PROTEIN EVOLUTION IN MICROBIAL EXTREMOPHILES

Protein Evolution in Microbial Extremophiles

By

Nicholas Waglechner, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2007) McMaster University

(Biochemistry and Biomedical Sciences) Hamilton, Ontario

TITLE: Protein Evolution in Microbial Extremophiles

AUTHOR: Nicholas Waglechner, B.Sc.

SUPERVISOR: Dr. Paul G. Higgs

NUMBER OF PAGES: x, 84

Abstract

Two separate but related projects make up the work of this thesis. The growing amount of sequence data available in public databases provides an opportunity to compare species in new ways. It can be shown that there is a systematic change in amino acid composition in a dataset of sequences from 69 species possessing a range of optimal growth temperatures. By creating a phylogenetic tree of all available Archaea, pairs may be selected that contain a relatively closely related mesophile and (hyper)thermophile. In addition, pairs may be selected from Bacteria to include psychrophiles as well as other thermophiles. An evolutionary model is derived here that detects amino acid asymmetries in these species pairs beyond what might be expected to be caused by differences in GC content. This amino acid asymmetry can then be plausibly explained by temperature adaptation occurring in these species since they diverged from a common ancestor.

In the second part, similarity searches using molecular sequences are drawn as networks, where open reading frames in one species may be linked to a corresponding sequence in another species if the similarity search score is above a given threshold. This process is similar to that used to identify orthologous sequences for use in evolutionary models. When drawn as a network of distinct clusters of similarity, patterns emerge that can be spurious or have some biological relevance. This work identifies the need to develop better methods of analyzing these network clusters.

Acknowledgements

I would like to thank my supervisor Dr. Paul Higgs for his supportive advice, invaluable direction and help. I would also like to thank Dr. Brian Golding, Dr. Radhey Gupta and Dr. Daniel Yang for their comments and suggestions regarding my work. I would like to thank Dr. Radhey Gupta and Beile Gao for providing archaeal sequences.

Thanks to Meng Wu and William Coletto for their help with software, and thanks to Wenqi Ran and Wenli Jia for sharing their experience.

I would also like to thank my wife for her patience and my family and friends for their support.

dedicated to Susan

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Genomes of Prokaryotes

The work described in this thesis explores the evolution of extremophile prokaryotes through the development of a phylogenetic model that makes use of protein sequences from candidate organisms. The availability of these sequences is due to the completion of various genome sequencing projects and the techniques used to infer data from the completed genomes. An example of the growth in the number of completed genomes is provided along with an example of a method used to derive protein sequence information from them. The motivation behind comparative genomics is discussed briefly and the Archaea, which include many examples of extremophile species, are introduced. Finally, a description of extremophily and its practical applications is provided as well as a brief outline of the specific problem being addressed and how it is approached in this work.

The genome of an organism consists of the entire complement of genetic material in that organism. This includes both coding and non-coding DNA on chromosomes and extra-chromosomal elements. Prokaryotic chromosomes are double-stranded and typically circular, though linear chromosomes are known for

some species (Madigan et al 2003, pg 177). They feature defined origins of replication, where the process of DNA replication begins during cell division, and are considered gene-dense compared to eukaryotic organisms since most of the sequence is comprised of coding regions separated by typically <10% non-coding DNA consisting of binding sites, spacers, remnants of past functional DNA and regulatory elements (Madigan et al. 2003, pg 178-9). Chromosome sizes are defined in terms of base-pairs (bp), referring to the number of paired DNA bases in the length of the chromosome. According to the current list of completed prokaryotic genomes available from the NCBI microbial genome database, the average length of bacterial and archaeal genomes is 3.58 million bp (MB) with the smallest being 0.16 MB (*Candidatus Carsonella ruddii* PV) and the largest currently being 13.03 MB (*Sorangium cellulosum* 'So ce 56') (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi Accessed July 2008).

The major technique used to produce this sequence information is the chain termination method developed by Sanger et al. (1977). In this method, a mixture of 2'-deoxy- and 2`,3`-dideoxyribonucleotides are used in a DNA synthesis reaction such that synthesis is terminated when a nucleotide lacking a 3' hydroxyl end is incorporated at a small fraction of sites. Over the entire course of the reaction random termination of the DNA synthesis reaction produces DNA fragments that differ in length by one base pair. By separating the reactions for the four bases and using a small amount of the respective labeled 3'-deoxyribonucleotide, the DNA sequence can be 'read' by observing the order of

DNA fragments proceeding from smallest to largest on an electrophoretogram after the DNA fragments are separated by electrophoresis. Refinements to this technique include the development of fluorescently labeled dideoxyribonucleotides as chain terminators to ease reading and eliminate the need to use isotopes for labelling, and thermostable polymerases and the polymerase chain reaction (Mullis et al. 1986; Prober et al. 1987; Saiki et al. 1988). Enhanced length of read was achieved through the use of capillary electrophoresis, allowing the production of longer sequences per reaction, to minimize the number of reactions required for the same length of read (Madabhushi et al. 1998).

The Whole-genome shotgun method was used to produce the first complete genome of a free-living organism, *Haemophilus influenza* Rd (Fleischmann et al. 1995). In this method a large number of mechanically sheared DNA fragments between 1.6 and 2.0 kb were created from *Haemophilus* genomic DNA, the size chosen to minimize the number of complete genes present in each fragment. These fragments were then digested with the Sma I restriction enzyme and inserted into suitable cloning vectors to make a library of single-insert containing plasmids to be propagated in *E. coli* cells deficient in recombination and restriction to preserve the randomness of the fragments. This is essential because the theory of shotgun sequencing follows a Poisson distribution of randomly generated bases of a given coverage factor to determine the percentage of the bases unsequenced. The authors' goal of 6x coverage of the 1.83 MB genome

follows from the calculation that 5x coverage would theoretically result in 0.67% of bases left unsequenced. 16 240 forward sequencing and 7 744 reverse sequencing reactions were successfully performed to generate 11 631 485 bases of data. These data were eventually assembled into 140 contiguous assembly fragments, or contigs, using 30 hours of computation by identifying overlapping sequences in the reads produced by individual sequencing reactions. These contigs were then ordered by designing primers for the ends of each contig to be used in four strategies to identify adjacent contigs and close gaps. One strategy used Southern analysis using labeled primer oligonucleotides hybridizing to common restriction fragments followed by targeted PCR reactions. 6-frame translations of contig sequences were used to query a peptide database to identify contig ends matching a given peptide, and therefore provisionally adjacent contigs. The construction of two λ phage libraries of *Haemophilus* genomic DNA containing inserts of 15 to 20 kb in size were probed with contig oligonucleotides and sequenced to determine adjacent contigs in the λ insert. Finally, PCR reactions and sequencing were used to confirm each gap identified between adjacent contigs by the other methods, and to completely close the other gaps. The use of the different strategies are necessary because contigs containing regions of repeat sequences, like the rRNA operons, are difficult to assemble due to the relatively small size of individual reads. Complete genome sequences obtained by techniques similar to this are a major source of new data and the number of complete genomes being made available is increasing every year.

1.2 Data Explosion

The sequences derived from the various sequencing projects provide the massive amount of data deposited into the public databases over the last 15 years. Currently, as of July 2008, the National Center for Biotechnology Information lists 728 complete microbial genomes, 676 of which are bacterial and 52 archaeal (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Specifically, the growth in the number of complete microbial genomes has been exponential, as can be seen in Figure 1.1.

Landmark microbial genomes include the very first, *Haemophilus influenzae* (Fleischmann et al 1995), the ubiquitous model organism *Escherichia coli* (Blattner et al. 1997), the human pathogen *Mycobacterium tuberculosis* (Cole et al. 1998), and the first Archaea *Methanocaldococcus jannashcii* (Bult et al. 1996).

Though the generation of raw DNA sequences provides the basic sequence data and general information like GC/AT content and size, more information about the species can be determined through the identification of open reading frames (ORFs) to determine the presence of individual genes in each genome.

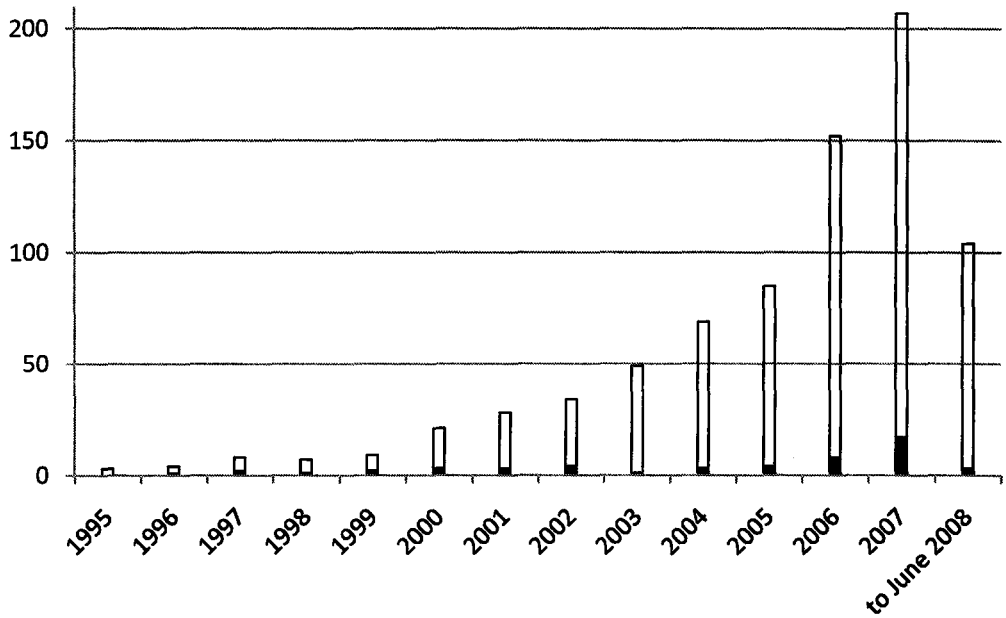

Figure 1.1 - Number of completed microbial genomes per year. Open bars - Bacteria, Solid bars – Archaea.

One of the most common software packages for identifying these ORFs is Glimmer (Salzberg et al. 1997). Glimmer uses interpolated Markov models (IMMs), which allow switching between different order Markov models depending on weighting and evidence criteria rather than relying on a fixed-order Markov model like those used in other software such as GeneMark (Borodovsky et al. 1995). This depends on the assumption that in-frame coding sequences contain patterns detectably different from those in out-of-frame or non-coding sequences. Where a fixed-order model, for example 5th order, will predict the base at a given position given the identity of bases observed in the preceding 5 bases

for every site in the training set, the IMM models used in Glimmer are weighted differently based on the available frequency of observed *k*-mers (sequence fragments of length *k*) (Salzberg et al. 1997). Because the chance of observing every possible combination of *k* bases decreases exponentially as *k* increases, the authors note there are cases where an insufficient number of *k*-mers are available in the training data to accurately predict the base following a given *k*-mer. In such cases the order of the model can be raised or lowered depending on the available evidence in the training set in order to use the order in each case with the maximum available evidence. Models for each of the 6 frames as well as a model for non-coding DNA are trained and used to score regions of the genome by calculating the linear combination of probabilities for all *k*th order IMMs (Salzberg et al. 1997). An ORF is identified as a region with a score above a given threshold for one of the 6 frames, and a corresponding low score for non-coding DNA model. Where overlapping ORFs are predicted, the longest ORF with the highest score is judged to be correct, and cases with unclear results are left for manual examination. (Salzberg et al. 1997).

Translated conceptual ORFs provide the protein sequence data necessary for detailed analysis and comparison between species. The genes encoded by these ORFs may be grouped into clusters, like the Clusters of Orthologous Groups (COGs) based on what function they provide the species, such as housekeeping functions like DNA replication and repair, gene expression proteins, signaling, the components of metabolic pathways, structural elements, motility proteins,

membrane functions, sporulation, virulence and antibiotic resistance factors (Tatusov et al. 1997, Tatusov et al. 2003). Some of these genes will be necessary for the organism to live in its particular environment, and some will be particular to a related group of organisms.

1.3 Comparative Genomics

As the availability of completed genomes increases, the ability to produce comparisons between species to answer biological questions likewise increases. Prior to the era of complete genome sequencing, growth of functional knowledge about genes was achieved as new sequences and individual experimental data were made available through the public databases. Newly obtained sequences are checked for similarity against those already deposited for clues regarding structure, function and evolution, as in the case of *Methanocaldococcus jannaschii* and nearly every newly obtained genome sequence (Bult et al. 1996). A library of tools and methods were developed, among them the various BLAST algorithms, which allow searching through the databases quickly and efficiently based on measures of similarity (Altschul et al. 1990). Various versions of the BLAST algorithm have been implemented for applications requiring raw DNA, translated DNA or protein queries against DNA, translated DNA or protein databases and have incorporated the creation of position-specific-scoring matrices for iterated BLAST searches to identify sequences matching specific patterns

(Altschul et al. 1997). The ability to identify sequences similar to one another within and between species is a crucial first step towards more complex analyses.

In addition to the basic tools used to interact with different databases, the development of various other analytical techniques was initiated to make use of the inflowing data. Comparisons of the composition of DNA and proteins between different species were performed to quantify the differences between species and used to determine species signatures (for example, see Di Giulio 2000). Evolutionary models, introduced in chapter 2 and developed in chapter 3 of this thesis, are used to examine the effects of selection in the time since two species diverged from one another. The goal of other models is to describe the general nature and mechanisms of molecular evolution. One subject of chapter 2 is the construction of phylogenies to better clarify the history and evolutionary relationships between species, and the reasons these techniques sometimes do not produce clear answers.

1.4 Archaea

The Archaea were proposed as a natural domain of life separate from Bacteria by Woese et al. (1978, 1990). Physiological differences between Archaea and Bacteria include the presence of ether-linked membrane lipids and lack of typical bacterial peptidoglycan cell walls (Woese et al. 1978). Differences were also identified in the 5S rRNA structure and tRNA sequences such that Woese et al. note specific portions archaeal rRNA is as different from Bacteria as Bacteria are

from eukaryotes (1978). Though other superficial differences, such as Archaea being originally discovered in so-called 'extreme' habitats, were used to support the concept of the three domains of life, dissenting opinions were proposed based on other molecular evidence. Signature sequences, conserved insertion/deletion patterns, in conserved proteins such as elongation factors, ribosomal proteins and amino-acyl tRNA synthetases were used to argue that the evolutionary relationship between the Gram-positive and Gram-negative Bacteria and the Archaea was not clear when compared to the relationships proposed by rRNA phylogenies (Gupta 1998). Nevertheless, the three domain classification is the more generally accepted system today.

Within the Archaea are two main groups, the Crenarchaeota and Euryarchaeota. A recent phylogeny based on 31 proteins and the analysis of the distribution of group-specific proteins reinforces the classification of the major groups within the Archaea (Gao and Gupta 2007). Specifically, the phylogeny shows the expected separation between Crenarchaeotes and Euryarchaeotes and, within the Euryarchaeotes, a possible polyphyletic history of methanogenesis as so-called class I and class II methanogens form separate groups apart from the non-methanogenic Euryarchaeota (Gao and Gupta 2007).

There also exists evidence to suggest a possible third group of Archaea, the Korarchaeota, of which the first genome of an organism termed "*Candidatus (Ca.)* Korarchaeum cryptofilum" has been completely sequenced very recently

(Elkins et al. 2008). Analysis of this composite genome produced via the whole-genome shotgun method, along with previous analysis of small subunit RNA obtained from environmental samples, suggest the proposed group Korarchaeota are deep-branching in the tree of Archaea, and may possess features of the earliest Archaea (Barns et al. 1996, Elkins et al. 2008).

The sequencing of a hyperthermophilic symbiotic archaeon, *Nanoarchaeum equitans*, along with its uncertain placement in the archaeal phylogeny led to the proposal of a fourth group of Archaea, the Nanoarchaeota (Huber et al. 2002, Waters et al. 2003). However, the degenerate nature of the *Nanoarchaeum equitans* genome at only 0.48 MB, coupled with high suspected evolutionary and lateral gene transfer rates, has been used to suggest that it is a highly derived Euryarchaeote rather than a member of a new kingdom of Archaea (Brochier et al. 2005).

1.5 Extremophiles

Regardless of the position the Archaea are placed in the tree of life, one of the characteristics that make them interesting as a group is the relatively large number of extremophiles counted among them. The classification of organisms as extremophilic depends on the nature of the environments they inhabit. Early on, the discovery of Archaea almost exclusively in extreme habitats meant that possessing such a lifestyle was considered a defining trait for these species (Woese et al. 1978). Since then, the discovery of other extremophiles in other

domains of life has become more common, though they remain dominantly prokaryotic.

Microbial extremophily has been studied as a rich source of basic and applied science. Since the description of *Thermus aquaticus* in 1969 (Brock and Freeze 1969), much effort has been made to describe the nature of extremophilic adaptations. The recognition of thermostable enzymes challenged what was understood of the basic principles of protein folding and stability with the promise of future of protein engineering applications (Argos et al. 1979). More immediately came the utilization of so-called 'extremozymes' as catalysts in new areas of biotechnology, most famously the use of a thermostable polymerase in the development of the ubiquitous polymerase chain reaction (Mullis et al. 1986; Saiki et al. 1988).

In addition to the various degrees of temperature dependence including thermophily (optimal growth temperature between 60 and 80°C, hyperthermophily (optimal growth temperature above 80°C) and psychrophily (optimal growth temperature below 15°C), extremophily includes acidophily and alkaliphily (requiring a pH below 3 or above 9 respectively). Barophily, also known as piezophily in some references, refers to requiring habitats under high gas or liquid pressure. Halophily, or more generally osmophily, refers to requiring 2M NaCl to live or some other high osmolyte concentration respectively. Xerophilic organisms are those living under dry or desiccating conditions. Some

authors feel it is important to remember that the definition of 'extreme' is relative to a human conception of 'normal' (Rothschild 2001). Bearing this in mind, preference of organisms for abundant or total lack of oxygen, a requirement of carbon dioxide known as capnophily, or other extreme metabolic requirements may also be described as a type of extremophily. As a group, the Archaea exhibit many types of these extremophilic lifestyles and are often polyextremophiles.

1.6 Aims of This Thesis

The basic problem of how life has adapted to various 'extreme' conditions has eluded explanation. It is well known that the proteins of mesophiles are not stable at the extreme high and low temperatures in which other organisms are capable of living. This thesis primarily focuses on temperature adaptation that has occurred in thermophilc and psychrophilic prokaryotes at the level of protein sequence. The comparison of sequences from either (hyper)thermophiles or psychrophiles to closely related mesophiles through the use of a phylogenetic model depends on the ability to choose valid pairs of species to compare. In Chapter 2 common methods for the construction of phylogenetic trees are described, and an example of how they were used to build an up-to-date tree of the Archaea is provided. After including every available completely sequenced Archaeal species at that time on a tree, pairs of closely related Archaea were identified for later analysis such that the optimal growth temperature difference in each pair was maximized. The chapter also describes the development of models of sequence evolution

beginning with a simple DNA model to the symmetric codon-based model employed as a starting point in chapter 3. Chapter 3 reviews the evidence for the existence of systematic differences in the protein composition of mesophiles versus extreme temperature organisms and an asymmetric model of sequence evolution that can incorporate these differences is derived. The asymmetric model is shown to be significantly better than the symmetric model at describing the evolution that has occurred between the mesophile-extremophile pairs selected. Finally, in chapter 4 the problem of identifying protein orthologues is revisited. The use of a common database search tool is illustrated via network graphs where links are drawn between query and search hits based on a common statistic. All-against-all local protein sequence searches are performed between two archaeal genomes and clusters are identified that represent 'units of similarity' as identified by the search algorithm. Applications of drawing the search results are discussed along with the interpretations of errors.

Chapter 2

Phylogenetic Methods and Model Building

2.1 Evolutionary Models

In this chapter some of the most common methods used to produce phylogenetic trees are described. Two methods make use of evolutionary models, and some simple examples are introduced. Following a previous work that employed these phylogenetic methods, an up-to-date tree for the Archaea is produced and the selection of pairs to be used in chapter 3 is described. The development of codon-based models is discussed. The motivation for this is that analysis of extremophile protein adaptation will require the selection of relevant pairs of species and a model that allows the frequencies of amino acids to vary between those species.

The purpose of evolutionary models is to describe evolution by creating mathematical terms that capture features of changing molecular sequences. The models differ in terms of complexity, which can be measured by the number of parameters required by each model and also the type of sequence information to which the models are fitted. Quantitative models of sequence evolution are important because they allow measurement of evolutionary distances or otherwise scoring changes between sequences which may then be used for scoring the

quality of sequence alignments and as the basis for building phylogenetic trees. Models are necessary because simply counting the observed number of changes between two sequences will always underestimate the true amount of evolution that has occurred. This is because mutations in sequences are assumed to occur at a given rate at random sites. As the amount of time increases there is an increasing chance that more than one mutation will have occurred at a single site though only the current state of the sequence is known. Models are used to estimate the actual amount of evolution that has occurred by fitting parameters to the observed sequence information. To introduce the topic of substitution models, the simplest case of the Jukes-Cantor model will be used (Higgs and Attwood 2005 pg. 60).

The Jukes-Cantor model describes the evolution of DNA sequences and makes use of a single parameter to describe the rates of substitution between nucleotides at a given site. A single rate parameter means the rate of substitution between bases is the same regardless of the identity of the bases. This may be expressed in the form of a 4x4 matrix where each element r_{ij} describes the substitution rate from any of the four bases, *i*, to any the four bases, *j*. The diagonal elements are the negative sums of the off-diagonal elements in each row such that each element r_{ii} represents the negative sum of the rate of substitution for the base i to any other base $j \neq i$ and the rates across the row sum to zero. Because the models describe the probability of observing substitutions in a

continuous time *t*, the rows of the rate matrix r_{ij} must sum to zero to satisfy the equation (Higgs and Attwood 2005 pg. 62):

$$\frac{dP_{ij}}{dt} = \sum_{k} P_{ik} r_{kj} \qquad (1)$$

Given an appropriate rate matrix the probability of observing a substitution from a state *i* to *j* in a time *t* may be calculated in a standard way. The details of each model come from which features are incorporated into the substitution rate matrix for each type of substitution. For example, instead of using a single rate parameter to describe the substitution rates between all nucleotides, transitions may be differentiated from transversions by using one parameter for each, denoted α and β, which are allowed to take different values. This specific example is called the Kimura two-parameter model (K2P) (Higgs and Attwod 2005 pg. 63; Kimura 1980). By increasing the number of parameters to distinguish between different types of substitutions the models gain complexity. Another common feature is to incorporate the different relative frequencies of each base, commonly denoted π, since intuitively the rate of substitution from one nucleotide to another depends on how often the nucleotides occur in a real sequence.

A model is said to be time reversible if it satisfies the condition that the frequency of base *i* multiplied by the rate of substitution from *i* to *j* is equal to the frequency of base *j* multiplied by the rate of substitution from *j* to *i*, or

$$\pi_i r_{ij} = \pi_j r_{ji} \quad (2)$$

This assumes the frequencies of bases remain constant and that evolutionary rates in the forward direction are equal to the rates in the reverse direction (Higgs and Attwood 2005, pg. 63). This property is important for phylogenetic methods, some of which require an evolutionary model in order to construct phylogenetic trees.

2.2 Phylogenetic Methods

As suggested in chapter 1.3 determining whether or not a set of species are evolutionarily related, and to what degree, is an important step in comparative genomics. Often the types of questions regarding the biology of a set of species that may be asked are dictated by the relationships of these species. Molecular phylogenetics is the extension of the theory of building trees by using observable shared phenotypic traits to describe evolutionary relationships between species to using the features of molecular sequences.

Part of the work embodied in this thesis included building a model to describe the evolution between pairs of related species that differ significantly by optimal growth temperature. Choosing the most relevant pairs required a method to determine the evolutionary closeness of a large group of species. The Archaea, as discussed in chapters 1.4 and 1.5, make an excellent group to investigate since they are a diverse group of organisms representing a wide range of lifestyles for

which more and more data are becoming available. Creating a phylogeny of the Archaea was therefore a useful starting point for this work.

Different tree-building methods are typically used because different methods of constructing trees result in different outcomes. The choice of method also determines what type of input is required and whether the tree being produced will be rooted with a defined ancestor, representing the earliest point of divergence between species, or otherwise unrooted.

One method of tree estimation is maximum parsimony. The principle of parsimony states that since observable changes in a molecular sequence should be expected to be rare the most plausible, or parsimonious, way of arranging species on a tree should be the one that minimizes the number of required changes. This criterion is not limited to molecular sequences as the definition of what constitutes a change can be any character, not necessarily the molecular identity of a base or amino acid at a given position in a sequence. There are no other parameters required for the construction of a maximum parsimony tree other than the number of changes counted between species (Higgs and Attwood 2005, pg. 177). A heuristic tree-search algorithm will attempt to search through the large possible number of unrooted trees until the tree with the fewest number of required changes is found (Higgs and Attwood 2005 pg. 177). A criticism of the maximum parsimony method is the problem of long-branch attraction, where

species with a large total number of changes relative to other species may be incorrectly grouped together (Wägele and Mayer 2007).

The neighbour-joining, or NJ, method is a distance matrix method, where all the species to be included on the tree are used to create a matrix of distances calculated from molecular sequences using some type of quantitative system (Higgs and Attwood 2005 pg. 166). For a simple example the distance between the same gene from any two species could be defined as the number of observed changes between their sequences. More complex types of evolutionary models could also be adapted through the use of a scoring system to generate a distance matrix. Once the distances are calculated, the two species with the smallest distance are joined together by creating a new node that becomes part of the tree, then the distances for the two species to the node are calculated, the distances of all the other species to the node are calculated, and finally the process is started over with the two closest joined neighbor species considered a single entity in the distance matrix (Higgs and Attwood 2005 pg. 167). As species are joined, the distance matrix becomes smaller and the algorithm is repeated until all species are joined on the tree. The NJ method produces unrooted additive trees, which are trees where the distance between two nodes is equal to the sum of the length of the branches between those nodes on the tree (Higgs and Attwood 2005 pg. 166). Since the distances between nodes is initially supplied by the distance matrix, and calculated distances between real sequences are not additive, the NJ method produces the tree of closest approximation to additive distances using the real

data. The closer the real data are to being additive, the closer the tree produced by the NJ method is to the ideal additive tree (Higgs and Attwood 2005 pg. 166).

The final method to be described is maximum likelihood (ML). This is the most complicated type of tree construction method since it requires an evolutionary model to judge the relatedness of species. The statistical models used to calculate relatedness are often substitution models that estimate the rates of substitutions between different elements of a sequence, which may be nucleotides, codons, or amino acids. For a given model with its associated parameters and a pair of species, the model describes the expected type and number of substitutions, and can be used to calculate the probability that the actual sequence data is observed (Higgs and Attwood 2005 pg. 173). When considering all data from all pairs of species to be included, the probability that the sequences are related according to a candidate tree with a specified topology and branch lengths may be calculated and the probabilities of different candidate trees may be compared to determine the most likely tree. This is another method that requires searching through many possible trees in so-called 'tree space' to find the best tree or trees, since there may be many candidates with similar probability. The method may be used to evaluate the probabilities of a limited set of specified tree topologies to determine, for example, the optimal tree when portions of the tree are known to have a reliable branching pattern and portions where there is ambiguity. When parts of a phylogeny are well known the number of candidate

trees to be searched may be reduced by assuming such parts are fixed, speeding up the overall search.

Bootstrapping may be applied to all three methods as a way to determine the level of support for the trees produced. Bootstrapping is used to overcome the effect of noise in the sequences on the final outcome of each tree by randomly sampling the dataset with replacement. The columns in the sequence alignments required to produce trees in each method are considered to be independent observations of the same site for each species. Each bootstrap iteration produces a dataset with the same length as the original but consisting of randomly sampled columns from the alignment with replacement such that some columns may be sampled many times, only once, or not at all. The data in some columns may not vary from species to species and therefore contains information about the phylogeny of the species. Some columns are extremely variable and represent noise in the sequence that obscures phylogenetic information. The information in different columns may produce different tree topologies. The method of sampling the data via bootstrapping results in the production of many trees, usually 100 or more, and provides information about how many times the most frequent topology was observed for every node and therefore which nodes are well-supported by the data and which are more ambiguous (Higgs and Attwood 2005 pg. 169).

In the case of the archaeal tree produced by Gao and Gupta (2007) the three different methods produced very similar topologies. To repeat this method, briefly, starting from the original dataset kindly provided by Beile Gao, additional sequences for the newly available species were identified through PSI-BLAST searches of public databases (Altschul et al. 1997) and added to the dataset. Alignments of these proteins were produced using the program ClustalW with default parameters (Chenna et al. 2003), and the TRANALIGN program of the EMBOSS package (Rice et al. 2000) was used to produce DNA alignments using the aligned protein coding regions. When repeating the phylogenetic work the NJ method was used, because of a computational speed advantage, to produce a phylogenetic tree bootstrapped 100 times using concatenated sequences stripped of poorly aligned regions of the same 31 proteins used previously. A total of 47 Archaea were included in this analysis, 18 more than before. Since that time five more complete genomes have been produced but have not been added to the dataset. The new tree possessed the same overall topology as the tree of the previous authors with the additional 18 species, and is shown in Figure 2.1. This tree was used to select pairs of species for further analysis. Pairs were selected such that relatively close neighbor species on the tree would represent related organisms where one member was an extremophile and the other a mesophile while being similar in other respects such as genome size and GC content. *N. equitans* was chosen as out-group only for the purpose of drawing the tree because of its uncertain placement and relatively long branch length.

Three pairs, referred to as A1, A2 and A3 were selected from the archaeal tree such that they differed in optimal growth temperature by 50, 28 and 23°C respectively, and represent thermophile pairs. Information regarding these pairs, and all other pairs to be discussed below, may be found in Table 3.1, while optimal growth temperature values with references and genomic GC content for all species used in this work may be found in Table 3.10.

Unfortunately, there were no acceptable cases where a mesophile and acidophile could be paired nor a mesophile and barophile. This is due to both the relatively small number of complete genomes available in the Archaea and the diversity of organisms they represent, and to lack of reliable information regarding the natural habitats of each organism. It is difficult to learn from the literature whether an organism isolated from the deep ocean is an obligate barophile or merely baro-tolerant. Information regarding optimal culture conditions obtained from various microbial culture collections is unfortunately vague regarding pressure requirements. For example, while *M. jannaschii* is reported to grow at up to more than 200 atm, the American Type Culture Collection (ATCC) specifies a minimum and maximum temperature and media requirement but no required pressure (Bult 1996, ATCC http://www.atcc.org/ *M. jannaschii* ATCC #43067).

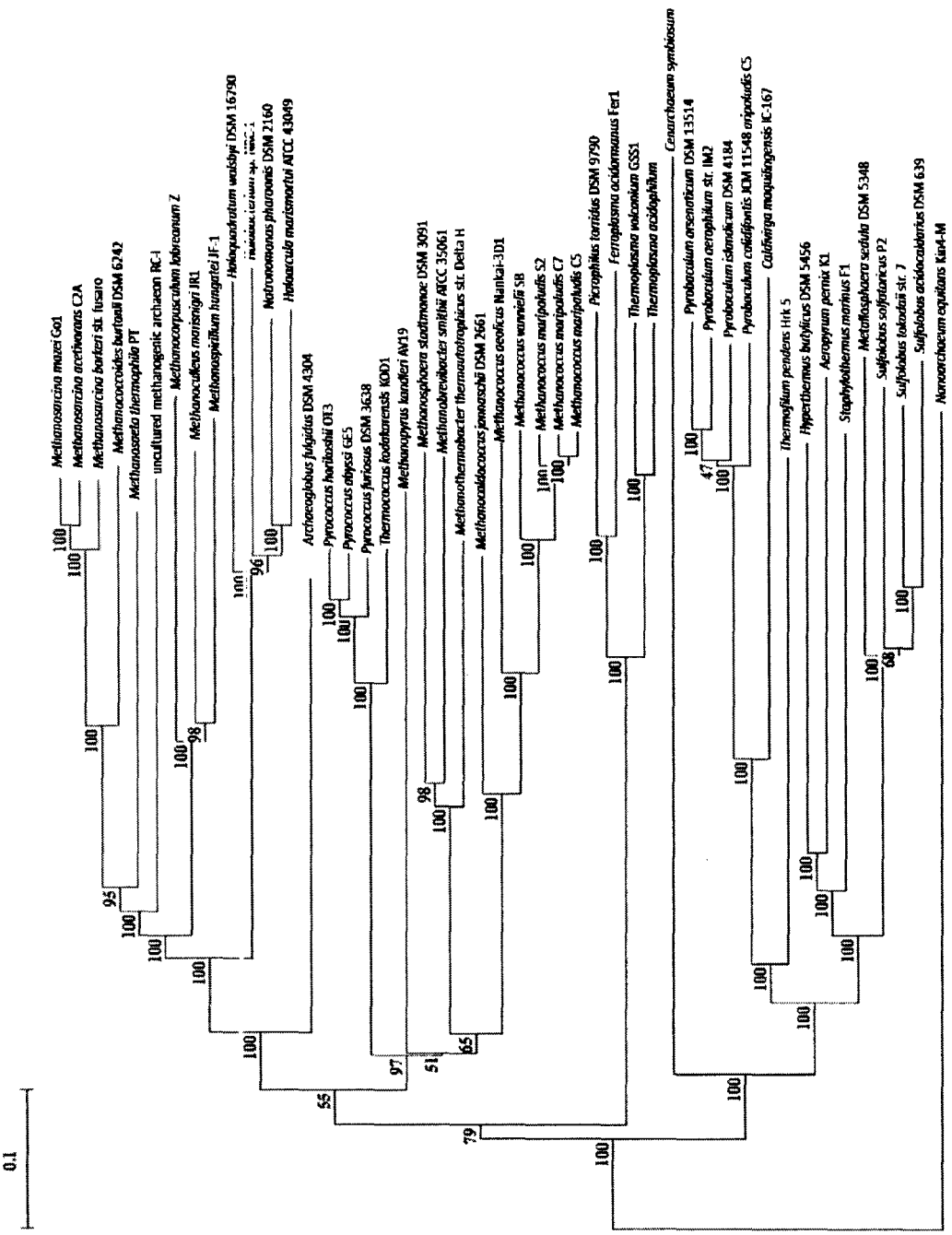

Figure 2.1 – Tree of the Archaea. Neighbour-Joining phylogenetic tree using 47 Archaeal species with *Nanoarchaeum equitans* used as outgroup. Node values indicate bootstrap support.

2.3 A General Symmetric Codon Model

The model used as a starting point for detecting asymmetric amino acid evolution between species pairs in chapter 3 was derived from a codon model used previously (Higgs et al 2007). This is referred to as the symmetric (S) model. There are numerous advantages for using codon-based models over either DNA or protein sequence models.

DNA models only require a 4x4 matrix and are therefore relatively small however the effect of natural selection on protein function is disregarded. Since protein function is based on structure and the structure of a protein depends on its amino acid sequence, when the amino acid sequence is not taken into account by DNA models this valuable information is lost. Similarly with protein-based models, only effects directly acting the amino acid sequence may be accounted for in the model and other information is not utilized. Effectively, only non-synonymous DNA substitutions are used by these types of models. Codon-based models are a compromise that can incorporate effects of selection on DNA and protein sequence simultaneously, albeit at the expense of requiring many more parameters than other types of models. The substitution matrix making use of every codon would be 64x64, or 4096 elements, significantly more than either the 4x4 or 20x20 DNA and amino acid models respectively. The size of the substitution rate matrix determines how difficult these models are to compute and also what format the sequence data needs to be tabulated in before it can be used

as input. A symmetric codon-based model is derived below. This model is to be applied to a pair of species, labeled A and B, to describe the evolution of their respective sequences in the time since these species diverged. The species pairs for the Archaea were chosen as described in chapter 2.1. Once model S is defined, along with the method to calculate maximum likelihood and the calculation of Akaike's Information Criterion (AIC) used for model discrimination, asymmetric terms may be added to the model, to be described in chapter 3.

		2nd position					
		U	C	A	G		
1st position	U	F (UUY)	S (UCY)	Y (UAY)	C (UGY)	Y	3rd position
		L (UUR)	S (UCR)		W (UGG)	R	
	C	L (CUY)	P (CCY)	H (CAY)	R (CGY)	Y	
		L (CUR)	P (CCR)	Q (CAR)	R (CGR)	R	
	A	I (AUY) I (AUA)	T (ACY)	N (AAY)	S (AGY)	Y	
		M(AUG)	T (ACR)	K (AAR)	R (AGR)	R	
	G	V (GUY)	A (GCY)	D (GAY)	G (GGY)	Y	
		V (GUR)	A (GCR)	E (GAR)	G (GGR)	R	

Table 2.1 – List of codon states. 32 Codon states used in all models with the amino acid each encodes. Stop codons are not included.

In model S codons are simplified into 32 states such that each is translated by a unique tRNA, disregarding stop codons. Most states include the codons of a given amino acid that differ by either a purine or pyrimidine in the third position. For example the phenylalanine codons TTC and TTU are included together under the codon state TTY. Single codons are represented by their actual codons, while four codon blocks are represented by their pyrimidine or purine ending codons. Table 2.1 summarizes the codon states used.

The model and its parameters are used to estimate relative substitution rates in the 32x32 matrix r_{ij} between codon states *i* and *j*. In model S each r_{ij} element can be described for non-synonymous substitutions as:

$$r_{ij} = \pi_j \alpha_{cat} K(i,j) e^{\frac{-d(i,j)}{D}} \quad (3)$$

Alternatively, each r_{ij} element for synonymous substitutions is described as:

$$r_{ij} = \pi_j \alpha_{cat} K(i,j) \quad (4)$$

for all $i \neq j$. The parameters π_j are measured from the dataset and represent the estimated equilibrium frequency of codon state *j*. The r_{ij} matrices for both species are normalized to an equilibrium rate of one by defining the diagonal elements r_{ii} to be equal to the negative sum of the remaining elements of each row, as with the Jukes-Cantor model described previously. Each element is divided by the negative sum over each state *i* of the product of the equilibrium frequency of state

i and the diagonal element r_{ii}. This is necessary to calculate the substitution probabilities $P_{ij}(t)$, where the parameter *t* is evolutionary time expressed in units of codon substitutions.

The α_{cat} parameters are introduced for each r_{ij} element to classify substitutions into different categories, therefore allowing different categories of substitutions to occur at different rates. Non-synonymous substitutions requiring 1 2 or 3 base changes between codon states *i* and *j* are accounted for by α_n, where n = 1, 2 or 3 respectively. Similarly, synonymous substitutions requiring 2 or 3 substitutions are accounted for by the α_5 and α_6 parameters. It is expected that synonymous substitutions involving one base change will occur most often and therefore α_4 is defined to be equal to one. Consequently α_1, α_2, α_3, α_5, and α_6 are expected to occur less frequently, and therefore be less than one. This is unlike other codon state substitution models, for example the model described by Liò et al. (1998), where substitutions requiring more than a single base change are explicitly forbidden and the rate set to zero.

The exponential term $\exp(-d(i,j)/D)$ for non-synonymous substitutions describes a decreasing function of the weighted distance $d(i,j)$ between the amino acids encoded by codon states *i* and *j*. This is important because the identity of the amino acid encoded by the codon at a given site may be related to one of its physical or chemical properties. Sites where an amino acid with a certain property is preferred will result in selection for amino acids with similar properties after a

mutation. Nine scales of physico-chemical properties of amino acids were transformed to have a mean of zero and variance of one to describe a 9-dimensional property space. The properties include volume, bulkiness, polarity, isoelectric point, hydrophobicity, hydrophobicity (alternate scale), water accessible surface area, fraction of surface area lost when folded, and the Woese polar requirement scale (Higgs and Attwood 2003 pg. 24; Higgs et al. 2007, and references therein). The d(*i,j*) function is the weighted Euclidean distance between amino acids encoded by states *i* and *j* in this space, under the constraint that the weights must sum to one. This may be written as:

$$d(i,j) = \left(\sum_{k=1}^{9} w_k (p_{ik} - p_{jk})^2 \right)^{\frac{1}{2}} \quad (5)$$

where *k* denotes the *k*th property of the amino acid encoded by codon states *i* and *j*. The weighted distance was previously shown to improve fit of the model starting from equal weights before optimization, at the cost of 8 additional parameters (Higgs et al. 2007). The parameter D controls the shape of the decreasing function. Once the model is fitted, the weighting of the different properties may provide insight as to which properties are more important than others

Hereafter, the terms described above to calculate r_{ij} will be referred collectively as Model 0. Model S (symmetric) will be Model 0 with the addition of the K(*i,j*) term described as follows. We expect transitions to occur more often

than transversions and therefore define a parameter κ which represents the transition/transversion ratio and should always be greater than 1. K(*i,j*) is a function equal to κ^n, where n = 0, 1, or 2 for substitutions from codon state *i* to *j* requiring 0, 1 or 2 transitions respectively.

The substitution rate matrix from the symmetric models can be used to calculate the substitution probabilities in time *t* since the sequences diverged as:

$$P_{ij}(t) = e^{(r_{ij} \cdot t)} \tag{6}$$

The relative substitution rate matrix is multiplied by a time factor measured in units of codon substitutions. The log-likelihood may be calculated as follows:

$$\ln L = \sum_i \sum_j n_{ij} \ln \left(\sum_k \pi_k P_{ki}(t) P_{kj}(t) \right) \tag{7}$$

where π_k, the frequency of state *k* in the ancestor, is estimated as the average frequency observed in species A or B. The term in parentheses is the sum over all codon states *k* at a site in the ancestral sequence of the probability of being substituted to state *i* in species A and to state *j* in species B in a time *t*. The log-likelihood is then the product of the number of times each *i* to *j* substitution is observed in the dataset and the log of the term described above, summed over all values of *i* and *j*.

Maximum likelihood is determined through a hill-climbing routine, evaluating the likelihood of the model starting with default initial parameters and

proceeding through a defined number of iterations. Following each iteration a randomly chosen parameter is increased or decreased by a small amount and the new likelihood is determined. Only changes that increase the likelihood are accepted. 50 000 iterations are used to achieve convergence in the parameters. Each model was repeatedly fitted 2 to 4 times, using different random seeds for each repetition.

The need to distinguish between models cannot be based on likelihood alone, as two models with a different number of parameters may possess the same likelihood. The best model is considered to be the one with the highest likelihood requiring the least number of free parameters. As in the previous work, Akaike's Information Criterion (AIC) is used to choose between models (Higgs et al. 2007). The AIC gives a score based on the maximum likelihood obtained for a given model penalized by the number of degrees of freedom in the parameters used by that model. This allows selection of a model that provides the best likelihood with the least complexity (extra parameters) to avoid fitting noise. This is calculated as

$$AIC = 2(-\ln \hat{L} + K) \tag{8}$$

where $\hat{L}$ is the maximum likelihood and K is the degrees of freedom in the parameters. The model to which all more complex models will be compared is Model S which possesses a total of 47 degrees of freedom (dof) in its parameters.

Chapter 3

Asymmetric Evolution of Prokaryotic Extremophiles

3.1 Introduction

Prokaryotic organisms have adapted to live in a very wide range of temperatures covering the full range from below 0 to just beyond 100 °C. Many of the proteins of typical mesophile organisms are not stable at extreme temperatures. Therefore, adaptation to extreme temperatures requires the evolution of unusual protein sequences. Several studies have identified significant statistical differences between sequences from mesophile and thermophile organisms (Haney et al. 1999; McDonald et al. 1999; McDonald et al. 2001). Di Giulio (2000) introduced a simple Thermophily Index that is a weighted sum of the frequencies of the amino acids in a sequence. He showed that this index correlates very strongly with the optimum growth temperature (OGT) of the organisms. Recently Zeldovich et al. (2007) have considered all possible subsets of amino acids to determine subsets whose frequencies strongly correlate with OGT. The high level of correlation of these simple indices with OGT suggests that there are selective forces on protein structure and function that act in similar ways in different thermophilic species. However, structural approaches have put forth many competing hypotheses for mechanisms of thermal adaptation. Simulated melting of proteins has been used to suggest greater entropic

stabilization at higher temperatures from lysine versus arginine due to the effect of the number of rotameric states accessible by each amino acid in the native state (Berezovsky et al. 2005). The examination of structures has also suggested that there may be multiple different effects at work in different situations, such as compactness increasing with temperature adaptation, or suggesting that small numbers of strong interactions between residues are responsible for thermal stability in some proteins (Berezovsky and Shakhnovich 2005). Comparisons between structures of a given set of mesophilic, thermophilic and psychrophilic variants of a single protein have been performed to examine denaturation and constraints in flexibility and rigidity (Bae and Phillips Jr. 2004). This has been extended to identifying substitution patterns in various environments in the 3D structure of proteins (Mizuguchi et al. 2007). Sequence-based methods have been used to look beyond substitutions towards specific patterns in the coupling of pairs of amino acids (Liang et al. 2005).

Although temperature adaptation is the most widely studied, adaptation of protein sequences to other factors has also been investigated. In the same way as for the thermophily index, indices have also been established to measure adaptation to high pressure (Di Giulio 2005a), acid pH (Di Giulio 2005b) and anaerobic versus aerobic conditions (Archetti and Di Giulio, 2007). In addition to the basic understanding of mechanisms of adaptation, which may allow directed engineering of functional proteins under extreme temperatures, this knowledge is important to understanding the evolution of current life in extreme conditions, and

to reveal useful information about conditions that existed at the time of the origin of life and the establishment of the genetic code.

In order to draw conclusions of this type, however, it is necessary to be sure that the variations in amino acids that are seen among different genomes are really the result of selection for the environmental condition in question, rather than some other factor that is specific to the species being compared. The aim of this project is to introduce models for protein sequence evolution that can be used to analyze sequences from pairs of related species that live in different environments, such as a mesophile and a thermophile. The method begins with an evolutionary model that is symmetric with respect to the two lineages, previously described in Chapter 2.3. In this model the expected numbers of substitutions between any two amino acids are equal in the forward and reverse directions, and the frequencies of the amino acids are constant in time. Asymmetric terms are then added to the model that allow forward and reverse substitution rates between two amino acids to differ. A systematic increase or decrease in frequency of an amino acid may then occur in one lineage with respect to the other. By comparing likelihoods of the data with the asymmetric and symmetric models, it is possible to ask whether the asymmetric terms are statistically significant and to identify the most important asymmetries in rates.

This method builds on that of Higgs et al. (2007), who used it to compare pairs of paralogous genes in *Saccharomyces cerevisiae* that have high and low expression levels. In that case, it was possible to distinguish between asymmetries

that arise due to selection for translational efficiency, translational robustness, and minimization of protein cost via reducing the use of metabolically expensive amino acids, all of which are expected to be stronger in highly expressed genes. A key advantage of the model is that it allows more than one asymmetric effect to be present at the same time and separates out the different effects. In the case of the mesophile/thermophile comparison the goal is to detect systematic selection on amino acid frequency arising from a need for the structural stability of proteins. However, an important confounding factor is that the GC content of the genome differs among species and this also causes changes in amino acid frequencies. The models used here allow for biased mutation rates that change the GC content and also for selective effects on amino acid usage.

3.2 Data Selection

Table 3.1 lists the pairs chosen for analysis in this work, including OGT for species A and B and the equilibrium GC content as measured by ϕ value derived from observed GC content at four-fold degenerate sites in the sequences used. The archaeal pairs were chosen based on the criteria outlined in Chapter 2.1.

Lacking a comparable bacterial phylogeny, selection of bacterial pairs was performed using taxonomy, via the NCBI taxonomy browser web server (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi). A list of candidate psychrophilic and (hyper)thermophilic bacterial species with completely sequenced genomes was created in May 2007, and subsequently

narrowed down by the removal redundant related species. For example, if two or more thermophiles in the same genus were present only one representative organism was chosen based on the ability to find a corresponding mesophile. Pairs were made by finding an organism within one or at most two higher levels of taxonomy with the goal of finding the closest pair possible. The sequences of the 31 proteins above were identified through PSI-BLAST searches of public databases for each species, downloaded for use in this work, and aligned as described in chapter 2.1. The thermophile (TB) and psychrophile (PB) pairs along with OGT values and ϕ measurements for each are included in Table 3.1. The bacterial species used in this work along with OGT values with references and genomic GC content are also included in Table 3.10.

In total, three archaeal and six bacterial pairs were chosen where one member is a mesophile and the other a (hyper)thermophile. Five bacterial pairs were chosen where one member was a mesophile and the other a psychrophile. In addition, two more archaeal pairs were chosen, labeled C1 and C2, such that their OGT values were the same. C1 are two species of *Methanococcus* with OGT of 37°C. C2 consists of a pair of *Pyrobaculum* species with OGT of 95°C. Both pairs are likewise listed in Table 3.1 and individual species along with OGT references are included in Table 3.10. The inclusion of these pairs was an attempt to test the assumption of the method when the members of a pair do not differ in terms of thermal lifestyle.

Pair	Species A	T_A	ϕ_A	Species B	T_B	ϕ_B
A1	*Methanococcus maripaludis* S2	*35*	*0.16*	*Methanocaldococcus jannaschii* DSM 2661	*85*	*0.12*
A2	*Methanosphaera stadtmanae* DSM 3091	*37*	*0.06*	*Methanothermobacter thermautotrophicus* str. Delta H	*65*	*0.49*
A3	*Ferroplasma acidarmanus* Fer1	*37*	*0.35*	*Picrophilus torridus* DSM 9790	*60*	*0.32*
TB1	*Clostridium difficile* 630	*37*	*0.05*	*Clostridium thermocellum* ATCC 27405	*55*	*0.38*
TB2	*Bacillus halodurans* C-125	*30*	*0.38*	*Thermoanaerobobacter tengcongensis* MB4	*75*	*0.32*
TB3	*Desulfitobacterium hafniense* Y51	*38*	*0.56*	*Carboxydothermus hydrogenoformans* Z-2901	*67*	*0.54*
TB4	*Deinococcus radiodurans* R1	*30*	*0.91*	*Thermus thermophilus* HB8	*85*	*0.95*
TB5	*Oceanobacillus iheyensis* HTE831	*28*	*0.18*	*Geobacillus thermodenitrificans* NG80-2	*65*	*0.64*
TB6	*Frankia* sp. CcI3	*28*	*0.90*	*Acidothermus celluloyticus* IIB	*55*	*0.87*
PB1	*Shewanella loihica* PV-4	*22*	*0.37*	*Colwellia psychrerythraea* 34H	*8*	*0.19*
PB2	*Acinetobacter* sp. ADP1	*37*	*0.27*	*Psychrobacter arcticus* 273-4	*22*	*0.32*
PB3	*Desulfovibrio desulfuricans* G20	*37*	*0.69*	*Desulfotalea psychrophila* LSv54	*10*	*0.32*
PB4	*Vibrio cholerae O1 biovar eltor str. N16961*	*37*	*0.36*	*Photobacterium profundum* SS9	*15*	*0.21*
PB5	*Marinobacter aquaeolei* VT8	*30*	*0.64*	*Psychromonas ingrahamii* 37	*5*	*0.27*
C1	*Methanococcus vannielii SB*	*37*	*0.35*	*Methanococcus aeolicus Nankai-3*	*37*	*0.35*
C2	*Pyrobaculum arsenaticum DSM 13514*	*95*	*0.66*	*Pyrobaculum islandicum DSM 4184*	*95*	*0.46*

Table 3.1 – List of species pairs. Pairs of related species with optimal growth temperature and GC content at fourfold-degenerate sites. Species A is a mesophile and species B is an extremophile in each case. A = Archaea; TB = thermophilic Bacteria; PB = psychrophilic Bacteria; C = control (equal growth temperatures).

3.3 Thermophily Indices

In this section the evidence that amino acid frequencies vary systematically with the OGT of the organism will be reviewed, and it will be shown that simple thermophily indices can be used as predictors of OGT. In general, these indices will be of the form:

$$I_i = \sum_a C_a f_{ai} \qquad (9)$$

where I_i is the value of the index in species i, f_{ai} is the frequency of amino acid a in species i (normalized so that the frequencies sum to 1 in each species), and C_a is a coefficient determining the importance of amino acid a to the index value. C_a is high for amino acids that increase in frequency in thermophiles and and low for those that decrease. The index is a useful predictor of OGT if there is a high correlation between I_i and T_i (the OGT of species i).

The scale of Di Giulio (2000) was derived from sequences of *Methanococcus* and *Bacillus*. It uses coefficients in the range 1 to 20. However, the range is not important because a linear transformation of these coefficients does not change the correlation. Therefore, to facilitate comparison with other scales discussed below, the coefficients of the Di Giulio scale were shifted so that the mean is zero and the standard deviation is 1 (shifted values shown in Table 3.2). Using these coefficients I_i was calculated for each genome in the data set.

There is a strong positive correlation with T_i, as shown in Figure 3.1 (Pearson correlation coefficient $R = 0.775$).

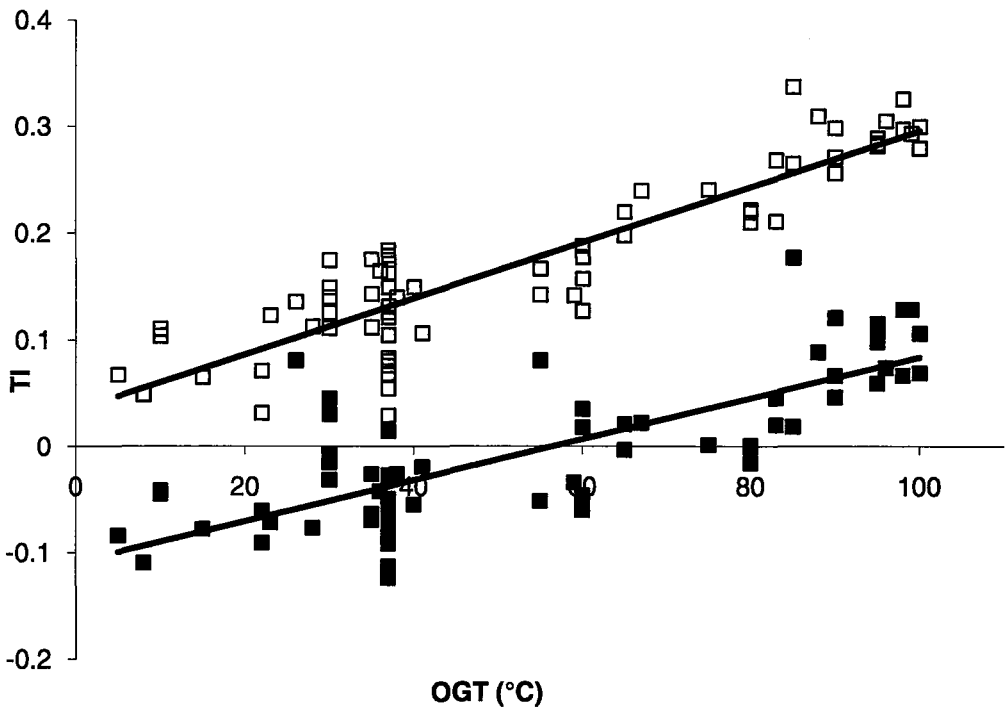

Figure 3.1 - Thermophily Index – Di Giulio and slope scale. Each point represents the TI calculated from amino acid frequencies obtained from all predicted open reading frames for each species plotted against OGT for that species. Solid squares – TI calculated using the normalized scale of DiGiulio (see text) R = 0.775. Open squares – TI calculated using the normalized slopes of amino acid frequencies versus OGT (see text) R = 0.901.

A linear regression was performed using each amino acid frequency f_{ai} against T_i. Another useful thermophily index can be obtained using the slopes of these regression lines. The mean value of the slope (averaged over the 20 amino

acids) is already constrained to be zero because the frequencies of the 20 amino acids sum to 1 for every species. The slopes were normalized so that the standard deviation was 1 (see Table 3.2) and used as the C_a coefficients. The plot of I_i versus T_i with these coefficients is also shown in Figure 3.1. The correlation is slightly higher ($R = 0.901$).

Table 3.2 – Amino acid scales and properties

a.a.	TI[a]	IVYWREL[b]	Ternary[c]	Normalized slope[d]	mean freq.[e] (%)
F	-0.05	0	-1	-0.08	4.49
L	0.63	1	0	0.99	11.13
I	0.97	1	0	0.35	8.22
M	-0.68	0	0	-0.40	2.76
V	0.49	1	1	1.62	8.59
S	-1.85	0	-1	-1.13	6.62
P	1.31	0	1	0.50	4.79
T	-1.07	0	-1	-1.26	5.77
A	0.10	0	-1	-0.58	9.10
Y	0.83	1	1	0.97	4.09
H	-0.24	0	0	-0.52	2.08
Q	-1.02	0	-1	-1.64	3.14
N	-1.60	0	-1	-1.05	4.37
K	-0.10	0	0	1.08	6.72
D	-0.88	0	-1	-1.35	6.01
E	0.15	1	0	1.27	8.15
C	0.63	0	-1	-0.36	1.07
W	1.51	1	1	0.19	1.22
R	1.75	1	0	1.38	6.07
G	-0.88	0	-1	0.03	8.23

a – normalized Thermophily Index scale from Di Giulio 2000
b,c – IVYWREL and ternary model from Zeldovich et al. 2007
d – slopes of amino acid frequency against OGT in the data set, normalized to mean zero and variance 1.
e – mean frequency of amino acids in all genomes in the data set.

Zeldovich et al. (2007) considered binary models in which each coefficient is 0 or 1, and ternary models in which each coefficient is 0, 1 or -1. Using a data set similar to the one used here, they found that the binary model that has the highest correlation with OGT has $C_a = 1$ for the amino acids IVYWREL. The best ternary model has $C_a = 1$ for VYWP and $C_a = -1$ for CFAGTSNQDH. Using these coefficients with the data set from this work, the correlation coefficients are $R = 0.945$ for the best binary model and $R = 0.934$ for the best ternary model. The plot of I_i versus T_i using these two sets of coefficients is shown in Figure 3.2.

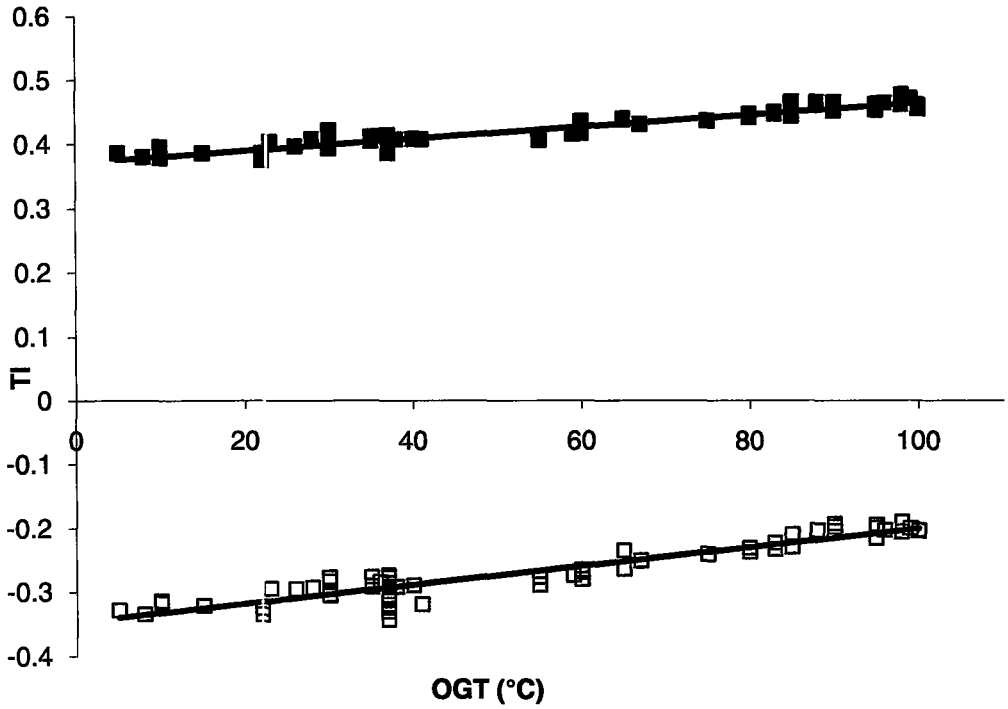

Figure 3.2 - Thermophily Index - alternate scales. Each point represents the TI calculated from amino acid frequencies obtained from all predicted open reading frames for each species plotted against OGT for that species. Open squares – TI calculated using the Ternary model scale of Zeldovich *et al.* (2007) R = 0.934. Solid squares – TI calculated using the IVYWREL scale of Zeldovich *et al.* (2007) R = 0.945.

3.4 Definition of Asymmetric Models

Asymmetric terms for amino acid and GC composition are introduced to the previously described model S as follows. The symmetric r_{ij} matrix must be calculated as in Chapter 2.3, with added terms for species A and B. The A and B species are defined such that A is the mesophile and B is either the thermophile or psychrophile being considered, depending on the pair. Amino acid asymmetry is

described as r_{ij} proportional to $(1 \pm \varepsilon_z \delta z_{ij})$ where the sign is positive for the r^B_{ij} matrix and negative for the r^A_{ij} matrix, ε_z is a parameter used as a scaling factor and δz_{ij} is the difference in value of coefficients $(z_j - z_i)$ for the non-synonymous substitution from codon state *i* to *j* on a given scale. This work uses a fixed scale derived from Di Giulio's original index (Di Giulio 2000), or free scales allowed to vary given the two constraints that the values must have a mean of zero and a variance of 1 at the cost of 18 additional parameter degrees of freedom to the model. When a fixed scale is being used, ε_z is allowed to be negative to accommodate the psychrophile pairs since the values in Di Giulio's scale described above are positive for increasing thermophily. This convention is retained when obtaining a scale as a set of free parameters, since species A is designated the mesophile and the positive sign is used for calculating the species B substitution matrix r^B_{ij}.

A term is needed to account for asymmetry in GC content which has been identified as a potential bias in species pairs with a large difference in GC content. Similarly to the asymmetric amino acid term above, the term $(1 \pm \varepsilon_{GC} \delta GC_{ij})$ is introduced where ε_{GC} is a parameter used as a scaling factor and δGC is the difference $(GC_j - GC_i)$ between the number of G or C bases in codon states *i* and *j*. Using the convention above, the sign is positive when calculating r^B_{ij}, and negative r^A_{ij} however the ε_{GC} value is allowed to be negative since either species

may have the greater GC content. When both effects are considered, we collect the terms as

$$r_{ij}^{A} = r_{ij}(1 - \varepsilon_{GC}\delta GC_{ij} - \epsilon_z \delta z_{ij}) \quad (10)$$

for species A and

$$r_{ij}^{B} = r_{ij}(1 + \varepsilon_{GC}\delta GC_{ij} + \epsilon_z \delta z_{ij}) \quad (11)$$

for species B. Because the asymmetric substitution rate matrices are now different, the substitution probability matrices are calculated separately. $P^{A}_{ij} = \exp(t \cdot r^{A}_{ij})$ and $P^{B}_{ij} = \exp(t \cdot r^{B}_{ij})$ describe the matrices P^{A}_{ij} and P^{B}_{ij} which denote the probabilities of going from codon state *i* to *j* for either of two species A and B in time *t*. We estimate the parameter π_k, the ancestral frequency of codon state *k* with the assumption that it is the average frequency of state *k* in species A and B. The pair frequencies f_{ij} predicted by the model can be calculated as

$$f_{ij} = \sum_{k} \pi_k P^{A}_{ki}(t) P^{B}_{kj}(t) \quad (12)$$

The pair frequencies describe the probability of codon state *k* in the ancestor of species A and B to be substituted to state *i* in species A and state *j* in species B in a time *t* since divergence, measured in units of codon substitutions.

The log-likelihood ln*L* given an aligned sequence and a model is calculated as:

$$\ln L = \sum_i \sum_j n_{ij} \ln f_{ij} \quad (13)$$

where n_{ij} is the matrix describing the observed number of substitutions from codon states i and j in the aligned sequences and f_{ij} are the pair frequencies as above.

Model F (Fixed Z-Scale) is defined as Model S with the addition of a term using a free ε_z parameter and a fixed Z-scale using the normalized scale of Di Giulio. Model V (Variable Z-scale) is Model F with the addition of allowing a free Z-scale and 18 additional parameters. Model GC is Model S with only the addition of an asymmetric GC parameter, ε_{GC}, as above. Model FGC is Model GC with the addition of a fixed Z-scale as in Model F. Model VGC is the same as Model FGC except the Z-scale is permitted to vary freely as in Model V. A summary of these models with a description and degrees of freedom (dof) for each is provided in Table 3.3.

Table 3.3 – Description of models

Model	Description	Degrees of freedom
S	Basic symmetric model	47
F	Model S + fixed asymmetric amino acid scale	48
V	Model V + variable asymmetric amino acid scale	66
GC	Model S + asymmetric GC	48
FGC	Model F + asymmetric GC	49
VGC	Model V + asymmetric GC	67

3.5 Asymmetry in Selected Pairs

Using the dataset of 31 concatenated proteins disregarding gaps and stop codons, codon substitutions for the described states were counted and entered into an appropriate n_{ij} matrix for each species pair. All models were fitted to these data. The values obtained for parameters in symmetric model S, for selected pairs, are provided in Table 3.4 and 3.5. These values are typical for each pair of species (data not shown).

Taking the AIC of model S to be a reference, we calculate the ΔAIC as:

$$\Delta AIC = (AIC_{model\ X} - AIC_{model\ S}) \quad (14)$$

where model X is the model under consideration. The ΔAIC values for the fitted models are provided in Table 3.6. Lower values of ΔAIC mean better performance of the model being compared to model S.

The α parameters were defined to explicitly account for the possibility of multiple base substitutions in a single time step, which has been shown in recent work to be a real feature of sequence evolution, though it was not attempted to show this is an improvement over a model that forbids multiple substitutions in this work (Kosiol et al. 2007, Higgs et al. 2007).

Table 3.4 - Amino acid weight parameters

Pair	*w1*	*w2*	*w3*	*w4*	*w5*	*w6*	*w7*	*w8*	*w9*
A1	0.000	0.272	0.000	0.076	0.094	0.000	0.291	0.127	0.141
A2	0.000	0.279	0.000	0.052	0.191	0.000	0.238	0.172	0.068
A3	0.000	0.249	0.000	0.039	0.128	0.000	0.297	0.197	0.090
Yeast	0.000	0.151	0.000	0.021	0.226	0.000	0.265	0.184	0.154

Table 3.5 - Parameters of model S

Pair	*t*	*α1*	*α2*	*α3*	*α5*	*α6*	*κ*	*D*
A1	0.930	0.040	0.014	0.005	0.579	0.080	1.496	0.661
A2	1.060	0.044	0.013	0.014	0.465	0.097	1.512	0.652
A3	0.785	0.042	0.017	0.008	0.708	0.037	1.474	0.661
Yeast	0.735	0.085	0.041	0.045	0.375	0.118	1.593	0.901

Table 3.4 shows physico-chemical property weight parameters: w_1 – volume; w_2 – bulkiness; w_3 – polarity; w_4 - isoelectric point; w_5 – hydrophicity; w_6 - hydrophobicity (alternate scale); w_7 - water accessible surface area; w_8 - fraction of area lost when folded; w_9 - polar requirement.

Table 3.5 contains the maximum-likelihood parameters for Model S for the three archaeal pairs in and the paralogous genes from yeast (taken from Higgs *et al.* 2007). Table 3.4 shows t – time in codon state changes per codon; α – rate parameters for each substitution category; κ – transition/transversion ratio; D – amino acid distance parameter.

The parameter κ behaves as expected, converging to a value greater than 1 in all cases. Allowing multiple substitutions required taking into account multiple transitions and transversions, and it was determined that the K(i,j) term in Model S is an improvement over Model 0 in every case (data not shown). This offers some additional support for the inclusion of multiple substitutions via the α parameters.

Model	A1	A2	A3	TB1	TB2	TB3	TB4	TB5	TB6	PB1	PB2	PB3	PB4	PB5	C1	C2
S	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
F	*-224.8*	*-135.9*	***-24.1***	*-18.5*	*-30.0*	*-36.8*	*-254.3*	*-94.2*	*-21.3*	*-7.7*	*-13.3*	*-5.9*	*-0.2*	*-1.5*	*-3.1*	*2.0*
V	*-287.0*	*-323.6*	*-13.3*	*5.1*	***-75.8***	***-59.0***	***-346.7***	*-95.1*	***-29.2***	***-13.2***	*-28.0*	***-14.8***	***-1.8***	***-55.8***	*-83.6*	*7.7*
GC	*-62.6*	*-490.9*	*2.0*	*-57.8*	*2.0*	*2.0*	*2.2*	*-59.0*	*2.0*	*2.0*	*2.0*	*2.1*	*2.0*	*2.0*	*-10.2*	***-26.2***
FGC	*-321.4*	*-557.2*	*-22.9*	*-68.1*	*-28.0*	*-34.8*	*-252.1*	***-125.3***	*-19.3*	*-5.7*	*-11.3*	*-3.9*	*1.8*	*0.6*	*-13.5*	*-24.3*
VGC	***-392.6***	***-674.4***	*-11.7*	***-156.3***	*-73.6*	*-56.9*	*-344.5*	*-115.0*	*-27.2*	*-11.2*	***-28.5***	*-12.7*	*0.2*	*-53.6*	***-115.9***	*-11.8*
ΔT	*50*	*28*	*23*	*18*	*45*	*29*	*55*	*37*	*27*	*-14*	*-15*	*-27*	*-22*	*-25*	*0*	*0*
Δϕ	*-0.04*	*0.43*	*-0.03*	*0.33*	*-0.05*	*-0.01*	*0.05*	*0.46*	*-0.03*	*-0.18*	*0.05*	*-0.37*	*-0.15*	*-0.37*	*0.00*	*-0.20*

Table 3.6 –ΔAIC values for fitted models relative to model S. The best fitting model in each case is in bold. Values of the difference in optimal growth temperature ($\Delta T = T_B - T_A$) and the difference in GC content at fourfold-degenerate sites ($\Delta\phi = \phi_B - \phi_A$) are also shown for comparison.

With a few exceptions, one of either Model V or Model VGC is regularly shown to have the largest ΔAIC and therefore describe the data best among the models considered. Cases where either of these models is not preferred can be explained by the strength of the asymmetric effect and the measured ϕ values from the DNA sequences of each species. In the absence of a strong asymmetric effect, the model with the least number of parameters dominates, explaining why model F is preferred for pair A3 (Table 3.6). In addition, models incorporating the GC term (Model GC, FGC and VGC) may dominate when the difference between ϕ of each species is large. The performance of each model depends on the strength of the asymmetry and strength of the GC effect. It is important to note that the GC models are not nested within Model V. When the GC effect is orthogonal to the asymmetric amino acid effect, the effects in Model VGC can be observed as approximately the sum of the GC or asymmetric effects in Model GC and F or V, which is true for most cases (see Table 3.6). When the GC effect is present, it is sometimes additive when combined with the asymmetric effect in Model VGC, or degenerate. Cases where the GC effect is strong may be predicted by looking at the magnitude of the difference between measured ϕ of the species in each pair (Δϕ, see Table 3.6). Cases where there is a large gain in ΔAIC for Models GC versus Model S are A1, A2, TB1, and TB5, and for all but A1, there is a greater than 30% difference in GC content as measured by ϕ (A2 $|\Delta\phi| = 0.4319$, TB1 $|\Delta\phi| = 0.3287$, TB5 $|\Delta\phi| = 0.4606$). It should be noted that even if a pair has a large

$|\Delta\phi|$, the GC effect may not necessarily be strong or preferred when fitting models.

3.6 Discussion

The parameters of Model S, listed in Table 3.5, are reasonably consistent. The *t* parameter is unique for each pair, describing roughly the distance between species in the pair. The D parameter describing the shape of the decreasing function of amino acid distance is remarkably consistent for all pairs (Table 3.5), and for all models (data not shown). Though the exact values vary slightly, the weight parameters for each amino acid property follow some general patterns (Table 3.4). The parameters w_1, w_3 and w_6, corresponding to the properties of volume, polarity and hydrophobicity (alternate scale), seem to tend towards zero in almost all models (data not shown). The highest weightings appear to be for w_2 and w_7, corresponding to bulkiness and water accessible surface area. It was expected that substitution rates would be related to specific physico-chemical properties since it is with respect to these properties that natural selection can act in the context of the protein.

Discussion in the literature regarding which properties are required for temperature adaptation are varied. Some explanation is provided by authors who have observed a trend to compactness for thermophilic proteins (Berezovsky and Shakhnovich 2005). Similarly, there is discussion over polarity being less important overall and hydrophobicity being only more important to the inner

cores of proteins where amino acids can contribute to overall compactness (Berezovsky et al. 2005, Mizguchi et al. 2007). Also, solvent exposed surface area may be an important factor for flexibility which has been discussed much in the context of thermal adaptation (Závodszky et al. 1998, Bae and Phillips Jr. 2004). Unfortunately, the parameters only describe relative weighting and as such provide little insight into the importance of each of these properties to specific types of adaptation. Also, the method treats each site of the input sequences equally, though there have been attempts by others to classify sites based on where they occur in the structure of a protein (Berezovsky et al. 2007, Mizguchi et al. 2007). Opinion in the literature seems to agree that physico-chemical properties of amino acids are important but in this work the method is unable to unambiguously show the importance of specific properties especially when some of the properties are correlated with each other such as volume and bulkiness.

To address the potential problem of multiple properties correlating with one another and therefore being redundant, models were modified to systematically exclude one of the nine properties at a time and fitted to pair A1. Changes in property weights versus models using all properties should then provide information regarding the potential correlation between properties. Results of these runs are provided in Table 3.7. The properties Volume, Polarity and alternate Hydrophobicity scale tend to zero in Model S. When each of these properties is systematically removed one by one, the ΔAIC values indicate they were contributing nothing to Model S except an unnecessary parameter.

	A1	Vol.	Bulk.	Pol.	pI	Hyd1	Hyd2	H_2O-SA	SA-Fold	WPR
Vol.	*0.000*	***0.000***	***0.192***	*0.000*	*0.000*	*0.000*	*0.000*	***0.246***	*0.000*	*0.000*
Bulk.	*0.272*	***-0.002***	***-0.272***	*0.000*	***-0.043***	***0.020***	*0.000*	***0.007***	***0.005***	***0.023***
Pol.	*0.000*	*0.000*	*0.000*	***0.000***	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	***0.020***
pI	*0.076*	*0.000*	***-0.020***	*0.000*	***-0.076***	***-0.004***	*0.000*	***0.005***	***0.011***	***0.022***
Hyd1	*0.094*	*0.000*	***0.119***	*0.000*	***-0.065***	***-0.094***	*0.000*	***0.045***	***0.089***	***0.050***
Hyd2	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	***0.000***	*0.000*	*0.000*	*0.000*
H_2O-SA	*0.291*	***0.001***	***-0.001***	*0.000*	***0.044***	***0.010***	*0.000*	***-0.291***	***0.002***	***0.004***
SA-Fold	*0.127*	*0.000*	***-0.027***	*0.000*	***0.040***	***0.051***	*0.000*	***-0.012***	***-0.127***	***0.023***
WPR	*0.141*	*0.000*	***0.008***	*0.000*	***0.101***	***0.016***	*0.000*	***0.001***	***0.019***	***-0.141***
ΔAIC	*0.00*	*-1.97*	*153.51*	*-2.00*	*86.40*	*6.65*	*-1.97*	*53.21*	*20.66*	*36.02*

Table 3.7 – Changes in property weights after removal of single properties. Amino acid property weights (w_1 through w_9) are reported for Model S in column A1. Each following column shows the change in property weights when the property heading each column was removed from the model. ΔAIC is also reported for the resulting model fit. Vol. – volume, Bulk. – bulkiness, Pol. – polarity, pI – isoelectric point, Hyd1 – hydrophobicity scale, Hyd2 – hydrophobicity scale (alternate), H_2O-SA - water accessible surface area, SA-fold. – surface area lost in folded state, WPR. – polar requirement. (See text).

	A1	A2	A3	C1	C2	TB1	TB2	TB3	TB4	TB5	TB6	PB1	PB2	PB3	PB4	PB5
Model S	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
DiG	***-224.78***	***-135.90***	***-24.15***	***-3.05***	*1.99*	***-18.45***	***-30.04***	*-36.84*	***-254.28***	***-94.22***	***-21.26***	***-7.68***	***-13.26***	***-5.89***	*-0.21*	*-1.48*
IVYWREL	*-75.13*	*-62.34*	*-18.59*	*-1.71*	*1.55*	*-1.56*	*-7.28*	*-18.12*	*-134.80*	*-18.19*	*-11.54*	*-6.15*	*-5.81*	*1.99*	*2.00*	*0.15*
Slope	*-49.89*	*-33.05*	*-7.93*	*0.65*	***0.98***	*-2.07*	*-25.99*	***-42.86***	*-233.85*	*-31.52*	*-8.74*	*-2.00*	*-12.80*	*2.05*	***-0.52***	***-4.24***

Table 3.8 – ΔAIC values for model F using alternate scales – Scales as in Table 3.2 (see text) for pairs listed in each column, ΔAIC calculated versus model S.

Alternatively, when a property with a relatively high weight is removed like bulkiness, or water accessible surface area, property weight is distributed in some cases to properties that received no weight before. This provides insight into which properties are contributing similar information to the model.

With regard to the scales used by Model F it is informative to see which scale performs best when supplied as a fixed amino acid asymmetry scale. Table 3.8 shows ΔAIC values for three runs of model F using Di Giulio's scale, the normalized slope scale and the IVYWREL scale. Di Giulio's scale performs best in 12 of 16 pairs, while the normalized amino acid slope scale is best in the remaining 4. It should be noted that in two of the four pairs where Di Giulio's scale was not best it must be concluded that no fixed scale is a significant improvement over model S. Despite the fact that Di Giulio's scale is the best performing fixed scale overall in model F, it is the weakest when used to calculate the TI relationship, while the IVYWREL scale which performs best when calculating TI is the worst scale supplied to model F. This suggests that any scale derived from model fitting parameters will not necessarily show a good correlation when used to calculate TI.

This is confirmed when model derived scales from the best variable model (V or VGC) are used as coefficients to calculate TI. Table 3.9 shows the correlation coefficients of these TI versus OGT plots. Only scales derived from fitting the data of 3 of 9 thermophile pairs have TI versus OGT correlation

coefficients greater than 0.80, while none of the scales derived from psychrophile pairs have correlation coefficients greater than ±0.30.

Pair	R
A1	0.346
A2	0.400
A3	**0.906**
TB1	0.141
TB2	0.500
TB3	0.660
TB4	**0.844**
TB5	**0.853**
TB6	0.532
PB1	-0.278
PB2	-0.295
PB3	-0.026
PB4	0.247
PB5	-0.136

Table 3.9 – Correlation of TI and OGT using model-derived scales – The scale from the best fitting variable model (V or VGC) fitted to each pair was used to calculate TI correlated against OGT as described in methods (see text)

When fitting this series of models to data from pairs C1 and C2, both with ΔOGT = 0°C some conflicting results are observed. The preferred model for hyperthermophile pair C2 was model GC, ΔAIC = –26.185, while ΔAIC for model F and model V were both positive. This suggests that the dominating effect in this pair is due to GC content and not amino acid asymmetry due to temperature adaptation, in agreement with the measurement of ϕ for these species ($\phi_{P.\ arsenaticum}$ = 0.663, $\phi_{P.\ islandicum}$ = 0.459). Positive ΔAIC for models F and V means that either a fixed or free Z-scale is worse than no Z-scale at all (model S),

which agrees with the assumption that since the ΔOGT for these two species is 0°C amino acid asymmetry aside from the GC effect should not be expected.

Mesophile pair C1 presents a different problem. The model best fitting the data is model VGC, ΔAIC = -115.893, the next best is model V ΔAIC = -83.582. The difference in ϕ between these species is effectively zero, and accordingly the GC effect is small. It must be concluded that there is an asymmetric effect present in these species even though they have the same OGT. Several explanations are possible. The criteria for pair selection were chosen so that any asymmetric effect observed between species would be most plausibly attributed to difference in lifestyle. This does not preclude the presence of an asymmetric effect in the absence of a significant difference in OGT. It may also be possible that there are relatively few ways for a sequence to be adapted to high temperature and relatively many ways for a sequence to be adapted to mesophile temperatures. If this is true, there may be fewer constraints on protein sequences from the mesophile pair C1 versus the hyperthermophile pair C2, and this may explain the unexpected amino acid asymmetry being detected in pair C1. The variable scale used in model V is designed to detect any asymmetry that may be present in the data, however the only conclusions that can be made based on the results of model fitting in the case of C1 are that the detected asymmetry is not due to adaptation to different temperatures and not attributable to differences in GC content.

Nevertheless, the results indicate that there is indeed detectable amino acid asymmetry in the species pairs analyzed here. In most cases this may be plausibly attributed to selection for thermal stability, even in cases where there is a simultaneous GC content effect. The models derived in this work are capable of discriminating between asymmetric GC and amino acid composition effects, however the scales derived from parameters used to fit these models are often not as good as other empirical sequence-derived scales when trying to correlate changes in a simple measurement such as the TI with the OGT values for a large group of species.

Table 3.10 - List of 69 species used

Organism	OGT (°C)	GC(%)*	OGT Ref
Aeropyrum pernix K1	95	56.3	Zeldovich et al. 2007
Archaeoglobus fulgidus DSM 4304	83	48.6	Zeldovich et al. 2007
Caldivirga maquilingensis IC-167	83	43.1	DSMZ
Cenarchaeum symbiosum	10	57.4	Preston et al. 1996
Ferroplasma acidarmanus Fer1	37	36.5	Macalady et al. 2004
Haloarcula marismortui ATCC 43049	37	61.1	Zeldovich et al. 2007
Halobacterium sp. NRC-1	37	65.9	Zeldovich et al. 2007
Haloquadratum walsbyi DSM 16790	37	47.9	JCM
Hyperthermus butylicus DSM 5456	99	53.7	DSMZ
Metallosphaera sedula DSM 5348	65	46.2	DSMZ
Methanobrevibacter smithii ATCC 35061	37	31.0	ATCC
Methanocaldococcus jannaschii DSM 2661	85	31.3	Zeldovich et al. 2007
Methanococcoides burtonii DSM 6242	23	40.8	Zeldovich et al. 2007
Methanococcus aeolicus Nankai-3	37	30.0	ATCC
Methanococcus maripaludis C5	37	33.0	ATCC
Methanococcus maripaludis C7	37	33.3	ATCC
Methanococcus maripaludis S2	35	33.1	Zeldovich et al. 2007
Methanococcus vannielii SB	37	31.3	ATCC
Methanocorpusculum labreanum Z	37	50.0	ATCC
Methanoculleus marisnigri JR1	30	62.1	ATCC
Methanopyrus kandleri AV19	98	61.2	Zeldovich et al. 2007
Methanosaeta thermophila PT	60	53.5	DSMZ
Methanosarcina acetivorans C2A	40	42.7	Zeldovich et al. 2007
Methanosarcina barkeri str. fusaro	35	39.2	Zeldovich et al. 2007
Methanosarcina mazei Go1	36	41.5	Zeldovich et al. 2007
Methanosphaera stadtmanae DSM 3091	37	27.6	Zeldovich et al. 2007
Methanospirillum hungatei JF-1	35	45.1	Zeldovich et al. 2007
Methanothermobacter thermautotrophicus str. Delta H	65	49.5	Zeldovich et al. 2007
Nanoarchaeum equitans Kin4-M	90	31.6	Huber et al. 2003
Natronomonas pharaonis DSM 2160	41	63.1	Zeldovich et al. 2007
Picrophilus torridus DSM 9790	60	36.0	Zeldovich et al. 2007
Pyrobaculum aerophilum str. IM2	100	51.4	Zeldovich et al. 2007
Pyrobaculum arsenaticum DSM 13514	95	55.1	JCM

Organism	OGT (°C)	GC(%)*	OGT Ref
Pyrobaculum calidifontis JCM 11548	90	57.2	JCM
Pyrobaculum islandicum DSM 4184	95	49.6	JCM
Pyrococcus abyssi GE5	96	44.7	Zeldovich et al. 2007
Pyrococcus furiosus DSM 3638	100	40.8	Zeldovich et al. 2007
Pyrococcus horikoshii OT3	98	41.9	Zeldovich et al. 2007
Staphylothermus marinus F1	90	35.7	ATCC
Sulfolobus acidocaldarius DSM 639	80	36.7	Zeldovich et al. 2007
Sulfolobus solfataricus P2	80	35.8	Zeldovich et al. 2007
Sulfolobus tokodaii str. 7	80	32.8	Zeldovich et al. 2007
Thermococcus kodakarensis KOD1	95	52.0	Zeldovich et al. 2007
Thermofilum pendens Hrk 5	88	57.6	DSMZ
Thermoplasma acidophilum	59	46.0	Zeldovich et al. 2007
Thermoplasma volcanium GSS1	60	39.9	Zeldovich et al. 2007
uncultured methanogenic archaeon RC-I	30	54.6	Ramakrishnan et al. 2001
Acidothermus cellulolyticus 11B	55	66.9	ATCC
Acinetobacter sp. ADP1	37	40.4	ATCC
Bacillus halodurans C-125	30	43.7	DSMZ
Carboxydothermus hydrogenoformans Z-2901	67	42.0	DSMZ
Clostridium difficile 630	37	29.1	ATCC
Clostridium thermocellum ATCC 27405	55	39.0	ATCC
Colwellia psychrerythraea 34H	8	38.0	Zeldovich et al. 2007
Deinococcus radiodurans R1	30	66.6	Zeldovich et al. 2007
Desulfitobacterium hafniense Y51	38	47.4	Zeldovich et al. 2007
Desulfotalea psychrophila LSv54	10	46.6	Zeldovich et al. 2007
Desulfovibrio desulfuricans G20	37	57.8	Zeldovich et al. 2007
Frankia sp. CcI3	28	70.1	Zhang et al. 1984
Geobacillus thermodenitrificans NG80-2	65	48.9	Feng et al. 2007
Marinobacter aquaeolei VT8	30	56.9	ATCC
Oceanobacillus iheyensis HTE831	28	35.7	DSMZ
Photobacterium profundum SS9	15	41.7	DSMZ
Psychrobacter arcticus 273-4	22	42.8	Zeldovich et al. 2007
Psychromonas ingrahamii 37	5	40.1	CIP
Shewanella loihica PV-4	22	53.7	ATCC
Thermoanaerobacter tengcongensis MB4	75	37.6	Zeldovich et al. 2007

Organism	OGT (°C)	GC(%)*	OGT Ref
Thermus thermophilus HB8	85	69.5	Henne et al. 2004
Vibrio cholerae O1 biovar eltor str. N16961	37	47.5	ATCC

* GC (%) is genomic GC content available from NCBI Microbial Genome Database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi Accessed July 2008)

Chapter 4

Network Relationships of Similarity Searches

4.1 Introduction

A problem not directly addressed in the data selection portion of chapter 2 revolves around how orthologous proteins are identified. A protein is orthologous in two species if it is present in the ancestor during the speciation event that gave rise to the two species in question. In other words, the protein in two related species was derived from the same ancestral protein when the species diverged. Because ancestral protein sequences are generally not available only measurements of similarity may be used to compare two or more modern day protein sequences to determine if they are orthologues.

The small dataset of 31 proteins used for the analysis in chapter 3 were identified by Gao and Gupta as conserved and widely distributed proteins (2007). Starting with a reasonable expectation that an orthologue for each of these proteins should be found in each species queried is not always possible. When a new genome is completed a host of new protein sequences may be identified and only some of them may show similarity to a protein of known structure and function. Often these new protein sequences are labeled 'conserved hypothetical' or just 'hypothetical' or 'unknown' depending on whether or not another protein

with similar sequence and unknown function has been identified previously. Reliably identifying protein orthologues via a method such as BLAST searching, as briefly described in chapter 1, can be somewhat difficult when scores for the quality of searches fall between the ranges of automatic rejection or acceptance.

Applications requiring large numbers of orthologous proteins for the purpose of detailed comparison between species, such as the development of the evolutionary model in chapter 3, would stand to greatly benefit from methods designed to use simple tools like BLAST to build such datasets. In this chapter, problems with using BLAST for this purpose are illustrated by the application of network theory to visualize the results of all-against-all protein-protein BLAST searches of every available predicted ORF of one species against another.

4.2 Data and Methods

The sequences of predicted open reading frames (ORFs) from 6 completely sequenced Archaeal genomes were obtained from the GenBank FTP site (http://www.ncbi.nlm.nih.gov/Ftp/ accessed August 2007). Sequences from extra-chromosomal elements were included with chromosomal sequences. The species were chosen such that one member of the pair is a mesophile, and the other a thermophile or hyperthermophile as per Table 3.1. The distinction between mesophile and thermophile is not important here as this chapter deals mainly with the results of BLAST searches and the significance of results rather than selection on the sequences though the original intention was to build a larger

dataset for a continuation of the work performed in chapter 3. The pairs are: mesophile *Methanococcus Maripaludis* S2, and the hyperthermophile *Methanocaldococcus jannaschii* DSM 2661 (Pair A1), the mesophile *Methanosphaera stadtmanae* DSM 3091, and thermpophile *Methanothermobacter thermautotrophicus* str. Delta H (Pair A2) and mesophile *Ferroplasma acidarmanus* Fer1 and thermophile *Picrophilus torridus* DSM 9790 (Pair A3).

A BLAST searchable database containing all predicted open reading frames was created for each species as per the documentation included with the standalone BLAST package (also available from the NCBI FTP referenced above). Sequences of one member of each pair were collected in a single file in FASTA format and supplied as queries to search against the database of sequences of the other member and vice versa. The default parameters for BLAST were used, including the Blosum62 scoring matrix, but an Expect-value, or E-value, cutoff of 10^{-1} was specified to generate a large list of search 'hits' between each pair of species. Practically, the Expect-value is a statistical measure that describes the probability of observing a match with the same or better score in the database being searched (Higgs and Attwood 2005, Chapter 7). The value of 10^{-1} is perhaps not a very strict cutoff, but the goal is to identify as many similarity matches, or 'hits', as possible.

A list of top hits for every species A ORF in species B and vice versa was generated. For example in pair A1, *M. maripaludis* S2 (species A) has a total of 1722 predicted ORFs which can be numbered from 1 to 1722, while *M. jannaschii* (species B) has 1786 numbered 1 to 1786. A program was written in C that accepts the list of 'hits' from species A against B, and B against A to cluster ORFs together to form networks based on similarity search results identified by BLAST at a specified Expect-value cutoff. The output of the cluster program was then used to draw a graph, using the Graphvis suite of programs, showing how the ORFs from this pair of species relates to each other in terms of similarity (Gansner and North 1999, http://www.graphvis.org/ accessed April 2008).

Every hit identified by BLAST can be represented as a directed link, which is a link that starts at one node and points at another node, drawn between two nodes where one is a species A ORF, and the other is a species B ORF. A cluster is identified as the set of nodes connected by all of their in- and out-links. A node can therefore only be part of a single cluster, as every other node pointing to a given node, or pointed at by another node will be identified as belonging to the same cluster. This is the basic methodology, though the number of links to draw to and from each node and the values of cutoffs may be changed in subsequent examples.

4.3 Results

The graph showing the network of BLAST hits between *M. maripaludis* S2 and *M. jannaschii* using an Expect-value cutoff of 10^{-1} is shown in Figure 4.1. Of the 1722 predicted ORFs in *M. maripaludis*, 1548 had 'hits' to *M. jannaschii* ORFs. Of 1786 ORFs from *M. jannaschii*, 1524 had similarity 'hits' in the reverse direction. ORFs with out-links represent 89.89% for species A, and 85.33% for species B. Of the 1524 'hits' from B to A, these represent only 1237 different ORFs of the total 1722 in species A, meaning that some nodes received multiple in-links. Similarly, of the 1548 'hits' from species A to B, only 1221 are represented of the total 1786 ORFs in species B.

The distribution of cluster sizes is shown in Figure 4.2. A cluster size of 1 represents an ORF without either an in or an out link, which can be interpreted as a sequence unique to its species or at least without any BLAST-identifiable similarity to a sequence in the other member of the pair using the specified parameters. A cluster size of 2 represents ORFs that either only have a single link between them, or ORFs with an in- and out- link and are therefore reciprocal top hits of each other. These represent the majority of clusters in Figure 4.1. There are still a large number of clusters of size 3, and after that the cluster size decreases rapidly. There are 5 clusters larger than 10, two of which are larger than 30. There are 1239 clusters in total.

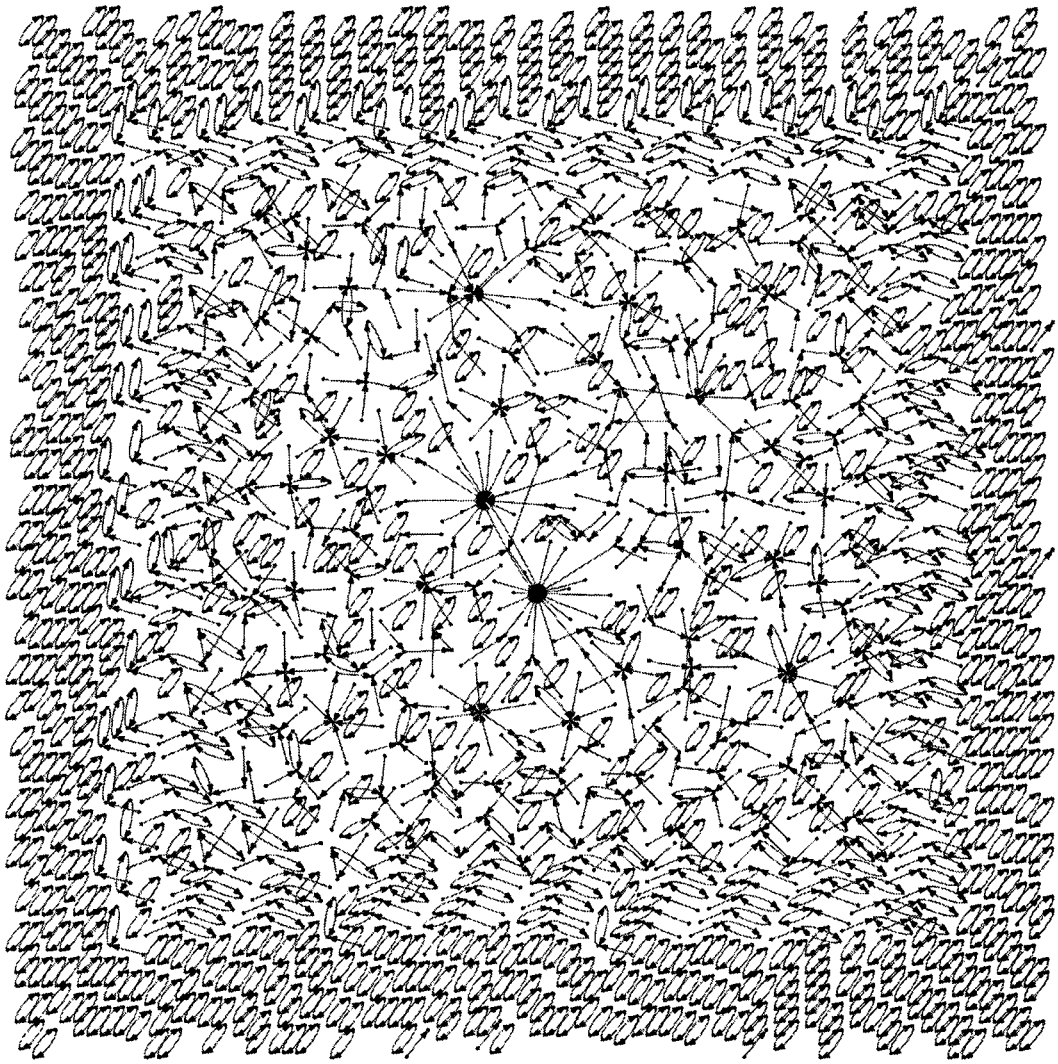

Figure 4.1 – Network of top BLAST results ($E < 10^{-1}$) for pair A1. Each node represents a predicted ORF from *M. maripauldis* or *M. jannaschii* and each link drawn represents a top BLAST 'hit' using an expect of $E=10^{-1}$ when all predicted open reading frames from one species are queried against the other and vice versa.

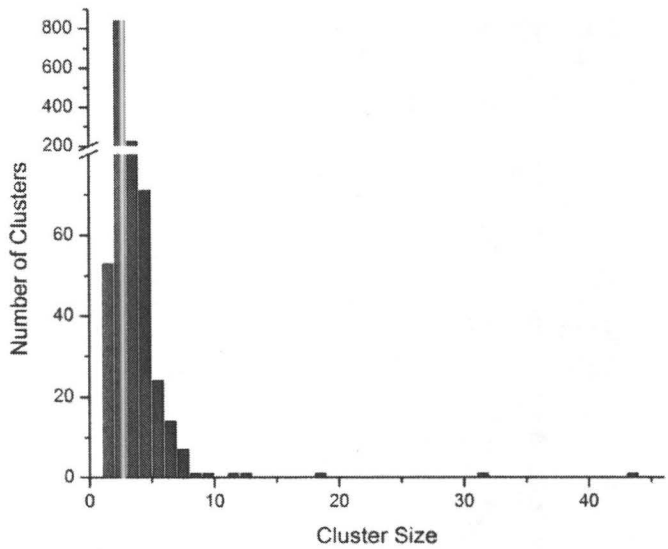

Figure 4.2. – Distribution of cluster sizes in Figure 4.1. Clusters are defined in figure 4.1 as the set of nodes joined together by a group of in- and out-links.

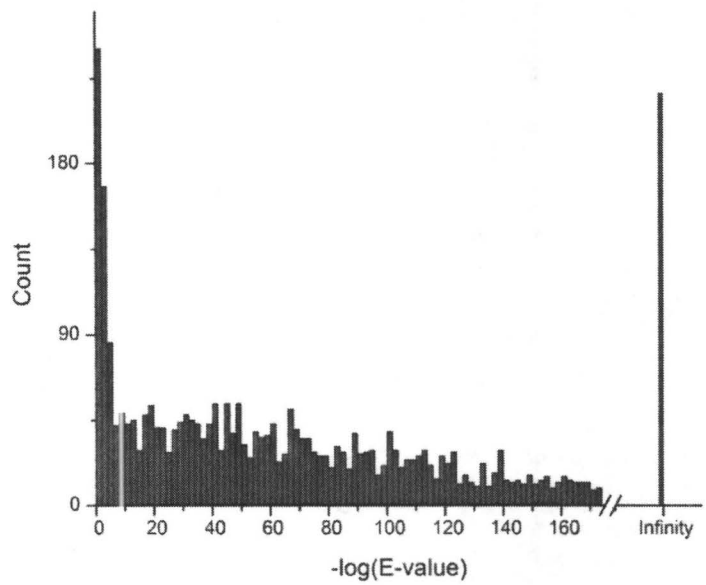

Figure 4.3. – Distribution of –log(E-value) of top 'hits' for pair A1 ($E < 10^{-1}$). All-against-all BLAST results using every predicted open reading frame from the genomes of *M. maripaludis* S2 (n = 1722) and *M. jannaschii* (n = 1786). Top 'hit' results with an E-value < 10^{-1} are shown. Results plotted as –log(E-value). Cases where E-value is effectively 0 are plotted on the x-axis at 'Infinity'

Modification to the method of drawing 'hits' can include drawing links for every BLAST 'hit' above a given threshold. For example, in the 10^{-1} E-value cutoff case where only the top hit for each query was ultimately used to draw links for pair A1 there were a total of 18 790 BLAST hits. The vast majority of 'hits' were discarded when drawing the graph. A network using a 10^{-10} E-value cutoff was produced, shown in Figure 4.4, for the same pair this time drawing links for every 'hit' above this threshold resulting in 4522 'hits' being identified. This is a comparable to the 10^{-1} case where 3072 top out-links were identified.

It was possible to draw the graph without explicitly clustering the nodes first, as the drawing software treats each link independently, however without clustering it is difficult to manually determine the average cluster size and number of clusters. In this case 4522 links need to be drawn between a total of only 3508 nodes, obviously meaning some nodes will end up being more highly connected than others and it might therefore be expected to observe at least one very large cluster. This is indeed observed in Figure 4.4, where at least two large highly linked clusters make discerning the individual nodes and links they encompass almost impossible. In this case, smaller clusters are still interesting because they will likely represent units of distinct similarity or uniqueness, while the large clusters will require further computation to be separated and analyzed.

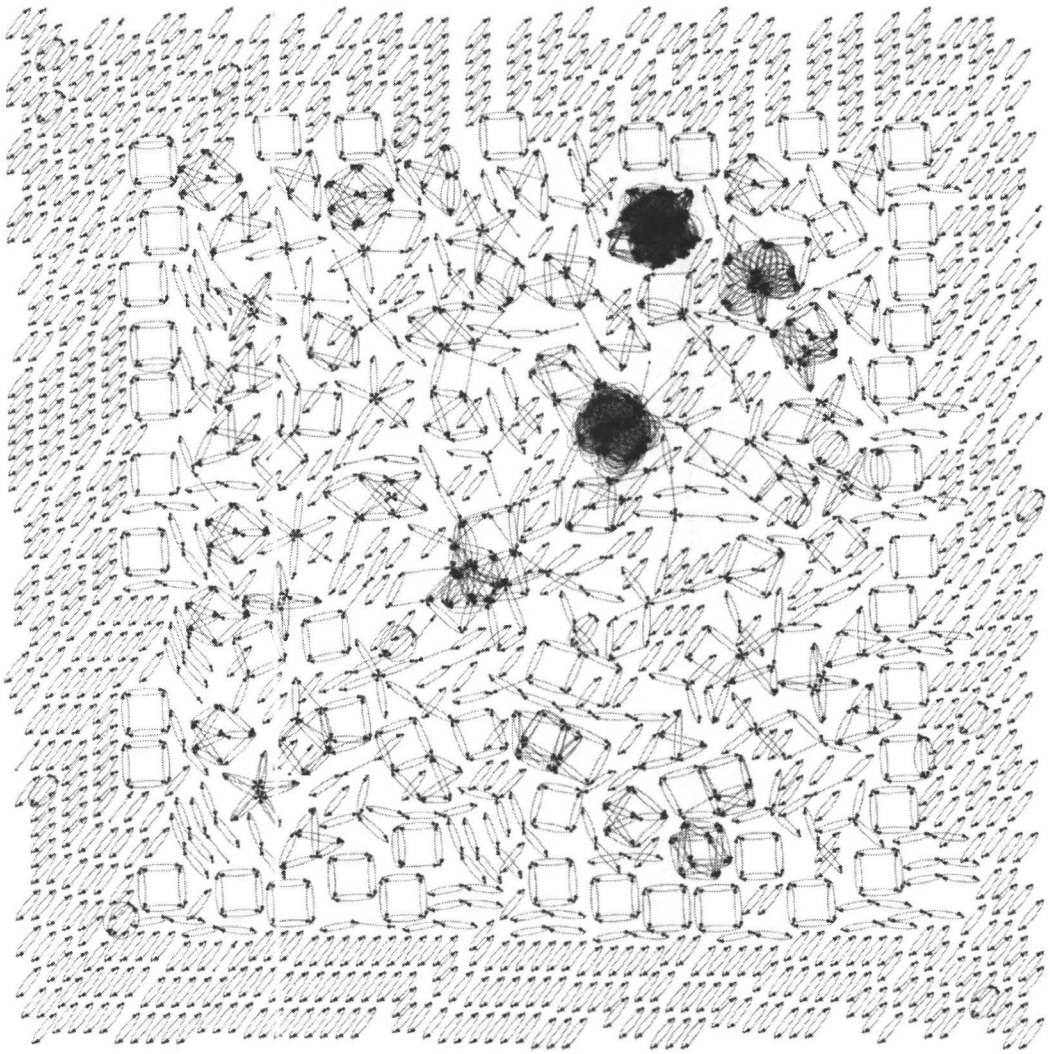

Figure 4.4. – Network of all BLAST results ($E < 10^{-10}$) for pair A1. As in figure 4.1 each node represents a predicted open reading frame from the species in pair A1, except instead of only top 'hits', every link representing a 'hit' with E-value < 10^{-10} is drawn.

The comparison of two example clusters from Figure 4.1 and 4.4 is illustrative. Figure 4.5 shows the cluster of size n = 18. Each node is labeled with the annotation of the ORF it represents and the E-value and percent identity for the BLAST result used to draw its outgoing link. Similarly, Figure 4.6 shows a cluster from Figure 4.4 of size n = 15. Each node is labeled with its associated annotation and length. Because there are multiple in- and out-links drawn, E-values and percent identity for each link are not provided.

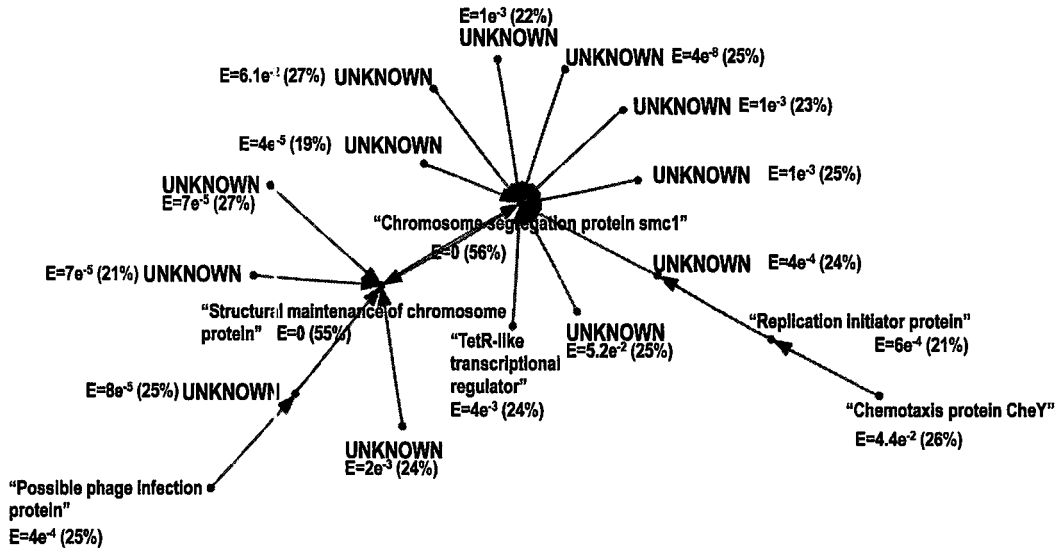

Figure 4.5 – Cluster of size n=18 taken from Figure 4.1. Annotations for each ORF are overlaid where available. Links represent top BLAST results. The E-value for these links is provided along with percent identity for the length of match.

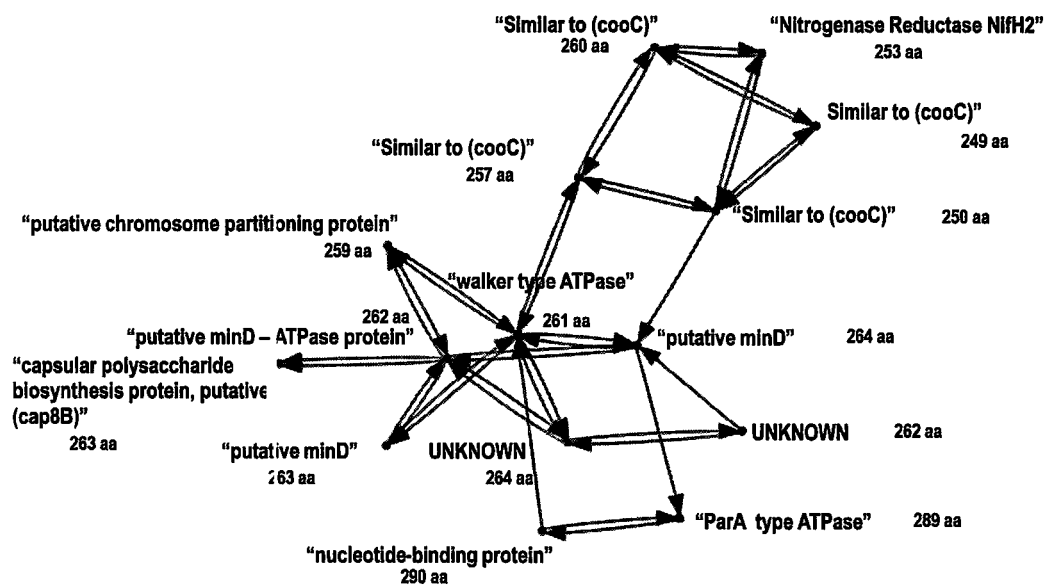

Figure 4.6 – Cluster of size n=15 taken from Figure 4.4. Each node represents an ORF from either species A or B. Annotations for each ORF are overlaid where available along with the length of each ORF.

4.4 Discussion

Preliminary interpretation of the biological meaning of the clusters must be considered with respect to the Expect-value cutoff. The E-value cutoff of 10^{-1} is quite high, and was initially chosen to ensure a balance between meaningful links and link number. Too low a cutoff and only the most very similar ORFs will be identified as 'hits'. Too high a cutoff and many biologically meaningless links may be drawn. Figure 4.3 shows the distribution of –log(E-values) for the list of 'hits' combined from both species. While there are a large number close to the cutoff, there are about as many with an effective E-value of 0, plotted at infinity on the x-axis. The majority of 'hits' lie at values larger than 20, suggesting that a large number of significant hits may be found with this criterion, judging by E-value alone.

With regard to cluster size in Figure 4.2, the majority of clusters are of size 2. If two identical sets of ORFs were used in the same procedure, it would be expected that every single ORF would have perfect similarity to itself, and only clusters of size 2 would be observed representing perfect top 'hits'. Given that the species pair was originally chosen based on relative evolutionary closeness the large number of size 2 clusters is not surprising. There are relatively few cases where a single out-link connects two nodes and this is likely an artifact of the BLAST parameters, cutoff value and potentially large difference in length between the two ORFs represented in the cluster. Not shown in Figure 4.1 are

ORFs with neither in- nor out-links. These ORFs can therefore be said to represent sequences with no appreciable amount of similarity to any ORF in the partner species. This may be because they are either pseudogenes, or otherwise unique genes in this pair resulting from a gain/loss event occurring at some point since the two species diverged.

The next step beyond showing that these graphs can be produced and pose interesting questions is to describe a method to analyze the networks produced by these procedures. For example, a cluster-by-cluster examination may further reveal the quality of the matches each link represents. Checking the annotation for the predicted ORFs in these clusters might reveal something interesting, where annotations are available. Nodes may be drawn on the network to indicate whether they have annotations or not, and other information may be overlaid, for example if a node contains a protein domain or motif of interest, or whether it may be assigned to a particular COG (Tatusov et al. 2003).

Two examples of this were provided in Figure 4.5 and 4.6. Annotation information is overlaid on the cluster structure along with some information regarding the BLAST results. In the case of Figure 4.5, most nodes are lacking annotation and it is clear that E-values for individual links are quite low except for the two central nodes identified as reciprocal top hits with E-values of 0.0. These two central nodes are also the largest, both approximately 1000 residues long.

With such a low cutoff used ($E < 10^{-1}$) the links from nodes to the central large sequences are likely due to chance rather than real homology.

Using a lower cutoff ($E < 10^{-10}$), the cluster depicted in Figure 4.6 displays much more biological relevance. The cluster can be divided into two portions, the upper-right which appears to be related to those ORFs possessing similarity to cooC, a nickel-insertion enzyme according to annotation, and the lower-left portion which according to annotation for those ORFs appears to be related to ATPase-like sequences. In this case, the lengths of all ORFs are shown to be relatively similar so there is little chance that these BLAST results are due to chance matches from short sequences to larger ones as in Figure 4.5.

A more immediate need includes the refinement of the clustering algorithm for use with nodes possessing multiple out-links as in Figure 4.4. The ability to correctly label clusters will allow the ability to sort, count and otherwise manipulate these clusters for a more in-depth examination. Other questions to ask of networks like that depicted in Figure 4.4 include: Do the shapes of the clusters reveal anything about the ORFs counted in the cluster? It is possible to classify every cluster based on the number of nodes and links included and the pattern of linkage. Different patterns may represent biologically meaningful events. For example there are several easily identified clusters in Figure 4.4 consisting of 4 nodes in the shape of a square where each node is reciprocally connected to its neighbours with an in- and out-link. This may possibly be a case where a gene

duplication in the ancestor of these two species occurred before a speciation event, so now both daughter species possess the duplicated ORFs and therefore might display a pattern of linkage between these ORFs in the square shape as described.

While to date only the first pair A1 have been used to produce the first graph, graphs for other pairs of species are easily produced. Pending the refinement of these methods it would be worthwhile to determine whether or not the same types of clusters are observed in multiple pairs of species, and if so, are the same orthologues residing in similarly shaped large clusters?

Network visualization represents a novel way to illustrate the results of similarity searches, in this case, those produced by BLAST though any similar method would be amenable. Given the appropriate criteria for determining which ORFs or nodes need be linked, and a robust clustering method, the similarity relationships may be sorted and counted and visualized in many different and interesting ways.

Bibliography

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

[2] Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nuc. Acids Res. 25:3389–3402.

[3] ATCC - American Type Culture Collection - http://www.atcc.org

[4] Archetti M, Di Giulio M. 2007. The evolution of the genetic code took place in an anaerobic environment. J. Theor. Biol. 245: 169-174.

[5] Argos P, Rossmann MG, Grau UM, Zuber H, Frank G, Tratschin JD. 1979. Thermal Stability and Protein Strucure. Biochemistry. 18(25):5698-5673.

[6] Bae, E and Phillips Jr, GN. 2004. Structures and Analysis of Highly Homologous Psychrophilic, Mesophilic, and Thermophilic Adenylate Kinases. J. Biol. Chem. 279:28202-28208.

[7] Barns SM, Delwiche CF, Palmer JD, Pace NR. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. USA. 93:9188-9193.

[8] Berezovsky IN, Chen WW, Choi PJ, Shakhnovich EI. 2005. Entropic stabilization of proteins and its proteomic consequences. PLoS Comput. Biol. 1(4):e47.

[9] Berezovsky IN, Shakhnovich EI. 2005. Physics and the evolution of thermophilic adaptation Proc. Natl. Acad. Sci. USA. 102(36):12742-12747.

[10] Blattner FR, Plunkett III G, Bloch CA, et al. (17 co-authors). 1997. The Complete Genome Sequence of *Escherichia coli* K-12. Science. 277(5331):1453-1462.

[11] Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Médigue C, Danchin A. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. Nuc. Acids Res. 23(17):3554-3562.

[12] Bult CJ, White O, Olsen G, et al. (40 co-authors). 1996. Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*. Science. 273(5278):1058-1073.

[13] Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P. 2005. Nanoarchaea: representatives of a novel archeal phylum or a fast-evolving euryarchaeal linear related to Thermococcales? Genome Biol. 6:R42.

[14] Brock TD, Freeze H. 1969. *Thermus aquaticus* gen. n. and sp. n, a Non-sporulating Extreme Thermophile. J. Bacteriol. 98(1):289-297.

[15] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nuc. Acids Res. 31(13):3497-3500.

[16] CIP - Collection de l'Institut Pasteur - http://cip.pasteur.fr/rech-bacteries-gb.html.

[17] Cole ST, Brosch R, Parkhill J, et al. (42 co-authors). 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 393:537-544.

[18] Di Giulio M. 2000. The late stage of genetic code structuring took place at a high temperature. Gene 261(1):189-95.

[19] Di Giulio M. 2005a. Comparison of proteins from *Pyrococcus furiosus* and *Pyrococcus abyssi*: barophily in the physicochemical properties of the amino acids and in the genetic code. Gene 346: 1-6.

[20] Di Giulio M. 2005b. Structuring of the genetic code took place at high pH. J. Theor. Biol. 237: 219-226.

[21] DSMZ - German Collection of Microorganisms and Cell Cultures - http://www.dsmz.de

[22] Elkins JG, Podar M, Graham DE, et al. (20 co-authors). 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc. Natl. Acad. Sci. USA. 105(23):8102-8107.

[23] Feng L, Wang W, Cheng J, et al. (12 co-authors). 2007. Genome and proteome of long chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. Proc. Natl. Acad. Sci. USA. 104(13):5602-5607.

[24] Fleischmann RD, Adams MD, White O, et al. (40 co-authors). 1995. Whole-Genome Random Sequencing and Assembly of *Haemophilus Influenzae* Rd. Science. 269(5223):496-512.

[25] Gansner ER, North SC. 1999. An open graph visualization system and its applications to software engineering. Softw. Pract. Exper. 00(S1):1-5.

[26] Gao B, Gupta RS. 2007. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics 8:86.

[27] Gupta RS. 1998. What are the archaeabacteria: life's third domain or monoderm prokaryotes related to Gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. Mol. Microbiol. 29(3):695-707.

[28] Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. Proc. Natl. Acad. Sci. U. S. A. 96(7):3578-3583.

[29] Henne A, Bruggemann H, Raasch C, et al. (20 co-authors). 2004. The genome sequence of the extreme thermophile *Thermus thermophilus.* Nat. Biotechnol. 22(5):547-553.

[30] Higgs PG and Attwood TK. 2005. Bioinformatics and Molecular Evolution. Malden MA: Blackwell Science Ltd.

[31] Higgs PG, Hao W, Golding B. 2007. Identification of Conflicting Selective Effects on Highly Expressed Genes. Evolutionary Bioinformatics. 2:1-13.

[32] Huber H, Hohn MJ, Stetter KO, Racher R. 2003. The phylum Nanoarchaeota: Present knowledge and future perspectives of a unique form of life. Res. Microbiol. 154:165-171.

[33] JCM - Japan Collection of Microorganisms - http://www.jcm.riken.go.jp

[34] Kimura M. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences. J. Mol. Evol. 16:111-120.

[35] Kosiol C, Holmes I, Goldman N. 2007. An Emprical Codon Model for Protein Sequence Evolution. Mol. Biol. Evol. 24(7):1464-1479.

[36] Liang HK, Huang CM, Ko MT, Hwang JK. 2005. Amino Acid Coupling Patterns in Thermophilic Proteins. Pro. Struct. Func. Bioinformatics. 59:58-63.

[37] Liò P, Goldman N. 1998. Models of Molecular Evolution and Phylogeny. Genome Res. 8(12):1233-1244.

[38] Macalady JL, Vestling MM, Baumler D, Boekelheide N, Kaspar CW, Banfield JF. 2004. Tetraether-linked membrane monolayers in *Ferroplasma* spp: a key to survival in acid. Extremophiles 8:411-419.

[39] Madabhushi RS. 1998. Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. Electrophoresis. 19:224-230.

[40] Madigan MT, Martinko JM and Parker J. 2003. Brock Biology of Microorganisms. 10th Edition. Upper Saddle River NJ: Pearson Education, Inc.

[41] McDonald JH, Grasso AM, Rejto LK. 1999. Patterns of Temperature Adaptation in Proteins from *Methanococcus* and *Bacillus.* Mol. Biol. Evol. 16(12):1785-1790.

[42] McDonald JH. 2001. Patterns of Temperature Adaptation in Proteins from the Bacteria *Deinococcus radiodurans* and *Thermus thermophilus.* Mol. Biol. Evol. 18(1):741-749.

[43] Mizuguchi K, Sele M, Cubellis MV. 2007. Environment specific substitution tables for thermophilic proteins. BMC Bioinformatics 8:S15.

[44] Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 51(1):263-273.

[45] NCBI Microbial Genome Database

http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

Accessed July 2008.

[46] NCBI Taxonomy browser

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

Accessed August 2007.

[47] Preston CM, Wu KY, Molinski TF, DeLong EF. 1996. A psychrophilic crenarchaeon inhabits and marine sponge: *Cenarchaeum* symbiosum gen. nov., sp. nov. Proc. Natl. Acad. Sci. USA. 93:6241-6246.

[48] Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. 1987. A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides. Science. 238(4825):336-341.

[49] Ramakrishnan B, Lueders T, Dunfield PF, Conrad R, Friedrich MW. 2001. Archaeal community structures in rice soils from different geographical regions before and after initiation of methane production. FEMS Microbiol. Ecol. 37:175-186.

[50] Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16(6): 276-277.

[51] Rothschild LJ, Mancinelli RL. 2001. Life in extreme environments. Nature. 409(6823):1092-1101.

[52] Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. Science. 239(4839):487-491.

[53] Salzberg SA, Delcher AL, Kasif S, White O. 1997. Microbial gene identification using interpolated Markov models. Nuc. Acids Res. 26(2):544-548.

[54] Sanger F, Nicklen S, Coulsen AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74(12):5463-5467.

[55] Tatusov RL, Koonin EV, Lipman DJ. 1997. A Genomic Perspective on Protein Families. Science. 278(5338):631-637.

[56] Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. BMC Bioinformatics. 4:41.

[57] Wägele JW, Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol. Biol. 7:147.

[58] Waters E, Hohn MJ, Ahel I, et al. (22 co-authors). 2003. The genome of *Nanoarcheum equitans*: Insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad Sci. USA. 100(22):12984-12988.

[59] Woese CR, Magrum LJ, Fox GE. 1978. Archaebacteria. J. Mol. Evol. 11:245-252.

[60] Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA. 87:4575-4579.

[61] Závodszky P, Kardos J, Svingor A, Petsko GA. 1998. Adjustment of conformational flexibility is a key even in the thermal adaptation of proteins. Proc. Nat. Acad. Sci. U.S.A. 95:7406-7411.

[62] Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007. Protein and DNA Sequence Determinants of Thermophilic Adaptation. PLoS Comp. Biol. 3(1):e5.

[63] Zhang Z, Lopez MF, Torrey JG. 1984. A comparison of cultural characteristics and infectivity of *Frankia* isolates from root nodules of *Casuarina* species. Plant and Soil 78:79-90.