Hydrophobic/Hydrophilic Profile of a Protein

Zasha Weinberg

March 7, 2000

Abstract

Regions of proteins can be hydrophobic or hydrophilic. Knowing the hydrophobicity of a protein gives clues as to the 3-D structure of the protein, and some likely behaviors of the protein.

The "Protein Hydrophilicity/Hydrophobicity Search and Comparison Server" graphically shows the hydrophobicity of a given protein. It is accessible via the Web.

The "Protein Hydrophilicity/Hydrophobicity Search and Comparison Server" is a tool to compute the hydrophobicity of proteins. It is available on the web[1].

This paper begins by describing what hydrophobicity is. This leads to a discussion of why hydrophobicity can be relevant in biological research. Then, the algorithm that the aforementioned site uses to calculate hydrophobicity is presented. It concludes with a description of the functionality of the Web site.

1 What is Hydrophobicity/Hydrophilicity?

Some amino acids are attracted to water. These are "hydrophilic". Similarly, amino acids that are repelled by water are called "hydrophobic". Not surprisingly, amino acids can be hydrophilic or hydrophobic to different degrees.

Regions of a protein also have hydrophilic or hydrophobic tendencies. This is dependent largely on the properties of the amino acids in that region, although this doesn't necessarily tell the whole story.

2 Relevance of Hydrophobicity/Hydrophilicity

Hydrophobicity indicates 3-D structure. Hydrophilic protein regions are attracted to water, while hydrophobic protein regions are repelled by water. In most cellular contexts, water surrounds the protein. Therefore, hydrophilic regions will tend to fold towards the outside of the protein, where the water is. By contrast, hydrophobic regions will try to avoid water by folding to the inside of the protein.

In this way, the predicted hydrophobicity of a region in a protein sequence is an indicator of whether that region will likely be on the outside or the inside of the protein, in that protein's 3-D shape.

3-D structure is relevant for other properties. The 3-D structure of a protein is notoriously difficult to discover. Computational methods do a poor job of predicting it, and the lab work involved in finding the correct structure is very extensive. However, for some purposes a researcher may only want to know if a given protein region will be folded outside or inside.

One example of this is in discovering phosphorylation sites of PKA in a protein, given the amino acid sequence of that protein. PKA phosphorylates sites on some proteins. This behavior is an important factor in the functionality of a protein. Currently, not much is known about how exactly PKA decides whether or not to phosphorylate a site on the protein.

However, we can presume that PKA is more likely to act upon regions on the exterior of a protein - which it can get to easily - rather than regions interior to a protein - which would be more difficult to reach. Suppose we also make the reasonable assumption that the predicted hydrophobicity of a region is correlated

[1] `http://bioinformatics.weizmann.ac.il/hydroph/index.html`

with it being interior to the protein. Then we can infer that highly hydrophobic subsequences are unlikely to be affected by PKA, while highly hydrophilic regions are more likely to be affected.

This type of argument is also relevant to at least one problem in drug design. In this scenario, we would like to design drugs that will target a virus's proteins. If the drug is to be effective, its target would presumably be something accessible, i.e., on the outside of the protein. Therefore, when searching for likely drugs, we would want to focus on targets that are significantly hydrophilic.

In general, when our interest in knowing 3-D structure is only to the extent that we want to know what is on the outside or the inside of a protein, then the hydrophobicity measure may be of practical value.

Hydrophobic/Hydrophilic factors in protein behavior. The hydrophobicity of regions of a protein relates to its behavior in the cell. One example of noteworthy behavior occurs in proteins that have highly hydrophobic regions that alternate with regions of high hydrophilicity. The interior of the cell has a lot of water. The cellular membrane, on the other hand, does not (being fatty). So, these kinds of proteins often will place themselves near the membrane of the cell. Their hydrophobic regions will lie in the membrane, while their hydrophilic regions will remain in the cell proper.

Therefore, when biologists see a hydrophobicity profile that matches this pattern, they may suspect this kind of behavior. This kind of pattern is fairly obvious given a graph of the hydrophobicity/-philicity of the protein.

3 Algorithm for Predicting Hydrophobicity/Hydrophilicity

The algorithm is very straightforward. First, we assign a hydrophilicity value to each amino acids, based on its chemical properties. Then, we assume that the actual hydrophilicity around some amino acid is based on the average hydrophilicity of surrounding individual amino acids.

The algorithm requires a constant "window size", which is an integer. Given an amino acid, we place it at the center of the window, and find the average individual hydrophilicity of all amino acids in the window with the given window size.

The hydrophilicity of a single amino acid is given by a table. There are two tables in use. They are

1. The Kyte-Doolittle table [2] and
2. The Hopp-Woods table [1]

By way of example, Table 1 shows the Kyte-Doolittle table.

4 Functionality of the Web Site

Broadly, a researcher can do the following on this Web site:

- Determine the hydrophilicity of one protein.
- Compare the hydrophilicity of two proteins.
- Find proteins whose hydrophilicity is similar to a given protein.

Protein sequences can be entered by the Swiss-Prot accession ID. In most cases their sequence can be cut & pasted into a form field.

4.1 The hydrophilicity of one protein

Given a protein, the Web site can show

- The predicted hydrophilicity around each position in the sequence
- A Fourier Transform of this information

Residue	Value
A	-1.8
F	-2.8
K	3.9
P	1.6
T	0.7
C	-2.5
G	0.4
L	-3.8
Q	3.5
V	4.2
D	3.5
H	3.2
M	-1.9
R	4.5
W	0.9
E	3.5
I	-4.5
N	3.5
S	-0.8
Y	1.3

Table 1: Kyte-Doolittle hydrophilicity values

- A report on transmembrane regions

For all of these functions, the user can select either the Kyte-Doolittle or Hopp-Woods values, and can select the window size. The default is Kyte-Doolittle with a window size of 17.

4.1.1 Predicted hydrophobicity

The site simply shows a graph of hydrophobicity as a function of position.

4.1.2 Fourier Transform

The Fourier Transform of the hydrophobicity by position is shown. Intuitively, it seems that this might be useful for analyzing whether or not the protein has certain characteristic properties. However, I am not aware of an example of such a property.

4.1.3 Transmembrane regions

Transmembrane regions are regions of a protein that cross the cellular membrane. The Web site will highlight the hydrophilicity of the transmembrane regions of a protein, and provide additional information showing how the hydrophilicity changes as it enters these regions.

The transmembrane regions themselves come from another database, although the Web site is not clear on where they come from. The documentation does not specifically say that the data comes from a database, but this seems the case, since (1) this is the only form that lacks a place to submit a sequence directly, which suggests that the site needs the Accession ID, and (2) the site sometimes complains that no transmembrane information exists for the given protein.

4.2 Comparing hydrophilicity profiles

The hydrophilicity of two proteins can be view simultaneously, to compare their hydrophilicity. It is left to the user to determine what, if anything, this means. This may be useful in understanding how two proteins are similar, or in explaining why two proteins share a common behavior.

4.3 Finding proteins with similar hydrophilicity

Given a protein, the Web site will find proteins (from GenBank) that have a similar hydrophilicity graph to the given protein. This is accomplished by discretizing the hydrophilicity values, and then running the Smith-Waterman algorithm to evaluate similarity.

This seems a reasonable heuristic, but it does not account for some notions of similarity that are relevant for hydrophilicity profiles. For example, if two proteins have the same sequence of hydrophilicity, except one of them is a constant amount greater at every position, we may want to think of these as being very similar. On the other hand, it is difficult to suggest a better measure of similarity.

One other issue with this is that we would expect that proteins with similar sequences will generally have similar hydrophilicity profiles. This follows because the hydrophilicity is based directly on the sequence. Therefore, a small number of changes to amino acids in the sequence will have a very small effect on the hydrophilicity.

For example, if you search for the protein P02945, the results of the hydrophilicity search are extremely similar to the results of a BLAST search for this protein on Entrez. There are differences, although it is unclear if these are because of differences between the BLAST and Smith-Waterman algorithms, differences between the default parameters on the Hydrophilicty/Hydrophobicity site versus those of the Entrez site, or because of proteins with dissimilar sequences that have similar hydrophilicity profiles.

It is also not clear that this is a drawback to the feature - when looking for proteins with similar hydrophilicity profiles, a researcher may not care if these are similar because of a similar sequence or not.

It should be noted that this feature of the site is described as being experimental.

5 Acknowledgement

Thanks to my project partner, Jesse Wiley, who showed me the Hydrophobicity/Hydrophilicity site, and suggested some possible uses for Hydrophobicity/Hydrophilicity measures in addition to our CSE 527 project on PKA.

References

[1] Hopp, T.P. and Woods, K.R., June 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Nat. Acad. Sci. USA*, 78(6): 3824-3828.

[2] Kyte, J. and Doolittle, R.F., 1982. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 157(6): 105-142.