The H-L Protein Folding Model

In the early 1990s, biochemists Ken Dill and Hue Sun Chan invented the H-L[1] protein folding model to explore some of the features of real proteins without worrying about all of the complexities [3]. This model and variations of it have been studied since then by many biochemists, mathematicians, engineers, and other scientists interested in protein folding. In 2001, Oswin Aichholzer, David Bremner, Erik Demaine, Henk Meijer, Vera Sacristán, and Michael Soss explored many of the questions presented here [1] [2]. Amino acids can be divided into two types:

- H stands for hydrophobic amino acids – they Hate water
- L stands for hydrophilic amino acids – they Love water

Amino Acid Types

Alanine	hydrophobic (H)
Asparagine	hydrophilic (L)
Aspartic acid	hydrophilic (L)
Glutamine	hydrophilic (L)
Glutamic acid	hydrophilic (L)
Histidine	hydrophilic (L)
Isoleucine	hydrophobic (H)
Leucine	hydrophobic (H)
Lysine	hydrophilic (L)
Methionine	hydrophobic (H)
Phenylalanine	hydrophobic (H)
Serine	hydrophilic (L)
Threonine	hydrophilic (L)
Tryptophan	hydrophobic (H)
Tyrosine	hydrophobic (H)
Valine	hydrophobic (H)

Model Key

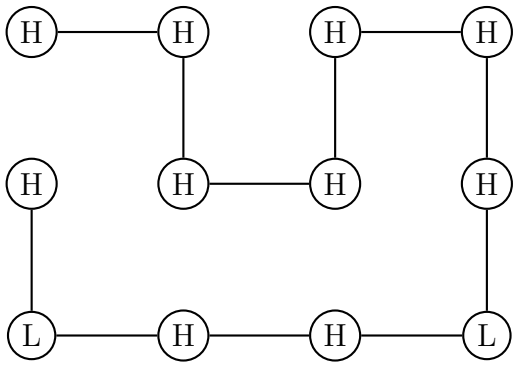
Hydrophilic amino acids (L)	Blue Circles
Hydrophobic amino acids (H)	Red Circles
Links in amino acid chain	White Bars
Hydrogen Bonds	Yellow Bars
Bond Score (B)	# of Yellow Bars

In the H-L model, a chain of amino acids is placed onto a square grid. This means that at each amino acid, the chain may continue straight ahead or may make a 90° bend to the left or right.

If two H amino acids are next to each other in the grid, but are not already connected by a link in the original chain, a hydrogen bond forms between the two H's. No diagonal bonds are allowed. The total number of hydrogen bonds is called the bond score (B) of the configuration. Amino acid chains fold themselves so that their bond scores are as large as possible, because these situations correspond to low-energy configurations for the protein and the water around it. The energy is lower in these configurations because when a chain has more hydrogen bonds, the hydrophobic amino acids are tucked inside the protein and are not exposed to the surrounding water molecules. When water molecules are disrupted by hydrophobic amino acids, nearby water molecules must orient themselves to create a water cage in order to be stable, which uses energy.

[1]The H-L model is usually called the H-P model in publications, but we use H to stand for amino acids that Hate water and L to stand for amino acids that Love water.

1. Here is an example of a folded protein chain. Using dashed lines, draw in the hydrogen bonds and determine the bond score (B) for the configuration. (Remember that diagonal bonds are not allowed and you may not draw hydrogen bonds between H's that are already next to each other in the chain.)



2. Translate the following chains of amino acids into strings of H's and L's:

   (a) Valine - Lysine - Tryptophan - Asparagine - Isoleucine - Histidine - Lysine - Phenylalanine

   (b) Leucine - Alanine - Aspartic acid - Glutamine - Tryptophan - Valine - Threonine - Tyrosine

   (c) Phenylalanine - Glutamic acid - Methionine - Glutamine - Alanine - Asparagine - Lysine - Valine - Threonine - Tryptophan - Aspartic acid - Glutamic acid - Phenylalanine - Isoleucine

3. Make a model of one of the chains in the previous section. Try to make a configuration that has as large a bond score as possible. Is there more than one way to get the maximal score?

4. Here are some other questions you might like to think about. While working, make a list of new questions, observations, and hypotheses that occur to you.

   (a) Make up your own strings of length 8, 9, 10, 11, or 12 and decide how many configurations they have which maximize B. Explore this question for other strings of the same length.

   (b) Under what conditions will a chain of odd length have a unique configuration maximizing B? What is different about chains of odd length versus chains of even length?

   (c) What is the largest bond score B that is possible for chains consisting of 8 H's? What about chains with 4 H's? Determine the maximal bond score for H-chains of many lengths, beginning with length 1 and going up from there. Can you give an upper or lower bound for the maximal B for a chain of an arbitrary number of H's?

   (d) Find an optimal folding for the chain L-H-L-L-H-L-L-H-L-L-H-L. We could write that chain as $(\text{LHL})^4$. Can you say anything about repeating pattern chains of the form $(\text{LHL})^n$ (meaning that the string L-H-L gets repeated n times)?

   (e) Make up your own repeating pattern chain along the lines of the previous question. Can you say anything about the maximal bond scores for your chain?

   (f) Consider a chain configuration picture and its mirror image. How do their corresponding H-L strings relate? Consider an H-L string and a string written in the reverse order. How do their chain configuration pictures relate? What does this suggest about how people exploring chains of a given length should count configurations?

   (g) How would you describe a particular configuration to someone without drawing the picture? Can you invent notation that would describe the picture using words or symbols? Does every possible configuration correspond to a description in your system? Is there more than one description in your system for any configuration?

   (h) Choose a chain with a unique maximal bond score configuration, and call its specific bond score b. How many other chains of the same length would have bond score b if they were placed in that configuration? Is b the maximal bond score for all of those other chains or do they have better configurations? If b is the maximal score, is this configuration unique for the other chains, or do they have other configurations that also have bond score b?

References

[1] O. Aichholzer, D. Bremner, E. D. Demaine, H. Meijer, V. Sacristán, and M. Soss. Long proteins with unique optimal foldings in the h-p model. In *Abstracts from the 17th European Workshop on Computational Geometry (EuroCG 2001), Berlin, Germany*, 2001.

[2] O. Aichholzer, D. Bremner, E. D. Demaine, H. Meijer, V. Sacristán, and M. Soss. Long proteins with unique optimal foldings in the h-p model. *Computational Geometry*, 25(1-2):139–159, May 2003.

[3] H. S. Chan and K. A. Dill. The protein folding problem. *Physics Today*, 29(31):7133–7155, February 1993.