \hat{V}_{Coulomb} + (\hat{T} - \hat{T}_f) + (\hat{V}_{ee} - \hat{V}_{Coulomb}) = \hat{V}_{Ne} + \hat{V}_{Coulomb} + \hat{V}_{XC} \qquad \text{eq. 41.}$$

$\hat{V}_{Ne}$ is the electron-nuclei interaction, and is known if we know the electron density. Also, the Coulomb repulsion is known for a certain density. All that is unknown (the difference of the kinetic energy of the interacting and non-interacting system, and the remaining part of the electron-electron correlation above the Coulomb repulsion) is put into one term, the exchange functional ($\hat{V}_{XC}$).

Although we do not know the exact form of the exchange-correlation operator, so far all the equations are mathematically correct.

As this exchange functional depends on the electron density, to carry out the calculations again the recursive, self-consistent methodology has to be used.

There has been much work done to determine the best form of this exchange-correlation operator. Physicists use the model of the free electron gas to derive a form, and the resulting theory is called Local Density Approximation (LDA), because the functional depends only on

the density at the coordinate where the functional is evaluated. It works well for metals, but not at all useful for molecules. The local spin-density approximation (LSDA) is a generalization of the LDA to include electron spin. Generalized gradient approximations (GGA) are still local but also take into account the gradient of the density at the same coordinate. Using the latter (GGA) very good results for molecular geometries and ground state energies have been achieved. Potentially, more accurate than the GGA functionals are the meta-GGA functionals. These functionals include a further term in the expansion, depending on the density, the gradient of the density and the Laplacian (second derivative) of the density.

Difficulties in expressing the exchange part of the energy can be relieved by including a component of the Hartree-Fock exchange energy. Functionals of this type are known as hybrid functionals. The functional we used for our calculations, the B3LYP is also a hybrid functional.

Along with the component exchange and correlation functionals, three parameters define the hybrid functional, specifying how much of the exact exchange is mixed in. The adjustable parameters in hybrid functionals are generally fitted to a "training set" of molecules. That is the reason why these functionals cannot be systematically improved (unlike MP or other post-Hartree-Fock methods). As the machinery of DFT calculations is the same as of HF calculations, DFT methods do not require longer computational times, even though they include parameters for the electron correlation. In summary it can be said, that if the targeted molecules are described more or less accurately by the chosen "training set", then DFT results can reach the performance of the computationally much more demanding post-Hartree-Fock methods. However, for molecules that do not fall into this category DFT can produce fundamentally wrong results.

Peptides consist of H, C, N, and O, that are all small elements, and the B3LYP functional we used is thoroughly tested for them. Therefore the results are accurate enough to draw chemically significant conclusions from them.

1.3.2 Periodic boundary condition (PBC) calculations

First the molecular dynamics simulations used periodic calculations to decrease the side effects of having a finite box. Today there are several quantum chemical programs (e.g. SIESTA[76], Gaussian03[77]) that can also apply periodic boundary conditions during

calculations. During PBC (periodic boundary condition) calculations a unit cell is infinitely repeated. Therefore when calculating the interactions of one atom the nearest others are considered, even if they are in another cell. Therefore the radii of the "interaction ellipsoid" must be smaller than the edges of the unit cell.

1.3.3 Basis sets

The functions that stand for the atomic orbitals (that are later linearly combined to form molecular orbitals) are called basis sets. The first type of used functions is called *Slater Type Orbitals* (STO). These functions have an exp(-a*r*) term, and therefore at $r \rightarrow 0$ they exhibit a cusp. At $r \rightarrow \infty$ they converge to zero, as required. These types of functions describe the electrons quite well because they mimic the eigenfunctions of the hydrogen atom, however, they have a drawback: the integral of the product of two STO-s can only be computed numerically, as no analytical techniques are available. The program ADF (Amsterdam Density Functional) uses STOs.

Gaussian Type Orbitals have the form:

$$f^{GTO} = x^l y^m z^n \exp(-\alpha r^2), \qquad \text{eq. 42.}$$

which means that the function still converges to zero as $r \rightarrow \infty$ (although too rapidly), but it does not reproduce the cusp as $r \rightarrow 0$, (it is continuously differentiable at $r = 0$). Calculating the integrates of their products is much faster than that of STOs, as they can be expressed analytically and there are several efficient algorithms for speeding it up. To improve their behavior at zero and at infinity, most computational programs use Contracted Gaussian Functions (CGF), where GTOs with different exponential parameters are linearly combined with fixed coefficients. This way the computational costs do not rise too much, and these contracted orbitals resemble much more to the Slater type orbitals.

Carrying out calculations on a molecule that is made up of atoms, the usual chemical instinct says that the closed (core) shell of the electrons do not contribute too much to the bond formation, while the outer, valance shell electrons do. This concept is used when the number of functions describing an atom is decided. In the basis sets we used (that belong to the family of Pople basis sets) there is one contracted Gaussian orbital for the core electrons, and more orbitals for the valance electrons. Namely, the 6-31 stands for the following: a

contracted Gaussian (composed of six orbitals) for the core, and one single and a contracted Gaussian (composed of three orbitals) for the valance electrons.

For better describing the bonds sometimes orbitals of lower symmetry are also added to the atoms. These are called polarization functions, and are denoted with the (unfilled) atomic orbital they resemble the most. For example, 6-31G(d) means that a d type of orbital is also placed on the heavy (not hydrogen) atoms. (d,p) means that the heavy atoms have one set of d functions, and hydrogens have one set of p functions.

Another improvement is to add diffuse functions, that is, s-type orbitals with a lower exponent, which means that they decrease less rapidly. These kinds of functions are denoted with a + in the sign of the basis set, e.g. 6-31+G(d). They are most useful for describing anions and long range interactions.

1.3.4 Basis set superposition error (BSSE)

Calculations using the LCAO-MO approximation can sometimes lead to noticeable errors, e.g. when we want to determine the weak interaction between two or more molecules. The most common method to determine the weak interaction between molecules A and B is to calculate the energy of the AB complex, and then subtract from it the energy of the individual A and B. However, in the complex to describe A the basis functions of B are used as well, and vice versa. Therefore in the complex we use a seemingly larger basis set to describe the parts that leads to the decrease of energy according to the variation theory, and subsequently to bigger interaction energy between the two parts. This error is called Basis Set Superposition Error, BSSE.[78,79] There are several ways to calculate its value, one of the most used is the counterpoise method.[80] In this approximation the energy of the individual parts are calculated with all the basis functions of the complex. Therefore it is known how much the energies of the individual parts change when calculated with more basis functions, and so the artificial change in energy as well.

When the functions describing the electrons of one molecule already do it quite accurately (the basis set is large enough) the neighboring functions cannot improve much on it. Therefore the larger basis sets are used, the smaller is the BSSE. Of course, the computational requirements usually severely limit the size of the basis set.

2 Methods

The Gaussian03 software program was used for all calculations.[77]

Precision and accuracy of the applied methods is also discussed here, at the end of the chapter.

2.1 Potential Energy Surface calculations

The potential energy surface (PES, called a Ramachandran map in the case of peptides) of selected for N- and C-protected amino acids was calculated. The central amino acid residue was elongated with an acetyl-group at the N-terminal, and an N-methyl group at the C-terminal. The potential energy surface of protected glycine, alanine, proline and sarcosine (N-methyl glycine) was calculated at the B3LYP/6-31G(d) level of theory. For all amino acids except for proline the ϕ and ψ angle was scanned by 15 degrees. For proline the ϕ angle was varied only between 200º and 360º, but by a step size of 10º; as it is known (and can be seen on the map) that other regions are inaccessible because of the ring structure. The ψ angle of proline was varied between the usual 0º and 360º. The protocol was the following: the dihedral angles were set to a desired value, but all remaining properties of the molecule were allowed to relax. After that the energy minimum had been reached one of the dihedral angles was changed by 15º (or 10º). Using this step-by-step method the whole Ramachandran-map can be explored. For proline the map consists of 400 points, while for the other amino acids 625 points were determined.

2.2 Crystal structure calculations: models for amyloid

We carried out 1- and 2-dimensional periodic boundary condition calculations on short (3 residue long) peptides. These calculations require enormous memory and disc space, furthermore they need an excessive amount of CPU time even at a very low level of theory. (At HF/3-21G for 224 basis functions one structure optimization took 10 days with 6 CPU-s.) Therefore we optimized the crystal structure only at HF/3-21G level, and subsequently carried out single-point calculations at B3LYP/6-31G(d) level. During these optimizations not only

the structure of the molecule, but the length and the angle of the transition vectors were also optimized.

We performed calculations for two types of peptides: Ac-$(Ala)_3$-NHMe (standing for amino acids that have side chains and are chiral) and Ac-$(Gly)_3$-NHMe (representing itself, thus an amino acid without a side chain). In our calculations we examined all possible conformations where all the amide groups have two peptide bonds. The resulting structures are described in **Table 3**, and three of them are shown in **Figure 8**. In these structures, an important parameter is the angle closed by two ensuing peptide bonds that is called the tilting angle, and is further explained on **Figure 9**.

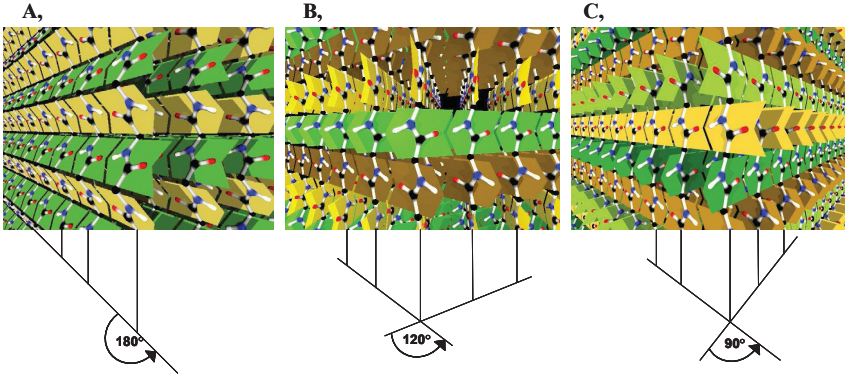


Figure 8. Schematic representation of the cross sections of the calculated endless structures. **A)** Multiple planes of 2D β-layers, thus 3D, with horizontal hydrogen bonds only between strands. **B)** A 3D aggregate where the horizontal hydrogen bonds are heading towards three directions due to the 120º tilt angle of the adjacent amide planes in the backbone. **C)** A 3D aggregate where the horizontal hydrogen bonds are heading towards four perpendicular directions due to the 90º tilt angle of the adjacent amide planes in the backbone.

Table 3. Number of neighboring strands and the corresponding tilting angles of the peptide chain in a crystal

	Resulting Structures					
Number of Neighbor Strands	**2**			**4**	**6**	
Tilting angle of peptide groups (θ)	θ=180º	θ=180º	θ=180º	θ=90º	θ=60º	θ=120º
Common description	2D parallel β-sheet, (one layer)	3D parallel β-sheet (multiple layers)	2D *anti*parallel β-sheet (one layer)			3D ε_L-conformation, collagen-like strand conformations

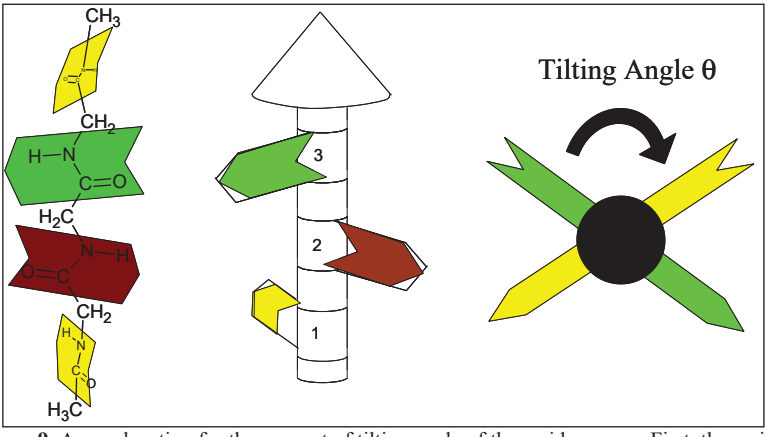


Figure 9. An explanation for the concept of tilting angle of the amide groups. First, the amide groups of a peptide chain can be regarded as arrows that have a plane. Second, when more or less elongated, this peptide chain can be regarded as a multi destination sign-post, where the signs are the arrows of the peptide bonds. Third, the tilting angle is the angle between the planes of the ensuing peptide planes.

2.3 Collagen models

We used two basic types of collagen models: first without bound water molecules to examine the stabilizing-destabilizing effects of different amino acids on the backbone of the triple helix. When examining the water binding strength at different positions we used models that contain explicit water molecules.

2.3.1 Waterless models for examining the backbone stability

To examine the backbone preferences and the triple-helix stabilizing effects the following model systems were constructed and subjected to full *ab initio* geometry optimizations. Six residue long N- and C-protected collagen triple helix models composed of a total of 18 amino acid residues were generated: *i)* glycine only (GGG collagen helix model), *ii)* from L-alanine only (AAA collagen helix model), *iii)* from L-alanine and glycine (AAG collagen helix model), *iv)* from L-alanine and D-alanine, (AAa collagen helix model) *v)* from L-prolines and glycine (PPG collagen helix model), *vi)* from L-prolines and D-alanine (PPa collagen helix model), *vii)* from sarcosine and glycine (SaSaG collagen helix model) and *viii)* from L-proline, L-hydroxyproline and glycine (POG collagen helix model). In all of these cases where more than a single type of amino acid residue is involved, glycines are placed appropriately and the polypeptide chains are suitably shifted: *e.g.* the -Gly-Xxx-Yyy-Gly-Xxx-Yyy-, -Yyy-Gly-Xxx-Yyy-Gly-Xxx- and –Xxx-Yyy-Gly-Xxx-Yyy-Gly- where chains are adjusted head to head. The AAG model contains alanine both at the Xxx and at Yyy positions, while the POG model has proline at the Xxx and hydroxyproline at Yyy positions. To test the effect of incorporating more than a single D-amino acid residue, the AAa and PPa model were used where both the Xxx and Yyy positions contained L-Ala or L-Pro but D-Ala was introduced where Gly should have been. These different models were created to model different "types" of collagen, to be able to follow the stabilizing effects on collagen. POG, PPG and PPa collagen helix models were supposed to mimic the "core" of collagen and so model the POG and PPG triplet containing parts of a triple helix. It is called the core, because according to melting point measurements[57] these natural triplets provide the highest stability to a triple helix. The appropriate experimental PDB structures are the 1V7H[63] and the 2CUO[64], respectively.

Also, there are numerous amino acid residues evenly spread throughout collagen none of the imino acid type, that are typical binding or cleavage sites. The GGG- and AAG-triplet

containing models could be used as candidates for studying such an Xxx-Yyy-Gly collagen type subunit, where neither Xxx nor Yyy is proline (or hydroxyproline). To test how well structural properties of the calculated collagen triple-helix correlates with experimental data of these binding-sites, we have used conformational values extracted from 1BKV.[58] This is a unique crystallized collagen-like triple helical molecule that contains a longer non-imino acid sequence, namely the -Ile-Thr-**Gly**-Ala-Arg-**Gly**-Leu-Ala-**Gly-** subunit.

The SaSaG triplet is used to examine a situation where Xxx and Yyy are both imino acids, but they do not have a ring to freeze the ϕ angle in the backbone. There is no appropriate X-ray or NMR structure for this triplet type.

Finally, AAA collagen model stands for the collagen having G-->A mutation (causing a bulge), and is compared to the 1CGD PDB structure[53].

These X-ray structures are either the only available of their kinds or the ones having the best resolution. All these models are shown in **Figure 10**.

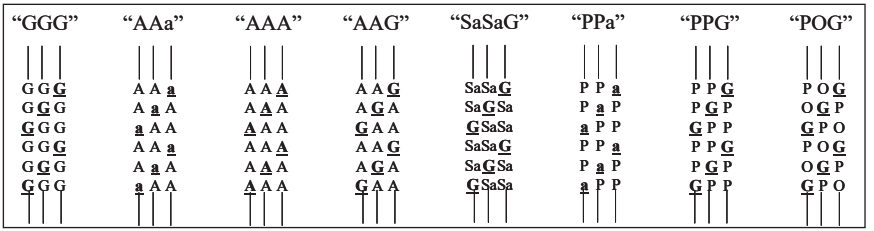


Figure 10. 3x6 residue containing models of different amino acid constitution that are used to examine the different backbone stabilizing effects on the triple helix.

All models of the same type of amino acid compositions were also optimized by adopting a triple-stranded parallel β-pleated sheet structure, and also as three individual strands. (see **Figure 11**)

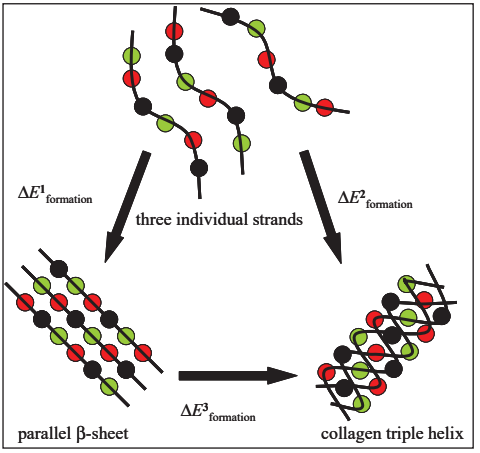


Figure 11. A hypothetical scheme of the collagen triple helix and the parallel triple-stranded β-sheet formation. There are three possibilities of how to calculate relative stability:
i) $\Delta E^1_{\text{sheet-form}} = E_{\text{sheet}} - [E_{\text{strand1}} + E_{\text{strand2}} + E_{\text{strand3}}]$,
ii) $\Delta E^2_{\text{collagen-form}} = E_{\text{triple helix}} - [E_{\text{strand1}} + E_{\text{strand2}} + E_{\text{strand3}}]$,
iii) $\Delta E^3_{\text{conversion}} = E_{\text{triple helix}} - E_{\text{sheet}}$
These three possibilities also exist for relative ΔH, ΔG and TΔS calculations.
Coloring identifies different sequential positions.

A molecular system of this size seems adequate for comparison with biochemical data, as the middle four amino acid residues behave very similar to those located in endless collagen triple helices, thus seen as the true building unit or "lego part" of this secondary structure.

All structures were fully optimized first at the RHF/3-21G and subsequently at the B3LYP/6-31G(d) levels of theory. Finally, energy calculations were completed both at the B3LYP/6-311++G(d,p) and B3LYP/PCM/6-31G(d) levels of theory on *a priory* optimized B3LYP/6-31G(d) structures. For the PCM calculations we used the IEF-PCM method[81] and water as a solvent, meaning that ε=78.39 was set.

Also, on all of the energy minimized structures frequency calculations within the harmonic approximation were carried out at the B3LYP/6-31G(d) level of theory. Gibbs free energy and entropy data are direct results of these calculations.

For model systems of similar type (*e.g.* multiple stranded β-sheets), BSSE (basis set superimposition error)[80] was found to be at subchemical ranges when computed at the B3LYP/6-311++G(d,p) level of theory.[82] To confirm the above results for three stranded

systems also, further BSSE studies were carried out on shorter models composed of alanine and glycine residues: *i*) a collagen triple helix composed of 9 amino acid residues ("short AAG collagen helix" model), and *ii*) a β-sheet model also composed of 9 amino acid residues ("short AAG sheet" model) were studied both at the RHF/3-21G and B3LYP/6-31G(d) levels of theory. Calculations for approximating the BSSE[80] were carried out using the counterpoise correction method[83], considering the three strands as three different subsystems.

2.3.2 Hydrated collagen models

To determine stabilities of water molecules connected to the tropocollagen *via* H-bonding(s), the following three types of model systems were optimized. The first one (**Figure 12**) was designed to examine the binding of the internal structural water, the so called ζ-type water, for which neither proline nor hydroxyproline can be at position Xxx. Therefore, in the 3x6 amino acid containing model both Xxx and Yyy positions are occupied by alanines. All six essential water molecules were placed in and the overall system was fully optimized at B3LYP/6-31G(d) level of theory.

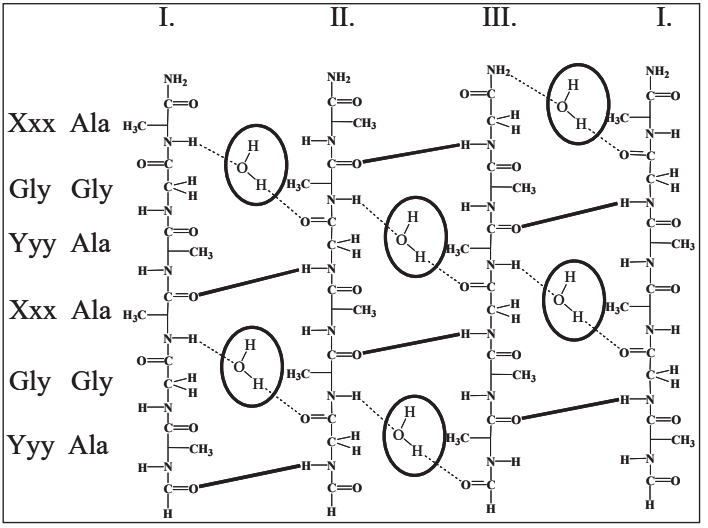


Figure 12. The model system characterizing the ζ-type water. The water molecules connect the amide hydrogen atom of residue Xxx (here Ala) with the C=O of Gly (dashed line). For the present model a total of 6 binding places exists, all marked.

The second type of model system (**Figure 13**) was designed to characterize the horizontal water thread. (**Figure 18** and **Figure 19**) Therefore, in this model both the α- and β-bridges were in focus. Some additional water molecules from neighboring bridges had to be added to the model to ensure the well-defined positions of the two bridges of interest and to avoid their shifting. There are two α^1 positions in these model systems, therefore we notate one of them with $\alpha^{1\alpha}$ and the other with $\alpha^{1\beta}$. $\alpha^{1\alpha}$ is part of the central α-bridge that has changing number of constituting water molecules, and $\alpha^{1\beta}$ is part of the central β-bridge that has changing number of constituting water molecules.

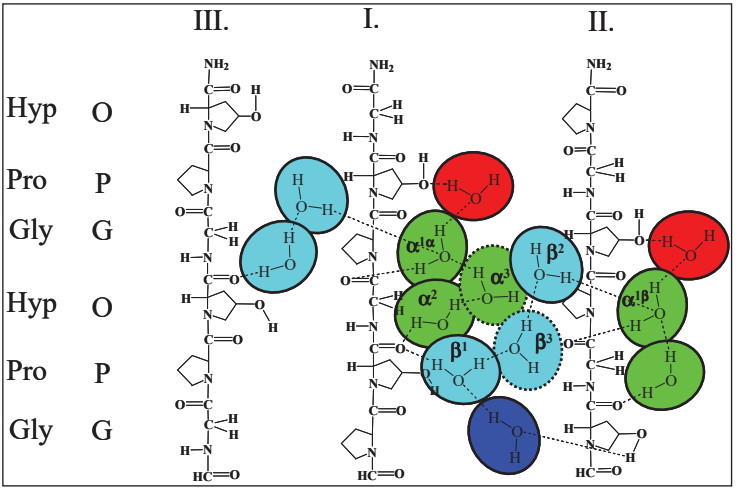


Figure 13. The model system characterizing the horizontal water thread. Note that both γ- and δ-bridges were also introduced to ensure that the α- and β-bridges remain unchanged. There are two optional waters, α^3 and β^3. Furthermore, $\alpha^{1\alpha}$ is an α^1 type water molecule considered an integrated part of the α-bridge and $\alpha^{1\beta}$ is an α^1 type water molecule considered as a part of the β-bridge.

The third type of supramolecular system (**Figure 14**) is to characterize the vertical water thread. (**Figure 18** and **Figure 20**) Thus, here the γ- and δ-bridges were in focus. Furthermore, we used this model to examine the effect of fluorine atom, replacing the γ-hydroxyl group of hydroxyproline, as the γ- and δ-bridges are both attached to this amino acid.

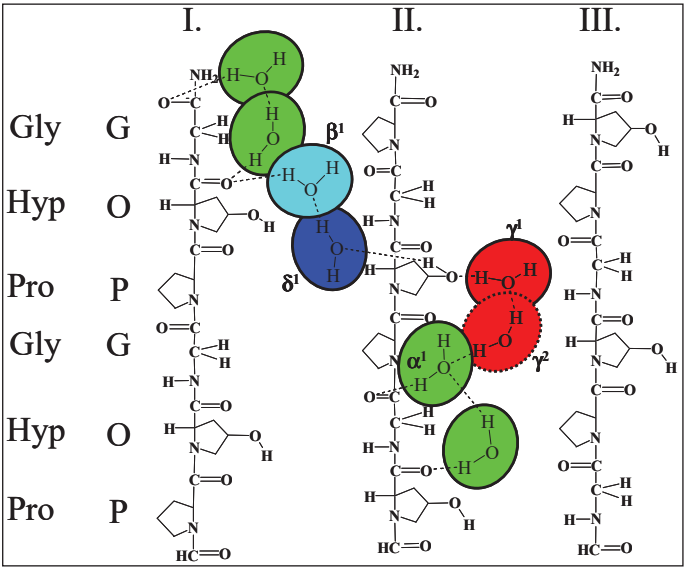


Figure 14. The model system characterizing the vertical water-threads. γ^2 is an optional water molecule. (In a set of these models the γ-OH group of the central hydroxyproline is changed to a fluorine atom.)

In each and every bridges, where X-ray crystallography data indicated (see **Table 1**) the total number of water molecules was altered, namely 2-3 in the α-, 3-4 in the β- and 2-3 in the γ-bridge. To test the effect of the OH → F substitution, two additional molecular structures were built, with a fluorine atom replacing the OH group of hydroxyproline. These models were also optimized both with two and three H_2O(s) forming the γ-bridge. The brief summary of these models are reported in **Table 4**.

Table 4. Number of water molecules in each of the model systems optimized

Type of model system:	Number of water molecules in water bridges						Calc. [a]
	α	β	γ	δ	OH↔F	ζ	Opt.+SP.
1st: AAG [b]	NA	NA	NA	NA	NA	6	1+6
2nd: horizontal thread [c] (α and β bridges)	2-3	3-4	2	2	OH	NA	4+24
3rd: vertical thread [d] (γ bridges and OH→F subst.)	2	3	2-3	2	OH-F	NA	4+18

a, Total number of supramolecular complexes optimized (Opt.) and subjected to energy calculation (SP.)
b, 18 residue (alanine and glycine) incorporating model (see **Figure 12**)
c, 18 residue (proline, hydroxyproline and glycine) incorporating models (see **Figure 13**)
d, 18 residue (proline, hydroxyproline and glycine) incorporating models (see **Figure 14**)

For the first type Ala and Gly, for the second and the third types of model systems a Pro, Hyp and Gly residues were incorporated within the collagen model (**Table 4**). All supramolecular systems are composed of 3x6 amino acid residues with water bridges of different length as required for modeling the X-ray data (see **Figure 13** and **Figure 14**). The structure of all these hydrated polypeptide models were fully optimized first at RHF/3-21G level of theory and subsequently reoptimized with the ONIOM method, where the collagen base was still calculated at RHF/3-21G, while all the waters and the carbonyl and hydroxyl groups attached to them were considered at B3LYP/6-31G(d) level of theory. This was carried out to ensure that hydrogen bonds of interest are treated at a higher level of computation. The first type of model system composed of alanine, glycine and six water molecules were entirely optimized at B3LYP/6-31G(d).

For the determination of the binding energy of each water molecule a counterpoise, BSSE, single-point calculation was carried out at the B3LYP/6-31G(d) level of theory, where one part (group of atoms) was the water molecule in focus and the other one was all the other atoms of the supramolecular system. So, for example, to determine the binding strength of the 3rd water molecule in the AAG water model (**Table 2** and **Figure 12**) the third water molecule was taken as the first part, while the overall triple-helix with the remaining five water molecules was treated as the second part of the system. The counterpoise single-point calculation gives a corrected total energy for the molecular ensemble of two subsets, from which the energies of the two individual systems are subtracted. The remaining energy is

what we call water binding energy or interaction energy of the system composed of the above two subsets:

$$\Delta E_{\text{binding}} = E_{\text{BSSE corrected}} - (E_{\text{water, part1}} + E_{\text{part2}}) \qquad \text{eq. 43.}$$

For example, applying this equation on the 3rd water of the AAG model the following calculations were completed: E of part 1 is the total electronic energy of the 3rd water molecule counted from the N-terminal. E of part 2 is the total electronic energy of the "rest" of the supramolecular complex, namely that of the 3 peptide chains plus the remaining 5 water molecules together. The energy of a single and isolated water molecule ($E_{\text{water, part1}}$) is –76.4088080059 Hartree ($E_{\text{water, part1}}$), while for the rest of the molecular system alone (without this water molecule) has E= –5107.93328851 Hartree (E_{part2}). Altogether this system has an energy of –5184.37342967 Hartree, which changes to –5184.36408998063 Hartree upon BSSE correction. The latter term is the $E_{\text{BSSE corrected}}$. Thus, the binding energy according to eq.43 is $E_{\text{BSSE corrected}} - (E_{\text{water, part1}} + E_{\text{part2}})$ = –5184.3640900 – [–76.4088080 + (–5107.9332885)] Hartree = –0.0219935 Hartree, which is –13.80 kJ·mol^{-1}.

In this way, we ensure that the interaction between the water molecule in focus and the tropocollagen model system is determined at an acceptable level of accuracy. The BSSE corrected binding energies, and their H-bond number normalized figures are reported in **Table 12**, **Table 16**, **Table 17** and **Table 18**.

The binding energy of a water molecule in a reservoir of waters, H_2O ensemble, is highly dependent on the calculation approach taken:[84] different methods provide different explicit results. Thus, water stabilities and binding energies presented here are to determine relative orders rather than absolute values.

2.4 Precision and accuracy

2.4.1 Periodic peptide models

As they were directly designed to overcome the drawbacks of dealing with finite systems the capping effect does not play a role in the case of the endless structures. However, as we used a very small basis set, the BSSE problem must be examined.

First of all, to reduce the effect, single point energy calculations were carried out at the B3LYP/6-31G(d) level on the HF/3-21G optimized structures. These calculations with the increased basis set did not change the energy order. Therefore we can conclude, that although BSSE is surely present, it does not affect the results too much in this situation, where each of the structures is closely packed.

To test the size of the unit cell a calculation pair was carried out: the single layer β-pleated sheet (180° tilting angle) was calculated twice. First the unit cell contained only one tripeptide, as for the calculations with other tilting angles. For the second time the unit cell contained two tripeptides in a parallel arrangement. This structure with the doubled unit cell was also optimized at the HF/3-21G level, using the periodic boundary conditions. The results show that the energy and the conformation of one tripeptide unit remains the same, an unlikely event for non-periodic systems. Therefore the size of the model system does not affect the results, which is the desired situation.

2.4.2 Collagen models without hydration

To confirm the small magnitude of the BSSE, the relative stability of the collagen-like helix and the triple-stranded extended-like backbone structure were computed for shorter models composed of 9 amino acid residues at different levels of theory. (**Table 5**)

Table 5. The relative stability ΔE[kcal·mol^{-1}] of the short-AAG collagen helix and AAG β-strand models, and the error introduced by the neglect of counterpoise correction in calculations.

	Type of Superstructure	**Stability** (ΔE_{form}) in kcal·mol^{-1} for a triplet (for the whole system)	
		B3LYP/6-31G(d)	B3LYP/6-311++G(d,p)// B3LYP/6-31G(d)
Without Counterpoise Correction: **BSSE is present**	$\Delta E_{collagen\text{-}form}$ (short-AAG col-helix)[a]	−13.99 (−41.97)	−9.25 (−27.74)
	$\Delta E_{sheet\text{-}form}$ (short-AAG sheet)[b]	−18.85 (−56.56)	−13.89 (−41.68)
	Stability difference: $\Delta E_{conversion}$[c]	+4.86 (+14.59)	+4.65 (+13.94)
With Counterpoise Correction: **No BSSE**	$\Delta E_{collagen\text{-}form}$ (short-AAG col-helix)[a]	−7.00 (−21.01)	−8.16 (−24.48)
	$\Delta E_{sheet\text{-}form}$ (short-AAG sheet)[b]	−12.22 (−36.65)	−12.90 (−38.69)
	Stability difference: $\Delta E_{conversion}$[c]	+5.21 (+15.64)	+4.74 (+14.21)
Error introduced if BSSE neglected		**−0.35 (−1.06)**	**−0.09 (−0.27)**

a, $\Delta E_{collagen\text{-}form} = E_{triple\ helix} - [E_{strand1} + E_{strand2} + E_{strand3}]$
b, $\Delta E_{sheet\text{-}form} = E_{sheet} - [E_{strand1} + E_{strand2} + E_{strand3}]$
c, $\Delta E_{conversion} = E_{triple\ helix} - E_{sheet}$

In the case of the AAG-type short model the destabilization energy of the collagen helix with respect to β-strand at the B3LYP/6-31G(d) level of theory, with and without BSSE is: $\Delta\Delta E_{BSSE}$ = 4.86 and $\Delta\Delta E_{no\text{-}BSSE}$= 5.21 kcal·mol^{-1}, respectively. Thus, the error introduced when BSSE is ignored is clearly below the chemical precision, namely ~ –1 kcal.mol^{-1} [4.86 –

5.21 = –0.35 kcal·mol^{-1}] **Table 1**. This difference becomes significantly smaller (−0.09 kcal·mol^{-1}) at a higher level of theory {B3LYP/6-311++G(d,p)//B3LYP/6-31G(d)}. Note that the latter measure is −0.03 kcal·mol^{-1} per amino acid residue, which is much lower than the range of the so-called chemical accuracy. Therefore, although BSSE is present, the magnitude of the error introduced is minimal when using B3LYP/6-311++G(d,p)//B3LYP/6-31G(d) results for stability comparisons, concordantly calculations at this level are adequate, even though they disregard BSSE.

For every calculated triple helix structure the starting and closing three amino acids (two from each strand, altogether a total of six) have somewhat distorted local conformations with respect to the "ideal", triple helical conformation called as capping effect.

To test the magnitude of such a "capping effect" on stabilization (the association) energy (**Figure 3**), we have carried out endlessly repeated ("crystal") calculations on a seven amino acid long POG model, at RHF/3-21G level of theory. This model was chosen because for crystals calculations the symmetry of the helix has to be known, and it is certain only for the POG and PPG models (7_5). $\Delta E^2_{\text{collagen-form}}$ (= $E_{\text{triple helix}} - [E_{\text{strand1}} + E_{\text{strand2}} + E_{\text{strand3}}]$) of the endless structure was compared to that of the POG collagen helix model (also computed at RHF/3-21G level of theory), where the latter model has serious capping effect. The result reveals that although the POG triple helix has capping effect at both ends, this structural flaw does not affect dramatically the collagen formation energy; discrepancy of about 1.5% is observed. Unfortunately this is not so nice for the PPG triplet. Carrying out the above mentioned procedure for the PPG triplet, the discrepancy between the endless structure and the finite model is 12%! However, these crystal calculations are not used here, because the HF/3-21G method cannot describe the fine H-bonding interactions that have high effect on the energy of the structure. (In the previous part there were the same number of hydrogen bonds in every structure, therefore the effect was more or less the same, unlike here.) This suggests that even though crystal calculations have already proven to be very useful, now with the use of a finite model system much higher level of theory can be used [*e.g.* B3LYP/6-311++G(d,p)//B3LYP/6-31G(d) instead of RHF/3-21G] and therefore much more accurate stability data can be obtained.

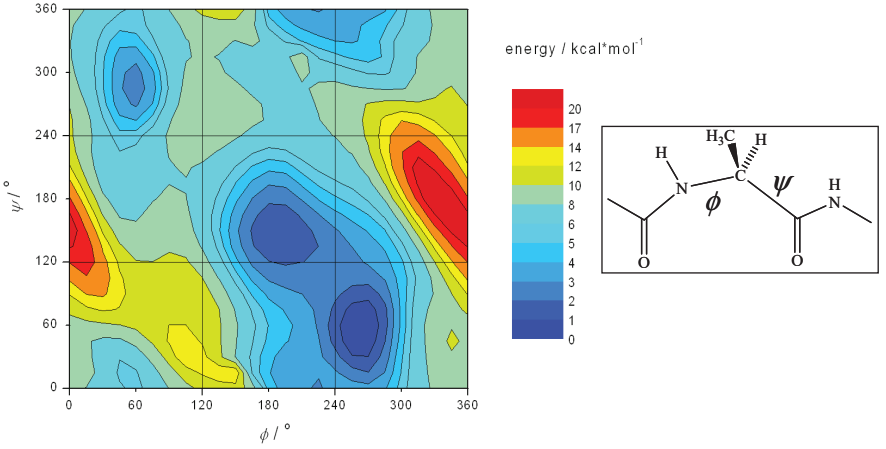
3 Results

3.1 Ramachandran maps

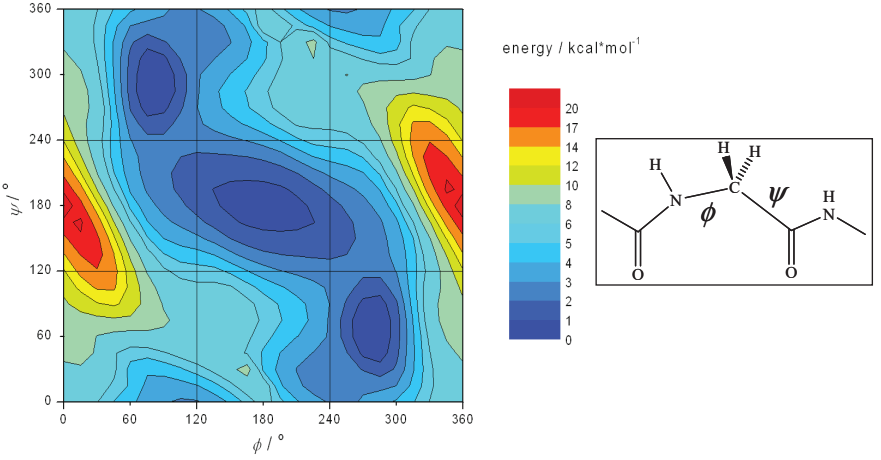
The potential energy surface (PES) maps of peptides are sometimes called Ramachandran-maps. They describe the energy of a residue in different configurations. For amino acids the ϕ and ψ angles are usually varied.

The Ramachandran-map of protected L-alanine (Ac-Ala-NHMe), glycine (Ac-Gly-NHMe) sarcosine (Ac-sarcosine-NHMe) and L-proline (Ac-Pro-NHMe) were constructed (for proline the ϕ angle was varied only between 190° and 360°). (**Figure 15**) Of course, these are not entirely new results, the topology of these Ramachandran-maps are already known[34,85,86,87,88], nevertheless they are presented here as they are extremely helpful for understanding the conformational and energetic properties of the supramolecular complexes discussed below.

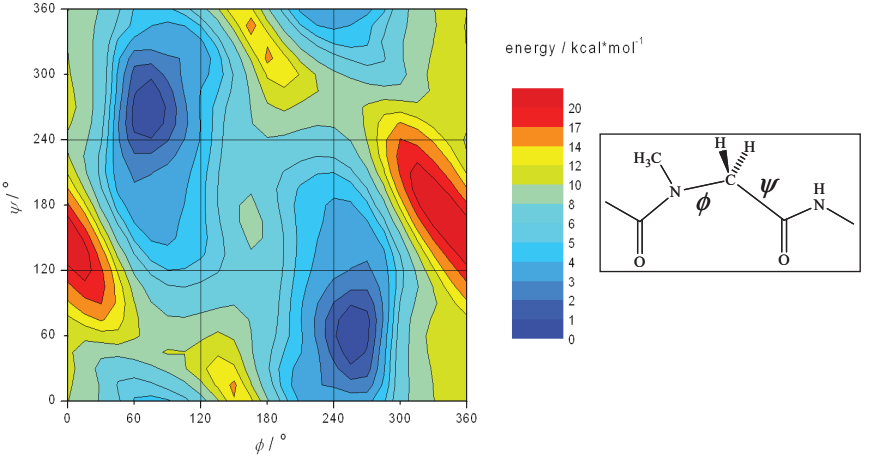
a, Ac-alanine-NHMe



b, Ac-glycine-NHMe



c, Ac-sarcosine-NHMe



d, Ac-proline-NHMe

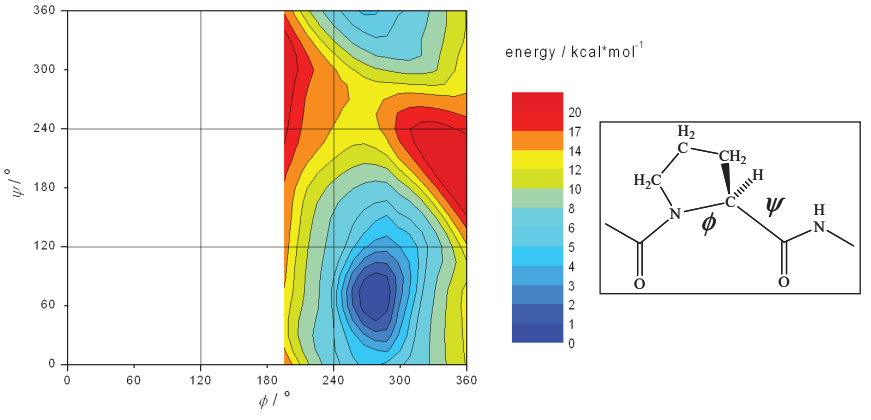


Figure 15. Ramachandran-maps of a, L-alanine (Ac-Ala-NHMe), b, glycine (Ac-Gly-NHMe), c, sarcosine (Ac-sarcosine-NHMe) and d, L-proline (Ac-Pro-NHMe) calculated at B3LYP/6-31G(d)

It can be seen that the L-alanine derivative has 3 main conformations, at the γ_D, β_L and γ_L regions (the lowest energy corresponds to the γ_L). It has one very high energy region (E > 10 kcal·mol^{-1}) around $\phi = 0°$, $\psi = 180°$. The diamide unit containing glycine and sarcosine residues both have a quite symmetric potential energy surface (as it is normal for a non-chiral amino acid). The map of the glycine derivative has very small area of energies higher than 10 kcal·mol^{-1}, and it is also around $\phi = 0°$, $\psi = 180°$. It has also three main minimal energy regions, γ_D, β_L and γ_L, as alanine, but they are all symmetric. The map of the sarcosine derivative has only two minimal energy regions: γ_L and γ_D, as it has not got a minimum in the β_L part (normal for an imino acid). Similar to the L-alanine derivative, the protected sarcosine has quite low energies at half of the ε_L part. The PES of protected proline has only a single minimum, in the γ_L region, but is has to be noted, that in around one third of the ε_L region (**Figure 3**) the total electronic energy is quite low: E< 6 kcal·mol^{-1}.

3.2 Crystal calculations

3.2.1 Calculated structures

This part of the results is to answer the question: "How do so many different peptides and proteins form amyloid-like plaques, and why do plaques have the β-pleated sheet structure?" Other members of our research group (András Perczel and Péter Hudáky) carried out calculations on dimer peptides, and showed that among the known secondary structure elements (β-pleated sheet, α-helix, β- and γ-turns) the antiparallel β-pleated sheet is the most stable one.[89] Here we show that this is applicable to all the structures, where the polypeptide chain is replicated periodically. We commence with a deduction that is supported by quantum-chemical calculations afterwards.

If one wants to fill the space (*tile*) periodically with "peptide units", and wants to find the conformation of the peptide in which the whole crystal is energetically optimal, it can be supposed, that

- each peptide group should be connected to another by hydrogen bonding,
- peptide or protein chains are laid parallel to each other, otherwise there would be parts where the hydrogen bonding with another peptide group is not guaranteed.

Furthermore, when the same peptides are used it can be assumed that the arrangement of the molecules is the same. Describing a polypeptide as several peptide groups along a chain, one can use such molecule as a tile, to cover the 3D-space with. (see **Figure 9** and **Figure 16**) Therefore, as the molecule itself is placed into one dimension, from hereby the problem is reduced to tiling the plane with polygons.

To tile the plane periodically with identical polygons triangles, squares and hexagons, or – as an extremum – parallel lines can be used. (**Figure 16** a,) In these cases the central polygon has three, four, six, and two adjacent neighbors, respectively.

When we substitute the polygons with the peptides they represented, it can be imagined, (see **Figure 16** c,) that having a side chain on the central peptide does not let the neighbors to come close enough to form strong hydrogen bonds. The only possible position of the strands is when they have two neighbors, in other words when the peptides that form hydrogen bonds are connected in a straight line. This molecular structure is the well-known β-pleated sheet.

However, when there are no side-chains on the central peptide unit the other possibilities cannot be ruled out. That is the most that can be said by theoretical deduction.

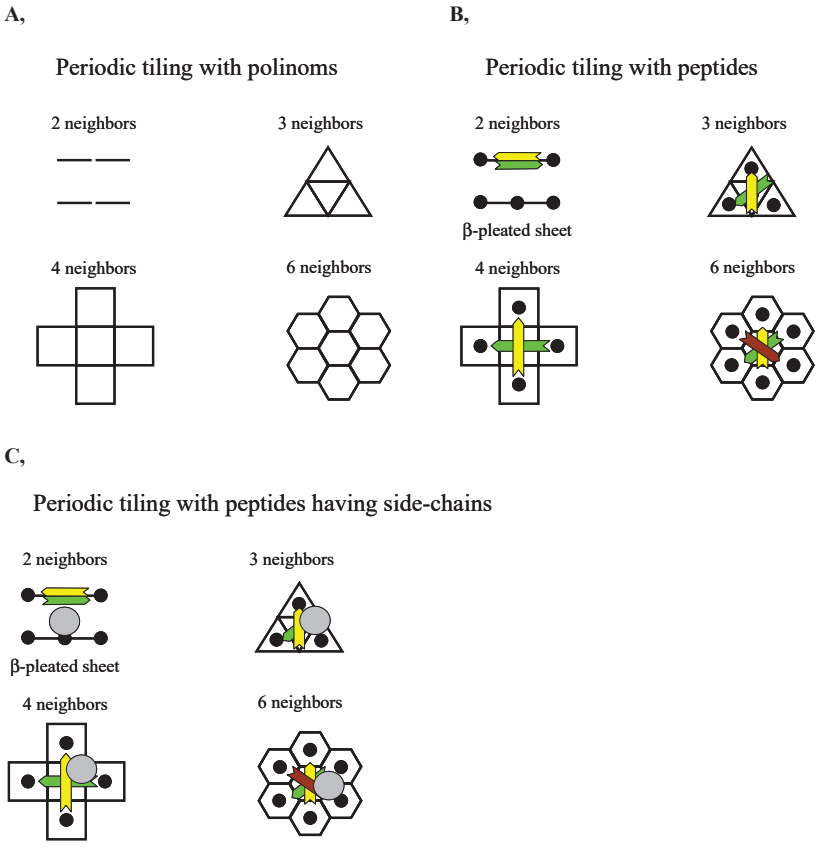


Figure 16. A, Tiling the plane (2D-space) with identical polygons. **B,** Tiling the space with oligopeptides composed of glycine (no side-chain) residues only: putting one peptide chain perpendicular to the plane into the center of each polygon. (Each oligopeptide is marked as a large black spot.) The colored arrows represent hydrogen binding between the peptides. **C,** Tiling the space with peptides composed of residues having side-chain (grey circle on the figure): putting one peptide chain perpendicular to the plane into the center of each polygon. The colored arrows represent hydrogen binding between the peptides.

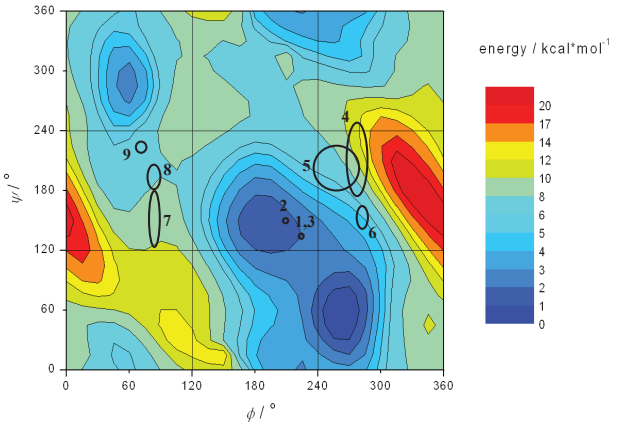
For supporting and extending this theory to those structures that we cannot compare by this simple deduction quantum-chemical calculations on modelpeptides were accomplished that form the above structures.

The planes of the peptide groups were set in directions so that the central molecule can form peptide bonds with the desired number of neighbors. When the number of neighbors is two or four, than the tilting angles between the peptide planes are 180° and 90°, respectively. (see **Figure 16** and **Table 3**) To have a central peptide that is connected to three other peptide units, the tilting angle must be set to 120°. However, in this case we can realize that one peptide group can form hydrogen bond with only one another molecule, as there is no possible partner on its "other side". Therefore this case can be ruled out, as here the number of hydrogen bonds formed would be only half of the maximum possible numbers. The last possibility is when one molecule has six neighbors, and here also one peptide group can form hydrogen bonds with two others, therefore it is also a configuration to be examined. It should be noted, that in case of six neighbors there are two possibilities for the tilting angles: 60° and 120°, both are to be examined by calculation as well.

Furthermore, it is to be considered that peptides are typically composed of amino acids that have a side chain of a given size. These types of residues are chiral. Therefore the direction of the tilting angle becomes important, as a −60° tilting of the peptide bonds does not result in the same structure as the +60° tilting, unlike for the achiral polyglycine.

The conformation (dihedral angles) and the average C_α distances of the neighboring peptide units are listed in **Table 6**. To emphasize these results, the resulted dihedral angles are also put onto the corresponding Ramachandran-map (**Figure 17**).

A,



B,

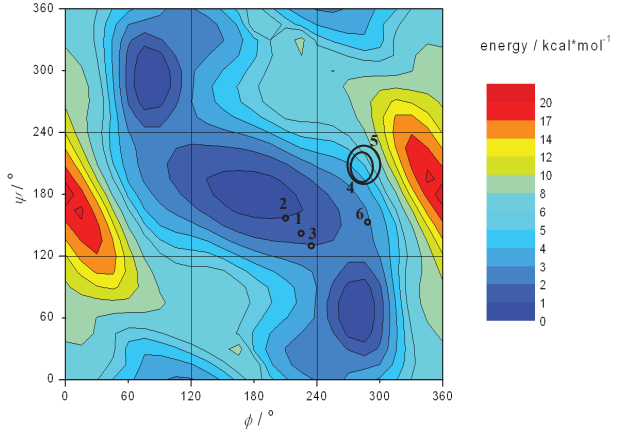


Figure 17. Ramachandran-maps of **A,** L-alanine diamide (Ac-Ala-NHMe), **B,** glycine diamide (Ac-Gly-NHMe), with the average calculated dihedral angles of the periodic structures. **A,** 1: 2D parallel β-layer; 2: 2D *anti*parallel β-layer; 3: 3D parallel β-layer; 4: 3D packing, $\theta = +60^{o}$; 5: 3D packing, $\theta = +90^{o}$; 6: 3D packing, $\theta = +120^{o}$; 7: 3D packing, $\theta = -60^{o}$; 8: 3D packing, $\theta = -90^{o}$; 9: 3D packing, $\theta = -120^{o}$. **B,** 1: 2D parallel β-layer; 2: 2D *anti*parallel β-layer; 3: 3D parallel β-layer; 4: 3D packing, $\theta = \pm 60^{o}$; 5: 3D packing, $\theta = \pm 90^{o}$; 6: 3D packing, $\theta = \pm 120^{o}$.

Concerning the Ac-$(Ala)_3$-NHMe modelpeptide, from **Table 6** it is clear, that the peptide chains are the closest to each other in one of the β-sheet structures (the average distance is 4.72 Å for the parallel and 4.67 Å for the antiparallel form). All the other 3D structures have much higher (>5 Å) C_α distances and consequently longer H-bonds that are significantly weaker[40]. The situation is completely different for the glycine containing model structures. Here the modelpeptides stay approximately at the same distance in all the structures (C_α distances vary from 4.68 Å to 4.75 Å only). These conformations are also put onto the corresponding Ramachandran-map, see **Figure 17/B**. Another interesting point is the uniformity of the repeated peptide that is reflected in the standard deviations of the dihedral angles. Where it is small the residues have the same conformation and the screwing of the chain is steady. This is the case for all types of β-sheet structures (for both the glycine and alanine containing peptides), and for the 3D ε-layer, built from glycines. These dihedral angles are in low energy regions of the Ramachandran-map, therefore the amino acids can take up these conformations easily. However, where a structure has large standard deviations for the two dihedral angles (e.g. 3D packing with tilting angles of ±90°, **Figure 17**), it means that the tilting angle of the amino acids is maintained, but this causes a significant strain in the residues and each of them try to lessen it in different ways. If we look at those angles on the Ramachandran-maps (**Figure 17**), we can see that these conformations belong to higher energy regions.

Table 6. Conformation (dihedral angles) and the average C_α distances of the neighboring peptides (the central peptide has hydrogen-bonding with) in the different crystal structures

Constitution of the building blocks $(Xxx)_3$	Property	Structural Parameters of the Type of Lattice					
		2D parallel β-layer, $\theta = 180°$	**2D *anti*parallel β-layer, $\theta = 180°$**	**3D β-layer untilted: $\theta = 180°$**	**3D packing H-bonds tilted by $\theta = \pm 60°$**	**3D packing H-bonds tilted by $\theta = \pm 90°$**	**3D ε-layer H-bonds tilted by $\theta = \pm 120°$ [b]**
Achiral (Xxx = Gly)	**Dihedral angles**	$\phi = -132.7 \pm 1.2$ $\psi = -132.2 \pm 2.6$	$\phi = -144.8 \pm 1.8$ $\psi = 147.9 \pm 2.3$	$\phi = -125.8 \pm 0.7$ $\psi = 122.2 \pm 0.6$	$\phi = -91.3 \pm 17.0$ $\psi = -164.9 \pm 28.7$	$\phi = -85.7 \pm 26.6$ $\psi = 193.2 \pm 38.0$	$\phi = -79.3 \pm 1.4$ $\psi = 154.8 \pm 1.4$
	Average C_α distance[a]	4.71	4.68	4.69	4.72	4.68	4.75
Chiral (Xxx = L-Ala)	**Dihedral angles**	$\phi = -131.9 \pm 1.4$ $\psi = 128.9 \pm 2.9$	$\phi = -148.8 \pm 1.6$ $\psi = 149.3 \pm 3.0$	$\phi = -130.6 \pm 3.0$ $\psi = 126.7 \pm 5.6$	$\phi = -84.2 \pm 13.0$ $\psi = -147.8 \pm 72.0$	$\phi = -99.9 \pm 39.5$ $\psi = 198.2 \pm 45.8$	$\phi = -87.3 \pm 8.6$ $\psi = 150.8 \pm 17.8$
	Average C_α distance[a]	4.72	4.67	4.72	5.49	5.01	5.51
	Dihedral angles	N.A.[b]	N.A.	N.A.	$\phi = 78.4 \pm 7.6$ $\psi = 155.7 \pm 60.9$	$\phi = 78.1 \pm 10.0$ $\psi = -168.7 \pm 20.4$	$\phi = 70.6 \pm 8.8$ $\psi = -144.7 \pm 8.2$
	Average C_α distance[a]	N.A.	N.A.	N.A.	5.44	5.24	5.36

a, distance in Å

b, N.A.: not applicable, as the conformation with $\theta = -180°$ is the same as the conformation with $\theta = +180°$

3.2.2 Stability of the calculated structures

The energies of the examined periodic structures are reported in **Table 7**. Peptides that have side chains are represented by an alanine containing tripeptide Ac-$(Ala)_3$-NHMe. The three most stable structures are some forms of the β-pleated sheet. The most stable form is the antiparallel, single layer β-pleated sheet, then the parallel single layer β-pleated sheet, and than the parallel multiple layered (3D) β-pleated sheet, that is still only 1.7 kcal·mol^{-1} less stable than the antiparallel form. It has to be noted that the 2 dimensional β-pleated sheet (single layer) is slightly more stable than the one where this layer is periodically repeated. As the peptides have a methyl side chain only, there is not much possibility to form interlayer connections and thus lower the energy of this type of structure. However, as seen for short peptides[13], the situation is likely to change when the system incorporates other amino acids as well.

The least stable structure is where the tilting angle is −60º. It is interesting, that the energy gap between "2-neighbored" and 4- or 6- neighbored structures is quite large, more than 10 kcal·mol^{-1}. That means that the calculations eventually follow the deductions we made by common sense. That is, the presence of a side chain forces such a big distance (as already seen in **Table 6**) among the peptide chains that the H-bonding becomes too weak.

Peptides that do not have side chains are built up of glycine (as this is the only amino acid without a side chain). The energy values for the optimized poly-Ac-$(Gly)_3$-NHMe can be seen in **Table 7**. The most stable structure is where one peptide is surrounded by six others, and the tilting angle of the amide groups is 120°. In this superstructure each molecule has a polyproline II-like backbone conformation which was also described by Crick and Rich[90]. The least stable structure is the other hexagonal, where the amide planes are tilted by 60°. In all these structures the distance of the peptide chains remain more or less the same. Therefore, as the strength of the H-bonds cannot differ so much, here the selection might also be based on the conformation preference of a glycine residue. This also indicates that even if the structure with 60° tilting angle has six neighbors (the most), and consequently its energy is most lowered by the BSSE, this error is more or less the same in the other structures and does not cause a change in the energy order of the model systems. That is quite relieving, as we have carried out the calculations on quite a simple level (B3LYP/6-31G(d)//HF/3-21G).

In summary we can state that the most stable conformation for peptides which have side chains is one type of β-sheet, while for polyglycine the most stable aggregation is hexagonal and the amino acids have ε_L conformations, the same as in collagen.

Table 7. The energies of the infinite model structures calculated at B3LYP/6-31G(d)//HF/3-21G level of theory

	Type of the lattice (structural properties of the aggregate)					
Constitution of the building blocks $(Xxx)_3$	2D parallel β-layer, $\theta = 180^o$	2D *anti*parallel β-layer, $\theta = 180^o$	3D β-layer untilted: $\theta = 180^o$	3D packing H-bonds tilted by $\theta = \pm 60^o$	3D packing H-bonds tilted by $\theta = \pm 90^o$	3D ε-layer H-bonds tilted by $\theta = \pm 120^o$ [b]
Achiral (Xxx = Gly)	8.68 [a]	7.03	3.96	8.84	7.36	0.00[c]
Chiral (Xxx = L-Ala)	1.17	0.00 [d]	1.73	12.79 ($\theta = +60^o$)	13.23 ($\theta = +90^o$)	17.57 ($\theta = +120^o$)
	N.A.	N.A.	N.A.	28.31 ($\theta = -60^o$)	25.55 ($\theta = -90^o$)	23.79 ($\theta = -120^o$)

a, Relative energies are in kcal·mol^{-1}.
b, Hexagonal arrangement resulting in a polyproline II structure for the molecule
c, E= −872.6138729 Hartree
d, E= −990.5514393 Hartree
e, N.A.: not applicable, as the conformation with $\theta = -180^o$ is the same as the conformation with $\theta = +180^o$

3.3 The internal stability of collagen

3.3.1 Calculated structures

Collagen's triple helical structure consists of amino acid residues of the same conformation, sometimes called PPII, or ε_L. Both notations stand for a residue having the dihedral angles: $\phi \approx -70^{\circ}$, $\psi \approx +150^{\circ}$. In the following parts we will try to shed some light on how collagen manages to maintain this conformation in three strands, which is otherwise not an energy minimum for any of the amino acids, as seen in the Ramachandran-maps (**Figure 15**).

For these purposes we will compare the stabilities of three individual strands with that of the collagen triple helix and a triple stranded parallel β-sheet for several amino acid compositions, but first we compare the calculated triple helical structures and the measured X-ray structures. The RMSD (root mean square deviation) of the dihedral angles of all the calculated and X-ray structures from each other are shown in **Table 8**.

3.3.1.1 Triple helices

The calculated overall average of the dihedral angles of both types of secondary structures can be seen in **Table 9**. The structural properties of the calculated triple helices are further detailed in **STable 1** (The average dihedral angles change a little according to the sequence position of the amino acid (Xxx, Yyy or Gly), therefore in **STable 1** the averages are shown according to positions.) As in the β-pleated sheet there are no such positions only the overall average is shown for this superstructure. As the beginning and ending amino acids do not always have a conformation that is appropriate in the relevant secondary structure (capping effect), all these averages are calculated only from the middle four amino acids in one strand.

Looking at the RMSD of the X-ray structures from each other in **Table 8**, it can be seen that the POG and PPG measured triple helices are quite close to each other, as their RMSD is only 4°. Both of them differ from the AAG-type X-ray structure (RMSD 16° and 28°, respectively). It is interesting that the PPG measured structure is much "further" from the AAG-type that the POG, although the POG is said to be the ideal triplet for a collagen triple helix, and the AAG-type represents a somewhat loosened, slightly "unfolded" triple helix. All these previously mentioned three structures differ much from the POA (RMSD 28°, 31° and

24°, respectively), where the structure has a bulge, and it absolutely does not belong to a well-folded triple helix. It is not a surprise that from the three folded structures the AAG-type is the closest to the POA. Again, the PPG is further from the POA than the POG, although the differences between the RMSD values are not so large as it was for the AAG-type structure.

Table 8. The RMSD of the average backbone values of all the calculated and measured collagen structures[a]

		X-ray				calculated							
		POG	**PPG**	**AAG**	**POA**	**POG**	**PPG**	**PPa**	**SaSaG**	**AAG**	**AAa**	**GGG**	**AAA**
X-ray	**POG**	0											
	PPG	4	0										
	AAG	16	28	0									
	POA	28	31	24	0								
calculated	**POG**	11	12	21	26	0							
	PPG	4	7	14	25	11	0						
	PPa	8	10	15	27	14	7	0					
	SaSaG	25	22	30	42	23	27	30	0				
	AAG	15	17	17	21	20	12	14	33	0			
	AAa	31	31	16	28	34	29	27	40	27	0		
	GGG	28	28	37	38	21	29	31	21	32	46	0	
	AAA	51	51	51	54	49	50	51	44	47	52	38	0

a, Root mean square deviation of the calculated and measured dihedral angles in degrees

There are also a number of interesting aspects about the RMSD of the calculated structures concerning these four measured structures (**Table 8**). The calculated POG model is the closest to the POG X-ray structure (RMSD = 11°, the rest are 12°, 21° and 26°, respectively). Both the PPG and PPa calculated models are also closer to the measured POG than to the PPG X-ray structure (4° < 7° and 8° < 10°, respectively). However, they are still closer to the measured PPG structure than the calculated POG to its experimental counterpart.

The rest of the computed structures resemble much less to any of the measured structures, indicating that these helices are slightly (or somewhat more) unfolded. The closest to experimental structures are the AAG and AAa calculated ones. The calculated AAa clearly resembles most to the AAG-type measured structure (RMSD = 16°, and the rest are above 27°). The calculated (only alanine and glycine containing) AAG model is close to the AAG-type X-ray structure (RMSD = 17°) (containing other amino acids as well), but it is that close to the measured POG and PPG structures as well (RMSD = 15° and 17°, respectively). Therefore regarding the structure the AAa triplet forms nearly as good triple helix as the AAG triplet.

The helix formed from the SaSaG triplet is closest to the PPG measured structure with an RMSD value of 22°, but it is still far enough. The GGG model resembles as much to the POG as to the PPG measured structure, although it is quite far even from them (with an RMSD of 28°). The AAA model forms the least folded triple helix, as it is very far from any of the measured structures (RMSD = 51° and 54°). Interestingly it is farthest from the POA bulge, that it is supposed to model.

The rest of **Table 8** is about the differences of the calculated triple helices. The calculated POG helix is (not surprisingly) closest to the PPG and PPa helices (RMSD = 11° and 14°). The PPG is close to the PPa helix (RMSD = 7°), and, interestingly, to the AAG helix (RMSD = 12°). This AAG helix is close to the PPa helix (RMSD = 14°), and not so far away from the POG model (RMSD = 20°). The SaSaG helix is the closest to the GGG and POG structures (RMSD = 21° and 23°), respectively, but these can only be regarded as similar structures, they are still quite different from each other. Not surprisingly, the GGG helix is thus closest to the calculated POG and SaSaG structures (both RMSDs are 21°). Although the calculated AAa helix seemed to be close to the AAG-type X-ray structure, it is quite different from all of the calculated helices, it resembles most the calculated PPa and AAG helices (both RMSDs are 27°). The furthest from all the other structures is the AAA helix, as its minimal RMSD is 38° from the GGG helix.

The average dihedral angles of the calculated collagen helices and the calculated β-sheets are shown in **Table 9**. The φ angles of the triple helices all lie in the range of –60° > x > –100°. The POG, PPG, PPa, AAG and AAa triplet containing triple helices have their φ value around –68°. The GGG and SaSaG triple helices have a φ angle about 10° lower (~ –79°). The AAA structure differs mostly, as its φ is ~ –93°. It is not surprising, as a non-glycine L-amino acid at the glycine position is known to disrupt the triple helix. The φ angle is not affected much from the amino acid substitution; it varies from 150° to 170° for all structures.

The average dihedral angles according to the amino acid position of the measured and calculated structures are further detailed in **STable 1**. The most important data here is that the φ angle of the AAA model at the glycine position is as high as –117°, which is not surprising as this is the place where the substitution affects the structure most.

Therefore both the RMSD (**Table 8**) and the dihedral angle values (**Table 9**) indicate, that the POG, PPG, PPa and AAG triplet containing calculated structures are close to their experimental counterparts. The AAa, SaSaG and GGG triplet containing triple helices are not so tightly bound, they are a bit unfolded. The AAA triplet can never take up the triple helical structure (as it was expected), because alanine in the place of glycine elongates the chains from each other. Looking at the Ramachandran-maps it can be seen that all amino acids have a part of ε_L region that has relatively low energy. That is why all these triplets (except for AAA) can eventually adopt the collagen triple helix structure.

Table 9. Structural data of the different collagen and β-pleated sheet models

(XYZ-XYZ)$_3$		Structure [a]	
		φ	ψ
Tropocollagen triple-helix	**POG**	–70.7 ± 13.8	163.3 ± 16.7
	PPG	–67.6 ± 7.9	161.5 ± 11.3
	PPa	–67.1 ± 8.0	158.8 ± 14.8
	SaSaG	–78.7 ± 3.3	169.4 ± 10.4
	AAG	–66.6 ± 10.6	159.0 ± 12.7
	AAa	–69.5 ± 5.4	149.9 ± 12.0
	GGG	–79.9 ± 18.7	167.0 ± 24.2
	AAA	–92.9 ± 24.7	158.1 ± 19.6
Parallel 3-stranded β-pleated sheet	**POG**	–98.4 ± 33.8	147.0 ± 30.1
	PPG	–89.9 ± 23.9	147.9 ± 33.9
	PPa	–93.1 ± 40.2	148.9 ± 46.5
	SaSaG	–109.0 ± 25.0	130.2 ± 49.6
	AAG	–143.4 ± 26.7	150.8 ± 20.9
	AAa	–151.2 ± 46.5	157.4 ± 31.9
	GGG	–137.4 ± 9.5	147.4 ± 8.4
	AAA	–134.4 ± 10.9	135.8 ± 14.1

a, Average backbone conformational parameters and standard deviation, all in degree

3.3.1.1 β-sheets, individual strands

The disintegration of the triple-stranded collagen helix into three "individual" and isolated extended-like polypeptide chains can be performed and the resulted polypeptide conformers can be used as energy reference structures.(**Figure 11**) This is useful also because these formation energies are the most relevant when compared with melting point measurement data.[57]

Depending on the amino acid sequence of the model, a somewhat different elongated backbone conformer is obtained, when fully optimized. These isolated polypeptide chains form β-strands, with every amino acid residues adopting a β_L-like local conformation for all the GGG, AAA and AAG triplet containing models. Strands formed by AAa triplets, systems incorporating D-Ala(s) and L-Ala(s) have also an elongated or $(\beta_L)_n$ type main chain fold. The average torsion angles associated with L-Ala are around $\phi = -160^\circ$, $\psi = +168^\circ$, while those of D-Ala are close to their mirror image ($\phi = +160^\circ$, $\psi = -168^\circ$). The conformation of the isolated polypeptide chains formed by POG, PPG, PPa and also by SaSaG are peculiar, since imino acids, such as proline, hydroxyproline and sarcosine can only adopt ε_L- or γ_L-type elongated local conformers,[91,92] (**Figure 3** and **Figure 15**) with ϕ around –70° (and the mirror

image, ϕ around 70˚ for sarcosine). Furthermore, in the POG strands, various H-bonds are formed between the OH group of hydroxyprolines and the CO groups of some other residues.

The backbone folds of these isolated single strands show a clear amino acid sequence dependence presenting the first steps on different oligomerization paths.

All β-sheets, as the collagen triple helices, consist of a total of 18 amino acid residues; 6 residues per strand. The GGG, AAA and AAG sheet models are regular triple-stranded parallel β-pleated sheets. Their average ϕ and ψ torsion angles vary to some extent with their amino acid composition, –135±11° and +147±8°, –134±11° and +136±14°, and –113±27° and +121±21°, respectively, all forming the usual interstrand S12 type H-bonding pattern[36]. The average H-bond lengths are again very similar, namely 1.99Å, 1.98Å and 2.00Å, respectively. Although the POG, PPG and PPa triplet containing β-strand models are composed of three parallel extended-like backbone structures, except for glycine they do not adopt β_L-type local conformations, also loosing the characteristic S12 H-bonding pattern too. Indeed Gly is the only residue which adopts β_L-type backbone conformation both in the PPG sheet and in the POG sheet models. The D-alanine in the PPa model adopts β_L or γ_D conformations (with torsion angles e.g. $\phi = +174°$ and $\psi = –186°$). In the AAa triplet containing three-stranded β-sheet model the "unnatural" configuration of the methyl groups of the D-alanine inhibits the formation of a conventional β-strand. Nevertheless, the three polypeptide chains stay close enough to consider their assembly as a proper reference structure, so much more, as the S12 H-bonding pattern is maintained, with 2.03Å average H-bond length.

In spite of their structural variability they all can be used as structures modeling the β-pleated sheet. However, the overall structural properties of these β-sheet models signal that the higher the imino acid content of the system is the more the β-sheet structure will be distorted.

Isolated strands are not restricted at all, and the conformational preferences of the amino acids can be observed the most here. These preferences can also be examined (although at a much lesser extent) at the other two secondary structures.

3.3.2 Stability of the calculated structures

The stability of the collagen triple helix with various amino acid constitutions is summarized in **Table 10** and **Table 11** (compared to the parallel β-sheet). The stability of these secondary structures with respect to the three isolated strands are shown in **STable 2** and **STable 3**. Values are scaled for a triplet (three amino acid residues) first, as this is the unit of the triple helix. Furthermore, concerning the crystal calculations, these results can be compared with the previous ones.

Energy calculations were carried out at the B3LYP/6-311++G(d,p)//B3LYP/6-31G(d) level of theory in vacuum and at the B3LYP/PCM/6-31G(d)//B3LYP/6-31G(d) level of theory in aqueous phase. Generally, the relative energy order of secondary structure elements do not change upon switching level of theory, however the absolute values are the smallest for PCM and largest for B3LYP/6-31G(d) results. (**Table 10**, **Table 11**, **STable 2** and **STable 3**)

The collagen triple helix stability compared to three individual strands ($\Delta E^2_{\text{collagen-form.}}$, **Figure 11**) as function of the primary sequence is: AAA<PPa<AAa<AAG<GGG <POG<PPG<SaSaG, **(STable 2** and **STable 3)** where the formation is favored for the GGG, AAG, POG, PPG and SaSaG triplets. The latter tendency is in agreement with general expectations based on experimental melting point data of various triplets embedded in a $(POG)_3$-XYG-$(POG)_4$ sequence.[57] According to Ackerman *et al*.,[57] the thermodynamic stability of a triple helix can be well-characterized by its melting temperature (T_m), which is the temperature where the triple helix ↔ individual chains transition occurs. Melting temperatures can be as low as 29°C for a model where the AAG sequence is embedded. The same T_m is 43°C when there is a PPG triplet at the centre of the POG oligopeptide. The melting temperature rises to 45°C for a POG triplet. As a comparison, the melting temperature of human type I collagen is around 36°C.[93] The G → A mutation is known to cause a destabilization in a POG sequence,[53] therefore we can assume the same for the AAG sequence. In conclusion, stability of the experimental models grow in the AAA → AAG → PPG → POG direction, a trend based on the increase of melting temperatures. (There is no available experimental data for the GGG, AAA and AAa-triplet collagen models.) As for the theoretical models the stability order is the following: AAA → AAG → POG → PPG.

The similarity between computed and measured stability data are very good, there is a slight discrepancy concerning the stability order of "POG" and "PPG" type models. (**STable 2**) That is why we have carried out frequency calculations, to try to find out whether it is the caused

by the fact that we have calculated only energies instead of Gibbs free energies. $\Delta G^2_{\text{collagen-form}}$ give the quite different GGG<PPa<AAA<AAG<PPG<POG<AAa<SaSaG order. The POG triplet has become more favored now as the PPG triplet, reproducing the measured stability order. However, the placement of the AAa triplet is quite unexpected. So much the more, because PPa became much less stable than the AAa triplet, and PPa is supposed to form a more stable triple helix than AAa, as the first have helix formation promoter prolines in it, and the second does not. Therefore these calculations do not provide the expected order neither, meaning that there are still more things (e.g. solvation effects) to be considered when trying to obtain the experimental stability order results with calculation.

Secondary structural preference or stability of collagen triple helix with respect to β-sheets as function of their primary sequence is reported in **Table 10** and **Table 11**. For the non-imino acid containing models, AAA, GGG, AAa and AAG, the triple-stranded parallel β-pleated sheet structure is stable over the collagen-like triple helical structure both with respect to energy and with respect to Gibbs free energy results. For these four models, by using the PCM solvent model, the relative energies are as follows: $\Delta E^{\text{B3LYP/PCM/6-31G(d)//B3LYP/6-31G(d)}}$ = +6.4, +3.8, +4.3 and +4.7 kcal·mol^{-1}, respectively.(**Table 10**) The most stable β-sheet is formed by the "alanine only" model, AAA, at all levels of theory.

Conformation selection is reversed for the other four (SaSaG, PPa, PPG- and POG) models. In fact, the collagen triple helix of the PPG model becomes more stable by 4.8 kcal·mol^{-1} in vacuum and by 3.8 kcal·mol^{-1} in water. (**Table 10**) The POG model in its triple helical form is more stable than the POG-sheet model by 3.4 kcal·mol^{-1} in vacuum and 2.0 kcal·mol^{-1} in water. Regarding the Gibbs free energy data, it is interesting that again the PPG model is more stable than the POG triplet. The SaSaG triplet has around the same amount of Gibbs free energy difference between the two secondary structures as the POG (**Table 11**), although regarding the energy data it is less stable. For the PPa triplet the triple helix formation is only slightly preferred regarding both the simple energy and the Gibbs free energy data.

Table 10. Stability of the triple helical structures compared to β-pleated sheet models Stability differences or $\Delta E_{conversion}$ between the triple-stranded β-sheet and the collagen triple helix models (both containing 18 amino residues) as a function of their amino acid composition, calculated at different levels of theory

	Energy Differences Between Secondary Structures (Collagen triple helix *vs.* β-sheet, $\Delta E_{conversion}$/kcal·mol^{-1} [a] per triplet [b]) (The same for the complete molecular system):			**Structural Differences [c] Between Secondary Structures (collagen triple helix *vs.* β-sheet) (Δζ/°):**
Type of Model	B3LYP/ 6-31G(d)	B3LYP/6-311++G(d,p)// B3LYP/6-31G(d)	B3LYP/PCM/ 6-31G(d)// B3LYP/6-31G(d)	B3LYP/6-31G(d)
POG	–4.3 (–25.9)	–3.4 (–20.4)	–2.0 (–11.9)	39
PPG	–5.3 (–31.6)	–4.8 (–28.6)	–3.8 (–22.9)	34
PPa	–2.5 (–15.1)	–0.9 (–5.4)	–0.3 (–1.6)	43
SaSaG	–2.7 (–16.3)	–2.5 (–15.1)	–2.9 (–17.4)	52
AAG	+7.1 (+42.9)	+6.8 (+40.8)	+4.7 (+28.4)	59
AAa	+6.2 (+37.5)	+6.0 (+35.9)	+4.3 (+26.2)	69
GGG	+6.3 (+38.1)	+6.3 (+37.6)	+3.8 (+22.6)	51
AAA	+8.8 (+53.0)	+8.6 (+51.6)	+6.4 (+38.5)	38

a, $\Delta E_{conversion}$: $E_{triple\ helix}$–E_{sheet} (**Figure 3**)

b, The formation energy is divided by the number of triplets: ($E_{triple\ helix}$–E_{sheet})/6

c, Structural differences (Δζ/°) are the quadratic mean of the differences of the relevant dihedral angles. For example, the ϕ angle of P in the OPGOPG strand of the β-sheet of POG models is applicable to the ϕ angle of P in the OPGOPG strand of the triple helix structure.

Table 11. Stability of the triple helical structures compared to β-pleated sheet models

Type of model	Stability[a]				
	ΔE[b]	ΔU[c]	ΔH	$T\Delta S$[d]	ΔG
POG	-23.9	-24.9	-24.9	-10.5	–14.4
PPG	-29.9	-30.6	-30.6	-9.1	-21.5
PPa	-12.9	-13.8	-13.8	-10.4	-3.4
SaSaG	-16.5	-16.4	-16.4	-3.4	-13.0
AAG	41.6	41.1	41.1	-7.7	48.8
AAa	36.1	36.6	36.6	-2.1	38.7
GGG	38.1	37.6	37.6	-6.1	43.6
AAA	50.1	50.5	50.5	-2.2	52.6

a, All values (in kcal·mol^{-1}) are relative to the appropriate extended like and isolated N- and C- protected hexapeptide (see method)
b, Electronic energy and Zero-Point Vibrational energy
c, Electronic energy, Zero-Point Vibrational energy, vibrational, rotational and translational energy
d, T = 298.15 K

The structural differences between the collagen triple helix and β-sheet are the smallest for the POG, PPG, PPa and AAA models, namely 39°, 34°, 43° and 38° respectively. For the SaSaG and GGG $\Delta\zeta$ is higher and the largest values are for the AAG- and AAa- models. A small value of $\Delta\zeta$ indicates that either the β-strand or the collagen helix secondary structure is not ideal because of one or more of the component amino acid residues have a locally distorted backbone structure. On the contrary a larger value signals that the amino acid composition is equally suitable for both types of secondary structures. Therefore the AAa, AAG, GGG and even SaSaG triplets could form both type of secondary structures, while POG, PPG and PPa have a strong preference for the triple helix, and AAA prefers the β-strand.

In total, the calculated stability order regarding the energy data of the collagen triple helices with respect to the parallel three stranded β-sheet is as follows:

AAA<AAG<AAa≤GGG<PPa<SaSaG<POG<PPG. Note that the relative stability order of AAa and AAG is reversed here relative to the order established previously when individual or isolated strands were used as the reference state. Interestingly, the formation of the GGG triple helix seems slightly more favorable than that of the AAG with respect to both reference states.

Regarding the Gibbs free energy the stability order changes to AAA<AAG<GGG<AAa<PPa<SaSaG<POG<PPG. The first four have positive $\Delta G^3_{conversion}$, therefore they would form β-sheet instead of triple helix. The PPa sequence containing triple helix is only slightly more stable than its sheet form, it would not make a good collagen. However, the formation of a SaSaG triple helix is favored both with respect to individual strands and with respect to a sheet structure, therefore it would be interesting to test it experimentally. The formation of a triple helix from sheets is preferred both for the PPG and POG triplets.

Also, the primary sequence dependent secondary structure preference obtained by *ab initio* calculations also verify that (L)-Ala instead of "Gly" destabilizes the collagen triple helix, and only with prolines (or hydroxylprolines) in the Xxx and Yyy position is the triple helical fold more stable than the β-sheet.

Although potential energy surfaces do not have a minimum at the ε_L region, half of that region has quite low energy (the exact amount depending on the amino acid residue), that is why amino acids can easily be incorporated into a collagen triple helix. Only D-alanine (that has a Ramachandran-map that is central symmetric to the map of alanine) has an ε_L region representing a higher energy, that is why this amino acid distorts slightly more the structure, as it can be observed most particularly on the AAa model.

Let us summarize the results obtained so far. From crystal calculations it is clear, that all amino acids that have side chains prefer the sheet structure in an aggregate. Polyglycine in a crystal mostly prefers a hexagonal arrangement than any sheet structure. However, reducing the number of chains to three (having collagen triple helix and triple stranded β-sheet), even for glycine chains the sheet becomes more stable. So what is needed to stabilize a triple helix? In crystal polyglycine (polyglycine II) every H-bonding capacity is fulfilled, therefore if we look at a triplet, it makes six H-bonds. In a triple helix the number of H-bonds is reduced to two per a capacity of six, therefore only one third of them are fulfilled. (In the sheet it grows to four per six.) Applying sarcosines instead of glycines where the position enables it (2 from

3 residues) the H-bonding capacities are reduced by two, as sarcosines do not have amide hydrogens. So the occupied/capacity ratio changes to 2/4. This ratio remains the same for the PPG and POG triplets as well. As it can be seen from collagen crystal structures (e.g. 1V7H), this remaining two hydrogen-bonding capacity is fulfilled by bonding water molecules for the POG and PPG triplets. It could be done for the SaSaG triplet as well, so there should be another explanation for why nature did not use sarcosine. There are two possible reasons. The first is the possibility of misfolding. A sheet structure made of SaSaG triplets is more stable than three individual strands, therefore it can be a folding trap. The second is, it might occur that collagen does not need to be as stable as a triple helix from SaSaG is. The C-terminal region of the protein strands helps the triple helix to fold, and tropocollagen is always further assembled into fibrils, therefore the individual stability of a triple helix might not be so important. Also, since collagen is a protein that sometimes needs to be quickly dismantled (e.g. when healing wounds), the individual stability of the triple helix does not have to be too high.

3.4 The first hydration layer of collagen

3.4.1 Reference bulk water molecules

The binding energies of the water molecules in the reference state turned out to be around –4.3±0.2 kcal·mol^{-1} per one hydrogen bond (**Table 12**). The central water molecule has different number of surrounding or neighboring water molecules, thus establishing hydrogen bonds of different numbers and types. In all cases the geometry of the water system is maintained as it was in a periodic ice. The energy difference between a water molecule attached to the surface of tropocollagen as an element of a given water-bridge (*e.g.* α–, β–, γ– or δ–) is to be compared to the stability of a molecule embedded among additional water molecules (**SFigure 1**). Note that, at present, this stability data differences are to be considered as semi-quantitative values due to several uncertainties described above.

Table 12. The binding energies of a central water surrounded by two, three or four hydrogen bonded neighbors all in a tetrahedral ice (**SFigure 3**) as calculated at the B3LY/6-31G(d) level of theory

Number of H-bonded Neighbors	**Binding Energy[a] (per H-bond)**
2	–9.3[b] (–4.6)
3	–12.6[c] (–4.2)
4	–16.8[d] (–4.2)
Average / H-bond	**–4.3±0.2**

a, kcal·mol^{-1}
b, the total energy of the system of 3 water molecules (without counterpoise correction) is –229.242838048 Hartree
c, the total energy of the system of 4 water molecules (without counterpoise correction) is –305.657568759 Hartree
d, the total energy of the system of 5 water molecules (without counterpoise correction) is –382.075133279 Hartree

3.4.2 Topology of the water threads

We have discovered that in the 1V7H[63] the above described four different types of water-bridges consist of two different sets of left-handed spirals wound around tropocollagen. The first one, (**Figure 18a** and **Figure 19**) which consists of the α– and β–bridges, has a shallow slope and makes one full turn around tropocollagen per triplet. The second one comprises α–, γ– and δ–water bridges and has a steeper ascend and makes one full turn only per 4 triplets. (**Figure 18b** and **Figure 20**) To sum it up, in total there are five different threads of water spirals around a tropocollagen fiber that wrap the protein like a net. These two types of threads, horizontal and vertical are interconnected (**Figure 21**) and one water molecule can be part of more water bridges. For example the β^1 water (**Figure 6**) is an integrated element of both the β– and γ– bridges and, so, it is the element of both the horizontal and vertical threads.

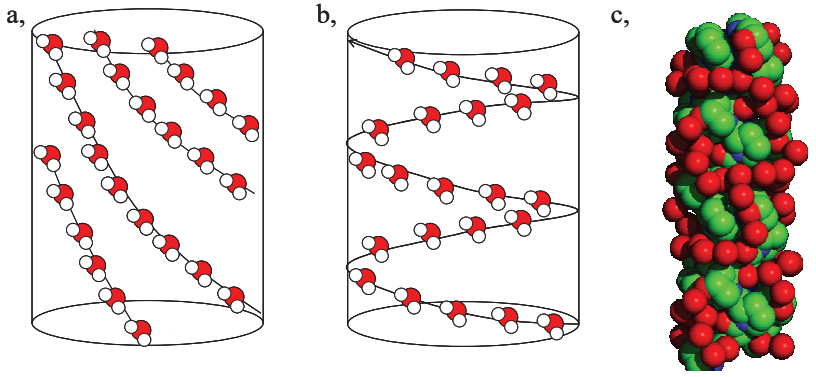


Figure 18. Schematic representation of **a,** the four parallel and quasi-horizontal as well as of **b,** the single and quasi-vertical water-threads forming the principal elements of the first hydration layer around tropocollagen. **c,** these five water-threads, as can be seen in the 1V7H X-ray structure[63]. The oxygens of the water molecules and of the tropocollagen helix are marked with red, the carbon atoms are green and the nitrogens are blue. No hydrogen atoms are shown.

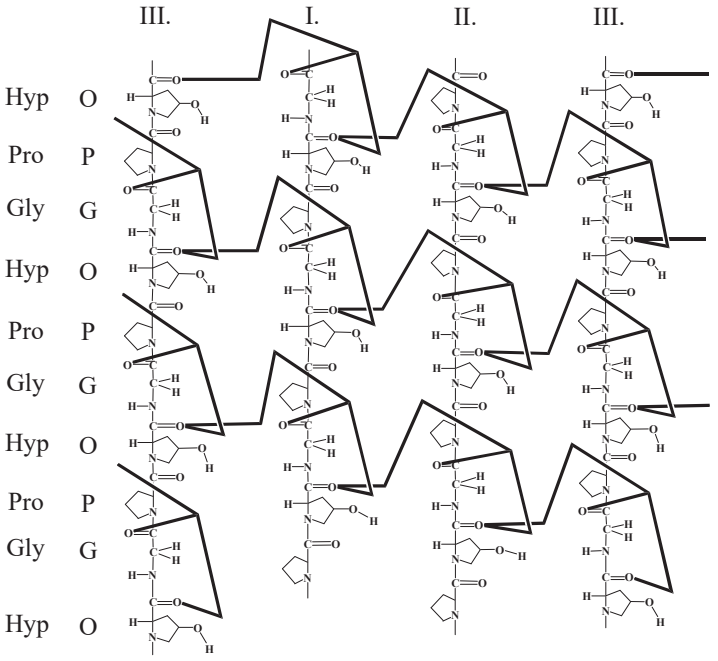


Figure 19. A schematic representation of the network of the horizontal water-thread. The water molecules are at the edges of the solid lines, otherwise representing a hydrogen-bond. This type of left-handed spiral water-thread does not depend on the amino acid composition of tropocollagen. However, if the residues are not the "normal" proline or hydroxyproline and glycine residues, the side chains may get into the way of these water-bridges (e.g. 1BKV), or even substitute one of the water molecules (1BKV)[58].

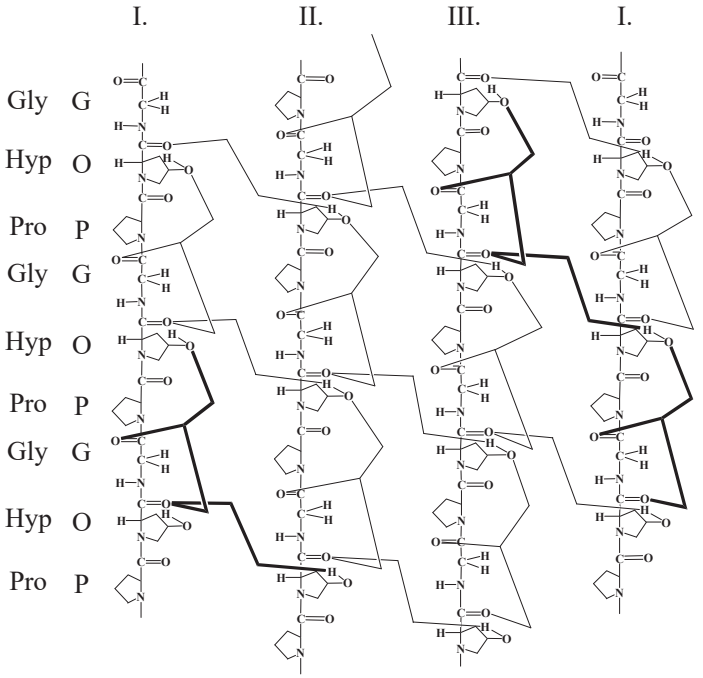


Figure 20. A schematic representation of the network of the vertical water-thread. (The water molecules are at the edges of the solid lines, otherwise representing a hydrogen-bond.) As significantly steeper than its horizontal counterpart, a total of four separated horizontal water-threads are needed to cover the tropocollagen completely.

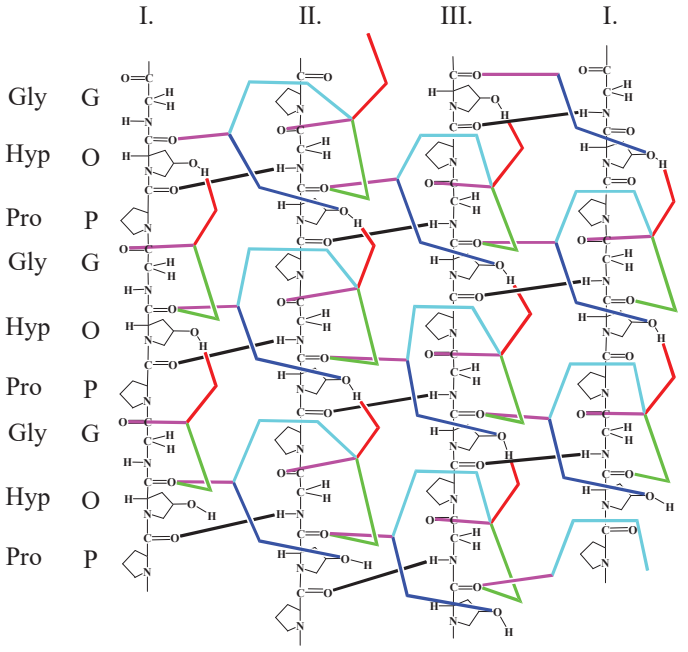


Figure 21. A schematic representation of the H_2O network of all the horizontal and the vertical water-threads. (The water molecules are at the edges of the solid lines, otherwise representing an H-bond.)

3.4.3 Calculated structural properties of the first hydration shell

The structural properties of the 3x6 amino acid containing "collagen base" model is to be compared to the waterless or "dry" collagen part made up of POG triplets[94] and to that of the X-ray structure 1V7H[63], which contains multiple layers of structural waters, too. The backbone geometrical properties of the *ab initio* optimized hydrated model systems are closer to the X-ray structure which contains explicit waters around collagen, than to that of the waterless or "dry" structural model. In fact, the RMSD of the average dihedral angles are around 4° smaller when compared to 1V7H[63], rather then to the vacuum optimized tropocollagen model. (*e.g.* RMSD of the POG-model with 2α- and 3β-waters with respect to the "dry"-model is 20°, while for the X-ray structure it is 16°.) (For RMSD calculations see reference 94.) Larger than 20° dihedral angle shifts are only present at the C- and N-termini of the model, a clear sign of the well-known "capping-effect", typical of any finite model system.

The structure comparison of the X-ray determined (measured) and of the *ab initio* calculated tropocollagen models, with respect to the connecting hydrogen-bond parameters are shown (**Table 13** and **Table 14).**

Table 13. Hydrogen-bond distances between collagen atoms and water molecules of the first hydration layer in the AAG-type collagen parts and in the AAG model with 6 water molecules (first type of model system) (**Figure 12**), at the B3LYP/6-31G(d) level of theory.

Type of H-bond	***r*(O…O) (Å)**	
	Measured: X-ray (standard deviation)	Calculated [b] (standard deviation)
Gly C=O...O	3.02 (0.21) [a]	2.82 (0.01)
Xxx NH...O	2.85 (0.17) [a]	2.89 (0.01)

a, Data are from the 1BKV PDB structure[58]
b, calculated values are taken from the AAG model system, see **Figure 12** and **Table 4**.

For the first two types of model systems (AAG-type water and α- and β-bridges) the calculated hydrogen-bond lengths are slightly shorter than those retrieved from the solid state measurements: *e.g.* for the AAG-type system $r^{measured}$= 3.02±0.21Å > $r^{calc.}$= 2.82±0.02Å. Note that the standard deviations of the experimentally determined H-bond distances for the internal waters are relatively large. The standard deviances of the measured data are by a

magnitude larger than the values obtained from calculations, indicating the inherent homogeneity of the calculated geometries (**Table 13** and **Table 14**).

In the third type of model system (**Table 15**) the situation is reversed; the measured H-bond distances are slightly shorter than the calculated ones, *e.g.* $r^{measured}$= 2.79±0.08Å < $r^{calc.}$= 2.87Å. However, the above mentioned structural differences are minor. In general, for all three model types the calculated and the experimentally determined r(O…O) values are rather close to each other: $\Delta r < 0.15$Å.

We cannot compare the fluorine atom containing models with any experimental data, as to the best of our knowledge there is no such information available. However, the RMSD of the flouroproline containing model from both the "dry" POG model and the hydrated X-ray structure also stays beneath 20°. It should also be noted that the Pro-F…HOH distances are slightly longer than that of the hydroxyproline containing models. On the other hand, there is no significant change for this distance as a function of the number of water molecules (2.96 Å and 2.93 Å for 2 and 3 water molecules in the γ-bridge, respectively).

In conclusion, both in terms of backbone conformation and hydrogen-bond parameters, the structural properties of the 18 residue containing tropocollagen models are good representatives of the X-ray measured structures (**Table 13, Table 14** and **Table 15**). Therefore, the associated stability data could be treated with confidence.

Table 14. Hydrogen-bond distances between the appropriate collagen atoms and water molecules for the α- and β-water bridges of the horizontal water thread (second type of model system), at the B3LYP/6-31G(d) level of theory.

Type of H-bond	*r*(O…O) (Å)				
	Measured: X-ray (standard deviation)	Calculated[b]			
		2α 3β	2α 4β	3α 3β	3α 4β
Gly C=O...$\alpha^{1\alpha}$	2.77 (0.05) [a]	2.70	2.70	2.69	2.69
Gly C=O...$\alpha^{1\beta}$	2.77 (0.05) [a]	2.67	2.73	2.74	2.74
Hyp C=O...α^{2}	2.79 (0.07) [a]	2.73	2.70	2.81	2.79
Hyp C=O...β^{1}	2.84 (0.10) [a]	2.69	2.71	2.73	2.70

a, Data are from the 1V7H PDB structure[63]
b, calculated O...O values were taken from the second type of model systems, see **Figure 13** and **Table 4**. As there are only one set of calculated distances, there is no standard deviation attached to them.

Table 15. Hydrogen-bond distances between the appropriate collagen atoms and water molecules for the γ- and δ-water bridges of the vertical water thread (third type model system), at the B3LYP/6-31G(d) level of theory.

Type of H-bond	*r*(O…O) (Å)				
	Measured: X-ray (standard deviation)	Calculated[b]			
		X=OH, 2γ	X=OH, 3γ	X=F, 2γ	X=F, 3γ
Pro-X...δ[1]	2.72 (0.12) [a]	2.70	2.82	2.96	2.93
Pro-X...γ[1]	2.79 (0.08) [a]	2.87	2.86	2.98	2.97

a, Data are from the 1V7H PDB structure[63]
b, calculated O...O values were taken from the third type of model systems, see **Figure 14** and **Table 4**. As there are only one such calculated distances, there is no standard deviation attached to them.

3.4.4 Stability properties of the internal and of the first hydration shell

3.4.4.1 Internal water molecules in tropocollagen

Recall that the AAG-model system stands for those collagen subunits where binding or enzymatic cleavage occurs in nature. The gradual analysis of the water binding energies of the six internal waters present in the AAG-model system (**Table 16**) shows that they do not differ so much from each other. In fact binding energies of the waters located either at the N- or at the C-terminus of the model system are very similar to those in the middle. The average binding energy is –4.9 kcal·mol^{-1} per hydrogen-bond, if all six molecules are taken into account. This average stabilization energy increases slightly to –5.1 kcal·mol^{-1} when the four central water molecules are considered only, with a halved standard deviation value (0.4 kcal·mol^{-1} → 0.2 kcal·mol^{-1}).

Table 16. Binding energies of water molecules in the AAG-model – with 6 bound molecules. As our model is 6 amino acid long, there are six binding sites. Each of these water molecules connect the amide NH of residue Xxx with the C=O of Gly (**Figure 7**). Water molecules are numbered along the model system, starting from the N terminus.

N° of water molecule (total No. of participating hydrogen-bonds)	**Binding energy[a] (per hydrogen-bond)**
1st (3)	–13.8 (–4.6)[b]
2nd (3)	–15.5 (–5.2)
3rd (3)	–14.7 (–4.9)
4th (3)	–15.2 (–5.1)
5th (3)	–15.8 (–5.3)
6th (3)	–12.9 (–4.3)
Averages	–14.6±1.1 (–4.9±0.4)
Average of the central 4 water molecules only	–15.3±0.5 (–5.1±0.2)

a, kcal·mol^{-1}, calculated at the B3LY/631G(d) level of theory by the counterpoise method.
b, the total energy of the system (without correction) is −5184.37342967 Hartree.

Thus, by removing structural distortions introduced by the capping effect, the average stabilization energy of the embedded waters seems constant: –5.1±0.2 kcal·mol^{-1}. Since this stabilization energy is somewhat higher than that of the cubic ice (–4.3±0.2 kcal·mol^{-1} per hydrogen bond), from a thermodynamic point of view the formation of these internal type hydrogen-bonds are preferred. The homogeneity of the above water-binding energies shows that this model is principally correct: it predicts uniform binding energies for all binding sites along the triple-helix. Note that the standard deviation of this water-binding energy is as low as 0.4 kcal·mol^{-1}, even if all six places are taken into account (**Table 16**). Based on the success of this model system, for all the following α–, β–, γ– and δ–bridges we have generated and optimized only a single suitable water-bridge and we have concentrated on placing the examined water molecules in the middle of a water "patch" and also in the middle of the triple helix structure.

3.4.4.2 The α– and β– type water bridges (the quasi-vertical water-thread of tropocollagen)

The water position specific H-bonding energies of both the α- and β-bridges formed by different number of water molecules are reported in **Table 17**, while **Figure 22** is designed to help to understand the connection between the model systems.

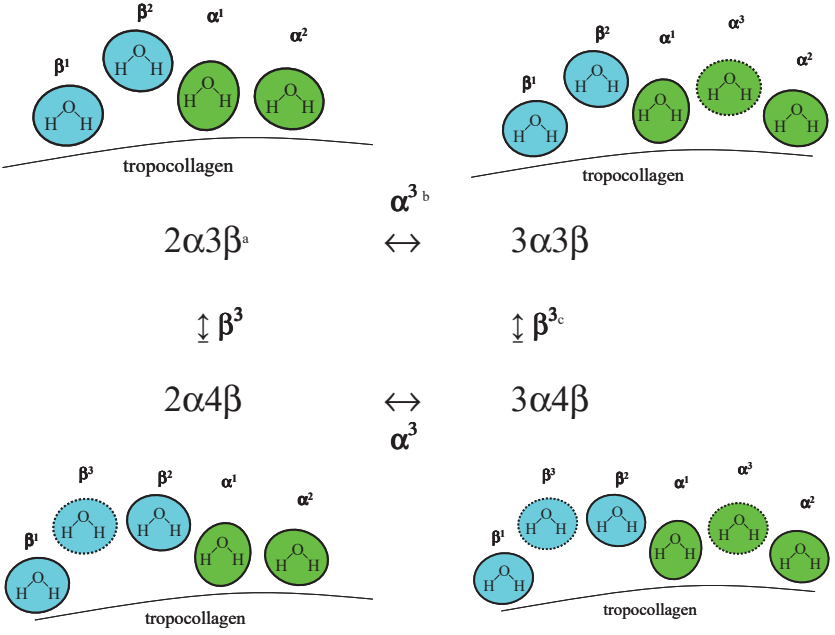


Figure 22. The explicit water content in the horizontal type of model systems, focusing on the α- and β-bridges. The α^1 water molecule is considered as a part of the α- and the β-bridge as well.(**Figure 6**)
a, 2α3β stands for the POG-model system, composed of 18 amino acid residues with 2 α and 3 β water molecules in focus (**Figure 13**)
b, α^3 stands for the optional α-type water molecule of the α-bridge (**Figure 13**)
c, β^3 stands for the optional β-type water molecule of the β-bridge (**Figure 13**)

For all sites the water binding energy is higher than it would be in a cubic ice –4.3±0.2 kcal·mol^{-1}/ hydrogen-bond (**Table 12**), thus, hydration is thermodynamically preferred. For all the other α- and β-type water binding places (*e.g.* $\alpha^{1\alpha}$, α^2,.., β^1, β^2,..) the normalized water binding energies change with the number of incorporated waters. For example, upon inserting the β^3 water while having two water molecules in the α–bridge, [(2α3β) → (2α4β)], the normalized water binding energy at the β^2 position becomes very different, namely in increases from –4.4 to –6.7 kcal·mol^{-1} (**Table 17**). The insertion of the β^1 water has similar tendencies: its stability also increases from –5.5 to –6.1 kcal·mol^{-1} upon the expansion of the β-bridge. This stability increase is clearly associated with the appearance of another water molecule in the β^3 position, also detected for the 3α3β and 3α4β systems.

Table 17. The position specific binding energy of waters located at the horizontal water thread (second type of model system) (**Figure 13**)

		Binding energy of the waters in the α–β combined bridges [a] (the same normalized for H-bonds)			
Water-bridge type	**Type of water (No. of H-bonds** [f]**)**	**2α3β** [b]	**2α4β** [c]	**3α3β** [d]	**3α4β** [e]
α	$\alpha^{1\alpha}$ (4)	–31.3 (–7.8)	–30.5 (–7.6)	–32.8 (–8.2)	–32.6 (–8.2)
	α^{2} (3 or 4)	–21.8 (–7.3)	–27.1 (–6.8)	–26.9 (–9.0)	–28.1 (–9.4)
	α^{3} (3) (optional H_2O)	---	---	–26.0 (–8.7)	–26.7 (–8.9)
minimal α bridge: (α^1+α^2)		–53.1 (–7.6)	–57.6 (–7.2)	–59.7 (–8.5)	–60.7 (–8.7)
β	β^{1} (3 or 4)	–16.5 (–5.5)	–24.2 (–6.1)	–22.1 (–5.5)	–28.3 (–7.1)
	β^{2} (2)	–8.7 (–4.4)	–13.4 (–6.7)	–10.0 (–5.0)	–14.4 (–7.2)
	β^{3} (2) (optional H_2O)	---	–12.3 (–6.1)	---	–12.6 (–6.3)
	$\alpha^{1\beta}$ (4)	–28.4 (–7.1)	–29.2 (–7.3)	–29.2 (–7.3)	–28.6 (–7.1)
minimal β-bridge: (β^1+β^2+$\alpha^{1\beta}$)		–53.6 (–6.0)	–66.8 (–6.7)	–61.3 (–6.1)	–71.3 (–7.1)
minimal combined α- and β-bridge: (α^1+α^2+β^1+β^2+$\alpha^{1\beta}$)		–106.7 (–6.7)	–124.4 (–6.9)	–121.0 (–7.1)	–132.0 (–7.8)

a, kcal·mol^{-1}, calculated at the B3LY/631G(d) level of theory by the counterpoise method.
b, the total energy of the system (without correction) is –7023.14841613 Hartree
c, the total energy of the system (without correction) is –7099.60677187 Hartree
d, the total energy of the system (without correction) is –7099.61128846 Hartree
e, the total energy of the system (without correction) is –7176.03985369 Hartree
f, total number of H-bonds formed by a particular water molecule

The insertion of the extra water molecule into the β-bridge at the β^3 position increases the normalized stability of the entire (β^1+β^2 +$\alpha^{1\beta}$) water bridge as well: $\Delta E(2\alpha3\beta)$= –6.0 kcal·mol^{-1} > $\Delta E(2\alpha4\beta)$= –6.7 kcal·mol^{-1}, respectively (**Figure 22** and **Table 17**). Furthermore, with a water molecule at β^3 position the binding energy of both of the "essential" β^1 and β^2 waters rises: $\Delta E(\beta^2)_{2\alpha3\beta}$= –4.4 kcal.mol^{-1} > $\Delta E(\beta^2)_{2\alpha4\beta}$= –6.7 kcal·mol^{-1} and $\Delta E(\beta^1)_{2\alpha3\beta}$= –5.5 kcal·mol^{-1} > $\Delta E(\beta^1)_{2\alpha4\beta}$= –6.1 kcal·mol^{-1}, respectively (**Table 17**).

On the contrary, the very same "surplus" β^3 water has no or only a slightly destabilizing effect on the normalized H-bonding energy of the waters forming the α-type water-bridge. For example, at the $\alpha^{1\alpha}$ positions no significant changes are detected: $\Delta E(\alpha^1)_{2\alpha3\beta} \sim \Delta E(\alpha^1)_{2\alpha4\beta} \sim -7.7\pm0.1$ kcal·mol^{-1} while $\Delta E(\alpha^1)_{3\alpha3\beta} = \Delta E(\alpha^1)_{3\alpha4\beta} \sim -8.7$ kcal·mol^{-1}, respectively. Unlike for α^1, the insertion of the β^3 water decreases more dramatically the normalized binding energy of the α^2 water: $\Delta E(\alpha^2)_{2\alpha3\beta} = -7.3$ kcal·mol^{-1} < $\Delta E(\alpha^2)_{2\alpha4\beta} = -6.8$ kcal·mol^{-1}, respectively (**Table 17**). Although the appearance of the β^3 water increases the total binding energy from –21.8 to –27.1 kcal·mol^{-1} (**Table 17**), the total number of H-bonds formed by the α^2 water also increases from 3 to 4. Thus, the H-bond normalized water binding energy decreases slightly at α^2 when the β-bridge is enlarged by β^3.

The normalized water binding energy of the $\alpha^{1\alpha}$- and $\alpha^{1\beta}$-type molecules, both of 4 H-bonded partners, is about the same in all the four models: ~–7.9±0.3 kcal·mol^{-1}, and ~–7.2±0.1 kcal·mol^{-1}, respectively. On the one hand, it is interesting that the stability data of both types of waters seems not to be affected by the total number of waters forming the α– or β–bridges. On the other hand, the actual difference in figures indicates that the $\alpha^{1\alpha}$ and $\alpha^{1\beta}$ positions are not quite equivalent. The same types of water binding sites have slightly different stabilities in the models. So we will use the H-bonding energy associated with the $\alpha^{1\alpha}$ position, as it is located in the middle of the hydration "patch", and so probably provides a better description of how tropocollagen is hydrated at these positions.

In general, the insertion of a single H_2O molecule either into the α– or into the β–bridge increases the stability of the overall system. For example, the incorporation of an H_2O at the β^3 water site results in a small normalized energy decrease: $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{2\alpha3\beta} = -6.7$ kcal·mol^{-1} > $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{2\alpha\mathbf{4}\beta} = -6.9$ kcal·mol^{-1} as well as $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{3\alpha\mathbf{3}\beta} = -7.1$ kcal·mol^{-1} > $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{3\alpha\mathbf{4}\beta} = -7.8$ kcal·mol^{-1}, respectively (**Table 17**). The insertion of an other H_2O molecule at the α^3 water site also results in a small stability increase: $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{2}\alpha3\beta} = -6.7$ kcal·mol^{-1} > $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{3}\alpha3\beta} = -7.1$ kcal·mol^{-1}, while for the four water molecule containing β-bridge the change is much more significant: $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{2}\alpha4\beta} = -6.9$ kcal·mol^{-1} > $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{3}\alpha4\beta} = -7.8$ kcal·mol^{-1} (**Table 17**). The "simultaneous" insertion of two water molecules, one at the α^3- and another one at the β^3-places, increases the normalized overall stability of the system: $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{2}\alpha\mathbf{3}\beta} = -6.7$ kcal·mol^{-1} >

$\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{3\alpha4\beta}}$= –7.8 kcal·mol^{-1}, respectively (**Table 17**). As a consequence, the binding energy of all component H-bonds ($\alpha^{1\alpha}$, α^2, β^1, β^2 and $\alpha^{1\beta}$) also increases (**Table 17**).

Finally, the swapping of a single water molecule from the α- to the β-bridge, H_2O "moves" from α^3-position to β^3-site, and *vice versa*, seems to be almost energy neutral: $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{2\alpha4\beta}}$= –6.9 kcal·mol^{-1} ≈ $\Delta E(\alpha^1+\alpha^2+\beta^1+\beta^2+\alpha^{1\beta})_{\mathbf{3\alpha3\beta}}$= –7.1 kcal·mol^{-1} (**Table 17**) allowing water to displace relatively freely. Even though in the latter site-exchange the stability of the overall system looks the same, the local water-binding energy of the component water sites can be rather different: $\Delta E(\alpha^2)_{\mathbf{2\alpha4\beta}}$= –6.8 kcal·mol^{-1} > $\Delta E(\alpha^2)_{\mathbf{3\alpha3\beta}}$= –9.0 kcal·mol^{-1} (**Table 17**).

3.4.4.3 The γ- and δ-type water bridges (the vertical water threads of tropocollagen)

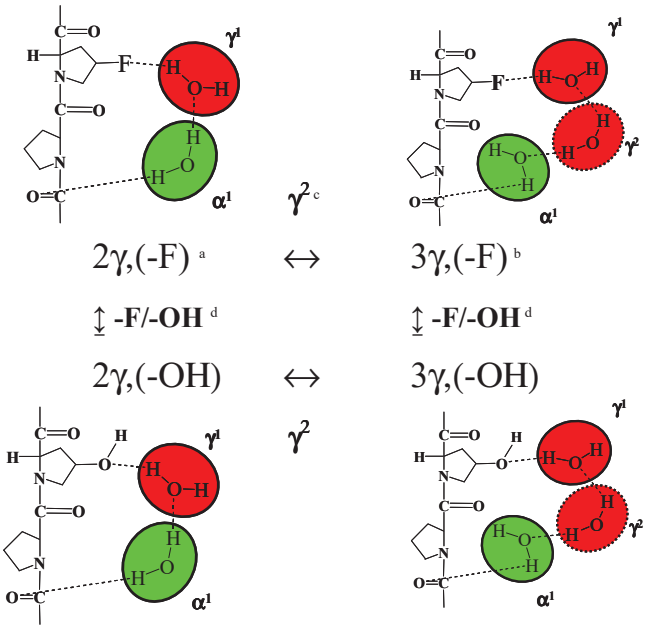


Figure 23. The explicit water content in the vertical type of model systems, focusing on the γ-bridges, and on the –OH ↔ –F substitution

a, 2γ,(-F) stands for the POG-model system, composed of 18 amino acid residues with 2γ-waters and with fluoroproline now in focus (**Figure 14**)

b, 3γ,(-F) stands for the POG-model system, composed of 18 amino acid residues with 3γ-waters and with fluoroproline now in focus (**Figure 14**)

c, γ^2 stands for the optional water molecule of the γ-bridge (**Figure 14**)

d, -F/-OH stands for the substitution of the hydroxyproline in focus to flouroproline and vice versa.

Table 18. The position specific binding energy of waters located at the vertical water thread (third type of model system) (**Figure 13**)

		Binding energy of the water molecules in the γ-δ combined bridges [a,b] (the same normalized for H-bonds)			
		Pro–OH		**Pro-F**	
Water-bridge type	**Type of water[b] (Number of H-bonds [g])**	**2γ** [c]	**3γ** [d]	**2γ** [e]	**3γ** [f]
γ	**α^1 (3)**	–19.5 (–6.5)	–23.1 (–7.7)	–22.1 (–7.4)	–23.7 (–7.9)
	γ^1 (2)	–12.6 (–6.3)	–13.7 (–6.8)	–6.9 (–3.5)	–11.9 (–6.0)
	γ^2 (2) (optional H_2O)	---	–17.3 (–8.7)	---	–19.4 (–9.7)
minimal γ-bridge: $\alpha^1+\gamma^1$		–32.1 (–6.4)	–36.8 (–7.4)	–29.0 (–5.8)	–35.7 (–7.1)
δ	**δ^1 (3)**	–20.4 (–6.8)	–23.2 (–7.7)	–18.5 (–6.2)	–19.2 (–6.4)
	β^1 (2)	–13.2 (–6.6)	–19.8 (–9.9)	–15.5 (–7.8)	–15.3 (–7.6)
δ-bridge: $\delta^1+\beta^1$		–33.6 (–6.7)	–43.0 (–8.6)	–34.0 (–6.8)	–34.5 (–6.9)
minimal combined γ- and δ-bridge: $\alpha^1+\gamma^1+\delta^1+\beta^1$		–65.7 (–6.6)	–79.7 (–8.0)	–63.0 (–6.3)	–70.2 (–7.0)

a, kcal·mol^{-1}, calculated at the B3LY/6-31G(d) level of theory by the counterpoise method.
b, please note that one water molecule of the γ-bridge is called α^1, and one molecule of the δ-bridge is called β^1
c, the total energy of the system (without correction) is –6717.39110930 Hartree
d, the total energy of the system (without correction) is –6793.82194994 Hartree
e, the total energy of the system (without correction) is –6741.41161108 Hartree
f, the total energy of the system (without correction) is –6817.84202998 Hartree
g, total number of H-bonds associated with a particular water molecule

The insertion of a single water molecule into the γ-bridge (γ^2) increases greatly the overall H-bonding energies of the system: $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{2}\gamma,OH}$= –6.6 kcal·mol^{-1} > $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{3}\gamma,OH}$= –8.0 kcal·mol^{-1}. The energy lowering can be traced for every water-binding places, however, the greatest change is detected for the β^1 place. $\Delta E(\beta^1)_{\mathbf{2}\gamma,OH}$= –6.6 kcal·mol^{-1} > $\Delta E(\beta^1)_{\mathbf{3}\gamma,OH}$= –9.9 kcal·mol^{-1}. Interestingly, this water-binding position (β^1) is not even at the same bridge as is the water molecule in focus (γ^2) that gives a hint to how much extent the binding of these molecules is interconnected. (**Figure 23** and **Table 18**) Also, the binding energy of the γ^2 molecule itself is quite high, $\Delta E(\gamma^2)_{\mathbf{2}\gamma,OH}$= –8.7 kcal·mol^{-1}.

The insertion of the same molecule (γ^2) into the fluoroproline containing model strengthens the overall binding of the others as well: $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{2}\gamma,F}$= –6.3 kcal·mol^{-1} > $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{3}\gamma,F}$ = –7.0 kcal·mol^{-1}. Looking at the other water molecules individually, it can be seen that the binding of the H_2Os in the same bridge (γ^1, δ^1), and even one in the other bridge (δ^1) is stronger. However, for the (β^1) place, where the highest change was observed for the hydroxyproline containing model, in the fluoroproline containing model the situation is reversed. Here the addition of the extra water molecule into the γ-bridge (γ^2) eventually decreases the binding energy, even if only slightly. $\Delta E(\beta^1)_{\mathbf{2}\gamma,F}$= –7.8 kcal·mol^{-1} < $\Delta E(\beta^1)_{\mathbf{3}\gamma,F}$= –7.6 kcal·mol^{-1}.

The analyses of the γ- and δ-bridges (**Table 18**) with respect to the –OH- or –F content shows that those water molecules that are directly connected to fluorine atom (γ^1 and δ^1) have a lower binding energy with respect to those which are associated with the hydroxyl group of hydroxyproline. For example, the normalized water binding energy at γ^1 position for the fluorinated compound is only $\Delta E(\gamma^1)_{\mathbf{2}\gamma,F}$= –3.5 kcal·mol^{-1}, while for the OH containing partner the same stability value is $\Delta E(\gamma^1)_{\mathbf{2}\gamma,OH}$= –6.3 kcal·mol^{-1}. This difference holds not only for the two- but also for the three-water containing models: $\Delta E(\gamma^1)_{\mathbf{3}\gamma,F}$= –6.0 kcal·mol^{-1} > $\Delta E(\gamma^1)_{\mathbf{3}\gamma,OH}$= –6.8 kcal·mol^{-1}.

Although the other two water molecules (α^1 and β^1) are slightly more stable in the fluorinated molecule, the sum of the binding energies shows that the molecules are altogether slightly less bound in the case of the fluorinated collagen. ($\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{2}\gamma,F}$= –6.3 kcal·mol^{-1} > $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{2}\gamma,OH}$= –6.6 kcal·mol^{-1} and $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{3}\gamma,F}$= –7.0 kcal·mol^{-1} > $\Delta E(\alpha^1+\gamma^1+\delta^1+\beta^1)_{\mathbf{3}\gamma,OH}$= –8.0 kcal·mol^{-1}.

Table 19 summarizes the H-bonding energies of the characteristic water molecules in the first hydration shell of collagen. This is to be compared directly (second column) with the binding energy of a water that can be found in bulk water instead of being located at the surface of tropocollagen. For the purpose of direct comparison the reference bulk water has the same number of neighbors as assigned on the surface of tropocollagen.

Table 19. Summary of the binding energies of water molecules per hydrogen bond in all the model systems

Binding site (number of neighbors)	Average binding energies on the appropriate site of tropocollagen[b]	Average binding energies inside or on the surface of bulk water[b]	Scaled stability (the difference of the H-bonding energies of the two places) [b]
α_t^1 (4)[a]	–7.6	–4.2	–3.4
α^2 (3-4) [a]	–8.1	–4.2	–3.9
α^3 (3)	–8.8	–4.2	–4.8
β^1 (3-4) [a]	–6.1	–4.2	–1.9
β^2 (2)	–5.8	–4.7	–1.1
β^3 (2)	–6.2	–4.7	–1.5
δ^1 (3) [a]	–7.3	–4.2	–3.1
γ^1 (2) [a]	–6.6	–4.7	–1.9
γ^2 (2)	–8.7	–4.7	–4.0
ζ (3) [a]	–5.1	–4.7	–0.4

a, only these water molecules are directly attached to tropocollagen, all the other H_2Os are attached *via* these ones

b, kcal·mol^{-1}, calculated at the B3LY/6-31G(d) level of theory by the counterpoise method (eq.14).

The stability order of the different water binding places is as follows: (α^3), (γ^2), α^2, α^1, δ^1, γ^1, β^1, (β^3), (β^2), ζ (or internal) (without calibration). α^3 is the strongest while β^2 is the weakest water binding site on the POG-type collagen. The ζ (or internal) binding site is weaker than any of the sites of the POG-type collagen. This "preference" of the above water binding sites does not change significantly after the subtraction of the relevant water-embedded energies (third column, **Table 19**). Only the $\gamma^1$and β^1 binding sites become equally "attractive".

The most strongly bound water molecules are α^3 and γ^2, E^{HB}= –8.8 kcal·mol^{-1} and E^{HB}= –8.7 kcal·mol^{-1}, respectively. This is rather interesting since neither of them are directly bound to the surface of tropocollagen (**Figure 6**), but rather to other structural waters. The strongest "direct" H-bonds formed between tropocollagen and water are associated with the α^2 and α^1 places: E^{HB}= –8.1 kcal·mol^{-1} and E^{HB}= –7.6 kcal·mol^{-1}, respectively. (**Table 19**) In addition, δ^1 is also a relatively strong binding site for H_20 (E^{HB}= –7.3 kcal·mol^{-1}). Interestingly, β^2 is the second least strongly bound water, even though it can always be found in the X-ray structures. This β^2 type water is also attached to tropocollagen only *via* other water molecules (**Figure 6**). The simplest network and the least strongly bound water molecules are those called as the "internal" or ζ-type water molecules, associated with the natural enzymatic recognition sites of tropocollagen.

It has to be noted, that although the binding energy values suggest here that water molecules are rather strongly bound to the surface of tropocollagen, these are only energy values. As seen for the backbone stabilities, entropy contributions can change the situation. Therefore these binding site preferences can only be regarded as indications.

Each of the polypeptide chains of collagen forms a left-handed screw, but the triple helix formed from these turns out to be a right-handed supramolecular complex. As Orgel *et al.*[25] have described a filament built up from seven tropocollagen units evolves as a right-handed screw. Therefore, the two types of left-handed water threads, the four quasi-horizontals and the single vertical ones, described here nicely fits into the "gap" between the right-handed triple helical tropocollagen and the collagen filament. (**Figure 24**) This counter-twist is what provides the stability and attachment between strands for twisted ropes, it is interesting to observe the same thing for a "molecular rope". Also, it further underlines the necessity of a counter-twisted layer between the two zones that apparently have the same helicity.

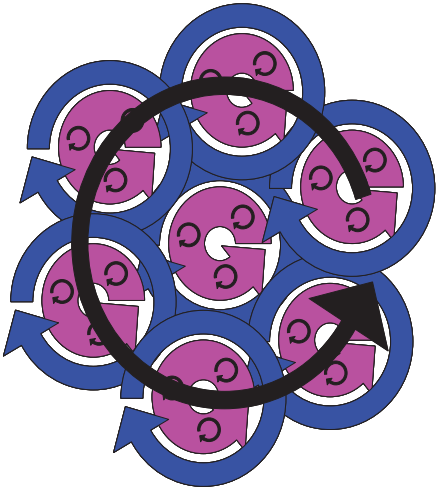


Figure 24. Schematic representation of a collagen filament that forms a right-handed screw[25] (large black circular arrow). The filament is made up of seven tropocollagen triple helices (pink circular arrows, forms a right handed helix). The tropocollagen consists of three polypeptide chains, (small black circular arrows, left handed) and surrounded by its hydration shell (blue circular arrow, left handed).

Raines *et al.*[48] have stated that the collagen triple helix is more stable when fluorinated with respect to the natural OH group containing Hyp. (These measurements were carried out in two solvents: 50 mM acetic acid, which stabilizes triple helices by protonating the C-terminal carboxylate groups and thereby eliminating unfavorable Coulombic interactions, and phosphate-buffered saline (PBS), which mimics a physiological environment.) In both solvents the fluoroproline containing triple helix was the most stable. However, the question arises that for what reason nature did not use fluorine atom instead of OH groups. The answer might be that the stability of the tropocollagen triple helix in itself is also important, but it is more important to maintain a strong hydration shell around the tropocollagen as further on they are connected by the water threads. Therefore, it is apparently useful to have OH groups containing collagen, as the hydration shell is slightly weaker around the fluorinated collagen. (see **Table 18**)

In summary we can say that by the analysis of X-ray data we have discovered two types of water threads typical of collagen built from POG-type triplets that are both forming a left-handed helix. We have designed suitable *ab initio* models to explore the stability of water

molecules in the above hydration network. We have found that the OH→F substitution destabilizes in some extent the binding of the water bridges around tropocollagen.

The function of the extra water molecules in the α, β and γ-bridges can be various. First, their places might serve as "water-hole-conducting" places, meaning that these places can be used by water molecules, when they are flowing between tropocollagen triple helices. That can be reason why Henkelman *et al.* have found that the flow of water molecules in collagen is fluid-like.[95] Second, these "extra" water molecules can serve as molecular "buffers": at high water concentration molecules occupy these places, whereas in case of low water concentration these places might be left vacant. In this case, of course, the remaining water molecules arrange themselves to have the best H-bonding contacts. Therefore, collagen can act as a sponge: take up and store water, and release it if necessary at a relatively low energy cost, without deteriorating the global fold of the macromolecule. However, it seems that there is a minimum required number of water molecules (4-5) that are necessary to obtain the regular fold.[72] Therefore the "sponge" can only work above a certain water content. This nicely corresponds to the observation that in the case of a POG structure there are five places where water molecules can bind directly to the tropocollagen molecule.

4 Summary

Protein oligomers and aggregates can have multiple roles. Amyloid and similar type of aggregates are formed without strict self-regulation and are associated with several illnesses. On the other hand, associated multichain nanosystems formed by different polypeptide chains such as collagen are vital in cell and tissue formation. Furthermore, the latter type of nano-associates, formed in a strictly controlled manner such as collagen fibers can be produced or dismantled according to the needs of the living organism.

Plaque formation from different peptides or proteins is the cause of many illnesses[6]. Also, several other proteins were found[5] to form plaques under the proper cellular conditions. These plaques have the same macroscopical forms, and are therefore thought to possess the same microscopical structure: the β-pleated sheet. This leads to the question of why are these β-pleated sheet structures so much preferred. To answer this question we have proposed a deduction scheme and for the confirmation of it we have designed and accomplished suitable theoretical calculations on periodical model systems. This research appears to be the first investigation that carried out periodical calculations on peptides.

The results show that for a glycine containing peptide the most stable form is a two-dimensional superstructure that was already described by Crick and Rich[90], where the residues have the ε_L local backbone conformation. The collagen forming residues have the same conformation as these glycine residues. The alanine containing peptide models, however, prefer the form of β-pleated sheet, as these types of structures are more stable by ~10 kcal·mol^{-1} than any of the others. As the side chains do not let the peptide chains get close enough to each other, the hydrogen bonds get weaker, and thus the stability decreases, resulting in the presented energetics of the structures. Consequently, the above quantum-chemical calculations have shown that the deduction holds and that polypeptides having side chains can only align in the form of β-sheets when closely packed. That is why all proteins or peptides that are allowed to adopt a closely packed structure form amyloid-like fibrils.

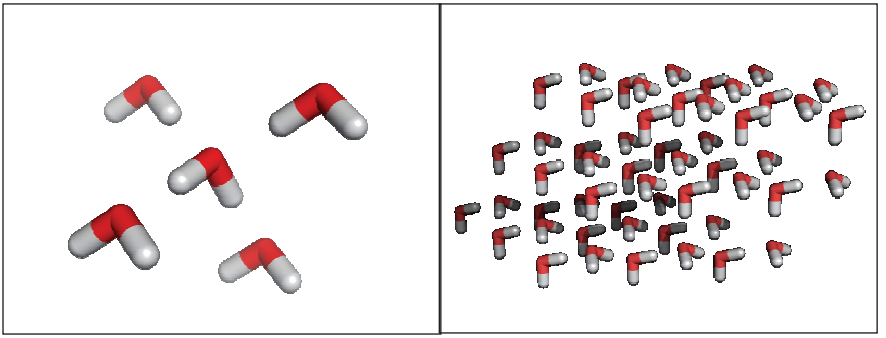
The conformation that is the most stable for glycine only [$(Gly)_n$] containing peptides is preserved in the triple-helical collagen structure. We have studied the effect of selected amino acids residues on the stability of the tropocollagen structure. For this reason we have carried out quantum chemical calculations and compared the stability of the collagen triple helices with those of the β-pleated sheets, the building unit of amyloid. In this way we could analyze the different contributing factors arising from different types of amino acids. For the

experimentally already studied amino acids these theoretical comparisons provided the same results as experimental ones about which amino acids or triplets strengthen collagen formation. Sarcosine, an amino acid that is abundant in our body but is not among the encoded 20 residues, helps to form a triple helix apparently stronger than even those built from proline or hydroxyproline. Therefore sarcosine containing artificially synthesized collagens can have future in medical or cosmetical applications.

From our results a simple hypothetical deduction emerges that describes that collagen stabilizing amino acids are needed to reduce the number of H-bonding capacities that are oriented away from the triple helix. This way the importance of bound water molecules is put into a different perspective, as they form hydrogen bonds with the remaining carbonyl groups that are still oriented away from the triple helix.

We have determined the individual binding strength of the water molecules forming the first hydration layer, to the best of our knowledge for the first time in the literature. According to our calculations a synergistic effect can be observed between the bound waters that are interconnected with strong hydrogen bonds. Furthermore, these water molecules around collagen form threads that wrap it as a net. Tropocollagens are protected by these water nets and are also first contacted to each other through them. Therefore, some biological or medical applications (for example for decreasing the effects of osteoporosis imperfecta) could be based in the future on the maintenance of a strong hydration network around the triple helices.

5 Appendix



SFigure 1. Water molecules in a crystal, as calculated at the B3LYP/6-31G(d) level of theory.

STable 1. The average backbone values (ϕ and ψ) of the different collagen models, as measured in the solid state and calculated for the gas phase at the B3LYP/6-31G(d) level of theory. Also, for each calculated structure the difference from one experimentally measured X-ray structure is shown.

Sequence Position		Xxx		Yyy		Gly	
Backbone Torsions		ϕ	ψ	ϕ	ψ	ϕ	ψ
		aver.[a] ± stdev.[b]	aver.[a] ± stdev.[b]	aver.[a] ± stdev.[b]	aver.[a] ± stdev.[b]	aver.[a] ± stdev.[b]	aver.[a] ± stdev.[b]
POG	X-ray[c]	–72.6 ± 1.6	163.7 ± 1.4	–57.4 ±1.7	151.8 ±1.4	–72.5 ±2.1	174.2 ±1.5
	calc.[d]	–76.1 ± 4.0	169.5 ± 7.4	–57.5 ±12.0	148.3 ±6.1	–78.5 ±13.4	169.7 ±21.6
	diff.[e]	–3.5	5.8	–0.1	–3.5	–6.0	–4.5
PPG	X-ray[f]	–74.5 ± 2.9	164.3 ± 4.1	–60.1 ± 3.6	152.4 ± 2.6	–71.7±3.7	175.9 ± 3.1
	calc.[d]	–71.1 ± 3.7	161.3 ± 7.0	–57.5 ± 3.0	150.3 ± 6.6	–74.0 ± 3.3	173.2 ± 4.7
	diff.[e]	3.4	–3.0	2.6	–2.1	–2.3	–2.7
PPa	calc.[d]	–72.3 ± 2.4	159.8 ± 6.6	–57.3 ± 1.5	144.5 ± 7.6	–71.7 ± 2.0	175.0 ± 3.4
	diff.[e]	0.3	–3.9	0.1	–7.3	0.8	0.8
SaSaG	calc.[d]	–80.3 ± 1.7	175.0 ± 5.9	–75.8 ± 2.6	158.2 ± 7.9	–80.0 ± 3.8	176.0 ± 4.4
	diff.[e]	–5.8	10.7	–15.7	5.8	–8.3	0.1
AAG	X-ray[g]	–69.9 ± 6.1	155.4 ± 8.2	–66.8 ± 4.8	147.8 ± 3.2	–67.7 ± 4.5	166.7 ± 5.6
	calc.[d]	–62.3 ± 4.4	155.6 ± 7.0	–58.6 ± 5.9	148.8 ± 10.0	–79.0 ± 6.7	172.6 ± 6.8
	diff.[e]	7.6	0.2	8.2	1.0	–11.3	5.9
AAa	calc.[d]	–65.8 ± 3.3	149.8 ± 3.9	–75.8 ± 3.6	136.8 ± 6.0	–67.1 ± 1.7	163.0 ± 4.7
	diff.[e]	4.1	–5.6	–8.9	–11.0	0.7	–3.8
GGG	calc.[d]	–78.0 ± 11.3	175.8 ± 25.5	–65.5 ± 1.5	150.1 ± 9.9	–96.2 ± 22.7	175.5 ± 28.7
	diff.[e]	–8.1	20.4	1.4	2.3	–28.5	8.8
AAA	X-ray[h]	–61.4 ± 4.1	158.6 ± 9.1	–57.0 ± 4.2	142.6 ± 6.3	–81.2 ± 19.2	152.8 ± 15.4
	calc.[d]	–83.8 ± 6.6	153.5 ± 12.5	–76.4 ± 3.6	164.2 ± 6.2	–117.1 ± 3.2	171.7 ± 5.0
	diff.[e]	22.5	5.1	19.5	–21.6	35.9	–18.9

a, Average values in degrees
b, Standard deviations in degrees
c, Okuyama et al.[63]
d, Due to the "capping effect", all values are the average of the four "middle" residues
e, Differences between measured and computed values, for the AAa and GGG models the reference is the 1BKV X-ray structure[58]
f, Hongo et al.[64]
g, Kramer et al.[58]
h, Bella et al.[53]

STable 2. Relative stability data of the different β-pleated sheet and collagen models compared to three individual strands, calculated at different levels of theory

		Energy Differences Between Secondary Structures and Three Individual Strands[a] (Δ*E*/kcal·mol^{-1} per triplet) [b] (for the whole system):		
	Type of Model	**B3LYP/6-31G(d)**	**B3LYP/6-311++G(d,p)// B3LYP/6-31G(d)**	**B3LYP/PCM/6-31G(d)// B3LYP/6-31G(d)**
Triple helix	**POG**	–7.5 (–45.2)	–4.8 (–28.7)	–3.0 (–18.3)
	PPG	–8.1 (–48.7)	–5.5 (–33.1)	–3.6 (–21.9)
	PPa	–3.5 (–21.2)	+0.4 (+2.3)	+1.7 (+10.0)
	SaSaG	–10.8 (–64.6)	–7.7 (–46.1)	–4.8 (–28.9)
	AAG	–5.9 (–35.4)	–3.1 (–18.7)	–1.0 (–5.8)
	AAa	–4.5 (–27.1)	–1.2 (–7.0)	+1.0 (+6.2)
	GGG	–6.7 (–40.1)	–3.3 (–19.8)	–1.1 (–6.6)
	AAA	–0.6 (–3.6)	+1.6 (+9.9)	+3.1 (+18.4)
Parallel 3-stranded β-pleated sheet	**POG**	–3.2 (–19.3)	–1.4 (–8.3)	–1.1 (–6.4)
	PPG	–2.9 (–17.1)	–0.8 (–4.6)	+0.2 (+1.0)
	PPa	–1.0 (–6.1)	+1.3 (+7.7)	+1.9 (+11.6)
	SaSaG	–8.1 (–48.3)	–5.2 (–31.0)	–1.9 (–11.5)
	AAG	–13.1 (–78.3)	–9.9 (–59.5)	–5.7 (–34.2)
	AAa	–15.8 (–94.8)	–12.0 (–72.1)	–6.9 (–41.1)
	GGG	–13.0 (–78.1)	–9.6 (–57.4)	–4.9 (–29.2)
	AAA	–13.6 (–81.7)	–10.6 (–63.9)	–6.5 (–39.2)

a, $\Delta E_{\text{formation}} = E_{\text{triple helix}} - [E_{\text{strand1}} + E_{\text{strand2}} + E_{\text{strand3}}]$ (**Figure 3**)
b, The formation energy is divided by the number of triplets: $(E_{\text{triple helix}} - [E_{\text{strand1}} + E_{\text{strand2}} + E_{\text{strand3}}])/6$

STable 3. Relative stability data of the different β-pleated sheet and collagen models compared to three individual strands

Type of model		Stability[a]				
		ΔE [b]	ΔU [c]	ΔH	$T\Delta S$ [d]	ΔG
Tropocollagen triple-helix	**POG**	–41.1	–14.9	–40.6	–39.3	–1.3
	PPG	–40.8	–40.1	–41.1	–40.9	0.3
	PPa	–15.8	–14.9	–16.0	–41.8	25.8
	SaSaG	–60.4	–59.5	–60.7	–44.8	–15.9
	AAG	–27.1	–27.8	–29.0	–48.4	19.5
	AAa	–51.0	–50.4	–51.6	–43.3	–8.2
	GGG	–27.2	–29.7	–30.9	–58.1	27.2
	AAA	-22.4	-21.8	-23.0	-42.7	19.8
Parallel 3-stranded β-pleated sheet	**POG**	–17.2	–14.5	–16.0	–28.8	13.1
	PPG	–10.9	–9.5	–32.0	–31.8	21.8
	PPa	–2.9	–1.1	–2.3	–31.5	29.2
	SaSaG	–43.9	–43.1	–44.3	–41.4	–2.9
	AAG	–68.7	–68.9	–70.1	–40.7	–29.3
	AAa	–87.1	–87.0	–88.2	–41.2	–46.9
	GGG	–65.3	–43.1	–68.5	–52.1	–16.4
	AAA	-72.4	-72.2	-73.4	-40.6	-32.9

a, All values (in kcal·mol^{-1}) are relative to the appropriate extended like and isolated N- and C- protected hexapeptide (see method)
b, Electronic energy and Zero-Point Vibrational energy
c, Electronic energy, Zero-Point Vibrational energy, vibrational, rotational and translational energy
d, T = 298.15 K,

6 References

(1) Anfinsen, C. B. *Nature* **1973**, *181*, 223-230.
(2) Nelson, R.; Sawaya, M. R.; Balbirnie, M.; Madsen, A. O.; Riekel, C.; Grothe, R.; Eisenberg, D. *Nature* **2005**, *435*, 773-778.
(3) Dobson, C. M. *Trends Biochem Sci* **1999**, *24*, 329-332.
(4) Fandrich, M.; Forge, V.; Buder, K.; Kittler, M.; Dobson, C. M.; Diekmann, S. *P Natl Acad Sci USA* **2003**, *100*, 15463-15468.
(5) Dobson, C. M. *Nat Struct Mol Biol* **2006**, *13*, 295-297.
(6) Dobson, C. M. *Nature* **2005**, *435*, 747-749.
(7) Taylor, J. P.; Hardy, J.; Fischbeck, K. H. *Science* **2002**, *296*, 1991-1995.
(8) Sunde, M.; Blake, C. *Advances in Protein Chemistry,* **1997**, *50*, 123-159.
(9) Nelson, R.; Eisenberg, D. *Curr Opin Struc Biol* **2006**, *16*, 260-265.
(10) Chan, J. C. C.; Oyler, N. A.; Yau, W. M.; Tycko, R. *Biochemistry-Us* **2005**, *44*, 10669-10680.
(11) Wasmer, C.; Soragni, A.; Sabate, R.; Lange, A.; Riek, R.; Meier, B. H. *Angew Chem Int Ed Engl* **2008**, *47*, 5839-5841.
(12) Wasmer, C.; Lange, A.; Van Melckebeke, H.; Siemer, A. B.; Riek, R.; Meier, B. H. *Science* **2008**, *319*, 1523-1526.
(13) Sawaya, M. R.; Sambashivan, S.; Nelson, R.; Ivanova, M. I.; Sievers, S. A.; Apostol, M. I.; Thompson, M. J.; Balbirnie, M.; Wiltzius, J. J. W.; McFarlane, H. T.; Madsen, A. O.; Riekel, C.; Eisenberg, D. *Nature* **2007**, *447*, 453-457.
(14) Eakin, C. M.; Berman, A. J.; Miranker, A. D. *Nat Struct Mol Biol* **2006**, *13*, 202-208.
(15) Jimenez, J. L.; Nettleton, E. J.; Bouchard, M.; Robinson, C. V.; Dobson, C. M.; Saibil, H. R. *P Natl Acad Sci USA* **2002**, *99*, 9196-9201.
(16) Guijarro, J. I.; Sunde, M.; Jones, J. A.; Campbell, I. D.; Dobson, C. M. *P Natl Acad Sci USA* **1998**, *95*, 4224-4228.
(17) Bucciantini, M.; Giannoni, E.; Chiti, F.; Baroni, F.; Formigli, L.; Zurdo, J. S.; Taddei, N.; Ramponi, G.; Dobson, C. M.; Stefani, M. *Nature* **2002**, *416*, 507-511.
(18) Fandrich, M.; Dobson, C. M. *Embo J* **2002**, *21*, 5682-5690.
(19) Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M. *Nature* **2003**, *424*, 805-808.
(20) Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. *P Natl Acad Sci USA* **2006**, *103*, 16623-16633.
(21) Wright, C. F.; Teichmann, S. A.; Clarke, J.; Dobson, C. M. *Nature* **2005**, *438*, 878-881.
(22) Otzen, D. E.; Kristensen, O.; Oliveberg, M. *P Natl Acad Sci USA* **2000**, *97*, 9907-9912.
(23) Richardson, J. S.; Richardson, D. C. *P Natl Acad Sci USA* **2002**, *99*, 2754-2759.
(24) Bhattacharjee, A.; Bansal, M. *IUBMB Life* **2005**, *57*, 161-172.
(25) Orgel, J. P. R. O.; Irving, T. C.; Miller, A.; Wess, T. J. *P Natl Acad Sci USA* **2006**, *103*, 9001-9005.
(26) Myllyharju, J.; Kivirikko, K. I. *Ann Med* **2001**, *33*, 7-21.
(27) Byers, P. H.; Wallis, G. A.; Willing, M. C. *J Med Genet* **1991**, *28*, 433-442.
(28) Fratzl, P. (ed.) *Collagen: Structure and Mechanics*; SPRINGER, 2008.
(29) Sorensen, L. T. *Hernia* **2006**, *10*, 456-461.
(30) Baum, J.; Brodsky, B. *Fold Des* **1997**, *2*, R53-R60.
(31) Buevich, A. V.; Baum, J. *Journal of the American Chemical Society* **2002**, *124*, 7156-7162.

(32) Liu, X. Y.; Kim, S.; Dai, Q. H.; Brodsky, B.; Baum, J. *Biochemistry-Us* **1998**, *37*, 15528-15533.
(33) Perczel, A.; Jakli, I.; Csizmadia, I. G. *Chem-Eur J* **2003**, *9*, 5332-5342.
(34) Perczel, A.; Angyan, J. G.; Kajtar, M.; Viviani, W.; Rivail, J. L.; Marcoccia, J. F.; Csizmadia, I. G. *Journal of the American Chemical Society* **1991**, *113*, 6256-6265.
(35) Fasman, G. D. *Current Science* **1990**, *59*, 839-845.
(36) Perczel, A.; Gaspari, Z.; Csizmadia, I. G. *J Comput Chem* **2005**, *26*, 1155-1168.
(37) Burjanadze, T. V. *Biofizika+* **1992**, *37*, 231-237.
(38) Brodsky, B. *P Indian as-Chem Sci* **1999**, *111*, 13-18.
(39) Bella, J.; Berman, H. M. *J Mol Biol* **1996**, *264*, 734-742.
(40) Tang, T. H.; Deretey, E.; Jensen, S. J. K.; Csizmadia, I. G. *Eur Phys J D* **2006**, *37*, 217-222.
(41) Schumacher, M.; Mizuno, K.; Bachinger, H. P. *Journal of Biological Chemistry* **2005**, *280*, 20397-20403.
(42) Doi, M.; Nishi, Y.; Uchiyama, S.; Nishiuchi, Y.; Nishio, H.; Nakazawa, T.; Ohkubo, T.; Kobayashi, Y. *Journal of Peptide Science* **2005**, *11*, 609-616.
(43) Persikov, A. V.; Ramshaw, J. A. M.; Brodsky, B. *Biopolymers* **2000**, *55*, 436-450.
(44) Improta, R.; Benzi, C.; Barone, V. *Journal of the American Chemical Society* **2001**, *123*, 12568-12577.
(45) Improta, R.; Mele, F.; Crescenzi, O.; Benzi, C.; Barone, V. *Journal of the American Chemical Society* **2002**, *124*, 7857-7865.
(46) Vitagliano, L.; Berisio, R.; Mastrangelo, A.; Mazzarella, L.; Zagari, A. *Protein Sci* **2001**, *10*, 2627-2632.
(47) Hongo, C.; Noguchi, K.; Okuyama, K.; Tanaka, Y.; Nishino, N. *J Biochem* **2005**, *138*, 135-144.
(48) Raines, R. T. *Protein Sci* **2006**, *15*, 1219-1225.
(49) Parthasarathi, R.; Madhan, B.; Subramanian, V.; Ramasami, T. *Theor Chem Acc* **2003**, *110*, 19-27.
(50) Tsai, M.; Xu, Y. J.; Dannenberg, J. J. *Journal of the American Chemical Society* **2005**, *127*, 14130-14131.
(51) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *Biochemistry-Us* **2000**, *39*, 14960-14967.
(52) Bhate, M.; Wang, X.; Baum, J.; Brodsky, B. *Biochemistry-Us* **2002**, *41*, 6539-6547.
(53) Bella, J.; Brodsky, B.; Berman, H. M. *Structure* **1995**, *3*, 893-906.
(54) Lazarev, Y. A.; Lazareva, A. V.; Khromova, T. B.; Grechishko, V. S. *Biofizika+* **1997**, *42*, 326-333.
(55) Feng, Y. B.; Melacini, G.; Taulane, J. P.; Goodman, M. *Journal of the American Chemical Society* **1996**, *118*, 10351-10358.
(56) Kishimoto, T.; Morihara, Y.; Osanai, M.; Ogata, S.; Kamitakahara, M.; Ohtsuki, C.; Tanihara, M. *Biopolymers* **2005**, *79*, 163-172.
(57) Ackerman, M. S.; Bhate, M.; Shenoy, N.; Beck, K.; Ramshaw, J. A. M.; Brodsky, B. *Biophys J* **1999**, *76*, A174-A174.
(58) Kramer, R. Z.; Bella, J.; Mayville, P.; Brodsky, B.; Berman, H. M. *Nat Struct Biol* **1999**, *6*, 454-457.
(59) Song, I. K.; Kang, Y. K. *Journal of Physical Chemistry B* **2006**, *110*, 1915-1927.
(60) Benzi, C.; Improta, R.; Scalmani, G.; Barone, V. *Journal of Computational Chemistry* **2002**, *23*, 341-350.
(61) Miles, C. A.; Avery, N. C.; Rodin, V. V.; Bailey, A. J. *J Mol Biol* **2005**, *346*, 551-556.

(62) Bailey, A. J.; Paul, R. G.; Knott, L. *Mechanisms of Ageing and Development* **1998**, *106*, 1-56.
(63) Okuyama, K.; Hongo, C.; Fukushima, R.; Wu, G. G.; Narita, H.; Noguchi, K.; Tanaka, Y.; Nishino, N. *Biopolymers* **2004**, *76*, 367-377.
(64) Hongo, C.; Nagarajan, V.; Noguchi, K.; Kamitori, S.; Okuyama, K.; Tanaka, Y.; Nishino, N. *Polym J* **2001**, *33*, 812-818.
(65) Berisio, R.; Vitagliano, L.; Mazzarella, L.; Zagari, A. *Protein Sci* **2002**, *11*, 262-270.
(66) Ramachandran, G. N.; Chandrasekharan R. *Biopolymers* **1968**, *6*, 1649
(67) De Simone, A.; Vitagliano, L.; Berisio, R. *Biochem Biophys Res Commun* **2008**, *372*, 121-125.
(68) Melacini, G.; Bonvin, A. M. J. J.; Goodman, M.; Boelens, R.; Kaptein, R. *Journal of Molecular Biology* **2000**, *300*, 1041-1048.
(69) Fullerton, G. D.; Nes, E.; Amurao, M.; Rahal, A.; Krasnosselskaia, L.; Cameron, I. *Cell Biology International* **2006**, *30*, 66-73.
(70) Fullerton, G. D.; Amurao, M. R. *Cell Biology International* **2006**, *30*, 56-65.
(71) Fullerton, G. D.; Rahal, A. *J Magn Reson Imaging* **2007**, *25*, 345-361.
(72) Boryskina, O. P.; Bolbukh, T. V.; Semenov, M. A.; Gasan, A. I.; Maleev, V. Y. *Journal of Molecular Structure* **2007**, *827*, 1-10.
(73) P. Echenique and J. L. Alonso, *Molecular Physics* **2007**, 105, 3057-3098.
(74) A Chemist's Guide to Density Functional Theory, Second Edition, Wolfram Koch, Max C. Holthausen, 2001, Wiley-VCH Verlag GmbH
(75) Hohenberg, P.; Kohn, W. *Phys. Rev B* **1964**, *136*, 864.
(76) Ordejon, P.; Artacho, E.; Soler, J. M. *Phys Rev B Condens Matter* **1996**, *53*, R10441-R10444.
(77) Gaussian 03, Revision C.02, M. J. Frisch G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, Gaussian, Inc., Wallingford CT, 2004.
(78) Jansen, H. B.; Ross, P. *Chem Phys Lett* **1969**, *3*, 140
(79) Liu, B.; McLean, A. D. *J Chem Phys* **1973**, *59* 4557
(80) Boys, S. F.; Bernardi, F. *Molecular Physics* **1970**, *19*, 553.
(81) Cances, E.; Mennucci, B.; Tomasi, J. *J Chem Phys* **1997**, *107*, 3032-3041.
(82) Perczel, A.; Hudaky, P.; Fuzery, A. K.; Csizmadia, I. G. *J Comput Chem* **2004**, *25*, 1084-1100.
(83) Simon, S.; Duran, M.; Dannenberg, J. J. *J Chem Phys* **1996**, *105*, 11024-11031.
(84) Geerke, D. P.; Thiel, S.; Thiel, W.; van Gunsteren, W. F. *Physical Chemistry Chemical Physics* **2008**, *10*, 297-302.
(85) Enriz, R.D.;Morales, Mirta.E.;Baldoni, H.A.; *Journal of Molecular Structure: THEOCHEM* **2005,** *731*, 177-185

(86) Baldoni, H.A.;Rodriguez, A.M.;Zamora, M.A.;Zamarbide, G.N.;Enriz, R.D.;Farkas, O.;Csaszar, P.;Torday, L.L.;Sosa, C.P.;Jakli, I.; *Journal of Molecular Structure: THEOCHEM* **1999**, *465*, 79-91
(87) Hudaky, I.;Baldoni, H.A.;Perczel, A.; *Journal of Molecular Structure: THEOCHEM* **1999**, *582*, 233-249
(88) Tran, T. T.; Treutlein, H.; Burgess, A. W. *Protein Eng Des Sel* **2006**, *19*, 401-408.
(89) Perczel, A.; Hudaky, P.; Palfi, V. K. *Journal of the American Chemical Society* **2007**, *129*, 14959-14965.
(90) Crick, F. H. C. & Rich, A. Structure of polyglycine II. *Nature* **176**, 780-781 (1955)
(91) Hudaky, I.; Baldoni, H. A.; Perczel, A. *Journal of Molecular Structure: THEOCHEM* **2002**, *582*, 233-249.
(92) Hudaky, I.; Perczel, A. *Journal of Molecular Structure: THEOCHEM* **2003**, *630*, 135-140.
(93) Leikina, E.; Mertts, M. V.; Kuznetsova, N.; Leikin, S. *P Natl Acad Sci USA* **2002**, *99*, 1314-1318.
(94) Palfi, V. K.; Perczel, A. *J. Comput Chem.* **2008**, *29*, 1374-1386
(95) Henkelman, R. M.; Stanisz, G. J.; Kim, J. K.; Bronskill, M. J. *Magnetic Resonance in Medicine* **1994**, *32*, 592-601.