Spring 6-14-2019

# WAVELET ANALYSIS OF SHORT GLOBULAR HOMOLOGOUS PROTEINS IN MESOPHILE AND THERMOPHILE PROKARYOTES

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

## Recommended Citation

# WAVELET ANALYSIS OF SHORT GLOBULAR HOMOLOGOUS PROTEINS IN MESOPHILE AND THERMOPHILE PROKARYOTES

A Thesis

Presented in

Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

June, 2019

BY

John B. Linehan

PHYSICS DEPARTMENT

College of Liberal Arts and Sciences

DePaul University

Chicago, Illinois

## ABSTRACT

This study looked to identify features related to thermal stability and function in the amino acid chains of short globular proteins from mesophile and thermophile species, within the constraint that the protein fold to perform a specific function. To do so 540 homologous pairs of proteins were studied. The amino acid chains were converted to hydrophobicity signals by assigning a hydropathy score to each residue in the polypeptide. The hydrophobicity signals were passed through a wavelet packet transform and the resulting spectra analyzed. Bootstrapping was used to generate a control data set to determine if the true ordering of amino acids codes for a non-random fluctuation in hydropathy along the length of the polypeptide. A method to relate the spectral characteristics to the function of a protein making use of gene ontologies was developed as a proof of concept. As a group, mesophile and thermophile proteins have very similar total power. However, on a protein-to-protein basis the thermophile contains a greater total power in 489 of the 540 pairs (90.56%). The hydrophobicity scale used in this study is strongly correlated with Gibbs free energy. The total power of a protein is also strongly correlated to the Gibbs free energy, so that the thermophile protein contains a greater free energy than its corresponding mesophile partner. It has been noted in the experimental literature that thermophile proteins are stabilized by increasing their Gibbs free energy. The statistical measures skew and kurtosis were adapted so that a spectrum of skew and kurtosis values were generated for each protein. These values indicate that the fluctuation in hydropathy is non random and position dependent. Thermophile proteins have larger power at frequency bands 21 through 31 (average intervals of 100 to 77 amino acids), and 44 to 56 (on average 46 to 19 amino acids), which may contribute to their having greater total power in 90.56% of the pairs. Increases to the fluctuation in hydropathy within certain lengths throughout the total amino acid chain of a protein may be a means of raising the temperature at which a protein denatures.

## TABLE OF CONTENTS

## TABLE OF CONTENTS – *Continued*

## LIST OF FIGURES

## LIST OF FIGURES – *Continued*

## LIST OF FIGURES – *Continued*

## LIST OF FIGURES – *Continued*

## LIST OF FIGURES – *Continued*

## LIST OF FIGURES – *Continued*

## LIST OF TABLES

## LIST OF TABLES – *Continued*

# CHAPTER 1

# Introduction

## 1.1 Life in Extreme Conditions

A protein that maintains its conformation, or structure, through alterations in the temperature of its environment is thermal stable [1]. The focus of this study is to determine how the major driving force behind protein folding, the hydrophobic effect, is encoded into the linear amino acid chain of homolog short globular proteins in organisms with non-overlapping optimal growth temperatures. This problem is interesting since proteins with similar function are expressed in prokaryotic organisms occupying different ecological niches, characterized by drastically different temperature regimes. This study looks to determine the way natural selection has acted on the proteins of the thermophiles to allow for their occupation of ecological niches inaccessible to their mesophile counterparts through a comparative analysis of similarly(homologous) functioning proteins. Determining the relationship between amino acid content, ordering, and the hydrophobic effect can help investigators design proteins that denature at higher temperatures.

Hydrophobicity describes a molecule's response to suspension in a liquid water solution. The molecule acts to change the hydrogen bond network of the surrounding water molecules [2]. If the molecule has neither hydrophobic or hydrophillic qualities, then the surrounding water molecules adapt a cage-like structure around the molecule. If the molecule is hydrophobic, the number of hydrogen bonds shared by each water molecule decreases as the molecule folds over itself, occupying a larger volume [2]. If the molecule is hydrophylic, the average number of hydrogen bonds increases. This process is entropic, and the magnitude of hydrophobic effect is directly related to temperature. The hydrophobic molecule rearranges the surrounding water molecules by adjusting the hydrogen bond network [2].

The hydrophobic effect is temperature dependent, and acts to reduce the number of conformations accessible to the protein [3]. In this way the total number of configurations accessible to the linear amino acid chain is reduced, and the net force of other interactions causes the protein to fold into its native state [3]. These additional forces, excluded in this study, include the electrostatic and Van Der Waals forces as well as hydrogen bonding [3].

Short globular proteins are composed of a single chain of bonded amino acids. This chain folds over itself to reach its native state. The native state is the conformation, or folding pattern, coded by the DNA as a result of natural selection [4]. The proteins used in this study come from two general groupings of organisms: mesophile and thermophile. Mesophile proteins are taken from species of bacteria living in environments with temperatures between 20 and 45 degrees Celsius (68 - 113 degrees Fahrenheit). Thermophile, and hyper-thermophile organisms, members of the Domains Archea and Bacteria, occupy ecological niches characterized by temperatures between 41 and 122 degrees C (105-251 degrees Fahrenheit) [1]. The proteins of mesophile organisms unfold, referred to as denaturing, in the thermophile's environment. The hypothesis of this study is that the net hydrophobic effect, resulting from the hydropathy of each amino acid residue in the polypeptide, is encoded into the linear amino acid chain through the ordering and prevalence of each amino acid. The encoding is thought to reflect both the protein's function and thermal stability.

This study looks to see how the net hydrophobic effect is encoded into proteins with similar function in prokaryotes occupying niches with dichotomous temperature regimes. It may be that these proteins take close to or the same conformation, or entirely different conformations to perform the same function. Through use of the wavelet packet transform and spectral analysis measures, this study looks to identify characteristics in the hydrophobic free energy of the primary chains of short globular proteins.

## 1.2 Protein Overview

The proteins used in this study consist of one long amino acid chain that folds over itself to achieve a native conformation. Proteins are long strands of amino acids linked by bonds formed between the amide group of one amino acid with the carboxyl group of another. There are twenty amino acids, each composed of an amide group (-CONH2) and a carboxyl group (COOH) [5]. The distinguishing feature for these molecules is its residue (R), this feature dictates the amino acid's response in an aqueous solution. The residues are used to group amino acids into five categories: the polar group (uncharged R group), negatively charged group, non-polar aliphatic group, the aromatic group, and the positively charged group [5]. Figure(1) provides illustrations of each amino acid's chemical structure. The amino acids in this figure are broken up by group, and the reisude of each amino acid is highlighted in pink [5]. The residue is the distinguishing feature of the amino acid molecule.

A protein's conformation is a complex pattern of folds and twists that can involve interactions with other proteins and molecules expressed inside a cell. In this study we focus on short globular proteins consisting of a single primary chain of amino acids under 600 residues in length. The study ignores the possibility of interactions that may occur beyond those of the primary chain with itself and a liquid water solution. Protein structure is categorized in four tiers: the primary, secondary, tertiary and quaternary. The primary structure consists of the indexed order of amino acids from the n-terminal to c-terminal ends of the polypeptide. Proteins are long chains of amino acids that bind to one another from carboxyl group to amide group. The polypeptide chain starts with an exposed amide group and ends with an exposed carboxyl group. This is the level of interest in this study. The secondary level considers the folds and twists the primary chain makes. This level contains two major features, alpha helixes and beta sheets. The alpha helix consists of the primary chain coiling over itself to resemble a helix (like that of Euckaryote DNA). If the primary chain folds itself into a sheet, so that the primary chain is folded into

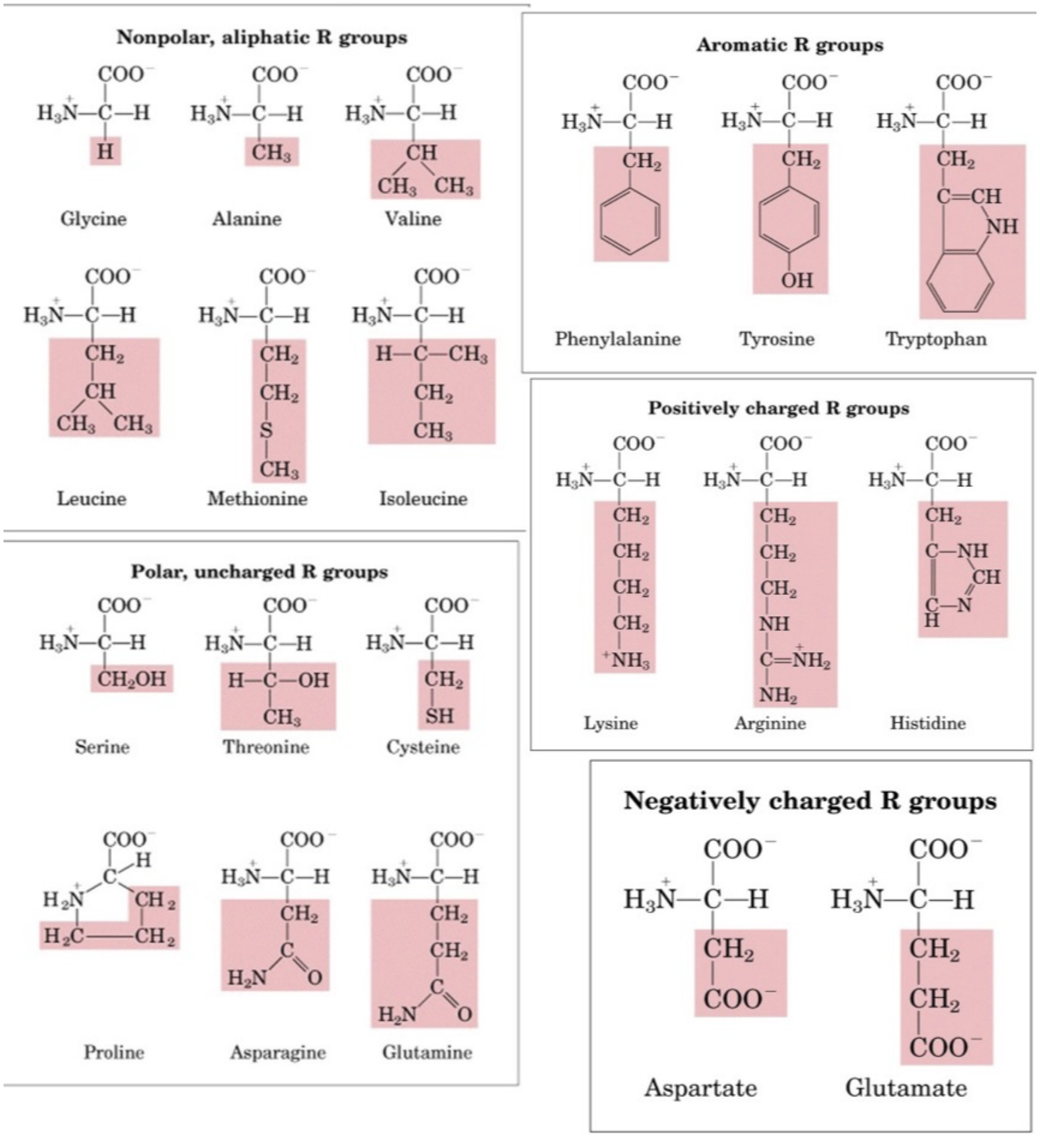


Figure 1.1: Structure of 20 amino acids. Includes the residue highlighted in pink. Broken up into: polar group, negatively charged group, non-polar aliphatic group, aromatic group, and positively charged group [5].

similar lengths lying along a plane, it is in a beta-sheet.

The tertiary structure consists of the protein in its native state, the conformation consisting of a three dimensional folded pattern containing within it the secondary structure. The quaternary structure describes the shape taken by the protein when its native structure interacts with other primary chains in their native state. In this way the protein is made up of multiple primary chains, referred to as sub-units. Proteins with these complex structures are excluded from this study.

## 1.3 Natural Selection and the Thermodynamic Hypothesis

Proteins are not randomly generated molecules. They are acted on by evolution through natural selection to fold into a specific conformation that determines the protein's function. The term native structure, or native state, describes the pattern of twists, turns, and folds a polypeptide takes that result in its ability to perform useful activity. This state is only reached when the polypeptide is expressed in the solution in which natural selection has acted on it. This concept is referred to as the thermodynamic hypothesis [4]. The conformation of the protein acts as a constraint for evolutionary processes. Mutations can increase interactions within the chain resulting in increased internal packing. If mutations made in the genome result in a protein that folds improperly, then this molecule is denatured, and does not result in a phenotype on which the environment can act.

In this study a temperature adaptation is a characteristic that raises the denaturing temperature of a protein. This characteristic must be encoded by the DNA, and thus is an evolutionary adaptation. Such adaptations allow a prokaryote to occupy niches that require a higher optimal growth temperature, above 41 degrees celsius. Temperature is considered to be a selective measure that either allows for or restricts the development of microbial communities by acting on the proteins within these organisms. In addition to this, individual organisms can implement various molecular mechanisms in response to fluctuations in the temperature of the environment [1]. These molecular mechanisms are crucial for the organism to maintain a lifestyle suited to high temperature environments, and go hand in hand with the thermal adaptations made to their proteins [1].

Temperature adaptations go beyond thermo-tolerant proteins. For an organism to live in any environment it must be able to handle changes in temperature. These adaptations are critical for maintaining the cellular membrane and in preventing denaturation of DNA [1]. Mechanisms common to both mesophiles and thermophiles include the expression of heat shock proteins. These proteins are termed chaperones and aid in the folding of proteins immediately after transcription, and rescue denatured proteins after activity or in conditions stressful to the organism. These proteins consist of large complexes made up of multiple subunits lending to their stability in times of stress [1].

To maintain the integrity of the membrane a single-celled organism will increase the expression of branched fatty acids. The cell membrane is a lipid bilayer populated by transmembrane proteins and lipid rafts that help the cell communicate with its environment and exchange materials. To handle rising temperature bacteria will alter the make up of the membrane by increasing production of branched fatty acids [1][6][7]. The extent to which a bacteria can survive alterations to the temperature of its environment depends on the types of branched fatty acids it is able to produce. This makes it possible to identify differences in the make up of mesophile and thermophile cell membranes, and can help to determine those fatty acids that are produced by organisms that have adapted to life in high temperatures [1][6][7].

The ability to both protect and repair DNA is necessary and apparent in all domains of life. However demand is increased in thermophiles, since DNA and RNA structure are temperature dependent [1]. To repair damaged DNA organisms utilize homologous recombination [1]. This process is used to repair breaks in the DNA by swapping nucleotide sequences between identical strands. Recombination events are measured in terms of the ratio of nucleotide changes introduced by recombination to point mutations. Higher levels of homologous recombination, ranging from 24 - 100 nucleotide changes to point mutations, have been observed in the thermophile genus *Thermotoga* [1][8]. Measures made in mesophiles range from 0.02 - 64 nucleotide changes to point mutations. Generally instances over 10 for this measure are considered to be high [1]. Higher rates of recombination have been observed in thermophiles *Pyrococcus furiosus* [8], *Sulfolobus islandicus* [9], and *Persephonella* [10][1]. This is evidence that thermophiles have increased rates of homologous recombination allowing for survival in higher temperature regimes.

Protective methods include raising the melting temperature of DNA through the expression of thermostable proteins that bind to DNA [1]. This process has been observed in *Thermococcus Kodakaraensis* to raise the melting temperature of DNA by 20°C [1][11]. Observations of *Thermotoga* species has shown that concentrations of the polyamines (molecules containing an amide group at both ends) caldopentamine and caldohexamine increase with temperature [1][12]. In addition, it has been found that the expression of the reverse gyrase protein, which binds to DNA, is predominately expressed in thermophile organisms. Knockout studies of this gene show that growth rates are inhibited for knockout genotypes as compared to wild type in *Thermococcus Kodakaraensis.* The study showed that knockout *Thermococcus Kodakaraensis* were unable to grow at temperatures above 90°C, where wild types exhibit continued growth [1][11].

Since the thermophiles are living in environments of extreme temperature, the entirety of their proteome (those proteins expressed by the organism), is thermally stable at high temperatures. Adaptations to thermal stability have been made at all levels or protein structure and are specific to the function of the protein [1]. The proteins of mesophile and thermophile prokarytoes overall tend to be very similar from analysis of the alignment of their chains. It is thought that thermophiles tend to have a larger percentage of their primary chain composed of aspartic acid, glutamic acid, lysine and arginine and decreased numbers of asparagine, glutamine, serine, and threonine [1]. It is believed that this may show that ionic interactions between amino acids contribute more to stability at higher temperatures than hydrogen bonding [1]. At the secondary level, thermostable proteins tend to more $\alpha$-helices than mesophiles. At the tertiary level, thermophile proteins appear to be more tightly packed, with reduced entropy of unfolding [1].

Additionally, it has been noted that proteins most affected by thermal adaptation are involved with transcription factors for chaperones and other stabilizing molecules. A transcription factor is a protein that controls the expression of a gene. This has been shown by lowering the optimal growth temperature of *Bacillus subtilis* 168 by replacing its copy of the *groEL*, which codes for a chaperone, with a copy from another extremophile living in mesophilic conditions [1][13]. As well, inserting a heat shock protein gene from the thermophile species *Caenorhabditis elegans* into *E. Coli* increased the range of its growth temperature 3°C [1][14].

Prokaryotes have adapted a variety of mechanisms to handle fluctuations in the temperature of their environment and other forms of stress. Although these mechanisms are expressed in both mesophiles and thermophiles, alterations have been made that reflect the conditions of their respective niches. For thermophiles to live they must be totally adapted to living in these extreme conditions. By comparing similar mechanisms, for instance the composition of the cell membranes, between groups further insights can be made into adaptations that result in these organisms occupying these niches.

## 1.4 Hydrophobic Effect

The hydrophobic effect is commonly used to described a molecules affinity for water. Either the molecule is attracted to (hydrophilic) or repelled from (hydrophobic) water molecules. Solutes that are polar and contain charged groups will be attracted to water, and apolar solutes (even distribution of charges, so that there is no polarity) will be repelled from water [2].

The mechanism behind the hydrophobic effect consists of two physical components. The water molecules surrounding the molecule, or solvent, must be close to their phase coexistence with vapor[2]. This has to do with the spacing between water molecules and the likelihood of a hydrogen bond forming between any two molecules. The second is that the water molecules are more attracted to one another than they are to the solute[2].

A large hydrophobic molecule will induce an interface around itself by reorienting the surrounding water molecules[2]. The solvation free energy of this interface is proportional to the surface area of the molecule. Solvation free energy is a macroscopic measure of the free energy required to transform a system from one state to another[2]. The free energy is the amount of work the system did to reorient the water molecules. The solvation free energy is calculated as the change in the Gibbs free energy between two states, where the partition function $Z$ is introduced to quantify the effects of molecular interactions between water molecules and solute

$$\Delta G = G_2 - G_1 = -k_B T \ln(Z_2/Z_1) = -k_B T \ln < \exp(-\Delta E/k_b T) >_1 \approx < \Delta E >_1 \text{ [2].} \tag{1.1}$$

The $\Delta E$ term is the difference in the energies of macro-states 1 and 2 where $< ... >_1$ indicates the equilibrium ensemble average over the micro-states of macro-state 1 [2]. The Boltzmann weight $\exp(-E/k_B T)$ is proportionately equivalent to the probability that a system is in a state with energy $E$ in thermal equilibrium[2]. The term $k_B T$ has units of energy and is used to compare quantities as being large or small. A system that has similar energy levels across its micro-states will have a large entropy, this is indicated by small values for $\Delta E$ and $< \Delta E >$. Since each micro-state has a similar energy, there are many ways to configure the micro-states for a given macro-state. $Z$, the partition function, is computed for a macro-state by summing over all micro-states as

$$Z = \sum_{i=1}^{n} \exp(-E_i/k_B T). \tag{1.2}$$

In this way the solute rearranges the surrounding water molecules into an interface [2].

Water molecules can form a maximum of four hydrogen bonds. In an ice state, water molecules will have four hydrogen bonds. In liquid state, the average number of hydrogen bonds drops to 3.4. Around a solute, water will adapt a cage like structure, and the number of hydrogen bonds will change depending on the nature of the molecule, or in the case of proteins the volume of its conformation [2]. Therefore the water molecules surrounding a hydrophobic molecule are close to their phase coexistence with vapor, since the average number of hydrogen bonds of the water molecules in the interface is less than 3.4 [2].

## 1.5 Hydrophobicity Scales

The hydrophobicity scale used in this study comes from the work of Kyte and Doolittle [15]. This scale was determined using a computational method calculating the average hydropathy of a moving segment along a protein from the n to c terminal ends. These values were compared with the structures of those proteins to determine if the average hydropathy of each section calculated agreed with data. For instance, Kyte and Doolittle checked to see if a segment that was calculated to be hydrophilic matched with structural data showing the segment to reside on the exterior of the protein. The hydropathy of each amino acid was determined by taking the average of three different scales. Two that used structural information, and one based off the calculation of free energy [15]. The authors of the hydrophobicity scale made adjustments to one of these scales, as well as their own calculated values. The method and reasoning they followed is detailed below.

Hydrophobicity scales attempt to quantify the hydropathy, the affinity for water, of an amino acid residue. Proteins are amphipathic molecules, meaning that they have both hydrophobic and hydrophilic regions. Hydrophobicity scales are generated off the assumption that the hydrophobic regions of the protein are found on the interior of its native structure, while hydrophilic regions will reside on the exterior [15]. Three methods are described to determine these quantities. One of which involve the calculation of the hydrophobic free energy of an amino acid [15].

The hydrophobic free energy for an amino acid is determined by calculating the work done to move the amino acid from water to another solution that has similar properties to the interior of a folded protein. It is determined by the free energy, equations 1.1, 1.2, 1.3. The first method uses ethanol as a solvent representing the interior of a protein. Kyte and Doolittle were critical of this approach. To calculate the free energy one must consider ethanol to be a neutral (pH of 7), non-interacting media (no internal forces due to interactions between particles). Kyte and Doolittle did not feel that this criterion can be met by any substance and that interactions between particles in the media will affect the hydrophobic free energies calculated. This will prevent them from being an adequate measure of the hydropathy of each amino acid, because these values will include the energy of the two solutions, water and ethanol, mixing together [15].

The second method instead uses the transfer free energy for water from liquid to vapor. This allows for the elimination of energy due to mixing, and is a simpler way of computing the hydrophobic free energy. The energy added from the mixing of the two solutes is computed as the product of the increased entropy and temperature. To perform this calculation Kyte and Doolittle note the following. The transfer of the residue between states has to occur at standard pressure and temperature. This is done so that the entropy of mixing between states can be calculated. The amino acid must be moving between equal volumes, so that the only interactions considered in the free energy coefficient are due to the amino acid and the water. To account for this in their hydrophobicity scale, Kyte and Doolittle made a correction to the work done by Wolfden *et al* 1979 in computing the transfer free energy for fourteen of the amino acids' residues [15]. The correction consists of the following adjustment. The free energy should be calculated from the relation of changing volume of an ideal as

$$\Delta G = -RT\ln(V_2/V_1). \tag{1.3}$$

The formula used to implement this correction is

$$\Delta G = -RT\ln(\frac{N_w V_g}{N_g \phi}) = -RT\ln(\frac{18.07\gamma}{\phi}). \tag{1.4}$$

$N_w$ is the equilibrium mole fraction in aqueous phase, $N_g$ is the equilibrium mole fraction in vapor phase, and $V_g$ is the volume of the vapor phase, at standard temperature and pressure [15]. $\gamma$ is the partition coefficient for that residue, and

$\phi$ is the apparent molar volume, or the volume of the amino acid residue given the condition of standard temperature and pressure. Kyte and Doolittle remark that this equation for the given states allows for the calculation of free energy exclusively for the amino acid residue and the water [15].

The third method looks at the actual position of each amino acid in the three dimensional structure of a protein. This method utilizes the known structures of proteins by taking the atomic coordinate of each amino acid in the structure. It checks to see if any amino acid appears to have a preference for location, either inside the conformation or on the outside. Those amino acids that are most likely found on the exposed regions of a protein are considered to be hydrophilic, while those amino acids found on the interior of the protein are considered hydrophobic [15].

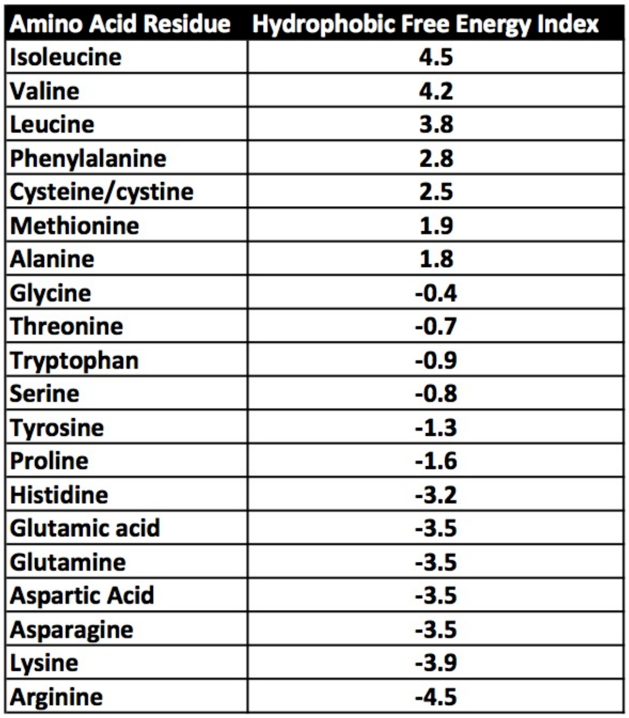
Chothia determined the ensemble average of amino acid location in 12 globular proteins in 1976 [16]. These values are presented as the fraction of amino acids that are more than the 95% and 100% buried. Meaning that for the total number of occurrences of a given amino acid from n to c terminal ends of the protein, if the amino acid is on the outside facing the solution, subtract one, if its facing another amino acid add 1. Divide that number by the total number of amino acids that are buried, for the fraction that are 100% buried, and by 95% for that fractional percentage. If a specific amino acid occurs mostly on the exterior of the conformation, this value will be negative, since we've driven that running sum into the negative values. This indicates that the amino acid is hydrophilic. Large positive numbers indicate that this amino acid might be forcing the proteins conformation to fold inwards [15].

Kyte and Doolittle note that the hydropathy character of an amino acid is far less impacted by the tenths and one hundredths places, than the ones place. Because of this, Kyte and Doolittle made adjustments to their scale subjectively. The values for valine, phenylalanine, threonine, serine, and histidine were taken as the average value for the hydrophobic free energy, 95% and 100% buried values. Essentially they were left alone after calculation. Kyte and Doolittle decided that glutamic acid, aspartic acid, asparagine, and glutamine all had the same hydropathic character, and determined their values by taking the the average of all their values across each scale. Tryptophan was assigned its values based off its normalized free energy due to lack of structural information. Glycine's value was determined by taking the weighted value of all sequences measured in the data set, they felt it is neither hydrophobic or hydrophylic. The hydropathy value for alanine was lowered, because it was originally determined to be more hydrophobic than leucine. This decision was made because leucine has four methyl groups and alanine has only one, Kyte and Doolittle felt that leucine should be more hydrophobic because it has more methyl groups. The value assigned to alanine is half the average of glycine and the original value determined for alanine. For reference, lysine has a hydropathy value of 3.8, glycine -0.4, and alanine 1.8. Proline was made more hydrophobic than its original value because it has 3 methyl groups. Arginine was made the lowest point on the scale. The values for tyrosine and leucine were negligibly raised, and the value for lysine was lowered in the same way [15].

The Kyte-Doolittle hydropathy index is presented in figure 1.2. This index assigns a value for the tendency of each amino acid to be either attracted to or repelled from water. This scale is used to generate hydrophobicity signals for the linear amino acid chains of homologous short globular proteins in mesophile and thermophile proteins [15].

It should be noted that to date there are over 90 hydrophobicity scales available in the literature [17]. The scale used in this study was published in 1982 and since then, new experimental and computational methods have been refined and developed. These more modern scales have made use of advances in x-ray crystallography, molecular dynamic simulations (*i.e.* increased computational power), and Grid Inhomogenous Solvation Theory to name a few. However they have a tendency to be designed around a very specific criterion, structure identification, for instance finding $\alpha$-helicies and $\beta$-sheets within the primary structure [17].

A hydrophobicity scale's effectiveness at identifying secondary structure determines whether the authors of those scales make additional refinements to the values



| Amino Acid Residue | Hydrophobic Free Energy Index |
|---|---|
| Isoleucine | 4.5 |
| Valine | 4.2 |
| Leucine | 3.8 |
| Phenylalanine | 2.8 |
| Cysteine/cystine | 2.5 |
| Methionine | 1.9 |
| Alanine | 1.8 |
| Glycine | -0.4 |
| Threonine | -0.7 |
| Tryptophan | -0.9 |
| Serine | -0.8 |
| Tyrosine | -1.3 |
| Proline | -1.6 |
| Histidine | -3.2 |
| Glutamic acid | -3.5 |
| Glutamine | -3.5 |
| Aspartic Acid | -3.5 |
| Asparagine | -3.5 |
| Lysine | -3.9 |
| Arginine | -4.5 |

Figure 1.2: Kyte-Doolittle Hydrophobic Free Energy Scale

they come up with. This study is not focused on developing a method to identify secondary or tertiary structures from a protein's amino acid chain. Neither is it focused on generating a molecular dynamics simulation to understand folding pathways. This study is designed to be a comparison between homologous proteins with well-separated melting points to identify evolutionary adaptations in the hydrophobic free energy of a protein that raises its melting point within the constraint that it must fold to perform a specific function. As such, the Kyte-Doolittle hydropathy index takes into consideration those factors most important to this study. Even more importantly it presents a line of clear reasoning I can follow to interpret my data.

# CHAPTER 2

# Methods

## 2.1 Introduction

This project's focus is to identify aspects of the linear amino acid chain of short globular proteins related to thermal stability and function. To do so, a method has been developed that uses the amino acid chains of homolog short globular proteins from prokaryotic organisms occupying ecological niches with different temperature regimes. The primary chains of both organisms are converted to hydrophobicity signals through a mapping using the Kyte-Doolittle hydropathy index that scores the hydrophobic tendency of each residue. The proteins with similar function, one from mesophile and the other from thermophile, are considered to be a homologous pair.

The first step in this process is to develop a measure to classify one of the members of a homologous pair as a thermophile, a protein expressed by an organism living in an environment with a temperature regime between 41 and 122 degrees Celsius [1]. This measure is a *pair specific* and *distinguishing* measure. *Pair specific* in this study is used to describe features of a specific homologous pair. *Distinguishing* is a measure that is different for each of the members of a homologous pair.

Next is to identify features common to both hydrophobicity signals within the pair. These features are considered to be *pair specific* and *similar*, where *similar* is a measure whose value is about the same between the members of a pair. To determine if the features identified in this study thought to be related to the function of a protein are *function specific* all pairs found to have the same *pair specific* and *similar* features are grouped. A gene ontology is then performed on these groups to determine if they carry out similar functions. *Function Specific* is a term used in this study to describe those features found at the *pair specific* level that are *similar*.

A pair specific similar measure is only function specific if it can be found through a literature search using gene ontology that the pairs grouped by this measure have similar molecular function. Finally a *thermal specific* feature is one that can be found to have a similar value in one of the two temperature classifications used in this study (mesophile and thermophile). This measure is pair specific and distinguishing.

To achieve these goals an algorithm has been developed considering both the amino acid content of a protein, and the ordering of amino acids within the chain (or hydrophobic content of each residue within the hydrophobicity signal, and the ordering of the hydrophobic values within the signal). This project began with hypothesis tests, scatter plots of average hydrophobicity, and the alignments of proteins in a pair. It was determined that these methods are insufficient to classify the temperature regime in which the proteins were expressed. Also, they did not provide information related to the function of the proteins in a pair in comparison to other pairs within the data set.

The next step implemented a wavelet packet transformation of the hydrophobicity signals. Earlier work was carried out evaluating a number of wavelet variants, in which the study concluded that the wavelet packet transformation presented the best representation of localized hydrophobic content for the aims of this study. Spectral analysis methods are then applied to the wavelet decomposed hydrophobicity signals. An earlier study was carried out to develop a pair specific distinguishing feature, but the measure presented in that study failed to distinguish thermophile from mesophile.

## 2.2 Alignments

A first test of similarity is to compare the alignment of the amino acids in each chain of the proteins composing a homologous pair. For every location in the sequence we check to see if the same amino acid occupies the corresponding location in the other chain. Because the lengths of the proteins are often different, we look for sequence alignment within the length of the shorter of the two proteins. Alignment is computed as a percentage, for each location in the sequence where there is a matching amino acid, a score of one is given, while a mismatch is given a score of zero. The total score is computed as a sum of all the matches, and divided by the length of the shorter of the two proteins.

## 2.3 Hypothesis Testing

Initial work evaluated several traditional methods of statistical analysis to compare the hydrophobicity signals of the homologous pairs. This included using statistical hypothesis tests to establish a baseline of general inferences. Tests included scatter plots of the ratio of mean hydrophobicity, the Anderson-Darling test of normality [22], and the two sample Kolmogorov-Smirnov test [18].

A quick measure of information content in the hydrophobicity signals is to find the average hydrophobicity in each signal within a pair. These values were then plotted against one another to make a scatter plot. The spread of values in this plot can be used to qualitatively asses the correlation between the two values throughout the entire data set. Homologous pairs with similar values will be clustered in the scatter plot. Because the mean value is so often close to zero, the ratio of mean hydropathy of mesophile and thermophile is excluded.

Figures 2.1 and 2.2 provide examples of typical scatter plots. The usefulness of a scatter plot is that it allows for a comparison of the distribution of two values in comparison to one another. Figure 2.1 provides an example of the spread of values for two normally distributed values $x$ and $y$. The number density of values decreases as you move away from zero in any direction. The number density in the scatter plot decreases with the percentiles of the normal distributions used to generate this figure. Sixty eight percent of values in a normal distribution fall within one standard deviation on either side of the expected value. The percentage of values two standard deviations away is twenty six percent. Thus the number of values either more negative or more positive than expected decreases moving away from the center of the cluster.

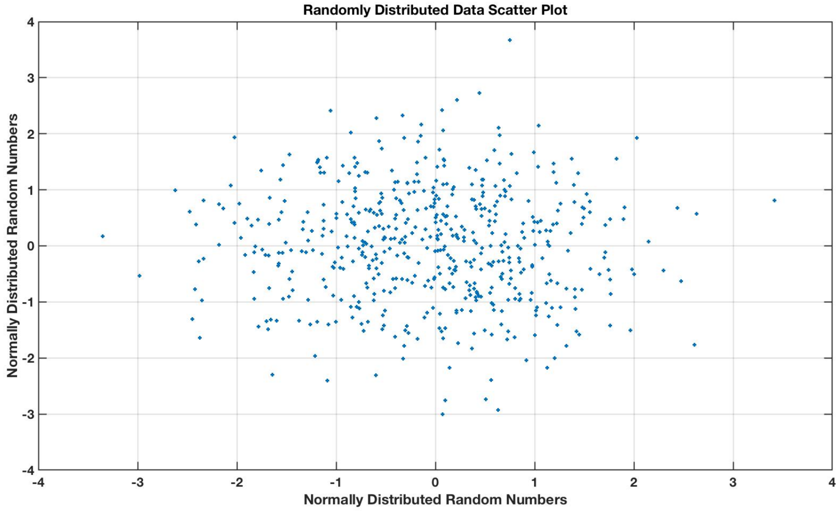


Figure 2.1: Scatter plots are useful tools to determine the spread of sets of values. This figure presents a scatter plot of two normally distributed values. Scatter plots emphasize the spacing between data points. Points closer together are more likely related. In this case the values are normally distributed around zero.

Figure 2.2 provides an example of clusters. This figure was produced by generating three different normal distributions for both $x$ and $y$. The spread of values in $x$ is determined by the standard deviation of the normal distribution. The spread in $y$ is determined by that variables normal distribution's standard deviation. Scatter plots allow for a qualitative analysis of the spread of values, their center, and if sections of the data have similar values. In this example it can be seen that the data appears in three groupings. These groupings have values dispersed around a central value given by the mean of both the $x$ and $y$ normal distribution used to generate them. The spacing between groups shows that they were generated by different processes, in this case, by normal distributions with different means and standard deviations. The statistical parameters used to describe these distributions are the same, but their values are different.

The Anderson Darling (AD) test is a hypothesis test used to determine if a set of values are normally distributed [22]. For normally distributed values, the difference in the ordered values from smallest to largest is approximately uniform.

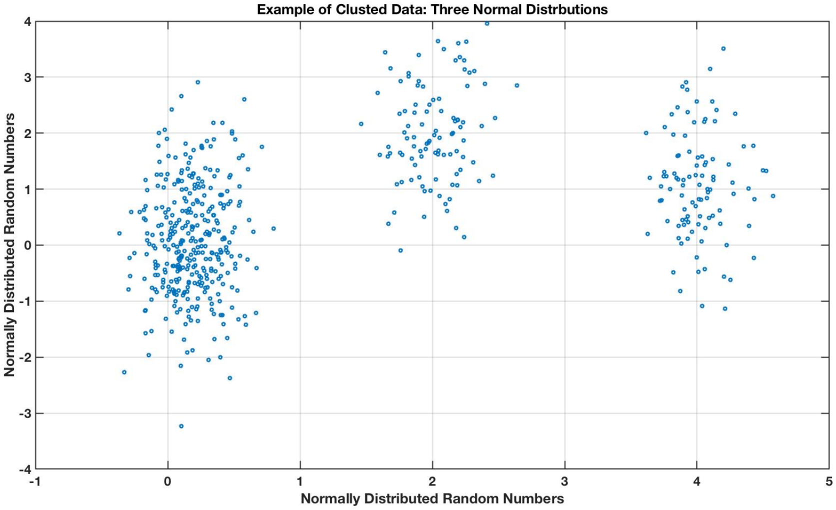


Figure 2.2: This scatter plot presents an example of clustered data. Each of the three clusters were generated from normal distributions with different means and standard deviations. Scatter plots are useful to determine the number of groupings that might exist in a data set. This plot was generated using six separate normal distributions, three for the $x$ variable and 3 for the $y$.

By assuming the test data comes from a normal distribution, the AD test is able to measure the distance between consecutive values of the cumulative distribution function for a data set. The test statistic is calculated as

$$A_n^2 = -n - \Sigma_{i=1}^{n} \frac{2i-1}{n}[\ln(F(X_i)) + (1 - F(X_{n+1-i}))] \tag{2.1}$$

where $n$ is the number of values in the set, $F(X_i)$ is the cumulative distribution function evaluated at sample $X_i$, and $X$ is the set of values sorted from lowest to highest [23]. The decision to reject or accept the null hypothesis is determined by comparing a $p$-value to a significance level (taken to be 0.05). The first step to determine the $p$-value is to compute a test statistic

$$D = A(1 + \frac{.75}{n} + 2.25n) \tag{2.2}$$

with the assumption that the values in the data set are normally distributed [23]. The value of $D$ determines the way the $p$-value is computed. Table 2.1 presents

| D | p-value |
|---|---|
| D $\leq$.2 | 1 - exp $(-13.436 + 101.140D - 223.73D^2)$ |
| .2 < D $\leq$ .34 | 1 - exp $(-8.318 + 42.796D - 59.938D^2)$ |
| .34< D < .6 | exp $(0.9177 - 4.279D - 1.38D^2)$ |
| D $\geq$ .6 | exp $(1.2937 - 5.709D + 0.0186D^2)$ |

Table 2.1: Calculations of $p$-value given some value of $D$. The Anderson Darling test of normality assumes that a set of values is normally distributed. The alternative hypothesis is that the data are not normally distributed [23].

the different equations used to compute the $p$-value for the AD test based of the D test statistic. The method of computing the $p$-value depends on the value of $D$. Once the $p$-value is computed it is compared to a specific significance level (0.05). If $p$ is greater than 0.05 the null hypothesis that the data is normally distributed is accepted [23].

The two sample Kolmogorov-Smirnov takes two sets of values and determines whether or not they come from the same underlying distribution [18]. To do this the test computes the empirical distribution functions for each of the data sets. The empirical distribution function is a function that collects a running tab on the number of values less than or equal to some value in the data set [24]. For example take a set of values Y = $Y_i$,...,$Y_n$. Compute

$$E_n = \frac{n_i}{N} \tag{2.3}$$

where $n_i$ is the number of elements of $Y$ less than or equal to $Y_i$. The empirical distribution is a step function, so that $E(i+1) > E(i)$. Perform the same routine for a set of values X = $X_i$, ... , $X_n$ to compute the empirical distribution function for that data set [24]. Now find the distance between the empirical distribution functions for $Y$ and $X$ as

$$D = max|E_Y(i) - E_X(i)| \tag{2.4}$$

to determine the distance between the values of the empirical distribution functions for $X$ and $Y$. The absolute difference in the empirical distribution functions of each

| $N(0,1)$ | $Gamma(1,1)$ | $E_n(i)$ | $E_g(i)$ | $\lvert E_n(i) - E_g(i)\rvert$ |
|---|---|---|---|---|
| 2.02 | 0.00 | 1 | 0.1 | 0.9 |
| 0.68 | 1.84 | 0.8 | 0.8 | 0 |
| 1.02 | 0.35 | 0.9 | 0.5 | 0.4 |
| 0.25 | 0.19 | 0.7 | 0.3 | 0.4 |
| −1.01 | 0.23 | 0.1 | 0.4 | 0.3 |
| 0.01 | 0.14 | 0.5 | 0.2 | 0.3 |
| −0.69 | 1.11 | 0.2 | 0.7 | 0.5 |
| −0.06 | 0.45 | 0.4 | 0.6 | 0.2 |
| 0.05 | 1.99 | 0.6 | 1.0 | 0.4 |
| −0.39 | 1.92 | 0.3 | 0.9 | 0.6 |

Table 2.2: Two sample KS Test example. Column 1 is normal distribution, column 2 is gamma distribution. Empirical distribution functions for both sets are calculated (column 3 and 4) and the absolute difference taken (column 5).

of the two data sets is the test statistic. The $p$-value for the two sample Kolmogorov-Smirnov test is the measure D which is compared to a significance level, $\alpha = 0.01$. If D is greater than $\alpha$ the null hypothesis that the two data sets have the same underlying distribution is rejected [24].

For instance consider the following two sets of data generated from two different distribution functions presented in Table 2.2. The first comes from a normal distribution centered about zero with a standard deviation of 1. The second was generated using a gamma random number generator with shape parameter one and scale parameter one. The steps involved in calculating the two sample KS test are provided in each column of the table. The first column shows the normally distributed values. The second column shows the gamma distributed values. The third and fourth are the associated empirical distribution functions. The fifth is the absolute difference in the empirical distribution functions for each set of values. The maximum value of the absolute difference in empirical distribution functions 0.9, well above the significance level of $\alpha = 0.01$. Thus we reject the null hypothesis that the two sets of values have the same underlying distribution.

## 2.4 Signals and their Analysis

The proteins in this study are represented as hydrophobicity signals. This means that for every amino acid in a polypeptide a signal was generated using the Kyte-Doolittle scale. Each amino acid is represented by a numeric value from the Kyte-Doolittle hydrophobicity scale rather than an abbreviation. It is this mapping that creates a discrete signal. A signal is a recording of events that occur over some period of time, known as a time series. For instance, a song consists of specific notes played in an exact sequence over some period of time, which causes the listener to react. Thus a piece of music is a signal with localized frequency information. The hydrophobicity signals considered in this study are analogous to electronic signals. The specific ordering of hydrophobicity values mimics the way time series information is recorded in a manner similar to reading values off of a voltmeter, displayed on an oscilloscope, or produced by a speaker. It is the index ordering of values in a sequence that is considered to be the time component in the hydrophobicity signals.

Since proteins are amphipathic molecules, containing both hydrophobic and hydrophilic regions, it would be useful to determine localized hydropathy frequency components. This study makes use of the wavelet transformation to do this. Wavelets have been developed from other methods of signal analysis such as the Fourier transformation. With the Fourier method, frequency components can be extracted from a signal. A Fourier analysis of some piece of music could tell you how often each note is played but does not provide any information regarding when the notes were played. This changes the meaning of the signal, by only seeing the amount of each frequency component, information regarding when each note is played is lost. Two pieces of music may contain the same frequency components but elicit different emotional responses in the audience based of the ordering of those notes. This is the difference between hearing Don't Stop believing by Journey and Barbie Girl by Aqua.

For instance consider the signal presented in Figure 2.3. This signal is the combination of frequency components given by

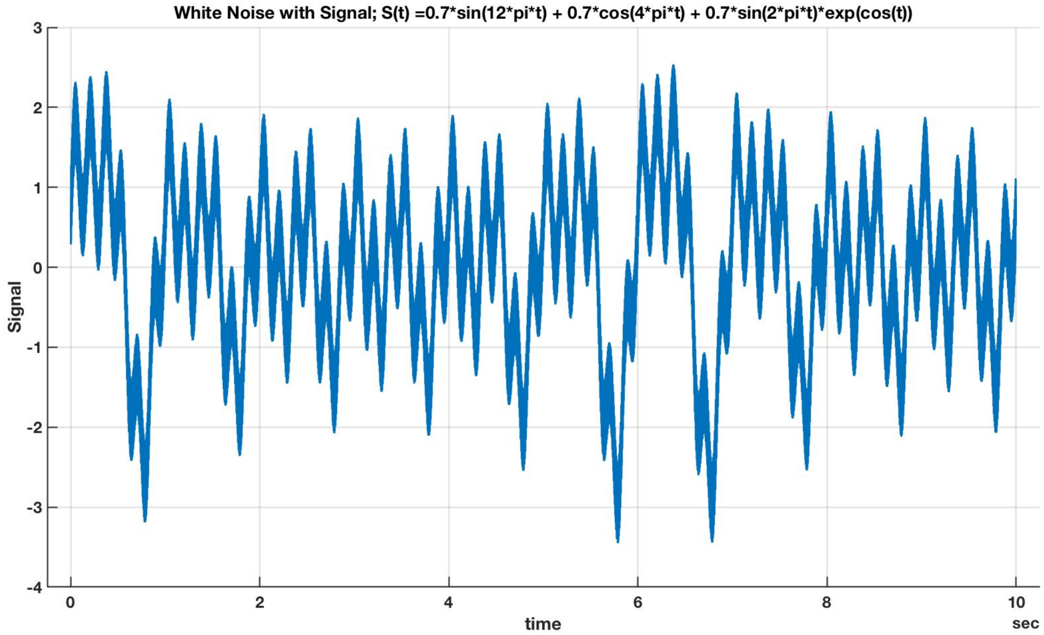


Figure 2.3: This figure displays a signal made up of a single function $S$ with added white noise. The first three terms in Equation 2.5 generate the shape of the function. The noise contributes to oscillations about that function's shape. The presence of noise makes it look as if a highlighter has been used to extend the width of the function. This width is referred to as the envelope of the function. The purpose of this example is to show that through the use of the Fourier transform the noise can be identified and removed. This should result in a decreased envelope in the reconstructed signal.

$$S' = \underbrace{0.7 \sin 12\pi t + 0.7 \cos 4\pi t + 0.7 \sin 2\pi t \exp \cos t}_{S} + \epsilon \tag{2.5}$$

where $\epsilon$ represents a white noise process. $S'$ is the signal with noise and $S$ is the true signal without the $\epsilon$ term. This signal will be considered to be an audio sample, containing very clear sine and cosine components. However the information contained in the signal is obstructed by the presence of noise. In this example, the Fourier transformation will be used to clean the signal, removing the fuzzy noise from the audio. When played aloud the noise can act to essentially mask those frequency components in $S'$ due to $S$. The Fourier transformation can be used to extract $S$ from $S'$. To do so one must compute the Fourier coefficients in the form

$$Y(k) = \Sigma_{j=1}^{n} S' j W_n^{(j-1)(k-1)} \tag{2.6}$$

where $Y(k)$ is the Fourier coefficient in frequency space, $n$ is the length of the signal, $k$ is the wave number, and $S'$ is the signal with noise [25]. $W_n$ is an exponential function defined as $W_n = \exp(-2\pi i/n)$. The exponential can be expanded in terms of sine and cosine functions as $\exp(-2\pi i/n) = \cos(2\pi/n) + i \sin(2\pi/n)$. The sine and cosine functions span all of space and are the bases used to extract frequency components. In this way we sum over the entirety of the signal's length to compute each frequency component.

Figure 2.4 represents the Fourier transformation of the signal $S'$. The figure was computed using the *fft* function in MatLab. The $x$-axis is frequency and the $y$-axis is the single sided spectrum of values. The Fourier transform is limited by the Nyquist frequency, which is the upper most sampling rate of a signal (anything greater results in oversampling, and this creates redundancy). The redundant information has been excluded from this figure, this is indicated by the term single sided [25]. The components of each frequency was determined by taking the absolute value of the Fourier transformation and dividing by signal length [25].

The figure shows the dominant frequencies, the $S$ part of $S'$. It also shows the white noise process as those frequencies with components approximately zero. A white noise process is one in which there is no dominant frequency, each contributes an equally small part to the signal. With the information presented in Figure 2.4 $S$ can be recovered by setting those frequency components that do not have a strong peak to zero, and performing the inverse Fourier transform. Only the first 25 frequencies are included in Figure 2.4. This is done because it is clear from Figure 2.3 that the signal only contains lower frequency components.

To filter the white noise from the signal, and improve audio quality, identify the value of the lowest peak that represents a component from $S$ (this is approximately .15). Now set all frequencies with components less than .15 to zero and compute the inverse Fourier transform given by

$$X(k) = \frac{1}{n}\Sigma_{k=1}^{n} YjW_{n}^{-(j-1)(k-1)} \tag{2.7}$$

where $X$ gives back a signal close to $S$ from $S'$, and $Y$ is the frequency components computed with the Fourier transform [25]. In this way the Fourier transformation can be used to determine the dominant frequencies contained in a signal. In this example, noise was removed from a signal presented in Figure 2.3, the components of that signal were analyzed using Figure 2.4. Those frequencies with components below .15 were set to zero and the inverse Fourier transform was taken to produce a scrubbed version of the original signal. Figure 2.5 contains the signal resulting from the *ifft* function in Matlab and the original function $S$ is included for comparison. The cleaned signal contains fewer oscillations around the waveform produced by $S$ than the waveform produced in Figure 2.3. The envelope of the function has been shrunk. The envelope is used to describe the width of the oscillations about the shape of the function $S$, the first three terms in equation 2.5. The cleaned function more closely matches $S$ than the original function $S'$.

The Fourier Transformation can be adjusted to provide localized information through windowing. This method is useful for analyzing non-stationary signals, those signals that have different statistical features throughout their lengths. For instance, the average value of a signal can be computed along equally sized sections of its length. Suppose a discrete signal is 25 elements in length. Then the average

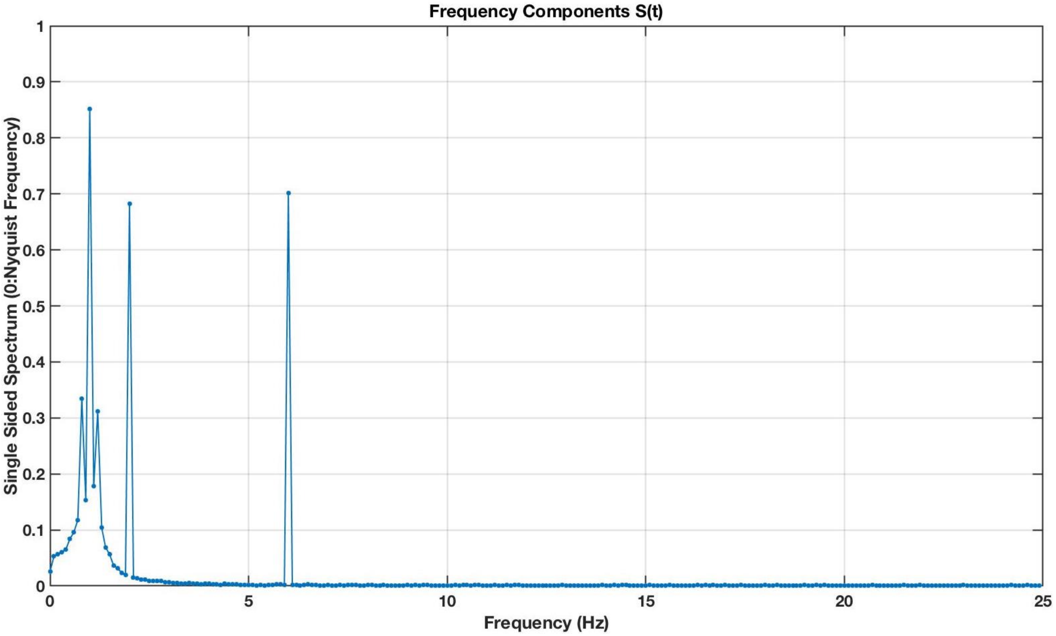


Figure 2.4: Results of the Fourier transformation on the white noise signal with added frequency components. The $x$-axis is the frequency spectrum and the $y$-axis is the amount that the frequency contributes to the signal. The sharp peaks are those frequencies that were added into the signal, and the white noise are the frequency components close to zero.

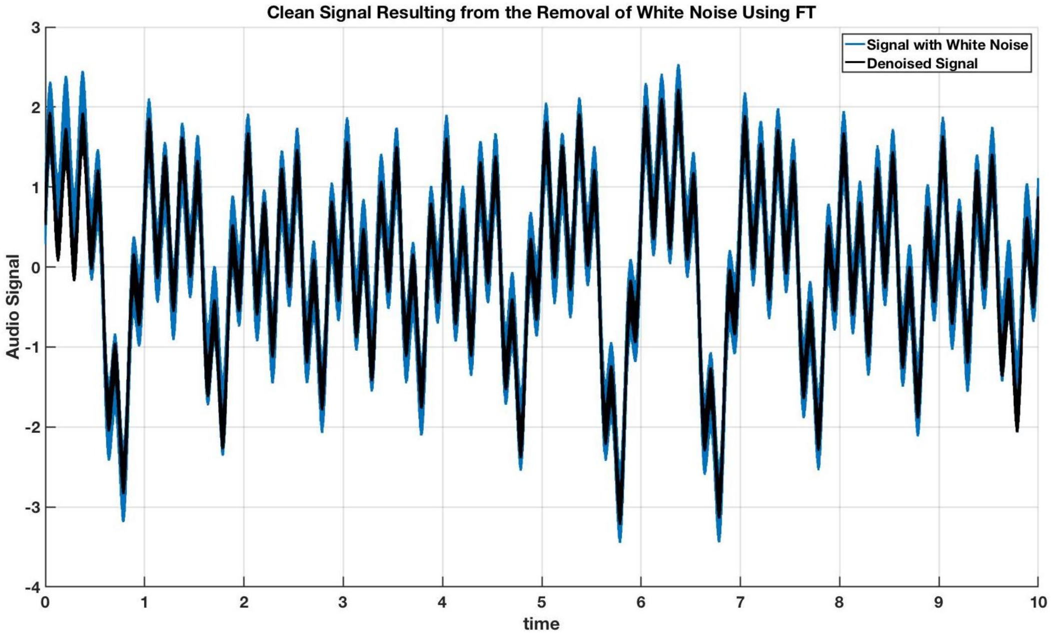


Figure 2.5: Signal with noise removed using the Fourier and Inverse Fourier transforms is in black. The blue waveform is the signal $S'$. By using the Fourier transform, white noise components were identified and excluded from the frequency spectrum. These corrections result in a new signal with fewer oscillations about the main waveform that more closely matches the true signal given by the first three terms in equation 2.5. This can be seen as original signal appears to be a backdrop on the new filtered signal.

of each fifth of the signal can be computed without redundancy (*i.e* the first fifth and second fifth contain information about different parts of the signal). If any of these sections have a mean value that is different than the others, the signal is non-stationary. If each section's mean value is the same, the signal is stationary in regards to the moving average taken over each fifth.

Since the Fourier bases are sine and cosine these signals span all space. In the Fourier transform, one takes the convolution of a signal

$$\int_a^b g(t)FWdt \tag{2.8}$$

where *g(t)* is the signal, $F$ its Fourier transform, and $W$ the window that determines the limits of integration. The Fourier coefficients consist of an infinite series of sine and cosine terms that span all space. The choice of window is arbitrary, meaning the user selects the shape and length of the window to apply. The wavelet transformation generates localized bases, making it better suited to capturing frequencies in certain sections of the signal. The wavelet performs a multiresolution analysis naturally. Both methods provide the same information, the wavelet transformation presents a structure within which the information is stored in a more useful form.

## 2.5 Discrete Wavelet

The discrete wavelet is a signal analysis tool used to resolve localized frequency components within a signal. This study deals in one dimensional hydrophobicity signals generated by mapping each amino acid of a short globular protein to its corresponding hydropathy value as determined by the Kyte-Doolittle scale [15]. The discrete wavelet transformation takes a signal and applies a low and high pass filter to it. The application of each filter produces two new sets of coefficients. Each set of coefficients is a frequency band containing a section of the span of frequencies from zero to pi. Each coefficient within the frequency band represents the contributions to the signal from each frequency within a section of signal. The filters are then applied to the low pass frequency band, producing two new sets of coefficients, increasing resolution in the frequency domain [21][26].

The filters are given by

$$\phi(x) = \sqrt{2}\sum_{\tau} h_{\tau}\phi(2x - \tau) \tag{2.9}$$

$$\psi(x) = \sqrt{2}\sum_{\tau} g_{\tau}\phi(2x - \tau), \tag{2.10}$$

where $\phi$ is the low pass filter, and $\psi$ the high pass [21]. $\tau$ is the translation parameter, moving the wavelet along the signal in discrete steps, and $h_{\tau}$ and $g_{\tau}$ are coefficients representing the type of wavelet transformation used [26]. In this study the coefficients represent the average value, and average difference of successive elements of the initial signal. This study makes use of the Daubechies One (db1) wavelet transformation bases, which takes the average value, and average difference in value between successive elements of the initial signal, and its resulting frequency bands. The application of the filters to the signal results in frequency bands, each containing a number of coefficients that is half the original signal length. The application of the filters requires that the initial signal be of length power 2 ($2^n$) [21].

In this study hydrophobicity signals are zero padded, meaning that zeros are added to the end of the signal till they are of length power 2. Each of the frequency bands resulting from a pass of the filters will contain coefficients affected by this edge extension. The zeros can make regions of the signal appear to fluctuate less than they actually do. This edge extension is uniformly applied to every hydrophobicity signal in the data set. Coefficients within the frequency bands computed over the extended edge can be identified and removed from the frequency bands. This study did not remove coefficients due to time constraints, but zero padding is used uniformly throughout the homologous pairs.

To achieve a multiresolution analysis, these filters are applied to the resulting low pass frequency band. The functions for each filter are adapted as

$$\phi_{j\tau}(t) = 2^{\frac{j}{2}}\phi(2^{j}t - \tau) \tag{2.11}$$

and

$$\psi_{j\tau}(t) = 2^{\frac{j}{2}}\psi(2^j t - \tau). \tag{2.12}$$

to represent application at a subsequent level. $j$ indicates what is called the level of the decomposition, where level indicates the resulting low and high pass frequency bands after a pass through the filters. Passing the signal through the filters once results in level $j = 1$. $t$ is used to indicate the filters are now being applied to the coefficients within the low pass frequency band. Figure 2.6 shows an example of how the filters extract information from a signal [21].

The signal in Figure (2.6) is fourteen elements long. Each filter is applied to the signal to extract localized frequency components along its length, without overlap. This is achieved by fitting the signal to the bases defined by the transformation. In this study, the average value is taken by the low pass filter, and the average difference in successive values is determined by the high pass. The figure shows that application of the filters takes a signal and generates two new sets of coefficients. These coefficients represent the average value and average difference of successive pairs along the length of the signal. This is indicated by the width of the bars for the low and high pass filter coefficient values. The high pass filter contains seven coefficients even though the visual makes it seem there are fourteen. The visual shows the values used to reconstruct the original signal by taking the sum of the coefficients in the low pass with the high. The first coefficient in the low pass filter is added to the first coefficient in the high pass filter twice. The first coefficient of the high pass filter is shown to account for this. For example add the first bar in the high pass with the first bar in the low pass to reproduce the first value in the original signal. To compute the second, one must take the negative of the high pass coefficient and add it to the low pass value, returning a value of two, the magnitude of the second value in the signal. This representation allows for the visualization of the reconstruction process.

The wavelet is a short oscillation in time that rapidly collapses to zero outside its support. Whereas sine and cosine span the entirety of the space, the wavelet is finite. This allows for the wavelet to be stretched or compressed in space. A

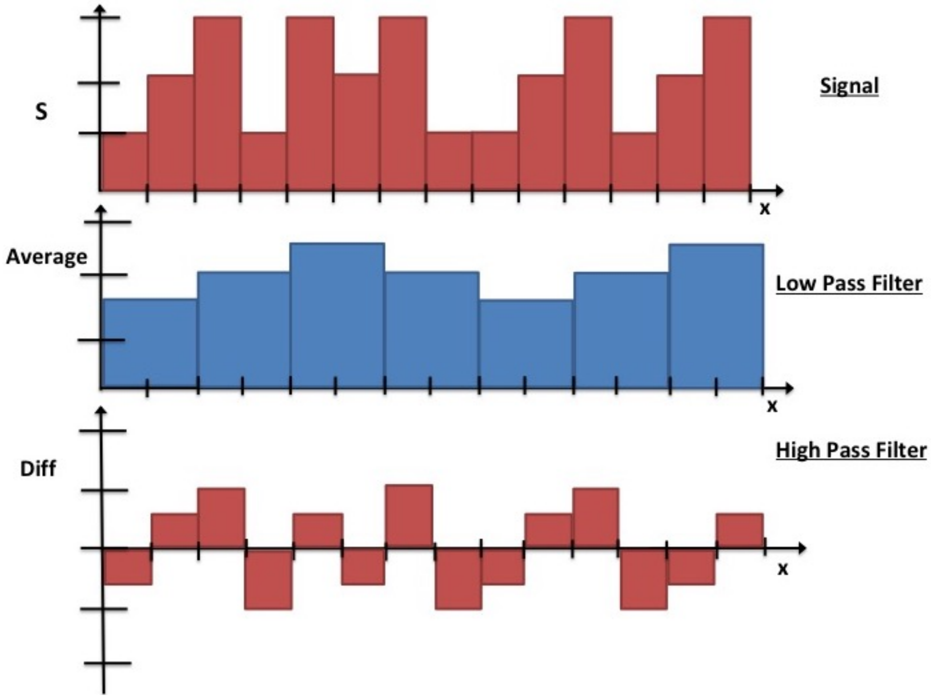


Figure 2.6: Application of low and high pass filters to a signal. The initial signal is presented at the top of the figure. The application of the filters results in two sets of coefficients. The first set is the result of the low pass filter. This study uses the Daubechies' One wavelet bases. The low pass filter finds the average value of a pair of signal values, and the high pass takes the average of the difference. The resulting sets are half the length of the original signal. The high pass filter shows the values needed to reconstruct the initial signal. The original signal can be obtained by adding the low and high pass coefficients.

stretched wavelet is used to determine oscillations that occur over longer intervals, while a compressed wavelet can capture rapid oscillations [21]. Since the wavelet's support is finite, it can be shifted across the length of the signal. In this case the shift occurs in steps, so that the first coefficient in either frequency band resulting from the initial application of the filters is taken over the first and second element, third and fourth, fifth and sixth, and so on. This shows that the bases used is orthogonal [21].

Each of the coefficients within and between frequency bands are orthogonal to one another. In this context orthogonal means that each application of the filter to the coefficients making up a signal comes with zero overlap. The coefficients in each frequency band do not contain the same information about the signal. The translation ensures that previous sections of the signal that have been covered in a previous application of filters are not double counted as the filter moves along the signal. This is a result of shifting, or translating, the bases of the transform. The bases of the transform is the wavelet. This study uses the Daubechies 1 wavelet basis. This bases finds the average value and average difference of a signal along its length [21].

In this study the hydrophobicity signals were considered a pair between the homologous proteins from mesophile and thermophile. To prevent oversampling, the homologous pair was decomposed to one level above the maximum decomposition level of the shorter of the two proteins. The maximum decomposition is determined to be the number of applications of filters that results in a frequency band containing a single coefficient. The Wavelet Transformation requires a signal to be of length power 2. In this study hydrophobicity signals were zero padded. This means that zeros were added to the end of a signal extending it to some length power 2. This edge extension was used throughout the study.

## 2.6 Wavelet Packet Transformation

The wavelet is a powerful tool in this investigation. The method provides information about localized hydrophobic content. This offers the opportunity to look at the hydrophobic content of the polypeptides in terms of specific regions within the sequence, or more globally in terms of the information content within each similarly sized section of the protein.

The Wavelet Packet Transformation (WPT) is a signal analysis tool used to resolve localized frequency information [21]. The WPT is useful in that it performs a multi-resolution analysis of the hydrophobic free energy signals, achieving a purpose similar to a windowed Fourier Transform. It differs from the windowed Fourier transformation by expressing the original signal in terms of localized bases. The wavelet transformation determines localized frequency components by fitting bases to the signal through the use of a wavelet. Fitting means that the wavelet forces the values within a specific region of the signal to take on its shape, stretching or compressing the form of the signal. This is achieved through a convolution with the signal. Convolution means that the integral of the product of the initial signal with the wavelet is taken with limits determined by the translation parameter $\tau$ [21].

In the WPT the same filters as given by Equations (2.9) and (2.10) are applied to the original signal. This is followed by the application of the filters given by Equations (2.11) and (2.12) to both of the resulting frequency bands. Each application of the filters is given by $j$. At each level there are $2^j$ frequency bands, each containing a set of coefficients quantifying the fluctuation in hydrophobic free energy within a unit length along the protein. These frequency bands are referred to as leaves, or nodes, of the WPT tree. The low and high pass filters are applied to the signal. This results in two frequency bands each with their own set of coefficients. Subsequent applications of the filters are applied to both frequency bands. This decreases the range of frequencies in each band, while decreasing localization. The entire wavelet packet tree structure is recorded.

Figure 2.7 details the steps necessary to perform a wavelet packet transformation on an example signal. The first step is to determine the length of the signal. If the signal is not of length power 2, then it must be extended. In this study signals are extended using zero padding, and an example is provided. The next step is to determine the max decomposition level, which is the number of times the filters need to be applied to result in leaf with a single coefficient. To apply the filters to each frequency band, the set of coefficients within that band must be of power 2 as well. For longer signals that can be decomposed to higher levels, this requires that subsequent applications of zero padding be applied to the frequency bands, in addition to zero padding the signal. The orange brackets indicate the orthogonality of the wavelet bases. A bases is applied to each of the brackets separately with zero overlap. The filters with their specified coefficients $g$ and $h$ then act to take the average and average difference of each of the bracketed values. This results in the set of coefficients at the bottom of the figure.

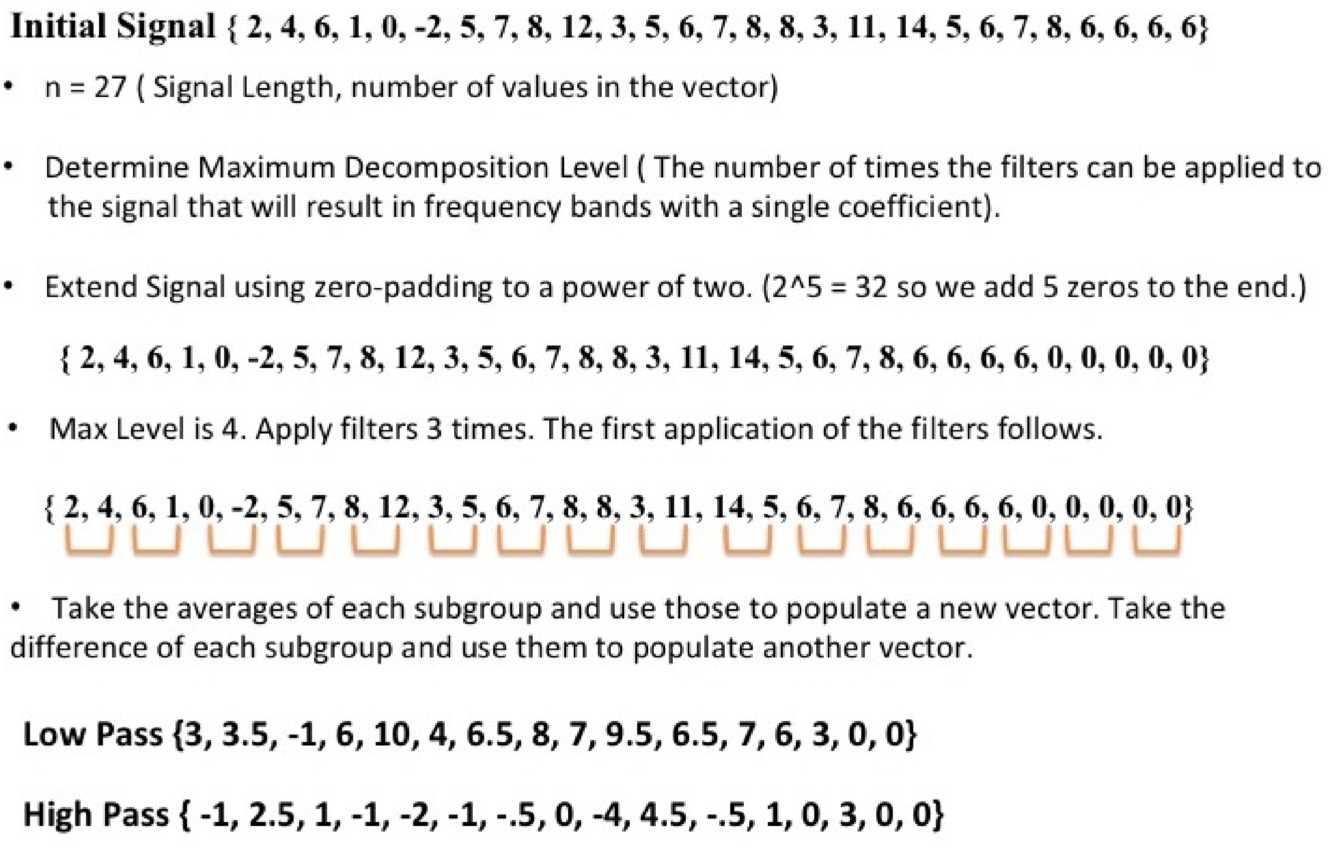
Figure 2.8 presents the entire wavelet packet tree structure and the corresponding frequency range contained within each leaf. A color coding scheme is used to help explain the wavelet's ability to extract localized frequency information, and the orthogonality of the bases. The initial signal is presented at the top of the figure, as indicated by $j = 0$. The first application of the low and high pass filters act on the pairs of values indicated by orange brackets. The coefficients of the average and average difference of these values are printed in orange and immediately follow below. The range of frequencies captured by the application of the filters is indicated by black brackets next to the filter value.

The second application of the filters act on the values bracketed in purple. The coefficients resulting from the application of the filters are in purple, the frequency spectrum range captured by the filters is in black. The level is indicated by $j = 2$. The first element of each frequency band corresponds to the first bracket of its corresponding color and so on. At level $j = 2$ it can be seen that successive application of the filters results in a rearrangement of the frequency bands, they are no longer in rank order (from lowest to highest frequency). Each application of the filters, increasing $j$, further deviates the order of leafs from rank order. To account for this, the terminal leaves of a WPT applied to the hydrophobicity signal were placed in rank order using the *otnodes* function in Matlab.

The rank ordering of frequency bands is lost as the wavelet is stretched and compressed through the application of the filters. A stretched wavelet captures lower level frequencies while a compressed wavelet captures higher frequencies. When the high pass filter is applied, it uses a compressed wavelet to capture high frequencies in a section of the signal. To increase the resolution of that filters frequency spectrum, both the low pass and high pass filters are applied. So that a new wavelet is applied and stretched to a section of the signal that was just compressed. In this way the rank ordering of frequency bands is lost.

The successive application of wavelet bases to the signal causes it to be reshaped. This reshaping results in the subsequent applications of the filters to pick out frequency ranges that deviate from rank order. For example, in Figure 2.8 it can be seen that the deviation from rank order occurs at level $j = 2$. The coefficients within the high pass filter at level $j = 1$ represent the compressed signal. Those coefficients are then stretched when convoluted with the next iteration's low pass filter. This results in the low-pass filter picking out high frequency components.

At the third level of the decomposition presented in Figure 2.8 it can be seen that increasing resolution in frequency space decreases resolution in physical space. At the lowest level $j = 3$ each of the leaves capture a portion of the span from zero to $\pi$ of size $\frac{\pi}{8}$ in width. To achieve this resolution required that each coefficient of the terminal nodes be computed over a fourth of the original signal with zero padding. In addition to this, at every level there are coefficients that take into account edge effects. Their values represent the addition of zeros and not characteristics of the original signal. Other methods of edge extension are available, for instance period wrapping, which repeats the signal until it is of length power 2. Zero padding was used in this study since the amino acid chains of proteins are finite in length. In addition to this, the study goal is to understand how the hydrophobic effect is encoded into the linear amino acid chains to reflect thermal stability and function through a comparative analysis of homologous proteins expressed in prokaryotes with non-overlapping optimal growth temperatures. Since the same edge extension is used for both the mesophile and thermophile proteins in each pair, the effects of edge extension are uniformly represented in each of the homologous pairs.



**Initial Signal** { 2, 4, 6, 1, 0, -2, 5, 7, 8, 12, 3, 5, 6, 7, 8, 8, 3, 11, 14, 5, 6, 7, 8, 6, 6, 6, 6}

- n = 27 ( Signal Length, number of values in the vector)

- Determine Maximum Decomposition Level ( The number of times the filters can be applied to the signal that will result in frequency bands with a single coefficient).

- Extend Signal using zero-padding to a power of two. (2^5 = 32 so we add 5 zeros to the end.)

  { 2, 4, 6, 1, 0, -2, 5, 7, 8, 12, 3, 5, 6, 7, 8, 8, 3, 11, 14, 5, 6, 7, 8, 6, 6, 6, 6, 0, 0, 0, 0, 0}

- Max Level is 4. Apply filters 3 times. The first application of the filters follows.

  { 2, 4, 6, 1, 0, -2, 5, 7, 8, 12, 3, 5, 6, 7, 8, 8, 3, 11, 14, 5, 6, 7, 8, 6, 6, 6, 6, 0, 0, 0, 0, 0}

- Take the averages of each subgroup and use those to populate a new vector. Take the difference of each subgroup and use them to populate another vector.

**Low Pass {3, 3.5, -1, 6, 10, 4, 6.5, 8, 7, 9.5, 6.5, 7, 6, 3, 0, 0}**

**High Pass { -1, 2.5, 1, -1, -2, -1, -.5, 0, -4, 4.5, -.5, 1, 0, 3, 0, 0}**

Figure 2.7: Steps to carrying out a Wavelet Packet Transformation of a signal. The figure also includes descriptions of the maximum decomposition level, orthogonality (orange brackets), and edge extensions. In addition to this, it presents a calculation of the coefficients in the first two frequency bands of each leaf resulting from application of the low and high pass filters to the zero padded signal.

## 2.7 Spectral Analysis

The conversion of the amino acid chains to hydrophobicity signals produces a discrete time series that, when passed through the WPT, produces a spectrum of values. These values represent the fluctuation in hydrophobic free energy along the signal. This data type can be analyzed by looking at the mathematical analogs of physical quantities such as energy and power. The energy of a signal is a means of measuring its activity, or how quickly its values change. This quantity is traditionally applied directly to the signal, where the signal takes the form $x(t)$ over all space. To measure the energy in a signal define an interval of time such that $-T \leq t \leq T$. Then integrate so that

$$Energy\ of\ x(t) = \int_{-T}^{T} x^2(t)dx \tag{2.13}$$

determines the energy of the signal [20].

The energy of a signal can be informative so long as the signal is of finite length and non-stationary, meaning that different sections of the signal have unique statistical properties. Often a signal is stationary, it has the same statistical properties throughout its length, and can be considered to come from a signal that spans throughout time. In this case, the energy of the signal will be infinite [20]. To adjust for this, one can compute the average energy per unit time, or power of the signal within a given region, or over its entire length. Mathematically this value gives us units of energy per unit time, or watts which is the unit of power[20]. In the continuous case the power is computed as

$$Power\ of\ x(t) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x^2(t)dx[20]. \tag{2.14}$$

The spectral analysis used in this study examines the terminal nodes of the wavelet packet decomposition. Figure 2.8 is a representation of the wavelet packet tree structure where the lowest level, the leaves of the wavelet packet tree, are the average fluctuation in the difference in hydrophobic free energy per residue per unit width. At this level, we apply the concept of power from signal analysis to these coefficients. This measures the average variance in the fluctuation of the difference

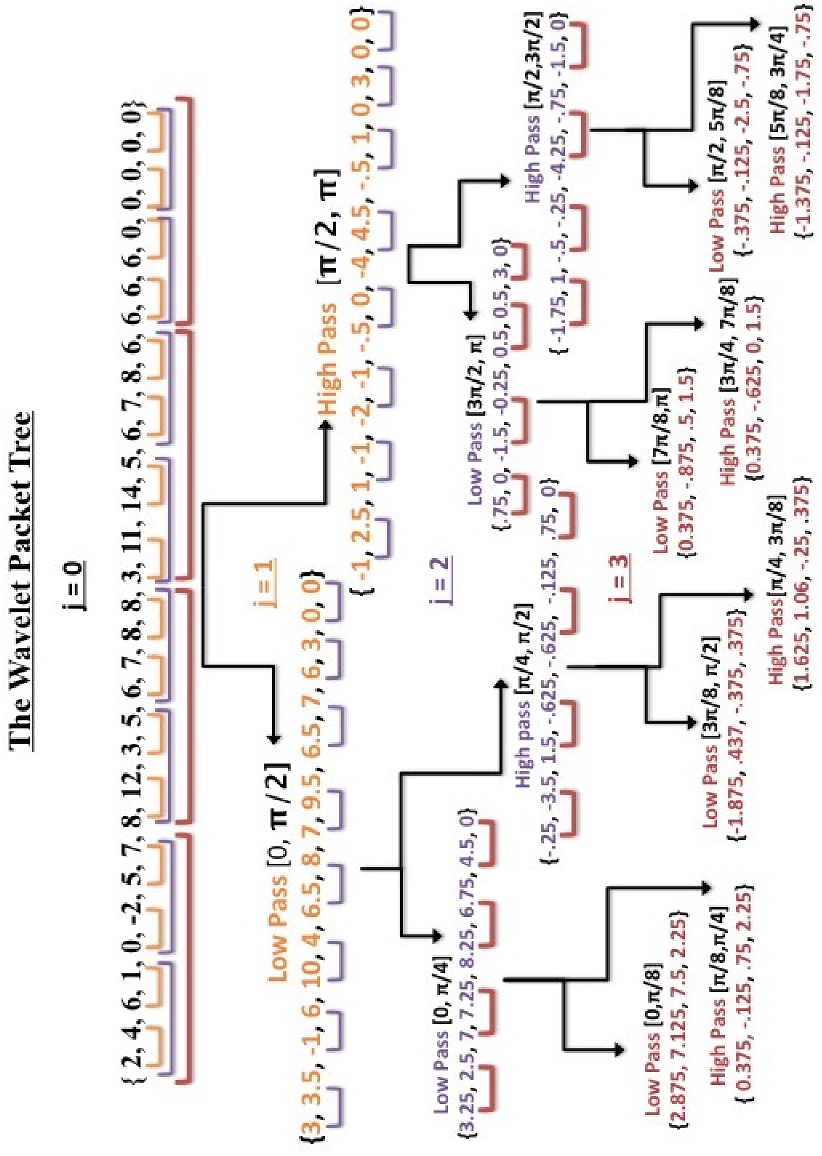


Figure 2.8: This figure goes through the application of both the low and high pass filters for a signal using the db1 transformation. This transformation results in the average and difference of sections of the signal. The WPT is limited by the Nyquist frequency, for instance if a signal contains frequency components from zero to $2\pi$, the WPT measures components from 0 to $\pi$. The orange brackets indicate the orthogonality of the wavelet bases on the first pass. The values bracketed in orange are used to compute the first sets of coefficients. The purple brackets indicate the range of values used to compute the frequency bands at $j = 2$, the second level. Red indicates those values used to compute the leaves of the third level of the decomposition. Through the use of color coding and bracketing the location of each frequency component can be identified at each level and between leaves. It can be seen that as resolution is increased in frequency space, resolution is lost in physical space.

in hydrophobic free energy per residue per unit length squared. This measure can be used to compare the fluctuation in hydrophobic free energy per unit length, across all unit lengths (or frequencies) throughout the entirety of the signal. Plotting these values in respect to frequency ordering allows for a qualitative comparison of how the fluctuation of hydrophobic free energy changes with frequency band.

In this study, we compute the power of the wavelet packet tree coefficients two levels above the maximum decomposition level. The method used here is implemented in the same manner as that presented in Equation (2.14), which deals with the power of the actual signal, not its frequency components. The extension of power from time series to frequency was shown by the French mathematician Marc-Antoine Parseval for the Fourier transform. Parselval's theorem shows that the integral of the square of a Fourier transform coefficient is the same as the integrated square of the original function. Parselval's theorem can be extended to the WPT to show that the integrated square of each frequency band is the same as the integrated square of the original function.

The sum of the power series coefficients gives the total power of the signal at the WPT level. This measure is interesting and different from power at each frequency in that it is independent of the ordering of the values within the original signal.

## 2.8 Spectral Methods as Measures of Similarity and Differences

Statistical tests were also run on the terminal nodes of the wavelet packet tree. These tests were skew and kurtosis. Skew is used to determine if there are regions of the hydrophobicity signal contributing more to the power of that frequency than others. The skew of each terminal node can then be compared to one another to see if some frequency bands have larger contributions being made by certain sections of the hydrophobicity signal [27]. Kurtosis was used to compare the number of outliers present in each terminal node's coefficients with one another [27]. In this study, skew, kurtosis, and power are functions of frequency. In this way, additional information is provided for the frequencies within the hydrophobicity signals.

Skew measures how spread out the values in a set are from the average. This value was calculated for each leaf of the WPT using the skewness function in MatLab. This calculation takes the form

$$s = \frac{E(x - \mu)}{\sigma^3} \tag{2.15}$$

$\mu$ is the average of the set of values $x$, and $\sigma$ is the standard deviation of $x$ [27]. Since each of the coefficients within the leaf (frequency band) are determined by localized wavelets, the skew can be used to determine the balance of the frequencies along the length of the signal [27].

The kurtosis of the terminal leaves was computed as well. This measure was used to determine if the power measured at some frequency was influenced by outlying values. Kurtosis is computed as

$$k = \frac{E(x - \mu)^4}{\sigma^4} \tag{2.16}$$

where $\mu$ is the mean value of the set $x$, $\sigma$ is the standard deviation of $x$, and $E$ indicates the expected value [27]. The kurtosis of a normal distribution is 3, for a set $x$ with a larger number of outlying values $k > 3$ and with fewer outliers, or a sharper curve $k < 3$ [27].

Because the wavelet has a value of zero outside its support, and is translated across the length of the signal in non-overlapping steps, each of the coefficients within the frequency bands are orthogonal to one another, meaning they contain information about different parts of the signal. This relationship holds with successive applications of the filters so that each of the frequency bands at every level are orthogonal. Because of this, the skew and kurtosis of each band can be taken as a function of frequency. This results in a skew and frequency spectra for each protein in the data set.

The power of each terminal leaf of a WPT decomposed hydrohpobicity signal was computed. Both proteins in the homologous pair were decomposed to the same level so that they would have the same number of terminal leaves. Each terminal leaf contains the amount of a certain frequency range present in a portion of a signal.

By finding the square of each coefficient in the leaf, and then taking the average, the power within that frequency range for the entire signal is calculated.

To determine how similar the power is at each frequency band within the homologous pair, the absolute difference was taken between the proteins in the pair at that frequency. If the difference in power was less than .99 joules, the value was recorded. The same procedure was performed on the bootstrapped control data (detailed in the following section). In this way, the number of homologous pairs that have similar power at any given frequency range within the entirety of the data set can be determined and compared with control data testing for the effects of amino acid ordering. In addition to this, the total power of a signal was computed. This measure is determined by summing the power of every frequency band, and is independent of the ordering of amino acids, and decomposition level.

## 2.9 Confidence Intervals

A control data set was constructed to determine the effect of amino acid ordering on the wavelet packet decomposed hydrophobicity signal. This was done by generating 299 permutations of each signal, maintaining amino acid content while removing position dependence. The sequence of hydrophobicity values was reordered using the randperm function in MatLab. Each permutation was determined using the Mersenne Twister pseudo-random number generator.

Each of these 299 permutations is passed through the WPT, and the coefficients at each frequency recorded. These data were then treated to the same spectral analysis methods as applied to the “native” ordering of amino acids. For instance, each frequency band had 299 permutations of power along the entirety of the decomposed signal.

These control data were then used to construct confidence intervals. To determine the confidence intervals, histograms were generated of the values computed at each frequency band and were used to eliminate those distributions that by visual inspection did not fit the data. Cumulative distribution functions were generated for the best looking fits to the true distribution, and used as the expectation value in a chi-square test against the values recorded at each frequency band. In this way, the distribution function was tested against each of the bootstrapped spectra. The best fitting distribution was determined to be the one that accepted the chi square test most often.

Three spectral analysis methods were applied to the terminal leaves of the WPT of each hydrophobicity signal, power, skew and kurtosis. The Gamma distribution was determined to be the best fitting CDF to describe the distribution of power measured at each frequency band for the bootstrapped data. For skew, the normal distribution fit the bootstrapped data. For kurtosis the lognormal distribution was used to construct confidence intervals.

Expectation values and confidence intervals were computed in Matlab based of the best fitting distributions as determined above. This was done using the *fitdist* distribution function to generate descriptive parameters for the distributions. The *paramci* function was then used to determine the upper and lower level confidence intervals at each frequency.

A second control data set was generated where the number of iterations was unique to each hydrophobicity signal. Each signal was shuffled ten times its length to generate a sample of the set of all possible amino acid combinations within the length of that protein. The purpose of this bootstrapping routine was to determine the distribution function of the power for each frequency band.

## 2.10 Summary

This study focuses on identifying adaptations made to the linear amino acid chains of short globular proteins related to thermal stability in prokaryotes with non-overlapping optimal growth temperatures, within the constraint that the protein must fold to perform a specific function. In this way alterations made to the hydropathy of a protein related to thermal stability are identified. The analysis methods aim to identify pair specific similar and pair specific distinguishing features.

Pair specific similar features are thought to be due to the homology of function between the proteins in a pair. Identification of these features leads to the formation of subgroups that can include all pairs with that feature. These groupings are then used to perform a gene ontology to determine if the feature is function specific. Pair specific distinguishing features are thought to be a result of thermal adaptation.

The first method used in this study was an alignment of the amino acid chains of the proteins composing a homologous pair. This was followed by tests of the hydrophobicity signal of each amino acid chain. The tests used included a scatter plot of the mean hydropathy of each signal pair, and the two sample Kolmogorov-Smirnov test. The Anderson Darling test was used to see if the content of hydropathy values in either hydrophobicity signal were normally distributed. These tests were followed by the application of the Wavelet Packet Transformation to each of the hydrophobicity signals in a pair. Spectral analysis methods were then applied to the resulting frequency bands produced by the transformation. Spectral analysis methods included finding the power of each frequency band, and then comparing the power of each band in rank order between the two members of a pair to determine if any had similar values. The same procedure was performed on the bootstrapped control data to determine a baseline likelihood of randomly encountering a matching power for some frequency band across the data set. Next, the total power was computed and compared between each of the members of a pair. This measure is invariant of the ordering of hydropathy values in a signal. Confidence intervals were determined for each frequency band's power for each protein in a pair using bootstrapped control data.

The skew of each of the frequency bands was determined, and confidence intervals were calculated using bootstrapped control data. The kurtosis of each frequency band was determined, and bootstrapped control data was again used to calculate confidence intervals. In both the skew and kurtosis, confidence intervals were calculated to determine if the ordering of hydropathy values in the signals determined the calculated value.

Figure 2.9 presents a flow chart consisting of the analysis methods used in this

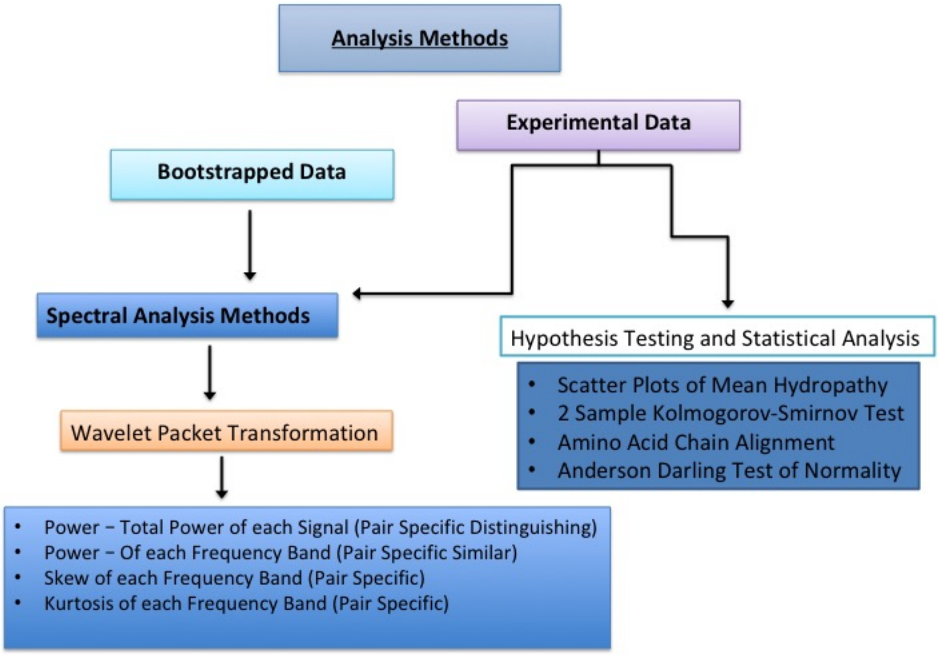

Analysis Methods
Experimental Data
Bootstrapped Data
Spectral Analysis Methods
Hypothesis Testing and Statistical Analysis
Scatter Plots of Mean Hydropathy
2 Sample Kolmogorov-Smirnov Test
Amino Acid Chain Alignment
Anderson Darling Test of Normality
Wavelet Packet Transformation
Power – Total Power of each Signal (Pair Specific Distinguishing)
Power – Of each Frequency Band (Pair Specific Similar)
Skew of each Frequency Band (Pair Specific)
Kurtosis of each Frequency Band (Pair Specific)

Figure 2.9: A flow chart of the analysis methods is presented. This chart details the methods used, and the experimental groups. The methods include alignment, scatter plots of mean hydropathy, the Anderson Darling test of Normality and the Two Sample Kolmogorov-Smirnov test. Spectral Methods were applied to the results of a wavelet packet transformation and include the total power of each signal, the power of each frequency band, and the skew and kurtosis of each frequency band.

study. Largely the methods can be broken up into two groups, statistical and hypothesis tests, and spectral analysis methods. The study uses both the amino acid chains of the proteins, the hydrophobicity signals themselves, and the frequency bands that result from a wavelet packet transformation applied to the hydrophobicity signals. In Chapter 4, these methods are applied to the hydrophobicity signals of short globular proteins from mesophile and thermophile prokaryotes.

# CHAPTER 3

# Data Overview

## 3.1 Overview

The data used in this study consists of the primary amino acid chains of homologous proteins in mesophile and thermophile prokaryotes. The polypeptides were converted to hydrophobicity signals by mapping each amino acid residue to a hydropathy value using the Kyte-Doolittle scale, maintaining the ordering of the amino acids. The species and protein function were unknown at the start of the study and were queried using Basic Local Alignment and Search tool from the National Center for Biotechnology Information through packages in R. In this way, the species and protein name were recorded. With this information, a gene ontology was performed on a subset of grouped homologous pairs.

This study consists of a comparative analysis of homologous pairs of proteins. In total there are 1080 short globular proteins analyzed in this study. These proteins are used to construct 540 homologous pairs, so that both mesophile and thermophile classifications contribute 540 proteins to the data set. The pair is constructed so that a procedure can be developed to identify characteristics in the linear amino acid chain related to thermal stability and function.

Both pre and post processing steps are performed on the data. These steps are used to test common assumptions made about thermal stability present in the literature, evaluate the use of spectral analysis methods and the results they produce, and determine what organisms contributed to the data set.

## 3.2 Data Structure

The proteins used in this study are short globular proteins, consisting of a single primary amino acid chain less than 600 amino acids (residues) in length. The proteins in this study were collected by collaborators at the University of Colorado Denver. The data were stored in two text files, one containing the amino acid sequence of all mesophile, and another containing the amino acid chains of the thermophiles. Amino acids chains were separated by white space within the text file. The files were organized so that the first amino acid chain in the mesophile data set was the partner of the first amino acid chain in the thermophile text file, and so on.

The data consists of various species of Bacteria and Archea. This information can allow for species specific groupings and better contextualizes the results of this study. This information also allows for the implementation of a Gene Ontology.

Table 3.1 presents those species contributing proteins to the thermophile data set. Six species of prokaryotes were identified, four from domain Bacteria, and two from domain Archea. In total there are 540 thermophile proteins. There are 16 proteins in the data set that could not be matched to a species or genus. For instance one of these proteins is classified as ”RecName: Full = Uncharacterized protease MJ0090”. There are 98 proteins that could only be associated with the genus *Thermotoga*. Overall 426 of the proteins have both species and genus information associated with them.

Table 3.2 presents those species contributing proteins to the mesophile data set. Thirteen species of bacteria were identified. There are 540 proteins in the mesophile data set. These thirteen species contribute 310 proteins. There are 7 entries containing only the genus of the organism a protein comes from. These 7 genus contribute 208 proteins to the data set. There are 22 proteins that could not be associated with a species or genus. For instance one of these proteins in the mesophile data set is classified as ”MULTISPECIES: NADH-quinone oxidoreductase subunit E”.

| —Thermophile Species— | —Domain— | —Number of Proteins— |
|---|---|---|
| *Pyrococcus abyssi* | Archea | 62 |
| *Aquifex aeolicus* | Bacteria | 197 |
| *Thermotoga maritima* | Bacteira | 128 |
| *Thermotoga*(unclassified) | Bacteria | 98 |
| *Thermotoga neapolitano* | Bacteria | 1 |
| *Thermotoga naphthophila* | Bacteria | 1 |
| *Methanocaldococcus jannaschii* | Archea | 37 |
| No Species Info | — | 16 |

Table 3.1: This table displays the Species and Domain of each of the thermophile proteins. The third column contains the number of proteins each species contributes to the data set. The best represented species comes from *Aquifex aeolicus* which is a member of Domain Bacteria. *Thermotoga* is a genus of bacteria and is only used because no further species specific information was provided by the Blastp query.

### 3.3 Amino Acid Sequencing

This study relies on the biochemistry of protein sequencing. The general idea behind sequencing a protein consists of three parts. The first is to determine the amino acid content within the polypeptide. This is done by denaturing the protein at high temperatures and then hydrolyzing it [28]. Hydrolyze means to separate a molecule into its constituent components by stress in water. The next step is to determine the n-terminal amino acid. Polypeptides are formed by linking the amide group of one amino acid to the carboxyl group of another, producing a water molecule. In this way, the start of an amino acid sequence is considered to be the exposed amide group [28].

The process of identifying the n-terminal amino acid is known as Edman degradation. This entails removing one amino acid from the polypeptide at a time. *Dabsyl chloride* is commonly used to perform this operation. The process is not 100% efficient, and for proteins containing more than 50 amino acids, an additional step is made [28]. The protein is cleaved into smaller pieces through a cleavage enzyme that acts at certain sites along the polypeptide chain. A common cleavage enzyme is *Cyanogen bromide* which splits the polypeptide on the carboxyl side of a methionine.

| —Mesophile Species— | —Domain— | —Number of Proteins— |
|---|---|---|
| *Mycobacterium tuberculosis* | Bacteria | 24 |
| *Bacillus halodurans* | Bacteria | 164 |
| *Streptococcus pneumoniae* | Bacteria | 46 |
| *Streptococcus* | Bacteria | 12 |
| *Escheria coli* | Bacteria | 55 |
| *Proteobacteria* | Bacteria | 45 |
| *Bacillus* | Bacteria | 128 |
| *Corynebacterium* | Bacteria | 15 |
| *Corynebacterium glutamicum* | Bacteria | 11 |
| *Mycobacterium shimoidei* | Bacteria | 1 |
| *Bacillaceae* | Bacteria | 2 |
| *Bacillus okuhidensis* | Bacteria | 2 |
| *Shigella sonnei* Ss046 | Bacteria | 2 |
| *Shigella flexneri* | Bacteria | 1 |
| *Enterobacter hormaechei* | Bacteria | 1 |
| *Enterobacteriaceae* | Bacteria | 5 |
| *Escherichia fergusonii* | Bacteria | 1 |
| *Salmonella enterica* | Bacteria | 1 |
| *Cronobacter sakazakii* | Bacteria | 1 |
| *Enterobacterales* | Bacteria | 1 |
| No Species Info | — | 22 |

Table 3.2: This table displays the Species and Domain of each of the mesophile proteins. The third column contains the number of proteins each species contributes to the data set. The best represented species come from *Bacillus*. This study differentiates between *Bacillus* and *Bacillus halodurans* so that proteins are not over counted. *Bacillus* is a genus of bacteria, and is only used because no further species specific information was provided by the Blastp query.

This step is followed by the use of a second cleavage enzyme that acts at another amino acid [28]. Edman degradation is followed by peptide overlap. The strands produced by the various clevage enzymes will contain overlaps in the amino acid sequence. By aligning these overlaps, the full sequence of the polypeptide can be deduced. The various segments are separated using chromotography. This process causes different sized molecules to move in solution at different speeds, separating them [28].

## 3.4 Polypeptide Chains

The total power of a signal is dependent on its length. To determine if this measure is biologically significant as a means of distinguishing mesophile from thermophile, it is necessary to know both the length of the proteins in the data, and the difference in the lengths of each protein in a homologous pair. The proteins used in this study are under 600 amino acids long, and only one protein is less than 50 amino acids long. Table 3.3 contains the number of proteins that fall into a range of lengths. Most proteins in the data set are between 300 and 350 amino acids in length.

Figure 3.1 shows the distribution of the difference in the lengths of amino acid chains in a homologous pair. This difference was calculated by finding the length of the thermophile and subtracting from it the length of its corresponding mesophile protein. The histogram appears to be normally distributed, however the distribution deviates from normality at its tails. The Anderson Darling test showed that this data is not normally distributed. The median of this distribution is -1 amino acid, meaning that the mesophile is one amino acid longer then the thermophile. The skew is -0.05, while the skew of an ideal normal distribution is zero. Since the skew is so close to zero the distribution of the difference in length is balanced. Further more, the 95 percentile of the absolute value of the difference in length is 15 amino acids. 513 of the pairs are less than 15 amino acids different in length. The 75 percentile is 8 amino acids, and the 50 percentile is four amino acids. Since neither thermophile or mesophile groupings have a tendency to be longer than their partner protein, the total power of a hydrophobicity signal is computed. In addition to this, the histogram shows that increasing the length of a protein does not add to its thermal tolerance within the constraint that the protein must fold to perform a specific function. Percentiles were determined using the *prctile* function in Matlab.

Table 3.4 presents the one letter code for amino acids. It appears that cysteine (C), methionine (M), asparagine (N), proline (P), arginine (R), leucine (L) and tryptophan (W) contribute equally to the total length of mesophile and thermophile proteins. Figure 3.2 presents the average percentage each amino acid contributes to the length the mesophile and thermophile proteins. These values were computed by determining the amino acid content in each protein in the data set. In this way the percentage of length occupied by a specific amino acid is calculated for each protein. The average percent length is taken for all mesophile proteins and all thermophile proteins. The $x$-axis is the amino acid and the $y$-axis is the average percentage length for thermophile and mesophile.

It has been noted that thermophiles tend to have an increased number of aspartic acid (amino acid one letter code: D, KD scale value: -3.5), glutamic acid (E,-3.5), lysine (K,-3.4), and arginine (R,-4.5), in their amino acid chains as compared to mesophiles [1]. In addition to this it is believed that they have decreased numbers of asparagine (N,-3.5), glutamine (Q,-3.5), serine (S,-0.8), and threonine (T,-0.7) [1]. Because of this, the amino acid content of each protein was computed, and taken as a percentage of that proteins total length. The average percent length was then computed for each amino acid within mesophile and thermophile groupings. This data is displayed in Figure 3.2.

The average content of the four amino acids thought to be over expressed in thermophiles follows. Aspartic acid on average makes up 5.71% of a mesophile protein and 5.04% of a thermophile. Glutamic acid had a value of 7.88% in mesophiles and 9.70% in thermophiles. Lysine makes up on average 6.02% of mesophiles and 9.14% of thermophile chains. Arginine makes up 5.50% of mesophile chains and 5.58% of thermophiles. For those amino acids that are thought to make up a smaller percentage of the chain it was found that: asparagine makes up 3.52% of mesophiles and 3.51% of thermophiles. Glutamine makes up 3.65% of mesophiles and 1.95% of thermophiles. Serine makes up 4.95% of mesophiles and 4.40% of thermophiles. Threonine makes up 5.52% of mesophiles and 4.23% of thermophiles.

This study shows that aspartic acid on average composes a smaller portion of the amino acid chain of thermophiles than originally thought. Arginine was found to have the same level of expression in thermophiles as mesophiles. Glutamic acid and lysine appear to support the original conclusion. Glutamic acid makes up 1.81% more of the average thermophile amino acid chain and lysine makes up 3.11% more of the thermophile chain. These data show that on average aspartic acid and arginine do not make up a significantly larger portion of a thermophile's amino acid chain in comparison to a mesophile.

Of those amino acids thought to be under-expressed in thermophiles, aspargi-nine on average made up the same amount of a proteins length in both mesophile and thermophiles. The number of serine amino acids in a amino acid chain appears to be very similar between the two gropus, with a difference of 1.55% between mesophile and thermophile. Threonine on average makes up 1.24% less of a thermophile's amino acid chain than a mesophile's. Glutamine makes up 1.71% less of a thermophile's amino acid chain than it does a mesophile's.

This data set shows that of those amino acids thought to be more prevalent in thermophiles, only glutamic acid and lysine made up a larger percentage of the total amino acid chain length in thermophiles. For those amino acids thought to make up a smaller portion of a thermophile's protein, the largest difference was found to be in glutamine. Asparagine on average makes up the same amount of a thermophile and mesophiles' chain length. Serine and threonine made up only slightly less of the average thermophile amino acid chain than in mesophiles. Comparing the average amino acid content of the thermophile proteins to the mesophile proteins shows that amino acid content is fairly similar between the two groups. There is a small amount of variation in the amino acid content, however this difference is below 5%.

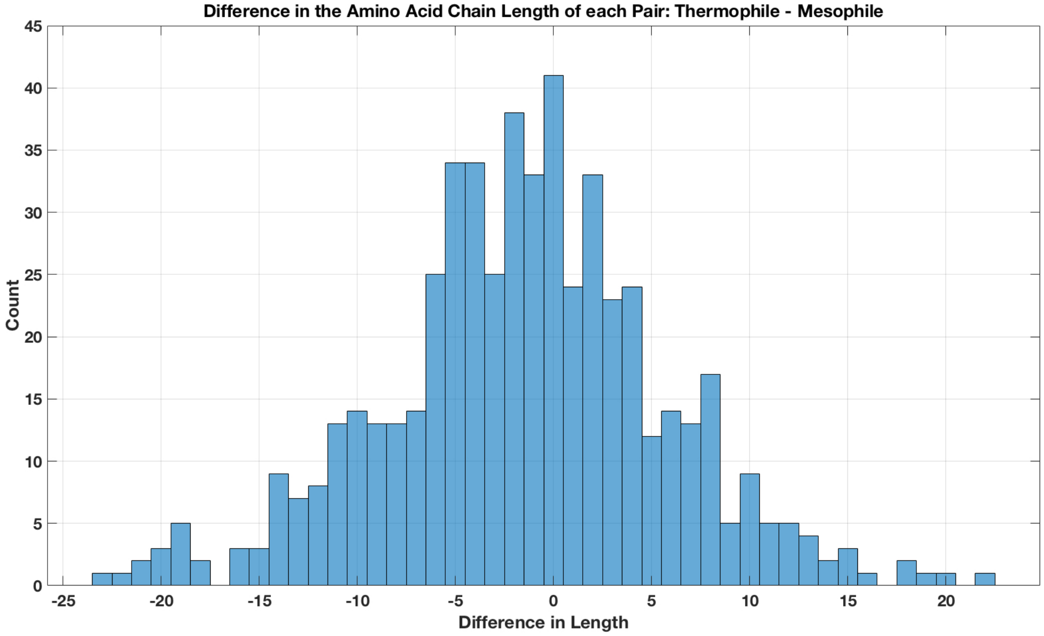


Figure 3.1: Histogram of the difference in amino acid chain length between the mesophile and thermophile members of a homologous pair. The histogram appears to be normally distributed, however it deviates from normality in its tails. Use of the Anderson Darling test confirms this. The skew of the histogram is -0.05. There is no trend for either a mesophile or thermophile protein to be longer than its partner in the homologous pair. Length does not appear to be a feature related to thermal stability.

| Length Range (Residues) | Mesophile Proteins | Thermophile Proteins |
|---|---|---|
| 0-50 | 1 | 0 |
| 50-100 | 23 | 26 |
| 100-150 | 57 | 52 |
| 150-200 | 74 | 79 |
| 200 - 250 | 63 | 62 |
| 250-300 | 66 | 70 |
| 300-350 | 82 | 77 |
| 350-400 | 46 | 48 |
| 400-450 | 56 | 60 |
| 450-500 | 43 | 43 |
| 500-550 | 18 | 15 |
| 550-600 | 11 | 8 |

Table 3.3: The length of each protein in amino acids is presented. The first column is a range of values for the number of amino acids in the chain. The second column is the total number of mesophile proteins that fall within these lengths. The third column is the number of thermophile proteins that fall within these lengths.

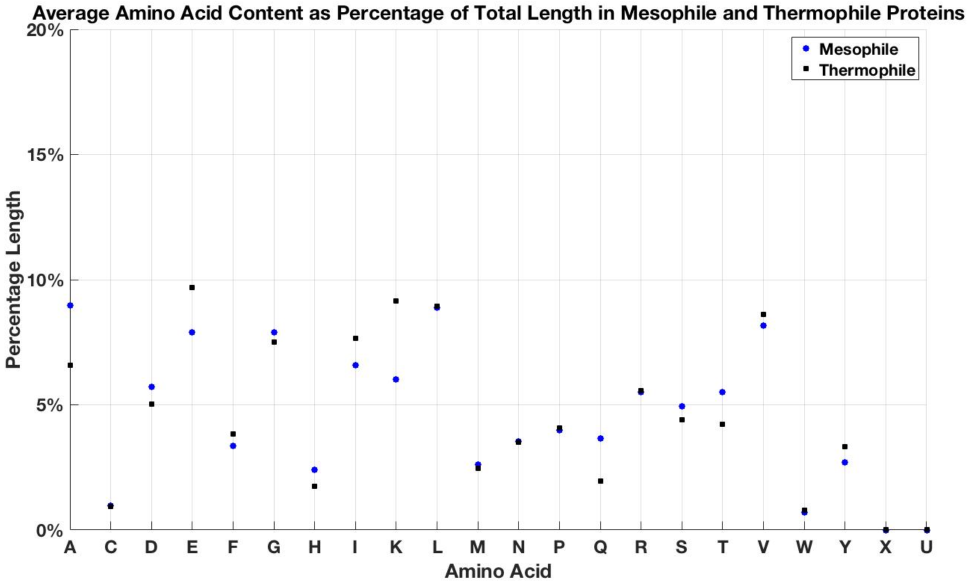


Figure 3.2: Average percentage chain length of each amino acid across mesophile and thermophile groupings. X and U are used to indicate that the amino acid at some location is unknown, there are only two of these in the entire data set.

| | |
|---|---|
| Alanine | A |
| Arginine | R |
| Asparagine | N |
| Aspartic acid | D |
| Cysteine | C |
| Glutamic acid | E |
| Glutamine | Q |
| Glycine | G |
| Histidine | H |
| Isoleucine | I |
| Leucine | L |
| Lysine | K |
| Methionine | M |
| Phenylalanine | F |
| Proline | P |
| Serine | S |
| Threonine | T |
| Tryptophan | W |
| Tyrosine | Y |
| Valine | V |

Table 3.4: One letter amino acid code. The letters X and U are used to indicate that the identity of the amino acid at the location is unknown. There is only one instance of X and one of U in the entire data set. The hydropathy for these two entries was set to zero.

## 3.5 Bioinformatics

The species and protein function were unknown and had to be looked up using the Protein Basic Local Alignment Tool (BLASTp) provided by NCBI. This was done in R using the blastsequences function from the annotate package. BLASTp queries a database containing known amino acid chains of proteins, and aligns the queried amino acid chain sequence with those in its database [29]. It then returns a list of hits, containing those proteins that have high scoring alignments, making it likely that the proteins are either the same or have similar function [29]. The top scoring result from this list was taken for the protein name and species information.

BLAST stands for the basic local alignment search tool and is used to determine sequence similarity [29]. BLAST is operated by the National Center for Biotechnology Information and offers sequence alignment for both amino acid and nucleotide sequences. The search algorithm operated by BLAST consists of three phases, the first is the set up. This involves reading in the query, search parameters and database. It then reduces the input sequence into "words" by breaking it into predefined lengths based on the query. A word is just a short sequence of amino acids. The preliminary search takes these words and compares them to the database for matching words. It takes that sequence of amino acids and looks for similar sequences in other amino acid chains [29]. Proteins have conserved regions called domains that correspond to specific function. The words are designed to find sequences with similar domains. If in the search the program finds a partial alignment, one containing a gap between the word and sequence from the library, a penalty is applied to the alignment score. The alignment score is used to rank potential matching sequences. The final phase is traceback. This phase determines insertions or deletions within the sequence. This study made use of the BLASTp algorithim to perform a protein-protein sequence comparison. [29]

The open source bioconductor project was used to query the BLASTp database. This software project provides statistical packages in R for bioinformatics work. The annotate package provided by bioconductor was used to run the *blastsequence* function [30]. This function takes a sequence and submits it to BLASTp database. To do so, the user has to specify the database, program and output formats. In this study, the first output was collected from querying the Prokaryotic RefSeq database using BLASTp. These data were returned in XML format and saved to a matrix in R. The information was then saved to two separate text files, one for the mesophiles and another for the thermophiles.

The text files were used to determine the different species present in each of the data sets manually. This information was then used in a Matlab program to determine the number of proteins present from each species. To do this, the program converted the text information provided by BLASTp containing the protein function and species information to an aggregate character array. The name of each species present was then counted using the *find* function in Matlab. This process was used to determine the total number of proteins each species contributed to the data set.

### 3.6 Gene Ontology

The Gene Ontology was developed to aid investigations like this, where similarly functioning genes, and their products (proteins), are compared between organisms. A gene ontology (GO) is a controlled vocabulary used to structure biological information. Its focus is to take information from disparate databases and combine them. In doing so, the hope is that newly discovered genes can be better categorized. For instance, if a new gene is discovered in the *Tardigrade* (water bear), then that gene can be compared to the human genome to infer its function. This can help the investigator develop a study to further knowledge of that gene, and evaluate its use as a model for human health.

An Ontology is created to determine relationships between entities and a specific interest [31]. It creates a hierarchy from general to specific terms populated at each level by classes. These classes within the structure of the hierarchy are used to organize data, and can help aggregate information across platforms. To do so relationship types were generated to better describe the interconnections that exist within the hierarchy of terms. The first relationship is **part_of**, which is a relation of parenthood. The level follows from the one above. The second is **derives_from** which means that one entity is made up of the product of another that occurred earlier in time [31]. The third is **has_participated** which relates entities contributing to or apart of the same process. This entity would be linked back to a more general term. For instance, L-plastin and actin participate in regulating the cytoskeleton, so they would both contain the has participated link to cytoskeleton regulation. The final and most specific level is **has_function**. This relation links the entity to its supposed function [31].

The concept of function is one that needs definition. GO defines function as a *selected effect function* meaning that the function of a biological entitity is the one for which it was selected during evolution. The second is *causal role function*, that the function is determined in regards its contribution to a system. *Selected effect function* considers an entity's function within evolutionary constraints and considers the function to be selected by evolution for the survival of the organism, making it more biologically meaningful. In addition to this investigations performed to understand a gene product are structured to determine its role in the survival of an organism. The literature GO uses to describe a gene already uses the *selected effect function* definition [31].

There are three aspects to GO: biological processes, molecular function, and cellular components. These aspects are built around the dogma that a gene codes a protein to perform some function. GO was developed so that a common language could be used to describe similar genes across species. Biological processes is the largest of the three aspects, and can contain information regarding molecular actions of a gene product, to the gene products' role in complex processes like cell migration. A biological process is defined by its end product, and contains specific processes to reach that end. The annotation for a protein coming from some gene product in the biological process category would say that the gene product does something that is a part of a process [31].

Molecular function is the activity of a gene product that can be carried out by a single protein by directly interacting with other entities. For instance, the gene LCP1 codes for L-plastin, an f-actin binding protein, whose GO molecular function terms include: actin binding, integrin binding, and protein binding. The cellular components aspect details the location of the protein relative to organelle or other structure inside the cell, where the gene product is thought to carry out its function. For instance, cell component terms for L-plastin include podosome, cytosol, cytoplasm, and cytoskeleton.

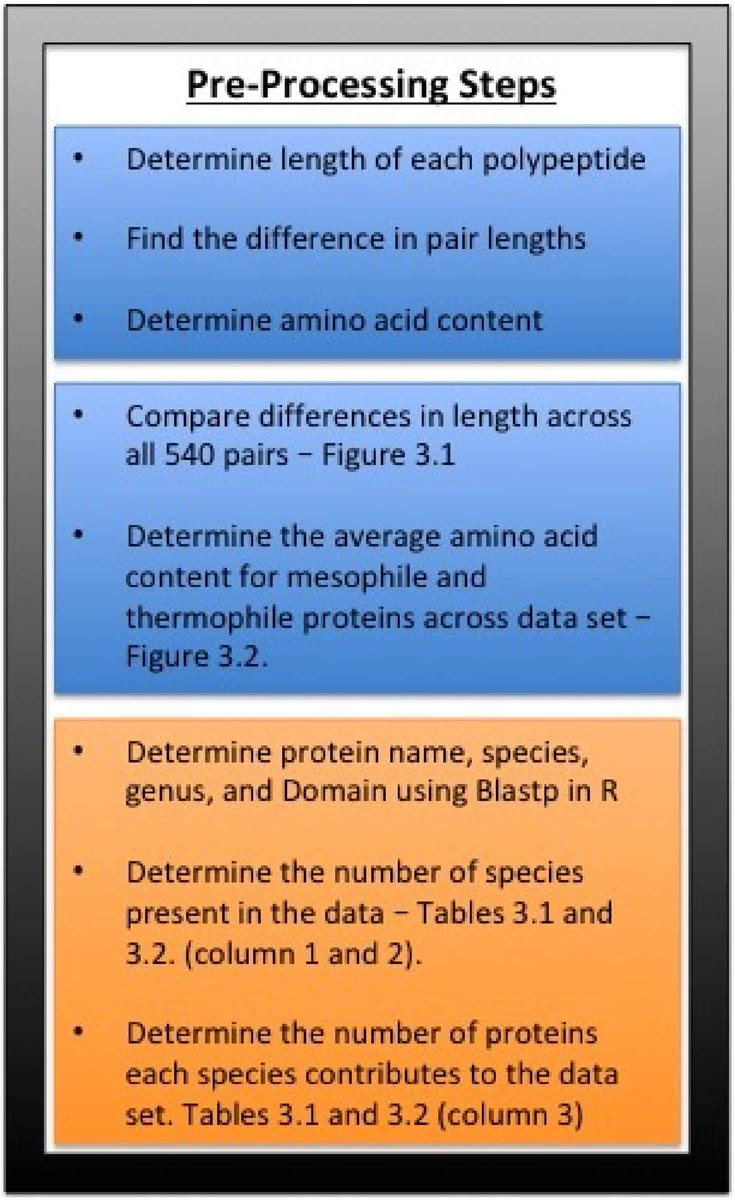
In this study, GO is used to determine if homologous pairs found to have similarities in the power spectrum of their wavelet decomposed hydrophobicity signals are known to have similar function. To determine this, a gene ontology must be carried out to gather information about the function of each of the homolog pairs, and determine if the groupings are biologically relevant.

The GO was carried out for this study by looking up the gene IDs of a protein using the Ensembl Bacteria. Emsembl is a genome browser that provides detailed information about a region of DNA. Since proteins are gene products, GO annotations are associated with the IDs of the protein that gene encodes. Ensemble bacteria contains annotations for the genomes of both Bacteria and Archea. The protein name was queried to determine the gene ID. This returns a list of potential matches, and the option coming from either the exact species, or if not species then genus, was selected. This allows the user to navigate to the genes biological process, molecular function and cellular component annotations.

## 3.7 Summary

Figure 3.3 provides information regarding the pre-processing steps performed on the data before application of the methods detailed in chapter 2. The length of each protein was determined, and the difference in the lengths of each protein in a pair was calculated. There is no tendency for one member of a homologous pair to be longer than the other. This allows for the calculation of total power as a pair specific distinguishing feature. The amino acid content of each protein was determined as a percentage of total polypeptide length. It was found that glutamic acid and lysine make up a larger percentage of the amino acid chain of thermophiles than mesophiles, and that glutamine makes up a smaller portion of the thermophile polypeptide than the mesophile.

The species, genus and domain of the proteins were looked up using the protein basic local alignment search tool provided by the National Center for Biotechnology Information. Figure 3.4 provides information regarding the post-processing procedures, those steps carried out after the application of methods detailed in Chapter 2. A pair-specific similar feature was found by comparing the power of the ranked order frequency bands resulting from the WPT between members of a homologous pair. Pairs where similar power was found were grouped by the frequency band. These steps would be included in the purple box, under the label *Spectral Analysis Methods*. Using the protein name, species, and genus information, a gene ontology can be carried out to determine if this feature is function specific. This is done by collecting gene IDs for the proteins in the subgroup and looking up GO annotations for biological process and molecular function. In this way the groupings are evaluated based of the characterization of the proteins from the literature. The literature definition of function is *selected effect function*, that defines the molecular function of a protein to be the one for which it was selected by evolution for the survival of the organisms. These annotations are used to determine if the groupings are biologically meaningful and function specific.



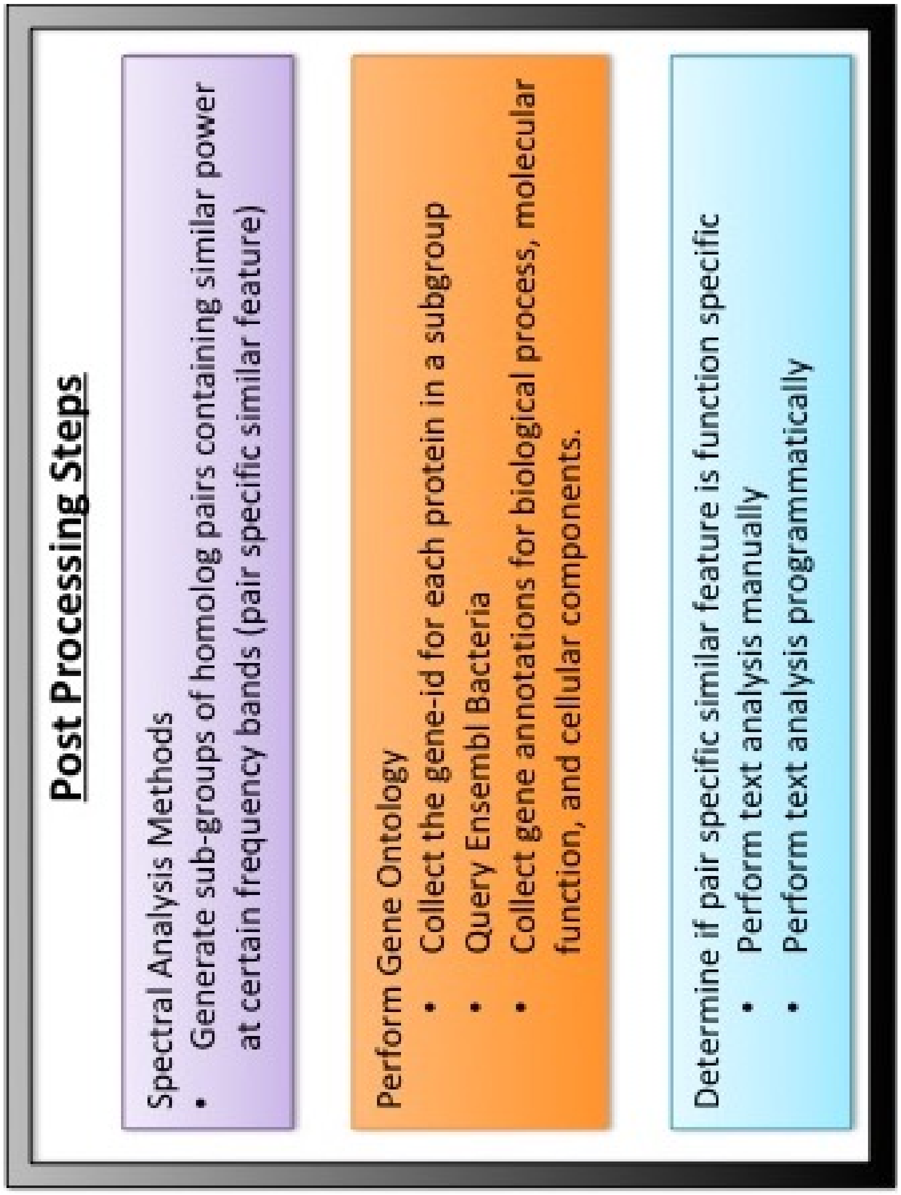
**Pre-Processing Steps**

- Determine length of each polypeptide
- Find the difference in pair lengths
- Determine amino acid content

- Compare differences in length across all 540 pairs – Figure 3.1
- Determine the average amino acid content for mesophile and thermophile proteins across data set – Figure 3.2.

- Determine protein name, species, genus, and Domain using Blastp in R
- Determine the number of species present in the data – Tables 3.1 and 3.2. (column 1 and 2).
- Determine the number of proteins each species contributes to the data set. Tables 3.1 and 3.2 (column 3)

Figure 3.3: The data was pre-processed to determine the amino acid content and length of each protein. The difference in the lengths of each protein in a homologous pair was also determined. This measure is used to determine if the total power in a hydrophobicity signal can be used as a pair specific distinguishing measure. The protein name, species and genus which produces that protein were determined using the Protein Basic Local Alignment Search Tool from the National Center of Biotechnology Information.



**Post Processing Steps**

Spectral Analysis Methods

- Generate sub-groups of homolog pairs containing similar power at certain frequency bands (pair specific similar feature)

Perform Gene Ontology

- Collect the gene-id for each protein in a subgroup
- Query Ensembl Bacteria
- Collect gene annotations for biological process, molecular function, and cellular components.

Determine if pair specific similar feature is function specific

- Perform text analysis manually
- Perform text analysis programmatically

Figure 3.4: Homologous pairs were grouped together based of a pair specific similar feature. The methods used to create the groupings are described in Chapter 2. To determine if this feature is function specific, a gene ontology is performed to gather information about each protein in the sub-groupings molecular function and biological process. These data are then analyzed either manually or pragmatically to determine if the proteins share anything in common beyond the pair specific similar measure.

# CHAPTER 4

# Results

## 4.1 Introduction

The focus of this study is to identify features related to thermal stability and function in the amino acid chains of short globular proteins from mesophile and thermophile prokaryotes. To do so, 540 homologous pairs of proteins, one from mesophile and the other from thermophile, were constructed. The proteins composing a pair perform similar functions in different temperature regimes. In this way, adaptations made to the linear amino acid chain related to thermal stability can be identified, since the proteins must fold within the constraint that they perform a certain function, through a comparison between members of a pair. Characteristics related to thermal stability are considered to be pair specific and distinguishing, while aspects related to function are pair specific and similar.

The hydrophobic effect is the variable being tested in this study. Each amino acid contains a residue with a hydropathy value scored by the Kyte-Doolittle hydropathy scale. This scale was developed taking into account both structural data and thermodynamic calculations. A description of the scale and the methods used to derive it can be found in Chapter 1 Section 4. A hydrophobicity signal is generated by assigning a value to each of the amino acids in a linear chain. To determine features related to thermal stability and function, both the amino acid content (hydorpathy content) and ordering of amino acids (hydropathy values) within the chain are considered.

An alignment was run to determine how similar the linear amino acid chains of each protein are to one another within a pair. Following this, the Anderson Darling test was used to see if the hydrophobic values within the chains are normally distributed. Next, the two sample Kolmogorov-Smirnov test was used to determine if the hydrophobicity signals within each pair have similar distributions of values within their lengths. These tests were followed by a scatter plot of the mean hydropathy of the mesophile versus the thermophile.

The hydrophobicity signals were then treated to a Wavelet Packet Decomposition. This methods generates localized frequency bands containing the fluctuation in the hydropathy along various lengths of the hydrophobicity signals. Spectral analysis methods were then used to determine how the coefficients within each band were distributed. These measures included finding the skew and kurtosis of coefficients within each leaf. This was followed by calculating the power, or average variance in the fluctuation in hydrophobic free energy, of each frequency band. The total power of each hydrophobicity signal was computed by summing the powers for each of the frequency bands resulting from the Wavelet Packet Transformation.

A comparison of total power was then performed by finding the difference in the total value of power between members of each pair. This was done by subtracting the mesophile total power away from the thermophile. Next, the power of each frequency band was compared between mesophile and thermophile. This was done by finding the difference in the power of each frequency band. Pair specific similar features were used to create groupings that included all pairs with those values. These subgroups were then used to perform a gene ontology to determine the molecular function and biological process of each protein in the group. This is done to determine if the pair specific similar measure is related to the function of the protein.

| Percentage Aligned | Count | Percentage Aligned | Count |
|---|---|---|---|
| 0 - 5 | 32 | 40 - 45 | 9 |
| 5 - 10 | 322 | 45 - 50 | 6 |
| 10 - 15 | 59 | 50 - 55 | 6 |
| 15 - 20 | 27 | 55 - 60 | 5 |
| 20 - 25 | 19 | 60 - 65 | 7 |
| 25 - 30 | 17 | 65 - 70 | 2 |
| 30 - 35 | 10 | 70 - 75 | 6 |
| 35 - 40 | 12 | 75 - 80 | 1 |

Table 4.1: Results of the Alignment. The majority of pairs have dissimilar amino acids chains, 322 pairs are less than 10% similar. Only one pair is between 75% and 80% similar. 459 of the pairs are less than 25% similar. This means that ordering of amino acids within each of the polypeptides of a pair are dissimilar.

## 4.2 Hypothesis Tests

Alignment is a measure of the similarity in the ordering of amino acids between proteins composing a homologous pair. This measure checks to see if the n-terminal amino acid in the mesophile is the same as the n-terminal amino acid in the thermophile and so on through the length of the shorter of the two proteins. If the amino acids are the same at that location a score of one is given, if they are different a score of zero is assigned. The total score is computed by summing the number of ones assigned, and that number is divided by the length of the shorter of the two proteins.

Table 4.1 shows the number of pairs that fall within a given percentage alignment. The vast majority of pairs have dissimilar orderings of amino acids, slightly more than 60% (322) of the proteins are less than 10% similar. 459 of the pairs are less than 25% similar. Different orderings of amino acids can be used to generate proteins with a certain function.

The Anderson Darling test (discussed in Chapter 2 Section 3) was used on the hydrophobicity signals generated for each polypeptide to determine if the hydropathy values within each signal were normally distributed. If the hydropathy values are normally distributed, then certain hydropathys contribute more to the linear amino

| — | Accept | Reject |
|---|---|---|
| Anderson Darling | 0 | 540 (mesophile and thermophile) |
| Two-Sample Kolmogorov-Smirnov | 530 | 10 |

Table 4.2: Results of Anderson Darling and Two Sample Kolmogorov-Smirnov tests. No hydrophobicity signal is normally distributed. 530 of the homolog hydrophobicity scales accepted the null hypothesis of the KS test. The distribution of hydropathy values in both members of a pair have the same underlying distribution. 10 pairs rejected the null hypothesis for the test, and have different distributions of hydropathy values within their lengths. The Anderson Darling and two sample Kolmogorov Smirnov tests are discussed in Section 2.3.

acid chain than others. Meaning that certain values of hydropathy may contribute more to thermal stability and function than others.

The first row of Table 4.2 presents the results of the Anderson Darling test, as discussed in Chapter 2 Section 3. The hydropathy values within the amino acid chains of the proteins are not normally distributed, making it unlikely that there is a specific value that can be used as a pair specific measure.

The Kolmogorov Smirnov Test was used to determine if the distribution of hydropathy values in the amino acid chains composing a homologous pair had similar distributions. The second row of Table 4.2 presents the results of the two sample KS test. 530 of the pairs accept the null hypothesis that the two hydrophobicity signals have the same distribution of hydropahty values. Only ten proteins (1.8% of the data set) rejected the null hypothesis that the two signals have the same underlying distribution of hydropathy values within their lengths. There is no significant difference in the distribution of hydrophobic content in the hydrophobicity signals within each pair. The distribution of hydropathy values within the hydrophobicity signals cannot be used as a pair specific distinguishing feature.

Figure 4.1 displays a scatter plot of the mean hydropathy of the thermophile hydrophobicity signal versus the mean hydropathy values for the mesophile signal. The distribution appears to be centered at $(-0.2, -0.2)$. The majority of points lie in the third quadrant, indicating that both the mesophile and thermophile average value are negative. The data appear to follow a linear trend, indicating that the

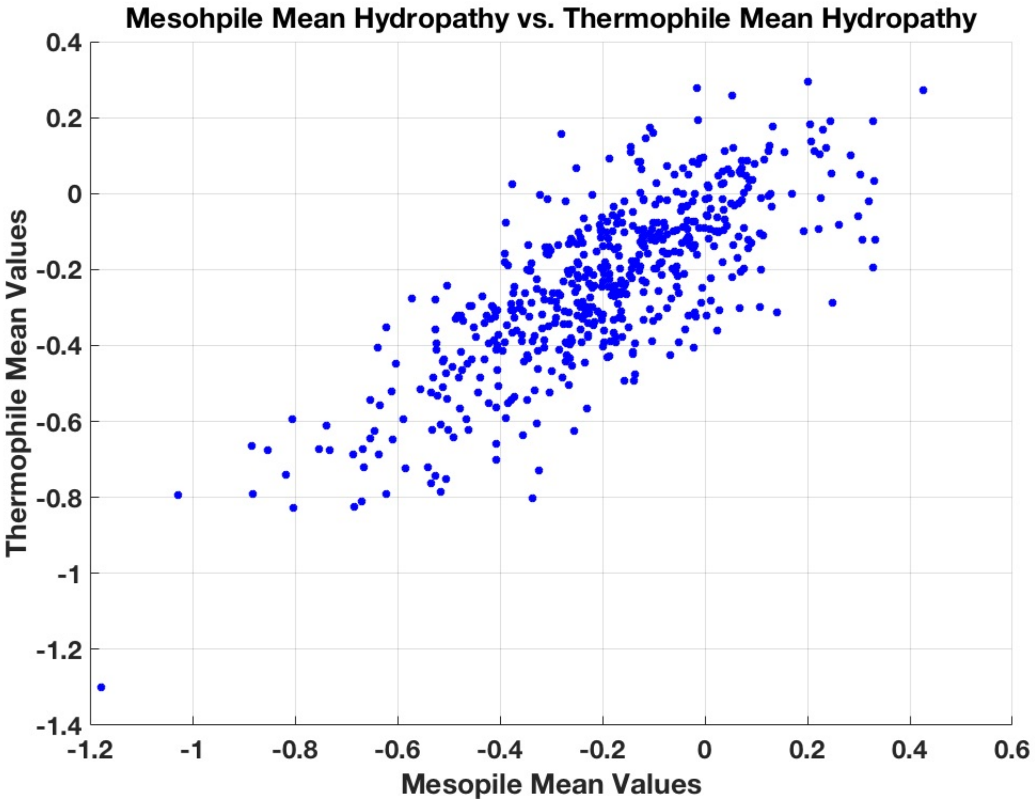


Figure 4.1: Thermophile mean hydropathy versus Mesophile mean hydropathy scatter plot. There is a single cluster of values centered around $(-.2, -.2)$. There is no tendency for either a mesophile or thermophile to have a larger average hydropathy.

average hydropathy of the mesophile and thermophile are similar.

## 4.3 Total Power

Spectral analysis methods were applied to the resulting frequency bands of a wavelet packet transformation (WPT) applied to each of the members of the homologous pairs. The wavelet and wavelet packet transform are detailed in Sections 2.5 and 2.6 of Chapter 2. The first measure was to determine the power of each frequency band resulting from the WPT. Next, the sum of each of the powers was taken to determine the total power of the hydrophobicity signal. This measure was used to determine the difference in power between members of a pair, and to see if it could be used as a pair specific distinguishing feature.

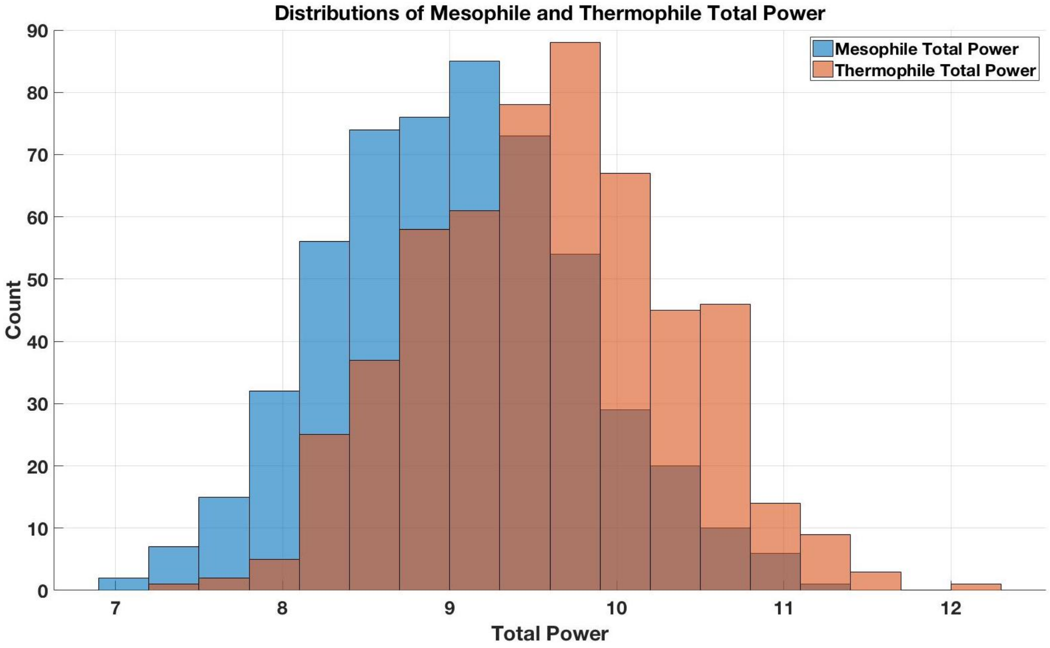


Figure 4.2: Histograms for the total power in the mesophile and thermophile hydrophobicity signals. The mesophile total power distribution is in blue, and the thermophile distribution is in red. The two distributions overlap, the thermophile distribution is translucent, and makes it appear that there is a third distribution. There is no tendency for either a mesophile or thermophile protein to take on a specific value for total power.

Figure 4.2 displays the distribution of the total powers in the mesophile and thermophile proteins. The mesophile total power distribution is in blue, and the thermophile distribution is in red. The two distributions overlap, and there is no tendency for a mesophile or thermophile protein to have a specific value for total power.

Figure 4.3 is a histogram of the difference in total power of each pair, computed as the total power in the thermophile minus the total power in the mesophile. This figure shows that the thermophile member of each pair has a tendency to contain more power in its hydrophobicity signal than the mesophile. There are 489 values greater than zero, meaning that the thermophile contains a greater total power in 90.56% of the pairs. Total power is a pair specific distinguishing feature. It is interesting to note that the largest positive difference is 2.37 while the most negative difference is $-0.69$. This shows that the total power of the mesophile and thermophile in each pair is very similar, but that there is a dominant trend for the thermophile to be more powerful than the mesophile.

Figure 3.3 in Chapter 3: Data Overview, shows that there is no trend for either the mesophile or thermophile protein to be longer than the other member of its pair. The difference in total power is then due the amino acid content within each of the polypeptides. The magnitude of the total power can only be used to distinguish members of a pair, there is no tendency for thermophiles to take on a specific value. It is possible that the total power is encoded to reflect thermal tolerance within the constraint that the protein folds to perform a certain function, and is not directly related to thermal stability. The distribution of total power of the mesophile proteins overlaps with the distribution of the total power of the thermophile. The total power of the thermophile is only larger in comparison to a similarly functioning mesophile protein, not just any mesophile protein.

Figure 4.4 provides a scatter plot of total power for each of the pairs, thermophile versus mesophile. The values seem to be increasing linearly so that the value of the total power in the thermophile is matched by the mesophile. The data in this scatter plot shows that the total power tends to be very similar between the mesophile and thermophile. There appears to be a single cluster of values centered around (10, 10.5). This shows that neither mesophile nor thermophile has a tendency to contain a certain value for total power. Rather total power seems to be related to the function of the protein.

The data presented in Figures 4.2 and 4.3 show that the thermophile hydrophobicity signal contains a greater total power in 90% of the data. However the thermophile contains only slightly more power, at most by a value of 2.37, and in some cases the mesophile contains more power (in one instance by 0.69). The distributions of the total power in the thermophile proteins overlaps with the distribution of total power in the mesophile proteins, as shown in Figure 4.2. This makes total power a pair specific distinguishing feature, meaning that a thermophile protein only contains a greater total power than a similarly functioning mesophile protein. To

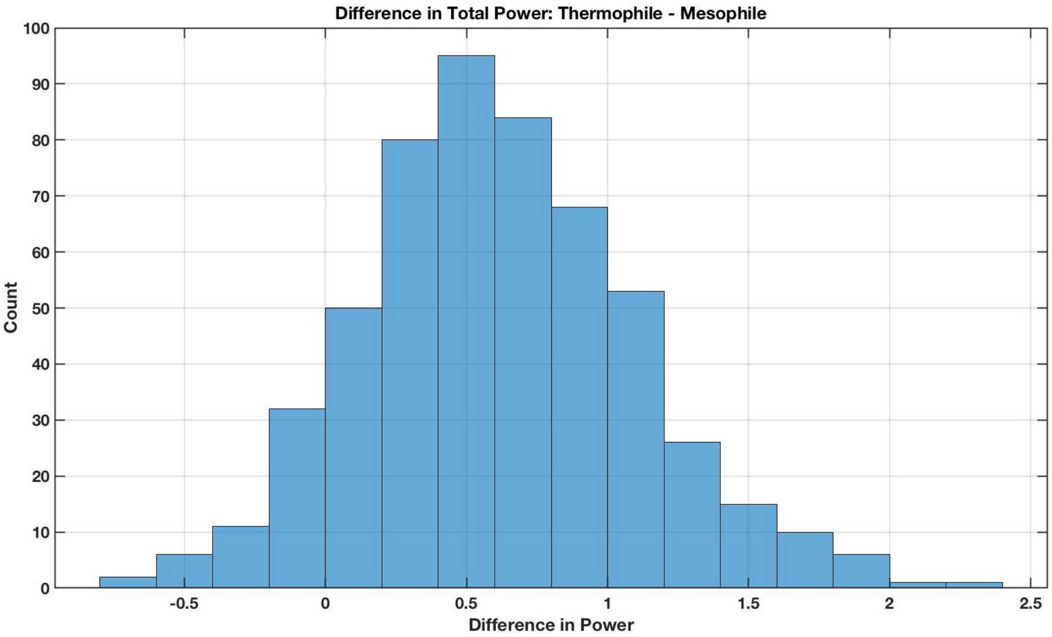


Figure 4.3: Difference in the total power of each hydrophobicity signal in a pair. The value was computed by subtracting the total power of the mesophile signal from the thermophile. Positive values indicate that the thermophile is the more powerful signal. The thermophile is the more powerful signal in 489 of the pairs (90.5%).

better understand the reason for this small difference in total power, a skew spectra, kurtosis spectra, and power spectra were generated for each of the proteins.

## 4.4 Spectral Analysis

To better understand how hydrophobic information is encoded into the hydrophobicity signals, a spectral analysis is performed. This allows for an analysis of the various frequencies making up the signal at locations along its length. Three spectral analysis methods were used in this study: power, skew, and kurtosis. For each of these three measures confidence intervals were generated by shuffling the ordering of hydropathy values within the signal and performing the wavelet packet transformation on that iteration. This procedure was run 299 times for each hydrophobicity signal. Each spectral analysis method was then preformed on the frequency bands resulting from the WPT so that each frequency band for every protein has 299 bootstrapped replicates. In this way a distribution function was chosen by analyzing the distribution of values at each frequency. The expectation value was computed using the descriptive parameters for the specific distributions, and the lower and upper intervals were determined by using the corresponding values of the descriptive parameters.

Figure 4.5 provides the power of each frequency band for the proteins of pair ten. The top graph is the mesophile hydrophobicity signal and the bottom graph is the thermophile. Both hydrophobicity signals were decomposed to the same level, determined to be two above the maximum decomposition level of the shorter of the two proteins, so that they contain the same number of frequency bands. The confidence intervals for the proteins were determined using the gamma function, whose expectation values is computed as the product of its shape and scale parameter. The gamma function was chosen using the method detailed in Section 2.9. The power of each frequency band was computed as the mean of the square of each coefficient in the frequency band, so that power is a positive quantity. Normalization was performed so that the values of power at each level could be compared. This is done by

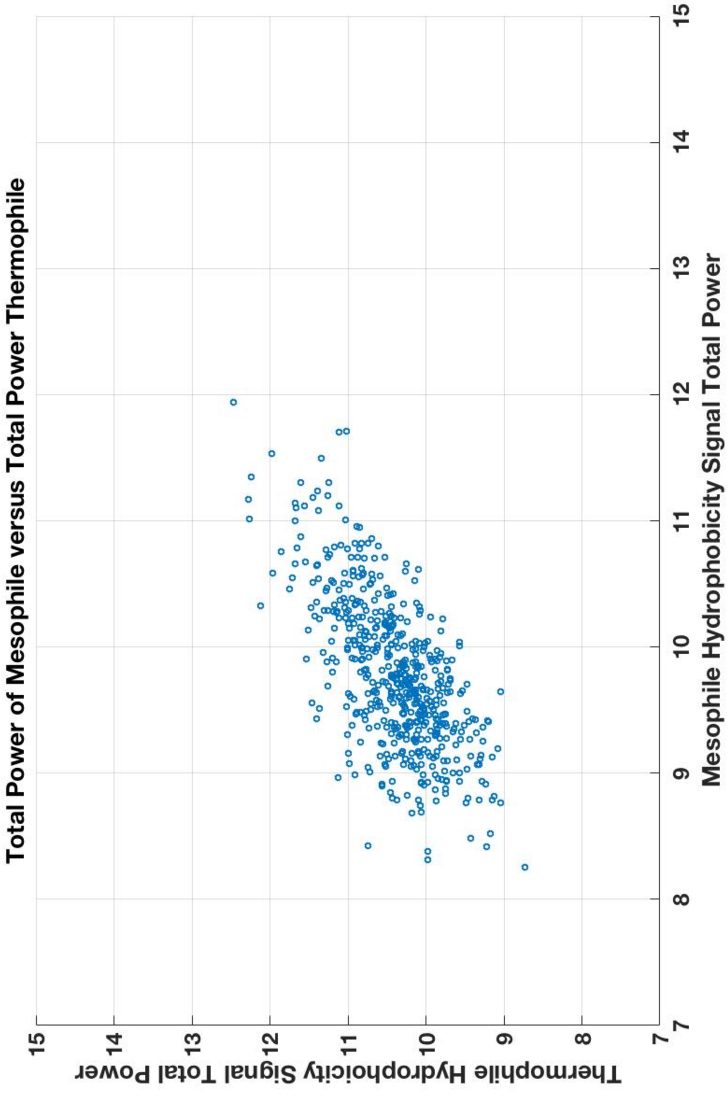


Figure 4.4: Total Power Scatter Plot, Thermophile versus Mesophile. The thermophile member of each homologous pair is more powerful than the mesophile member in 90% of the pairs. This scatter plot appears to have similar characteristics as the scatter plot of mean hydropathy of mesophile versus thermophile. However the mean hydropathy could not be used to determine a pair specific distinguishing feature.

multiplying the power of each leaf by $\frac{1}{2^j}$, where $j$ indicates the decomposition level.

Comparing the confidence intervals to the actual value for power in Figure 4.5 shows that the ordering of amino acids within the polypeptide codes for a specific value of power that depends on the ordering of amino acids. Power, as computed in this study, is the average of the squared fluctuation in hydropathy at some frequency, for the entire protein. So each value of power describes the average fluctuation in hydropathy for that frequency band through the length of the polypeptide. A random sequence of amino acids produces a power that is close to zero. While the sequence of amino acids in the proteins used in this study, most often encodes a value that is significantly non-zero. It appears that the fluctuation in hydropathy along the length of the protein is non-random, and may encode information related to function and thermal stability.

Figure 4.6 provides the skew of the values within each frequency band. The skew was computed on the frequency bands to determine the balance of the values of the coefficients within that leaf. Skew is discussed in Section 2.7. In this study skew becomes a function of frequency. In this way additional information is provided regarding the symmetry of the distribution of the fluctuation in hydropathy along the length of the protein. The number of coefficients within the leaf depends on the level of the decomposition, which was determined by the length of the shorter of the two proteins in the pair. The decomposition level varies across the 540 groupings, but at a minimum there are at least 11 coefficients in the leaf where skew is being measured. Skew is used to see if the distribution function of the coefficients within the leaves is asymmetric. If there are more coefficients that have values less than the expected value, the distribution is negatively skewed. If there are a larger number of coefficients with values greater than expected, the skew is positive.

The expectation values and confidence intervals for the skew of each frequency band were generated using a normal distribution. These data show that a random sequence of amino acids produces a frequency band with a skew value given by a normal distribution centered about zero. The ordering of amino acids in the polypeptides produces a distribution of coefficients within the leaves that are most

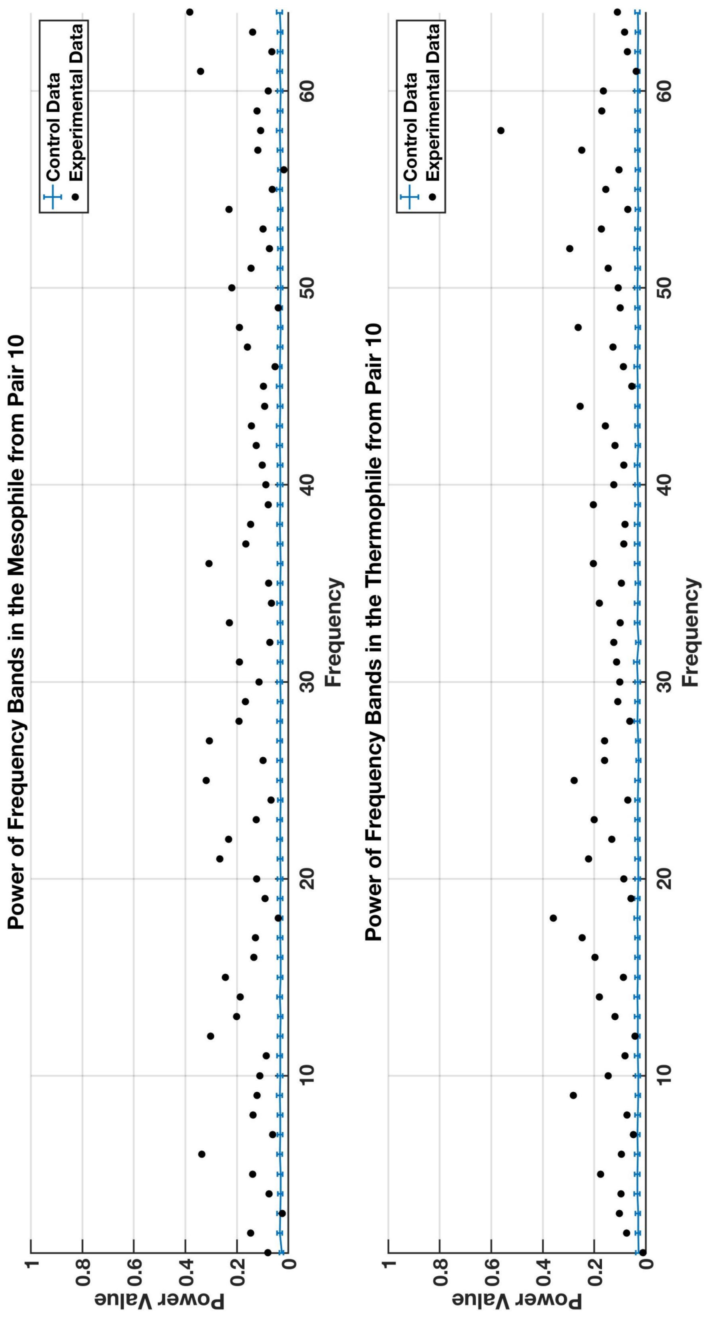


Figure 4.5: Power of each frequency band for the members of homologous pair 10. A control data set was generated by shuffling the ordering of amino acids in the linear chain 299 times, and then treating each iteration with the WPT. A random ordering of amino acids will most often produce a value for power that is near zero. The amino acid chain of each of the polypeptides codes for a non-random value of power. The variance in the fluctuation of hydropathy is non random and encoded by the linear amino acid chain.

often asymmetric. The asymmetry indicates that the fluctuation in hydropathy varies, and the value of the coefficients depends on the locations along the length of the protein at which they were calculated. This shows that fluctuation in hydropathy varies along the different lengths over which the coefficients were calculated.

Figure 4.7 provides the kurtosis of each of the frequency bands for the proteins of pair ten (kurtosis is discussed in Section 2.7). This study takes the kurtosis of each frequency band that results from the WPT. In this way kurtosis becomes a function of frequency, and additional information becomes available in regards to the spread of the value of the fluctuation in hydropathy within the hydrophobicity signal. The top graph is the mesophile and the bottom is the thermophile. Confidence intervals were generated based off a log normal distribution as determined by the values generated from the control data. The expected values generated from the bootstrapped data indicate that a random order of amino acids will produce a distribution of values in the frequency bands that has fewer outliers than normal. Most of the coefficients are similar in value, and are close to the expected value. The kurtosis values generated from the frequency bands of the actual proteins are non-random. The ordering of the amino acids in the polypeptides produce frequency bands with kurtosis indicating that the fluctuation in hydropathy changes along the amino acid chain. Values of kurtosis that are greater than two and less than or equal to three indicate that the spread of coefficients in the frequency bands approach normality. Values less than two indicate that the values in that frequency band are fairly similar. While values of kurtosis greater than three indicate that the coefficients in the frequency band have a wider range of values. For those values less than two, the coefficients are more similar, meaning that the fluctuation in hydropathy is fairly consistent at that frequency band across the length of the protein. Those values of kurtosis greater than two indicate that the fluctuation in hydropathy within that frequency band varies more across the length of the protein.

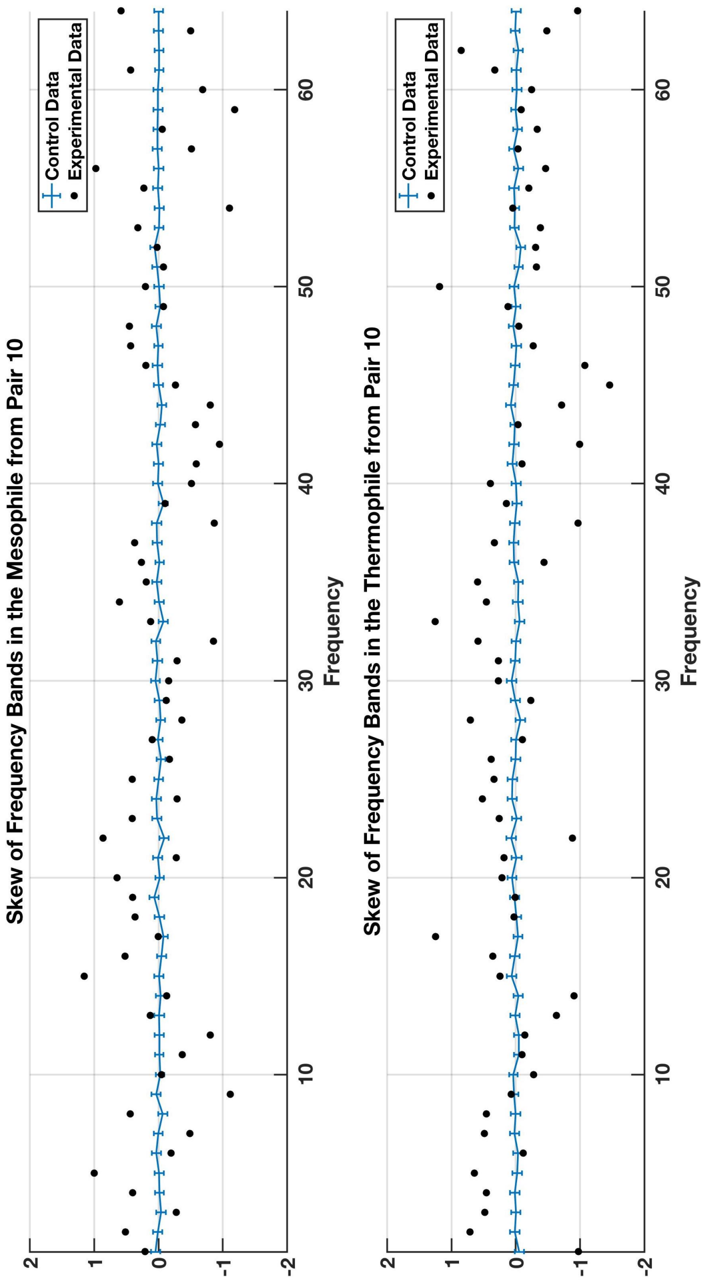


Figure 4.6: Skew of each frequency band for the members of homologous pair 10. Skew in this study is used to determine the symmetry of the distribution of coefficients within each frequency band. A non-zero skew indicates that the fluctuation in hydropathy varies within the protein, and that the values of the coefficients depends on the location within the hydrophobicity signal where the fluctuation was calculated.

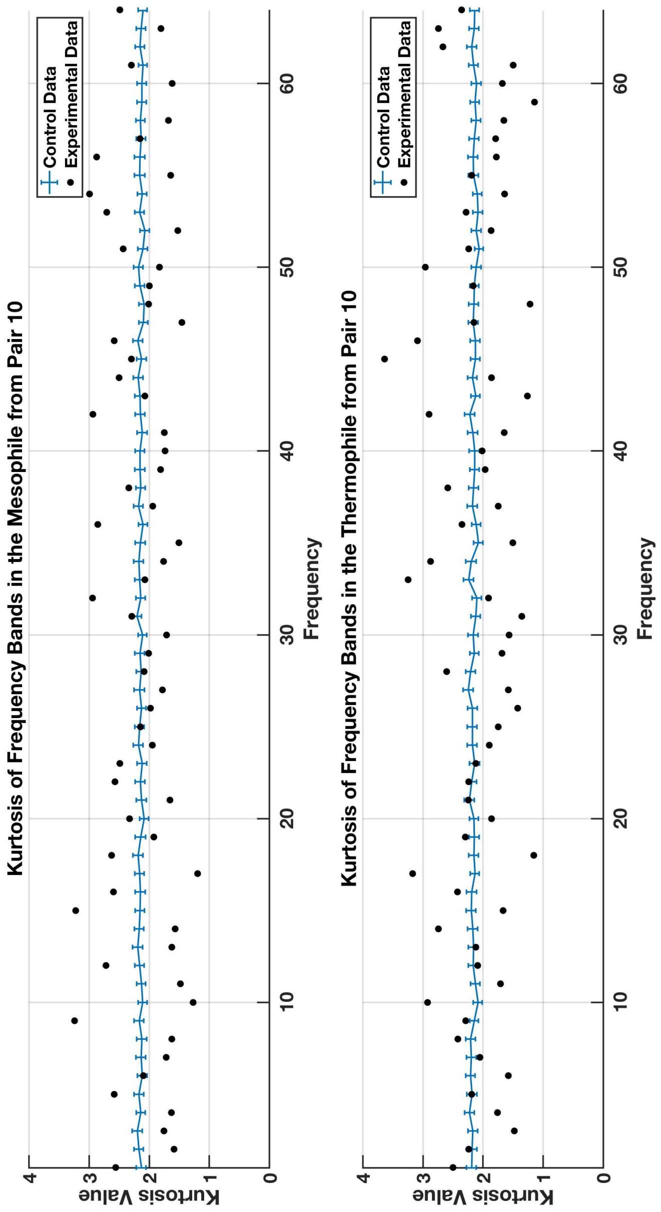


Figure 4.7: Kurtosis of each frequency band for the members of homologous pair 10. Confidence intervals were determined using a control data set consisting of 299 bootstrapped iterations of kurtosis for each frequency band. Kurtosis is used to understand the range of values included in each of the frequency bands. Values of kurtosis less than three indicate that the values within the frequency bands are similar. Values of kurtosis greater than three indicate that there is a greater spread in the values within a frequency band. Kurtosis is a measure of how the fluctuation in hydropathy changes along the length of the protein given some unit interval of amino acids.

## 4.5 Power and Thermo-tolerance

Knowing that thermophiles contain a slightly greater total power than their corresponding mesophile partner, the next step is to determine how each of the frequency bands contribute to this total. The goal is to see how the power of each frequency band varies with temperature. To do so, the powers of each hydrophobicity signal were collected for a specific frequency band, and then ranked in ascending order. So that there are 540 values of power coming from the mesophile ranked from smallest to largest, and 540 values of power from the thermophile ordered in the same way. These values are then used to generate plots of the mesophile power versus thermophile power at a single frequency band. This analysis tests how the temperature of a prokaryote's environment affects the power in its protein at a specific frequency band, ignoring the function of the proteins. This analysis was run for the first 64 frequency bands. Not all of the pairs contribute to every frequency band's plot since one of the members of the pair is not long enough to be decomposed to the sixth level.

Figure 4.8 shows the plots of thermophile power versus mesophile power ranked in ascending order for frequency bands 7, 15, 16, and 20. These plots include a line with a slope of one for reference. The frequency bands included were chosen at random to provide an overview of the method. The x-axis in these figures is the mesophile power, and the y-axis is the thermophile power. It can be seen in all four of the graphs that as the power in the mesophile increases there is an associated increase in power in the thermophile. The plots show that the majority of data follow a linear relationship. At higher levels of power the data deviate from linearity, in frequency band 7 the data follows a straight line with a slope of one until the mesophile power reaches 0.4. After this point the mesophile power increases more rapidly than the thermophile. The slope of the best fit line for this frequency band is 0.98.

In frequency band 15, the data deviate from a straight line with slope one around a power with magnitude .18. This shows that the thermophile power increases more rapidly than the mesophile. The slope of the best fit line for this is 0.96. Frequency band 16 shows similar features as frequency band 15. At first the values seem to track one another, and then the thermophile power grows more quickly than the mesophile. The slope of the best fit line is 1.11. Frequency band 20 shows a similar trend, where the values begin to diverge from linearity around 0.2. The thermophiles appear to have slightly more power at this frequency band than the mesophile. The slope of this frequency band's best fit line is 1.08.

The slopes of the best fit line for each frequency band are displayed in Figure 4.9. There are 46 frequency bands with slopes greater than one, indicating that the thermophile proteins at those bands have slightly more power than the corresponding mesophile protein. 18 of the slopes are less than one, indicating that the mesophile proteins have greater power than the thermophile. The smallest value of slope is 0.72 and the largest is 1.28. There are 23 slopes that have a value between 1 and 1.1 (about 36% of the frequency bands), 15 slopes that are between 1.1 and 1.2 (23% of frequency bands), and 8 that fall between 1.2 and 1.3 (12.5% of frequency bands). There are 13 slopes whose value fall between .9 and 1 ( 20% of the frequency bands), 3 that fall between 0.8 and 0.9 (4.7% of frequency bands) and two that fall between 0.7 and 0.8 (3.1% of frequency bands).

In addition to this, it appears that thermophile proteins consistently contain more power at certain ranges of frequency bands. There appears to be two general groups in this figure that show this. Those groups include frequency bands from 21 to 31 (specifically frequency bands $22, 24, 25, 26, 27, 28, 29, 30$, and 31) corresponding to roughly 100 to 77 amino acid long unit intervals and between frequency bands 44 to 56 (specifically frequency bands $44, 46, 47, 48, 49, 50, 52, 53, 54, 55, 56$) corresponding to roughly 44 to 19 amino acid long unit intervals. There appears to be one group of frequency bands where the mesophile contains more power, between 18 and 22 (specifically frequency bands $18, 19, 21, 22$) corresponding to unit lengths of 107 to 97 amino acids.

An additional test was run to examine the power of each frequency band that considers the function of the protein. This was done by taking the difference in power

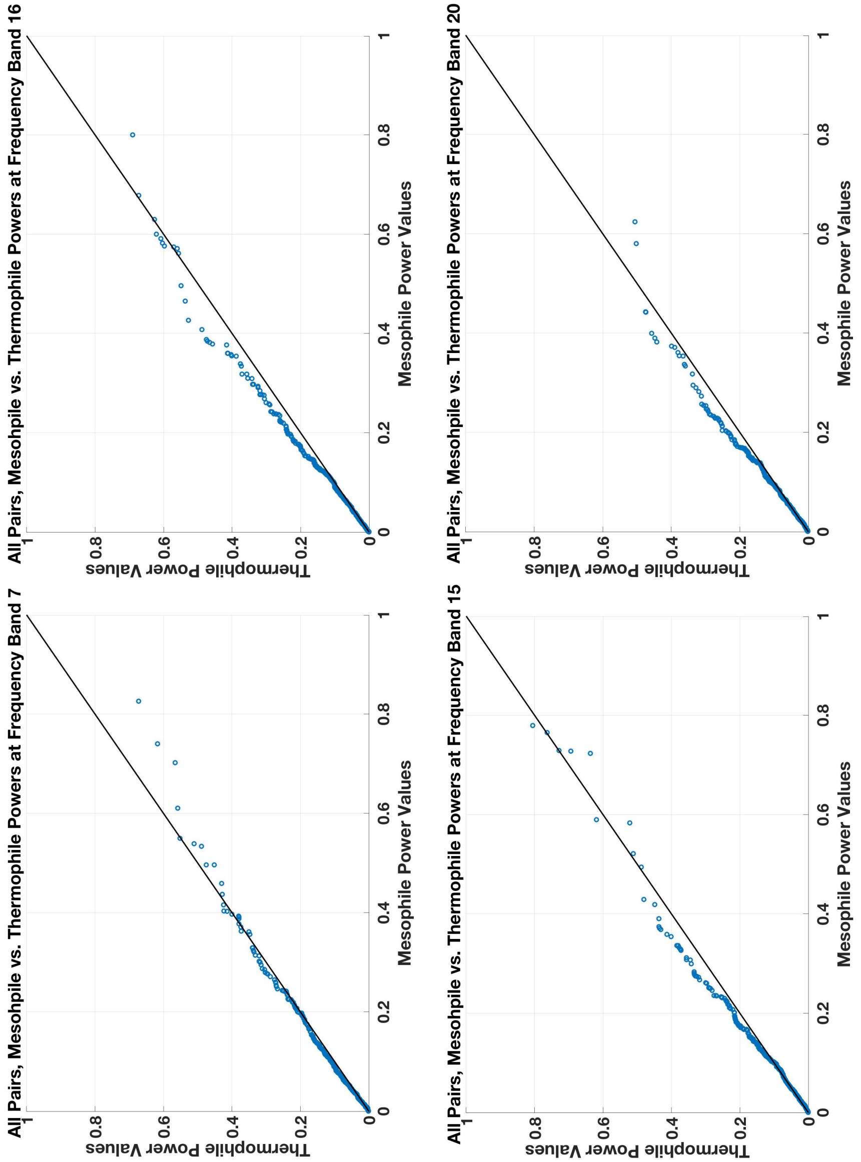


Figure 4.8: Four scatter plots of the value of power for each mesophile and thermophile protein. Frequency bands include bands 7, 16, 15, and 20. The values of power were ranked in ascending order and plotted. The purpose of these plots was to determine if the slopes of the line of best fit is one. A slope that is greater than one indicates that the thermophile tends to have a greater power than the mesophile at that frequency band. A slope less than one indicates that the mesophile contains a larger power.

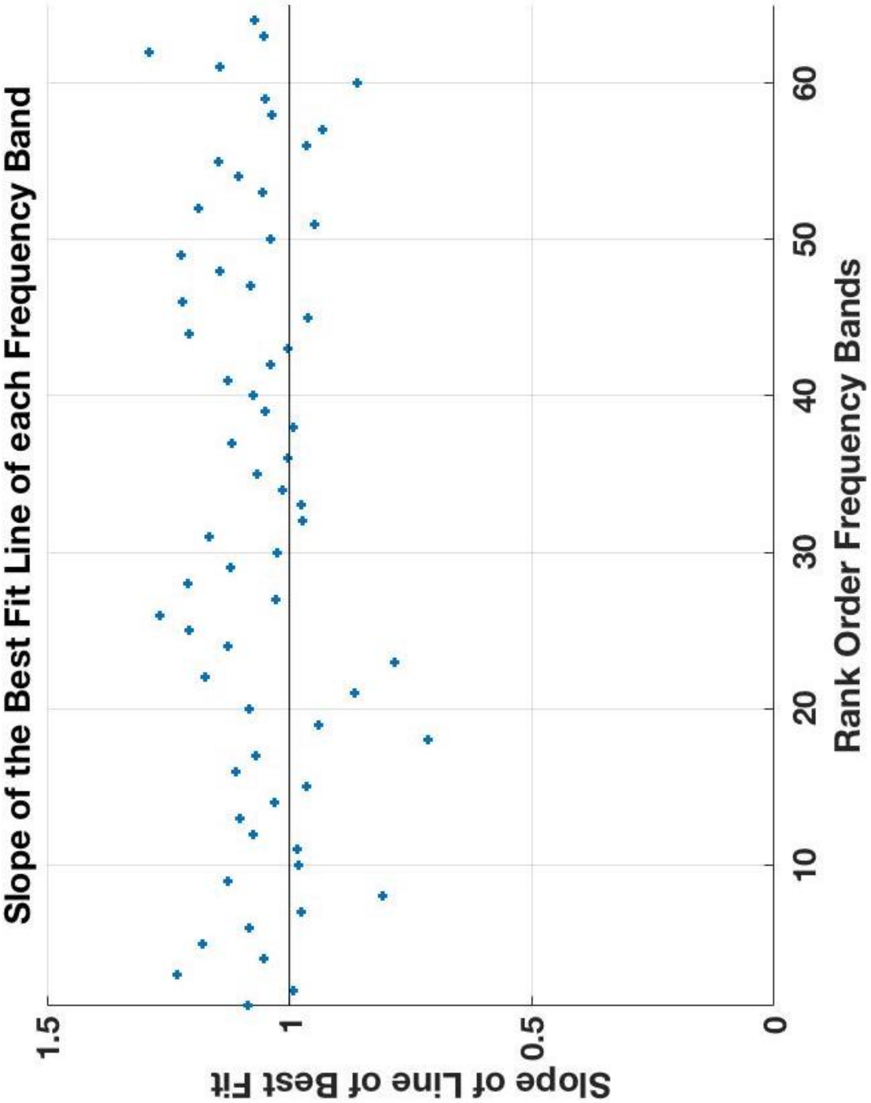


Figure 4.9: A plot of the slope of the best fit line for each frequency band. Thermophile proteins have a larger power than the mesophile in 46 of the frequency bands, as indicated by the slope being greater than one. There appear to be two groups of frequency bands, 22 to 31 and 44 to 56 where the thermophile consisting has more power.

at each frequency and finding the number of positive values at each frequency. A positive difference indicates that the thermophile contains a greater power at that frequency band. The number of positive difference in power at each frequency band is presented in Figure 4.10.

The orange line in Figure 4.10 indicates that in 270 of the pairs the mesophile contained a greater power than the thermophile. If the ratio is above 0.5 than the thermophile partner contained a greater power at that frequency band in more than 270 of the pairs. The largest ratio is about .58, and the smallest is about .47. The ratios are scattered around .5, however there is a dominant trend for the ratios to be above .5, indicating that the thermophile partner of the homologous pair contained more power than the mesophile.

## 4.6 Pair Specific Similar Features

To determine if the power of any frequency band could be used to infer the function of the proteins, the difference in non-normalized power was taken between the members of the pair at each of the frequency bands. This method compares the power of the frequency bands in each pair by taking the difference between the two. If the difference in power was less than one, the index order of that frequency band, *i.e* 1,2,3,..., was recorded. This was done to test whether certain frequency bands in the wavelet packet decomposed hydrophobicity signals were similar, and if this similarity could be used to infer function. In this way, power is evaluated as a pair specific similar measure related to the function of the homologous pair.

Figure 4.11 provides the number of pairs that had similar power at every frequency band, where each protein pair was decomposed to one level above maximum. The orange line was generated using the bootstrapped control data. It shows the number of matches at each frequency band that can be expected by a random ordering of amino acids. The orange line, which is the baseline likelihood that a random ordering of amino acids will result in similar power between members of a pair, was generated using control data. The same treatment was applied to the control data,

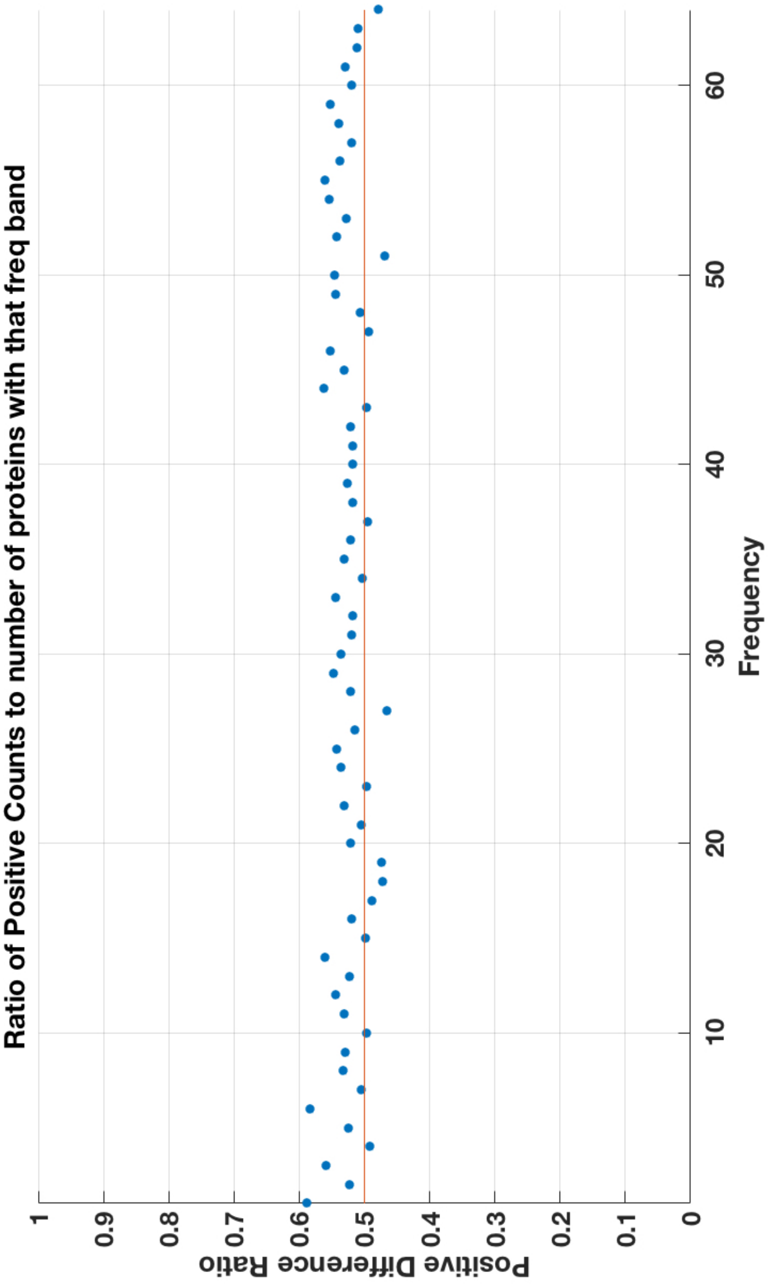


Figure 4.10: Ratio of the number of positive differences in power for 64 frequency bands. The difference in power at each frequency band was taken between members of a homologous pair. A value of 0.5 indicates that 270 of the pairs had a mesophile member with more power at that frequency, and in 270 the thermophile contained more power. This test is included to see how power at each frequency band compares when the function of the protein is taken into account.

and the number of counts was divided by the number of iterations run on this data.

For instance, each protein was shuffled 200 times, and treated to the wavelet packet decomposition. For each of those iterations the comparison was run. For each iteration of the bootstrapping procedure, the frequency bands that had similar power was recorded. This was done for every pair in the data set. The total number of matches for all bootstrapped iterations for every pair was then broken down by frequency band. So that the total number of matches across the 540 pairs for each of the 200 iterations was recorded. This number was then divided by 200 to account for the number of iterations. Figure 4.11 was generated using an earlier experimental design that ran the bootstrapping routine 200 times.

The black dots in Figure 4.11 are the number of pairs in the experimental data that had similar power at the frequency band given by the x-axis. The true ordering of the amino acids in the proteins generates a larger number of matches in power than if the amino acids were ordered randomly. In addition to this, at the highest frequency bands (above 120) there are instances where the experimental data do not contain any matches. There are only 16 homologous pairs long enough to be decomposed to this level. The homologous pairs tend to have a number of frequency bands with similar power, but that number is still a fraction of the total number of frequency bands they contain.

To generate groupings of homologous pairs containing similar power at a frequency band, an additional control data set was run. This was done to prune out those pairs that matched due to random chance. For instance, at frequency band 1 there are over 60 matches in the experimental data set. The control data, given by the orange line, indicates that about 20 of these matches are due to random chance. To flush out those pairs that match due to chance, the likelihood of finding the magnitude of the power at that frequency band was determined. This was done by generating an additional control data set.

This control data was run through a bootstrapping routine with a number of iterations given by ten times the length of the protein. This was done so that a sample could be generated containing the same characteristics as the population of

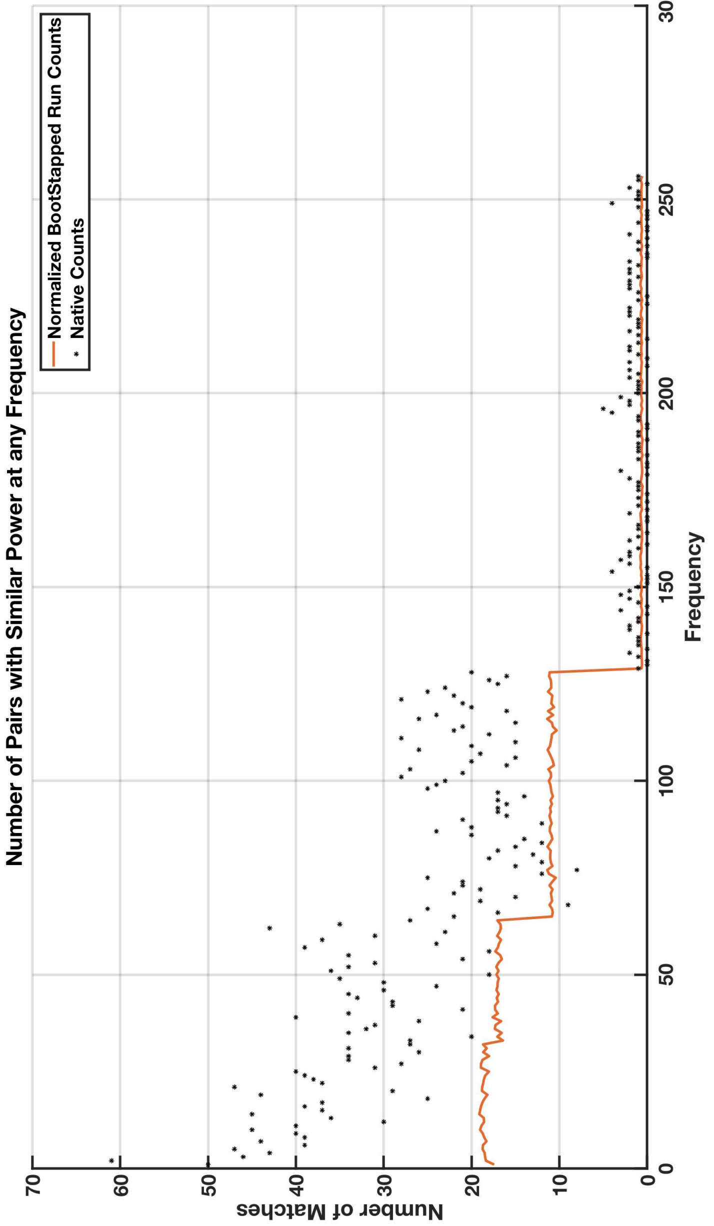


Figure 4.11: Number of instances at each frequency where members of a homologous pair have similar power.

all possible arrangements of hydropathy values within the signal. In this way, each protein had a number that is ten times its length for the power of that frequency band where there was a match. This was done so that the likelihood of measuring the value of power in the hydrophobicity signal at that frequency band could be determined. This was done by counting the number of times that specific value of power was calculated in the bootstrapped control data. If the likelihood of measuring that value for the power of the frequency band was less than 7% in either the mesophile or the thermophile, the pair was discarded from the subgroup. The most frequent value for power in the bootstrapped data had a likelihood of being found about 18% of the time for each protein.

A threshold of 7% was chosen as the cutoff for the following reason. If the magnitude of the power of a certain frequency band is related to the function of the protein, then evolution would act so that point mutations, which act similarly to the reordering of amino acids used in the bootstrapping routine, will have a negligible effect on the magnitude of the power at that frequency band. Otherwise once the protein is expressed, it will either denature, or have functioned differently. The mutation would then either prevent the protein from functioning, and would result in a phenotype onto which selective pressures would not act, or, be adapted for another function.

To determine if similarity in power at a specific frequency band is related to the function of the homologous pair, all the pairs that had similar power at the same frequency band were grouped together. This was done by generating a spreadsheet in excel that contained all the pairs, after pruning, that had been found to have similar power at a certain frequency band. A gene ontology was performed on two of these subgroups. Those pairs that had similar power at frequency bands 1 and 10.

The gene ontology was carried out manually. The first step involved querying the protein name and species info in Ensembl Bacteria. This resulted in a list of genes and species that were most similar to the query. The ID for the gene that codes for the protein was recorded. This was then used to query Ensembl Bacteria to look up molecular function and biological process gene annotations. This procedure was performed for the mesophile protein since the study assumes that members of the pair have the same function.

## 4.7 Gene Ontology: Frequency band 1 and 10

A gene ontology (GO) was performed on two of the subgroups generated from the matching power experiment. These subgroups are the proteins, after pruning, that had similar power at frequency band 1 and 10. The GO was performed to determine if similarity in power at certain frequency bands can be used to infer the function of the proteins. Figures 4.12 and 4.13 contain the information for those homolog pairs that had similar power at frequency band 1. Figure 4.12 provides the pair number, protein name, species, genus, and gene ID for that grouping. The gene ontology was performed on the mesophile species, since the thermophile has a homologous function. The species, genus, and protein name were obtained by querying the amino acid chain sequence in the Protein Basic Local Alignment Search Tool (BLASTp) maintained by the National Center for Biotechnology Information.

There are 35 pairs in the frequency band 1 grouping. There were 61 pairs that had similar power at this frequency band initially and 26 of these pairs were removed through pruning. Figure 4.13 contains the pair number, molecular function annotation, and biological process annotations for these proteins. Some proteins had multiple annotations for different processes, each of these terms were collected. In addition to this, an annotation for function rests inside a tree structure of other related terms. The terms immediately above the protein's annotation were included as well. In some instances, no information was available for a protein's molecular function and/or biological process. There are 3 proteins that could not be associated with a gene ID.

The second column of Figure 4.11 contains the molecular function of those proteins in the frequency band 1 group, six of them do not contain annotations for molecular function. There are 13 proteins that contain annotations for binding, this includes RNA binding, DNA binding, ion binding, and metal ion binding. There are 5 proteins that are structural constituents of ribosome, and one of these includes an annotation for binding activity. The third column of Figure 4.11 contains the biological process annotation for the proteins in this group. There are 10 proteins that do not have annotations for biological process. There are 5 proteins that have annotations for biosynthetic process. Five of the proteins have annotations for translation.

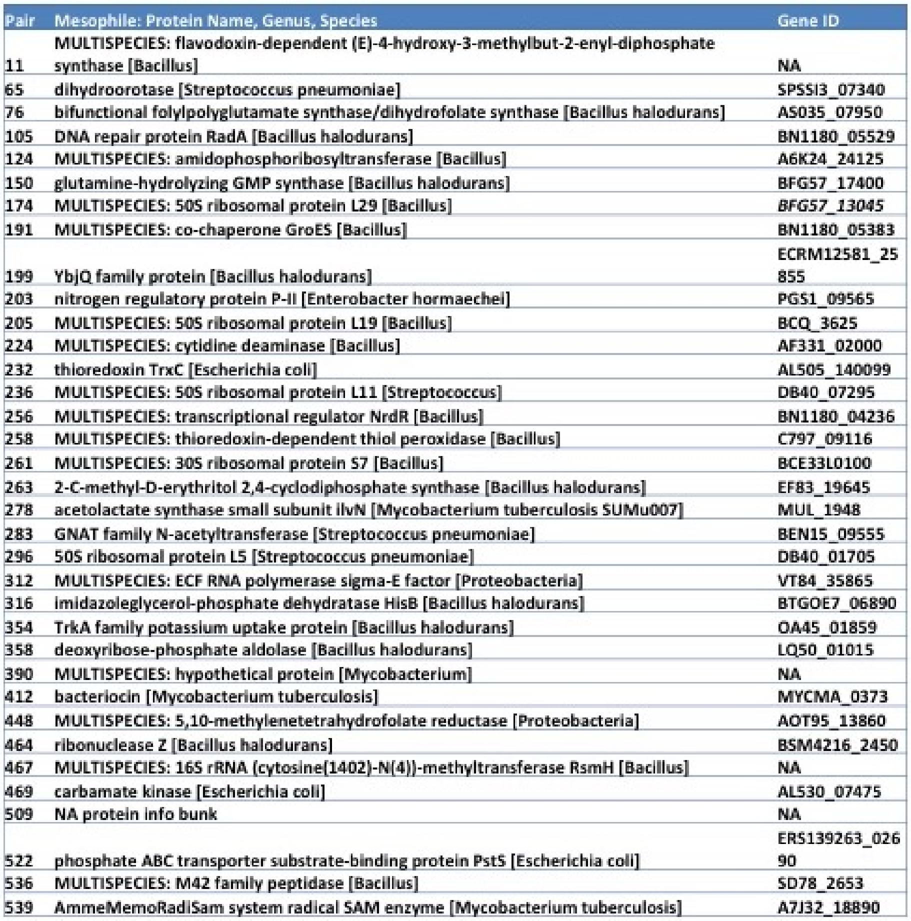
Figure 4.14 contains the pair number, protein name, genus, species, and gene id for the frequency band 10 grouping. There were 45 initial matches in this group, 19 of these pairs were pruned out. The gene ontology was performed on 26 proteins. Two of the proteins in this group, pair numbers 323 and 491 were not associated with a gene ID. Figure 4.15 contains the pair number, molecular function, and biological process annotations for these proteins. Three of the proteins in this group did not have annotations for molecular function. There are 12 proteins associated with the term binding, these annotations include ATP binding, pyridoxal phosphate binding, 4 iron, 4 sulfur cluster binding and nucleotide binding. Six proteins have annotations associated with lyase activity; these include hydro-lyase activity and lyase activity. Three of the proteins are associated with structural constituent of ribosome.

The third column of Figure 4.15 contains the biological process annotations for the proteins in the frequency band 10 grouping. Six of the proteins do not have biological process annotations. There are five proteins with annotations for translation, and four with annotations related to RNA. There are 8 proteins with annotations for biosynthetic process. There are 7 proteins with annotations for metabolic process.

There are a number of annotations that overlap between the two groups. This may mean that there is little variation in the 540 homologous pairs used in this data set in terms of function. Further work must be done to fully flesh out the gene ontology process included in this study. For instance, many of the biological process annotations are a part of a much larger tree containing multiple branches. As well there are four relationship types used in GO. These are **part_of**, **derives_from**,

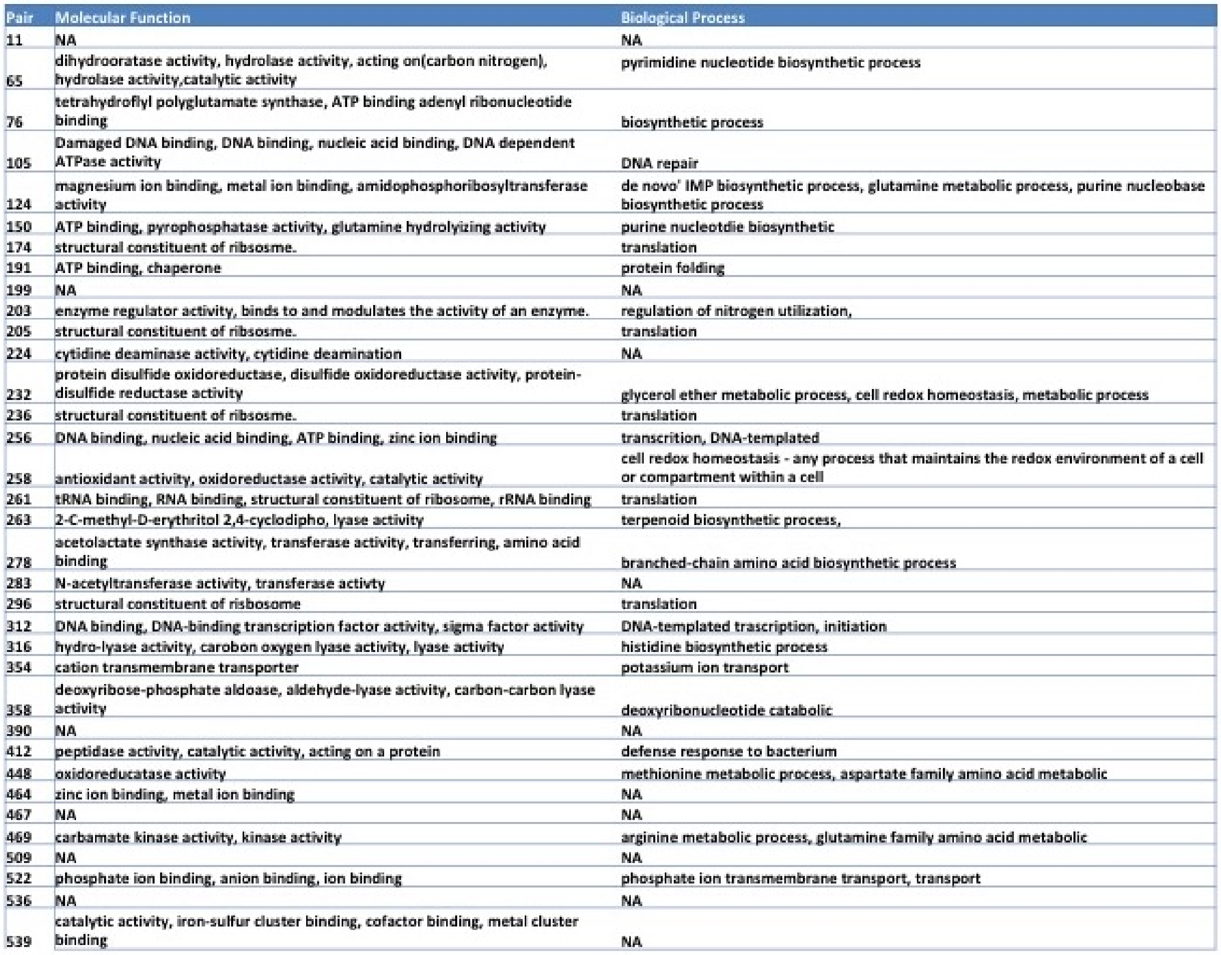
**has_participated**, and **has_function**. Including these relationships may help to infer the function and biological context for which these proteins are expressed in the prokaryotes.

The data presented in this gene ontology is the first iteration of the process of identifying characteristics in the spectra of the proteins related to function. The purpose of the gene ontology was to evaluate its ability to provide information related to the proteins used in this study. It appears that the gene ontology is a useful tool and should be expanded moving forward. The process of identifying similarities in the spectra of WPT decomposed protein pairs will require further computational methods, and a more developed gene ontology. This should further the aim of identifying pair specific similar features related to the function of these proteins.



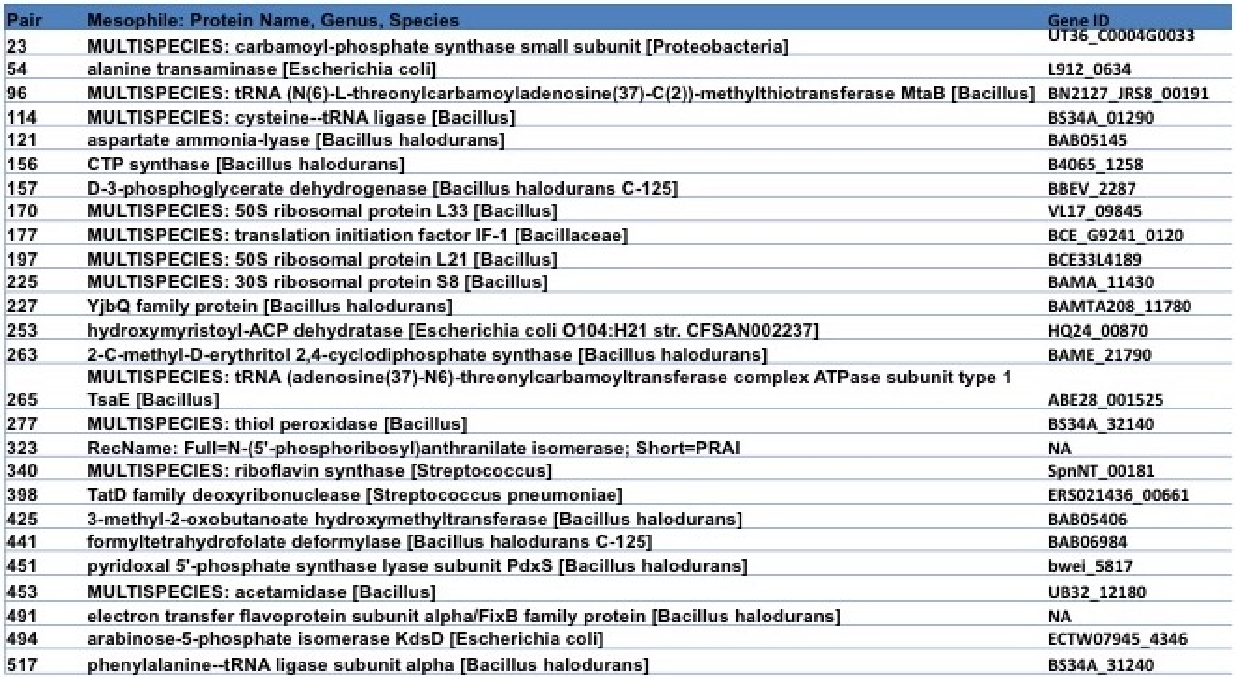
| Pair | Mesophile: Protein Name, Genus, Species | Gene ID |
|---|---|---|
| 11 | MULTISPECIES: flavodoxin-dependent (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [Bacillus] | NA |
| 65 | dihydroorotase [Streptococcus pneumoniae] | SPSSI3_07340 |
| 76 | bifunctional folylpolyglutamate synthase/dihydrofolate synthase [Bacillus halodurans] | AS035_07950 |
| 105 | DNA repair protein RadA [Bacillus halodurans] | BN1180_05529 |
| 124 | MULTISPECIES: amidophosphoribosyltransferase [Bacillus] | A6K24_24125 |
| 150 | glutamine-hydrolyzing GMP synthase [Bacillus halodurans] | BFG57_17400 |
| 174 | MULTISPECIES: 50S ribosomal protein L29 [Bacillus] | *BFG57_13045* |
| 191 | MULTISPECIES: co-chaperone GroES [Bacillus] | BN1180_05383 |
| 199 | YbjQ family protein [Bacillus halodurans] | ECRM12581_25855 |
| 203 | nitrogen regulatory protein P-II [Enterobacter hormaechei] | PGS1_09565 |
| 205 | MULTISPECIES: 50S ribosomal protein L19 [Bacillus] | BCQ_3625 |
| 224 | MULTISPECIES: cytidine deaminase [Bacillus] | AF331_02000 |
| 232 | thioredoxin TrxC [Escherichia coli] | AL505_140099 |
| 236 | MULTISPECIES: 50S ribosomal protein L11 [Streptococcus] | DB40_07295 |
| 256 | MULTISPECIES: transcriptional regulator NrdR [Bacillus] | BN1180_04236 |
| 258 | MULTISPECIES: thioredoxin-dependent thiol peroxidase [Bacillus] | C797_09116 |
| 261 | MULTISPECIES: 30S ribosomal protein S7 [Bacillus] | BCE33L0100 |
| 263 | 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase [Bacillus halodurans] | EF83_19645 |
| 278 | acetolactate synthase small subunit ilvN [Mycobacterium tuberculosis SUMu007] | MUL_1948 |
| 283 | GNAT family N-acetyltransferase [Streptococcus pneumoniae] | BEN15_09555 |
| 296 | 50S ribosomal protein L5 [Streptococcus pneumoniae] | DB40_01705 |
| 312 | MULTISPECIES: ECF RNA polymerase sigma-E factor [Proteobacteria] | VT84_35865 |
| 316 | imidazoleglycerol-phosphate dehydratase HisB [Bacillus halodurans] | BTGOE7_06890 |
| 354 | TrkA family potassium uptake protein [Bacillus halodurans] | OA45_01859 |
| 358 | deoxyribose-phosphate aldolase [Bacillus halodurans] | LQ50_01015 |
| 390 | MULTISPECIES: hypothetical protein [Mycobacterium] | NA |
| 412 | bacteriocin [Mycobacterium tuberculosis] | MYCMA_0373 |
| 448 | MULTISPECIES: 5,10-methylenetetrahydrofolate reductase [Proteobacteria] | AOT95_13860 |
| 464 | ribonuclease Z [Bacillus halodurans] | BSM4216_2450 |
| 467 | MULTISPECIES: 16S rRNA (cytosine(1402)-N(4))-methyltransferase RsmH [Bacillus] | NA |
| 469 | carbamate kinase [Escherichia coli] | AL530_07475 |
| 509 | NA protein info bunk | NA |
| 522 | phosphate ABC transporter substrate-binding protein PstS [Escherichia coli] | ERS139263_02690 |
| 536 | MULTISPECIES: M42 family peptidase [Bacillus] | SD78_2653 |
| 539 | AmmeMemoRadiSam system radical SAM enzyme [Mycobacterium tuberculosis] | A7J32_18890 |

Figure 4.12: Mesophile Proteins from pairs with similar power at frequency band 1. Included are the pair number, protein name, species and genus information, and gene ID.



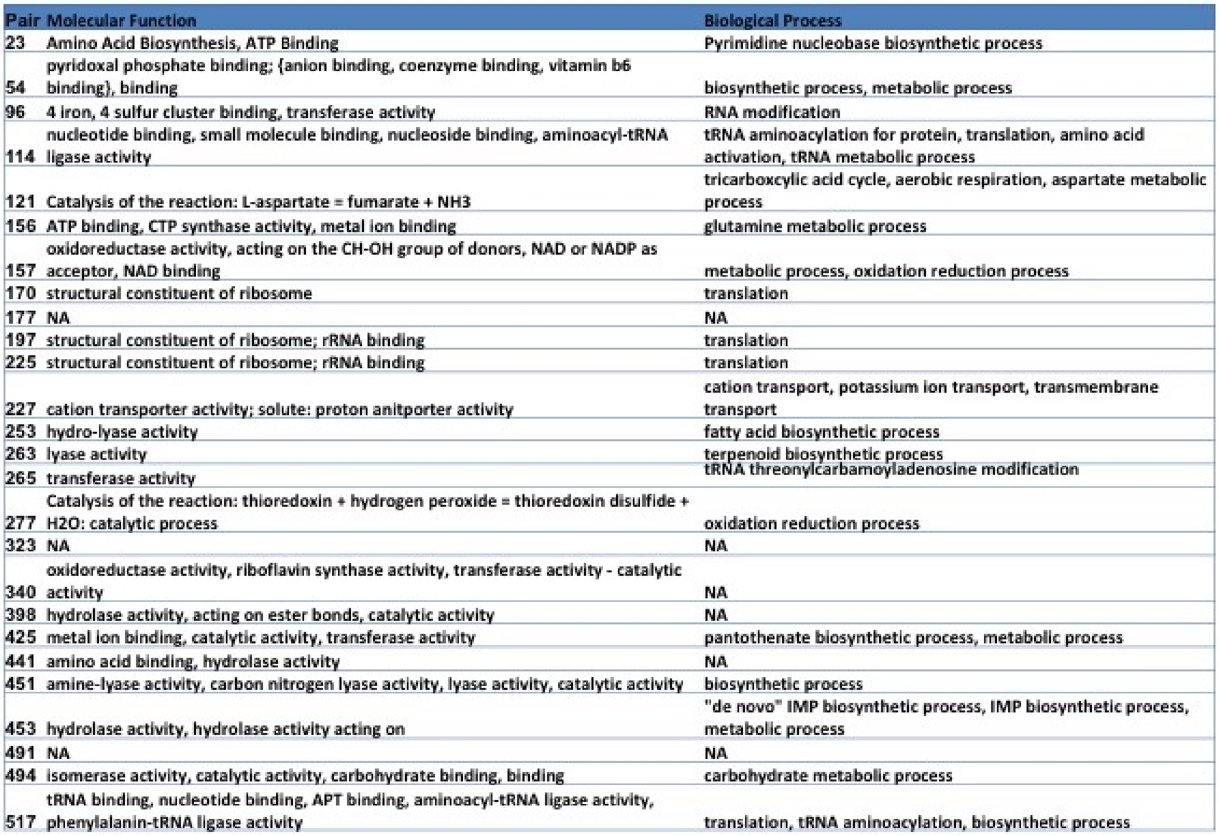
| Pair | Molecular Function | Biological Process |
|---|---|---|
| 11 | NA | NA |
| 65 | dihydrooratase activity, hydrolase activity, acting on(carbon nitrogen), hydrolase activity,catalytic activity | pyrimidine nucleotide biosynthetic process |
| 76 | tetrahydroflyl polyglutamate synthase, ATP binding adenyl ribonucleotide binding | biosynthetic process |
| 105 | Damaged DNA binding, DNA binding, nucleic acid binding, DNA dependent ATPase activity | DNA repair |
| 124 | magnesium ion binding, metal ion binding, amidophosphoribosyltransferase activity | de novo' IMP biosynthetic process, glutamine metabolic process, purine nucleobase biosynthetic process |
| 150 | ATP binding, pyrophosphatase activity, glutamine hydrolyizing activity | purine nucleotdie biosynthetic |
| 174 | structural constituent of ribsosme. | translation |
| 191 | ATP binding, chaperone | protein folding |
| 199 | NA | NA |
| 203 | enzyme regulator activity, binds to and modulates the activity of an enzyme. | regulation of nitrogen utilization, |
| 205 | structural constituent of ribsosme. | translation |
| 224 | cytidine deaminase activity, cytidine deamination | NA |
| 232 | protein disulfide oxidoreductase, disulfide oxidoreductase activity, protein-disulfide reductase activity | glycerol ether metabolic process, cell redox homeostasis, metabolic process |
| 236 | structural constituent of ribsosme. | translation |
| 256 | DNA binding, nucleic acid binding, ATP binding, zinc ion binding | transcrition, DNA-templated |
| 258 | antioxidant activity, oxidoreductase activity, catalytic activity | cell redox homeostasis - any process that maintains the redox environment of a cell or compartment within a cell |
| 261 | tRNA binding, RNA binding, structural constituent of ribosome, rRNA binding | translation |
| 263 | 2-C-methyl-D-erythritol 2,4-cyclodipho, lyase activity | terpenoid biosynthetic process, |
| 278 | acetolactate synthase activity, transferase activity, transferring, amino acid binding | branched-chain amino acid biosynthetic process |
| 283 | N-acetyltransferase activity, transferase activty | NA |
| 296 | structural constituent of risbosome | translation |
| 312 | DNA binding, DNA-binding transcription factor activity, sigma factor activity | DNA-templated trascription, initiation |
| 316 | hydro-lyase activity, carobon oxygen lyase activity, lyase activity | histidine biosynthetic process |
| 354 | cation transmembrane transporter | potassium ion transport |
| 358 | deoxyribose-phosphate aldoase, aldehyde-lyase activity, carbon-carbon lyase activity | deoxyribonucleotide catabolic |
| 390 | NA | NA |
| 412 | peptidase activity, catalytic activity, acting on a protein | defense response to bacterium |
| 448 | oxidoreducatase activity | methionine metabolic process, aspartate family amino acid metabolic |
| 464 | zinc ion binding, metal ion binding | NA |
| 467 | NA | NA |
| 469 | carbamate kinase activity, kinase activity | arginine metabolic process, glutamine family amino acid metabolic |
| 509 | NA | NA |
| 522 | phosphate ion binding, anion binding, ion binding | phosphate ion transmembrane transport, transport |
| 536 | NA | NA |
| 539 | catalytic activity, iron-sulfur cluster binding, cofactor binding, metal cluster binding | NA |

Figure 4.13: Molecular function and biological process annotations for those proteins in the frequency band 1 group.



| Pair | Mesophile: Protein Name, Genus, Species | Gene ID |
|---|---|---|
| 23 | MULTISPECIES: carbamoyl-phosphate synthase small subunit [Proteobacteria] | UT36_C0004G0033 |
| 54 | alanine transaminase [Escherichia coli] | L912_0634 |
| 96 | MULTISPECIES: tRNA (N(6)-L-threonylcarbamoyladenosine(37)-C(2))-methylthiotransferase MtaB [Bacillus] | BN2127_JRS8_00191 |
| 114 | MULTISPECIES: cysteine--tRNA ligase [Bacillus] | BS34A_01290 |
| 121 | aspartate ammonia-lyase [Bacillus halodurans] | BAB05145 |
| 156 | CTP synthase [Bacillus halodurans] | B4065_1258 |
| 157 | D-3-phosphoglycerate dehydrogenase [Bacillus halodurans C-125] | BBEV_2287 |
| 170 | MULTISPECIES: 50S ribosomal protein L33 [Bacillus] | VL17_09845 |
| 177 | MULTISPECIES: translation initiation factor IF-1 [Bacillaceae] | BCE_G9241_0120 |
| 197 | MULTISPECIES: 50S ribosomal protein L21 [Bacillus] | BCE33L4189 |
| 225 | MULTISPECIES: 30S ribosomal protein S8 [Bacillus] | BAMA_11430 |
| 227 | YjbQ family protein [Bacillus halodurans] | BAMTA208_11780 |
| 253 | hydroxymyristoyl-ACP dehydratase [Escherichia coli O104:H21 str. CFSAN002237] | HQ24_00870 |
| 263 | 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase [Bacillus halodurans] | BAME_21790 |
| 265 | MULTISPECIES: tRNA (adenosine(37)-N6)-threonylcarbamoyltransferase complex ATPase subunit type 1 TsaE [Bacillus] | ABE28_001525 |
| 277 | MULTISPECIES: thiol peroxidase [Bacillus] | BS34A_32140 |
| 323 | RecName: Full=N-(5'-phosphoribosyl)anthranilate isomerase; Short=PRAI | NA |
| 340 | MULTISPECIES: riboflavin synthase [Streptococcus] | SpnNT_00181 |
| 398 | TatD family deoxyribonuclease [Streptococcus pneumoniae] | ERS021436_00661 |
| 425 | 3-methyl-2-oxobutanoate hydroxymethyltransferase [Bacillus halodurans] | BAB05406 |
| 441 | formyltetrahydrofolate deformylase [Bacillus halodurans C-125] | BAB06984 |
| 451 | pyridoxal 5'-phosphate synthase lyase subunit PdxS [Bacillus halodurans] | bwei_5817 |
| 453 | MULTISPECIES: acetamidase [Bacillus] | UB32_12180 |
| 491 | electron transfer flavoprotein subunit alpha/FixB family protein [Bacillus halodurans] | NA |
| 494 | arabinose-5-phosphate isomerase KdsD [Escherichia coli] | ECTW07945_4346 |
| 517 | phenylalanine--tRNA ligase subunit alpha [Bacillus halodurans] | BS34A_31240 |

Figure 4.14: Mesophile Proteins from pairs with similar power at frequency band 10. Included are the pair number, protein name, species and genus information, and gene ID.



| Pair | Molecular Function | Biological Process |
|---|---|---|
| 23 | Amino Acid Biosynthesis, ATP Binding | Pyrimidine nucleobase biosynthetic process |
| 54 | pyridoxal phosphate binding; {anion binding, coenzyme binding, vitamin b6 binding}, binding | biosynthetic process, metabolic process |
| 96 | 4 iron, 4 sulfur cluster binding, transferase activity | RNA modification |
| 114 | nucleotide binding, small molecule binding, nucleoside binding, aminoacyl-tRNA ligase activity | tRNA aminoacylation for protein, translation, amino acid activation, tRNA metabolic process |
| 121 | Catalysis of the reaction: L-aspartate = fumarate + NH3 | tricarboxcylic acid cycle, aerobic respiration, aspartate metabolic process |
| 156 | ATP binding, CTP synthase activity, metal ion binding | glutamine metabolic process |
| 157 | oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor, NAD binding | metabolic process, oxidation reduction process |
| 170 | structural constituent of ribosome | translation |
| 177 | NA | NA |
| 197 | structural constituent of ribosome; rRNA binding | translation |
| 225 | structural constituent of ribosome; rRNA binding | translation |
| 227 | cation transporter activity; solute: proton anitporter activity | cation transport, potassium ion transport, transmembrane transport |
| 253 | hydro-lyase activity | fatty acid biosynthetic process |
| 263 | lyase activity | terpenoid biosynthetic process |
| 265 | transferase activity | tRNA threonylcarbamoyladenosine modification |
| 277 | Catalysis of the reaction: thioredoxin + hydrogen peroxide = thioredoxin disulfide + H2O: catalytic process | oxidation reduction process |
| 323 | NA | NA |
| 340 | oxidoreductase activity, riboflavin synthase activity, transferase activity - catalytic activity | NA |
| 398 | hydrolase activity, acting on ester bonds, catalytic activity | NA |
| 425 | metal ion binding, catalytic activity, transferase activity | pantothenate biosynthetic process, metabolic process |
| 441 | amino acid binding, hydrolase activity | NA |
| 451 | amine-lyase activity, carbon nitrogen lyase activity, lyase activity, catalytic activity | biosynthetic process |
| 453 | hydrolase activity, hydrolase activity acting on | "de novo" IMP biosynthetic process, IMP biosynthetic process, metabolic process |
| 491 | NA | NA |
| 494 | isomerase activity, catalytic activity, carbohydrate binding, binding | carbohydrate metabolic process |
| 517 | tRNA binding, nucleotide binding, APT binding, aminoacyl-tRNA ligase activity, phenylalanin-tRNA ligase activity | translation, tRNA aminoacylation, biosynthetic process |

Figure 4.15: Molecular function and biological process annotations for those proteins in the frequency band 10 group.

# CHAPTER 5

# Discussion

## 5.1 Life In Extreme Conditions

A thermophile is a prokaryotic organism occupying an ecological niche with temperatures ranging from 41°C to 122°C. A mesophile is a prokaryote living in environments with temperatures ranging from 20°C to 45°C. The proteins expressed by a mesophile organism would denature in the environment of the thermophile. Thus thermophiles have adapted to life in extreme temperature. This study looked to identify features related to thermal stability and function in the amino acid chains of short globular proteins from mesophile and thermophile species, within the constraint that the proteins fold to perform a specific function.

To do so, 540 homologous pairs of proteins were generated. A homologous pair consists of one protein from the mesophile and another from the thermophile that have similar function. The amino acid chains are then converted to hydrophobicity signals by assigning a hydropathy score from the Kyte-Doolittle hydropathy scale to each of the amino acids in the chain. The hydrophobic effect is temperature dependent and is the driving force behind protein folding. It is thought that the hydrophobic effect essentially limits the number of conformation accessible to the amino acid chain of a short globular protein. This allows for the sum of all other forces acting on the molecule, such as the electrostatic force, to dictate the native conformation. The linear amino acid chain of a short globular protein contains all the information necessary for the polypeptide to fold in the proper conditions. This statement is referred to as the Thermodynamic Hypothesis.

This study looked to identify features related to the temperature of the environment in which the protein is expressed. The study consisted of a comparative analysis of the amino acid chains of short globular proteins in mesophile and thermophile bacteria and Archea with similar function. The experiments performed were aimed at generating a procedure allowing for the identification of certain features related to thermal stability and function. These features are classified as *pair specific* and *distinguishing* or *similar*. These are features that are found within a pair. The next classification is *function specific* this feature is generated from pair specific similar features through computational methods, and is confirmed through the use of a gene ontology. The next measure is *thermal specific*, this is a measure found to be similar throughout one of the two temperature classifications.

## 5.2 Methods

Several methods were used to determine the relationship between hydropathy, function, and thermotolerance. The first of which was to determine how similar the amino acid chains of each pair are with one another. The next step was to determine the amino acid content within each protein. These were followed by statistical hypothesis tests to determine if they hydropathy values in the signals were normally distributed, and if members of a pair had the same underlying distribution of hydropathy values in their signals. This was followed by a scatter plot of the mean hydropathies of the mesophile proteins versus the mean hydropathy of the thermophiles.

Both of the hydrophobicity signals composing a homologous pair were passed through a wavelet packet transformation. This was done to determine the fluctuation in hydropathy along different regions of the hydrophobicity signals at various frequency bands. This allowed for the use of spectral analysis methods to determine the average variance in the fluctuation in hydropathy at certain frequency bands along the length of the entire hydrophobicty signal, referred to in this study as power.

Each of the frequency bands resulting from the WPT contained a number of coefficients containing localized information about the fluctuation in hydropathy within certain regions of the protein. The skew of the frequency bands was taken to see if the distribution of coefficients within that band was asymmetric. The presence of an asymmetry within the bands would indicate that the fluctuation in hydropathy along a unit length throughout the protein varies depending on location. A skewed distribution would also indicate that the power of that frequency band was being driven by outliers. The kurtosis of each frequency band was calculated to determine the degree to which the fluctuation in hydropathy varies throughout the length of the signal.

A control data set was generated by shuffling the order of hydropathy values within the signals 299 times. Each iteration of this bootstrapping procedure was passed through the wavelet packet transformation, and spectral analysis methods were applied to the resulting frequency bands. This data set was used to generate confidence intervals for the power, skew, and kurtosis of each leaf.

The total power was calculated for each hydrophobicity signal, and the difference was taken between members of a homologous pair. This was followed by collecting the powers from every protein for each frequency band, and ranking them in order from least to greatest and generating a scatter plot of mesophile power versus thermophile power. A linear regression was performed to determine the relationship between power and temperature regime.

The non-normalized power of each frequency band was compared between members of a homologous pair to determine if they were similar. The same procedure was applied to the bootstrapped control data to determine a baseline likelihood of finding similar power at any frequency band across the data set. Homologous pairs with similar power at the same frequency band were then grouped together. A second control data set was generated to prune out those pairs that matched due to random chance.

## 5.3 Discussion of Results

The results of the amino acid chain alignment show that majority of protein pairs have dissimilar orderings of amino acids. This indicates that evolutionary processes have produced proteins with similar function using different orderings of amino acids. It appears that the information necessary to satisfy the thermodynamic hypothesis can be encoded into proteins using different amino acid combinations.

Four amino acids have been referenced in the literature as making up a larger percentage of the amino acid chains of thermophiles than mesophiles. These amino acids are aspartic acid, glutamic acid, lysine and arginine. This study calculated the percent make up of each amino acid in every protein, and then calculated the average percent make up for both the mesophile and thermophile groups. Comparing the average percentage of the thermophile to the mesophile it was found that aspartic acid made up 0.67% less of the average thermophile protein than the average mesophile protein. Glutamic acid makes up 1.81% more of the average thermophile protein than mesophile, lysine makes us 3.11% more of the average thermophile, and arginine makes up 0.08% more of the average thermophile.

There are four amino acids that are thought to make up less of a thermophile protein in comparison to mesophile. These amino acids are asparagine, glutamine, serine, and threonine. Asparagine was found to make up 0.01% less of the average thermophile than the average mesophile. The percentage difference for Glutamine is 1.71% less, serine 0.55% less and threonine 1.29% less. It appears that of those amino acids thought to make up a larger percentage of thermophile proteins, only glutamic acid and lysine had percentage differences between the average thermophile and mesophile greater than 1.00%. For those amino acids thought to make up a smaller portion of the average thermophile, glutamine and threonine made up more than 1.00% more of the thermophile than the mesophile.

The results of this test show that the variation in amino acid content of the average thermophile and mesophile protein are greatest in lysine, glutamic acid, glutamine, and threonine. While aspartic acid, arginine, asparagine, and serine showed differences that were less than 1.00%. In addition to this, alanine was found to make up 2.39% more of the average mesophile than the average thermophile. Because of the similarity in the percent make up of each amino acid in the chain of the average mesophile and thermophile protein, amino acid content cannot be used as a thermal specific identifying feature.

A comparison of the lengths of the proteins did not reveal a tendency for either the mesophile or thermophile protein to be longer than its partner. It does not appear that thermal stability is achieved through the addition of amino acids to the polypeptide. Adaptations made to thermal stability must occur within the constraint that the protein fold to a conformation specific to its function. The addition of amino acids to extend the length of the protein is not a strategy employed in this data set. There is no tendency for one member of a homologous pair to be longer than its partner, and there is little difference in the amino acid content between the average mesophile and the average thermophile protein. This means that the ordering of the amino acids within the polypeptide encodes information critical to the function and thermal tolerance of that protein.

The results of the Anderson Darling test showed that none of the proteins have normally distributed hydropathy values within their lengths. In addition to this, 530 of the pairs have the same underlying distribution of hydropathy values within their lengths. The distribution of hydropathy values in the hydrophobicity signals cannot be used to distinguish the mesophile protein from the thermophile. Neither can the prevalence of a single hydropathy value be used to distinguish the two.

The scatter plot of mean hydropathy of the mesophile versus thermophile, Figure 4.1, showed that the average value could not be used as a means of distinguishing the proteins. Each data point appeared to be clustered about a single value, indicating that the average hydropathy is not an indicator of function either. The values appear to be distributed about $(-.2, -.2)$. There appears to be a linear relationship between the pairs, an increase in the average hydropathy of a mesophile is matched by an increase in average hydropathy in its thermophile partner.

The total power of each hydrophobicity signal was calculated and the difference taken between members of a pair. It was found that 90.56% of thermophile proteins contain more power in their hydrophobicity signals than their mesophile partner. The spread of the difference in total power indicated that the total power between the members of a homologous pair is quite similar. This shows that the total power can only be used as a pair specific distinguishing feature. A thermophile protein has a larger total power when compared to a similarly functioning mesophile protein. This indicates that variation in hydropathy along the length of a protein is related to thermal stability, but is adapted within the constraint that the protein fold to perform a specific function. Thus increasing the fluctuation in hydropathy along the length of a protein will only raise the denaturing temperature of a protein if it is encoded in a way to reflect the function of that protein.

At any frequency band, a random ordering of the amino acids making up a protein will produce a near zero power and skew. The true ordering of amino acid produces a value of power that is significantly non zero. Thus the ordering of amino acids throughout the length of the protein is significant to the structure of that protein. The amino acids are ordered in a way that produces a non-random fluctuation in hydropathy. In addition to this, the ordering within the amino acid chain is arranged so that various regions of the protein differ in regards to the fluctuation in hydropathy. The fluctuation in hydropathy at any frequency band and along the length of the protein varies in a way that is related to the function of that protein.

For instance Figure 4.6 contains the power of each frequency band for the proteins in homologous pair 10. A control data set was run to determine the value of power measured from 200 random permutations of the amino acids for both proteins at each frequency band. The results of the control data were used to generate confidence intervals showing that a random ordering of amino acids will produce a near zero value of power. The actual powers calculated from the experimental data are significantly non-zero, indicating that the fluctuation in the hydropathy of the amino acid chain is not due to random chance.

The kurtosis of each frequency band was calculated to compare the magnitude of the fluctuation in hydropathy along the length of the signal. The results of this test show that frequency components fluctuate in value along the length of the signal, and that the degree to which they vary is dependent on the location within the amino acid chain, and the number of amino acids taken into account. The ordering of amino acids is organized in a non-random way, and different regions of the protein contain different fluctuations in hydropathy.

The power at every frequency band was collected from all the mesophile and thermophile hydrophobicity signals and ranked from smallest to largest. These values were then used to generate scatter plots of the power in the mesophile frequency band versus the power in the thermophile frequency band for the first 64 frequency bands. A linear regression was performed on each of these plots and the slopes collected. It appears that the thermophiles contain more power at frequency bands between 21 and 31 and 44 through 56. While the mesophiles appear to contain more power at frequency bands between 18 and 22. It appears that the trend for the thermophile to contain more power than its mesophile partner is due to increased power at frequency bands between $[21 - 31]$ and $[44 - 56]$. It may be that increasing the fluctuation in hydropathy at these unit lengths will result in increased thermal tolerance. By analyzing the distribution of power at these frequency bands it may be possible to develop a numeric value that can be used as a thermal specific indicator.

The difference in power at each frequency band was taken between members of the homologous pairs. The differences at each frequency band were collected, and the number of positive values counted. The ratio of positive values to the number of pairs that contained that frequency band was then computed. This method looked at the similarity in power at each frequency between a pair across the entire data set incorporating the function of the proteins into the measurement. The majority of ratios are close to but slightly greater than 0.5. This indicates that a larger power at certain frequency bands may be a means of increasing denaturing temperature, but that the specific frequency band is dependent on the function of the protein.

The hydrophobicity scale used in this study is strongly correlated with the Gibbs free energy of moving an amino acid from a liquid water to vapor state. Because of this the total power of a protein is also strongly correlated with Gibss free energy. This indicates that thermophile proteins have a greater Gibbs free energy in comparison to a similarly functioning mesophile protein.

It has been noted in the literature that thermophile proteins implement strategies to increase their Gibbs free energy to increase their denaturing temperature. The authors Abbas Razvi and J. Martin Scholtz make note of this in the review article *Lessons in stability from thermophilic proteins*. In this paper Razvi and Scholtz note a variety of mechanisms to increase the denaturing temperature of a protein that involve increases to free energy [32]. The total power is an informatics method that can distinguish members of a pair, and its physical significance is related to the Gibbs free energy of the protein. Thermophile proteins require a larger amount of energy to unfold in comparison to similarly functioning mesophile proteins.

A comparison of the non-normalized power between the frequency bands of a homologous pair showed that the experimental data is more than twice as likely to generate similar power than a random bootstrapped sequence. This may indicate that the power of a certain frequency band is related to the function of a protein and can be used as a function specific measure. To evaluate the use of this method in identifying similarly functioning proteins a gene ontology was carried out on those pairs, after pruning, that had similar power at frequency bands 1 and 10. It was found that there was a large degree of overlap between these two groups in the annotations for both molecular function and biological process.

It may be that the 540 protein pairs used in this study only fall within a small number of protein families. To improve the procedure, the pruning method must be refined, for instance by analyzing the distribution of powers at that frequency band where there is a match generated by the bootstrapped control data to determine a less arbitrary threshold value. The majority of pairs in the data set had similar non-normalized power at multiple frequencies. So after pruning, it may be better to group homologous pairs that have similar power at multiple frequency bands, rather than just one. In addition to this, only the first few annotations were collected for the proteins within the grouping. GO presents a larger framework within which the molecular function and biological process of a protein can be understood. Expanding the gene ontology to take advantage of this additional information will help to contextualize the role of each of the proteins in every grouping, and will be the natural complement to the computational methods used in this study.

## 5.4 Summary

This study looked to determine how the hydrophobic effect is encoded into the linear amino acid chains of proteins from mesophile and thermophile prokaryotes to reflect thermal stability and function. Five hundred and forty homologous pairs were generated containing a protein from each group with similar function. It was found that different orderings of amino acids can be used to generate proteins with similar function. A pair specific distinguishing feature was identified in the difference in total power between members of a pair. A thermophile protein contains more power than a mesophile protein with similar function. The difference in power is small and appears to be due mainly to certain frequency bands contained within the chains. The amino acid chains of both proteins code for specific fluctuations in hydropathy along the lengths of the protein.

The hydrophobicity scale used in this study is strongly correlated with Gibbs free energy. It appears that the method presented in this thesis can be used to determine the way Gibbs free energy is encoded into the amino acid chain of a protein to reflect its function and denaturing temperature.

It appears that the fluctuation in hydropathy is a characteristic related to thermal stability and function. It may be that the difference in total power changes the number of conformations accessible to the linear amino acid chain, and that increasing the fluctuation in hydropathy within the length of the protein can be used to increase thermal tolerance. Because the fluctuaion in hydropathy is different, and it is clear that the ordering of amino acids codes for a specific variance in the fluctuation in hydropathy, the amino acid chains between each pair are dissimilar. This may mean that the information most necessary to the folding of a protein is the fluctuation in hydropathy along various unit lengths throughout the amino acid chain. It may be possible to raise the denaturing temperature of a protein by increasing the fluctuation in its hydropathy along certain unit lengths throughout its amino acid chain, and maintain its function.

# REFERENCES

[1] G. L. N. Stephen M.J. Pollo, Olga Zhaxybayeva. Insights into thermoadaptations and the evolution of mesophily from the bacterial phylum *Thermotogae*. *Canadian Journal of Microbiology*, vol. 61, pp. 655-670, 1951.

[2] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, vol. 05, 2005.

[3] K. Dill. Dominant forces in protein folding. *Biochemistry: Perspectives in Biochemistry*, vol. 29, no. 31, 1990.

[4] C.B. Anfinsen. Studies on the principles that govern the folding of protein chains. *Nobel Lecture*, 1972.

[5] A. L. Lehninger, D. L. Nelson, M.M. Cox. Principles of Biochemsitry. *New York: Worth Publishers*, 2000.

[6] Mansilla, M.C., Cybulski, L.E., Albanesi, D., Mendoza, D.D.. Control of membrane lipid fluidity by molecular thermosensors. *Journal of Bacteriology*, vol. 42, pp. 6681-6688, 2004.

[7] Zhang, Y., and Rock, C.O.. Membrane lipid homeostasis in bacteria. *Nature Reviews Microbiology*, vol. 6, pp. 222-233, 2008.

[8] DiRuggiero, J., Santangelo, N., Nackerdien, Z., Ravel, J., and Robb, F.T.. Repair of extensive ionizing-radiation DNA damage at 95 degrees C in the hyperthermophilic archaeon *Pyrococcus furiosus*. *Journal of Bacteriology*, vol. 179, pp. 4643-4645, 1997.

[9] Whitaker, R.J., Grogan, D.W., Taylor, J.W.. Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Molecular Biology and Evolution*, vol. 22, pp. 2354-2361, 2005.

[10] Mino, S., Makita, H., Toki, T., Miyazaki, J., Kato, S., Watanabe, H., et. al. Biogeography of *Persephonella* in deep-sea hydrothermal vents in the western pacific. *Frontiers in Microbiology*, vol. 4, pp. 1-12, 2013.

[11] Atomi, H., Matsumi, R., Imanaka, I.. Reverse gyrase is not a prerequisite for hyperthermophilic life. *Journal of Bacteriology*, vol. 186, pp. 4829-4833, 2004.

[12] Zellner, G., Kneifel, H.. Caldopentamine and caldohexamine in cells of *Thermotoga* species, a possible adaptation to the growth at high temperatures. *Archives of Microbiology*, vol. 159, pp. 472-476, 1993.

[13] Endo, A., Sasaki, M., Maruyama, A., and Kurusu, Y.. Tempearture adaptation of *Bacillus subtilis* by chromosomal *groEL* replacement. *Bioscience, Biotechnology, and Biochemistry*, vol. 70, pp. 2357-2362, 2006.

[14] Ezemaduka, A.N., Yu, J., Shi, X., Zhang, K., Yin, C., Fu, X., et al. A small heat shock protein enables *Escherichia coli* to grow at a lethal temperature of 50°C conceivably by maintaining cell envelope integrity. *Journal of Bacteriology*, vol. 196, pp. 2004-2011, 2014.

[15] R. F. D. Jack Kyte. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, vol. 157, pp 105-132, 1982.

[16] Chothia, C. *Journal of Molecular Biology*, vol. 105, pp.1-14, 1976.

[17] Simm, S., Einloft, J., Mirus, O., Schleiff, E.. 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological Research*, 2016.

[18] F. J. M. Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, vol. 46:253, pp.68-78, 1951.

[19] K. Bury. Statistical Distributions in Engineering. *Cambridge University Press*, 1999.

[20] L. Koopmans. The spectral Analysis of Time Series. Academic Press, 1974.

[21] B. Vidakovic. Statistical modeling by wavelets. John Wiley and Sons, Inc., 1995.

[22] T. W. Anderson D. A. Darling. A Test of Goodness of Fit. *Journal of the American Statistical Association*, 49:268, 765-769, 1954.

[23] SPC for Excel. Anderson-Darling Test for Normality. June 2011. April 2019. https://www.spcforexcel.com/knowledge/basic-statistics/anderson-darling-test-for-normality

[24] Charles Zaiontz. Real Statistics Using Excel. Two Sample Kolmogorov-Smirnov Test. April 2019. http://www.real-statistics.com/non-parametric-

tests/goodness-of-fit-tests/two-sample-kolmogorov-smirnov-test/

[25] MathWorks, Inc. Documentation, fft. April 2019. https://www.mathworks.com/help/matlab/ref/fft.html

[26] C. Valens. A Really Friendly Guide to Wavelets. 1999 c.valens@mindless.com

[27] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, date.

[28] Berg, J.M., Tymoczko, J.L., Stryer, L.. Biochemistry. 5th Edition, section 4.2. W.H. Freeman, New York, 2002.

[29] Thomas Madden. The NCBI Handbook, second edition (internet). The BLAST Sequence Analysis Tool. https://www.ncbi.nlm.nih.gov/books/NBK153387/

[30] Gentleman R (2019). annotate: Annotation for microarrays. R package version 1.62.0.

[31] Lewis S.E. (2017) The Vision and Challenges of the Gene Ontology. In: Dessimoz C., Škunca N. (eds) The Gene Ontology Handbook. Methods in Molecular Biology, vol 1446. Humana Press, New York, NY

[32] Razvi, A., Scholtz, J.M. Lessons in stability from thermophilic proteins. *Protein Science*, 15(7):1569-78, 2006.