DePaul University
University Libraries

DePaul Discoveries

Volume 5 | Issue 1 Article 20

2016

Wavelet Packet Analysis of Amino Acid Chain Sequences in the Proteins of Mesophile and Thermophile Bacteria

John B. Linehan
DePaul University, jack.linehan.18@gmail.com

Wavelet Packet Analysis of Amino Acid Chain Sequences in the Proteins of Mesophile and Thermophile Bacteria

John Linehan*

Department of Physics

Jesus Pando, PhD; Faculty Advisor

Department of Physics

ABSTRACT In this project, proteins from mesophile and thermophile bacteria with similar functions are compared. Initially it is assumed that the differences between these two bacteria are substantial to be recognized in the amino acid sequences of their proteins. These differences would then lead to the creation of a statistical measure, which would allow the classification of a protein to its corresponding bacteria. By assigning hydrophobicity values from three well-known scales, a discrete numeric signal is produced for each protein, which is analyzed using wavelet packets. The result of this method indicates that the overall hydrophobic tendencies of these two bacteria's proteins are very similar. As such, no identifying characteristic is readily apparent to classify a protein as belonging to specific bacteria.

INTRODUCTION

Mesophile bacteria live in moderate environments with temperatures ranging from $15 - 40$ degrees Celsius. Thermophile bacteria are extremophiles found in environments with temperatures ranging from 41 to 125 degrees Celsius [1]. Although these two bacteria have followed different evolutionary tracks, they perform similar biological functions.

Since both of these classes of bacteria have similar biological functions, it follows that their proteins performing these tasks are also similar in function. Since the two classes of bacteria evolved under different environmental pressures, it is likely that these proteins are made up of different amino acid chains. This study looked to find whether any characteristics exist that distinguish proteins with similar biological functions from the different classes of bacteria. Analysis of the proteins was performed in two steps. The first was to determine whether a statistical difference existed between the proteins using a statistical analysis of the proteins' amino acid sequences. The second was to study whether the different evolutionary patterns of the proteins resulted in different statistical features.

To answer these questions, the amino acid chain sequence of each protein was made into a signal.

*Linehan.jack@yahoo.com
Research Completed in Winter 2016

This will allow for analysis of a protein to be done in respect to a certain parameter. Signal analysis tools were then used to determine the differences and/or similarities between proteins with similar biological function from the different classes of bacteria. The hydrophobicity of each amino acid was used to construct the signal for each of the proteins. Hydrophobicity was used because it is an important factor in determining the secondary and tertiary structure of proteins. Since these levels of structure are considered to be a major factor in determining the overall function of a protein, using hydrophobicity to distinguish amino acids is reasonable.

METHODS

Data & Signal

Collaborators at the University of Denver provided the data for this project. The data contained the amino acid sequence for each of the 540 proteins per bacteria class used in this work. Data used in this project can be found at the Protein DataBase (PDB) [2].

As stated earlier, the proteins used in this research come from bacteria with evolutionarily distinct lineages. The first, mesophile, exists in environments with temperatures ranging between 15 and 40 degrees Celsius. Thermophile bacteria live in environments with temperatures ranging from 41 to 125 degrees Celsius. The differences in the temperatures of the bacterias' environment are assumed to be great enough so that the bacteria followed very different evolutionary tracks [1]. Proteins with similar functions were selected from both bacteria. This resulted in 540 proteins from both bacteria, a total of 1080 different proteins.

Amino acids are organic compounds composed of an amino and carboxyl group. They are distinguished from one another by their specific R group, bonded to the central carbon atom joining the amino and carboxyl groups. There are four different types of R groups: hydrocarbon (6 amino acids), neutral (7 amino acids), acid (1 amino acid) and base (6 amino acids). The R group is used to determine an amino acid's hyrdrophobicity [3].

The amino acids of each protein contain information about how they react to their environment. To discern this information, hydrophicity values were assigned to each amino acid. Hydrophobicity is the tendency to bend towards or away from water. The hydrophobicity of the 20 amino acids has been experimentally determined and are listed below for three well-known hydrophobicity scales.

Table 1. Hydrophobicity Scales, adapted from [4]

Amino Acid	Abbreviation	Engelman and Steiz [4]	Kyte Doolittle [5]	Hopp Woods [6]
phenylalanine	Phe	3.7	2.8	-2.5
methionine	Met	3.4	1.9	-1.3
isoleucine	ile	3.1	4.5	-1.8
leucine	Leu	2.8	3.8	-1.8
valine	Val	2.6	4.2	-1.5
cysteine	Cys	2	2.5	-1
tryptophan	Trp	1.9	-0.9	-3.4
alanine	Ala	1.6	1.8	-0.5
threonine	Thr	1.2	-0.7	-0.4
glycine	Gly	1	-0.4	0
serine	Ser	0.6	-0.8	0.3
proline	Pro	-0.2	-1.6	0
tyrosine	Tyr	-0.7	-1.3	-2.3
histidine	His	-3	-3.2	-0.5
glutamine	Gln	-4.1	-3.5	0.2
asparagine	Asn	-4.8	-3.5	0.2
glutamic acid	Glu	-8.2	-3.5	3
lysine	Lys	-8.8	-3.9	3
aspartic acid	Asp	-9.2	-3.5	3
arginine	Arg	-12.3	-4.5	3

Three common hydrophobicity scales are used: Engelman-Steiz [5], Kyte-Doolittle [6], and Hopp-Woods [7]. Each scale computes hydrophobic values differently and assigns different values to each of the amino acids [4]. Once hydrophobicity values are assigned to the amino acids, the amino acid content of a protein can be treated as a discrete numeric signal.

A fundamental concept to this study is that the hydrophobic signal of proteins is not random. This idea was tested with the use of artificial data generated using a bootsrap method. Replicating each protein's signal and shuffling the order of amino acids in the sequence produced random data. The control sets maintained the frequency of the amino acid while removing position dependence [8]. This

data was treated using the same methods as the actual data.

Wavelet Packets & Power Spectra

Wavelets are a mathematical tool that allows for the analysis of data in respect to both scale (or frequency) and position (or time). Wavelets are useful in studying proteins because they preserve local information, allowing for the study of the amino acid sequences. By using wavelets one is able to look for patterns that may emerge at varying scale along the sequences.

It is useful to describe the mathematics of wavelets before introducing the wavelets packet technique. Wavelets are constructed using two functions, a scaling function, ϕ, and the corresponding wavelet function ψ, where

$$\phi(x) = \sqrt{2} \sum_k h_k \phi(2x - k) \quad (1)$$

$$\psi(x) = \sqrt{2} \sum_k g_k \phi(2x - k) \quad (2)$$

with coefficients g_k and h_k. The multiresolution analysis offered by wavelets is as follows. The shifted and dilated wavelet and scaling functions are defined as:

$$\phi_{jk}(t) = 2^{\frac{j}{2}} \phi\left(2^j t - k\right) \quad (3)$$

$$\psi_{jk}(t) = 2^{\frac{j}{2}} \psi\left(2^j t - k\right) \quad (4)$$

This allows a function, $f(t)$, to be analyzed by the translation parameter, k, at a scale j via a convolution with ϕ. The convultion produces a function split into two parts. One part is a filtered, coarser version of the original function. The other part captures local fluctuations about the function locally. The multiresolution process occurs when the wavelet is applied to the coarser approximation of the original function. This again results in a splitting into a coarsed and fluctuation of the initial approximation. The process continues until only the average of the original function remains.

Wavelet packets are an evolution of wavelet transformations. The transformation passes a signal through two filters producing average and difference coefficients [8]. Each splitting results in the transform being applied to both the coarse and detail parts of the function. Figure 1 gives a graphical representation of the wavelet packet multiresolution process. The box, *S*, is the original function, A_1 and D_1 are the first approximation and detail coefficients. The wavelet is then applied to both the $A_1$and D_1 coefficients. This results in the next line in the figure. The process then continues similarly. Each box is referred to as a leaf, and the entire structure is called the wavelet packet tree.

Wavelet packets are similar to wavelet transformations in that they both analyze a signal by computing local approximation and difference coefficients. However, wavelet packets will iteratively apply the convoluted function to both sets of data. This process, known as downsampling, is continually applied to the previous sets of smoothed and differenced coefficients, as displayed in Figure 1 [9].

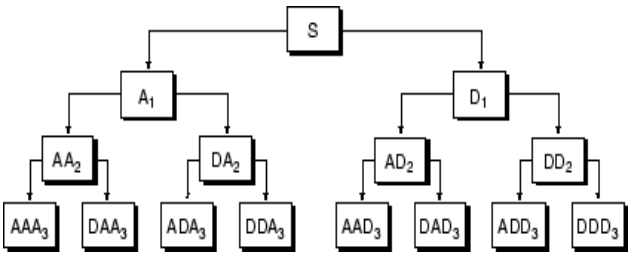

Figure 1. Wavelet packet decomposition levels with detail coefficients (high pass) given by D_m and approximation coefficients (low pass) given by A_m [9].

Wavelet packet fast algorithms produce leaves ordered by Gray code that must be sorted. Once correctly sequenced, the mean value of the coefficients squared was taken of each leaf, reducing a leaf from a frequency band to a single coefficient. This resulted in the production of the power spectrum vector used to describe each protein.

$$p = \begin{pmatrix} < A_{j,k}^2 > \\ < D_{j,k}^2 > \\ \vdots \end{pmatrix} \quad (5)$$

Statistical Analysis

A two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test (KS) was applied to the proteins signals created using the Kyte-Doolittle scales [6]. The KS test checks to see if the power spectra of similar proteins belong to the same underlying population [10]. KS is a nonparametric test comparing cumulative distributions of two data sets. KS looks to determine whether two data sets were sampled from the same distribution; if so, then its null hypothesis cannot be rejected. [10].

The power spectra for each protein were normalized to allow for their comparison. Normalization was performed by finding the quotient of each element in the power series array with the sum of its elements. For a vector p with n elements representing the power spectrum,

$$p_{normalized} = \frac{A_i}{\sum_i^n A_i} \quad (6)$$

Normalization allows for the comparison of power spectra. This comparison was done by taking the mean value of the difference squared of power series corresponding to proteins of similar function. For two power spectra given by vectors P_{meso} and P_{therm} with n elements, τ is given by:

$$\tau = \frac{1}{n}\sum_i^n \left(P_{meso_i} - P_{therm_i}\right)^2 \quad (7)$$

The vector τ is the mean value of the difference-squared coefficients for power spectra of the same transformation and proteins of similar function.

This process was carried for the three different hydrophobicity scales. Also, four different transformation functions were used from the Daubechies family of wavelets, Daubechies 1 (db1), Daubechies 2 (db2), Daubecheis 3 (db3), and Daubechies 4 (db4). Each daubchies transformation function corresponds to different coefficients used for g_k and h_k. This was done to determine the impact of the transformation functions on the signal. The number associated with the transformation function corresponds to the number of vanishing points. A larger number of vanishing points allows for the analysis of more complex signals [9].

Twenty-five sets of randomly distributed hydrophobicity signals were created for each hydrophobicity scale under the db2 transformation. The mean value of all 25 realizations was taken after performing the wavelet packets transformation and creating the power spectra and τ vectors.

The centers of the histograms were determined by finding the mean value of the center of each bin.

RESULTS

The methods listed above allow for the creation of histograms that display the total number of coefficients from τ. Daubechies second wavelet transformation is included below. Transformations performed using db1, db3, and db4 were similar to db2 and offer no additional information relating to the study goals.

Kyte-Doolittle Scale:

The center of the db2 histogram displayed in Figure 2 is located at 0.0128. The mode of this distribution is at 0.0038 with 185 values. The median of this histogram is at 0.0053. The skewedness of the db2 histogram is 1.5273.

The random distribution, shown in Figure 3, has the predictable Gaussian distribution with a few outliers. The center of the Gaussian distribution is at 0.008; its mode is at 0.0072 with 175 values. The skewedness of the random distribution is 1.2357.

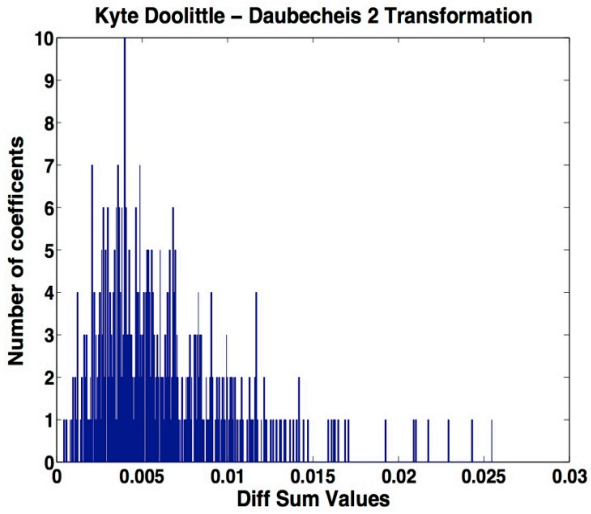

Figure 2. Kyte-Doolittle hydrophobicity scale signals transformed using db2.

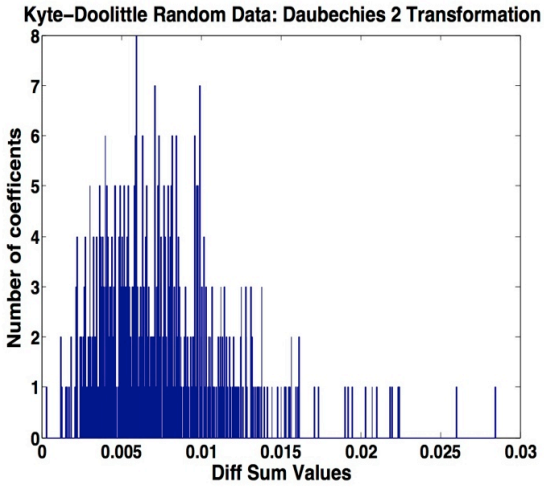

Figure 3. Kyte-Doolittle Scales signals transformed with db2, random sequence.

Hopp-Woods

As shown in Figure 4, the center of the Hopp-Woods db2 histogram is located at 0.0145. The mode is 0.0043 with 202 values. The median is located at 0.0056. All four transformations have skewed right distributions and similar spreads in range. The skewedness of the Hopp-Woods db2 histogram is 1.4760.

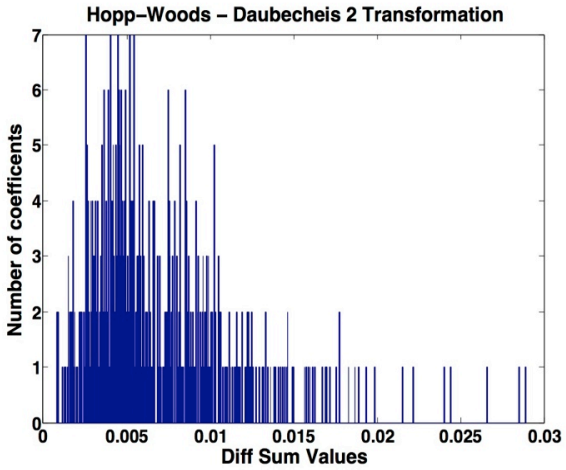

Figure 4. Hopp-Woods hydrophobicity scaled signals with transformation db2.

As shown in Figure 5, the randomly generated sequence histogram has a Gaussian distribution with outlying data near its maximum value 0.0173. The center of this distribution is at 0.0091 with a mode of 0.0045 with 185 values. The skewedness of this histogram is 1.2606.

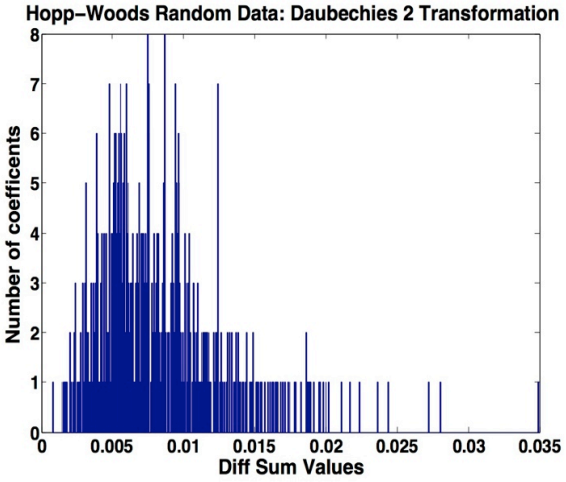

Figure 5. Hopp-Woods db2 random sequence histogram.

Engelman-Steiz

As shown in Figure 6, the center of the Engelman-Steiz db2 histogram is at 0.0119. The mode of the distribution is 0.0012 with 332 values. The median value for this distribution is

at 0.0018. The skewedness of the histogram is 3.1603.

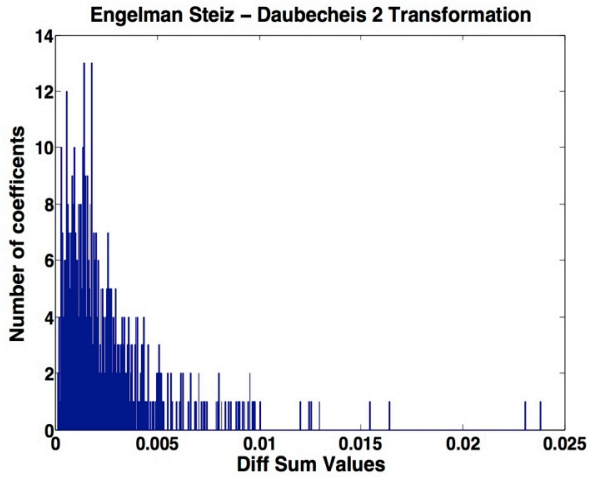

Figure 6. Engelman-Steiz hydrophobicity scaled signal with transformation db2.

Figure 7 shows the randomly generated sequence histogram, which has a skewed right distribution. The center of this distribution is at 0.0081 and has a mode of 0.0024 with 196 values, and median value at 0.0032. The skewedness of this histogram is 1.9400.

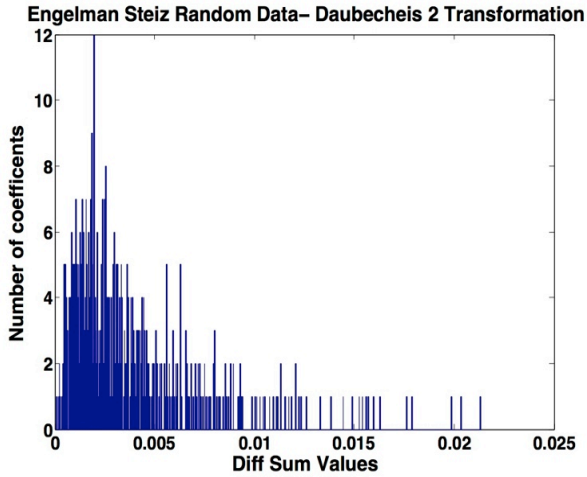

Figure 7. Engelman-Steiz db2 random sequence distribution.

Kolmogorov-Smirnov

The KS test was applied to signals produced using the Kyte-Doolittle, Hopp-Woods, and Engelman-Steiz scales. The power series of mesophile and thermophile proteins were compared.

For Kyte-Doolittle, 527 proteins were unable to reject the null hypothesis, 13 did reject the null hypothesis. Similarly, for Hopp-Woods, 516 proteins were unable to reject the null hypothesis, while 24 rejected the null. For Engelman-Steiz, 516 proteins failed to reject the null hypothesis, while 24 were able to reject it. Only one protein was rejected under all three scales. Hopp-Woods and Engelman-Steiz both failed to reject the null for the same 8 proteins.

DISCUSSION

The goals of this research were to investigate whether it is possible to distinguish proteins using statistical analysis of the amino acid chains and to determine whether the proteins had similar statistical features. This was done by converting the amino acid sequences from the proteins of the two classes of bacteria into signals. By applying experimentally determined hydrophobicity values to each amino acid, a discrete numeric signal is generated.

Daubechies second wavelet transformation function was of interest here, with the other three types used as controls. Little variation was seen across the histograms in respect to frequency when comparing Daubechies 1 through 4 transformations. Because of this, analysis was focused on the data generated using Daubechies second wavelet. Daubechies second wavelet has two vanishing points, which is suitable for this research [8].

The wavelet packets technique is an important tool in analyzing the local information of any signal. In this case, local information is considered to be the frequency bands (the leaves of the wavelet packet tree) within the signal of the amino acid chain as the result of analysis at some scale *j*. The hydrophobic content of any frequency band has been maintained through wavelet packet analysis.

In this study, the magnitude of local hydrophobic information was compared between the two types of bacteria. The varying

magnitudes of hydrophobic content that is associated with scale, *j,* are the statistical features of interest.

To determine these statistical features, the difference was taken between power spectrum coefficients of the two types of bacteria. The difference between the two proteins power spectra was then averaged out, and is included in the τ vector. By taking this difference it is possible to identify frequency bands, in respect to scale, that vary in hydrophobic value between the two proteins.

Each hydrophobicity scale was paired with a corresponding control set of data. These controls were used to test the concept of position dependence in hydrophobic signals. Through bootstrapping, these control sets randomized amino acid order while maintaining frequency.

Kyte-Doolittle

The mode of the Kyte-Doolittle control data was 0.0072 with 175 values. The actual data had a mode of 0.0038 with 185 values. The most frequent data values were a thousandth the size of the values used in the Kyte-Doolittle scale. This does not imply that mesophile proteins had a higher hydrophobic magnitude because the difference between the two coefficients was squared, but it does show that the hydrophobic content is very similar.

The modes of the control and actual data are very similar. The centers of the two data sets, 0.0128 for actual and 0.008 control, and skewedness (1.527 for actual and 1.2357 control) are very similar as well.

Hopp-Woods

The mode of the Hopp-Woods actual data was 0.0043 (202 values) and 0.0045 (185 values) for control. Also the center of the actual data was 0.0145 and 0.0091 for control. The skewedness of the actual data set was 1.4760 and 1.2606 for control.

Engelman-Steiz

The mode of the Engelman-Steiz actual data is 0.0012 (332) and 0.0024 (196 values) for control. The center of the actual data is 0.0119 and 0.0081 for control. The skewedness of the actual data set is 3.1603 and 1.94 for control.

The skewedness of each histogram directly affects its center. All six histograms have a rightward skew. This tendency is in part due to scale shifting that was done before hydrophobicity values were even assigned to amino acids. Each of the three hydrophobicity scales were shifted up by a tenth more than the most negative number to avoid zeros and negatives.

However the skewedness of each graph is interesting because it highlights the influence of frequency bands that have a greater difference in hydrophobic magnitude. Because of this, the mode is considered to provide the best estimate of the variance in the majority of proteins. But the skewedness shows that in some proteins there is a greater difference in hydrophobicity.

The best example of this is in the Engelman-Steiz data sets. With similar modes, the center of the actual data's bin is far larger than the controls. This can be seen in Hopp-Woods (actual center and skew: 0.0145, 1.4760; control center and skew: 0.0091,1.2606) and Kyte-Doolittle (actual center, skew: 0.0128, 1.5273; control center, skew: 0.008, 1.2357) as well.

The majority of frequency bands are very similar between protein pairs in respect to hydrophobic magnitude. However the difference in the centers and its corresponding relationship with skewedness infers that some proteins may have a significantly larger difference in hydrophobic content. Future research may be able to use this data to determine a statistical measure that allows for classificiation of proteins from amino acid sequence alone.

The results of the KS test indicate that the majority of hydrophobicity signals come from similar distributions. By rejecting the null, the KS test states that the two protein's signals come from different distributions. As such, the majority of proteins failed to reject the null.

If the two classes of bacteria were statistically different, then the KS test would have rejected the null more often. However this is not the case, under the Kyte-Doolittle scale only 2% of protein pairs rejected the null hypothesis, while

close to 98% accepted the null. Both Hopp Woods and Engelman-Steiz rejected the null approximately 4% of the time.

Looking at the modes of each histogram, it can be seen that the majority of protein pairs are very similar in hydrophobic content. Additional support for this similarity is seen in the results of the KS test, which show that most proteins in this study likely share a common distribution. However, the differences in the center of actual and control histograms infer that a small portion of proteins have a greater difference in hydrophobic magnitude. The skewedness of these graphs show that outlying data has a large effect on these centers. However, the modes of the actual data histograms fall below their centers, such that the majority of values are comparatively small.

There does not seem to be a difference in the hydrophobic content of proteins with similar biological functions from mesophile and thermophile bacteria. The differences in the centers of the control and actual data histograms imply that amino acid sequencing is fundamental in determining hydrophobic content.

ACKNOWLEDGEMENTS

Funding Provided By DePaul University College of Science and Health, Undergraduate Summer Research Program (USRP). Research Advisor: Dr. Jesus Pando.

REFERENCES

[1] Boundless Biology. *Classification of Microorganisms by Growth Temperature.* WWW Document. (https://www.boundless.com/microbiology/textbooks/boundless-microbiology-textbook/culturing-microorganisms-6/temperature-and-microbial-growth-64/classification-of-microorganisms-by-growth-temperature-388-5509/).

[2] National Center for Biotechnology Information. NCBI. WWW Document. (ncbi.nlm.nih.gov/pubmed/).

[3] Nave, C.R. *Amino Acids*, WWW Document, (hyperphysics.phy-astr.gsu.edu/hbase/organic/amino.html#c1).

[4] Gallik, S. *Amino Acid Hydrophobicity | Cell Biology Olm*, WWW Document, (cellbiologyolm.stevegallik.org/node/32).

[5] Engelman, DM, Steitz, TA, Goldman, A. " Identifying Nonpolar Transbilayer Helicies in Amino Acid Seqeuences of Membrane Proteins". *Am. Rev. Biophyps. Chem.* 1986. 15: 321-353.

[6] Kyte, J, Doolittle, R. "A Simple Method of Displaying the Hydropathic Character of a Protein". *Journal of Molecular Biology*. 1982. 157: 105-132.

[7] Hopp, T and Woods, K.R. "A Computer Program for Predicting Protein Antigenic Determinants". *Mol Immunol*, 1983. 20: 483-489.

[8] Pando, J, Sands, L, and Shaheen, S. "Detection of Protein Secondary Structures via the Discrete Wavelet Transfrom." *Physical Review E* 80.5 (2009).

[9] The Mathworks Inc. *Wavelet Packets*, WWW Document, (mathworks.com/help/wavelets/ug/wavelet-packets.html).

[10] Press, W.H, Teukolsky S.A, Vetterling, W.T. and Flannery, B.P, *Numerical Recipies in Fortran*, 2nd Edition. (Cambridge University Press, 1996) pp. 604-620.