Introduction
One of the major challenges in molecular biology is the complete knowledge of how genes have their levels of expression regulated. In other words how does the sequence of bases in DNA become transformed into the proteins essential for cell structure and function. The genetic code operates in a similar way in both the simplest and most complex organisms. Therefore, information that is useful across the animal kingdom can be determined from simpler organisms. The bacterium *Haloferax volcanii* is a useful model system not only because it is a singled cell organism but as one of the earliest living organisms it is an important source of information for evolutionary genetics.

The Domain Archaea
Until recently all organisms were divided into two categories, single cell or multi-cellular organisms, the prokaryotes and eukaryotes, respectively. However, the analysis of gene sequences from bacteria has revealed a third category of prokaryotes, the Archaea, which have DNA that resembles both the bacteria and the eukaryotes. These microorganisms inhabit harsh environments thought to be similar to the conditions on earth when life first emerged.

The *Archaea* are classified essentially into two phyla, the *Euryarchaeota* and the *Crenarchaeota*. The *Euryarchaeota* are the largest phylum. These include methanogenic, extremely halophilic and some hyperthermophilic bacteria. Similarly, the *Crenarchaeota* includes species that survive extreme environments, some live at high temperature in hot springs and others are able to withstand acidic and sulphurous conditions.

This paper is interested in the extreme halophile *Haloferax volcanii*, a much studied species of the *Archaea*.
**Extremely Halophilic Archaea**

The term extreme halophile is used to indicate organisms with a requirement for a very high salt environment. The halophiles comprise a diverse group of prokaryotes that inhabit highly saline environments such as solar salt evaporation ponds and natural salt lakes in hot, dry areas of the world (Garrity, 1984).

Salt lakes vary in ionic composition with the predominant ions being dependant on the surrounding topography, geology and general climatic conditions (Madigan et al, 2003). In the Great Salt Lake (USA) the predominant cation is sodium and chloride the predominant anion. However, the Dead Sea is relatively low in sodium but contains high levels of magnesium and chloride ions.

A number of extreme halophiles are currently recognized using specific genetic criteria (Madigan et al, 2003). These include the *Halobacterium*, *Halorubrum*, *Halobaculum*, *Haloferax*, *Haloarcula*, *Halococcus*, *Halogeometricum*, *Haloterrigena*, *Natronobacterium* and *Natronomonas*.

The usual archaeal genome contains a single, large, circular chromosome of double-stranded DNA of approximately 4200 kilobases (the term kilobases refers to how many thousands of nitrogenous base pairs form the bridges between the nucleotide strands) (Howland, 1995).

**Control sequences in Archaea**

A gene is a short sequence of base pairs in a chromosome coding for a specific protein. The information in the gene has to be changed into a protein by some mechanism. This process is referred to as gene expression and involves copying the sequence of DNA base pairs into RNA, which acts as an intermediary, then the translation of the copied sequence into a protein molecule. These events have to be managed in some way. The first step is handled by control sequences, which are short sequences of base pairs at specific points in the DNA. They direct the transcription of a gene (DNA) into RNA. RNA code is then translated into a protein.

Control sequences are of two types, promoter regions and termination regions. The promoter region is found at the start of the gene and surrounds the first base pair to be transcribed into RNA. Therefore, control sequences are the starting points of transcription, the first stage of gene expression. Promoters are also important because they are recognized by proteins known as transcription factors which bind to the control sequences in the promoter region in order to activate or repress the transcription of that gene.

The core promoter elements in Archaea are called box A and box B (Soppa, 1999). Box A consists of the sequence 5′-T/CTTAT/AA-3′. This code is read from left to right where the “5′” refers to the upstream end of the DNA sequence and “3′” the downstream end and the capital letters to the nitrogenous bases adenine, cytosine, guanine and thymine. This sequence is located 25bp upstream of the transcription starting point. Box B element contains the sequence 5′-T/CG/A-3′ (Soppa, 1999).

The promoter regions of Archaea are very similar to those of eukaryotes called the TBP and TFIIB (Soppa, 1999).

**Aims and Methods**

The specific aim of this study involved a search of a DNA database of *Haloferax volcanii* to determine the precise positions of the box A and B promoter sequences in the genome.

Two computer programs called Glimmer and AlignACE were applied to the databases. Glimmer and AlignACE are tools used for searching unknown genes and promoter regions (Salzberg et al., 1997; Hughes et al., 2000). Glimmer was used to identify gene sequences in the DNA (Salzberg et al., 1997). It was also used to identify sequences of base pairs 50 bp long upstream of the start points of the identified sequences. These 50 bp sequences contain the promoter regions.
Thereafter AlignACE was used to find the promoter regions by comparing gene sequences with known promoter sequences Box A and B.

**Results**

The first results from Glimmer returned the number of genes included in the genome of *H. volcanii* from the Scranton sequencing project file. It revealed 9154 genes.

After several runs of AlignACE a total of 92 promoter regions were returned. These regions were found in the 50 bp region upstream of the gene start sites. All 92 were examined to find similarities with known regulatory motifs (Box A and B). Three of these regions had closer similarities to the Box A sequence most notably Region 15 (out of 92). This was the most clearly recognized motif. Two other promoter regions 23 and 24 also demonstrated similarity with Box A (See Table 1).

<table>
<thead>
<tr>
<th>Motif No</th>
<th>Motif Logo</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td><img src="image1" alt="Motif 15 Logo" /></td>
</tr>
<tr>
<td>23</td>
<td><img src="image2" alt="Motif 23 Logo" /></td>
</tr>
<tr>
<td>24</td>
<td><img src="image3" alt="Motif 24 Logo" /></td>
</tr>
</tbody>
</table>

**Discussion**

The main criterion used in the study for identifying a promoter region was its similarity with the consensus of known elements Box A and B, and their upstream position suggested from the literature. The promoter sequence found in region 15 was found at 10 to 36 bp upstream of the start site. Motif 23 shows some similarities to Box A. This is the control sequence of the heat-shock genes that exist in the genome of *Haloferax volcanii* (Thomson & Daniels, 1998).

The study determined that computational methods were useful for finding control sequences. Glimmer was the first gene finder for microbial genomes and still holds a place in the top ranking gene-finding programs available (Salzberg *et al.*, 1997). AlignACE and its accessory programs have proven useful for the analysis of bacterial genomes (Hughes *et al.*, 2000). Further computational analysis could reveal other information concerning not only the known but also novel sequence control features within the genome of *Haloferax volcanii*. 
References

Garrity, M. G., 1984, Bergey’s Manual of Systematic Bacteriology, 16, 220-240


Further Reading


Tompa M. et al., 2005, Assessing computational tools for the discovery of transcription factor binding sites, Nat. Biotech, 23, 137-145


Author Profile:

Nicholas is 23 years old, studied in the School of Biosciences at the University of Nottingham graduating in 2007 with an upper second class BSc in Microbiology. Nicholas was particularly interested in microbial diseases and hopes to specialise in this area in the future.