Modeling Polyhelical Repeat Proteins:

A Mathematical Approach to a Chemical Problem

Lincoln Wurtz

Abstract

Proteins are the most abundant biological macromolecules and, based on their three-dimensional shape, perform life-sustaining functions. The process by which a protein assumes its folded shape remains an open question and has intrigued biologist and chemists for decades. Mathematicians have joined forces with the natural scientists and brought with them the tools of differential geometry, which prove powerful for modeling proteins. We explore the method of [3] to model a small subset of proteins using polyhelical space curves. We successfully modeled three alpha-helical repeat proteins. The developed model has demonstrated possible uses in predicting theoretical tertiary structures of proteins given a set of secondary structures—a step in the right direction of solving the protein folding problem. Additionally, we provide insight into the relationship between clashes and the model's stability calculator, which may improve the viability of their model.

1 Introduction

Proteins are large biomolecules typically regarded as the workhorses of biological systems. Their title is undoubtedly based on the fact that proteins have such a wide range of function. Proteins are responsible for immunity (e.g. antibodies), structural support (e.g. collagen), catalyzing slow biological reactions (e.g. enzymes), and molecular signaling (e.g. insulin).

Incredibly, proteins accomplish all of this simply by utilizing their three-dimensional shape. For example, antibodies can target and flag foreign materials likely to cause illness (antigens) based on a perfect fit between the antibody's binding pocket and the antigen. Much like a lock and key, the antibody and its target antigen must match for an immune response to be mounted. Even a small variation in shape can decrease the function of the antibody to select a specific antigen. In this example, we see that a protein's usefulness is highly dependent on proper shape. This property holds in general. Therefore, integral to a protein's function—and by extension biological life—is the process by which a protein assumes its active three-dimensional shape.

Proteins assume their final active shape through a process called protein folding. In order to understand folding, it is first important to understand the composition of proteins. Proteins consist of small monomer molecules called amino acids. Proteins can have anywhere from tens to tens of thousands of amino acids depending on their function. Each amino acid

consists of an amine group, a hydroxyl group, and a residue group attached to what is labeled as the "alpha-carbon." See Figure 1 .

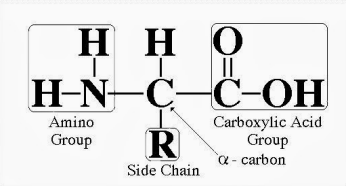

Figure 1: Molecular structure of an amino acid with all relevant features labeled. Image retrieved from http://dgbiochem.blogspot.com/2014/11/amino-acids-structure-and-nomenclature.html.

The residue group of a specific amino acid differentiates it from other amino acids. The residue groups vary in size, shape, electrostatic charge, and chemistry. This invariably influences protein structure and function.

The amino acids that make up a protein are linked together by a series of peptide bonds. Taken as a whole, these peptide bonds are considered the backbone of a protein, and the backbone does not include the variable residue regions. See Figure 2 for detail.

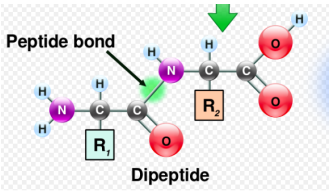

Figure 2: Molecular structure of two amino acids linked by a peptide bond labeled by the green circle. The backbone of a protein consists of a series of peptide bonds and includes all the atoms labeled above except the residues labeled R_1 and R_2. Image retrieved from wikipedia.org/wiki/Peptide_bond.

At this point in the synthesis, the protein has not yet assumed its final three-dimensional shape. Rather, the protein exists as a loose chain of amino acids. This is called the primary structure. See Figure 3. The protein has no definite shape and is free to twist and turn. The next step in the protein folding process is the development of secondary structures such as α-helices and β-sheets (Figure 3). As the name implies, α-helices are helical structures where roughly 3.6 amino acids constitute each turn in the helix. On the other hand, β-pleated sheets are a collection of amino acids that lie in a single plane. Both secondary structures work to stabilize the chain as it becomes more regulated and rigid. Nearby amino acids on the backbone interact with each other to form local fixtures.

The final stage relevant to our model is the tertiary structure of the protein. In this stage, the protein folds in on itself while preserving secondary structures. This is considered a global process since amino acids from very distant parts of the protein chain can fold close to each other. Guiding this process is the development of salt bridges and disulfide bonds, which are

created by the interaction of amino acid residues. Additionally, proteins are subject to their aqueous environments and tend to compact due to interactions with water molecules ([7]).

Interestingly, a folded protein can be completely denatured (i.e. unfolded), and when placed under appropriate conditions, refolded into the same exact functional shape. Therefore, the shape of the protein is dependent on the sequence of amino acids. Theoretically, the sequence of amino acids could be used to predict the three-dimensional shape ([7]).

This problem of finding the folded shape of a protein given its sequence of amino acids is called the protein folding problem. The problem's solution is highly sought after since protein structure is so fundamental to biological life, and the only way to deduce the structure is to experimentally crystallize the protein and gather X-ray diffraction data. The crystallization process is delicate, and collecting the necessary data to create an accurate protein structure model is labor intensive. Whereas determining protein structure experimentally is an involved process, determining the sequence of a protein is straightforward. Thus, there is a real advantage to developing methods to determine the elusive protein structure given the readily available amino acid sequence.

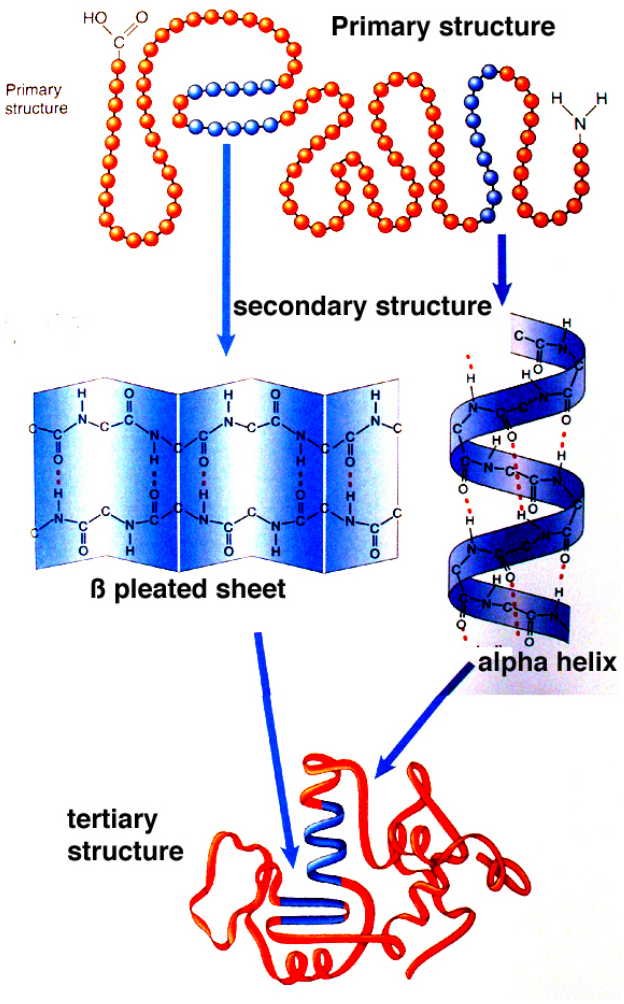

Figure 3: The sequence of protein folding starting from the primary structure to the final tertiary structure. Image retrieved from schoolworkhelper.net/protein-structures-primary-secondary-tertiary-quaternary.

As it turns out, the entire folding process is guided by thermodynamics, and the final protein structure minimizes total energy. Other researchers, like Lazaridis and Karplus in [5] or Lundgren in [6], have tried to understand protein folding by minimizing a complex energy function on protein models. The work in [6] is incomplete since it only considers the backbone of a protein with no regard for the residue amino acids. Additionally, the work was completed on randomly generated protein backbones and not real proteins.

Rackovsky and Scheraga in [8] and [9] describe their efforts to understand the folding of proteins by utilizing differential geometry. Using experimental data, they determined the impact each residue had on the torsion and curvature of the backbone of the protein. Continuing the use of differential geometry to study protein folding, Simmons and Weiner in [10] modeled proteins with a mathematical ribbon and obtained differential equations to

study folding behaviors.

Clearly there is an intersection between the fields of mathematics and chemistry. Hausrath and Goriely in [3] strike the perfect balance and develop a simple protein model using straightforward techniques in differential geometry. This paper works through the development of the model as laid out in [3] as well as expands on their work. In particular, we will study the formation of stable tertiary protein structures given specific secondary structures. To simplify this endeavor, we examine proteins that are composed entirely of α-helices. Such proteins can be accurately modeled by simple polyhelices using differential geometry.

2 Methods

The methods below were adopted in part from [3]. The theory behind the construction of polyhelices was obtained from Goriely and Hausrath's work in [2].

2.1 Differential Geometry and Development of Polyhelix

Consider a space curve $r = r(s)$ parameterized by arc length s. Next consider the Frenet Frame at each point on the curve, where $t(s)$ represents the tangent vector, $n(s)$ represents the normal vector, and $b(s)$ represents the binormal vector. We know that we can describe the change in these vectors by taking their derivatives with respect to s. This yields the Frenet equations listed below:

$$\begin{aligned} r' &= t \\ t' &= \kappa n \\ n' &= -\kappa t + \tau b \\ b' &= -\tau n, \end{aligned}$$

where κ and τ are the curvature and torsion of the space curve, respectively. Each vector in the Frenet Frame consists of three components and can be denoted via a subscript. For example, $t = \langle t_1, t_2, t_3 \rangle$. With this understanding, define Y as

$$Y = [t_1, n_1, b_1, t_2, n_2, b_2, t_3, n_3, b_3, r_1, r_2, r_3]^T.$$

Notice that we can express the Frenet equations as a differential matrix equation $Y' = M(s) \cdot Y$, where

$$M = \begin{bmatrix} F & 0 & 0 & 0 \\ 0 & F & 0 & 0 \\ 0 & 0 & F & 0 \\ V_1 & V_2 & V_3 & 0 \end{bmatrix}, F = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0. \end{bmatrix},$$

and V_i is the 3×3 matrix whose single entry is a 1 in row i column 1. The V_i submatrices ensure that the derivative of the position vector r is equal to the tangent vector. In other words, they ensure that $r' = t$.

If $M(s)$ is constant, then an exact solution to the matrix differential can be obtained. This requires curvature and torsion to be constant. Under these circumstances we know that

the solution is a helix. Now, if we suppose that $M(s)$ is piecewise constant (i.e. curvature and torsion are constant throughout intervals), then the solution to the differential equations can still be found, and it would describe a piecewise helix. The piecewise helix is a series of connected helical arcs and is referred to as a polyhelix.

Since each segment of the polyhelix consists of a constant curvature and torsion for some prescribed distance, we can describe polyhelices by a *curvature profile*. The curvature profile is a list of triples $P = \{(\kappa_i, \tau_i, L_i), i = 1..N\}$. The first segment of the polyhelix would be given by (κ_1, τ_1, L_1) where κ_1 and τ_1 are the curvature and torsion, respectively, over the length of the polyhelix L_1. The next segment would build off of the previous segment with a new set of curvature, torsion, and length. See Figure 4.

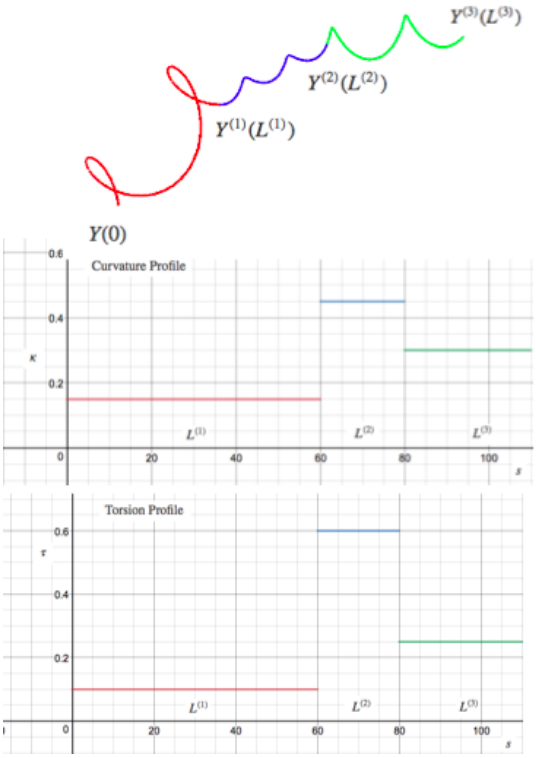

Figure 4: The curvature profile for an example three segment polyhelix. Each colored segment is itself a helix. Notice that in each segment the curvature and torsion are constant over some length $L^{(i)}$.

In Figure 4, we can completely describe the three segment polyhelix by the list of triples $P = \{(0.15, 0.1, 60), (0.45, 0.6, 20), (0.3, 0.25, 30)\}$. Each triple encodes all the necessary information to describe a segment. Taken together, they describe the red-blue-green polyhelix.

2.2 Solving the Matrix Differential Equation

We want to solve the equation $Y' = M(s) \cdot Y$ when $M(s)$ is piecewise constant. We find

$$Y(s) = A(\kappa, \tau; s) \cdot Y(0), 0 \leq s \leq L,$$

where $Y(0)$ encodes the initial conditions and $A(\kappa, \tau; s) = e^{sM}$ is the matrix exponential. We know that the matrix exponential is defined as the series $e^{sM} = sI + sM + \frac{sM^2}{2!} + \frac{sM^3}{3!} + \ldots$, where I is the identity matrix of appropriate size. This can be computed using Maple. Upon simplification using the Taylor series definition of sine and cosine, we find

$$A(\kappa, \tau; s) = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ b_1 & b_2 & b_3 & I_3 \end{bmatrix}, a = \begin{bmatrix} \frac{1}{\alpha^2}(\tau^2 + \kappa^2 \cos(\alpha s)) & \frac{\kappa}{\alpha}\sin(\alpha s) & \frac{\kappa\tau}{\alpha^2}(1 - \cos(\alpha s)) \\ -\frac{\kappa}{\alpha}\sin(\alpha s) & \cos(\alpha s) & \frac{\tau}{\alpha}\sin(\alpha s) \\ \frac{\kappa\tau}{\alpha^2}(1 - \cos(\alpha s)) & -\frac{\tau}{\alpha}\sin(\alpha s) & \frac{1}{\alpha^2}(\kappa^2 + \tau^2 \cos(\alpha s)) \end{bmatrix},$$

where $\alpha = \sqrt{\kappa^2 + \tau^2}$ and the 3×3 submatrices b_i have the single nonzero row i with entries

$$\begin{aligned} (b_i)_{i1} &= \frac{\alpha s \tau^2 + \kappa^2 \sin(\alpha s)}{\alpha^3}, \\ (b_i)_{i2} &= \frac{\kappa}{\alpha^2}(1 - \cos(\alpha s)), \\ (b_i)_{i3} &= \frac{\kappa\tau}{\alpha^3}(\alpha s - \sin(\alpha s)), i = 1, 2, 3. \end{aligned}$$

2.3 Constructing the Polyhelix

Let us examine the significance of this result. We have found the equations that completely describe the segments of a polyhelix. For each segment, however, we need to orient it so that the entire piecewise polyhelix is continuous. To accomplish this, we ensure that the Frenet Frames between segments align. By aligning the Frenet Frames at the connection points, we are guaranteed that at least the first and second derivatives agree.

To orient the segment, we need to multiply the A matrix by an initial condition vector $Y(0)$. In our case, we set the origin to be the starting point and position our Frenet Frame along the x, y, and $z-$axes for simplicity. This means that $Y(0) = [1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0]^T$. We can then define the first segment as

$$Y^{(1)} = A(\kappa_1, \tau_1; s) \cdot Y(0).$$

The parametric expression in arc length for the first segment of the curve $r(s)$ is given by the last three components of the $Y^{(1)}(s)$ vector. Using Figure 4 as a reference, $Y^{(1)} = A(0.15, 0.1; s) \cdot Y(0)$ and corresponds to the red segment.

Similarly, we can define the second segment as

$$Y^{(2)} = A(\kappa_2, \tau_2; s - L_1) \cdot Y^{(1)}(L_1).$$

This construction is similar to the first segment because we multiply the A matrix by a set of initial conditions. This time the initial condition vector is generated by evaluating $Y^{(1)}$ at its last point, L_1.

As discussed above, this ensures that the next segment picks up where the first segment left off and also ensures the Frenet Frames align for smoothness. Additionally, we evaluate the A matrix at $s - L_1$ so that the first point in the second segment occurs when we feed the value $s = L_1$. Again, this is simply a manifestation of making the segments align. In Figure 4, $Y^{(2)} = A(0.45, 0.6; s - 60) \cdot Y^{(1)}(60)$ and corresponds to the blue segment.

Continuing with this extended example in Figure 4, we can find the parametric equations for the third green segment. It is described by $Y^{(3)} = A(0.3, 0.25; s - (80)) \cdot Y^{(2)}(80)$. Of course, this method can be used to describe polyhelices of any number of segments, not just three. For our purposes, we will focus on polyhelices with six segments.

2.4 Modeling Proteins

Goriely and Hausrath in [3] provide the curvature profiles for three proteins listed by their PDB codes. As it turns out, all of these proteins can be modeled by a repeated six-segmented polyhelix.

PDB	1qqe			1b3u			1hz4		
segment	κ	τ	L	κ	τ	L	κ	τ	L
1	0.38	0.15	68.5665	0.38	0.15	65.1111	0.38	0.15	70.1281
2	0.2733	−0.0082	8.5389	0.2490	0.1380	8.7975	0.2258	−0.0432	6.0429
3	0.0	0.6629	5.6701	0.1569	−0.2731	8.9004	0.0425	−0.03288	6.3368
4	0.38	0.15	74.6944	0.38	0.15	66.5759	0.38	0.15	70.2449
5	0.9781	−1.000	8.2056	0.3019	0.1381	6.7962	0.1382	−0.2512	6.0209
6	0.2317	−0.8655	4.3622	0.1151	−0.4492	6.4457	0.0	−0.2688	4.0167

Figure 5: The curvature profile values for three helical repeat proteins listed by their PDB codes as given by Goriely and Hausrath in [3].

Two of the helical segments are used to approximate the regular secondary α-helical structures in the protein, and the other four are used to provide an accurate transition between α-helices. Thus, the pattern is {α-helix1, turn, turn, α-helix2, turn, turn}. As a note to discourage confusion, the turn segments (segments 2,3,5, and 6) are still described by helices, but they are not obviously recognizable in the model since they are shorter in length. This motif is repeated a number of times and accurately describes the three proteins. Below is a graphic of the protein 1qqe as determined by X-ray crystal diffraction.

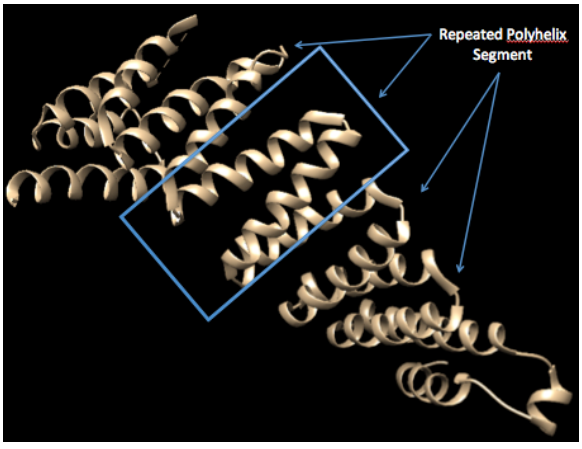

Figure 6: The protein 1qqe is considered a repeat polyhelix protein since the region in the blue box is repeated throughout the protein. Image generated using UCSF Chimera software.

In the above example, the protein 1qqe can be described by constructing the polyhelix in the blue box using the data from Figure 5 and repeating it a number of times. The other listed proteins also follow this scheme and can be generated accordingly.

2.5 Scoring Proteins

Once constructed, the proteins are quantitatively scored using a quality function. The quality function was built with biological and chemical foundations in mind with the hopes that the quality function encapsulates some of the crucial components of protein stability. This is a greatly simplified analogue of the complex energy functions seen in [5] and [6]. In this paper, we maintain the quality function used in [3]. While their quality function is fairly basic in construction, it provides a good start for quantitatively describing protein stability.

To score the proteins, a set of sample points $S = \{r(s_k)\}$ is needed, where r is the parameterized polyhelix. One chemical quality we want included in our model is the idea of compactness. Proteins, under the influence of solvents, tend to curl up on themselves. Therefore, we want proteins that are compact to be higher scoring than those that are loosely packed.

To accomplish this we create a measure called *contact order*. If two points $r(s_i)$ and $r(s_j)$ are within some prescribed distance d of each other in three-dimensional space, then we say that the two points form a contact. Of course, this occurs frequently when random sample points are taken from local neighborhoods. However, we promote the concept of compacting a protein by averaging the arc length distance between the sample points that produce a contact. Notationally we can write

$$CO(d) = \frac{1}{LN} \sum^{N} |s_i - s_j|,$$

where L is the number of sample points, and N is the number of contacts.

Notice that a summand is large if two points form a contact but are distant from one another by arc length. Therefore, the contact order scores proteins by measuring how well distant points on the backbone are brought within a small distance in three dimensional space.

However, we want to avoid points that are too close to one another. In other words, we want to punish proteins that have self-intersections or near intersections. To do this we define the quality function as

$$Q(c, d) = \frac{CO(d)}{2^{M(c)}},$$

where c is the clash distance, and $M(c)$ is the number of points that are within the clash distance. A clash is observed when two points on the model are within the prescribed clash distance c. Notice that for a protein with many clashes, the quality function exponentially drops toward zero regardless of the value of the contact order.

3 Results and Discussion

3.1 Resources Used to Develop Model

The methods described above follow the procedure in [3]. Maple 18 was used to develop and display protein models. Unless otherwise specified, for each protein, a set of 100 sample points was used.

3.2 Modeling Proteins

The first step in modeling proteins was to use the curvature profiles of [3] (see Figure 5) to generate parametric curves. Below is an image created in Maple that depicts part of the model of protein 1qqe (Figure 7). Each of the six segments is given a different color.

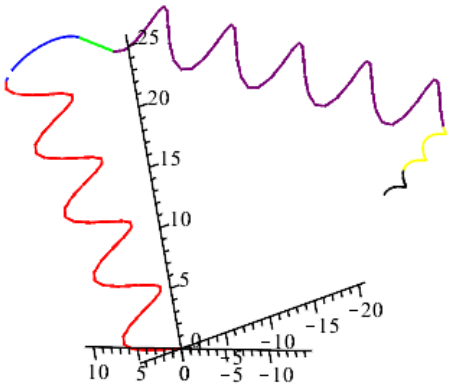

Figure 7: Part of the model for protein 1qqe.

Notice that the red segment models an alpha-helix and starts at the origin since it is the first segment. The blue, green, yellow, and black pieces are considered turn pieces and correspond to segments 2,3,5, and 6, respectively, in the table displayed in Figure 5. Finally, the purple piece models another alpha-helix given by segment 4. The other proteins were also generated.

To complete the model of these proteins, the six-segment polyhelix must be repeated a number of times. This capability was also coded into a Maple workbook. For example, the protein 1b3u can be described by thirteen repeated six-segment polyhelices. Figure 8 depicts the modeled backbone of the protein in comparison with experimentally obtained X-ray diffraction pattern.

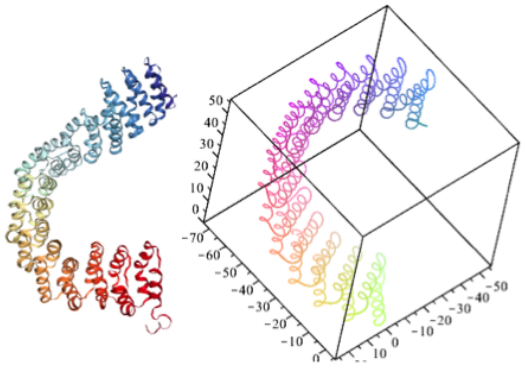

Figure 8: Comparison between experimentally determined structure of 1b3u and modeled 1b3u. On the left is the structure generated by X-ray diffraction methods retrieved from RCSB PDB, and on the right is the polyhelix model of the protein generated using our model in Maple 18.

Notice the remarkable similarities between the model and the actual protein. This indicates that in at least a few cases, our repeated polyhelix model is capable of generating the mathematical equations that accurately represent real proteins.

3.3 Exploring Curvature Space

Notice that all three polyhelices listed in Figure 5 have identical $\kappa_1, \tau_1, \kappa_4$, and τ_4. In other words, segments 1 and 4 are held constant (although their lengths can change) while the remaining segments can vary in all parameters. This was an intentional move by Hausrath and Goriely in [3]. In particular, they wanted to examine the space of all polyhelices of this type.

In some ways, this is an attempt to examine biological connections between proteins. In this example, all three have the same modeled alpha-helix segments while allowing the "turn segments" of the model to vary. This addresses the possible folding patterns of proteins. In particular, it models the last step in folding from the secondary to tertiary structure of a protein. This model allows us to start with a secondary structure set, hold these fixtures constant, and vary the connection pieces to simulate folding.

However, exploring this entire space is rather difficult given the freedom between all the parameters. If we were to look at all polyhelices of this type, we would need to search through a 14 parameter space. There are 14 parameters since each segment is described by a triple, and there are six triples for each polyhelix. Thus, there are 18 parameters to describe all six-segment polyhelices, and we hold four parameters constant ($\kappa_1, \tau_1, \kappa_4$, and τ_4), which yields a 14 parameter space.

Hausrath and Goriely looked at a two dimensional subspace of that 14 parameter space. They accomplished this by formulating the 14 parameters in a vector. Next, v_0 was defined to be the vector with parameters to specify the polyhelix segment of protein 1qqe. The vectors v_1 and v_2 were similarly constructed to represent proteins 1hz4 and 1b3u, respectively. Then, they analyzed the space generated by the two new parameters a and b, which acted like scalars

to produce a new set of parameters, v. Below is their construction:

$$v = v_0 + a(v_1 - v_0) + b(v_2 - v_0).$$

Thus, the protein 1qqe lives at the point $(a, b) = (0, 0)$ on the a-b plane. In addition, $(1, 0)$ corresponds to 1hz4, and $(0, 1)$ corresponds to 1b3u.

The points in this plane represent a protein and can be scored by the quality function. This provides a simple way of visualizing connections between polyhelical proteins. Since the parameters a and b are scalars, small distances in the a-b plane translate to small perturbations in protein structure. This provides a nice way to generate new polyhelical proteins. For example, we can "watch" the transformation of the protein 1qqe to 1hz4 by slowly stepping through a values. See Figure 9.

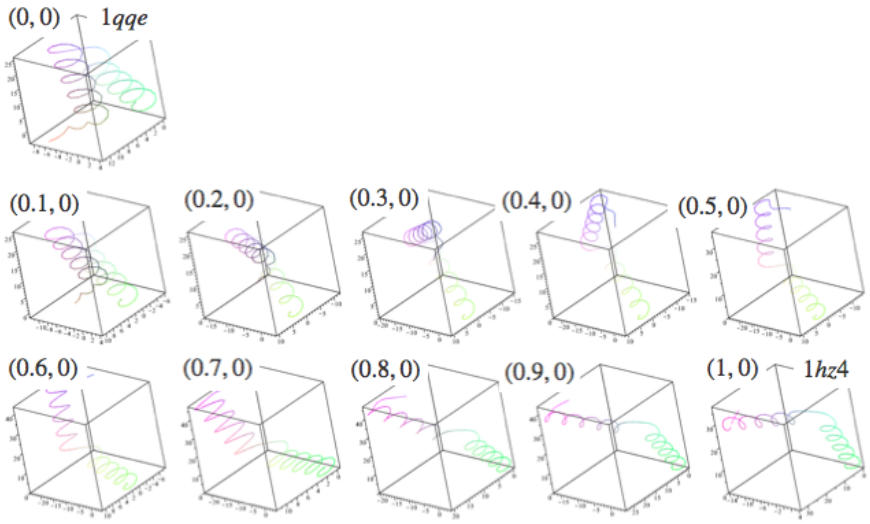

Figure 9: Pictures of the transformation of the protein 1qqe into 1hz4 by incrementing the a component with a step size of 0.1.

While the two parameter subspace provides an easy way to modify many different parameters of a protein model, the computing power to determine anything quantitatively using the quality function is substantially higher than Maple can produce in a reasonable timeframe. Therefore, we instead looked to modifying one parameter in an attempt to gain insight to the quality function.

3.4 Understanding the Quality Function: Contact Distance

As we attempt to score folding proteins, we need to be confident that our quality function is properly functioning. One way to address this concern is to test results with expectations. For example, as we increase the contact distance on the quality function, we expected to observe more contacts and a higher quality function.

We observed the average number of contacts on the protein 1qqe as we varied the contact distance from $d = 0.1$ to $d = 20$ by a 0.1 step size. As a note of reference, 100 sample points were used, and we averaged the results over 10 trials.

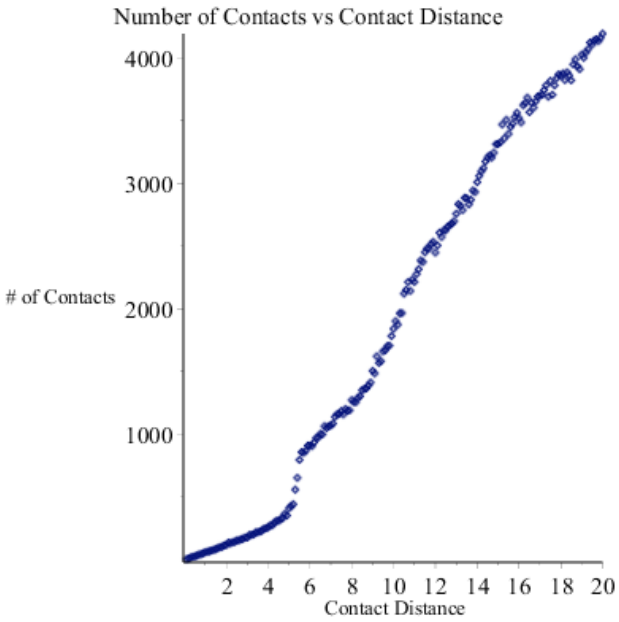

Figure 10: The average number of contacts over 10 trials is graphed against the contact distance.

As expected, the number of contacts increases as the contact distance is increased. Moreover, it appears mostly linear except for a jump between $d = 5$ and $d = 6$. This sharp increase in number of contacts is most likely due to the distance between the two modeled alpha-helices (segment 1 and segment 4 from Figure 5). Since these helices account for the bulk of the total length in the polyhelix, the majority of sample points will reside on them. Therefore, when the contact distance is larger than the distance between the two helices, we expect to find a jump in the number of contacts formed. This is supported by analyzing one distance between the helices on the protein 1qqe. See Figure 11 below.

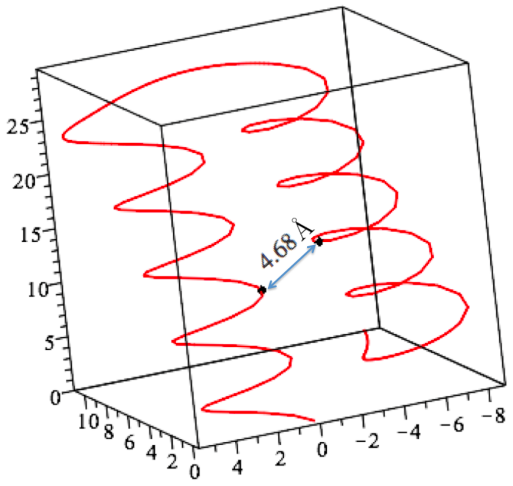

Figure 11: A scale to indicate the distance between two points on opposite alpha-helices. Units are likely in Angstroms, Å.

Therefore, we expect that soon after the contact distance increases past five units, the number of contacts should increase dramatically since contacts can now form across helices instead of only within. The importance of this detail can easily be understated, but it may help inform our model. Specifically, helical stacking helps to stabilize proteins, and we can build that feature into our model by setting the contact distance equal to the distance between helices. More work is needed to determine the optimal distance between helices; however, this result at least confirms that the contact order mechanism is capable of capturing realistic properties.

3.5 Understanding the Quality Function: Relationship with Length

The first parameter that we looked at was the length of the first helix. Using 1qqe as a base, we modified the length of the first helical segment and observed the quality function. We used increments of 1 unit from a starting length of 0 to a final length of 568.

The measurements were taken with a contact distance of $d = 21$ and a clash distance of $c = 0.1$. While there was a positive trend in the line of best fit, there was great variability in the output of the quality function. In an attempt to smooth out the noise, for each length, we observed the quality function 10 times and averaged the results. The final trend is displayed in Figure 12 below.

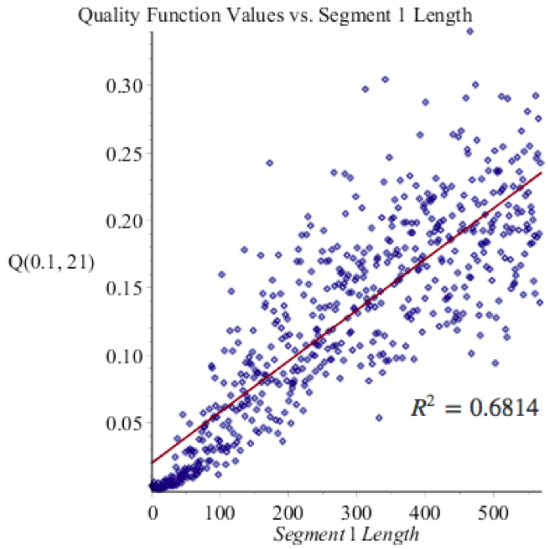

Figure 12: Scatter plot of the quality function versus the length of the first segment after averaging the quality function over 10 trials. The line of best fit is also included.

Simply averaging the values 10 times greatly improved resolution. We tried to push our resolution to the maximum by attempting a 100 trial average run, but the processing time in Maple was unrealistic (i.e. on the order of a week).

From this data it is clear that our quality function has a bias for longer proteins given our choice of parameters. This can be rationalized by considering the potential to create more long-ranged connections, which bolster the contact order value and the quality function. However, we suspected there would be a critical distance where the model protein was so long and the sample points so spread out that fewer contacts would be made thereby causing the quality function to suffer. In an attempt to find that critical distance, we hyperextended segment one to 568 units, which is far past its natural length of 68.56. Our results in Figure 12 are inconclusive at best with regards to finding this critical distance. Perhaps we could have observed this phenomenon if we used fewer samples points or had a smaller contact distance.

3.6 Understanding the Quality Function: Relationship with Clashes

In order to examine the relationship of the quality function with clashing events, we constructed a protein that very obviously clashes with itself. See Figure 13 below.

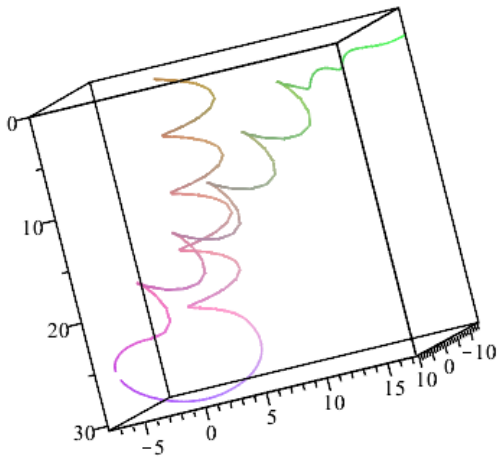

Figure 13: Depiction of a protein that nearly intersects with itself.

In fact, the protein was constructed in such a way that the first segment intersects with the rest of the protein. Therefore, we would expect that as we increase the length of this segment from 0 to its proper length of 68, the number of clashes would go up and the quality function would go down.

However, this is not the result we found. Instead we found that the average number of clashes steadily decreased as the length of the segment grew despite the fact that it was crossing itself (Figure 14). As a result the protein quality function rose.

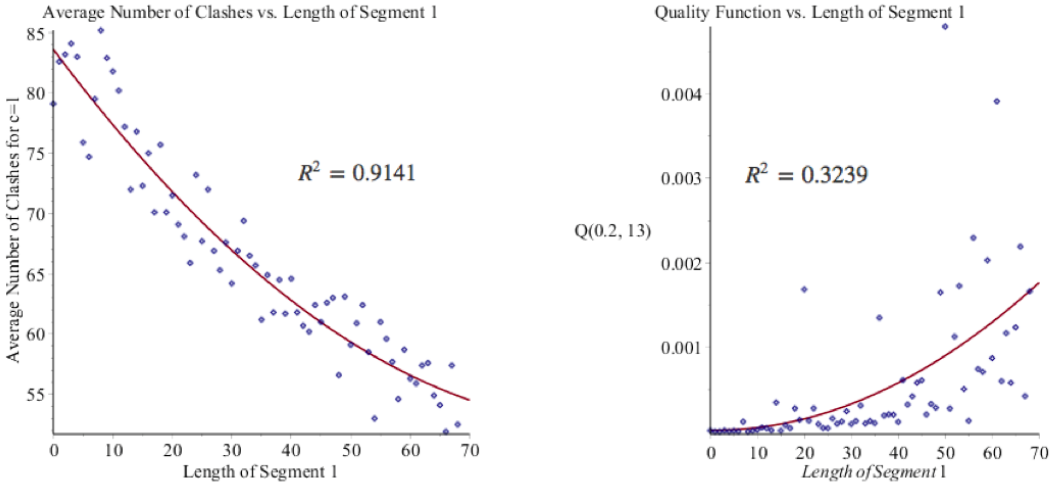

Figure 14: The average number of clashes and the quality function is analyzed as a function of Segment 1 length. Both experiments were averaged among ten trials.

Clearly, this issue exposed one weakness of the simple quality function that the authors in [3] used to score protein stability. The most disturbing issue about the construction was

the way in which contacts and clashes were defined. In particular, the randomly generated set of sample points may frequently distribute points nearby on the same segment. If the points were close enough, they would be considered a clash, despite the reality that they were simply generated near each other and were never in any real danger of intersecting. In other words, while the intention of the clash counter was meaningful, the actual implementation did not reflect the physical reality it was meant to describe as indicated by our previous result.

In an effort to quickly remedy the previous fault in the model, we explored a new clash counter. The new counter would only consider two points on separate segments to be a clash if they were within the clash distance. This modification produced the desired results.

After applying the new clash counter, the same test was run on the self-intersecting protein of Figure 13. Now we found that as the length of Segment 1 increased, the number of clashes rose and the quality function plummeted (Figure 15).

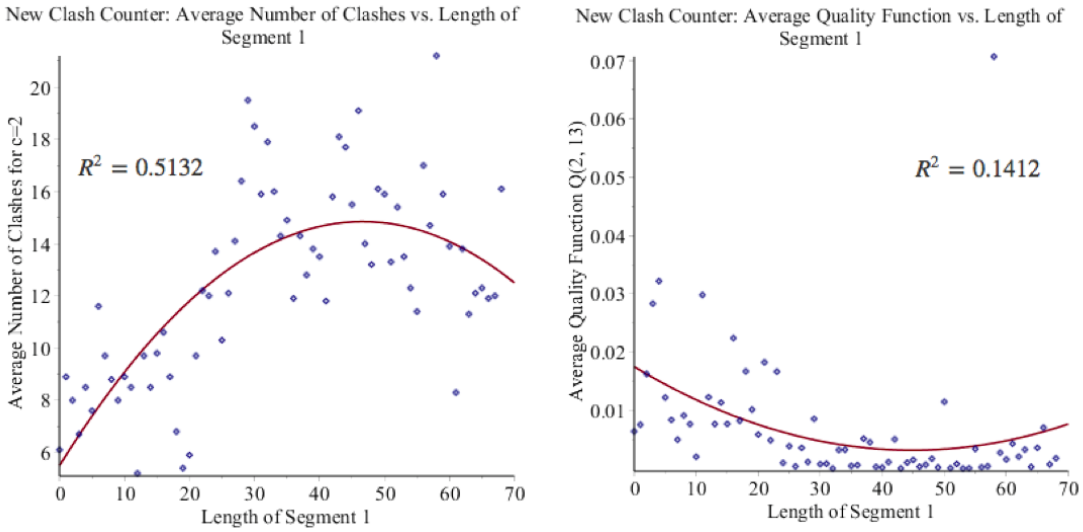

Figure 15: The average number of clashes and the quality function is analyzed as a function of Segment 1 length using the new clash counter. Both experiments were averaged among ten trials.

A neater result is displayed in Figure 16, which depicts the number of clashes when the threshold is increased to a clash distance of 3 as opposed to 2.

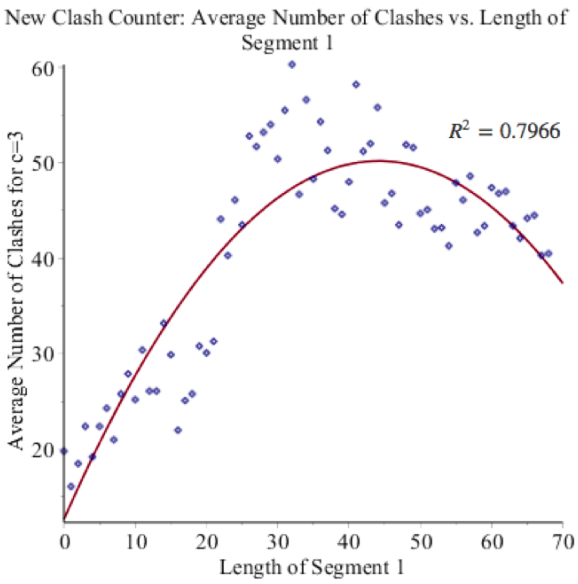

Figure 16: The average number of clashes is analyzed as a function of Segment 1 length using the new clash counter when the $c = 3$. This result is the average of ten trials.

Thus our modification to the clash counter produced the desired effects. This illustrates the flexibility of the model, but also indicates that there could be other shortcomings in the meaningfulness of the quality function.

4 Limitations and Future Work

While this model has successfully modeled three real proteins and is capable of describing countless theoretical proteins, there are a number of limitations. The most glaring limitation is the fact that we are not even addressing the protein folding problem directly. Instead of starting with a chain of amino acids, we begin with secondary structures. However, researchers like Katti et al ([4]) have contributed information about what types of sequences create what secondary structures. This work makes possible the creation of databases that would link short sequences to corresponding secondary structure. Therefore, it is not unrealistic to start the protein folding problem from a set of secondary structures as that information may become easier and easier to obtain.

Another limitation is that we are focused only on helical repeat proteins. Moreover, we are not even exploring the global stability of the protein since we are only focused on the quality of the unit cell that is repeated some number of times. For example, in Figure 6, we explore the quality or stability of the region inside the blue box, but our model does not consider the interactions between units (i.e. the entire protein). The repeat proteins were chosen because of their simplicity, and moving forward we could attempt to model other non-repeating proteins such as hemoglobin. The framework is in place. The only limiting factor is obtaining the curvature profile data.

Additionally, it is possible to generalize even further and move away from alpha-helices to explore other secondary structures. In particular, Hausrath and Goriely, who developed the

mathematics to describe alpha-helices in [3], also developed similar equations for beta-sheets in [1]. This expansion would open up the possibility of modeling nearly all proteins since alpha-helices and beta-sheets are by far the most common secondary structures ([7]).

Instead of expanding into new territory, we suggest that we perfect our given model first. In particular, the quality function is still the weakest aspect of our model. One issue is that, given a random distribution of sample points, the quality function is wildly inconsistent. Any definitive data is obtained only by averaging the quality function tens or hundreds of times at the cost of hundreds of hours of computation. We suspect this could be easily overcome by coding in Python or another programming language rather than Maple.

There is great potential in this model, and small refinements in the quality function will yield promising results. As shown above, our small change in the clash counter helped to make the model more realistic. We believe that by simply trying the model on different proteins and adjusting outcomes to expectations, the model's reliability and integrity will drastically improve.

For instance, we can imagine implementing data about the residue sequences to improve the accuracy of the quality function in determining protein stability. As mentioned in the introduction, the amino acid residues are not considered as part of the backbone; however, their chemistry and geometry influence the folding patterns of a protein. For example, electrostatic charges may hold a protein in a contorted position that it would not otherwise inhabit (and not otherwise be predicted by our quality function). We think that it would be possible to model the chemistry of the residues by coding in premium rewards in the quality function for bringing together points on the backbone that correspond to oppositely charged residues. This would represent a real step forward in the protein folding problem because from the literature surveyed, very few authors even consider the residue data due to the complexity their incorporation adds to models.

References

[1] Goriely, A., Hausrath, A., Neukirch, S. (2008). The differential geometry of proteins and its applications to structure determination. *Biophysical Reviews and Letters.* 3,1:77-101.

[2] Goriely, A., Neukirch, S., and Hausrath, A. (2009). Polyhelices through n points. *International Journal of Bioinformatics Research and Applications.* 5,2.

[3] Hausrath, A., Goriely, A. (2006). Repeat protein architectures predicted by a continuum representation of fold space. *Protein Science.* 15:753-760.

[4] Katti, M., Sami-Subbu, R., Ranjekar, P., and Gupta, V. (2000). Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Science].* 9:1203-1209.

[5] Lazaridis, T. and Karplus, M. (2000). Effective energy functions for protein structure prediction. *Structural Biology.* 10:139-145.

[6] Lundgren, M. (2007). A survey into Protein Folding: Curves and Energy Functions. *Uppsala Universitet.*

[7] Nelson, D. and Cox, Michael. (2013).Lehninger Principles of Biochemistry. *W. H. Freeman and company*, New York. Print.

[8] Rackovsky, S. and Scheraga (1978). Differential Geometry and Polymer Conformation Comparison of Protein Conformations. *American Chemical Society*. 11-6.

[9] Rackovsky, S. and Scheraga (1984). Differential Geometry and Protein Folding. *American Chemical Society*. 17:209-214.

[10] Simmons, W., Weiner, J. (2008). Protein Folding: A New Geometric Analysis. *arXiv*:0809.2079v1.