Published in final edited form as:

Biopolymers. 2015 July ; 104(4): 360–370. doi:10.1002/bip.22644.

Empirical Estimation of Local Dielectric Constants: Toward Atomistic Design of Collagen Mimetic Peptides

Douglas H. Pike and Vikas Nanda*

Center for Advanced Biotechnology and Medicine, and the Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers University, Piscataway, New Jersey 08854

Abstract

One of the key challenges in modeling protein energetics is the treatment of solvent interactions. This is particularly important in the case of peptides, where much of the molecule is highly exposed to solvent due to its small size. In this study, we develop an empirical method for estimating the local dielectric constant based on an additive model of atomic polarizabilities. Calculated values match reported apparent dielectric constants for a series of *Staphylococcus aureus* nuclease mutants. Calculated constants are used to determine screening effects on Coulombic interactions and to determine solvation contributions based on a modified Generalized Born model. These terms are incorporated into the protein modeling platform protCAD, and benchmarked on a data set of collagen mimetic peptides for which experimentally determined stabilities are available. Computing local dielectric constants using atomistic protein models and the assumption of additive atomic polarizabilities is a rapid and potentially useful method for improving electrostatics and solvation calculations that can be applied in the computational design of peptides.

Phrases for indexing

collagen; electrostatics; dielectric; polarizability; solvation; protein design

INTRODUCTION

Peptides serve a number of biological and biotechnological roles as hormones, signaling molecules, antimicrobials, therapeutics and components of natural and synthetic biomaterials. They are of tremendous interest as therapeutics, inhabiting a pharmacologic ‘Goldilock’s zone’ between protein-based biologics and small molecules, combining favorable properties of large and small molecules. There is a need to develop molecular engineering methods for modulating peptide structure, stability and function. Most approaches begin with an existing biological structural template, followed by rational design or elaboration of combinatorial libraries of related sequences. The short lengths of target peptides normally allow for efficient exploration of related compounds, affording a reasonable chance of success. However, this approach is less effective in cases where

* nanda@cabm.rutgers.edu.

biological analogs are not available for mimicry of structure or function. In such instances, a *de novo* design approach is required. The design of new structures and functions has benefited significantly from advances in computational protein design, which allows rapid searches through sequence/structure space, evaluating fitness based on a combination of atom-level and residue-level energy functions.

Despite being composed of amino acids, peptides are not simply ‘little proteins’, and many successful computational approaches for protein design do not translate well to peptides. Linear peptides generally have minimal tertiary structure, providing few opportunities for engineering stabilizing long-range contacts. Related to this, peptides tend to be highly dynamic, adopting a broad ensemble of structures. This makes it challenging to estimate changes in free energy of folding caused by amino acid substitutions, which would ostensibly require energetic sampling of the substitution in the context of an ensemble of near-native structures.

Another consequence of peptide size is the high degree of amino acid solvent exposure. The net hydrophobicity of a protein correlates strongly with its chain length for lengths shorter than fifty amino acids¹. The inability of small chains to form compact globular structures results in minimal or non-existent hydrophobic cores. Even model mini-proteins, which exhibit two-state folding behavior such as the extensively characterized Trp-Cage^{2,3}, have minimal cores consisting of one to two amino acids. This complicates energetics calculations where estimates of protein-water interactions either require implicit solvation approximations or extensive conformational sampling in explicitly modeled waters.

A related problem is the molecular design of fibrous proteins such as collagen. Much like peptides – collagens are highly dynamic and their fold exposes most amino acid positions to solvent. Many groups are pursuing the design of collagen mimetic peptides (CMPs) as in vivo matrix labeling reagents⁴, as self-assembling biomaterials^{5–13}, and as models for collagen folding and disease^{14–18}. Although natural fibrillar collagens can be over 1000 amino acids long, properties relevant to molecular design are in several ways similar to much smaller peptides due to unique structural features of the collagen triple-helix. The triple-helix is composed of three chains, each adopting a polyproline-II type conformation with a repeating (Gly-Xxx-Yyy)_n sequence motif where the X and Y positions are frequently proline and hydroxyproline respectively. Unlike most protein folds, which primarily utilize hydrophobic residues for stabilizing tertiary contacts, the three chains in collagen are held together by a network of main chain hydrogen bonds between glycine amide protons and X-position carbonyls of the adjacent chain (Figure 1). Core glycines are required for the close approach of the three chains in order to form the network of hydrogen bonds. Side chains at X and Y positions project away from the triple-helix axis and are highly solvent exposed.

Existing molecular design tools cannot be effectively applied ‘out-of-the-box’ to the collagen triple-helix. Aforementioned challenges such as conformational dynamics, lack of a hydrophobic core and high degree of water accessibility must be addressed when developing rules for fibrous protein design. Much of the progress in this area is driven by an appreciation of naturally occurring collagens. The unusual structure of collagen results in a unique amino acid composition profile. Eliminating critical glycine and proline residues

required for maintaining a triple-helix fold, the remaining sequence is rich in charged and polar uncharged amino acids. Nonpolar and aromatic side chains make up less than 25% of the sequence (Figure 1). In contrast, a globular protein such as hemoglobin is over 40% nonpolar. Fibrillar α -helical proteins such as the extended tropomyosin coiled-coil dimer are intermediate in their composition, due to high solvent accessibility of the protein chains.

Both collagen and tropomyosin have a high frequency of charged amino acids, which make up half of the sequence of these fibrous proteins. In α -helical coiled-coils, a significant fraction of charged residues make interchain salt bridges across the helix-helix interface. In engineered α -helical oligomers, these networks can be used to control topology and heteromeric interactions – a strategy referred to as *molecular velcro*¹⁹. Similarly, X and Y positions in the (GXY)_n sequence of the triple-helix of collagen are able to form interchain electrostatic interactions. Networks of these interactions contribute to stability and specificity of collagen folding^{20–23}. Synthetic networks have been designed rationally and computationally resulting in charged-residue-rich polyelectrolytes^{20,24–28}.

The energy functions used for computational sequence design of CMPs have thus far been simplistic, and consist of tallying favorable and unfavorable charge pair interactions^{25,27,28}. These functions depend on a discrete sequence-based model where charge groups at Y-X and/or Y-X' (in G-X-Y-G-X'-Y' positions) positions are scored based on their propensity to interact (Figure 2). The linear structure of the triple-helix makes this a reasonable assumption and the simple sequence-level simulation protocols and scoring functions makes it possible to rapidly screen and sometimes fully explore all possible sequences of CMPs held together by charge-pair interactions. In particular, this approach has been used to design ABC-type CMP heterotrimers where three distinct peptide chains assemble specifically as a 1:1:1 complex^{24,27}.

While parameterized scoring functions can rationalize the stability of CMP sequences, they require a limited amino acid alphabet – i.e. designs are constructed solely of combinations of glycine, hydroxyproline, one basic amino acid type (Lys or Arg) and one acidic amino acid type (Asp or Glu). They also require prior knowledge of more detailed rules – i.e. Lys-Asp pairs are stronger than Arg-Glu interactions^{23,29}, and Y-X' interactions are stronger than Y-X interactions³⁰. Ideally, a design force field would specify these rules based on first principles.

Collagen stability has also been studied through extensive host-guest CMP experiments conducted by the Brodsky laboratory³¹. This information was integrated in the development of a sequence-level collagen stability predictor that took into account all 20 amino acids and their relative effects on stability at either X or Y positions in the triple-helix¹⁷. The scoring function also included terms for pairwise electrostatic interactions between chains. It performed well on an extensive set of studied CMPs including both *de novo* sequences and fragments of natural collagens. Other than a few special cases, it generally does not consider interchain pairwise interactions and has not been tested as a scoring function in computational design.

One strategy for getting around specialized design protocols for different classes of collagen and CMP molecules is to develop design rules that incorporate atomistic detail and can be generalizable to CMPs and perhaps even more broadly to most classes of peptides. Atomistic simulations of CMPs have used molecular dynamics (MD) software platforms to understanding the molecular basis of structure and stability. MD simulations have been used to recapitulate host-guest peptide studies³², understanding folding pathways³³ and to identify how interchain charge pairs stabilize the triple helix³⁴. While MD has provided valuable insight, it is not a feasible platform for design. Even the fastest MD simulation scenarios would require hours to days to effectively sample a single sequence. This would prove a significant bottleneck in a design protocol where anywhere from 10^5 to 10^{20} sequences may need to be considered to identify a reasonable set of targets. As such, atomistic design tools use a number of strategies to significantly reduce conformational sampling and the burden of energetics calculations such that large numbers of sequences can be rapidly screened.

One of the key challenges in protein design - one that is even more evident in peptides - is effective treatment of solute-solvent interactions. Explicit water models, such as those in MD simulations, are not compatible with most design algorithms, which constrain conformational degrees of freedom by focusing on relevant collective motions within the structure of the target. Instead, design tools rely on continuum solvation approaches, knowledge-based potentials, or hybrid treatments such as solvated-rotamers that focus on first shell waters surrounding surface-exposed polar amino acids^{35,36}. In the case of CMPs, nearly all sites in the chain are highly solvated³⁷. Networks of charge pairs that characterize both natural and *de novo* collagen sequences are significantly impacted by solvent, through dielectric screening of electrostatic forces, as well as solvation effects that consider the energetic consequences of changes in amino-acid water interactions in the native and unfolded reference state. In this work, we describe a rapid, empirical method for estimating the local dielectric on the surface of a peptide, toward improving both electrostatic force and solvation calculations in the context of a computational peptide design protocol. These tools have been incorporated into the protCAD molecular engineering software platform and can calculate the stability of a broad range of existing CMPs with experimentally characterized stabilities, and can reconstruct the conformations of networks of electrostatic and hydrophobic interactions on the surface of CMPs of known structure.

RESULTS AND DISCUSSION

protCAD development platform

The tools developed in this report have been implemented in the protCAD modeling environment. protCAD (protein Computer Aided Design) is a library of modules for manipulating protein structure and calculating energetics. protCAD was first developed in 2000 in William DeGrado's group by Christopher Summa³⁸. It has since served as a powerful platform for rapid prototyping of new molecular design methods and has been used in over twenty published molecular design projects. Applications include the design of template folds for metalloproteins³⁹⁻⁴⁵, development and testing of membrane protein design potentials⁴⁶⁻⁵³, and the incorporation of D-amino acids in peptides⁵⁴⁻⁵⁶. Current

versions of the source code are available for academic use (see Methods for details on accessing the software).

Rationale for estimating dielectric constants

We focused on estimating dielectric constants in specific protein environments as a strategy to model the extensive hydration in small peptides. The dielectric constant parameterizes the polarizability of the protein microenvironment, reflecting the ability of dipoles in the medium to respond to an applied electric field. High dielectric mediums such as water can significantly reduce the strength of ionic interactions such as salt-bridges and hydrogen bonds on the surfaces of proteins. The same interactions are significantly stronger in protein interiors where a low dielectric exists due in part to rigidity of the molecular environment and exclusion of water. This is reflected in the Coulombic component of the energy function:

$$E_{coul}(kcal/mol) = 332 \frac{q_1 q_2}{\epsilon r} \quad \text{eq. 1}$$

in which two charges q_1 and q_2 are separated by distance r in units of Ångstroms. Often, ϵ is estimated based on the type of environment, ten to twenty for a protein core and up to eighty for more solvated positions. Another approach is to use a distance-dependent dielectric constant where $\epsilon = \epsilon_s r^{57}$, which models the assumption that electrostatic interactions are only significant when atoms are in close proximity.

We developed an empirical approach that calculates a local dielectric constant for each atom based on the number and type of adjacent atoms within a specific cutoff distance. Electronic polarizabilities can vary based on atom type – i.e. an alkane carbon has a weaker response to an applied electric field than an aromatic carbon. Assuming a homogenous, non-interacting medium, the dielectric constant can be estimated as ⁵⁸:

$$\epsilon = 1 + 4\pi N\alpha \quad \text{eq. 2}$$

where N is a density in atoms/Å³ and α is the isotropic atomic polarizability. N is calculated for a 9 Å radius sphere around each atom, where protein atoms are counted and excluded volume is populated by an appropriate number of waters. Assuming additivity of atom polarizabilities ⁵⁹, α is then defined as the sum of individual polarizabilities for protein atoms and waters in the 9 Å radius sphere (see Methods, Table 2 for atomic volumes and polarizabilities). Using these values, the calculated dielectric constant is 76 in a fully solvated sphere, consistent with the experimentally determined dielectric of bulk water ⁶⁰. This approach assumes that principles for estimating the dielectric constants in homogeneous, isotropic, non-interacting media – essentially ideal gases – can be reasonably applied to the complex microenvironment of a protein and associated solvent. Generally, this problem is addressed using grid-based calculations of the Poisson-Boltzmann equation, including structural optimization or elaboration of local molecular degrees of freedom to

determine the role of conformational ensembles in tuning the local dielectric environment^{61,62}.

Benchmarking calculations

To determine whether such assumptions allowed reasonable estimates of ϵ , we calculated dielectric constants for a series of *Staphylococcus aureus* nuclease (SNase) mutants developed in the Garcia-Moreno laboratory^{63,64}. Mutations generally take the form of the introduction of a non-native titratable amino acid. Mutations are distributed throughout the protein from the core to the surface. Collectively, studies of these mutants demonstrated that the protein microenvironment can significantly modulate the pKa of titratable amino acids. Based on observed pKa shifts of lysine, aspartate and glutamate substitutions, apparent local dielectric constants were estimated for each position. In order to compare our calculations to these values, mutations were modeled into a high-resolution structure of SNase⁶⁵ using protCAD. Side chain rotamer optimization and local backbone moves were allowed to relax unfavorable interactions. ϵ was calculated from the minimized structure at the site of the charged side chain atom(s) of the mutation (Figure 3). The correlation between ϵ determined from observed pKa shifts versus model structures of the mutant was significant ($R^2 = 0.85$) and spanned a range of values from 10 to 30. Estimates for surface-exposed groups ranged as high as 50 (Figure 4), indicating dielectrics approaching that of bulk water.

Solvation Potential

The energetics of protein-water interactions is another important component of an effective force field that takes on greater significance in peptides. Of the implicit solvation models used, Generalized Born (GB) is one that allows for rapid calculation and is frequently incorporated into MD and design calculations⁶⁶. A key parameter in GB models is the Born radius (b), which describes the degree of burial of a group within a solute – and can be considered as related to its distance from the molecular surface:

$$\Delta E_{GB} = -\frac{1}{2} \left(\frac{\epsilon_s - 1}{\epsilon_s} \right) \frac{q^2}{b} \quad \text{eq. 3}$$

where ϵ_s is the dielectric constant of the solvent and q is the atom partial charge.

The approach described here for calculating local dielectrics considers the number of solute atoms within a 9 Å radius sphere, which would correlate with the degree of burial. Other design potentials use similar atom density calculations for implicit solvation⁶⁷. Thus, we explored whether the estimated local dielectric could be incorporated into a GB-like term for modeling solvation, by evaluating the ratio of the local estimated dielectric to that of the solvent:

$$\Delta E_{solw} = -\frac{1}{2} \left(\frac{\epsilon_a}{\epsilon_s} \right) \frac{q^2}{9} \quad \text{eq. 4}$$

The local dielectric ϵ_a estimated from structure as described in eq. 2 is used to describe burial instead of a Born radius. The radius is a constant 9 Å reflecting the distance across which atom density is computed.

Calculating CMP Stabilities

Given the high degree of solvent accessibility of the collagen triple helix, estimating CMP stability from structures or structural models would be an appropriate test of the utility of local dielectric constant computations. Applying the same local dielectric calculations developed on SNase to a model CMP, (POG)₁₀ – O = 4R-hydroxy-proline, reveals a different distribution of dielectric constant values, reflecting the high solvent accessibility throughout the triple-helix fold (Figure 4). No atoms in collagen peptides experience a dielectric lower than 20, with the majority of the backbone atoms at 30 and side chain atoms from 35–45. Values as high as sixty are computed at the termini of the triple helix. Thus, all interactions in collagen from backbone interchain hydrogen bonds to side chain charge pair networks exist in moderate dielectric environments that would be expected to mitigate their contributions to stability.

A dataset of forty-three CMP sequences containing charged amino acids was assembled from various biophysical studies in the literature (Table 1). Peptides ranged in length from 21 to 31 amino acids and included sequences from host-guest studies, *de novo* designed molecules and fragments of natural collagens. In all of the studies, stability was reported in terms of melting temperatures (T_m), generally measured by thermal denaturation experiments in circular dichroism spectrometers. Some variability is expected in the experimental data set as CMP stabilities are sensitive to concentration, ionic strength and the presence of N- and C-terminal capping groups. Furthermore, the relatively slow folding kinetics of collagen requires that thermal denaturation profiles be conducted with a very gradual heating schedule⁶⁸, as faster melts can result in overestimation of T_m .

Sequences were threaded onto the high-resolution X-ray diffraction structure of a (POG)₁₀ homotrimer and subject to side chain optimization in protCAD. The energy function consisted of a van der Waals packing term, and the Coulombic (eq. 1) and solvation (eq. 4) terms that incorporated calculated local dielectric constants. The folding energy of the triple helix was defined as the difference in the computed stability of the trimer and those of the component monomer chains, modeled in a relaxed extended conformation.

A comparison of the computed and experimental stabilities revealed a strong correlation for all but eleven sequences (Figure 5a). Each of these outliers had a significant number of hydrophobic amino acids (four or more per chain). It is possible that the lack of a term for solvation of hydrophobic surfaces in the current energy function results in overestimating the stability of these sequences. The function adequately reflects stability differences identified in host-guest studies in those cases where changes are significant. For example, at Y positions in a model (POG)₇ homotrimer, the relative stability of charged amino acids (R>E>K>Y)¹⁷ which span T_m s over 15°C is recapitulated (Figure 5b). The same mutations at the X position have a minor impact on peptide stability, spanning on 3°C. In this case, no correlation is observed with computed stabilities. This may be due to precision of experimental measurements as well assumptions regarding the conformations of folded and

unfolded states. Models of peptides start with the assumption of an idealized backbone conformation matching that of (POG)₁₀, which may bias sampling and contribute to the accuracy and precision of the calculation. If a uniform dielectric of 30, reflecting the mean calculated value across the triple helix (Figure 4), is imposed on all positions in the peptide, the correlation between calculation and experiment is very poor ($R^2 < 0.1$).

An atomic-level scoring function should be capable of making predictions at the atomic level. For nearly all of the CMPs studied, high-resolution structures were not available. We therefore applied the scoring function in protCAD to recapitulate a few CMPs for which solved structures were available. The structure of human Type III collagen, residues 991–1032 was solved to 2.3 Å resolution⁶⁹. Of particular interest was a cluster of interchain charge pairs between arginine, glutamate and asparagine residues, and whether the current scoring function would recover this interaction network.

In protCAD, side chains were removed and replaced with the same amino acid in an idealized conformation. Side chain rotamers were then optimized with a constant temperature Monte Carlo sampling. The best scoring models showed significant overlap with the experimental structure (Figure 6a). In many cases alternate rotamers were observed that brought the charged ends of the side chains into equivalent proximity to the structure, suggesting that despite extensive flexibility on the peptide surface, the favorable electrostatic interactions were promoting interchain charge pairs over protein-solvent interactions. Charged amino acids that were not participating in charge pair interactions in the experimental structure were less likely to have similar conformations in the model, suggesting that these electrostatic interactions were the primary constraint on the model. In another part of the same structure, a histidine-hydroxyproline hydrogen bond was recovered in the model (Figure 6b). The correspondence between model and structure, while encouraging, should be interpreted with caution as the proline-poor regions of the structure were poorly resolved and had high B-factors, presumably due to the ability of flexible amino acids to adopt multiple conformations.

In another fragment of Type III collagen whose structure was experimentally resolved⁷⁰, a network of hydrophobic interactions was recovered between phenylalanine, valine and methionine residues. As with charge pair interactions involving flexible arginine side chains, flexible methionine groups would often adopt different rotamers in the model and experimental structure, but maintain the same interchain contacts. While there is no explicit term for burial of hydrophobic surface, van der Waals packing interactions between side chains would be expected to dominate the energy function when side chain-solvent interactions would be weak (due to small partial charge values for side chain atoms in eq. 4).

CONCLUSIONS

Using an additive model of atomic polarizability, one can reasonably estimate local apparent dielectric constants in proteins based atomic structures or models. These calculations are expected to be valuable in computing electrostatics and solvation contributions to the stability of small peptides, where solute-water interactions dominate the free-energy of folding. A comparison of computed and experimental stabilities for a series of CMPs shows

some success, but highlights the need for appropriate treatment of hydrophobic interactions. Next steps in the development of these tools include building a hydrophobic burial term, testing their extensibility to a broader set of peptides that include other topologies, and evaluating speed and accuracy relative to GB and other fast solvation models.

The ability to design collagen at the atomic level is expected to drive development of the triple-helix as a structural scaffold for molecular recognition, and lead to better predictive models for CMP stability and structure. More generally, tools for peptide design will accelerate therapeutics development and allow us to broadly explore the sequence-structure-function space accessible to this class of molecules.

METHODS

Calculating the local dielectric

We estimate the electrostatic free energy of peptides by solving for a local atomic dielectric environment using the dielectric constant form⁵⁸ (see eq. 2), where N is a density in atoms per $\text{\AA}^3$ and α is the additive isotropic atomic polarizability⁵⁹, for all atoms within $9\text{\AA}$. Volumes for each atom were used to estimate how many water molecules can occupy the remaining volume. The exact volume used to calculate the remaining volume for waters was optimized fit to experimental effective dielectrics. Polarizabilities and volumes used for calculation are listed in Table 2 and are categorized into amber unified atom van der Waals types for calculation. Polarizabilities are additive within $9\text{\AA}$ sphere around an atom and the polarizability of water is then added for the estimated number of water molecules in the remaining volume to result in final predicted effective dielectric product for each atom.

Energy calculation

The resulting atomic dielectric is then used to estimate the solute-solvent energy in the form (see eq. 4), where ϵ_s is the dielectric constant of the bulk solvent, q is the the AMBER 2003 forcefield united atom charge⁷¹ and r is the radius of the dielectric sphere ($9\text{\AA}$). This is a modification of the Generalized Born solvation free energy model for a variable dielectric environment. An average of the local atomic dielectrics is used to calculate the Coulombic pairwise interactions (eq. 1), where ϵ is the average atomic dielectric of the atom pair and r is the distance between the atoms. The van der Waals energy E_{vdw} is calculated using the standard Lennard-Jones with parameters from the AMBER 2003 united atom force field. The total energy for each atom is then:

$$E_{tot}(kcal/mol) = E_{vdw} + E_{elec} + E_{solv} \quad (\text{eq. 5})$$

Building the peptide model

Models were built by using the starting crystal structure of (POG)₉ (PDB ID: 3B0S)⁷² as a structural repeating unit that was extended to desired size. A new sequence was then threaded on the structure and then optimized using protCAD. The optimization protocol samples rotamers for each amino acid⁷³, applying the first rotamer that improves the E_{tot} for the protein system iteratively until a defined successful sampling plateau is reached. The

dihedral backbone angles are also adjusted one degree iteratively when the structure nears the sampling plateau, while the total energy improves. After each structural change the dielectrics are recalculated for the entire structure to capture minor solvation changes and further optimize the structure. The final E_{tot} sum for all residues and all atoms final dielectrics is the final E_{complex} .

Calculating triple helix stability

The E_{tot} is summed for all atoms in the triple helix and then subtracted from the sum of E_{tot} for each chain, where the dielectrics of each chain are calculated in the absence of the other two chains.

$$E_{\text{fold}} (\text{kcal/mol}) = E_{\text{complex}} - (E_{\text{chain1}} + E_{\text{chain2}} + E_{\text{chain3}}) \quad (\text{eq. 6})$$

The backbone geometry of the solvated chains are assumed to be the same as in the folded state.

Accessing and using the code

protCAD³⁸ is available as an open source software library for linux at <https://github.com/protCAD/protcad>. Installation and usage instructions are included in the main github folder. Code (example.cc) that runs the optimization program used to generate the final triple helix models generated for this publication is available in the projects directory.

Acknowledgments

This work was supported by NIH NIGMS grant R01 GM-089949.

References

1. Shen MY, Davis FP, Sali A. Chemical Physics Letters. 2005; 405:224–228.
2. Streicher WW, Makhataдзе GI. Biochemistry. 2007; 46:2876–2880. [PubMed: 17295518]
3. Neidigh JW, Fesinmeyer RM, Andersen NH. Nature Structural Biology. 2002; 9:425–430. [PubMed: 11979279]
4. Li Y, Foss CA, Summerfield DD, Doyle JJ, Torok CM, Dietz HC, Pomper MG, Yu SM. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:14767–14772. [PubMed: 22927373]
5. O’Leary LE, Fallas JA, Bakota EL, Kang MK, Hartgerink JD. Nature chemistry. 2011; 3:821–828.
6. Jiang T, Xu C, Zuo X, Conticello VP. Angewandte Chemie. 2014
7. Rele S, Song Y, Apkarian RP, Qu Z, Conticello VP, Chaikof EL. Journal of the American Chemical Society. 2007; 129:14780–14787. [PubMed: 17985903]
8. Cejas MA, Kinney WA, Chen C, Vinter JG, Almond HR Jr, Balss KM, Maryanoff CA, Schmidt U, Breslav M, Mahan A, Lacy E, Maryanoff BE. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:8513–8518. [PubMed: 18559857]
9. Pires MM, Chmielewski J. Journal of the American Chemical Society. 2009; 131:2706–2712. [PubMed: 19182901]
10. Przybyla DE, Chmielewski J. Journal of the American Chemical Society. 2010; 132:7866–7867. [PubMed: 20499839]

11. Xu F, Khan IJ, McGuinness K, Parmar AS, Silva T, Murthy NS, Nanda V. *Journal of the American Chemical Society*. 2013; 135:18762–18765. [PubMed: 24283407]
12. Xu F, Li J, Jain V, Tu RS, Huang Q, Nanda V. *Journal of the American Chemical Society*. 2012; 134:47–50. [PubMed: 22171825]
13. McGuinness KN, Nanda V. *Journal of Biomolecular Structure & Dynamics*. 2013; 31:109–109.
14. Bella J, Eaton M, Brodsky B, Berman HM. *Science*. 1994; 266:75–81. [PubMed: 7695699]
15. Hwang ES, Thiagarajan G, Parmar AS, Brodsky B. *Protein science: a publication of the Protein Society*. 2010; 19:1053–1064. [PubMed: 20340134]
16. Kar K, Ibrar S, Nanda V, Getz TM, Kunapuli SP, Brodsky B. *Biochemistry*. 2009; 48:7959–7968. [PubMed: 19610672]
17. Persikov AV, Ramshaw JA, Brodsky B. *The Journal of biological chemistry*. 2005; 280:19343–19349. [PubMed: 15753081]
18. Gauba V, Hartgerink JD. *J Am Chem Soc*. 2008; 130:7509–7515. [PubMed: 18481852]
19. Oshea EK, Lumb KJ, Kim PS. *Current Biology*. 1993; 3:658–667. [PubMed: 15335856]
20. Gauba V, Hartgerink JD. *J Am Chem Soc*. 2007; 129:2683–2690. [PubMed: 17295489]
21. Chan VC, Ramshaw JA, Kirkpatrick A, Beck K, Brodsky B. *J Biol Chem*. 1997; 272:31441–31446. [PubMed: 9395477]
22. Venugopal MG, Ramshaw JA, Braswell E, Zhu D, Brodsky B. *Biochemistry*. 1994; 33:7948–7956. [PubMed: 8011657]
23. Persikov AV, Ramshaw JA, Kirkpatrick A, Brodsky B. *Biochemistry*. 2005; 44:1414–1422. [PubMed: 15683226]
24. Fallas JA, Hartgerink JD. *Nature communications*. 2012; 3:1087.
25. Giddu S, Xu F, Nanda V. *Proteins*. 2013; 81:386–393. [PubMed: 23042255]
26. Xu F, Silva T, Joshi M, Zahid S, Nanda V. *J Biol Chem*. 2013; 288:31616–31623. [PubMed: 24043622]
27. Xu F, Zahid S, Silva T, Nanda V. *J Am Chem Soc*. 2011; 133:15260–15263. [PubMed: 21902217]
28. Xu F, Zhang L, Koder RL, Nanda V. *Biochemistry*. 2010; 49:2307–2316. [PubMed: 20170197]
29. Gauba V, Hartgerink JD. *J Am Chem Soc*. 2007; 129:15034–15041. [PubMed: 17988128]
30. Fallas JA, Dong J, Tao YJ, Hartgerink JD. *The Journal of biological chemistry*. 2012; 287:8039–8047. [PubMed: 22179819]
31. Brodsky B, Thiagarajan G, Madhan B, Kar K. *Biopolymers*. 2008; 89:345–353. [PubMed: 18275087]
32. Raman SS, Gopalakrishnan R, Wade RC, Subramanian V. *Journal of Physical Chemistry B*. 2011; 115:2593–2607.
33. Stultz CM. *Protein Science*. 2006; 15:2166–2177. [PubMed: 16943446]
34. Keshwani N, Banerjee S, Brodsky B, Makhatadze GI. *Biophysical Journal*. 2013; 105:1681–1688. [PubMed: 24094409]
35. Jiang L, Kuhlman B, Kortemme TA, Baker D. *Proteins-Structure Function and Bioinformatics*. 2005; 58:893–904.
36. Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:10147–10152. [PubMed: 16006526]
37. Bella J, Brodsky B, Berman HM. *Structure*. 1995; 3:893–906. [PubMed: 8535783]
38. Summa CM. Department of Biochemistry and Molecular Biophysics; University of Pennsylvania School of Medicine; Philadelphia: 2002.
39. Summa CM, Rosenblatt MM, Hong JK, Lear JD, DeGrado WF. *Journal of Molecular Biology*. 2002; 321:923–938. [PubMed: 12206771]
40. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL, Saven JG, DeGrado WF. *Journal of the American Chemical Society*. 2005; 127:5804–5805. [PubMed: 15839675]
41. Cochran FV, Wu SP, Wang W, Nanda V, Saven JG, Therien MJ, DeGrado WF. *Journal of the American Chemical Society*. 2005; 127:1346–1347. [PubMed: 15686346]

42. McAllister KA, Zou HL, Cochran FV, Bender GM, Senes A, Fry HC, Nanda V, Keenan PA, Lear JD, Saven JG, Therien MJ, Blasie JK, DeGrado WF. *Journal of the American Chemical Society*. 2008; 130:11921–11927. [PubMed: 18710226]
43. Grzyb J, Xu F, Weiner L, Reijerse EJ, Lubitz W, Nanda V, Noy D. *Biochimica Et Biophysica Acta-Bioenergetics*. 2010; 1797:406–413.
44. Braun P, Goldberg E, Negron C, von Jan M, Xu F, Nanda V, Koder RL, Noy D. *Proteins-Structure Function and Bioinformatics*. 2011; 79:463–476.
45. Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG. *Plos Computational Biology*. 2012;8.
46. Slovic AM, Summa CM, Lear JD, DeGrado WF. *Protein Sci*. 2003; 12:337–348. [PubMed: 12538897]
47. Lear JD, Stouffer AL, Gratkowski H, Nanda V, DeGrado WF. *Biophysical Journal*. 2004; 87:3421–3429. [PubMed: 15315956]
48. Stouffer AL, Nanda V, Lear JD, DeGrado WF. *Journal of Molecular Biology*. 2005; 347:169–179. [PubMed: 15733926]
49. Howard KP, Liu W, Crocker E, Nanda V, Lear J, Degrado WF, Smith SO. *Protein Science*. 2005; 14:1019–1024. [PubMed: 15741331]
50. Cristian L, Nanda V, Lear JD, DeGrado WF. *Journal of Molecular Biology*. 2005; 348:1225–1233. [PubMed: 15854657]
51. Adamian L, Nanda V, DeGrado WF, Liang J. *Proteins*. 2005; 59:496–509. [PubMed: 15789404]
52. Tatko CD, Nanda V, Lear JD, DeGrado WF. *Journal of the American Chemical Society*. 2006; 128:4170–4171. [PubMed: 16568959]
53. Hsieh D, Davis A, Nanda V. *Protein Science*. 2012; 21:50–62. [PubMed: 22031179]
54. Nanda V, DeGrado WF. *Journal of the American Chemical Society*. 2004; 126:14459–14467. [PubMed: 15521766]
55. Nanda V, DeGrado WF. *Journal of the American Chemical Society*. 2006; 128:809–816. [PubMed: 16417370]
56. Rodriguez-Granillo A, Annavarapu S, Zhang L, Koder RL, Nanda V. *Journal of the American Chemical Society*. 2011; 133:18750–18759. [PubMed: 21978298]
57. Gordon DB, Marshall SA, Mayo SL. *Curr Opin Struct Biol*. 1999; 9:509–513. [PubMed: 10449371]
58. Purcell, EM. *Electricity and Magnetism*. Cambridge University Press; 2011.
59. Mkadmh AM, Hinchliffe A, Abu-Awwad FM. *Journal of Molecular Structure-Theochem*. 2009; 901:9–17.
60. Malmberg CG, Maryott AA. *Journal of Research of the National Bureau of Standards*. 1956; 56:1–8.
61. Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B. *Proteins-Structure Function and Genetics*. 1993; 15:252–265.
62. Song YF, Mao JJ, Gunner MR. *Journal of Computational Chemistry*. 2009; 30:2231–2247. [PubMed: 19274707]
63. Isom DG, Castaneda CA, Velu PD, Garcia-Moreno B. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107:16096–16100. [PubMed: 20798341]
64. Isom DG, Castaneda CA, Cannon BR, Garcia-Moreno BE. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108:5260–5265. [PubMed: 21389271]
65. Castaneda CA, Fitch CA, Majumdar A, Khangulov V, Schlessman JL, Garcia-Moreno BE. *Proteins-Structure Function and Bioinformatics*. 2009; 77:570–588.
66. Still WC, Tempczyk A, Hawley RC, Hendrickson T. *Journal of the American Chemical Society*. 1990; 112:6127–6129.
67. Kono H, Saven JG. *Journal of Molecular Biology*. 2001; 306:607–628. [PubMed: 11178917]
68. Persikov AV, Xu Y, Brodsky B. *Protein Sci*. 2004; 13:893–902. [PubMed: 15010541]
69. Boudko SP, Engel J, Okuyama K, Mizuno K, Bachinger HP, Schumacher MA. *J Biol Chem*. 2008; 283:32580–32589. [PubMed: 18805790]

70. Boudko SP, Bachinger HP. *J Biol Chem*. 2012; 287:44536–44545. [PubMed: 23132862]
71. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong GM, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang JM, Kollman P. *Journal of Computational Chemistry*. 2003; 24:1999–2012. [PubMed: 14531054]
72. Okuyama K, Miyama K, Mizuno K, Bachinger HP. *Biopolymers*. 2012; 97:607–616. [PubMed: 22605552]
73. Bower MJ, Cohen FE, Dunbrack RL. *Journal of Molecular Biology*. 1997; 267:1268–1282. [PubMed: 9150411]
74. Yang W, Battineni ML, Brodsky B. *Biochemistry*. 1997; 36:6930–6935. [PubMed: 9188687]
75. Doss-Pepe E, Deprez P, Silva T, Inestrosa NC, Kirkpatrick A, Ramshaw JA, Brodsky B. *Biochim Biophys Acta*. 2004; 1698:187–195. [PubMed: 15134651]
76. Deprez P, Doss-Pepe E, Brodsky B, Inestrosa NC. *Biochem J*. 2000; 350(Pt 1):283–290. [PubMed: 10926855]
77. Parmar AS, Zahid S, Belure SV, Young R, Hasan N, Nanda V. *Journal of Structural Biology*. 2014; 185:163–167. [PubMed: 23603270]
78. Parmar AS, Joshi M, Nosker PL, Hasan NF, Nanda V. *Biomolecules*. 2013; 3:986–996. [PubMed: 24970200]
79. Jalan AA, Demeler B, Hartgerink JD. *J Am Chem Soc*. 2013; 135:6014–6017. [PubMed: 23574286]
80. Shah NK, Sharma M, Kirkpatrick A, Ramshaw JA, Brodsky B. *Biochemistry*. 1997; 36:5878–5883. [PubMed: 9153429]
81. Mohs A, Li Y, Doss-Pepe E, Baum J, Brodsky B. *Biochemistry*. 2005; 44:1793–1799. [PubMed: 15697204]
82. Buevich AV, Silva T, Brodsky B, Baum J. *J Biol Chem*. 2004; 279:46890–46895. [PubMed: 15299012]
83. Xu Y, Hyde T, Wang X, Bhate M, Brodsky B, Baum J. *Biochemistry*. 2003; 42:8696–8703. [PubMed: 12873129]
84. Kang YK, Jhon MS. *Theoretica chimica acta*. 1982; 61:41–48.
85. Gerstein, M., Richards, FM., Chapman, MS., Connolly, ML. *International Tables for Crystallography Chapter 221 Protein surfaces and volumes: measurement and use*. Wiley; 2006. p. 531-574.

Figure 1.

The unique amino acid composition profiles of fibrous and globular proteins reflect solvent accessibility and folding mechanisms. **(a)** Excluding proline and glycine, which comprise half of the human Type-I collagen sequence, most amino acids are charged and/or polar. This is consistent with the collagen triple helix structure, which lacks a hydrophobic core and is instead held together by a network of interchain hydrogen bonds – black arrows. **(b)** The muscle protein tropomyosin is an extended dimeric α -helical coiled-coil and has intermediate properties between fibrous and globular proteins. It folds along a hydrophobic core, but utilizes an extensive network of charged amino acids for stability and specificity of assembly. **(c)** Hemoglobin is typical of a globular fold with an extensive hydrophobic core and a smaller fraction of polar and charged residues.

Figure 2.

Networks of side chain charge pairs provide stability and specificity of folding to the collagen triple helix. For example, in the collagen fragment shown, two arginines flank a glutamic acid in the sequence. The triple helix fold does not allow for significant intrachain interactions. These groups participate in cross-strand interactions. Assuming a homotrimer, this sequence can contribute up to five charge-pair interactions to the triple helix. A structure of such a network is shown (PDB ID 3DMW)⁶⁹. For clarity, chain 1 is depicted twice to indicate both its interactions with chains 2 and 3 in the helix.

Figure 3.

Calculations of local dielectric constants were benchmarked against measured constants in a series of mutations to SNase. **(a)** Coloring of SNase structure according to calculated ϵ shows a range of values correlated to burial of atoms within the structure. Values match experimentally determined apparent dielectric constants^{63,64}. For example - site 1: L125K, predicted = 11.1, experimental = 10.0; site 2: L38E, predicted = 16.4, experimental = 17.3; site 3, predicted = 26.0, experimental = 26.0. At site 4, K53, the calculated ϵ at the terminal side chain amine is 47.2. **(b)** A reasonable correlation was observed across all reported mutations in references 63,64.

Figure 4.

Calculated local dielectric constants for a model CMP are consistent with a high degree of solvent accessibility throughout the structure. Normal distribution fits of ϵ show a higher mean and smaller environmental variability in the CMP versus the SNase globular fold.

Figure 5.

Correlation of computed and experimentally determined stabilities for a series of CMPs. **(a)** The scoring function reasonably recapitulates a broad range of peptide stabilities (solid circles), if those containing greater than four hydrophobic residues per chain are excluded (open circles). Computed energies are expressed relative to $(\text{POG})_{10}$. **(b)** For a subset of CMPs consisting of host-guest studies of amino acid substitutions at X and Y positions in the triple helix, the scoring function identifies the stability rank of Y-position substitutions (circles), but not the subtle effects of X-position substitutions (triangles).

Figure 6. Side chain conformation prediction on the surface of CMPs

(a) Networks of charge-pair interactions on the surface of a fragment of the structure of human Type-III collagen (PDB ID 3DMW)⁶⁹ (green) are recapitulated in the model (light brown). (b) A nearby histidine-hydroxyproline hydrogen bond in the same structure is identified by the model. (c) On another fragment of Type III collagen (PDB ID 4GYX)⁷⁰, packing interactions between hydrophobic Phe and Met side chains are preserved in the model.

Table 1

CMP data set

Sequence	T _m (exp)	T _m (calc)	ref
GARGPAGPQGPRGDKGETGPOGPOGPOGPOGV	30.8	34.6	74
GERGERGEKGERGEQQRDGPPOGPOGPOGPOGV	23.2	13.0	17
GPOGPOGPOGKRKGPOGPOGPOGPOGG	32.3	36.9	75
GPOGPOGROGRKGROGPOGPOGPOGPOGPOG	30	31.2	76
GPOGROGRRGROGLKGEQGPPOGPOGPOGPOGY	23.3	22.7	17
PKGPKGPKGKOGPDGDOGDOGDOGPKGPKG ^b	19	14.0	77
POGPOGPOGPOGPOGPOGPOGDDGDDGDDG	20.8	13.4	78
DDGDDGDDGPOGPOGPOGPOGPOGPOGPOG	17.5	7.3	78
PKGPKGPKGPKGPKGPKGPKGPKGPKGPKG ^b	37	28.5	79
PKGPKGPKGKOGPDGDOGDOGDOGPKGPKG ^b	29	16.7	27
KKGKKKKKGPOGPOGPOGPOGPOGPOGPOG	31.1	34.1	<i>a</i>
POGPOGPOGPOGPOGPOGPOGPOGPOGPOG	60	56.8	17
POGPOGPOGDOGPOGPOGPOGPOG	40.1	39.5	17
POGPOGPOGEOGPOGPOGPOGPOG	42.9	39.4	17
POGPOGPOGKOGPOGPOGPOGPOG	41.5	42.6	17
POGPOGPOGROGPOGPOGPOGPOG	40.6	43.2	17
POGPOGPOGPDGPOGPOGPOGPOGG	34	38.6	17
POGPOGPOGPEGPOGPOGPOGPOGG	39.7	45.8	17
POGPOGPOGPKGPOGPOGPOGPOGG	36.8	45.0	17
POGPOGPOGPRGPOGPOGPOGPOGG	47.2	47.1	17
POGPOGPOGPOGPOGPOGKKGKKGKKG	31.1	28.2	<i>a</i>
POGPOGPOGPOGPOGPOGEEGEEGEEG	17	15.2	<i>a</i>
POGPOGPOGPOGPOGPOGRRGRRGRRG	44.3	34.7	<i>a</i>
GPOGPOGROGKRKGQKQKPOGPOGPOGPOG	23.9	30.4	76
GDAGAOGERGPOGLAGAOGPOGPOGPOGPOGV	23.2	32.8	80
GLAGEOGKOGIOGLOGRAGPOGPOGPOGPOGV	25.4	24.0	17
GAOGERGPOGLAGAOLRGPOGPOGPOGPOGV	25.9	30.9	80
GINGFOGKDGRDGTKGEKGEOPGPOGPOGPOGG	23	11.2	81
GGKGDAGAOGERGPOGLAGPOGPOGPOGPOGV	20.9	30.7	80
GSOGAQGLQGPRGLOGTOGPOGPOGPOGPOGV	25	37.8	17
GPOGIRGPKGQKGEOLGPOGPOGPOGPOGY	28.4	29.2	17
GPOGPOGPOGAKGDAGPOGPAGPOGPOGPOGY	42.8	39.5	17
<i>High hydrophobic content CMPs</i>			
GPAGPAGPVGPAARGPAGPOGPOGPOGPOGV	26	52.6	74
GPOGPOGPOGPOGPAAGPVGPAARGPAGV	26	55.1	82
GPAGPAGPVGPAARGPAGPOGPOGAOGPOGV	24.1	51.6	83
GPAGPAGPVGPAARGPAGPOGPAAGPOGPOGV	23.2	55.9	83
GERGPOGLAGAOLRGAGPOGPOGPOGPOGV	16.5	36.4	80
GPOGLAGAOLRGAGPOGPOGPOGPOGPOGV	15.8	40.1	80

Sequence	T _m (exp)	T _m (calc)	ref
GLAGAOGLRGGAGPOGPEGPOGPOGPOGV	17.5	37.7	80
GPOGROGAAGAOGVRGPRGPOGPOGPOGPOG	18.6	35.9	76
GPOGAOGRKGROGVAGPAGPOGPOGPOGPOG	11	35.1	76
GPOGROGKRKGTGLKGDIGPOGPOGPOGPOG	9.8	29.9	76
GPOGLOGMLGQKGEMGPKGPOGPOGPOGPOG	7	34.0	76

^a unpublished data

^b abc heterotrimer: sequence of one of the three chains shown

T_m(calc) values are based on the linear correlation of computed energies to experimental values in Figure 5: $T_m = (142 - E_{\text{calc}})/2.5$

Table 2

Atomic parameters used in dielectric constant estimation

Atom Type	α - polarizability ($\text{\AA}^3$)	volume ($\text{\AA}^3$)
C	1.382	8.7
C*	1.382	8.7
C2	1.836	23.2
C3	2.222	36.7
CA	1.382	8.7
CB	1.529	8.7
CD	1.768	21.3
CN	1.529	8.7
CG	1.768	21.3
CH	1.45	20.4
N	1.48	13.6
N2	1.866	21.4
N3	2.252	22.7
NA	1.48	13.6
O	0.46	15.9
OH	1.05	18
O2	0.664	18
S	3.2684	29.2
SH	3.6643	36.7
HOH	1.4907	19.1

Atom types are selected from the AMBER 2003 force field and reflect properties of the heavy atom and bound protons, i.e. united atoms⁷¹. α values are from refs. 59,84. Atomic volumes from ref 85.