The goal of this project was to elucidate at the molecular level some of the features that
make archaebacteria unique and distinguish them from eubacteria and eucaryotes. Three types of
genes, encoding rRNAs, ribosomal proteins and superoxide dismutase are phylogenetically
conserved in all three kingdoms. The structure, organization, regulation and expression of these
genomes and their gene products has been examined in several archaebacterial species and compared
to their eubacterial and eucaryotic counterparts. The major findings from this work are:
i) The promoter recognition signals in halophilic archaebacteria consist of two sequence elements centered about -30 and -40 nucleotides upstream of the transcription initiation site. The consensus for these sequences are TTAA (-30) and TTCGA (-40). Transcription termination signals are tracts of T residues in the (+) strand of the template DNA and are often preceded by a G+C rich sequence, sometimes possessing inverted repeat symmetry. The transcripts of protein encoding genes are generally monocistronic, initiated close to the ATG translation initiation codon, and noticeably lack the Shine-Dalgarno ribosome binding sequence.

ii) Ribosomal RNA operons are transcribed from one up to nine tandem promoters in the 5' flanking region. The initial step in processing precursor 16S and 23S from the primary transcript involves cleavage by an endonuclease that is also used to excise the intron from certain tRNA gene transcripts. The primary sequence and secondary structure of the cleavage site is highly conserved among archaebacteria.

iii) Ribosomal protein genes are organized into eubacterial-like operons. The L1e-L10e-L12e halophilic operon contains a putative regulatory sequence in the 5' transcribed leader that could function in autogenous translational control.

iv) Archaebacterial ribosomal proteins exhibit substantial structural and sequence similarity to eucaryotic ribosomal proteins and less similarity to the eubacterial equivalent ribosomal proteins.

v) Halophilic archaebacteria contain a Mn containing superoxide dismutase that is homologous to the eubacterial Mn or Fe SOD and unrelated to the eucaryotic CuZn enzyme. The sod gene and a related gene, sgl, have been cloned and sequenced from H. cutirubrum and their regulation and expression has been characterized. These two sequences represent a unique example of divergent evolution driven by selection at the molecular level of a duplicated sequence within a genome.

Abbreviations:

Hcu - Halobacterium cutirubrum
Hha - Halobacterium halobium (same species as Hcu)
Hvo - Halobacterium volcanii
Hma - Halobacterium marismortui
Sso - Sulfolobus solfataricus
Eco - Escherichia coli
Sce - Saccharomyces cerevisiae

Research Objectives:

i) To characterize the principles of gene organization and regulation of gene expression in archaebacteria.

ii) To elucidate the evolutionary relationship between these novel organisms and the traditional eubacterial and eucaryotic organisms.

iii) To understand in biophysical and molecular terms some of the mechanisms that allow archaebacteria to inhabit extreme environments.
Accomplishments:

Ribosomal protein genes: A 5.2 Kbp Clal-BamH1 fragment of Hcu genomic DNA was cloned using synthetic oligonucleotides complementary to the coding regions of partially sequenced Hcu 50S subunit ribosomal proteins. The fragment was completely sequenced and found to contain the genes encoding the proteins equivalent to the Eco L11, L1, L10 and L12 ribosomal proteins. In addition, two open reading frames designated ORF and NAB were also detected. The transcripts from this region were extensively characterized by Northern hybridization, S1 nuclease protection and primer extension analysis. Four promoters were located in front of ORF, NAB, L11e and L1e genes, respectively. The first three initiate transcription at or adjacent to the respective ATG translation initiation codons whereas the fourth initiates to produce a transcript with a 75 nucleotide long untranslated leader sequence. Preceding each transcription initiation site are two motifs that appear to be conserved; they are the general archaebacterial motif TTAA centered at about position -30 and a halophile specific motif TTCA centered at about position -40. Transcription termination occurs at poly T tracts in the DNA (+) strand that are often preceded by GC rich sequences sometimes containing inverted repeat symmetry. Within the coding regions, poly T sequences are noticeably under represented and the TTT Phe codon is not used.

The tricistronic mRNA encodes the L1e, L10e and L12e ribosomal proteins that are produced in a 1:1:4 stoichiometry. Each of these cistrons is preceded by what appears to be the equivalent of the eubacterial Shine-Dalgarno ribosome recognition sequence. Monocistronic mRNAs lack these sequences and may initiate translation by a eucaryotic thread on type mechanism. The L1e binding sequence on 23S rRNA has been defined previously. We identified a region in the 5' untranslated leader to the L1e-L10e-L12e mRNA that resembles in primary sequence and secondary structure this rRNA binding site. We propose that this mRNA leader sequence is used to autogenously regulate translation of the mRNA by a mechanism similar to that employed in eubacteria. Thus halophilic archaebacteria retain the same gene order and possibly also the same regulatory mechanism for controlling ribosomal protein synthesis that is found in eubacteria.

Ribosomal protein structure: The complete amino acid sequences of the L11e, L1e, L10e and L12e ribosomal proteins from Hcu and Sso have been deduced from the respective gene sequences and compared to the amino acid sequences of the respective eucaryotic and eubacterial equivalent ribosomal proteins. On the large ribosomal subunit, the 1:4 complex of L10e and L12e in loose association with L11e binds to 23S rRNA at about nucleotides 1050-1120 to form a prominent stalk structure. This structure functions as a factor binding domain with associated GTPase activities during the protein synthesis cycle. The L1e protein binds to 23S rRNA near nucleotide 2150 and forms a prominent ridge on the large subunit opposite the stalk structure. The L1e protein stabilizes peptidyl tRNA binding to the P site and stimulates GTPase activity at the factor binding domain. Alignments of amino acid sequences of these four proteins from eubacteria, eucaryotes and archaebacteria indicate that (i) archaebacteria represent a distinct phylogenetic group and (ii) in general, the archaebacterial proteins more closely resemble their eucaryotic rather than the eubacterial equivalent proteins. Similarities between the L11e, L1e, L10e and L12e proteins of Eco, Hcu, Sso and Sce are presented in the accompanying table.

Our work has provided some valuable insights into the structure and evolution of the L10e-L12e complex and has the important potential to help us identify functional domains within the respective proteins. Eucaryotes including Sce possess two different L12e genes
designated type I and II; in Sce, each of these is again duplicated to give four L12e genes designated IA, IB, IIA and IIB genes. Within the universally shared factor binding domain of all L12e proteins (from eubacteria, archaebacteria and eucaryotes), the type I versus type II split appears to be the most ancient. This implies that the common ancestor contained two L12e genes and that one of these (probably the type II gene) was lost from the eubacterial and archaebacterial lines. Although the factor binding domain remains a constant feature of all L12e proteins, the eubacterial protein clearly differs in its overall structure from the archaebacterial and eucaryotic equivalents. In the Eco L12, the factor binding domain is shifted to the carboxy terminus of the protein and the amino terminus appears to have originated by a partial duplication event. The archaebacterial-eucaryotic L12e proteins have the factor binding domain at the amino terminus and a unique charged region at the carboxy terminus that is not present in the eubacterial protein. Furthermore, a unique gene fusion placing the distal 75% of the L12e gene at the 3' end of the L10e gene is evident in both archaebacteria and eucaryotes. This domain in the two proteins is highly conserved, implying a strong constraint on sequence divergence. In eubacteria, most of this fusion sequence has been removed from the contemporary L10 gene by subsequent deletion and there is virtually no homology left from the ancient fusion events. These and other detailed analyses of the structure, function and evolution of the domains of the L11e, L1e, L10e and L12e proteins is continuing.

Superoxide dismutase in the archaebacteria: When life originated about $3.5 \times 10^9$ years ago, the earth's atmosphere was anaerobic. Oxygen began to accumulate in the atmosphere about $2 \times 10^9$ years ago as a by-product of the photosynthetic process. Molecular oxygen is toxic to living organisms because of the reactivity of the superoxide anion. The enzyme superoxide dismutase carries out the dismutation of superoxide to molecular oxygen and peroxide. Two separate SOD activities have evolved independently during evolution, the CuZn enzyme of eucaryotes and the Mn or Fe enzyme of eubacteria (and eucaryotic organelles). We have shown that the halophilic SOD is homologous to the eubacterial enzyme. The enzyme activity has been purified to near homogeneity from Hcu and the gene encoding the activity has been cloned, sequenced and its activity has been characterized.

**Table 2. Similarities between the L11e, L1e, L10e, and L12e ribosomal proteins of E. coli, H. cutrubrum, S. solfataricus, and S. cerevisiae**

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Identities</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>L11e</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hcu/Sso</td>
<td>164</td>
<td>65 (40%)</td>
<td>1</td>
</tr>
<tr>
<td>Hcu/Eco</td>
<td>138</td>
<td>46 (33%)</td>
<td>1</td>
</tr>
<tr>
<td>Sso/Eco</td>
<td>139</td>
<td>45 (32%)</td>
<td>2</td>
</tr>
<tr>
<td>L1e</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hcu/Sso</td>
<td>214</td>
<td>66 (31%)</td>
<td>4</td>
</tr>
<tr>
<td>Hcu/Eco</td>
<td>211</td>
<td>58 (27%)</td>
<td>12</td>
</tr>
<tr>
<td>Sso/Eco</td>
<td>220</td>
<td>49 (22%)</td>
<td>10</td>
</tr>
<tr>
<td>L10e</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hcu/Sso</td>
<td>343</td>
<td>90 (26%)</td>
<td>1</td>
</tr>
<tr>
<td>Hcu/Eco</td>
<td>169</td>
<td>40 (24%)</td>
<td>6</td>
</tr>
<tr>
<td>Sso/Eco</td>
<td>169</td>
<td>35 (21%)</td>
<td>6</td>
</tr>
<tr>
<td>Hcu/Sce</td>
<td>329</td>
<td>76 (23%)</td>
<td>7</td>
</tr>
<tr>
<td>Sso/Sce</td>
<td>322</td>
<td>72 (23%)</td>
<td>7</td>
</tr>
<tr>
<td>Eco/Sce</td>
<td>163</td>
<td>27 (17%)</td>
<td>6</td>
</tr>
<tr>
<td>L12e</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hcu/Sso</td>
<td>110</td>
<td>46 (42%)</td>
<td>1</td>
</tr>
<tr>
<td>L12e</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Globular Domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hcu/Sso</td>
<td>56</td>
<td>28 (50%)</td>
<td>0</td>
</tr>
<tr>
<td>Hcu/Eco</td>
<td>44</td>
<td>13 (30%)</td>
<td>7</td>
</tr>
<tr>
<td>Sso/Eco</td>
<td>44</td>
<td>12 (27%)</td>
<td>7</td>
</tr>
<tr>
<td>Hcu/Sce IA</td>
<td>57</td>
<td>16 (28%)</td>
<td>2</td>
</tr>
<tr>
<td>/Sce IB</td>
<td>57</td>
<td>18 (32%)</td>
<td>2</td>
</tr>
<tr>
<td>/Sce IIA</td>
<td>54</td>
<td>9 (17%)</td>
<td>6</td>
</tr>
<tr>
<td>/Sce IIB</td>
<td>54</td>
<td>16 (30%)</td>
<td>6</td>
</tr>
<tr>
<td>Sso/Sce IA</td>
<td>57</td>
<td>18 (32%)</td>
<td>2</td>
</tr>
<tr>
<td>/Sce IB</td>
<td>57</td>
<td>17 (30%)</td>
<td>2</td>
</tr>
<tr>
<td>/Sce IIA</td>
<td>54</td>
<td>12 (22%)</td>
<td>6</td>
</tr>
<tr>
<td>/Sce IIB</td>
<td>54</td>
<td>17 (31%)</td>
<td>6</td>
</tr>
<tr>
<td>Sce IA/Sce IB</td>
<td>58</td>
<td>32 (53%)</td>
<td>2</td>
</tr>
<tr>
<td>/Sce IIA</td>
<td>55</td>
<td>8 (15%)</td>
<td>6</td>
</tr>
<tr>
<td>/Sce IIB</td>
<td>55</td>
<td>14 (25%)</td>
<td>6</td>
</tr>
<tr>
<td>Sce IB/Sce IIA</td>
<td>54</td>
<td>9 (17%)</td>
<td>6</td>
</tr>
<tr>
<td>/Sce IIB</td>
<td>54</td>
<td>11 (20%)</td>
<td>6</td>
</tr>
<tr>
<td>Sce IIA/Sce IIB</td>
<td>52</td>
<td>27 (52%)</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: The average lengths in amino acids over the region of comparison of the two sequences is determined as the length of the comparison region minus half of the total gaps. Identities are given as the number (and percentage) of perfect matches over the region of comparison. Gaps are given as the number of deletions (or insertions) required to achieve alignment of the two sequences within the region of comparison. The L12e globular domain is the region indicated in Fig. 4B in Hcu and Sso L12 this region extends from position 1 to 360 of Fig. 4A. Comparisons with Eco L12 are with the C domain and not the N-terminus in Fig. 4B.
In addition to the sod gene which encodes the authentic SOD activity that has been purified, we cloned a second gene designated slg from the Hcu genome which is highly similar to sod. This slg gene is actively transcribed; no sod-like activity corresponding to the product of the slg gene has been detected; the transcriptional regulation of slg is different from sod and the 5' and 3' flanking regions are unrelated. Within the coding region, the four codons used to specify the amino acid residues used to bind Mn in the protein are conserved in both genes. The genes have 87% nucleotide sequence identity whereas the proteins they encode have only 83% amino acid sequence identity. Mutations occur randomly at first, second and third codon positions and transversions outnumber transitions. Most mutational differences between slg and sod are confined to two limited regions; other regions totally lack differences. These two gene sequences are apparently in the initial stage of an unusual mode of divergent evolution. Presumably, this divergence is being driven by strong selection at the molecular level for either acquisition of new or partition and refinement of ancestral functions in one or both of the respective gene products.

An attempt to clarify the intriguing sod-slg relationship in Hcu was made by examining the homologous sequences from the related halophile Hvo. This organism also has two copies of a sod related sequence; amazingly both copies are nearly identical, differing only at codons two and three as follows:

<table>
<thead>
<tr>
<th>codon</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>199</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>amino acid</td>
<td>M</td>
<td>S</td>
<td>D</td>
<td>Y</td>
<td>F</td>
<td>E</td>
</tr>
<tr>
<td>sod 1</td>
<td>...</td>
<td>AACACCTAC ATG TCA GAC TAC ...</td>
<td>TTC GAC TAA CGCGTACGC ...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>****</td>
<td>****</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sod 2</td>
<td>...</td>
<td>GTTACACATT ATG AGC --- TAC ...</td>
<td>TTC GAG TAA CCGGATCAT ...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>S</td>
<td>-</td>
<td>V</td>
<td>F</td>
<td>E</td>
</tr>
</tbody>
</table>

The remaining 197 codons are identical and homology abruptly ends at the cistron boundaries. We suspect that the duplication of the primordial sod gene was an ancient event and that in Hcu selection at the molecular level is operating to produce two proteins with different (although probably related) function whereas in Hvo selection is operating to conserve structure and function. We are beginning to exploit the Hvo DNA transformation system to address questions relating to the function, regulation and evolution of the sod genes in halophilic archaeabacteria halophilic rRNA operons.

Ribosomal RNA genes: In halophilic archaeabacteria, rRNA operons have the following gene order: 16S, 23S and 5S. An ala tRNA gene is located in the 16S-25S intergenic space and a cys tRNA is sometimes positioned distal to the 5S gene. The 5' flanking region contains from one up to nine tandemly repeated promoter sequences each containing the TTAG (-30) and TTGAG (-40) conserved sequences. The 16S and 23S genes are flanked by inverted repeat processing sequences that are highly conserved in both primary sequence and secondary structure. Excision at these sites liberates precursor 16S and 23S rRNAs from the primary transcript; it appears likely that the endonuclease responsible is identical to enzymes used to excise introns from the 23S rRNA genes.

During a comparison of leader sequences from Hcu, Hvo and Hha, we have observed an 80 nucleotide long region that immediately precedes the 16S processing site that is more highly conserved than is the 16S gene sequence between these organisms. Only the last few nucleotides of this sequence are required for recognition by the endonuclease. The function of the remainder of this conserved sequence is being investigated.
The species Hma has two rRNA operons, designated HC8 and HH10. The first is preceded by four tandem promoters whereas the second is preceded by only a single promoter. In addition, the HH10 operon appears to lack the normal 16S rRNA processing sequence. We have begun a complete sequence comparison of these two operons. Remarkably, our results indicate extensive differences between the two 16S coding regions. These include base substitutions as well as insertion-deletion. In virtually all organisms, the sequence of multiple rRNA genes are virtually identical. The functional significance of this sequence variation in the two Hha operons is currently under investigation.

Publications:


**Training:**

Bruce May: MRC, predoctoral fellow, Canadian
Lawrence Shimmin: ONR, predoctoral student, Canadian
Craig Newton: ONR technician, Canadian
Diedre de Jong-Wong: ONR technician, Canadian
Phalgun Joshi: ONR predoctoral student (grant), Canadian
Peter Durovic: MRC predoctoral student (grant), Canadian
Janet Yee: NSERC predoctoral student, Canadian

STATEMENT "A" per Michael Marron
ONR/Code 1141
TELECON 4/16/90 VG
Distribution List for Annual and Final Reports

1. Put a cover page (Form DD 1473) on your report and attach a copy of the distribution list. Mail one copy of the report to each person on the contractor subset list attached on which your name appears. The other subset list is for your information only. Please don't forget to attach this distribution list to your report - otherwise the folks below think they have mistakenly received the copy meant for the Molecular Biology Program and forward it to us.

2. Mail two copies to (include a DTIC Form 50 with these two copies too)
   Administrator
   Defense Technical Information Center
   Building 5, Cameron Station
   Alexandria, VA 22314

3. Mail one copy to each of the following:
   (a) Dr. Michael Marron
       ONR Code 1141
       Molecular Biology Program
       800 N. Quincy Street
       Arlington, VA 22217-5000
   (b) Administrative Contracting Officer
       ONR Resident Representative
       (address varies - see copy of your grant)
   (c) Director,
       Applied Research Directorate
       ONR Code 12
       800 N. Quincy Street
       Arlington, VA 22217-5000
   (d) Director
       Office of Naval Technology
       Code 22
       800 N. Quincy Street
       Arlington, VA 22217-5000
   (e) Director
       Chemical and Biological Sci Div
       Army Research Office
       P. O. Box 12211
       Research Triangle Park, NC 27709
   (f) Life Sciences Directorate
       Air Force Office of Scientific Research
       Bolling Air Force Base
       Washington, DC 20332
   (g) Director
       Naval Research Laboratory
       Technical Information Div, Code 2627
       Washington, DC 20375
NEALSON, Kenneth H.
Center for Great Lakes Studies
University of Wisconsin-Milwaukee
600 E. Greenfield Avenue
Milwaukee, WI 53204

OLSEN, Gary J.
Indiana University
Department of Biology
Jordan Hall 138
Bloomington, Indiana 47405

PACE, Norman R.
Department of Biology
Indiana University
Bloomington, IN 47405

PREZELIN, Barbara B.
Marine Science Institute
University of California
Santa Barbara, CA 93106

REEVE, John N.
Department of Microbiology
Ohio State University
484 West 12th Avenue
Columbus, OH 43210-1292

ROSEMAN, Saul
Department of Biology
Johns Hopkins University
Baltimore, MD 21218

SEARCY, Dennis G.
Zoology Department
University of Massachusetts
Amherst, MA 01003

SILVERMAN, Michael
Agouron Institute
505 Coast Blvd. South
La Jolla, CA 92037

SMIT, John
Department of Microbiology
University of British Columbia
#300 - 6174 University Blvd
Vancouver, British Columbia
V6T 1W5 CANADA

SPUDICH, John L.
Dept of Anat and Structural Biology
Albert Einstein College of Medicine
1300 Morris Park Avenue
Bronx, NY 10461

STAHL, David A.
College of Veterinary Medicine
University of Illinois
Urbana, IL 61801

SWIFT, Hewson
Dept of Molec Genetics and Cell Biology
University of Chicago
1103 East 57th Street
Chicago, IL 60637

TAYLOR, Gordon T.
Hawaii Institute of Geophysics
University of Hawaii
2525 Correa Road
Honolulu, HI 96822

TOSTESON, Thomas R.
Department of Marine Sciences
University of Puerto Rico
Mayaguez, PR 00709

TRENCH, Robert K.
Marine Science Institute
University of California-Santa Barbara
Santa Barbara, CA 93106

WALEH, Nahid
Molecular Biology Department
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

WHITE, David
Institute of Applied Microbiology
University of Tennessee
P. O. Box X, Building 1503/6
Oak Ridge, TN 37831

WOESE, Carl R.
Genetics Department
University of Illinois
515 Morrill Hall
Urbana, IL 61801

YAYANOS, A. Aristides
Physiological Research Laboratory
Scripps Institution of Oceanography
University of California-San Diego
La Jolla, CA 92093

ZINDER, Stephen H.
Department of Microbiology
Cornell University
Stocking Hall
Ithaca, NY 14853