THE SUPEROXIDE DISMUTASE GENE FAMILY IN THE HALOBACTERIA: STRUCTURE, EXPRESSION and EVOLUTION

by

PHALGUN B. JOSHI

B.Sc. (Hons.), University of Guelph, 1986

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF BIOCHEMISTRY
GENETICS PROGRAM

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

SEPTEMBER, 1992

(Signature)

Department of Biochemistry

The University of British Columbia
Vancouver, Canada

Date March 24th 1993

ABSTRACT

The halophilic archaebacteria belong to a group of closely related organisms that evolved from the methanogens. Since the methanogens are strict anaerobes, the divergence of the aerobic halophiles from this lineage may have required the acquisition of protection against oxygen toxicity; a number of reactions involving oxygen produce highly reactive free-radical by-products. An example of such by-products is the superoxide radical, O_2^-. Most aerobic organisms possess the enzyme superoxide dismutase (SOD) to protect against this free radical.

It has previously been shown that the halophile, *Halobacterium cutirubrum* possesses a type of SOD that contains manganese. The genome of this organism contains two closely related genes; one that encodes the SOD enzyme, designated *sod*, and another sod-like gene designated *slg*. The two genes are 87% identical and their putative proteins are 83% identical. For most genes that are homologous, the DNA sequence identity initially decreases more rapidly than the corresponding amino acid sequence identity of the proteins they encode. The disparity in the identities between the *sod* and *slg* genes and between their corresponding proteins can be attributed to the almost even distribution of substitutions between the three codon positions. This pattern of substitutions results in a high incidence of non-synonymous nucleotide substitutions. The two genes also differ in their response to exposure to paraquat, a generator of superoxide free radicals; mRNA levels from the *sod* gene were seen to be elevated whereas those from *slg* were unaffected.

To investigate whether other halophiles contain similar paralogous genes (i.e., products of gene duplication in an organism) and whether they exhibit similar patterns of evolutionary divergence and gene expression, *sod* genes from three different halophiles were isolated and characterized.

The number of copies of genes homologous to *sod* from *Hb. cutirubrum* varies from one in *Haloarcula marismortui* to two in *Halobacterium* sp. GRB and *Haloferax volcanii.* The pattern of substitutions between the paralogous genes in *Hb.* sp. GRB (*sod* and *slg*) is almost identical to that observed between the *sod* and *slg* genes in *Hb. cutirubrum.* In contrast, the *sod1* and *sod2* genes in *Hf. volcanii* are 99% identical. Comparison of the nucleotide sequences reveals that all the genes are related (identities vary from 76% to 99%) and form a coherent family. Within the entire family, substitutions in the first and second positions are much more frequent than expected; this results in a large number of amino acid substitutions in the encoded protein.

Both *Hb.* sp. GRB and *Hf. volcanii* contain one gene each that is induced by paraquat and one that is unaffected. The single *sod* gene in *Ha. marismortui* is unaffected by paraquat treatment.

Comparison of the protein sequences encoded by the superoxide dismutase gene family in the halobacteria with sequences from other organisms has enabled the identification of halobacterial-specific residues.

Phylogenetic analyses were performed to determine the evolutionary relationship between the members of the *sod* gene family in the halobacteria. Results from these analyses together with the comparison of upstream flanking sequences and response to paraquat have enabled the postulation of possible evolutionary histories of the genes.

TABLE OF CONTENTS

ABSTRACT....................ii

TABLE OF CONTENTS....................v

LIST OF FIGURES....................viii

LIST OF TABLES....................ix

ABBREVIATIONS....................x

ACKNOWLEDGMENTS....................xiii

INTRODUCTION....................1

Early Evolution....................1

Molecular Phylogeny And The Three Kingdom Classification....................2

Molecular Phylogeny Methods....................9

Archaebacteria....................11

Halophilic Archaebacteria....................12

Oxygen Toxicity And Superoxide Dismutases....................15

Superoxide Dismutases In Archaebacteria....................20

Aims Of This Study....................23

MATERIALS AND METHODS....................25

Bacterial strains and growth conditions....................25

Isolation of DNA and RNA....................26

Preparation of radioactive probes....................26

Southern hybridization analysis....................27

Library construction....................27

DNA sequencing....................28

RNA transcript mapping....................28

Enzyme activity assay....................29

Sequence analysis 29

CHAPTER 1. Characterization of Paralogous and Orthologous members of the Superoxide Dismutase Gene Family from Genera of the Halophilic Archaebacteria..31

1.1. INTRODUCTION..31

1.2. LITERATURE CLARIFICATION AND NOMENCLATURE..........34

1.3. RESULTS AND DISCUSSION..36

1.3.1. Gene Isolation..36

1.3.2. Sequence determination and alignment....................40

1.3.3. Transcript characterization and regulation by paraquat..47

1.3.4. Transcription and translation signals...........................54

1.3.5. Regulation by paraquat..57

1.4. SUMMARY..63

CHAPTER 2. The Family of Superoxide Dismutase Proteins from Halophilic Archaebacteria: Structure, function and evolution..............65

2.1. INTRODUCTION..65

2.2. RESULTS AND DISCUSSION..67

2.2.1 Amino acid and nucleotide sequence alignments...67

2.2.2. Signature amino acid residues in SOD.........................70

2.2.3. Nucleotide sequence divergence..................................74

2.2.4. Phylogenetic analysis..83

2.2.5. The consensus tree..87

2.3 Summary..92

CONCLUSIONS 94

FUTURE RESEARCH PROSPECTS 97

REFERENCES 99

LIST OF FIGURES

Figure 1 Universal phylogenetic trees determined from the comparison of SS rRNA sequences. 7

Figure 2 Identification of *sod*-like sequences by Southern hybridization 37

Figure 3 Aligned nucleotide and amino acid sequences of the *sod* gene family and the proteins they encode. 41

Figure 4 Nucleotide sequences of the 5' and 3' flanking regions of the halophilic *sod* genes 45

Figure 5 Nuclease S1 and primer extension mapping of the 5' and 3'transcripts from the *sod* family of genes. 49

Figure 6 Alignment of the 5' flanking regions of the non-inducible *sod* family of genes. 59

Figure 7 Amino acid sequence alignments of SOD proteins 68

Figure 8 Relative rates of nucleotide substitutions in the *sod* and *slg* genes of *Hb. cutirubrum* 81

Figure 9 Phylogenetic relationships between members of the halophilic *sod* gene family or the proteins they encode 84

Figure 10 The consensus tree illustrating phylogenetic relationships between the halophilic *sod* genes (or proteins) 88

LIST OF TABLES

Table 1. Characteristics of the Primary kingdoms....................................5

Table 2. Transcripts from the *sod* family of genes...................................56

Table 3. Signature amino acid residues in eubacterial and halophilic SODs...71

Table 4. Comparison of paralogous and orthologous gas vacuoles genes...76

Table 5. Comparison of halophilic *sod* genes and proteins.....................78

ABBREVIATIONS

A_{600}	absorbance at 600 nanometers
A	adenine (in nucleic acids)
Ala or A	alanine
AMV	avian myeloblastosis virus
Arg or R	arganine
Asn or N	asparagine
Asp or D	aspartic acid
ATP	adenosine triphosphate
bp	base pairs
C	cytosine
Cys or C	cysteine
dATP	deoxyadenosine triphosphate
dCTP	deoxycytosine triphosphate
dGTP	deoxyguanosine triphosphate
DNA	deoxyribonucleic acid
dTTP	deoxythimidine triphosphate
EDTA	ethylenediaminetetraacetic acid
G	guanine (in nucleic acids)
g	grams
Glu or D	glutamic acid
Gln or Q	glutamine
Gly or G	glycine
His or H	histidine
Ile or I	isoleucine
kbp	kilobase pairs

kcal	kilocalories
Leu or L	leucine
Lys or Y	lysine
M	moles per liter
Met or M	methionine
mg	milligram
ml	millilitre
mRNA	messenger ribonucleic acid
mg	microgram
ml	microlitre
ng	nanogram
nt(s)	nucleotide(s)
Phe or F	phenylalanine
Pro or P	proline
RNA	ribonucleic acid
RNAse	ribonuclease
S	Svedberg unit
SDS	sodium dodecyl sulfate
Ser or S	serine
SOD	superoxide dismutase
T	thymine
Thr or T	threonine
Tris	tris(hydroxylmethyl)aminomethane
Trp or W	tryptophan
Tyr or Y	tyrosine
u	units of enzyme
Ura or U	uracil

Val or V	valine

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Patrick P. Dennis for his help and support during the entire course of graduate work. I also thank Drs. Caroline Astell and Rosemary Redfield for their advice and help during the same period of time. A special thanks goes to Lawrence Shimmin for sharing his wisdom, patience, enthusiasm and introducing me to the joys of "B" movies. Thanks also go to my colleagues, Peter, Janet, Damien, Shanthy, Deidre, Luc, Daiqing and Josephine for their cooperation and friendship in the lab. I thank Dr. W. Robert McMaster for his help during the compilation of this thesis.

Finally and most importantly, I thank my parents, my brothers and sisters, and my wife, Alison for their unconditional support and encouragement during the course of this work.

INTRODUCTION

Early Evolution

How life began and how it subsequently evolved into the vast array of extant forms presently inhabiting earth has intrigued scientists for decades. Until recently most clues to the history of cellular evolution originated from a combination of geological and paleontological sources. For example, the oldest fossils of any life form, the stromatolites from Warrawoona in Western Australia, place the origin of life at approximately 3.5 X 10^9 years ago (Walter, 1983). These and other fossils show characteristics similar to depositions (mats) formed by present day cyanobacteria. Large and more complex cells first appear as fossils in depositions dated to about 2.1 X 10^9 years ago. These cells were of sizes comparable to the smallest present day eucaryotes. Eucaryotes are distinct from the procaryotes; they contain organelles and a nuclear membrane that surrounds a complex genome. From the fossil data it had been concluded that evolution was a linear process beginning with simple procaryotic cells and eventually producing a larger and more complex eucaryotic cell.

It was also recognized that the procaryotes contained organisms, now grouped together and called the archaebacteria, that seemed unusual in their ability to survive in extreme environments. Although these organisms showed peculiar structural and physiological characteristics they were not considered different from the bacteria.

The dogma of linear evolution from simple procaryotes to complex eucaryotes prevailed prior to the early 1970s. It was generally regarded that the "demarcation between the eukaryotic and prokaryotic organisms (was) the largest and most profound single discontinuity in the contemporary biological world" (Stanier *et al.*, 1976). In the latter part of the decade this idea was radically altered due to results obtained from molecular phylogenetic analyses by Carl Woese and his colleagues.

Molecular Phylogeny And The Three Kingdom Classification

The study of evolutionary relationships among organisms using macromolecular sequence data is termed molecular phylogeny. In the mid 1960s, Zuckerkandl and Pauling (1965) had suggested the use of macromolecules as molecular chronometers; the comparison of primary structures of proteins and nucleic acids that are derived from a common ancestor and separated by speciation could provide clues to their phylogenetic relatedness. It is now generally accepted that an enormous number of nucleic acid (and polypeptide) sequences produce a nearly identical phenotype. Genetic variability unaccompanied by significant phenotypic change is not subject to selective pressure and therefore is a good measure of evolutionary relatedness. In spite of the fact that the real situation is more complex, macromolecular sequences are being used to estimate phylogenetic relationships in the presence of uncertainty using a probabalistic model of evolution (Felsenstein, 1988).

The evolutionary history of macromolecules or organisms (termed taxonomical units) is often presented graphically in the form of dendrograms or "trees". The branches of the trees denote the evolutionary relationship between the tips of the branches, or external nodes, representing the different taxonomical units. The branch points, or internal nodes, represent the ancestral states.

One of the best molecular chronometers is a constituent of the ribosome - the principal component in the process of translation in which messenger RNAs are decoded into polypeptides. Ribosomes are composed of a large and a small subunit designated according to their sedimentation coefficients (S values): the eucaryotes mostly contain 60S and 40S subunits and the archaebacteria and the eubacteria mostly contain 50S and 30S subunits. Each subunit consists of one or more RNA (rRNA) and a large number of proteins (r-proteins); for example, in the eubacterium *Escherichia coli*, the small subunit (SS) contains 16S rRNA and 21 r-proteins and the large subunit (LS) contains 23S and 5S rRNAs and 31 r-proteins. The SS rRNA is a very useful molecular chronometer because: 1) it is present in all cellular life forms, 2) it contains many independent positions that range from highly conserved to highly variable domains, and 3) it is easily sequenced (Woese, 1987).

The analysis and comparison of SS rRNA sequences resulted in an unexpected finding. The earliest life form, designated the progenote, apparently branched (prior to the appearance of the cyanobacteria 3.5 X 10^9 years ago) into three independent lineages, designated urkingdoms: the eubacteria, the eucaryota, and the archaebacteria (Woese, 1987). Subsequent endosymbiotic association between a primitive eucaryote and the bacteria accounted for the introduction of organelles into the former which gradually evolved into the modern eukaryotic cell (Margulis, 1970). The archaebacteria appear to be distinct although they share some properties with the eubacteria and some with the eucaryotes (Table 1). [It has recently been suggested that the terms eubacteria, archaebacteria and eucaryotes be changed to Bacteria, Archaea and Eucarya respectively to avoid the implication that the eu*bacteria* and the archae*bacteria* are more closely related than either one to the eucaryotes (Woese *et al.*, 1990).]

The universal phylogenetic tree based on SS rRNA sequence comparison is depicted in FIG. 1A. This tree is unrooted because the point that corresponds to the universal ancestor is undetermined.

Further comparison of sequences from other macromolecules such as ribosomal proteins and translational elongation factors suggest that the archaebacteria and the eucaryotes are more closely related to each other than to the eubacteria (Woese, 1990). Figure 1B shows a modified tree illustrating this relationship. The position of the root of the tree was determined by comparing the genes for the translation elongation factors Tu and G, and for the α and β subunits of ATPase. Both are pairs of paralogous genes (i.e. two genes of a common ancestry in an organism that have arisen by a duplication event) that diverged from each other before the emergence of the three organismal lineages from their common ancestor (Iwabe *et al.*, 1989). The root of a tree constructed by comparing the sequences of one member of the paralogous pair can be determined if the other member is used as an outgroup.

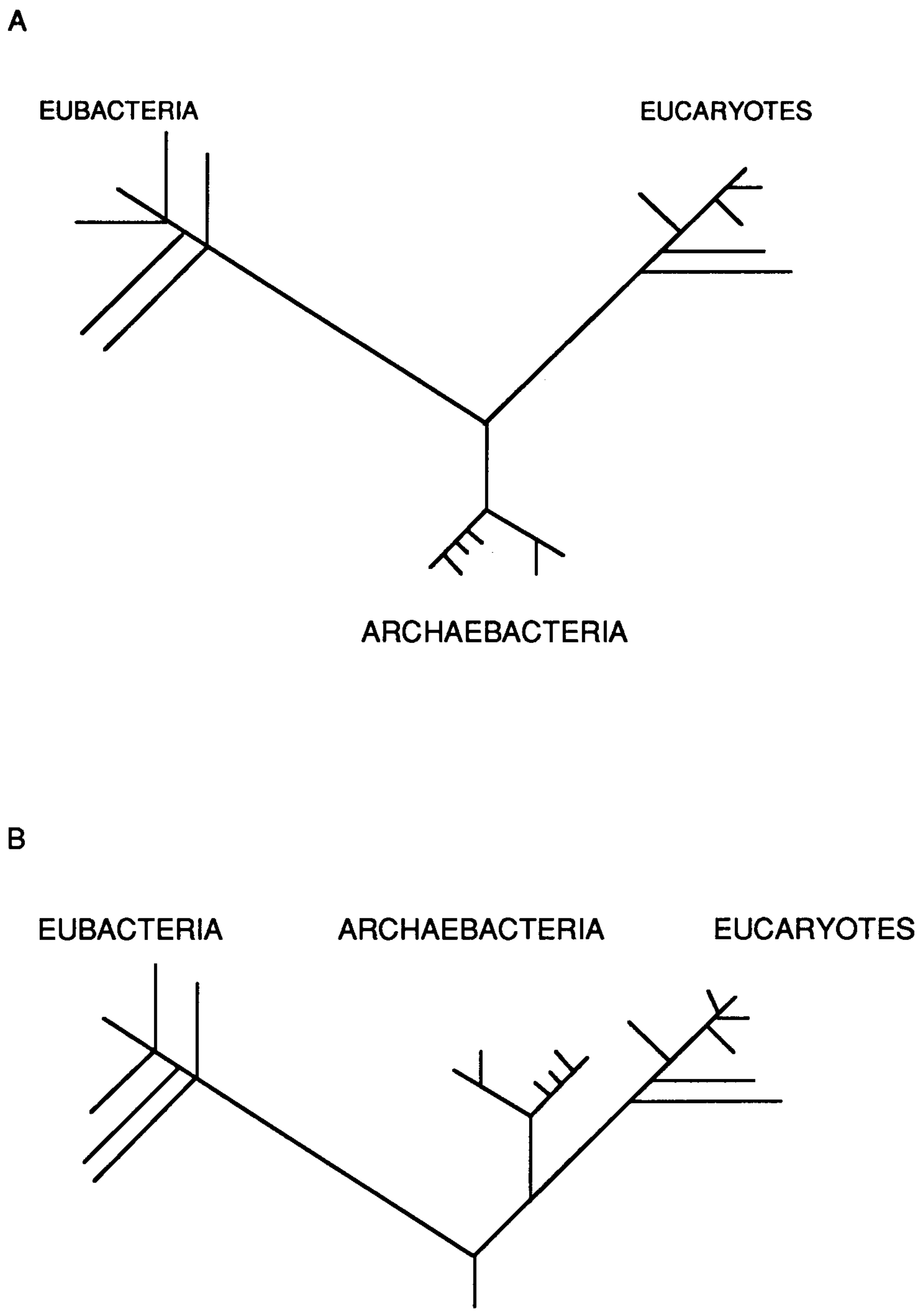
TABLE 1: Characteristics of the Primary kingdoms.

Distinguishing features of the eubacteria, archaebacteria and eucaryota are listed. Abbreviations are chloramphenicol (CM), anisomycin (Ani), Kanamycin (Kan), pseudouracil (y), a-amatin (Ama) and rifampin (Rif).

Characteristic	Eubacteria	Archaebacteria	Eucaryota
Cellular Organization	anucleate	anucleate	nucleated with organelles
Genome Size (bp)	$5x10^5 - 5x10^6$	$5x10^5 - 10^7$	$1.5x10^7 - 3x10^{11}$
Membrane Lipids	ester linked straight chain	ether linked branched chain	ester linked straight chain
Cell Walls	peptidoglycan	various but not peptidoglycan	various or none
Ribosomes			
rRNA	5S, 16S, 23S	5S, 16S, 23S	5S, 5.8S, 18S, 28S
diptheria toxin	insensitive	sensitive	sensitive
antibiotic sensitivity	CM^S Ani^R Kan^S	CM^R Ani^S Kan^R	CM^R Ani^S Kan^R
Transfer RNA			
TψC loop	TψCG	1 - methylψψCG	TψCG
1 - methyl adenine	absent	present	present
initiator tRNA	5' monophosphate	5' triphosphate	5' monophosphate
initiator amino acid	N - formyl methionine	methionine	methionine
RNA Polymerase			
number of types	1	1	3
subunits	5	6 - 13	12 or greater
antibiotic sensitivity	Ama^R Rif^S	Ama^R Rif^R	Ama (Pol II)^S (Pol I+III)^R Rif^R
mRNA	uncapped	uncapped	7 - methyl G cap and polyadenylation

Figure1: Universal phylogenetic trees determined from the comparison of SS rRNA sequences.

(A): An unrooted tree based on a matrix derived from distances between pairs of aligned SS rRNA sequences (Woese, 1987). (B): Tree from (A) above modified to contain a root which corresponds to the universal ancestor from which all extant life forms ultimately diverged. The position of the root was derived by comparing sequences of duplicated paralogous genes (two genes that have resulted from a duplication event in a given organism) for translation elongation factors Tu and G and the genes for α and β subunits of ATPase (Iwabe *et al.*, 1989). These duplicated genes diverged from each other before the three primary lineages evolved from the common ancestor. This rooting strategy utilizes one member of the pair of paralogous genes as an outgroup for the comparison of members of the other genes.


A
EUBACTERIA
EUCARYOTES
ARCHAEBACTERIA
B
EUBACTERIA
ARCHAEBACTERIA
EUCARYOTES

Molecular Phylogeny Methods

The refinement of techniques for obtaining sequences of biological macromolecules during the last 30 years has resulted in an exponential increase in the sequence data set available for phylogenetic analysis. Numerous methods allow the statistical comparison of these sequences; associated with each method are specific variables and assumptions (reviewed by Felsenstein, 1988). No single method can exclusively infer the phylogenetic tree closest to the "true tree" (i.e. the tree that represents the correct phylogenetic history). It is possible, however, that certain evolutionary relationships can be inferred with reasonable certainty if they consistently appear from all or most of the methods of analysis employed.

The most commonly used analyses for inferring phylogenies are parsimony and distance methods. The principle of parsimony methods is minimum evolution: trees that require the smallest number of changes between sequences are assumed to be closest to the true tree and are therefore preferred (Fitch, 1971; Fitch and Ferris, 1974). The method involves first assembling the sequences (nucleic acid or protein) into an alignment such that every position along the length of the alignment contains homologous members. Often, it is necessary to introduce gaps within the sequences in order to optimize the alignment. The gaps represent insertion or deletion events in the sequences that are assumed to have occurred at these sites since their divergence from the common ancestor. Next, "informative sites" on the sequence alignment are determined. A site is considered to be informative only if, taken in isolation, it favours one specific evolutionary pathway (tree) describing the relationship of the members at the site. A site where it is not possible to assign a specific tree as being the most likely is considered to be uninformative and not factored into the analysis. The minimum number of substitutions at each informative site is then calculated and a tree describing these substitutions is assigned to this site. Finally, the incidence of each of the trees over all the informative sites is calculated; the one that occurs most frequently is considered to be the most parsimonious and best representative of the evolutionary relationship of the sequences under consideration (Felsenstein, 1988).

In distance methods, the evolutionary tree is derived from matrices of pairwise differences, usually expressed as the proportion of sites differing between sequences. An example of such a method is the neighbour joining method of Saitou and Nei (1987). This method works as follows: First, the different sequences are aligned as described above. The distances (differences) between all pairs are then computed and a matrix is constructed. Next, all the sequences are placed on a star-shaped dendrogram with no internal nodes (i.e., all taxa are placed at the end of individual branches which all originate from a single point). To determine the first internal node, two most similar sequences are taken from the distance matrix and joined to the rest of the tree to give the shortest total branch length. The two joined sequences are then considered to be "neighbours" and merged into a single unit by averaging the distance between them. The distance of this average unit to each of the other taxa is then computed and a second matrix is compiled treating the first neighbour pair as a single sequence. Repeating the process then allows the assignment of two more sequences as neigbours which are also merged and subsequently considered as a single unit. This procedure is continued until no more internal branching is possible. This analysis results in a tree that infers the overall evolutionary history of all the sequences (or species) that are under consideration.

Since the methods described above are statistical, there are inevitable errors that occur. An example of such errors is the underestimation of distances between sequences due to multiple mutations at single sites. In order to estimate the uncertainty in the overall tree elucidation the data are subjected to *resampling* analysis. The most widely used resampling method in phylogenetic analyses is "bootstrapping" (Felsenstein, 1988). This consists of resampling with replacement; randomly selected samples in the set (correponding to sites in aligned sequences) are analyzed and returned to the set so that some sites of the sequences under comparison are sampled more than once and some are never sampled. If such resampling is performed frequently enough, it is possible to determine the confidence limits (P values) and hence the significance of any inferred evolutionary branch.

The Archaebacteria

The archaebacteria are composed of two major lineages: 1) the extreme thermophiles and 2) the methanogen-halophils (Woese, 1987). Woese *et al.*, (1990) have recently suggested the terms Crenarchaeota and Euryarchaeota to describe the two respective lineages. The extreme thermophiles are quite uniform in their phenotype; most species are anaerobic (an exception is *Sulfolobus,* which is aerobic), utilize sulfur as an energy source and grow optimally at extremely high temperatures (70°-110°C).

Based on SS rRNA sequence comparisons, the methanogens and the halophiles are more closely related to each other than either group is to the sulfur dependent extreme thermophiles (Woese, 1987). The methanogens are obligate anaerobes that possess a unique mode of energy metabolism that generates methane. They can utilize H_2, formate, acetate or methylamines with the aid of coenzymes (e.g. coenzyme M) that are not found in other organisms (Jones *et al.,* 1987). They occupy extremely diverse habitats and have been found wherever anaerobic biodegradation of organic compounds occurs, including freshwater and marine sediments, digestive tracts of animals, and anaerobic waste digesters of sewage treatment plants (Jones *et al.,* 1987).

Halophilic Archaebacteria

Analysis of SS rRNA indicates that the halophilic archaebacteria have emerged from one group of the methanogens (Woese, 1987). It has been proposed that this transition involved adaptations to high salt and oxygen containing environment (Woese, 1987). Halophilic methanogens, which may represent evolutionary intermediates between the methanogenic and halophilic lineages, have recently been described (Paterek and Smith, 1988 and Liu *et al.,* 1990). The combination of halophilicity and aerobiosis is the most significant phenotype that separates the halophiles from all the other archaebacteria. The requirement for salt varies between 2 to 4M NaCl; these organisms are found in salterns and along natural high salt-containing bodies of water including the Dead Sea in Jordan and Lakes Natron and Magadi in East Africa.

Halophiles exhibit variable morphological shapes; they can be pleomorphs (*Haloarcula marismortui*) (Oren *et al.,* 1988), box-shaped (*Halobacterium sp. GN*), rod-shaped *(Halobacteria salinarium* and its relatives including *Hb. cutirubrum* and *Hb. halobium)* or disc-shaped *(Haloferax volcanii)* (Mullakhanbhai and Larsen, 1975). They are often deep red in colour due to the presence of a membrane pigment, bacteriorubrin. The members of the genus *Halobacterium* possess a purple membrane containing bacteriorhodopsin, a light-activated proton pump used to make ATP.

The high salt environment of the halophiles poses a number of problems which are ingeniously solved by the organisms. First, the high ionic exterior places an osmotic pressure on the cell. This is balanced by a corresponding increase in the internal salt concentration; potassium ions are actively taken up raising the intracellular concentration to as high as 5M (Lanyi, 1979).

Salt, especially at high concentrations is destructive to biological macromolecules whose optimal configurations are based on weak, non-covalent interactions. 'Salting-out' compounds such as NaCl and KCl stabilize both the intra- and inter-molecular hydrophobic interactions among non-polar residues. This strengthening of hydrophobic bonds causes proteins to assume a more tightly folded conformation and induces intermolecular aggregation. To reduce such effects, halophiles possess proteins containing a smaller number of non-polar amino acids (Lanyi, 1974).

Another effect of a high intra-cellular salt concentration is the decreased availability of water for the solvation of the proteins. In order to compete effectively for available water molecules, halophilic proteins contain a high proportion of charged residues (Lanyi, 1974; May and Dennis, 1989). The acidic amino acids, aspartate and glutamate, are preferred because they bind at least twice as many water molecules per side chain compared to other amino acids (Saenger, 1987).

Although the halophilic archaebacteria grow in high concentrations of NaCl, the interior of the cells contains a high concentration of KCl. The K^+ ions are actively imported raising the internal KCl concentration to as high as 5M. Presence of K^+ is preferred over Na^+ probably because the hydration number of K^+ is half that of Na^+; consequently, less water is sequestered by the K^+ ions (Lanyi, 1974).

The genomes of the halophilic archaebacteria are extremely complex. The purple membrane-containing halophiles, for example, contain two major fractions. About 80% of the genome is composed of a chromosomal fraction that is 68% G+C and contains mostly single-copy sequences. The rest of the genome is heterogeneous and contains a variety of different covalently-closed circular plasmid DNAs as well as A+T rich islands from the chromosomes (Ebert *et al.*, 1984; Pfeifer and Betlach, 1985). In *Hb. halobium* a large number of repetitive elements, some of which have been shown to be insertion elements, are present in both chromosomal and extra-chromosomal fractions of its genome (Sapienza and Doolittle, 1982). These repetitive elements are the source of numerous rearrangements making the genome highly unstable.

The genera *Halobacteria* contain purple membranes and are physiologically and biochemically the best studied of the halophilic archaebacteria. However, due to their unstable genomes they have not been as amenable to genetic manipulations. Instead *Haloferax volcanii*, which possesses a more stable genome, has been exploited and an efficient transformation system has been developed (Cline *et al.*, 1989). For genetic manipulation, a number of *E. coli* - *Hf. volcanii* shuttle vectors have also been constructed (Lam and Doolittle, 1989; Holmes and Dyall-Smith, 1990 and Holmes *et al.*, 1991).

Because of the uniformity in 16S rRNA sequence divergence, it has been difficult to reconstruct the phylogenetic history within the halophilic archaebacteria. The data suggest rapid radiation to form the major genera within a short period of time (Mylvaganam and Dennis, 1992).

Oxygen Toxicity And Superoxide Dismutases

The appearance of the cyanobacteria in the earth's biosphere 3.5 to 4.0 X 10^9 years ago set into motion a chain of events that radically changed the existing anaerobic atmosphere. The cyanobacteria were the earliest appearing organisms capable of utilizing solar energy to derive reducing equivalents from water. This process, photosynthesis, generated molecular oxygen (O_2) as a byproduct. Initially, oxygen was sequestered by the ferrous "sink" within the earth's mantle. About 2.0 X 10^9 years ago the oxidation of ferrous to ferric compounds was complete and O_2 release into the atmosphere began (Chapman and Schopf, 1983). The continual release of O_2 has changed the earth's atmosphere from a reducing anaerobic to an oxidizing aerobic composition.

The presence of O_2 in the biosphere enabled other organisms to evolve a mode of respiration that utilized O_2 as a terminal electron acceptor. This mode is energetically more efficient than either fermentation or respiration using nitrate and sulphate as terminal electron acceptors.

Although oxygen provides advantages to organisms able to use it, its presence and use in respiration generates highly reactive free-radical derivatives. These free radicals react randomly, and often catastrophically with biomolecules including nucleic acids, proteins and lipids (as reviewed by Cadenas, 1989). Why these radicals are generated becomes obvious upon examination of the atomic structure of molecular oxygen.

Dioxygen, O_2, contains two unpaired electrons in its outer molecular orbital. When O_2 oxidizes another atom or molecule it accepts two electrons of the same spin. Since most molecules contain a pair of electrons of anti-parallel spin in their outer orbitals, the reduction of O_2 is restricted to one electron at a time. The primary intermediate in the univalent reduction of O_2 is the superoxide radical O_2^-:

$$O_2 + e^- \longrightarrow O_2^-$$

Further reduction of the superoxide radical then gives rise to hydrogen peroxide (H_2O_2):

$$O_2^- + 2H^+ + e^- \longrightarrow H_2O_2$$

The combined presence of the superoxide radical and hydrogen peroxide, through catalysis of a metal group, gives rise to the Haber-Weiss reaction and results in one of the most reactive and destructive of all free radicals, the hydroxyl radical (OH•) (Haber and Weiss, 1934):

$$O_2^- + H_2O_2 + H+ \longrightarrow OH\cdot + H_2O + O_2$$

This is, however, not the full explanation of oxygen toxicity since superoxide has been shown to be involved in damage independently of the presence of H_2O_2 (Fridovich, 1986a).

In nature there are numerous enzymatic and non-enzymatic sources of superoxide. The enzymatic sources include reactions catalyzed by xanthine oxidase, galactose oxidase and several flavin dehydrogenases (Fridovich, 1978). Non-enzymatic sources include the auto-oxidation of thiols and hydroquinones. The photolysis of molecular oxygen by solar radiation can also yield superoxide. As a result, even organisms not utilizing oxygen are potentially at risk.

Most organisms protect themselves from the destructive effects of oxygen radicals. Several modes of protection exist: 1) the presence of enzymes such as cytochrome oxidase which contain paramagnetic transition metals that enable the multivalent reduction of O_2 without releasing either O_2^- or H_2O_2 (Antonini *et al.*, 1970), 2) the absence of thiol and other groups on proteins that are especially sensitive to oxidation and 3) the presence of protective enzymes capable of converting reactive free radicals to less reactive species.

Superoxide dismutase (SOD) is one of the most important of these protective enzymes (Fridovich, 1986b). It contains a prosthetic metal atom which catalyses the transfer of an electron from one superoxide anion to another, producing molecular oxygen and hydrogen peroxide. The latter is then converted to water through the action of catalases or peroxidases.

$$2O_2^- + 2H+ \xrightarrow{\text{SOD}} H_2O_2 + O_2$$

$$H_2O_2 + H_2O_2 \xrightarrow{\text{catalase}} 2H_2O + O_2$$

$$\text{or} \quad H_2O_2 + RH_2 \xrightarrow{\text{peroxidase}} 2H_2O + R$$

where RH_2 is a reductant other than H_2O_2 (e.g. formate).

Depending on the type of prosthetic metal group, SODs can be divided into two classes: 1) the Cu/Zn type (containing both Cu and Zn ions) found in the cytosol of eucaryoteic cells and in a small number of eubacterial organisms and 2) the Fe or Mn type (containing only one or both of Fe and Mn ions) found in most eubacteria, archaebacteria and the organelles of eucaryotic cells

(Puget and Michelson, 1974; Steinman, 1982 and Bannister and Parker, 1985). The bovine erythrocyte Cu/Zn SOD was the first of these enzymes to be purified (McCord and Fridovich, 1969). It is a dimer of subunits of about 16,000 daltons. In contrast, the Fe- and Mn-SODs enzymes exist as homo-dimers or tetramers of about 20,000 daltons (Steinman, 1982). They show identity in both their primary sequence and tertiary structure (Stallings *et al.,* 1984 and 1985) and are distinct from the Cu/Zn type in their tertiary structure and in their sensitivity to inhibitors (Steinman, 1982). These dissimilarities suggest that the Cu/Zn and Mn and Fe enzymes arose by convergent evolution from two separate origins.

The importance of superoxide dismutase activity to organisms is supported by: 1) the separate origins for the Cu/Zn and the Fe and Mn enzymes, 2) the ubiquitous occurrence of SODs in aerobic organisms (Imlay and Fridovich, 1991), 3) the high degree of conservation of primary and secondary structures within each enzyme class throughout nature, indicative of a crucial function (Steinman, 1982), 4) the induction of Mn SOD by exposure of *E.coli* to increased levels of O_2 or superoxide generators such as paraquat, and 5) the increased sensitivity to oxygen radicals by mutants of *E.coli* and *Drosophila* that lack SOD (Carlioz and Touati, 1986; Phillips *et al.,* 1989).

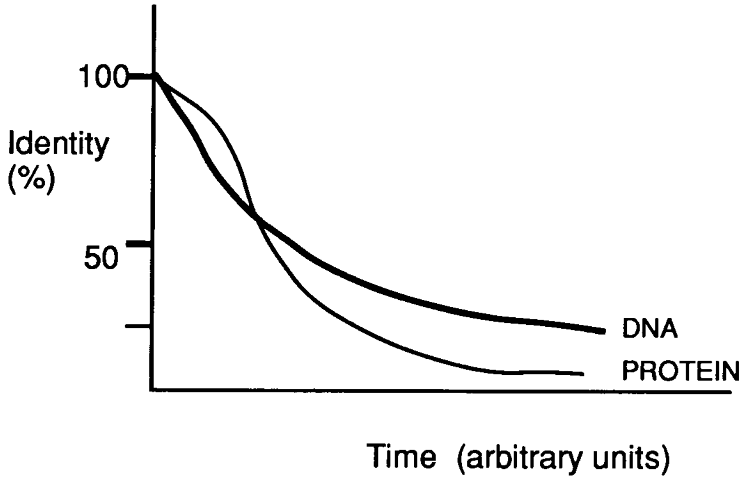
Superoxide Dismutases In Archaebacteria

All archaebacteria investigated to date appear to possess the Mn- or Fe-containing SOD enzyme. Methanogens and the thermophile *Thermoplasma* possess the Fe-containing SOD (Kirby *et al.*, 1981; Takao *et al.*, 1991; Searcy and Searcy, 1981). The Mn-type of the enzyme occurs in the halophiles *Hb. halobium* and *Hb. cutirubrum* (Salin and Oesterhelt, 1988; May and Dennis, 1987).

Genes encoding SOD from a methanogen and a halophile have been cloned and sequenced (Takao *et al.*, 1990; May and Dennis, 1989). These genes show high sequence identity to each other and to genes encoding Fe and Mn-SODs in other organisms (33%-42%) (Takao *et al.*, 1990).

In *Hb. cutirubrum*, the *sod* gene encodes a protein with superoxide dismutase activity. This protein has been purified to near homogeneity and characterized with respect to its metal content (Mn), size (25,000 daltons), subunit composition (tetramer) optimal salt requirement, susceptibility to specific inhibitors (resistant to azide and sensitive to inactivation by H_2O_2) and inducibility by paraquat (May and Dennis, 1989). In addition to *sod*, a second homologous gene designated *slg* (*sod* like gene) was detected within the genome of *Hb. cutirubrum* (May and Dennis, 1990). The *slg* gene is actively transcribed but its protein product has not been identified. The expression of the *sod* gene is enhanced by paraquat whereas the the *slg* gene is unresponsive.

At the nucleotide level *sod* and *slg* exhibit 87% identity whereas the proteins they encode exhibit 83% amino acid identity. This observation is unusual. Normally when duplicated genes begin to diverge, most of the nucleotide substitutions occur at the third codon position and are of the synonomous type (i.e. no amino acid replacement occurs at the position specified by the codon). Such changes are more easily fixed in the population because they are often not subject to significant negative selection. Non-synonomous mutations are usually subject to stronger negative selection and are fixed far less frequently in the population. As a consequence, nucleotide identity between the two sequences degenerates more rapidly than the amino acid identity between the two proteins. Only much later when a substantial number of first and second position changes accumulate, does the amino acid identity fall below the nucleotide identity. This relationship between the nucleotide and amino acid identities has been shown from the analysis of sequences of duplicated genes and their protein products and can be represented graphically (R. F. Doolittle, personal communication):



The point at which the nucleotide and amino acids identities converge (the crossover point) represents the stage in the divergence of duplicated genes when non-synonymous substitutions begin to outnumber synonymous ones. From the analysis of sequences of numerous duplicated genes and the proteins they encode, this crossover has been shown to be around 60% (R. F. Doolittle, personal communication). Beyond this point, amino acid identity changes faster than the corresponding nucleotide identity. Some examples of this relationship are 1) the *trpG* genes of *Escherichia coli* and *Salmonella typhimurium* (DNA and protein identities of 83.0% and 95.0%, respectively; Ochman and Wilson, 1987) 2) the *amy2* and *amy3* genes of *Drosophila* (DNA and protein identities of 98.6% and 99.6%, respectively; Boer and Hickey, 1986) and the amylase genes of *Drosophila* and mouse (DNA and protein identities of 57.0% and 55.4%, respectively; Boer and Hickey, 1986).

The *sod* and *slg* genes of *Hb. cutirubrum* are unusual in that their crossover point is above 83% and the amino acid identity is lower than the nucleotide identity. This is the result of the almost equal distribution of substitutions between the three codon positions. Most of these substitutions are non-synononomous and result in amino acid replacements in the protein. The effect of this is a higher degree of difference in amino acid identity between the respective proteins compared to the nucleotide differences between their DNA. It was also found that transversions outnumber transitions; transversions in the third codon position (compared to transitions), are more likely to be non-synonymous changes.

In summary, *sod* and *slg* are paralogous genes (duplicated genes within a species) that are 87% identical at the nucleotide level. The 5' and 3' flanking regions are totally devoid of any sequence similarity. The expression of the *sod* gene is enhanced by oxygen radicals whereas the *slg* gene is unresponsive. The two proteins are only 83% identical at the amino-acid level. These observations suggested that the divergence of *sod* and *slg* was being driven by strong selection for diverse function (May and Dennis, 1990)

Aims Of This Study

The genome of the extreme halophile, *Hb. cutirubrum* has been shown to possess two paralogous genes, *sod* and *slg* (May and Dennis, 1990). These genes are differentially regulated and show an unusual pattern of divergence.

The aims of the present study were:

1) To determine if other species of halophilic archaebacteria also possess similar paralogous genes and if they exhibit similar patterns of gene expression and divergence as those seen in the *Hb. cutirubrum sod* and *slg* genes.

2) To determine the phylogenetic relationship of these orthologous (genes seperated by a speciation event) and paralogous members of the superoxide dismutase gene family both within and between genera of the halophilic archaebacteria.

3) To determine the phylogenetic relationship between the superoxide dismutases from the halophilic archaebacteria and from non-halophilic organisms.

These studies provide useful insight into the mechanisms of molecular evolution in the halobacteria and important clues to the evolution of one form of protection against oxygen toxicity, the presence of the enzyme superoxide dismutase.

MATERIALS AND METHODS

Bacterial strains and growth conditions

All halobacterial strains were grown in enriched high salt media that were adjusted to pH 6.8-7.0 and sterilized by autoclaving. The *Hb. cutirubrum* and *Hb. GRB* were grown in medium which contained per liter 231 g NaCl, 24.6 g $MgSO_4 \cdot 7H_2O$, 2.2 g KCl, 1.35 g sodium citrate, 3 g yeast extract and 5 g tryptone (Bayley, 1971). For *Hf. volcanii*, the medium contained per liter 125 g NaCl, 45 g $MgCl_2 \cdot 7H_2O$, 10 g $MgSO_4 \cdot 7H_2O$, 10 g KCl, 1.34 g $CaCl_2 \cdot 2H_2O$, 3 g yeast extract and 5 g tryptone (Daniels et al., 1986). For *Ha. marismortui*, the medium contained per liter 206 g NaCl, 36 g $MgSO_4 \cdot 7H_2O$, 0.37 g KCl, 0.5 g $CaCl_2 \cdot 2H_2O$, 0.013 mg $MnCl_2$ and 5 g yeast extract (Oren et al., 1988). All cultures were grown at 37-40°C in rotary incubators.

Paraquat was added to early log phase cultures (A_{600}=0.2) to a final concentration between 1 and 2.5 mM. Incubation was continued for a further 24 h before cells were harvested for RNA preparation. In the presence of paraquat, the growth rate is substantially reduced and during the 24 h incubation, the cultures did not reach stationary phase.

The *Escherichia coli* strains DH5α and JM101 were used for plasmid propagation and generation of single stranded phagemids, respectively. The helper phage, R408, was used in conjunction with pGEM (Promega) and pEMBL (Dente and Cortese, 1987) phagemids. The *E. coli* strain KW251 was used as host for propagating the *Ha. marismortui* genomic library constructed in λGEM-11 (Promega). All strains were grown in YT medium and, when required, the antibiotics ampicillin and kanamycin were added to final concentrations of 100 and 50 μg per ml, respectively. For plate assays of β-galactosidase, IPTG (isopropyl-thio-galactoside) and X-gal (5-bromo-4-chloro-3-indolyl-βD-galactosidase) were added to 50 μM and 0.005% (w/v), respectively. To induce the receptor for phage λ, strain KW251 was grown in medium supplemented with 0.2% maltose and 10 mM $MgSO_4$.

Isolation of DNA and RNA

Total DNA was isolated from stationary phase cultures of halobacteria as described by Schnabel et al. (1982). Following equilibrium centrifugation in EtBr-CsCl density gradients, the DNAs were extracted with isopropanol, dialyzed against TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0) and stored at -20°C. Plasmid DNA was isolated by the alkaline lysis method (Maniatis et al., 1982). Total cellular RNA was isolated using the boiling SDS-lysis method described by Chant and Dennis (1986). Following ethanol precipitation, RNA was resuspended in TE buffer and stored at -20°C.

Preparation of radioactive probes

Restriction fragments were dephosphorylated with calf intestinal alkaline phosphatase as described by Maniatis et al. (1982). Fragments and oligonucleotides were 5' end-labelled with T4 polynucleotide kinase and γ-^{32}P-ATP. Restriction fragments containing 3' recessed ends were 3' end-labeled using the appropriate α-^{32}P-dNTP and the Klenow fragment of DNA polymerase I. Uniform labeling of restriction fragments was achieved using the random primer method (Freinberg and Volgenstein, 1982).

Southern hybridization analysis

Genomic, cosmid, plasmid or λDNA was digested with restriction enzymes, separated on agarose gels and blotted onto nylon membrane (Southern, 1975). The filters were prehybridized at an appropriate temperature in 5X Denhardt's solution (1 X Denhardt's is 0.2% BSA, 0.2% Ficoll, 0.2% polyvinyl pyrrolidone), 5 X SSPE (1 X SSPE is 0.18M NaCl, 10 mM Na_3PO_4, 1 mM EDTA, pH 7.7) and 0.1% SDS for one hour prior to the addition of radioactive probe. Hybridization was carried out for 6-12 h. Filters were washed for 15 min, twice at 20°C in 2 x SSPE, 0.1% SDS and once each at the hybridization temperature in 1 x SSPE, 0.1% SDS, and 0.1 x SSPE, 0.1% SDS. The filters were then subjected to autoradiography.

Library construction

Genomic DNA from *Ha. marismortui* was partially digested with Sau3AI to yield fragments primarily in the 12-18 kb range. The *Sau*3AI fragment ends were partially filled with dATP and dGTP and ligated to *Xho*I arms of λGEM-11 which were partially filled with dTTP and dCTP. The recombinant molecules were packaged into phage particles and propagated in the *E. coli* host strain KW251. Plaque lifts on Hybond N filters (Amersham) were screened by hybridization with a radioactive probe as recommended by the supplier.

DNA sequencing

For sequencing, short fragments were subcloned into phagemids and sequenced directly. For larger fragments, phagemid clones containing overlapping unidirectional deletions were generated using Exonuclease III (Henikoff 1987). Single stranded templates were generated using helper phage R408 and were purified as described (Sanger et al., 1977). Double stranded DNA templates were prepared according to Zhang et al. (1988). The sequencing reactions were carried out using either the Klenow fragment of DNA polymerase I or T7 DNA polymerase.

RNA transcript mapping

Nuclease S1 protection was used to identify 5' or 3' mRNA transcript end sites (Dennis, 1985). Briefly, either 5' or 3' end labelled restriction fragment probes (10^5-10^6 dpm) were hybridized to 5-10 μg of total RNA and the hybrids were digested with S1 nuclease. The DNA fragments protected from digestion by RNA were separated on 8% polyacrylamide-8M urea sequencing gels and detected by autoradiography. Appropriate molecular length size standards were used to determine the sizes of protected fragments.

The 5' ends of mRNA transcripts were precisely located by extension analysis using specific 5' end labeled oligonucleotide primers (Neuman, 1987). Molecular length markers were generated using the same 5' end labeled oligonucleotides as primers in DNA sequencing reaction using appropriate template DNA.

Enzyme activity assay

Superoxide dismutase activity was assayed as previously described (Markland 1977; May and Dennis 1987). Protein was assayed by the method of Lowry as modified by Peterson (1977).

Sequence analysis

The relative test (Sarich and Wilson, 1973) was used to estimate the number of substitutions that have occurred in the *sod* and the *slg* genes of *Hb. cutirubrum* since their divergence from a common ancestral sequence. This is accomplished by comparison to an outgroup sequence. In this instance, the *sod*1 sequence of *Hf. volcanii* and the *sod* sequence for *Ha. marismortui* were used in two separate tests. The test involves solution of simultaneous equations:

$$AO = (AB + AC - BC)/2$$
$$BO = (AB + BC - AC)/2$$
$$CO = (AC + BC - AB)/2$$

where A, B and C represent the two related and the outgroup gene sequences and AB, AC and BC are the number of nucleotide differences in the pairwise comparisons of the three sequences. The calculated values of AO and BO estimate the number of substitutions that have occurred in sequence A and sequence B since their divergence from the common ancestral sequence O. The value of CO represents the number of substitutions separating sequence O from sequence C. These relationships are depicted graphically in Figure 8.

Parsimony analysis was carried out using the PAUP analysis package devised by David Swafford (Illinois Natural History Survey, Champaign, IL, U.S.A.). Neighbour joining analysis was carried out using the application from the Clustal V analysis package by Des Higgins (European Molecular Biology Laboratory, Heidelberg, Germany). The actual operation of the software was performed by Dan Fieldhouse at York University on a Sun Sparc Workstation. For bootstrap resampling (Felsenstein, 1988) 100 and 1000 repetitions, respectively, were carried out in the PAUP and Neighbour joining analysis.

CHAPTER 1

Characterization of paralogous and orthologous members of the superoxide dismutase gene family from genera of the halophilic archaebacteria

1.1. INTRODUCTION

Evolution is driven by a plethora of molecular processes. For example, genome expansion occurs by duplication of genes or sequences; random mutational processes introduce variability into these sequences which ultimately either lead to fixation and diversification or to elimination. Two homologous genes are paralogous if they are derived from a duplication event and orthologous if they are derived from a speciation event. Homologous sequences (both orthologous and paralogous) can also participate in recombination and/or gene conversion events which have the effect of maintaining homogeneity and minimizing apparent divergence. From a mechanistic point of view, charting the evolutionary diversification of paralogous and orthologous members of a gene family within a closely related group of organisms should be an informative process to visualize. For this purpose, the superoxide dismutase (SOD) family of genes from four species representing three genera of halophilic archaebacteria have been characterized and compared.

Halophilic archaebacteria are a group of aerobic or microaerobic organisms that evolved from a strictly anaerobic and non-halophilic methanogen ancestor (Woese, 1987). The adaptation to high salt environments was achieved by raising the intracellular salt concentration to near saturation and altering macromolecular structure to function at high salinity (Lanyi, 1979). (Halophilic species of eubacteria are unrelated and fundamentally different from halophilic archaebacteria; they utilize active pumping mechanisms to maintain a low intracellular salt concentration and, because of this fundamental difference, are not relevant to the present study.) The additional adaptation of halophilic archaebacteria to an aerobic metabolism almost certainly has enhanced the importance of protective enzymes such as superoxide dismutase which are used to dissipate highly reactive oxygen containing radicals. The SOD enzyme dissipates the highly reactive superoxide radical (O_2^-) by catalyzing the dismutation to peroxide (O_2^{-2}) and diatomic oxygen (O_2) (McCord and Fridovich, 1969). In eubacteria, the response to oxidative stress is a complex and highly regulated process that only recently has begun to be analyzed (Demple and Amabile-Cuevas, 1991; Fee, 1991).

The genus *Halobacterium* represented by *Hb. cutirubrum* and *Hb.* sp. GRB are distinguished from other halobacteria by production of transmembrane proton and chloride pumping proteins: bacteriorhodopsin and halorhodopsin. (*Hb. cutirubrum* and its immediate relative *Hb. halobium* are independent isolates of a single species, *Hb. salinarium*; *Hb.* sp. GRB is a separate species; Ebert, 1984). The other two organisms examined in this study are *Haloferax volcanii* (Mullakhanbhai and Larsen, 1975) and *Haloarcula marismortui* (Oren et al., 1988).

The genome of *Hb. cutirubrum* contains two paralogous genes designated *sod* and *slg* (superoxide dismutase like gene; May and

Dennis, 1990). The *sod* gene encodes a protein with SOD activity; this protein has been purified to near homogeneity and has been well characterized (May and Dennis, 1987). The *slg* gene is actively transcribed but its protein product has not been identified. The regulation and the pattern of nucleotide sequence differences between the two genes are remarkable. The level of both *sod* gene mRNA and SOD protein is elevated in the presence of paraquat, a generator of superoxide anions, whereas the *slg* gene mRNA is not affected. At the nucleotide level, the two genes exhibit 87% sequence identity whereas the proteins they encode exhibit only 83% amino acid identity. The distribution of mutations is nearly even between first, second, and third codon positions and the majority of nucleotide substitutions cause amino acid replacement in the proteins. Transversions outnumber transitions. The 5' and 3' flanking regions of the two genes exhibit no sequence similarity. It is presumed that oxygen toxicity has played an important selective role in the divergence between these two paralogous genes and the orthologous *sod* genes of related halophilic species.

In this study, genes of the superoxide dismutase family from three related halophilic genera have been cloned and sequenced. For each gene, the putative transcription start and stop sites have been determined and the regulatory response of each gene to paraquat has been characterized. In the accompanying paper, the nucleic acid and protein sequences of the superoxide dismutase family are subjected to a detailed phylogenetic analysis.

1.2. LITERATURE CLARIFICATION AND NOMENCLATURE

As indicated above, *Hb. cutirubrum* and *Hb. halobium* are independent isolates of a single species, *Hb. salinarium.* Salin et al. (1988) have published the sequence of a putative "sod" gene from *Hb. halobium.* Examination of this sequence indicates that it corresponds to the *slg* gene of *Hb. cutirubrum* (i) since it is virtually identical in sequence both in the coding and flanking regions and (ii) since the predicted protein it encodes differs at eight of 26 positions from the N terminal sequence of the purified *sod* protein of *Hb. halobium* and *Hb. cutirubrum* (Salin et al., 1988; May and Dennis, 1989). Furthermore, the published nucleotide sequence of this putative "*sod*" gene appears to contain a number of errors. The first is a reading frame shift resulting from a deletion of a G residue at position 271 and an insertion of a C residue at position 368 (all numbering according to Figure 3). The second is a dinucleotide inversion at position 530-531 that generates an ACG Thr codon in place of the AGC serine codon. The predicted amino acid sequence in the region of the frameshift exhibits no detectable similarity to other SOD proteins from either eubacteria or archaebacteria, whereas the corrected sequence exhibits 46% identity to the SOD protein from *Bacillus stearothermophilus* and 91% and 100% identity to the SOD and SLG proteins of *Hb. cutirubrum.* None of the other genes from the halophilic *sod* family exhibits this reading frame shift. The serine codon at position 529-531 overlaps an *Alu*I (AGCT) restriction site in the DNA; the presence of this site has been confirmed by restriction analysis (data not shown). Takao et al. (1989) have also characterized this region in *Hb. halobium.* They too, designate the gene as *sod*, although it shows 100% identity, within both the coding and flanking regions, to the *slg* gene. Therefore, the genes that have been cloned from *Hb. halobium* do not encode the authentic SOD protein that has been purified and extensively characterized (Salin and Oesterhelt, 1988; May and Dennis, 1987). Rather, they encode the related SLG protein. The sequence as published by Salin et al. (1988) almost certainly contains the three errors as indicated above.

1.3. RESULTS AND DISCUSSION

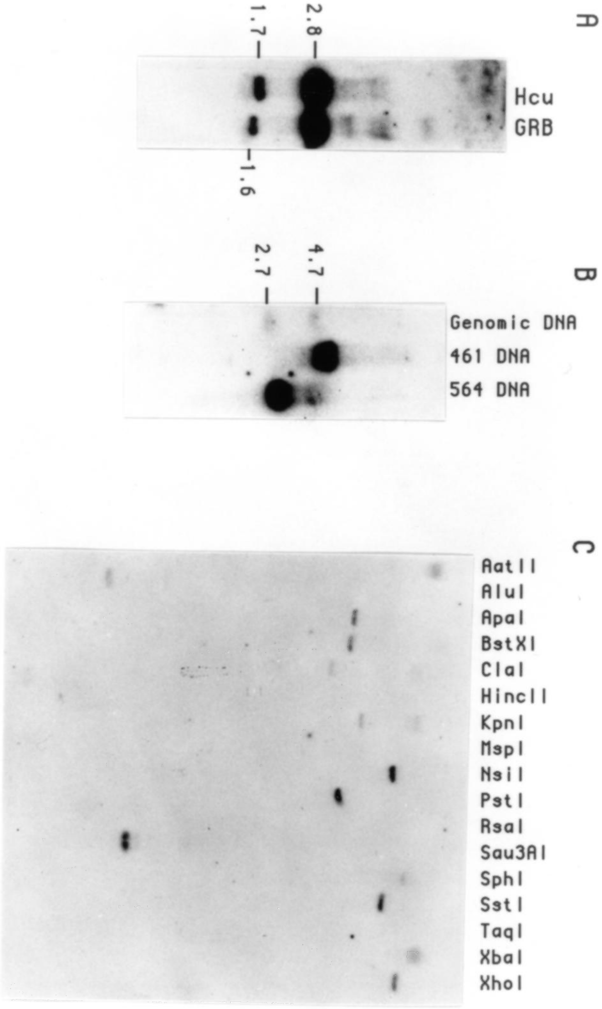
1.3.1. Gene Isolation

The *sod* gene of *Hb. cutirubrum* is 600 nucleotides in length and contained completely within a 1127 bp genomic *Sau*3AI restriction fragment. This fragment is located within a larger 2.8 kbp *Pst*I genomic fragment; the related but unlinked *slg* gene is contained within a 1.7 kbp *Pst*I genomic fragment. The 1.1 kbp *Sau*3AI fragment was used to probe a *Pst*I digest of genomic DNA from *Hb. cutirubrum* and *Hb.* sp. GRB (Figure 2A). The probe exhibited intense hybridization to a single 2.8 kbp fragment in both digests and less intense hybridization to 1.7 and 1.6 kbp fragments in the *Hb. cutirubrum* and *Hb.* sp. GRB digests, respectively. The larger fragments contain the orthologous *sod* genes and the smaller fragments contain the orthologous *slg* genes. A smaller 1.1 kb *Sau*3AI fragment that encoded the entire *sod* gene of *Hb.* sp GRB was contained within the 2.8 kb *Pst*I fragment. The 1.1 kbp *Sau*3AI and 1.6 kbp *Pst*I fragments, containing the *sod* and *slg* genes from *Hb.* sp. GRB, were cloned into the *Bam*HI site of pGEM7zf(+) and the *Pst*I site of pGEM5zf(+) to give pPD 1041 and pPD 1042, respectively.

An ordered library of overlapping cosmids representing more than 95% of the *Hf. volcanii* genome has been constructed (Charlebois et al., 1990). The 1.1 kbp *Sau*3AI fragment from *Hb. cutirubrum* that contains the *sod* gene was used to probe this library. Two non-overlapping cosmids, 564 and 461, that hybridized to the probe fragment were identified. (The latter is a derivative of cosmid B56; Schalkwyk, L., personal communication.) Genomic DNA from

Figure 2: Identification of *sod*-like sequences by Southern hybridization

Restriction enzyme digests of genomic or cosmid DNAs were probed using the 1.1 kbp *Sau*3AI fragment containing the authentic *sod* gene from *Hb. cutirubrum*. A: Genomic DNA was obtained from *Hb. cutirubrum* (Hcu) and *Hb.* sp. GRB (GRB), digested with *Pst*I and probed at intermediate stringency with radioactive probe. B: Genomic DNA from *Hf. volcanii* and from the *Hf. volcanii* cosmid clones 564 and 461 were digested with *Sau*3AI and probed at intermediate stringency. C: Genomic DNA from *Ha. marismortui* (Hma) was digested with a number of different restriction enzymes and probed at intermediate and high stringency. The enzymes used are indicated above each lane.


A
1.7
2.8
1.6
Hcu
GRB
B
2.7
4.7
Genomic DNA
461 DNA
564 DNA
C
AatII
AluI
ApaI
BstXI
ClaI
HincII
KpnI
MspI
NsiI
PstI
RsaI
Sau3AI
SphI
SstI
TaqI
XbaI
XhoI

Hf. volcanii and the DNAs from these two cosmids were digested with *Sau*3AI and reprobed with the 1.1. kbp *sod* fragment from *Hb. cutirubrum* (Fig. 2B). The probe exhibited equally intense hybridization to fragments of 4.7 and 2.7 kbp in length. The larger fragment contained within cosmid 461 mapped to pHV4, a 690 kbp mega plasmid of *Hf. volcanii.* The smaller fragment, contained within cosmid 564, mapped to the main 2.4 mbp chromosome. The 4.7 and 2.7 kbp *Sau*3AI fragments were cloned into the *Bam*HI site of plasmid pUC8 and pEMBL18(+), respectively, to give pPD 1038 and pPD 1039.

Genomic DNA from *Ha. marismortui* was digested with a number of different restriction enzymes. The fragments were separated and probed by Southern hybridization with the *sod* containing 1.1 kbp *Sau*3AI fragment from *Hb. cutirubrum* (Fig. 2C). In all digests, the probe hybridized to only a single restriction fragment; this implies that *Ha. marismortui*, unlike the other halophilic species examined, contains only a single gene belonging to the superoxide dismutase family. A partial *Sau*3AI genomic library of *Ha. marismortui* was constructed in λGEM-11. Three thousand plaques from the library, representing approximately 10 genome equivalents of DNA, were screened by hybridization to the 1.1 kbp *Sau*3AI fragment from *Hb. cutirubrum.* Six positive plaques were identified. The insert DNAs in the six clones were shown to be overlapping and all contain an identical 1.9 kbp *Sau*3AI fragment that hybridized to the *Hb. cutirubrum* probe. When the filter containing the genomic digests (illustrated in Figure 2C) was stripped and reprobed with the 1.9 kbp *Ha. marismortui Sau*3AI fragment at low as well as high stringency, an identical pattern of hybridization was observed. This confirms the existence of only a single gene of the *sod* family in the genome of *Ha. marismortui*. The 1.9 kbp *Sau*3 A fragment was cloned into the *Bam*HI site of the vector pGEM7zf(+) to give plasmid pPD 1040.

1.3.2. Sequence determination and alignment

The nucleotide sequences of the *sod* family of paralogous gene pairs from *Hb.* sp. GRB and *Hf. volcanii* and the single gene from *Ha. marismortui* were determined. The sequences of the coding regions of these genes along with the *Hb. cutirubrum sod* and *slg* sequences are aligned in Figure 3. Only positions that differ from the well characterized *Hb. cutirubrum sod* sequence are indicated; capitals indicate non-synonymous codon changes resulting in amino acid replacement and lower case indicates synonymous codon changes. Five of the genes are 600 nucleotides in length and encode 200 amino acid long proteins. The *sod2* gene of *Hf. volcanii* contains a three nucleotide long deletion which removes codon three and the *Ha. marismortui sod* gene contains a nine nucleotide insertion after codon four. A cursory examination of the aligned gene sequences and the proteins they encode indicates that an unexpectedly large proportion of the nucleotide substitutions are non-synonymous and result in amino acid replacements in the respective proteins. Furthermore, the substitutions are often clustered; this is particularly evident between nucleotide positions 568 and 603. These features are examined in more detail in the next chapter.

Figure 3: Aligned nucleotide and amino acid sequences of the *sod* gene family and the proteins they encode.

The nucleotide sequences of genes of the *sod* family are aligned to the *Hb. cutirubrum* sequence beginning at the ATG translation initiation codon. Species abbreviation are Hcu, *Hb. cutirubrum*; GRB, *Hb.* sp. GRB; Hvo, *Hf. volcanii*; Hma, *Ha. marismortui.* Genes with demonstrable superoxide dismutase are designated *sod* and with an unknown or undefined activity are designated *slg.* The corresponding proteins are SOD and SLG, respectively. Nucleotides identical to the *Hb. cutirubrum sod* sequence are represented by a dash *(-)*; deleted nucleotides are represented by a dot (•). Similarly, amino acids identical to the *Hb. cutirubrum* SOD protein sequence are represented by a dash (-) and deleted amino acids by a dot (•). In the nucleotide alignments (excluding the translation initiation and termination codons), large letters represent non-synonymous substitutions that result in amino acid replacements and small letters represent synonymous substitutions which do not change the amino acid. The position of restriction sites used in preparing probes for S1 nuclease protection experiments are overlined: *Ava*II, 514; *Xma*III, 134; *Msp*I, 188; *Eco*RI, 343 and *Bss*HI, 447

```
                                                                    30                                      60                                      90
GENE SEQUENCES     Hcu sod    ATG TCC GAA TAC ••• ••• ••• GAA CTC CCA CCG CTG CCG TAC GAC TAC GAC GCG CTC GAA CCA CAC ATC AGC GAG CAG GTG CTC ACG TGG
                   GRB sod    ATG --- --- --- ••• ••• ••• --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
                   Hcu slg    ATG AG- C-g C-- ••• ••• ••• --- --- --- T-- --- --- --- --- --- --- --a --- --- --- --- --- --t --- --- --- G-- --- ---
                   GRB slg    ATG AG- C-g C-- ••• ••• ••• --- --- --- T-- --- --- --- --- --- --- --a --- --- --- --- --- --t --- --- -C- G-- --- ---
                   Hvo sod1   ATG --a --C --- ••• ••• ••• --- --- GAc --- --t --- --- --G --- --- --c --- --- --g --- --- TC- --- --- --- --g --- ---
                   Hvo sod2   ATG AG- ••• --- ••• ••• ••• --- --- GAc --- --t --- --- --G --- --- --c --- --- --g --- --- TC- --- --- --- --g --- ---
                   Hma sod    ATG --- --- C-t TCA AAT CCC --- --g --- --- --c --- --- --- --- --- --a --g --g --- --- --- TC- --- --- --- --- --c ---

PROTEIN SEQUENCES  Hcu SOD     M   S   E   Y   •   •   •   E   L   P   P   L   P   Y   D   Y   D   A   L   E   P   H   I   S   E   Q   V   L   T   W
                   GRB SOD     -   -   -   -   •   •   •   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
                   Hcu SLG     -   -   Q   H   •   •   •   -   -   -   S   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   V   -   -
                   GRB SLG     -   -   Q   H   •   •   •   -   -   -   S   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   A   V   -   -
                   Hvo SOD1    -   -   D   -   •   •   •   -   -   D   -   -   -   -   E   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
                   Hvo SOD2    -   -   •   -   •   •   •   -   -   D   -   -   -   -   E   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
                   Hma SOD     -   -   -   H   S   N   P   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -

                                    120                 XmaIII              150                                     180         MspI                210
CAT CAC GAC ACC CAC CAT CAG GGC TAC GTG AAC GGC TGG AAC GAC GCC GAG GAG ACA CTA GCG GAG AAC CGT GAG ACC GGC GAC CAC GCC TCG ACA GCC GGC GCG CTC GGG GAC GTC ACG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --g --g --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--c --- --- --- --- --c --- A-- --- --- G-- --- CTC --- AG- --- --- --- --g --g --- --- --- --- --- --- --- --- --- --t --- --t --- --- --- --- --- --- --- ---
--c --- --- --- --- --c --- A-- --- --- G-- --- CTC --- AG- --- --- --- -Ac --g --- -G- --- --- --- --- --- --- --- --t --- --t --- --- --- --- --- --- --- ---
--- --- --- --- --- --c --- --- --- --c --- --- --- --- -CG -A- --C --- --c --c --c --- --- --c --a G-a --- --G TT- -G- --- T-c --- --- --c G-- C-c A-- --- --c
--- --- --- --- --- --c --- --- --- --c --- --- --- --- -CG -A- --C --- --c --c --c --- --- --c --a G-a --- --G TT- -G- --- T-c --- --- --c G-- C-c A-- --- --c
--c --- --- --- --- --- --- --g --- --a --- --- CTC G-A TCG --- --- --- --g --t --a --- --- --c --C G-- --t --- TT- -Gt --a -Gc --t -Ct --a TGg -Tc A-- --- --c

 H   H   D   T   H   H   Q   G   Y   V   N   G   W   N   D   A   E   E   T   L   A   E   N   R   E   T   G   D   H   A   S   T   A   G   A   L   G   D   V   T
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   S   -   -   D   -   L   -   S   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   S   -   -   D   -   L   -   S   -   -   -   N   -   -   G   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   A   D   D   -   -   -   -   -   -   -   -   A   -   E   F   G   -   S   -   -   -   V   R   N   -   -
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   A   D   D   -   -   -   -   -   -   -   -   A   -   E   F   G   -   S   -   -   -   V   R   N   -   -
 -   -   -   -   -   -   -   -   -   -   -   -   L   E   S   -   -   -   -   -   -   -   -   -   D   A   -   -   F   G   -   S   -   A   -   W   V   N   -   -

                                    240                                     270                                     300                                     330
CAC AAC GGC TCG GGG CAC ATC CTC CAC ACG CTG TTC TGG CAG TCC ATG AGC CCG GCG GGC GGC GAC GAG CCG TCC GGG GCG CTC GCC GAC CGC ATC GCG GCG GAC TTC GGC TCC TAC GAG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -T- --- --- --- --- --- --- ---
--- --- --- -Gt --- --- TA- --- --- --- A-- --- --- G-- CA- --- --t --c -Ac --g --- -G- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- -Gt --- --- TA- --- --- --- A-- --- --- G-- CA- --- --t --c -Ac --g --- -G- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --a --- --t --g --- GAc --- --- --- --- AA- --- TCG --- -Aa --- --- --- --- --c GAg --a CT- --- --a --G --- --- --- -A- --- --- --t --- --- ---
--- --- --- --- --a --- --t --g --- GAc --- --- --- --- AA- --- TCG --- -Aa --- --- --- --- --c GAg --a CT- --- --a --G --- --- --- -A- --- --- --t --- --- ---
--- --- --- -Gt --t --G GAt --- --- --- --- --- --- G-- AA- --- GA- --c AAc --- --- -Gt --- --- GAg --c -Aa --- CT- --- --- --- -A- -A- --- --- --- --- --- ---

 H   N   G   S   G   H   I   L   H   T   L   F   W   Q   S   M   S   P   A   G   G   D   E   P   S   G   A   L   A   D   R   I   A   A   D   F   G   S   Y   E
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   V   -   -   -   -   -   -   -
 -   -   -   C   -   -   Y   -   -   -   M   -   -   E   H   -   -   -   D   -   -   G   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   C   -   -   Y   -   -   -   M   -   -   E   H   -   -   -   D   -   -   G   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   -   -   D   -   -   -   -   N   -   -   -   E   -   -   -   -   -   E   -   L   -   -   E   -   -   -   E   -   -   -   -   -   -
 -   -   -   -   -   -   -   -   -   D   -   -   -   -   N   -   -   -   E   -   -   -   -   -   E   -   L   -   -   E   -   -   -   E   -   -   -   -   -   -
 -   -   -   C   -   Q   D   -   -   -   -   -   -   E   N   -   D   -   N   -   -   G   -   -   E   -   E   -   L   -   -   -   E   E   -   -   -   -   -   -

                EcoRI               360                                     390                                     420                                BssHI
AAC TGG CGG GCC GAG TTC GAG GCC GCC GCC AGC GCG GCC AGC GGC TGG GCG CTG CTC GTC TAC GAC TCC CAC AGC AAC ACG CTC CGG AAC GTG GCC GTG GAC AAC CAC GAC GAG GGC GCG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --t --a --- --- -Tg --g --- G-- --- --- --- --- --- --- --- --- --- --- --t C-g GT- GC- --G CA- --- --- --- --- --- --c --- --- --- --- --- --- ---
--- --- --- --t --a --- --- -Tg --g --- G-- --- --- --- --- --- --- --- --- --- --- --t C-g GT- GC- --A CA- --- --- --- --- --- --c --- --- --- --- --- --- ---
GCG --- AA- -Gt --a --- --a --g --g --- G-- --- --- GCG --- --- --- --t --g --g --- --- --g TT- TCG --- CA- --t --- --- --c -T- --c --- --G --- --- C-- --- ---
GCG --- AA- -Gt --a --- --a --g --g --- G-- --- --- GCG --- --- --- --t --g --g --- --- --g TT- TCG --- CA- --t --- --- --c -T- --c --- --G --- --- C-- --- ---
GG- --- AA- -G- --- --- --a --- --t --a TCA --t --t G-- --- --- --- --- --- --g --- --- C-- TGt GC- --G CAa --g --c --- --- C-g --c --- --G --- --- C-- --- ---

 N   W   R   A   E   F   E   A   A   A   S   A   A   S   G   W   A   L   L   V   Y   D   S   H   S   N   T   L   R   N   V   A   V   D   N   H   D   E   G   A
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   V   -   -   G   -   -   -   -   -   -   -   -   -   -   -   P   V   A   K   Q   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   V   -   -   G   -   -   -   -   -   -   -   -   -   -   -   P   V   A   K   Q   -   -   -   -   -   -   -   -   -   -   -   -   -
 A   -   K   G   -   -   -   -   -   -   G   -   -   A   -   -   -   -   -   -   -   -   -   F   -   -   Q   -   -   -   -   V   -   -   K   -   -   Q   -   -
 A   -   K   G   -   -   -   -   -   -   G   -   -   A   -   -   -   -   -   -   -   -   -   F   -   -   Q   -   -   -   -   V   -   -   K   -   -   Q   -   -
 G   -   K   G   -   -   -   -   -   -   -   -   -   G   -   -   -   -   -   -   -   -   P   C   A   K   Q   -   -   -   -   P   -   -   K   -   -   Q   -   -

                                    480                                     510     AvaII                           540                                     570
CTC TGG GGC AGC CAC CCC ATC CTC GCG CTC GAC GTC TGG GAG CAC TCC TAC TAC TAC GAC TAC GGT CCC GAC CGC GGC AGC TTC GTC GAC GCC TTC TTC GAG GTC GTC GAC TGG GAC GAG
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
--- --- --- --- --- --- --- --- --c --- --- --- --- --- --- --- --- --- --- --- --- --g --- --- --- --- --- --- --- --- --- --- --- --- --g A-- --- --- --- CCC
--- --- --- --- --- --- --- --- --c --- --- --- --- --- --- --- --- --- --- --- --- --g --- --- --- --- --- --- --- --- --- --- --- --- --g A-- --- --- --- CCC
--- --- --t --- --t --- --- --- --- --g --- --- --- --- --- --g --- --- C-- --- --- --c --g -C- --- --- GA- --- --- TC- --g --- --- --- --- --- --- --- --- --a
--- --- --t --- --t --- --- --- --- --g --- --- --- --- --- --g --- --- C-- --- --- --c --g -C- --- --- GA- --- --- TC- --g --- --- --- --- --- --- --- --- --a
--- --- --- TC- --t --- --- --- --a --- --- --- --- --- --- --- --- --- --- --- --- --- --- -C- --t --t GA- --- A-- --- --- --- --- --- --t --- --- --- --- A--

 L   W   G   S   H   P   I   L   A   L   D   V   W   E   H   S   Y   Y   Y   D   Y   G   P   D   R   G   S   F   V   D   A   F   F   E   V   V   D   W   D   E
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   I   -   -   -   P
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   I   -   -   -   P
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   H   -   -   -   -   A   -   -   D   -   -   S   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   H   -   -   -   -   A   -   -   D   -   -   S   -   -   -   -   -   -   -   -   -   -
 -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   A   -   -   D   -   I   -   -   -   -   -   -   -   -   -   -   K

                                    600
CCC ACC GAG CGC TTC GAG CAG GCG GCC GAG CGC TTC GAG TAA     Hcu sod
--- --- --- --- --- --- --- --- --- --- --- --- --- TAA     GRB sod
AT- G-g -C- AA- -A- --C G-C -T- -Tg TC- -Tg --- --- TGA     Hcu slg
AT- G-g -C- AA- -A- --C G-C -T- -Tg TC- -Tg --- --- TGA     GRB slg
--- G-- -CT --- -A- --- --- --c -T- --a -T- --- --- TAA     Hvo sod1
--- G-- -CT --- -A- --- --- --c -T- --a -T- --- --- TAA     Hvo sod2
G-t G-t --a GAg -A- --- A-- T-- -T- AGT -A- --- --- TAA     Hma sod

 P   T   E   R   F   E   Q   A   A   E   R   F   E          Hcu SOD
 -   -   -   -   -   -   -   -   -   -   -   -   -          GRB SOD
 I   A   A   N   Y   D   D   V   V   S   L   -   -          Hcu SLG
 I   A   A   N   Y   D   D   V   V   S   L   -   -          GRB SLG
 -   A   A   -   Y   -   -   -   V   -   L   -   -          Hvo SOD1
 -   A   A   -   Y   -   -   -   V   -   L   -   -          Hvo SOD2
 A   A   -   E   Y   -   K   S   V   S   H   -   -          Hma SOD
```

As expected, the orthologous *sod* gene pair and *slg* gene pair from *Hb. cutirubrum* and *Hb.* sp. GRB are nearly identical; the pairs differed by three (positions 147, 150 and 308) and five (positions 80, 146, 147, 155 and 408) nucleotide substitutions, respectively. This implies that *Hb. cutirubrum* and *Hb.* sp. GRB are indeed closely related and that most of the divergence between the *sod* and the *slg* genes of *Hb. cutirubrum* predated the species separation of *Hb. cutirubrum* from *Hb.* sp. GRB. The paralogous *sod1* and *sod2* genes of *Hf. volcanii* are also nearly identical in sequence. They differ only by the deletion of codon three from the *sod2* gene and by three nucleotide substitutions in codon 2 that converts one serine codon (TCA, *sod1*) to the unrelated serine codon (AGC, *sod2*).

The general pattern of serine codon utilization in the seven aligned sequences is noteworthy. Of the 20 positions where serine is represented two or more times, nine of these use both the TCN and AGY serine codons (e.g., see nucleotide positions 181-3 and 184-6). This implies that convergent evolution may have occurred at some or all of these positions (Brenner, 1988). At position 184-6, the TCN and AGY serine codons could have been derived by separate single nucleotide substitutions in an ancestral ACN threonine codon. The phylogenetic origin of the two serine codons at other positions is obscure. As an alternative to convergence, the substitutions at two adjacent nucleotide positions within the serine codon could be explained by mutagenic repair of a single UV induced cyclobutane pyrimidine dimer (Hsia et al, 1989). That is, AGY↔ TCY interconversion is possible in a single mutagenic event. Of the eleven remaining positions using only a single serine codon type, four are where there are only two serines at the alignment position and they occur in genes with high sequence similarity (e.g., see nucleotide positions 112-4 and 538-40 where the paired serines occur in the orthologous *slg* genes of *Hb. cutirubrum* and *Hb.* sp. GRB and the paralogous *sod1* and *sod2* genes of *Hf. volcanii,* respectively).

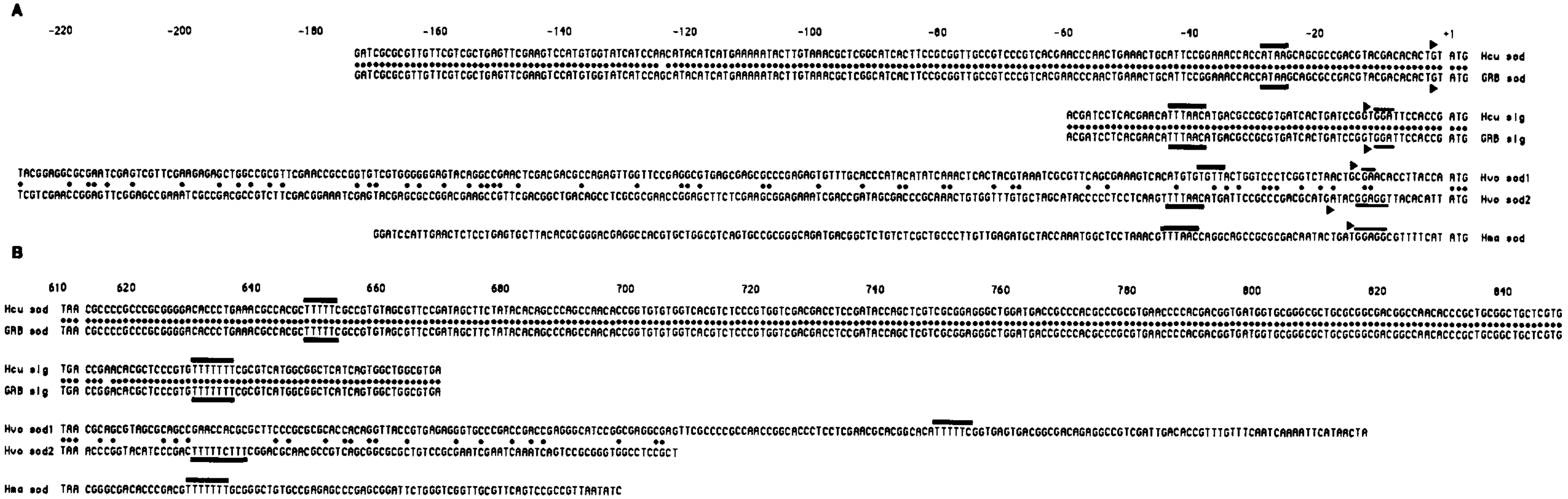
The 5' and 3' sequences flanking the members of the *sod* gene family are illustrated in Fig. 4. The 5' flanking sequences are aligned at the ATG translation initiation codons and numbered as a continuation of the coding sequence in a negative direction. The 3' flanking sequences are aligned at the termination codons, TAA or TGA, and numbered as a continuation of the coding sequence in a positive direction. It has been previously noticed that the sequences flanking the paralogous *sod* and *slg* genes of *Hb. cutirubrum* are unrelated (May and Dennis, 1990). This is also true of the paralogous *sod* and *slg* genes of *Hb.* sp. GRB, as well as the virtually identical *sod1* and *sod2* genes of *Hf. volcanii.* In contrast, the flanking regions of the orthologous *sod* gene pair and the *slg* gene pair of *Hb. cutirubrum* and *Hb. sp. GRB* are virtually identical (Fig. 4). This supports the presumption that the initial duplication to produce the progenitor *sod* and *slg* genes in the common ancestor was an ancient event and that the species separation of *Hb. cutirubrum* and *Hb.* sp. GRB was much more recent.

In the total of 527 nucleotides of 5' and 3' flanking sequence available for the two orthologous *sod* and *slg* gene pairs of *Hb.*

Figure 4: Nucleotide sequences of the 5' and 3' flanking regions of the halophilic *sod* genes.

Above, (A), the 5' flanking sequences of the *sod* family of genes are aligned beginning at the ATG translation initiation codon (position +1) and are numbered in a negative direction. The sequences are grouped according to the similarity of their coding regions. Identical nucleotides in the pairwise alignments of the paralogous genes are indicated by dots (•). The heavy under- and overlines indicate box A-like sequences and the arrowheads (▶) indicate prominent transcription initiation sites. Where present, complementarity to the 3' end of 16S rRNA (...AUCACCUCCU$_{OH}$) is indicated by a light under- or overline.

Below, (B), the 3' flanking sequences are aligned beginning at the respective translation termination codons (positions 610-612). Again, identities in the pairwise alignments are indicated by dots (•). The T tract sequences located at or near the site of transcription termination are under- or overlined.


A
-220
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
+1
ATG
Hcu sod
GRB sod
Hcu slg
GRB slg
Hvo sod1
Hvo sod2
Hma sod
B
610
620
640
660
680
700
720
740
760
780
800
820
840
TAA
TGA

cutirubrum and *Hb.* sp. GRB, there are only two nucleotide differences; one is in the 5' flanking region of the *sod* gene pair at position -125 and the other is in the 3' flanking region of the *slg* gene pair at position 616 (Fig. 4). Within the 1200 nucleotides of coding sequence for the same two gene pairs, there are eight nucleotide substitutions (Figure 3, positions 80, 146, 147, 147, 150, 155, 308, 408). From this somewhat limited data set it is apparent that the *sod* and *slg* coding sequences are still accumulating nucleotide differences; and the rate of accumulation in the coding regions is double that observed for the non-coding, flanking regions. At least half of the substitutions (4 of 7; position 147 is ambiguous) within the coding regions are non-synonymous and have resulted in amino acid replacements.

1.3.3. Transcript characterization and regulation by paraquat

The transcripts derived from the *sod* and *slg* genes of *Hb. cutirubrum* have been characterized by primer extension and S1 nuclease protection analysis and 7-methyl-G capping (May and Dennis, 1990). The *sod* transcript is initiated two nucleotides in front of the ATG translation initiation codon whereas the *slg* gene transcript is initiated 13 nucleotides in front of the initiation codon (Fig. 4, positions -2 and -13, respectively). The transcripts terminate in the poly T sequences that are centred 38 and 21 nucleotides beyond the respective termination codons of the *sod* and *slg* genes at positions 650 and 633, respectively (Fig. 4). Not surprisingly, the 5' and 3' end sites of the *Hb.* sp. GRB *sod* and *slg* gene transcripts were located at the same respective positions, and the expression of the *sod* gene, but not the *slg* gene, was enhanced by the addition of paraquat to the culture (Fig. 5A).

The location of the 5' and 3' end sites of the transcripts derived from the *sod*1 and *sod*2 genes of *Hf. volcanii* was more difficult to determine because of the perfect identity of the two sequences beyond position 9 of the coding region. To locate the 5' transcript end sites, a 355 bp *Msp*I fragment derived from the *sod1* region and a partially homologous 302 bp *Msp*I fragment derived from the *sod2* region were used; both fragments overlap the respective initiation codons and terminate at the common *Msp*I site at nucleotide position 188 within the coding regions. The two fragments were 5' end labelled on the (-) DNA strand at position 190 and hybridized to total *Hf. volcanii* RNA. Following digestion with S1 nuclease, the 355 bp *sod1* probe yielded protected products of about 196 and 172 nucleotides in length (Fig. 5B). The longer product corresponds to protection by the homologous *sod1* gene transcript and has a 5' end site near nucleotide position -15 in front of the ATG translation initiation codon. The shorter product corresponds to protection by the heterologous *sod2* gene transcript and has a 5' end near nucleotide position 9; this is the boundary of sequence divergence between the *sod1* and *sod2* genes. The 302 bp probe from the *sod2* gene yielded protection products of about 197 and 172 nucleotides in length. Again, the longer products result from protection of the probe by the homologous *sod2* transcript with a 5' end site at or near position -19 in front of the ATG translation initiation.codon. The

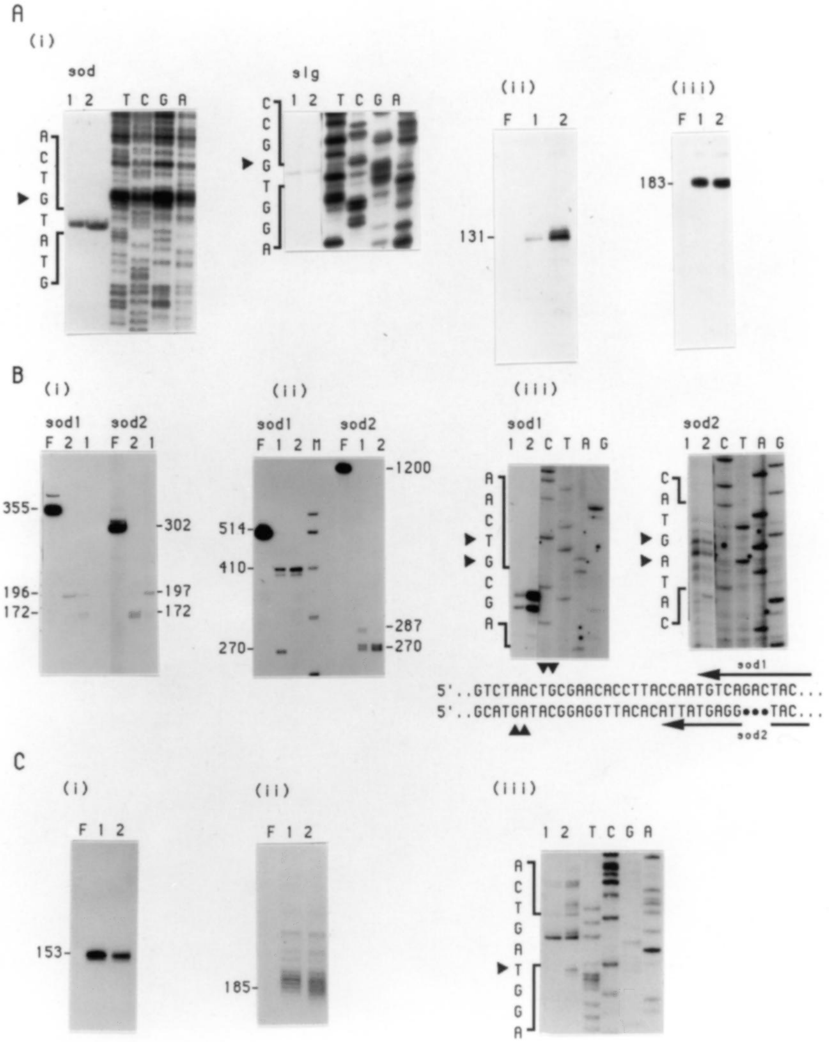
Figure 5: Nuclease S1 and primer extension mapping of the 5' and 3' transcripts from the *sod* family of genes.

Lane abbreviations are: F, end labeled S1 nuclease protection probe; 1, S1 or primer extension protection products generated using exponential RNA; 2, S1 or primer extension products generated using RNA from paraquat treated cells; M, molecular length markers; G, A, T and C, the products of the four DNA chain termination sequencing reactions.

A: Total RNA was prepared from exponential and paraquat treated cultures of *Hb*. sp. GRB. (i) For extension analysis primers complementary to position 236-219 of the *sod* gene and 140-121 of the *slg* gene were used. (ii) For 3' end mapping of the *sod* gene transcript, a 462 bp *Ava*II-*Eco*RI fragment was 3' end labelled at position 517 in the (-) DNA strand and hybridized to RNA from exponential and paraquat treated cultures. Because of their size, the labelled probes do not appear within the autoradiogram window.

B: Total cellular RNA was prepared from exponential and paraquat treated *Hf. volcanii* cultures. (i) For 5' mapping, partially identical 355 and 302 bp *Msp*I fragments from the *sod1* and *sod2* genes were 5' end labelled in the (-) strand at position 190 and used as probes in S1 nuclease protection assays. (ii) For 3' end mapping, partially identical 514 bp *Eco*RI and 1.2 kbp *Eco*RI-*Bam*HI fragments from the *sod1* and *sod2* genes were 3' end labelled in the (-) strand at position 346-347 and used as probes in S1 nuclease protection assays. (iii) For primer extension, the *sod1* and *sod2* specific primers were used to generate reverse transcription products and to generate the corresponding DNA sequence ladders. The sequence of the *sod1* and *sod2* genes overlapping the ATG translation initiation codons is presented under the autoradiogram and the regions of primer complementarity and positions of extension product termination is illustrated.

C: Total RNA was prepared from exponential and paraquat treated *Ha. marismortui* cultures. (i) For 5' end mapping, plasmid pPD 1040 was digested with *Bam*HI, 5' end labelled in the (-) strand at position 138 and used as probe for S1 nuclease protection assays. (ii) For 3' end mapping, a *Bss*HI-*Bam*HI fragment from plasmid pPD 1040 was 3' end labeled in the (-) strand at position 450 and used as probe in S1 nuclease protection assays. This fragment contains the 3' terminal portion of the insert and the entire vector of plasmid pPD 1040. Because of their size, the 5' and 3' end labeled probes do not appear within the window of the autoradiogram. (iii) For primer extension, a *Ha. marismortui* specific oligonucleotide primer complementary to the (+) strand DNA at position 106 to 88 was used to generate extension products and a DNA sequence ladder.


A
(i)
sod
1 2 T C G A
A C T G T A T G
slg
1 2 T C G A
C C G G T G G A
(ii)
F 1 2
131-
(iii)
F 1 2
183-
B
(i)
sod1
F 2 1
sod2
F 2 1
355-
-302
196-
172-
-197
-172
(ii)
sod1
F 1 2 M
sod2
F 1 2
-1200
514-
410-
-287
270-
-270
(iii)
sod1
1 2 C T A G
A A C T G C G A
sod2
1 2 C T A G
C A T G A T A C
sod1
5'..GTCTAACTGCGAACACCTTACCAATGTCAGACTAC...
5'..GCATGATACGGAGGTTACACATTATGAGG•••TAC...
sod2
C
(i)
F 1 2
153-
(ii)
F 1 2
185-
(iii)
1 2 T C G A
A C T G A T G G A

shorter product results from protection by the heterologous *sod1* gene transcript and has a 5' end site near position 9 within the coding region. Finally, treatment of the bacterial culture with paraquat results in an increase in the amount of *sod1* gene transcript relative to the amount of *sod2* gene transcript. For this regulatory feature, the *sod1* and *sod2* genes of *Hf. volcanii* are related respectively to the *sod* and *slg* genes of *Hb. cutirubrum*.

To confirm the above results, primer extension assays were carried out using separate primers capable of discriminating between the *sod1* and *sod2* gene transcripts (Fig. 5Biii). The *sod1* specific 18 mer oligonucleotide primer was complementary to the (+) strand of the *sod1* gene between nucleotides 29 and 2 (numbering according to Fig. 3); extension products terminate at nucleotide position -15 and -16. The *sod2* specific 18 mer oligonucleotide was complementary to the (+) strand of the *sod2* gene between nucleotides 28 and -3; the major extension products with this primer terminate at nucleotide positions -19 and -20. Again, paraquat treatment serves to enhance the amount of the *sod1* gene transcript as assayed by the abundance of extension product.

The 5' end of the transcript from the single *sod* gene in *Ha. marismortui* was located using plasmid pPD1040 linearized at the *Xma*III site and 5' end labeled in the (-) DNA strand at position 138 of the coding sequence (Fig. 5C). The fragment protected from S1 nuclease digestion was about 153 nucleotides in length and corresponds to a 5' transcript end site at or near position -15 in the 5' flanking sequence. This position was confirmed by primer extension using an 18 mer oligonucleotide complementary to position

106 to 89 within the coding region. Surprisingly, paraquat treatment did not seem to enhance the amount of mRNA; this implies that the regulation of the *Ha. marismortui sod* gene is similar to the *sod2* gene of *Hf. volcanii* and the *slg* genes of *Hb. cutirubrum* and *Hb.* sp. GRB. The response of SOD enzyme activity to paraquat was also measured in *Ha. marismortui.* The activity was 3.0 units per mg of protein in the absence of paraquat and 3.2 units per mg of protein after 24 hrs of paraquat treatment. Thus, in *Ha. marismortui*, paraquat is apparently not an inducer of either *sod* mRNA or SOD enzyme activity.

The 3' end sites of transcripts from the *sod* family of genes from *Hb.* sp. GRB, *Hf. volcanii* and *Ha. marismortui* were located by S1 nuclease protection of 3' end labeled restriction fragment probes originating within the coding sequences and terminating in the 3' flanking regions. In all cases, 3' transcript end sites were located to near or within T tract sequences ranging in length from 5 to 7 nucleotides. For most of the genes, these sequences are within 40 nucleotides of the translation termination codon; for the *sod*1 gene of *Hf. volcanii*, the T tract is located at position 749, about 140 nucleotides from the translation termination codon (see Fig. 4). Examination of the coding regions indicates that there are no tracts of three or more T residues in the (+) strand sequence. This absence may be a reflection of the role T tracts plays in the process of transcription termination.

The results of the S1 nuclease protection and primer extension analysis indicate that all seven members of the *sod* gene family in halophilic archaebacteria that have been analyzed are transcribed almost exclusively as monocistronic mRNAs. These mRNAs range in length from about 630 to 750 nucleotides. This result was independently confirmed by northern hybridization using gene specific restriction fragments as probes (data not shown). It has previously been reported that the photolyase gene of *Hb. halobium* (and *Hb. cutirubrum*) is positioned immediately upstream of the *slg* gene and that photolyase transcripts extend into the *slg* gene to produce a long bicistronic transcript (Takao et al., 1989; May and Dennis, 1990). Less than 5 percent of the *slg* gene transcripts are present in this larger species. A sequence related to the 3' end of the photolyase gene was also detected in front of the *slg* gene of *Hb.* sp. GRB but it has not been examined in detail. A related sequence has not been found in the region upstream of any of the *sod* genes of *Hf. volcanii* or *Ha. marismortui.*

1.3.4. Transcription and translation signals

Most archaebacterial promoters contain a moderately conserved hexanucleotide box A sequence (consensus TTTAWA) centred about 25 nucleotides in front of the transcription initiation site (Reiter et al., 1990). In addition, some promoters, including the rRNA promoters of halophilic archaebacteria, contain a box B (consensus AYGCGAA) that overlaps the start site (Dennis, 1985). The promoters of the halophilic *sod* genes seem to retain the box A like sequence but apparently lack the box B like sequence (see Fig. 4).

The precise mechanism for identifying the AUG translation initiation codon in the mRNA of halophilic archaebacteria has not yet been analyzed in detail. The transcripts from the *sod* family of genes are all predominantly or exclusively monocistronic and all contain relatively short 5' untranslated leader sequences of 2-19 nucleotides (Table 2). In all seven transcripts, the first AUG is utilized for translation initiation. This is a feature that is characteristic of most eucaryotic mRNAs (Kozak, 1978).

Translation initiation in eubacteria is facilitated by the presence of a purine-rich sequence centred about ten nucleotides in front of the translation initiation codon. This sequence forms a complementary structure with the highly conserved pyrimidine rich sequence at the 3' end of 16S rRNA (Shine and Dalgarno, 1974). In these organisms, this 16S sequence is 5'...AUCACCUCCU$_{OH}$. The leaders of the *Hf. volcanii sod2* and *Ha. marismortui sod* gene transcripts contain the complementary ribosome binding sequences GGAGG; the *Hb. cutirubrum* and *Hb.* sp. GRB *slg* gene transcripts contain only the three-nucleotide-long GGA complementary sequence. The *Hf. volcanii sod1* gene transcript has only a GA complementary sequence and the *Hb. cutirubrum* and *Hb.* sp. GRB *sod* gene transcripts have leaders too short to exhibit complementarity (Table 2). These observations suggest that halophilic archaebacteria may use one or a combination of both the eucaryotic scanning and eubacterial 16S rRNA complementarity mechanism to initiate translation.

Table 2: Transcripts from the *sod* family of genes.

Organism	Gene	5' leader[1]	3' trailer[1]	RBS[2]	Response toParaquat[3]
Hb. cutirubrum	*sod*	2 n	36 n	-	+
	slg	13n	18n	GGA	-
Hb. sp. GRB	*sod*	2 n	36n	-	+
	slg	13n	18n	GGA	-
Hf. volcanii	*sod1*	15n	140n	GA	+
	sod2	19n	18n	GGAGG	-
Ha. marismortui	*sod*	15n	17n	GGAGG	-

[1] The approximate length in nucleotides (n) of 5' leader and 3' trailer of the respective mRNA transcripts was determined by S1 nuclease protection and/or primer extension analysis. The *Hb. cutirubrum sod* transcript has been shown to possess a 5' triphosphate end (May and Dennis, 1989).

[2] The ribosome binding site (RBS) within the transcript 5' leaders complementary to the pyrimidine sequence at the 3' end of the 16S rRNA is indicated.

[3] Genes that exhibit elevated levels of mRNA and SOD enzyme activity in the presence of paraquat are indicated by (+); unaffected genes are indicated by (-)

1.3.5. Regulation by paraquat

The drug paraquat is a generator of superoxide anions and is known to reduce the growth rate and cause a selective induction of SOD activity in *Hb. cutirubrum* (Hassan and Fridovich, 1977; May and Dennis, 1987). Primer extension, S1 nuclease protection and northern hybridization experiments demonstrate that the amount of *sod* mRNA increases concomitantly with the increase in enzyme activity (May and Dennis, 1989). The four to five fold increase is gradual and occurs over a period of 24 hours. Presumably, the increase in the amount of *sod* mRNA results from increased transcription initiation although effects on mRNA stability cannot be completely ruled out. The relatively slow increase in *sod* mRNA suggests that neither paraquat nor superoxide radicals are the direct inducers; rather, the induction is more likely to result from the indirect accumulation of chemical damage caused by the increase in superoxide radical concentration. Surprisingly, the amount of the *slg* gene mRNA was not increased in the presence of paraquat.

When examined in the presence of paraquat, the expression of the *sod* gene of *Hb.* sp. GRB and the *sod*1 gene of *Hf. volcanii* was increased whereas the expression of the *slg* gene of *Hb.* sp. GRB, the *sod2* gene of *Hf. volcanii*, and the *sod* gene of *Ha. marismortui* was not affected (see above and Table 2). If the property of inducibility had a single ancestral origin for these genes and if it were mediated at the level of transcription initiation, one might expect to see conservation of a regulatory sequence element in the 5' flanking regions of the regulated genes of *Hb. cutirubrum* (or *Hb.* sp. GRB) and *Hf. volcanii.* When these regions were compared, no similarity was apparent even when gaps were introduced into the alignments. This implies either that the property of paraquat inducibility does not reside in the 5' flanking region or that the regulatory element is subtle and not easily detectable. If paraquat regulation is mediated through mRNA stabilization, one might expect to see unique sequence or structural elements in the 3' flanking region of the mRNAs; these 3' regions exhibit neither sequence similarity nor distinctive structural features.

The question of ancestral relationship between the 5' (and 3') flanking regions of the paraquat responsive *Hb. cutirubrum sod* gene and the paraquat responsive *Hf. volcanii sod1* gene remains uncertain. Either they are related, and nucleotide substitutions, insertions, and deletions have long since obliterated any significant sequence similarity, or they are not related. The latter might imply that at least one of the two genes was generated by reverse transcription of mRNA and reinsertion into the genome at a new location. This would not explain the common pattern of regulation by paraquat, however.

When the 5' flanking regions of the genes not induced by paraquat were examined, a substantial amount of hyphenated sequence similarity was detected. In Figure 6, the identical 5' flanking region of the *Hb. cutirubrum* and *Hb.* sp. GRB *slg* genes are aligned to the corresponding regions of the *Hf. volcanii sod2* genes and the *Ha. marismortui sod* gene. In the first 55 nucleotides (position -1 to -55), the pairwise sequences are between 54 and 61 percent identical, and 40 percent of the nucleotides are conserved in all three sequences. The alignments are interrupted only by the

Figure 6: Alignment of the 5' flanking regions of the non-inducible *sod* family of genes.

The 5' flanking regions of the non-inducible *Hb. cutirubrum slg*, *Hb.* sp. GRB *slg*, *Hf. volcanii sod2* and *Ha. marismortui sod* genes are aligned with the ATG translation initiation codons. The *Hb. cutirubrum* and *Hb.* sp. GRB sequences are identical; nucleotide identities between the sequences are indicated by dots (•) and the single gap required to maintain alignment is indicated by a dash (-). Upper case in the consensus indicates perfect conservation whereas lower case indicates conservation in only two of the three sequences. At the bottom, the 5' flanking region of the inducible *sodA* gene of *E. coli* (Takeda and Avila, 1986) is illustrated. Identities with the *Ha. marismortui* are indicated by dots and gaps required to maintain the alignment are indicated by dashes. The *E. coli* sequence exhibits slightly fewer identities with the other halophilic sequences.

```
          |         |         |         |         |         |         |         |
GRB slg   CGCGCCGACGAGTGAGCCCACGATCCTCACGAACATTTAACATGAC-GCCGCGTGATCACTGATCCGGTGGATTCCACCG ATG
             ●  ●          ●      ●●●●   ●●  ●●●●●●●●●●   ●●●●  ●●    ●●●● ●●● ●● ● ●      ●●●
Hvo sod2  ACTGTGGTTTGTGCTAGCATACCCCCTCCTCAAGTTTTAACATGAT-TCCGCCCGACGCATGATACGGAGGTTACACATT ATG
           ●  ●● ●  ●      ●  ●    ●●●●● ●●  ●●●●●●  ●    ●●●● ●●●●        ● ●●●●●●        ● ●●●
Hma sod   CCCTTGTTGAGATGCTACCAAATGGCTCCTAAACGTTTAACCAGGCAGCCGCGCGACAATACTGATGGAGGCGTTTTCAT ATG
            ●         ●●   ●●●     ●●●   ●●● ●●●●●●  ● ● ●●●●●● ●●  ●       ●● ●●  ●   ●   ●●●
Hcu slg   CGCGCCGACGAGTGAGCCCACGATCCTCACGAACATTTAACATGAC-GCCGCGTGATCACTGATCCGGTGGATTCCACCG ATG

CONSENSUS cccgtggt--g-tg---Ccaa---cCTCct-AAc-TTTAACatGac-gCCGCgcGAc-a-tga-acGGaGG-ttc--c-t ATG

              ●           ● ● ●●  ●●●          ●●●●● ●    ●●●●● ●●●●●●●●●●   ●●●●  ●  ●
Eco sodA  TTAATTAACTATAATGAACCAACTGCTTACGCGGCATTAACAATCG-GCCGCCCGACAATACT---GGAGATGAAT---- ATG
```

insertion of a single A residue into the *Ha. marismortui* sequence at position -34. Notably, two of the blocks of conservation within this region are the box A promoter element (-44 to -39) and the potential ribosome binding sequence (-14 to -10). The significance of other conserved nucleotides is not known.

The downstream coding region for these sequences (beginning at position +1) exhibits about 77% nucleotide sequence identity in the same pairwise comparisons, whereas the 3' flanking regions beyond the translation termination codons exhibit no significant homology outside of the T tract termination sequence. Similarly, the sequences upstream of position -55 exhibits no significant similarity. The translation termination codon of the upstream photolyase gene in *Hb. cutirubrum* and *Hb.* sp. GRB occurs at position -67; the equivalent gene is not present at the corresponding position in the *Hf. volcanii* and *Ha. marismortui* sequences. These comparisons suggest that the promoter region (defined by the 55 nucleotide of 5' flanking sequence) of the *slg* gene of *Hb. cutirubrum* and *Hb.* sp. GRB, the *sod2* gene of *Hf. volcanii* and the *sod* gene of *Ha. marismortui* are related and that sequence divergence within this region is only about two-fold more rapid than within the coding regions of the corresponding genes. Unfortunately, these comparisons shed little light on the differences between paraquat inducible and non-inducible genes and the mechanisms of induction.

The 5' flanking region of the non-inducible *sod* gene of *Ha. marismortui* was used to search the GENBANK data base; the highest similarity detected was to the 5' flanking region of the *Escherichia coli sodA* gene (Takeda and Avila, 1986). This is surprising for a non-coding region given (i) the evolutionary distance between eubacteria and halophilic archaebacteria and (ii) the fact that the *E. coli sodA* gene is inducible by paraquat. The 5' flanking regions of other eubacterial *sod* genes fail to exhibit this degree of sequence similarity. The significance of this limited sequence similarity is uncertain.

In only one case, *Hf. volcanii*, has the genomic context of the paralogous gene pairs been examined. The uninducible *sod2* gene is linked to the *rrnA* operon on the 2.4 mbp chromosome whereas the inducible *sod1* gene was mapped to the 690 kbp mega plasmid. As yet, no other genes have been mapped to this mega plasmid (Charlebois et al., 1991). Perfect identity is exhibited between the *sod1* and *sod2* genes beginning at codon four and extending to the translation termination codon at the end of the genes. This would seem to imply that the sequences of the two genes are being maintained by recombination or gene conversion mechanisms.

1.4. SUMMARY

Four species, representing three genera of halophilic archaebacteria were examined for the presence of genomic sequences that encode proteins of the superoxide dismutase (SOD) family. Three species, *Halobacterium cutirubrum*, *Halobacterium* sp. GRB, and *Haloferax volcanii* contain duplicated (paralogous) genes of the *sod* family; a fourth species, *Haloarcula marismortui*, contains only a single gene. These seven genes were cloned and sequenced, and their transcripts were characterized by northern hybridization, S1 nuclease protection, and primer extension.

The expression of one of the two genes in *Hb. cutirubrum*, *Hb.* sp. GRB and *Hf. volcanii* was shown to be elevated in the presence of paraquat, a generator of superoxide radicals. The other genes including the single gene from *Ha. marismortui* exhibited no elevated expression in the presence of paraquat.

The 5' and 3' flanking regions of all the genes contain recognizable promoter and terminator elements that are appropriately positioned relative to the 5' and 3' transcript end sites.

Between genera, the orthologous paraquat responsive genes exhibit no sequence similarity in either their 5' or 3' flanking regions whereas the orthologous non-responsive genes exhibit limited sequence similarity but only in the 5' flanking region. Within the coding region, the two paralogous genes of *Hf. volcanii* are virtually identical (99.5%) in spite of the absence of similarity in the flanking regions. In contrast, the paralogous genes of *Hb. cutirubrum* and *Hb.* sp. GRB are only about 87% identical. In the alignment of all seven sequences, there are nine codon positions where both the TCN and

AGY serine codons are utilized; some or all of these may well be examples of convergent evolution.

CHAPTER 2

The family of superoxide dismutase proteins from halophilic archaebacteria: structure, function and evolution

2.1. INTRODUCTION

Evolution has produced two unrelated proteins with superoxide dismutase activity. One is a Cu/Zn enzyme and is confined almost exclusively to the cytoplasm of eucaryotic organisms. The second contains Fe or Mn as metal ion cofactor and is found in eubacteria, in the eubacterial derived eucaryotic mitochondria and in archaebacteria (Steinman, 1982; May and Dennis, 1987; Takao et al., 1991; Kirby et al., 1981). These enzymes catalyze the dismutation of the reactive superoxide anion (O_2^-) to molecular oxygen and peroxide and thus protect organisms from chemical damage and inactivation (Fridovich, 1978, 1986a).

Halophilic archaebacteria are a collection of related organisms that evolved from an anaerobic methanogen ancestor and that now grow in nature either aerobically or microaerobically in high salt environments (Woese, 1987). Superoxide dismutase activity has been detected in at least two representatives of the ancestral methanogen group and in numerous halophilic species (Kirby et al., 1981 and Takao et al., 1991). The SOD enzyme has been purified and extensively characterized from both *Halobacterium cutirubrum* and its immediate relative, *Hb. halobium*; the enzyme contains Mn as a metal ion cofactor and is related to the Mn or Fe type enzymes of eubacteria (May and Dennis, 1987; Salin and Oesterhelt, 1988).

The *sod* gene, encoding the authentic SOD protein, was cloned from the genome of *Hb. cutirubrum* and its expression was shown to be elevated in the presence of paraquat, a compound that generates superoxide radicals (May and Dennis, 1989; May et al., 1989). In addition, a second related gene, *slg*, was also cloned and characterized. Although the *slg* gene is actively transcribed, its mRNA is not elevated by paraquat and as yet no enzymatic activity has been associated with the protein it encodes (May and Dennis, 1990). The pattern of nucleotide sequence divergence between *sod* and *slg* is highly unusual; silent third codon position substitutions are relatively infrequent and most substitutions are consequently nonsynonymous, resulting in amino acid replacement in the respective proteins. This pattern suggests that the SLG protein and possibly also the SOD protein are under intense and perhaps fluctuating selection for new and divergent function. This selection is almost certainly influenced by the concentration of the superoxide radical and possibly also by the composition and concentration of intra- and extracellular cations.

To extend this analysis, additional genes of the *sod* family have been cloned and characterized from other halophilic archaebacteria. In this chapter, the deduced halophilic protein sequences was compared to other Mn and Fe SODs and residues that characterize and distinguish these halophilic SODs were identified.

2.2. RESULTS AND DISCUSSION

2.2.1 Amino acid and nucleotide sequence alignments.

The deduced amino acid sequences of the SOD family of proteins from halophilic archaebacteria are aligned in Figure 7. To facilitate comparison with SOD sequences from other nonhalophilic organisms, the numbering system of Parker and Blake (1988) has been adopted. The three extra residues in the *Haloarcula marismortui* sequence near the amino terminus are accommodated between positions 2 and 3.

The most striking features of the halophilic SOD sequences are their high degree of sequence similarity and their high proportion of acidic amino acids. In pairwise comparison, the amino acid sequences are between 76 and 100% identical and 125 of the 199 common residues (62%) are conserved in all seven sequences (see the halophilic consensus sequence in Fig. 7). In contrast, any one of the seven halophilic proteins exhibits only about 35-40% sequence identity with eubacterial and eucaryotic mitochondrial SODs and about 40-45% identity to the SOD of *Methanobacterium thermoautotrophicum*. At the DNA level, 64 of the 125 codon positions, specifying amino acids conserved in the halophilic SOD proteins, use single unique triplets whereas the remaining 61 are degenerate (see chapter 1). Especially interesting is the use of both the TCN and AGY Ser codons at three positions (-1, 19 and 168) where Ser is conserved in all seven halophilic proteins. Also striking are the 74 positions of amino acid variability. At nearly half of these positions, connection between the encoding triplets requires

Figure 7: Amino acid sequence alignments of SOD proteins

The seven halophilic SOD sequences are aligned using the numbering system of Parker and Blake (1988). Abbreviations are: Hcu, *Hb. cutirubrum*; GRB, *Hb.* sp. GRB; Hvo, *Hf. volcanii*; Hma, *Ha. marismortui*; Mth, *M. thermoautotrophicum*; Eco, *E. coli.* The three extra residues near the amino terminus of the *Ha. marismortui* SOD are accommodated between positions 2 and 3. The *Hb. cutirubrum* sequence is given in its entirety; within the halophilic group, amino acids identical to the *Hb. cutirubrum* residues are indicated by a dash (-). The halophilic consensus sequence illustrates only residues conserved in all seven sequences; non-conserved residues are indicated by small dots (•). The halospecific signature residues are indicted by large dots (•) above the halophile consensus sequence. Deletions required to maintain alignments are indicated by the open circles (o). At the end of the alignment, the molecular weight, the number of acidic and basic amino acid residues, and the pI of the respective proteins, as determined from their primary sequences, are indicated. The *M. thermoautotrophicum* and *E. coli* sequences were obtained from Takao et al (1991) and Takeda and Avila (1986).

```
                          1                       10                  20                  30
Hcu Mn SOD    o o o o M S E V o o o E L P P L P Y D Y D A L E P H I S E Q V L T W H H D T H H Q G Y V N G
GRB SOD       o o o o - - - o o o o - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Hcu SLG       o o o o - - Q H o o o - - - S - - - - - - - - - - - - - - - - V - - - - - - - - - S - - D -
GRB SLG       o o o o - - Q H o o o - - - S - - - - - - - - - - - - - - - A V - - - - - - - - - S - - D -
Hvo SOD1      o o o o - - D - o o o - - D - - - - E - - - - - - - - - - - - - - - - - - - - - - - - - - -
Hvo SOD2      o o o o - - o - o o o - - D - - - - E - - - - - - - - - - - - - - - - - - - - - - - - - - -
Hma SOD       o o o o - - - H S N P - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
                                                                      •         •       •
HALOPHILIC CONSENSUS  M S . . o o o E L . . L P Y . Y D A L E P H I S E Q . . T W H H D T H H Q . Y V . G
Mth Fe SOD    N D L E K K F Y o o o E L P E L P Y P Y D A L E P H I S R E Q L T I H H Q K H H Q A Y V D G
Eco Mn SOD            M S Y o o o o T L P S L P Y A Y D A L E P H F D K Q T M E I H H T K H H Q T Y V N N

    40                  50                  60                  70                  80
W N D A E E T L A E N R E T o o o o o o o G D H A S T A G A L G D V T H N G S G H I L H T L F W Q S
- - - - - - - - - - - - - - o o o o o o o - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
L - S - - - - - - - - - - - o o o o o o o - - - - - - - - - - - - - - - - - C - - Y - - - M - - E H
L - S - - - N - - G - - - - o o o o o o o - - - - - - - - - - - - - - - - C - - Y - - - M - - - E H
- - A D D - - - - - - - - A o o o o o o o - E F G - S - - - V R N - - - - - - - - - - - D - - - - N
- - A D D - - - - - - - - A o o o o o o o - E F G - S - - - V R N - - - - - - - - - - - D - - - - N
L E S - - - - - - - - - D A o o o o o o o - - F G - S - A - W V N - - - - - C - Q D - - - - - - E N
                •   •                             •
. . . . . E . L A . N R . . o o o o o o o G . . . S . A . A . . . V T H N . . G . . L H . . F W . .
A N A L L R K L D E A R E S o o o o o o o D T D V D I K A A L K E L S F H V G G Y V L H L F F W G N
A N A A L E S L P E F A N T P V E E L I T K L D Q L P A D K K T V L R N N A G G H A N H S L F W K G

    90                  100                 110                 120                 130
M S P o o A G G D E P o S G A L A D R I A A D F G S Y E N W R A E F E A A A S A A o o S G W A L L V
- - - o o - - - - - - o - - - - - - - - V - - - - - - - - - - - - - - - - - - - - o o - - - - - - -
- - - o o D - - G - - o - - - - - - - - - - - - - - - - - - - - - - - V - - G - - o o - - - - - - -
- - - o o D - - G - - o - - - - - - - - - - - - - - - - - - - - - - - V - - G - - o o - - - - - - -
- - - o o E - - - - - o E - L - - E - - - E - - - - - - A - K G - - - - - - G - - o o A - - - - - -
- - - o o E - - - - - o E - L - - E - - - E - - - - - - A - K G - - - - - - G - - o o A - - - - - -
- D - o o N - - G - - o E - E - L - - - E E - - - - - - G - K G - - - - - - - - - o o G - - - - - -
        • •                                         •     •
M . P o o . G G . E P o . G . L . . R I . . D F G S Y E . W . . E F E . A A . A A o o . G W A L L V
M G P A D E C G G E P o S G K L A E Y I E K D F G S F E A F R K E F S Q A A I S A E G S G W A V L T
L K K o o o G T T L Q o o G D L K A A I E R D F G S V D N F K A E F E K A A A S R F G S G W A W L V

    140                 150                 160                 170           180
Y D o S H S N T L R N V A V D N H D E G A L W o o o o o o G S H P I L A L D V W E H S Y Y Y D Y G P
- - o - - - - - - - - - - - - - - - - - - - - o o o o o o - - - - - - - - - - - - - - - - - - - - -
- - o P V A K Q - - - - - - - - - - - - - - - o o o o o o - - - - - - - - - - - - - - - - - - - - -
- - o P V A K Q - - - - - - - - - - - - - - - o o o o o o - - - - - - - - - - - - - - - - - - - - -
- - o - F - - Q - - - - V - - K - - Q - - - - o o o o o o - - - - - - - - - - - - - - - - H - - - -
- - o - F - - Q - - - - V - - K - - Q - - - - o o o o o o - - - - - - - - - - - - - - - - H - - - -
- - o P C A K Q - - - - P - - K - - Q - - - - o o o o o o - - - - - - - - - - - - - - - - - - - - -
  •                 •       •         • • • •               • •                     •           • •
Y D o . . . . . L R N V . V D . H D . G A L W o o o o o o G S H P I L A L D V W E H S Y Y . D Y G P
Y C o Q R T D R L F I M Q V E K H N V N V I P o o o o o o H F R I L L V L D V W E H A Y Y I D Y R N
L K o o o G D K L A V V S T A N Q D S P L M G E A I S G A S G F P I L G L D V W E H A Y Y L K F Q N

    190                 200                 210                 PROTEIN         M.W.    ACIDIC BASIC    pI
D R G S F V D A F F E V V D W D E P T E R F E Q A A E R F E     Hcu SOD         22386     38      7     4.17
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -     GRB SOD         22414     38      7     4.16
- - - - - - - - - - - - I - - - P I A A N Y D D V V S L - -     Hcu SLG         22209     36      6     4.17
- - - - - - - - - - - - I - - - P I A A N Y D D V V S L - -     GRB SLG         22122     35      6     4.17
A - - D - - S - - - - - - - - - - - A A - Y - - - V - L - -     Hvo SOD1        22485     39      8     4.17
A - - D - - S - - - - - - - - - - - A A - Y - - - V - L - -     Hvo SOD2        22352     38      8     4.20
A - - D - I - - - - - - - - - - K A A - E Y - K S V S H - -     Hma SOD         22771     41      9     4.16
    •   •         • •     •                             • •
. R G . F . . A F F E V . D W D . . . . . . . . . . . . F E     HALO CONSENSUS
V R P D Y V E A F W N I V N W K E V E K R F E D I L o o o o     Mth SOD         24066     35     23     5.28
A R P D Y I K E F W N V V N W D E A A A R F A A K K o o o o     Eco SOD         23042     25     23     6.78
```

two or more nonsynonymous nucleotide substitutions. For example, at position 125 aspartic acid (GAC), serine (AGC and TCG) and alanine (GCG) all appear.

2.2.2. Signature amino acid residues in SOD

Comparison of the X-ray diffraction structures of the *Bacillus stearothermophilus* Mn SOD and a number of eubacterial and mitochondrial Mn and Fe SOD sequences resulted in the identification of 39 positions of high sequence conservation that appeared to be important in the structure or activity of the protein (Parker and Blake, 1988). These residues are listed in Table 3A. Four of these positions, His26, His81, Asp175 and His179 are essential for binding the metal ion cofactor (Stallings et al., 1984). In addition, the residues, His30, His31, Tyr34, Trp85, Trp133, and Tyr181 have been implicated in the formation of a hydrophobic pocket around the active site and in enzyme catalysis (Stallings et al., 1985). The remaining residues appear to have structural rather than functional roles. Of these 39 positions, 35 are conserved in the Fe SOD from *M. thermoautotrophicum* (Takao et al., 1991) In the halophilic SODs, only 26 of these positions are absolutely conserved; three, at residues 5, 39 and 131, are partially conserved; one at position 130 is deleted, and the remaining nine are substituted with the same amino acid in all seven sequences. None of the eleven positions of partial or complete replacement affect residues implicated in formation of the active site or in enzyme catalysis.

The SODs from halophilic archaebacteria contain 30 unique and conserved amino acid residues that are not found in any other SODs

Table 3: Signature amino acid residues in eubacterial and halophilic SODs

1. Amino acid positions in the protein alignments are presented in Figure 7.

   * indicates the four residues required for metal ion binding; ** indicates residues implicated in active site formation and/or catalysis.
2. The signature amino acid residues present in virtually all eubacterial (Eubact.) SODs are indicated (Parker and Blake, 1988; Smith and Doolittle, 1992).
3. Meth.: the single SOD sequence from methanogenic archaebacterium *M. Thermoautotrophicum* (Takao et al, 1991). Halo.: the seven halophilic archaebacterial SOD sequences. (+) denotes that the eubacterial signature residue is conserved in either the single methanogenic sequence or in all seven halophilic protein sequences. When the eubacterial signature residue is not conserved, the replacement(s) are indicated. The horizontal dashes connecting the Methanogen and Halophile columns highlight the same amino acid replacement in both archaebacterial groups. (o) denotes a gap (no amino acid) at this position in the protein alignment (see Figure 7).
4. The signature amino acid residues that occur in either halophilic archaebacterial or halophilic and methanogenic archaebacterial SOD proteins but not in any known eubacterial or mitochondrial SOD proteins are indicated. (V) denotes that the amino acid replacement in the eubacterial sequences is variable; where an amino acid is indicated in the eubacterial column, it represents a eubacterial signature residue (see Table 3A). (-) denotes that the signature residue in the seven halophilic SODs has been replaced in the methanogen sequence. The replacements can be identified in the alignment in Figure 7. The horizontal dashes connecting columns denote conservation of signature residues between eubacterial and the methanogenic or between the methanogenic and the halophilic sequences. (o) denotes a gap (no amino acid) at this position in the protein alignment (see Figure 7).

A. Amino acid residues conserved in eubacterial SOD proteins

Position[1]	Eubact.[2]	Meth.[3]		Halo.[3]
4	Leu	+		+
5	Pro	+		Pro/Asp
7	Leu	+		+
13	Ala	+		+
14	Leu	+		+
16	Pro	+		+
26 *	His	+		+
29	Lys	+		Thr
30**	His	+		+
31**	His	+		+
34**	Tyr	+		+
35	Val	+		+
39	Asn	+		Asn/Gln
80	Asn	Leu	---	Leu
81 *	His	+		+
85**	Trp	+		+
88	Leu	Met	---	Met
101	Gly	+		+
106	Ala	Tyr		Arg
107	Ile	+		+
111	Phe	+		+
112	Gly	+		+
130	Gly	+		0
131	Ser	+		Ser/Ala/Gly
133**	Trp	+		+
146	Leu	+		+
170	Pro	Ile		+
175 *	Asp	+		+
177	Trp	+		+
178	Glu	+		+
179 *	His	+		+
180	Ala	+		Ser
181**	Tyr	+		+
182	Tyr	+		+
187	Asn	+		Pro
192	Tyr	+		Phe
197	Trp	+		Phe
201	Asn	+		Asp
202	Trp	+		+

B. Amino acid residues conserved in halophilic SOD proteins

Eubact.		Meth.[4]		Halo.[4]	Position
V		-		Glu	20
V		Arg	---	Arg	24
V		-		Trp	25
Lys	---	Lys		Thr	29
V		-		Ala	46
V		-		Asn	48
V		Arg	---	Arg	49
V		-		Ser	63
V		Ala	---	Ala	67
V		-		Thr	72
V		-		His	73
V		Leu	---	Leu	80
V		Met	---	Met	88
Ala		Try		Arg	106
V		-		Tyr	114
V		-		Trp	117
V		Ala	---	Ala	128
V		-		Leu	135
V		Tyr	---	Tyr	138
V		-		Asn	148
V		Val	---	Val	151
V		-		Asp	152
V		His	---	His	154
V		-		Gly	157
V		-		Ala	158
V		-		Leu	159
V		-		Trp	160
V		-		Ser	168
V		-		His	169
Ala	---	Ala		Ser	180
V		-		Gly	186
Asn	---	Asn		Pro	187
V		-		Gly	190
Tyr	---	Tyr		Phe	192
Trp	---	Trp		Phe	197
V		-		Glu	198
Asn	---	Asn		Asp	201
0		0		Phe	216
0		0		Glu	217

at the corresponding alignment positions (Table 3B). Seven of these represent total replacement of eubacterial signature residues. In addition, there are nine positions where the single methanogenic and seven halophilic SODs exhibit a single amino acid not found in any known eubacterial SOD. These comparisons indicate in a qualitative way that the halophilic SOD proteins are almost as different from the SOD of their immediate methanogen relative as they are from the SODs of their more distant eubacterial cousins. Furthermore, the substitutions specific to the halophiles almost certainly have a substantial effect on the overall structure of the protein but leave the hydrophobic active site and the four metal ion binding residues intact.

Virtually all proteins from halophilic archaebacteria (in comparison to proteins from nonhalophilic organisms) exhibit a high content of acidic amino acids (Lanyi, 1974). These residues, usually located on the surface of the protein, are thought to sequester water molecules and create a hydration shell that allows the protein to function within the high ionic strength intracellular environment of the cell. For example, in the eubacterial SOD protein, the acidic and basic residues are equally prevalent whereas halophilic SODs contain more acidic residues and proportionately fewer basic residues. The pIs of these seven halophilic SODs range from 4.16-4.20 (Figure 7). One might predict that of the 30 halophile-specific signature residues listed in Table 3B, a substantial proportion would involve replacement of Lys or Arg by either Asp or Glu. That is, many of these substitutions should reflect the general adaptation of the enzyme to function in high salt. Surprisingly, this appears not to be the case. The 30 halophile-specific substitutions increase the number of acidic residues by only three (positions 20, 152 and 201) and an additional acidic residue (position 217) is generated within the halophile specific extension at the carboxy terminus of the protein. Only position 20 involves a lysine/arginine (*E. coli*/*M. thermoautotrophicum*) replacement by glutamic acid in the halophilic proteins.

In about 70% of the alignment positions where acidic residues are found in one or more of the halophilic SODs, an acidic residue is also found in at least one other non-halophilic SOD. Of the remaining positions, about half exhibit a basic residue in at least one non-halophilic sequence. Thus, the high acidity of the halophilic SODs has been achieved by insertion of Asp and Glu residues most often at positions where charged residues can be tolerated in non-halophilic SODs without adversely affecting enzyme activity. These sites are presumed to be on the surface of the protein molecule and well removed from the site of catalysis.

2.2.3. Nucleotide sequence divergence

When orthologous or paralogous gene sequences initially begin to diverge during evolution, mutational differences usually accumulate most rapidly in the third codon position and most substitutions are synonymous. As a consequence, nucleotide identity between two gene sequences initially degenerates more rapidly than the amino acid identity between the two corresponding proteins. Only later, when a substantial number of first and second position changes accumulate, does the amino acid identity fall below the nucleotide identity. Typical of this situation are the paralogous *p-vac* and *c-vac* genes of *Hb. halobium* and the orthologous *c-vac* gene of *Haloferax mediterranei* (Horne et al., 1988; Englert, 1990). For these, the DNA nucleotide identity is less than the amino acid identity and the vast majority of substitutions are synonymous and occur at the third codon position (Table 4).

In contrast, the halophilic *sod* genes do not conform to this expectation. The majority of substitutions are of the non-synonymous type and occur with an unexpectedly high frequency at the first and second codon positions (Table 5). Because of this, the amino acid identity of the proteins for most pairwise comparisons is less than the nucleotide identities of the genes.

Within the halophilic *sod* genes, substitutions by transversion outnumber transitions by almost two to one. The significance of this bias is uncertain. Compared to transitions, transversions have less of an effect on base composition and when they occur in the third codon position, they are more likely to be non-synonymous. This bias is not reflected in all genes from halophilic archaebacteria, however. For example, in the *vac* and 16S rRNA genes, transitions are more prevalent than transversions (Englert, 1990; May and Dennis, 1990; Mylvaganam and Dennis, 1992). Since the gas vacuole protein sequences in halophiles are highly conserved (Table 4), nonsynonymous transversion mutations in the third codon position would appear to be subject to strong negative selection and therefore less likely to become fixed in the population. Many of the 16S rRNA substitutions occur at compensatory positions within regions of RNA secondary structure. Transition mutations allow these compensatory

TABLE 4: Comparison of paralogous and orthologous gas vacuoles genes

1. Matrix comparison of the plasmid and chromosome encoded gas vacuole protein encoding genes (*p-vac* and *c-vac*, respectively). The species designations are: *Hha*, *Hb. halobium* and *Hme*, *Hf. mediterranei.*
2. Each entry in the matrix consists of three rows. The first row indicates the DNA (nucleotide) sequence identity/protein (amino acid) sequence identity, each expressed as a percent. The second row indicates the distribution of nucleotide substitutions in the comparison between first/second/third codon positions. The third row indicates the number of transition (Ts) substitution/transversion (Tv) substitutions. These designations are abbreviated in the boxed entry.

Species[1]	*Hha*	*Hha*
gene	*p-vac*	*c-vac*
Hha c-vac	84.7/97.4[2] 2/0/33 21/14	Nuc/AA 1st/2nd/3rd Ts/Tv
Hme c-vac	86.4/96.1 3/1/27 20/11	85.5/98.7 1/1/33 20/13

TABLE 5: Comparison of halophilic *sod* genes and proteins

1. Matrix comparisons of the members of the halophilic SOD encoding genes. They are as in the legend to Figure 7.
2. Matrix entries are as described in the legend to Table 4. The designations are abbreviated in the boxed entry.

TABLE 5: Comparison of halophilic *sod* genes and proteins

Species[1]	***Hcu***	***Hcu***	***GRB***	***GRB***	***Hvo***	***Hvo***
gene	*sod*	*slg*	*sod*	*slg*	*sod1*	*sod2*
Hcu slg	87.2/82.5[2] 25/21/31 31/46	— — —				
GRB sod	95.5/95.5 0/1/2 2/1	87.3/82.0 25/22/29 31/45	— — —			
GRB slg	86.7/81.0 25/24/31 33/47	99.2/98.5 0/3/2 3/2	86.7/80.5 85/25/30 33/47	— — —		
Hvo sod1	81.0/80.0 25/28/61 34/80	76.3/72.0 41/30/71 42/100	80.7/79.5 25/29/62 35/81	76.0/70.5 41/33/70 45/99	— — —	
Hvo sod2	80.9/80.4 26/29/59 34/80	77.1/72.4 39/29/69 42/95	80.6/79.9 26/30/60 35/81	76.7/70.9 39/32/68 45/94	99.5/100 1/1/1 0/3	— — —
Hma sod	77.7/74.5 36/35/63 62/72	77.2/76.0 37/30/70 57/80	78.0/75.0 36/34/62 61/71	76.3/74.5 37/33/72 60/82	76.3/78.0 29/26/87 45/97	76.2/78.4 30/27/85 45/97

Nuc/AA
1st/2nd/3rd
Ts/Tv

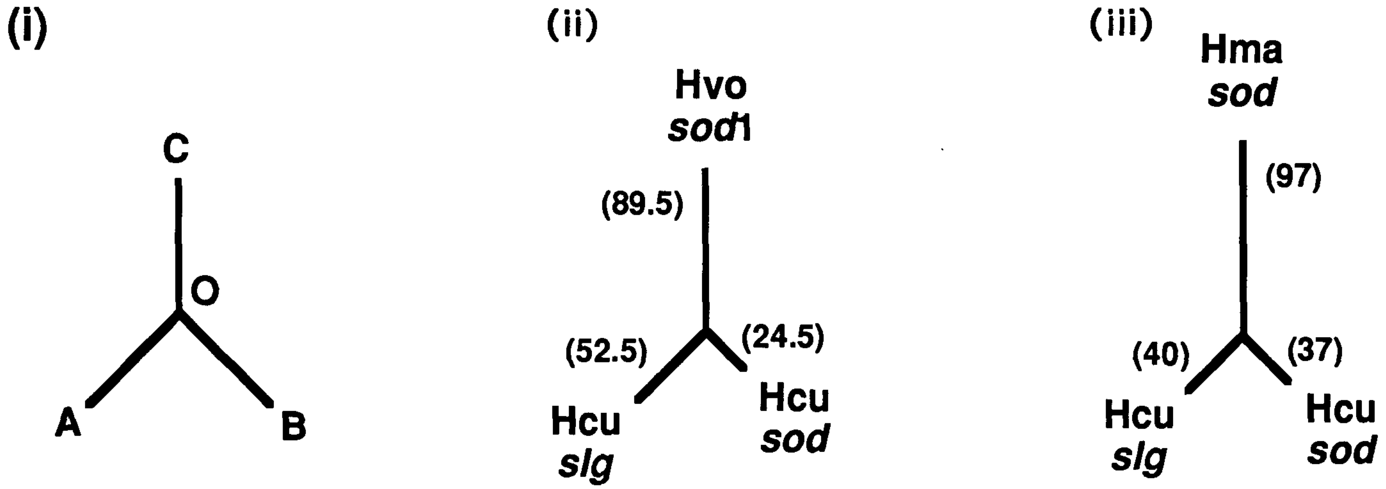
changes to proceed through stable G:U base pair intermediates and therefore could account for the transitional bias.

When this high proportion of nonsynonymous substitutions was first observed between the paralogous *sod* and *slg* genes of *Hb. cutirubrum*, it was suggested that the cause might be intense selection at the molecular level for new protein function (May and Dennis, 1990). Since the *sod* gene encodes the authentic superoxide dismutase activity, it was imagined that the superfluous *slg* gene was being used to generate a new and different enzymatic activity. That is, the *sod* gene was being conserved whereas the *slg* gene was being subjected to frequent nucleotide substitution.

The presence of additional sequences from the halophilic *sod* gene family sheds some light on this situtation. Clearly, not just the *sod* and *slg* genes of *Hb. cutirubrum* and *Hb.* sp. GRB, but all of the genes in the collection when compared pairwise appear to exhibit this unusual pattern of divergence. When the relative rate test (Sarich and Wilson, 1973) was applied to the paralogous *sod* and *slg* genes of *Hb. cutirubrum* (or *Hb.* sp. GRB) using either the *Hf. volcanii sod*1 gene or the *Hf. marismortui sod* gene as the outgroup, it was quite clear that both *sod* and *slg* have accumulated substitutions at a substantial rate (Figure 8). Relative to the *sod* gene of *Ha. marismortui,* the *sod* and *slg* genes of *Hb. cutirubrum* have accumulated mutations at essentially the same rate whereas relative to the *sod*1 gene of *Hf. volcanii* the *slg* gene has accumulated substitutions at twice the rate of the *sod* gene. This implies that both the *sod* and the *slg* genes and probably all the genes in this halophilic family have been subjected to intense but variable selective

Figure 8: Relative rates of nucleotide substitutions in the *sod* and *slg* genes of *Hb. cutirubrum*

The relative rate test is described in the Materials and Methods section. (I) depicts the generalized situation where A and B are contemporary gene sequences; O is the common ancestral sequence of A and B, and C is the outgroup sequence. The length of the three branches connecting at O are proportional to the number of nucleotide substitutions (in parenthesis) separating O from A, B and C. These are calculated using the three equations listed in the Methods. The distances (in nucleotide substitution) AB, BC and AC can be obtained from Table 3. (II) For the *Hcu slg* (A), *Hcu sod* (B), *Hvo sod*1 (C) calculations, AB, BC and AC are 77, 114 and 142, respectively. (III) For the *Hcu slg* (A), *Hcu sod* (B), *Hma sod* (C) calculations, AB, BC and AC are 77, 134, and 137, respectively.


(i)
C
O
A
B
(ii)
Hvo
*sod*1
(89.5)
(52.5)
(24.5)
Hcu
slg
Hcu
sod
(iii)
Hma
sod
(97)
(40)
(37)
Hcu
slg
Hcu
sod

pressures that have resulted in frequent amino acid replacements in each of the respective proteins. These replacements are confined to 74 of the 199 common amino acid positions (Figure 7). For the nearly identical paralogous *sod1* and *sod2* genes of *Hf. volcanii*, the forces and processes maintaining sequence homogeneity are not understood.

2.2.4. Phylogenetic analysis

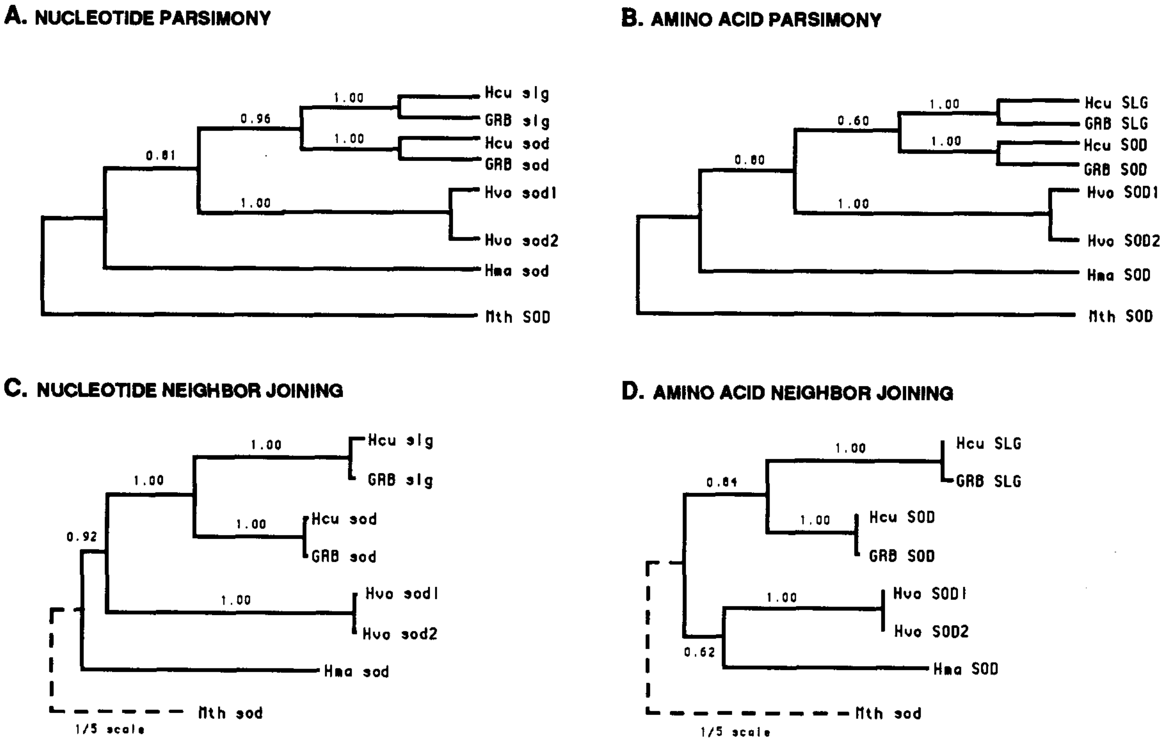
The two methods, maximum parsimony (Fitch, 1971) and neighbour joining (Saitou and Nei, 1987), were used to analyze the phylogenetic relationships among the halophilic *sod* genes and among the proteins they encode (Figure 9). The nucleotide and protein alignments used for the analysis were co-linear and are depicted in Figure 3 and in Figure 7. The use of the *sod* gene sequence from *M. thermoautotrophicum* as an outgroup allows for the positioning of the ancestral halophilic gene (or root) within the tree.

These phylogenetic analyses have been evaluated for significance at each branch point by using the commonly employed bootstrap re-sampling technique (Felsenstein, 1985). The consistency values indicated are the proportion in 100 or more resamplings for exclusive grouping of the taxonomic units (genes or proteins) within that branch of the tree, separate from all the other taxonomic units represented on the tree. A value greater than 0.95 is considered significant whereas values less than 0.95 are not necessarily significant.

The nucleotide parsimony, amino acid parsimony, and nucleotide neighbour joining trees exhibit identical topologies; the

Figure 9: Phylogenetic relationships between members of the halophilic *sod* gene family or the proteins they encode

Phylogenetic trees were constructed using maximum parsimony (A and B) or neighbour joining (C and D) methods. The neighbor joining method contains a correction for multiple substitutions occurring at a single position (Kimura 1980 and 1983). The length of the branches in the neighbour joining trees reflect evolutionary distance. The stippled branch leading to the outgroup is drawn at one-fifth scale. Trees A and C are based on DNA sequence alignments and trees B and D are based on protein sequence alignments. The numbers preceding each branch are bootstrap consistency values and indicate the proportion of replications that group all the taxonomic units within the branch and exclude all other taxonomic units represented in the tree. These consistency values were computed from 100 or more bootstrap resamplings. Abbreviations are as in Figure 8.


A. NUCLEOTIDE PARSIMONY
1.00
0.96
1.00
0.81
1.00
Hcu slg
GRB slg
Hcu sod
GRB sod
Hvo sod1
Hvo sod2
Hma sod
Mth SOD
B. AMINO ACID PARSIMONY
1.00
0.60
1.00
0.80
1.00
Hcu SLG
GRB SLG
Hcu SOD
GRB SOD
Hvo SOD1
Hvo SOD2
Hma SOD
Mth SOD
C. NUCLEOTIDE NEIGHBOR JOINING
1.00
1.00
1.00
0.92
1.00
Hcu slg
GRB slg
Hcu sod
GRB sod
Hvo sod1
Hvo sod2
Hma sod
Mth sod
1/5 scale
D. AMINO ACID NEIGHBOR JOINING
1.00
0.84
1.00
1.00
0.62
Hcu SLG
GRB SLG
Hcu SOD
GRB SOD
Hvo SOD1
Hvo SOD2
Hma SOD
Mth sod
1/5 scale

sequences from *Halobacteria* form a coherent group with the two *sod* genes (or proteins) on one branch and the two *slg* genes (or proteins) on another. The single *sod* gene (or protein) from *Ha. marismortui* branches early and the two nearly identical *sod* genes (or proteins) from *Hf. volcanii* branch later from the lineage leading to the *Halobacteria*. The bootstrap consistency values for most of the branch points is greater than 0.95. Only the amino acid neighbour joining tree exhibits a slightly altered topology. Here, the *Hf. volcanii* protein branches early from the lineage leading to *Ha. marismortui* rather than the lineage leading to *Halobacteria*. The bootstrap consistency for this grouping is 0.62.

In all these analyses, the *M. thermoautotrophicum* was clearly and unambiguously identified as the outgroup. This methanogen protein is apparently an Fe containing SOD (Takao et al., 1991) whereas the halophilic proteins are most likely all Mn containing enzymes (May and Dennis, 1987). The difference in metal ion certainly accounts for some of the amino acid differences between the methanogenic and halophilic proteins. When the *E. coli* Mn-*sod* gene and protein sequences were included, the parsimony analysis grouped the halophiles together but was unable to reliably distinguish between *Escherichia coli* or *M. thermautotrophicum* as the outgroup (not shown). This reinforces the impression that the halophilic proteins as a group are unique and different from all other SOD proteins (see Table 3).

2.2.5. The consensus tree

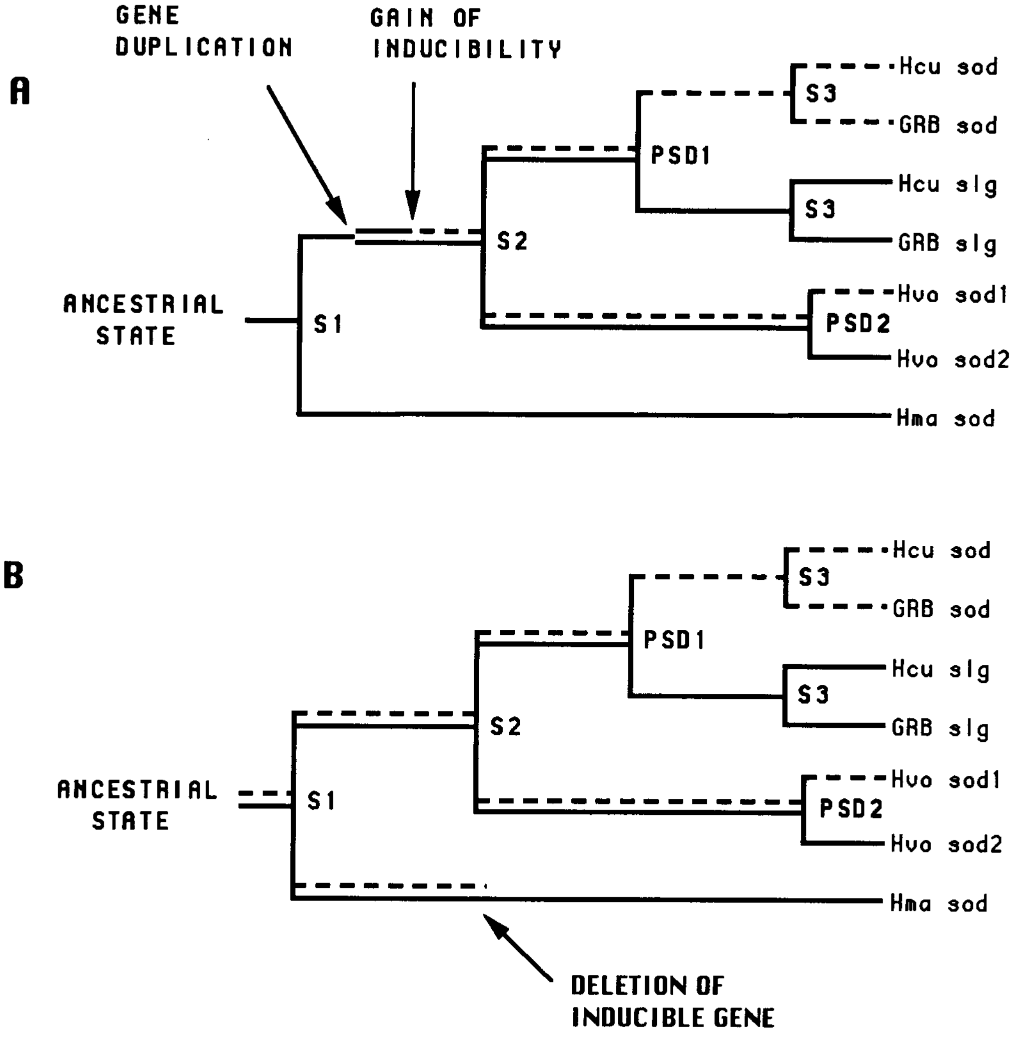
The consensus derived from these phylogenetic analyses is illustrated in Figure 10. The branch points represent either speciation events, designated S1, S2 and S3, or divergence of paralogous gene sequences, designated PSD1 and PSD2. The first speciation, S1, separates the lineage leading to *Haloarcula* spp. from other halophiles. The second event, S2, splits *Haloferax* spp. from *Halobacterium* spp., and the third more recent event, S3, splits *Hb. cutirubrum* from *Hb.* sp. GRB.

The first paralogous sequence divergence event, PSD1, represents the commencement of divergence between the paralogous *sod* and *slg* genes within the *Halobacterium* branch. The majority of the differences that have accumulated between these two sequences predate S3, the speciation of *Hb. cutirubrum* and *Hb.* sp. GRB. The second PSD2, is meant to depict the minor differences that exist between the *sod1* and *sod2* genes of *Hf. volcanii.*

Other features characteristic of halophilic *sod* genes can be superimposed on this diagram. One is *sod* gene copy number and another is response to paraquat. Both the methanogen ancestor represented by *M. thermoautotrophicum* and the early branching *Ha. marismortui* have only a single *sod* gene (Takao et al., 1991 and chapter 1). This suggests that the halophilic ancestor may also have had only a single *sod* gene in its genome. If this scenario is correct, the position for duplication would be early in the branch leading to *Halobacterium* and *Haloferax.* In the extant species examined that possess two *sod* related genes, one gene is inducible by paraquat and

Figure 10: The consensus tree illustrating phylogenetic relationships between the halophilic *sod* genes (or proteins)

The consensus tree was obtained from the four constructed trees illustrated in Figure 9. Speciation events are designated S1, S2 and S3. The commencement of paralogous gene sequence divergence is indicated by PSD1 and PSD2. (A) The positions of gene duplication and acquisition of inducibility by paraquat are indicated. Double lines indicate the presence of paralogous genes of identical coding sequence; solid lines are non-inducible genes; stippled lines are paraquat inducible genes. (B) The alternative scenario where the halophilic ancestor already possesses duplicated *sod* genes is depicted.


A
GENE DUPLICATION
GAIN OF INDUCIBILITY
ANCESTRIAL STATE
S1
S2
PSD1
S3
Hcu sod
GRB sod
S3
Hcu slg
GRB slg
PSD2
Hvo sod1
Hvo sod2
Hma sod
B
ANCESTRIAL STATE
S1
S2
PSD1
S3
Hcu sod
GRB sod
S3
Hcu slg
GRB slg
PSD2
Hvo sod1
Hvo sod2
Hma sod
DELETION OF INDUCIBLE GENE

the other gene is not inducible by paraquat (see chapter 1 and May and Dennis, 1990). This implies that shortly after duplication and prior to the separation of *Halobacterium* and *Haloferax*, one of the duplicated genes acquired, by an unknown mechanism, the property of inducibility. The *Ha. marismortui* branch retains the ancestral single copy non-inducible state (Figure 10A).

There is a second alternative scenario that is equally likely. The ancestral halophile might have already possessed duplicated *sod* genes, one of which was inducible and the other non-inducible by paraquat. To reach the current state would require simply the loss of the paraquat inducible gene from the branch leading to *Ha. marismortui*. This possibility is depicted in Figure 10B. Other more complex explanations are not considered here.

Although partially satisfying, these models for the evolution of *sod* related sequences within halophilic archaebacteria fail to explain a number of observations. The first is the absence of any detectable sequence similarity in the 5' flanking region of the *sod* related genes that were induced by paraquat (see Figure 5). This is especially enigmatic since the 5' flanking sequences of the uninducible genes that possibly have a deeper evolutionary origin (see Figure 10A), nonetheless exhibit easily identifiable sequence similarity. Second, the *sod1* and *sod2* genes of *Hf. volcanii*, although virtually identical within their coding regions, exhibit no detectable similarity in their flanking regions. This must mean that coding sequence homogeneity is maintained by concerted evolution and probably involves recombination or gene conversion type events; these events apparently do not involve or include either the 5' or 3' flanking regions.

In general, the flanking region of homologous genes, excluding conserved regulatory elements, accumulate nucleotide substitutions more rapidly than coding regions. This is in general true for the halophilic *sod* gene family. Comparison between genera indicate that only the non-inducible *sod* genes exhibit similarity and this is confined to a 50-55 nucleotide long region that contains the transcriptional promoter. Even here, however, sequence identity is substantially less than in the coding regions. This expected pattern is reversed within the two species of *Halobacterium*, *Hb. cutirubrum* and *Hb.* sp. GRB. Although the sample size is somewhat restricted, substitutions appear to be almost two fold more frequent in the coding than in the non-coding region.

In summary, nucleotide sequence divergence of the *sod* gene family from halophilic archaebacteria exhibits many unusual and remarkable features. These features indicate quite clearly that evolutionary processes are not uniform and predictable. Rather, they are complex and involve subtle but intense selection that is presumably exerted at the level of protein structure-function; somehow, this selection influences processes at the level of DNA that are only now becoming apparent from sequence data analysis. These processes include duplication, generation of variability by mutation, fixation or elimination of mutations by drift or selection, and recombination or gene conversion.

2.3 Summary

The protein sequences of seven members of the superoxide dismutase (SOD) family from halophilic archaebacteria have been aligned and compared with each other and with the homologous Mn and Fe SOD sequences from eubacteria and the methanogenic archaebacterium *Methanobacterium thermoautotrophicum*. Of 199 common residues in the SOD proteins from halophilic archaebacteria, 125 are conserved in all seven sequences and 64 of these are encoded by single unique triplets. The 74 remaining positions exhibit a high degree of variability and for almost half of these the encoding triplets are connected by at least two nonsynonymous nucleotide substitutions.

The majority of nucleotide substitutions within the seven genes are nonsynonymous and result in amino acid replacement in the respective protein, silent third codon position (synonymous) substitutions are unexpectedly rare. Halophilic SODs contain 30 specific residues that are not found at the corresponding positions of the methanogenic or eubacterial SOD proteins. Seven of these are replacements of highly conserved amino acids in eubacterial SODs that are believed to play an important role in the three dimensional structure of the protein. Residues implicated in formation of the active site, catalysis, and metal ion binding are conserved in all Mn and Fe SODs.

Molecular phylogenies based on parsimony and neighbour joining methods coherently group the halophile sequences but surprisingly fail to distinguish between the Mn SOD of *E. coli* and the Fe SOD of *M. thermoautotrophicum* as the outgroup.

These comparisons indicate that as a group the SODs of halophilic archaebacteria have many unique and characteristic features. At the same time, the patterns of nucleotide substitution and amino acid replacement indicate that these genes and the proteins they encode continue to be subject to strong and changing selection. This selection may be related to the presence of oxygen radicals and the inter- and intracellular composition and concentration of metal cations.

CONCLUSIONS

The orthologous and paralogous members of the superoxide dismutase gene family in the halophilic archaebacteria were cloned and sequenced. The number of copies of these genes present in the genome varies from one in *Ha. marismortui* to two in *Hf. volcanii* and *Hb.* sp.GRB. All the genes show high sequence identity to the *sod* and *slg* genes of *Hb. cutirubrum.*

The responses to paraquat by the five genes examined in this study are not constant; mRNA levels from *Hb.* sp. GRB *sod* and *Hf. volcanii sod1* were found to increase (as reported previously for *Hb. cutirubrum sod*) and those of *Hb.* sp. GRB *slg, Hf. volcanii sod2* and *Ha. marismortui* (as reported previously for *Hb. cutirubrum slg*). In *Ha. marismortui*, the presence of a single uninducible *sod* gene suggests that this organism may possess sufficient enzymatic SOD activity to counter the effects of intracellular superoxide radicals without the need to increase intracellular levels of the enzyme. An alternative possibility is that non-enzymatic SOD activity, similar to those reported for free Mn and Mn-complexes (Archibald and Fridovich, 1981; Rush *et al.,* 1991), may exist in this organism.

Analyses of the amino acid sequences of putative protein products of the *sod* and *slg* genes indicate that the proteins are typical of the other halobacterial proteins; they contain an excess of acidic residues over basic residues. Comparison of the halobacterial SOD sequences with the Mn/Fe SOD from other organisms has identified 30 residues that are specific for the seven halobacterial proteins. Only five of these residues are acidic, indicating that the acidic nature of the halobacterial proteins is due to the independent accumulation of acidic residues at sites where such residues are tolerated in other non-halobacterial SODs. This convergent evolution is one of the ways halobacterial proteins have evolved to withstand the high intracellular salt concentration without drastically compromising the overall structure-function of the proteins.

The comparison of the different *sod* and *slg* genes both within and between species of the halobacteria shows an unusual pattern of substitutions; there is no strong bias for third codon position substitutions. As a result, non-synonymous substitutions tend to be retained more frequently than synonymous substitutions and a more rapid change in amino acid sequence identity compared to the nucleotide identity is observed. This in contrast to the situation of other duplicated genes where, in the initail stages of divergence, nucleotide identity changes more rapidly than the amino acid identities of the proteins they encode (R. F. Doolittle, personal communication) It is also observed that transversions outnumber transitions in almost all pairwise comparisons of the members of the *sod* gene family. The significance of this bias is uncertain.

May and Dennis (1990) suggested that this pattern of substitutions reflects diverged, or diverging, protein functions. The results from the present study show that this pattern of substitutions also occurs between pairs of orthologous genes of almost certainly conserved function in different halophiles. It is difficult to establish whether this pattern is due to positive selection or random drift. It is possible that each protein is under strong selection for optimal activity in different (or changing) micro-environments. The environments might be related to the undetectable difference in internal and external salt compositions and concentrations.

In addition to this unusual pattern of substitutions, the halobacteria also possess, as seen in *vac* genes in *Hb. halobium* and *Hb. mediterranei*, the usual pattern that is found between closely related genes in the eubacteria and the eucaryotes. In these organisms, a strong bias exists for substitutions in the third codon position and transitions outnumber transversions. The pressures and mechanisms responsible for different patterns of mutation within different regions of the genome are currently not understood.

A consensus phylogenetic tree has been derived from the parsimony and neighbour joining analyses. When characteristic features of the genes of the halophilic *sod* family are superimposed onto the consensus tree, two scenarios describing the evolution of the genes can be formulated. The ancestral halophile might have possessed duplicated genes, one of which was inducible and the other non-inducible by paraquat. The present state would then result from the simple loss of the inducible gene from the branch leading to *Ha. marismortui*. Alternatively, the halophilic ancestor possessed only in single *sod* gene which was unaffected by paraquat. A gene duplication event accompanied by aquisition of paraquat inducibility early in the branching leading to Halobacterium and Haloferax would result in the presence of paralogous inducible and non-inducible genes in these genera.

This work has shown that nucleotide sequence divergence of the *sod* gene family from halophilic archaebacteria exhibits many unusual and remarkable features. These features indicate quite clearly that evolutionary processes are complex and involve subtle selection that is presumably exerted at the level of protein structure-function. This selection reflects itself in changes at the DNA level that are only now becoming apparent from sequence data analysis. These processes include duplication, generation of variability by mutation, fixation or elimination of mutations by drift or selection, and recombination or gene conversion.

FUTURE RESEARCH PROSPECTS

1) What is the function of the proteins that are encoded by the *slg* genes in *Hb. cutirubrum* and *Hb.* sp. GRB? It is possible that *slg* encodes SOD activity which was lost during purification or not detectable with the assay employed. In order to determine if *slg* does indeed encode SOD, *sod*$^-$ *Hb. cutirubrum* strains could be created by homologous recombination and assayed for SOD activity. In addition, complementation of *sod*$^-$ *slg*$^-$ strains with an expression plasmid containing the *slg* gene could be attempted. The construction of a *sod*$^-$ and/or *slg*$^-$ genotype by homologous recombination, using appropriate deletions, is now feasible with the availability of transformation systems for halobacteria.

2) What are the signals for paraquat inducibility in the halophilic *sod* genes? It may be possible to localize putative regulatory regions by deletion, by site specific mutational analysis of the flanking regions, and by construction of hybrids of paraquat-and non paraquat-inducible genes. For this study, *Hb. cutirubrum* or *Hb.* sp. GRB *sod*/*slg* genes with appropriate modifications of the flanking

regions contained on a shuttle plasmid could be utilized. Such constructs can be transformed into *Hf. volcanii* and the expression of the heterologous *sod* gene can be monitored by transcript analysis. By using appropriate constructs and controls, the paraquat responsive regions could then be delineated.

3) How are the other *sod* genes in the archaebacteria related in their primary structure and the regulation by paraquat? To answer this question, *sod* genes from other archaebacteria would have to be cloned. Since Mn-SOD polypeptides from different organisms have universally conserved regions, it should be possible to clone *sod* genes using degenerate oligonucleotides for amplification by the polymerase chain reaction. Subsequent phylogenetic analysis using a larger number of sequences will provide more insight into the evolution of the superoxide dismutase genes in the archaebacteria.

REFERENCES

Antonini, E., Brunori, M., Greenwood, C. and Malmstrom, B.G. 1970. Catalytic mechanism of cytochrome oxidase. Nature. 228: 936-937.

Archibald, F.S. and Fridovich, I. 1981. Manganese and defenses against oxygen toxicity in *Lactobacillus plantarum*. J. Bacteriol. 145: 442-451.

Bayley, S. T. 1971. Protein synthesis systems from halophilic bacteria. Methods Mol. Biol. 1:89-100.

Bannister, J. V., and Parker, M. W. 1985. The presence of a copper/zinc superoxide dismutase in the bacterium *Photobacterium leiognathi*: a likely case of gene transfer from eukaryotes to prokaryotes. Proc. Natl. Acad. Sci. 82:149-152.

Boer, P. H and Hickey, D. A. 1986. The alpha amylase gene in *Drosophila melonogaster*: nucleotide sequence, gene structure and expression motifs. Nucl. Acid. Res. 14:8399-8410.

Brenner, S. 1988. The molecular evolution of genes and proteins: a tale of two serines. Nature 334:428-430.

Cadenas, E.,1989. Biochemistry of oxygen toxicity. Annu. Rev. Biochem. 58: 79-110.

Carlioz, A., and Touati, D. 1986. Isolation of superoxide dismutase mutants of *Escherichia coli*; Is superoxide dismutase necessary for aerobic life? EMBO J. 5:623-630.

Chant, J., and Dennis, P. P. 1986. Archaebacteria: transcription and processing of ribosomal RNA sequences in *Halobacterium cutirubrum*. Embo. J. 5:1091-1097.

Chapman, D. J., and Schopf, J.W. 1983. Biological and biochemical effects of the development of an aerobic environment. In J. W. Schopf (Ed.) Earth's Earliest Biosphere: It's Origin and Evolution. Princeton Univ. Press: Princeton, N.J.

Charlebois, R. L., Schalkwyk, L. C., Hofman, J. D., and Ford Doolittle, W. 1991. Detailed physical map and set of overlapping clones covering the genome of the archaebacterium *Haloferax volcanii* DS2. J. Mol. Biol. 222:509-524.

Cline, S. W., Lam, W. L., Charlebois, R. L., Schalkwyk, L. C., and Ford Doolittle, W. 1989. Transformation methods for halophilic archaebacteria. Can. J. Microbiol. 35:148-152.

Daniels, C.J., McKee, A. H. Z., and Doolittle, W. F. 1984. Archaebacterial heat-shock proteins. EMBO J. 3: 745-749.

Demple, B., Amabile-Cuevas, C.F. 1991 Redox Redux: The control of oxidative stress responses. Cell. 67: 837-839.

Dennis, P. P. 1985. Multiple promoters for the transcription of ribosomal RNA gene cluster in *Halobacterium cutirubrum*. J. Mol. Biol. 186: 457-461.

Dente, L., and Cortese, R. 1983. pEMBL: A new family of single-stranded plasmids for sequencing DNA. Methods in Enzymology. 155:111-118.

Ebert, K., Goebel, W., and Pfeifer, F. 1984. Homologies between heterogeneous extrachromosal DNA populations of *Halobacterium halobium* and four new halobacterial isolates. Mol. Gen. Genet. 194:91-97.

Englert, C., Horne, M., and Pfeifer, F. (1990). Expression of the major gas vesicle protein gene in the halophilic archaebacterium *Haloferax mediterrane* is modulated by salt. Mol. Gen. Genet. 222: 225-232.

Fee, J. A. 1991. Regulation of sod genes in *Escherichia coli*: relevance to superoxide dismutase function. Mol. Microbiol. 5: 2599-2610.

Feinberg, A., and Volgelstein, B. 1982. A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. 132:6-13.

Felsenstein, J. 1988. Phylogenies from molecular sequences; Inference and reliability. Annu. Rev. Genet. 22:521-565.

Fitch, W. M. 1971. Toward defining the course of evolution; Minimal change for a specific tree typology. Syst. Zool. 20:406-416.

Fitch, W. M, and Ferris, S. D. 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. J. Mol. Evol. 3:263-278.

Fridovich, I. 1978. The biology of oxygen radicals. Science. 201: 875-880.

Fridovich, I. 1986a. Biological effects of the superoxide radical. Arch. Biochem. Biophys. 247:1-8.

Fridovich, I. 1986b. Superoxide dismutases. Adv. Enzymol. 58:68-97.

Haber, F., and Weiss, J. 1934. The catalytic decomposition of hydrogen peroxide by iron salts. Proc. R. Soc. London Ser. A. 147:332-351.

Hassan, H. M., and Fridovich, I. 1977. Regulation of the synthesis of superoxide dismutase in *Escherichia coli:* Induction by methyl viologen J. Biol. Chem. 252:7667-7672.

Henikoff, S. 1987. Exonuclease III generated deletions for DNA sequence analysis. Promega Notes. Promega Corp. No. 8. (pp. 1-3).

Hickey, D. A., Bally-Cuif, L., Abukashawa, S., Payant, V., and Benkel, B. F. 1991. Concerted evolution of duplicated protein-coding genes in *Drosophila*. Proc. Natl. Acad. Sci. 88:1611-1615.

Holmes, M. L., and Dyall-Smith M.L. 1990. A plasmid vector with a selectable marker for halophilic archaebacteria. J. Bacteriol. 172:756-761.

Holmes, M. L., Nuttall, S. D., and Dyall-Smith, M. L. 1991. Construction and use of halobacterial shuttle vectors and further studies on *Haloferax* DNA Gyrase. J. Bacteriol. 173:3807-3813.

Horne, M., Englert, C., and Pfeifer, F. 1988. Two genes encoding gas vacuole proteins in *Halobacterium halobium.* Mol. Gen. Genet. 213: 459-464.

Hsia, H., J. Lebkowski, P. Leong, M. Calos and J. Miller. 1989. Comparison of ultraviolet irradiation induced mutagenesis of the *lac* I gene in *Escherichia coli* and in Human 293 cells. J. Mol. Biol. 205:103-113..

Imlay, J.A. and Fridovich, I. 1991. Assays of metabolic superoxide products in *Escherichia coli.* J. Biol. Chem 266: 6957-6965.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. 86:9355-9359.

Jones, W. J., Nagle, D. P., and Whitman, W. B. 1987. Methanogens and the diversity of archaebacteria. Microbiol. Rev. 51:135-177.

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.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kirby, T. W., Lancaster, J. R., and Fridovich, I. 1981. Isolation and characterization of the iron-containing superoxide dismutase of *Methanobacterium bryantii.* Arch. Biochem. Biophys. 210:140-148.

Kozak, M. 1978. How do eucaryotic ribosomes select inititation regions in messenger RNA? Cell. 15:1109-1123.

Lam, W. L., and Ford Doolittle, W. 1989. Shuttle vectors for the archaebacterium *Halobacterium volcanii.* Proc. Natl. Acad. Sci. USA. 86: 5478-5482.

Lanyi, J. 1974. Salt dependence of proteins from extremely halophilic bacteria. Bacteriol. Rev. 38:272-290.

Lanyi, J. 1979. Physiochemical aspects of salt dependence in halobacteria. In M. Shilo (Ed.) Strategies of microbes in extreme environments. (pp. 93-107). Verlag Chemie: Weinheim, FRG.

Liu, Y. Boone, D., Choy, C. 1990. *Methanohalophilus orgonense sp. nov.*, a methylotrophic methanogen from an alkaline, saline aquifer. Int. J. Syst. Bacteriol. 40: 111-116.

Maniatis, T., Fritsch, E. F., and Sambrook, J. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbour Laboratory: Cold Spring Harbour, N.Y.

Margulis, L. 1970. Origin of eucaryotic cells. Yale Univeristy Press: New Haven, CT.

Marklund, S., and Marklund, G. 1974. Involvement of the superoxide anion radical in the autoxidation of Pyrogallol and a convenient assay for superoxide dismutase. Eur. J. Biochem. 47:469-474.

May. B. P., and Dennis, P. P. 1987. Superoxide dismutase from the extremely halophilic archaebacterium *Halobacterium cutirubrum*. J. Bacteriol. 169:1417-1422.

May. B. P., and Dennis, P. P. 1989. Evolution and regulation of the gene encoding superoxide dismutase from the archaebacterium *Halobacterium cutirubrum*. J. Biol. Chem. 264:12253-12258.

May, B., and Dennis, P. P. 1990. Unusual evolution of a superoxide dismutase-like gene from the extreme halophilic archaebacterium Halobacterium cutirubrum. J. Bacteriol. 172: 3725-3729.

May, B. P., Tam, P., and Dennis, P. P. 1989. The expression of the superoxide dismutase gene in *Halobacterium cutirubrum* and *Halobacterium volcanii*. Can. J. Microbiol. 35:171-175.

McCord, J. M., and Fridovich, I. 1969. Superoxide dismutase. An enzymic function for Erythrocuprein (Hemocuprein). J. Biol. Chem. 244:6049-6055

Mullakhanbhai, M. F., and Larsen, H. 1975. *Halobacterium volcanii* spec. nov.; a Dead Sea halobacterium with a moderate salt requirement. Arch. Microbiol. 104:207-214.

Mylvaganam, S., and Dennis, P. P. 1992. Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics, 130:399-410.

Neuman, A. 1987. Specific accessory sequences in *Saccharomyces cerevisiae* introns control assembly of pre-mRNA into apliceosomes. EMBO J. 6:3833-3839.

Ochman, H. and Wilson, A. C. 1987. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. 26: 74-86.

Oesterhelt, D.,1985., Light driven proton pumping in halobacteria. BioScience. 35: 18-21.

Oren, A., Lau, P.. P., and Fox, G. E. 1988. The taxonomic status of *Halobacterium marismortui* from the Dead Sea; a comparison with *Halobacterium vallismortis*. System. Appl. Microbiol. 10:251-258.

Parker, M. W., and Blake, C. C. 1988. Iron- and manganese-superoxide dismutases can be distinguished by analysis of their primary structures. FEBS Letters. 229:377-382.

Paterek, J. R., and Smith, P. H. 1988. *Methanohalphilus mahii* gen. nov., sp. nov., a methylotrphic halophilic methanogen. Inter. J. Syst. Bacteriol. 38:122-123.

Peterson, G. L. 1977. A simplification of the protein assay method of Lowry et al. which is generally more applicable. Anal. Biochem. 83. 346-356.

Pfeifer, F., and M. Betlach. 1985. Genome organization in *Halobacterum cutirubrum*: A 70 kb island of more (AT) rich DNA in the chromosome. Mol.Gen. Genet. 198: 449-455

Phillips, J. P., Campbell, S. D. , Michaud, D., Charbonneau, M., and Hilliker, A. J. 1989. Null mutation of copper/zinc superoxide dismutase in *Drosophila* confers hypersensitivity to paraquat and reduced longevity. Proc. Natl. Acad. Sci. USA. 86:2761-2765.

Puget, K. and Michelson, A.M. 1974. Isolation of a new copper-containig superoxide dismutase bacteriocuprein. Biochem. Biophys. Res. Commun. 58: 830-838.

Reddy, K.J., Webb, R. and Sherman, L.A. 1990. Bacterial RNA isolation with one hour centrifugation in a table top ultracentrifuge. BioTechniques. 8: 250-251.

Reiter, W. D., Palm, P., Voos, W., Kaniecki, J., Grampp, B., Schulz, W., and Zillig, W. 1987. Putative promoter elements for the ribosomal RNA genes of the thermoacidophilic archaebacterium *Sulfolobus* sp. strain B12. Nucleic Acids Res. 15:5581-5595

Rush, J.D., Maskos, Z. and Koppenol, W. 1991. The superoxide dismutase activities of two higher valent manganese complexes, Mn^{IV} desferrioxamine and Mn^{III} cyclam. Arch. Biochem.Biophys. 289: 97-102.

Saenger, W. 1987. Structure and dynamics of water surrounding biomolecules. Ann. Rev. Biophys. Chem. 16:93-114.

Saitou, N., and Nei, M. 1987. The neighbour joining method; A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Salin, M. L., and Oesterhelt, D. 1988. Purification of a manganese-containing superoxide dismutase from *Halobacterium halobium*. Arch. Biochem. Biophys. 260:806-810.

Salin, M. L., Duke, M. V., Oesterhelt, D., and Din-Pow, M. 1988. Cloning and determination of the nucleotide sequence of the Mn-containing superoxide dismutase *Halobacterium halobium*. Gene. 70:153-159.

Sanger, F., Nicklen, S., and Coulson, A. R. 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74:5463-5467.

Sapienza, C., and Ford Doolittle, W. 1982. Unusual physical organization of the *Halobacterium* genome. Nature. 295:384-389.

Sarich, V.M., and A.C. Wilson. 1973. Generation time and genomic evolution in primates. Science. 179:1144-1147.

Schiavone, J. R., and Hassan, H. M. 1987. Biosynthesis of superoxide dismutase in eight prokaryotes: effects of oxygen, paraquat and an iron chelator. FEMS Microbiol. Letters. 42:33-38.

Schnabel,H., Ziliig, W., Pfaffle, M., Schnabel, R., Michel, H.,Delius, H. 1982. *Halobacterium halobium* phage ϕH1. EMBO J. 1: 87-92.

Searcy, K. B., and Searcy, D. G. 1981. Superoxide dismutase from the archaebacterium *Thermoplasma Acidophilum.* Biochimica et Biophysica Acta. 670:39-46.

Shine, J. and Dalgarno, L. 1974. The 3' terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA. 71: 1342-1346.

Southern, E. M. 1975. Detection of specific sequences among DNA fragments seperated by gel electrophoresis. J. Mol. Biol. 98:503-517.

Stallings, W. C., Pattridge, K. A., Strong, R. K., and Ludwig, M. L. 1984. Manganese and iron superoxide dismutases are structural homologs. J. Biol. Chem. 259:10695-10699.

Stallings, W.C., Partridge, K.A., Strong, R.K., Ludwig, M.L. 1985. The structure of manganese superoxide dismutase from *Thermus thermophilus HB8* at 2.4-A resolution. J. Biol. Chem. 260: 16424-16432

Stanier, R. Y., Adelberg, E. A. and Ingraham, J. 1976. In The Microbial World. (pp. 65). Prentice -Hall, Inc. Englewood Cliffs, NJ.

Steinman, H. M. 1982. Superoxide dismutases: Protein chemistry and structure-function relatinships. In Superoxide Dismutases. Vol. 1. (pp. 11-68). L. W. Oberly (Ed.) CRC Press: Boca Raton, FL.

Takao, M., Kobayashi, T., Oikawa, A., and Yasui, A. 1989. Tandem arrangement of photolyase and superoxide dismutase genes in *Halobacterium halobium.* J. Bacteriol 171:6323-6329.

Takao, M., Oikawa, A., and Yasui, A. 1990. Characteristics of a superoxide dismutase gene from the archaebacterium *Methanobacterium thermoautotrophicum.* Arch. Biochem. Biophys. 283:210-216.

Takeda, Y.and Avila, H., 1986. Structure and gene expression of the *E.coli* Mn-superoxide dismutase gene. Nucl. Acid. Res. 14: 4577-4589.

Walter, M. R. 1983. Archean stromatolites; Evidence of the earth's earliest benthos. (pp. 187-212). In J. W. Schopf (Ed.) Earth's earliest biosphere; Its origin and evolution. Princeton University Press: Princeton, NJ.

Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. 51:221-271.

Woese, C. R., Kandler, O., and Wheelis, M. L. 1990. Towards a natural system of organisms: Proposal for the domains Archae, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. 87:4576-4579.

Wu, J., and Weiss, B. 1991. Two divergently transcribed genes, *soxR* and *soxS*, control a superoxide response regulation of *Escherichia coli*. J. Bacteriol. 173:2684-2871.

Zhang, Q., and Yonei, S. 1991. Induction of manganese-superoxide dismutase by membrane-binding drugs in *Escherichia coli*. J. Bacteriol. 173:3488-3491.

Zuckerkandl, E., and Pauling, L. 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8: 357-366.