STUDIES ON CHO hsp60 AND hsp10

GENOMIC SEQUENCES

STUDIES ON GENOMIC SEQUENCES FOR THE HEAT SHOCK PROTEINS hsp60 AND hsp10 FROM CHINESE HAMSTER OVARY CELLS

By

Joni Zurawinski, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Masters of Science

McMaster University

MASTER OF SCIENCE (1997) McMaster University
(Biochemistry) Hamilton, Ontario

TITLE: Studies on genomic sequences for the heat shock proteins hsp60 and hsp10 from Chinese hamster ovary cells

AUTHOR: Joni Zurawinski, B.Sc. (McMaster University)

SUPERVISOR: Professor R.S. Gupta

NUMBER OF PAGES: x, 90

ABSTRACT

Although the cDNA sequences for the 10 kDa (hsp10, hsp10) and the 60 kDa (hsp60, cpn60) heat shock proteins have been obtained for a number of mammalian species, until very recently information was not available on the functional genes encoding these proteins. The primary objective of this work was to clone and sequence the functional genes for these proteins from CHO, Chinese hamster ovary cells. Screening of a lambda EMBL3 CHO genomic library with the CHO hsp10 cDNA identified a clone containing the putative hsp10 functional gene. A ~5.5 kb fragment was isolated from one of these clones by enzymatic digestion and ~3.3 kb was sequenced. The clone was found to contain consensus regulatory sequences upstream of the putative transcription initiation site, +1, including two Sp1 binding sites, a CAAT box, and a single heat shock element, HSE, but lacked a TATA box. The coding region consists of four exons, identical to the hsp10 CHO cDNA sequence, separated by three introns, of 200 bp, 600 bp and 1600 bp in size, containing conserved splice sites.

Screening of the same EMBL3 CHO genomic library with the CHO hsp10 cDNA also resulted in isolation of a full length processed pseudogene with ~90 % identity to the cDNA. This pseudogene lacked introns, contained a poly(A) tract, as well as various single bp changes, additions and deletions. The upstream region of this pseudogene was found to contain similarity to the human LINE sequence, a DNA repetitive element. PCR amplification of CHO-WT genomic DNA resulted in isolation of five additional processed pseudogenes,

corresponding to the central ~270 bp of the CHO hsp10 cDNA. All the pseudogenes displayed a high degree of similarity to the CHO hsp10 cDNA sequence despite the presence of numerous mutations. Prior to this report, pseudogenes had not been found associated with hsp10. The identification of these pseudogenes suggests the presence of a multigene family for this heat shock protein in the CHO genome.

Previously, a semi-processed pseudogene, GC1, was identified for hsp60 from CHO cells which contained a single ~87 bp intron near its 3' end (Venner *et al.*, 1990). From this pseudogene, a fragment containing the ~87 bp intron was isolated for use as a probe to screen a lambda EMBL3 CHO genomic library. This resulted in isolation of several positive clones, two of which were purified, a ~1.0 kb fragment amplified by PCR and then sequenced revealing two additional semi-processed pseudogenes, designated λ4 and λ5. These pseudogenes were found to be homologous to the GC1 clone, containing many similar mutations as well as the ~87 bp intron. Utilizing CHO-WT genomic DNA, a separate PCR amplification resulted in isolation of a ~2.5 kb fragment which was partially sequenced and found to correspond to the putative hsp60 functional gene. The fragment contained one exon, which was identical to the CHO hsp60 cDNA in the region sequenced, and two introns of 800 bp and 1500 bp. This fragment can now provide an ideal probe for isolation of the CHO hsp60 functional gene.

ACKNOWLEDGEMENTS

I would first like to thank Dr. Gupta for acting as my supervisor, not only in the lab, but through his guidance and patience during many difficult times in this project.

I thank Dr. Singh for sharing his expertise and knowledge providing insight into alternative directions for my project.

I am grateful to McMaster University for providing me the financial support through scholarships allowing me to pursue my Masters of Science.

I am indebted to Dr. Wu, Aiping and Kevin for spending many hours of their time in offering technical support.

Special thank-yous to Jonathan for time spent in the darkroom helping to produce the wonderful figures in this thesis, and to Dr. Wright's lab for various "borrowed" materials.

To my fellow lab mates, Jonathan, Mary and Gillian, you have all been wonderful to work with and great friends whom I will always remember.

To my parents, you have always been there for me offering words of encouragement and love, your support has meant the world to me.

And especially to Shawn... for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ml	milliliters
mg	milligrams
Tris	tris (hydoxymethyl) aminomethane
EDTA	ethylenediamine tetraacetic acid
NaCl	sodium chloride
SDS	sodium dodecyl sulphate
vol	volumes
min	minutes
UV	ultraviolet
nm	nanometer
E.coli	*Escherichia coli*
LB	Luria-Bertani medium
μg	micrograms
mM	millimolar
M	molar
rpm	revolutions per minute
μl	microliters
NaOH	sodium hydroxide
DNA	deoxyribonucleic acid
cDNA	complementary DNA
RNA	ribonucleic acid
mRNA	messenger ribonucleic acid
tRNA	transfer ribonucleic acid
RNase A	ribonuclease A
hrs	hours
ddH_20	distilled deionized water
g	grams
kDa	kiloDalton
Mg^{2+}	magnesium ion
KCl	potassium chloride
Na_2HPO_4	disodium hydrogen orthophosphate
K_2HPO_4	dipotassium hydrogen orothophosphate
PCR	polymerase chain reaction
NaI	sodium iodide
TAE	Tris-acetate buffer
HCl	hydrochloric acid

$CaCl_2$	calcium chloride
$MgCl_2$	magnesium chloride
ATP	adenosine triphosphate
ADP	adenosine diphosphate
dNTPs	deoxynucleotide triphosphates
Taq	*Thermus aquaticus* (DNA polymerase)
Vent	*Thermococcus litoralis* (Vent DNA polymerase)
pmol	picomole
ng	nanogram
sec	seconds
cpm	counts per minute
SSC	sodium chloride/sodium citrate solution
SM	lambda dilution buffer
CHO	Chinese hamster ovary cells
ER	endoplasmic reticulum
HSF	heat shock factor
hsp	heat shock protein
HSE	heat shock element
cpn	chaperonin
Pod^R	podophyllotoxin resistant
pfu	plaque forming unit

1.0 INTRODUCTION

1.1 Heat Shock Proteins and the Heat Shock Response

A large group of proteins are present in the cell designed to provide protection against various forms of stress including sudden temperature increases, anoxia, heavy metals, as well as various teratogens, and mutagens. Collectively, these proteins are known as heat shock proteins or hsps. Studies involving numerous organisms have revealed that heat shock proteins are ubiquitous and their response to stress is universal and complex. The most extensively studied stress response of these proteins involves their ability to protect the cell from sudden increases in temperature; to provide the cell with thermotolerance. The response to sudden increases in temperature is rapid, producing large amounts of various heat shock proteins within minutes. Their synthesis is maintained as long as the heat shock persists and decreases slowly once conditions return to normal. These proteins can also provide protection against environmental stresses, because they possess the ability to bind to disorganized or partially unfolded regions of proteins, preventing their aggregation. As an extension of this binding, they also aid in restoring proteins to their native states. Additionally, this large group of proteins has also been shown to play a vital role in ordinary cellular functions involving protein synthesis and translocation. The heat shock proteins have been divided into families based on their molecular weights and their degree of sequence conservation. In terms of molecular weight, these families range from the small hsps of 10-30

kDa, to the largest of 80-110 kDa. Within this broad range, the most predominant classes are the hsp60, hsp70 and hsp90 proteins (see reviews by Lindquist, 1986; Lindquist and Craig, 1988; Becker and Craig, 1994; Buchner, 1996).

Heat shock proteins were first discovered during studies involving *Drosophila busckii* over twenty years ago. When the fruit fly larvae were subjected to heat shock puffs became visible on the chromosomes from salivary glands and other tissues within minutes (Tissiéres *et al.*, 1974). Further studies utilizing cultured *Drosophila* cells linked the appearance of these puffs with the production of new proteins, later identified as the first heat shock proteins (Lindquist- McKenzie *et al.*, 1975).

Control of the heat shock response occurs mainly at the transcriptional level. In eukaryotes, the heat shock element, HSE, is located upstream of the functional gene sequence and in some cases, within the first intron (Bienz, 1985; Rebbe *et al.*, 1989). The HSE consensus sequence consists of the repeat unit, GAA, commonly found as GAA--TTC or TTC--GAA, and is similar across species, from yeast to human, present in multiple copies (Amin *et al.*, 1988; Xiao and Lis, 1988). This element is bound by a transcription factor which becomes active in response to heat shock, in some cases by a post-translational modification, such as phosphorylation in yeast (Wu *et al.*, 1987; Sorger and Pelham, 1988). This trans-acting factor is known as heat shock factor, HSF, and in some higher eukaryotes it is present in multigene families (Schuetz *et al.*, 1991; Nakai *et al.*, 1997). In the more complex genomes of eukaryotes, regions on chromosomes containing heat shock proteins have an open structure for easier access by transcriptional machinery (Wu, 1980).

Additionally, heat shock mRNAs are translated more efficiently and as a priority during heat shock providing a translational regulation mechanism (Storti *et al.*, 1980; Lindquist, 1981).

In prokaryotes, regulation primarily involves control of the sigma factor, although upstream heat shock consensus sequences have also been identified by comparing a number of bacteria (Cowing *et al.*, 1985; Narberhaus and Bahl, 1992). In *E.coli*, the sigma^{32} factor itself is stabilized in response to heat shock, and synthesis of this 32 kDa protein is elevated by increases in both translation of its mRNA and transcription of its gene (Tilly *et al.*, 1986; Straus *et al.*, 1987). When conditions return to normal in the cell, the sigma^{32} factor becomes inactivated resulting in a decrease in heat shock protein production (Craig and Gross, 1991).

In both prokaryotes and eukaryotes, proteins from the hsp70 family have been suggested to be responsible for sensing temperature changes in the cell, to bring about their own induction upon heat shock or other stresses. A negative feedback model proposed by Craig and Gross, (1991), suggests that the hsp70 proteins are involved in the control of transcription factors, such as sigma^{32} in *E.coli* or HSF in eukaryotes. Upon heat shock, free levels of hsp70 proteins decrease as they bind an increasing population of denatured proteins. Interactions of hsp70 with the transcription factors decreases, allowing them to bind to DNA, increasing the production of the heat shock proteins. Upon a decrease in temperature and return to normal cellular conditions, free hsp70 proteins can rebind to the transcription factors resulting in a decrease in their own production.

As mentioned, an interesting property of the heat shock proteins is their high degree of sequence conservation between different species. For example, the hsp70 family, one of

the most abundant, has amino acid sequence identity between any two species of at least 45% with many changes constituting conservative substitutions (Lindquist and Craig, 1988; Gupta and Golding, 1993). Similarly, the hsp90 protein sequence demonstrates at least 50% identity when compared between any two eukaryotes and more than 40% when compared with *E. coli* (Bardwell and Craig, 1987; Lindquist and Craig, 1988). The smaller heat shock proteins demonstrate less sequence conservation between organisms, but show regions of high homology and maintain a certain degree of relatedness in terms of their structures (Lindquist, 1986; Lindquist and Craig, 1988).

Thus, these proteins first identified by their specialized role in the heat shock response, are now known to be constitutively expressed in cells where they are involved in folding of newly synthesized proteins leading to their dual designation as molecular chaperones.

1.2 Molecular Chaperones

The discovery of molecular chaperone proteins challenged the long standing theory that proteins contained all the necessary information for their spontaneous folding within their primary amino acid sequence (Anfinsen, 1973). In a review by Rothman, (1989), he supported the theory that polypeptide binding proteins or molecular chaperones are catalysts for many cellular processes involving proteins in the cell, including folding, assembly, translocation and repair. Additionally, he proposed specific properties that a protein should possess to be considered a molecular chaperone. These proteins must be able to bind to

regions of proteins that have become unfolded, the binding must be sufficiently strong in order to prevent aggregation, but also transient enough to allow for release once a native conformation has been achieved, and finally this release must utilize energy, usually in the form of ATP. Furthermore, proteins acting as molecular chaperones only aid in the folding process and do not become part of the final product (see reviews by Ellis, 1987; Rothman, 1989; Becker and Craig, 1994).

The first protein to be classified as a molecular chaperone was the nuclear protein nucleoplasmin, which is responsible for assembly of nucleosomes (Laskey *et al.*, 1978). It functions by binding to histone proteins thereby aiding in their association with each other and with DNA and at the same time preventing non-specific aggregation. However, the nucleoplasmin protein does not remain as part of the final nucleosome structure (Laskey and Earnshaw, 1980). Similar studies in plants revealed the existence of a chloroplast protein that functions in the assembly of the ribulose bisphosphate carboxylase-oxygenase enzyme, Rubisco. This enzyme is composed of two types of subunits, which are combined into a large oligomer of ~500 kDa. The larger of the two subunits is synthesized inside the chloroplast aided by the molecular chaperone, Rubisco subunit binding protein, while the smaller subunits are translated on cytoplasmic ribosomes. Eight of each type of monomer subunit is combined within the chloroplast to yield the holoenzyme (Barraclough and Ellis, 1980; Musgrove *et al.*, 1987). Numerous other studies concerning these molecular chaperone proteins have demonstrated their presence in all cells, with their major classifications also corresponding to the major groups of heat shock proteins, specifically the hsp60, hsp70 and hsp90 families.

However, since the main focus of this thesis is on the hsp60 and hsp10 families of proteins, further discussion is restricted to only these proteins.

1.3 GroEL/hsp60 Chaperonin Family

Protein from this family was initially identified in *E.coli* while studying the growth of different types of bacteriophage. Using mutant cells, Georgopoulos *et al.*, (1973), and Zweig and Cummings, (1973), found this protein to be essential for different aspects of bacteriophage growth, such as lambda head assembly, and T5 tail assembly. This protein has a molecular weight of 65 kDa and is referred to in *E.coli* as GroEL. It is constitutively expressed, but its amount is greatly increased upon heat shock, allowing it to bind misfolded or denatured proteins, preventing aggregation and aiding in their refolding to the native state. This family of proteins has also been shown to be essential for normal cellular growth because of their role in protein folding and assembly of oligomeric protein complexes. Hence the terms heat shock protein and molecular chaperone have become synonymous with GroEL and related proteins. The term "chaperonin" was first proposed by Hemmingsen *et al.*, (1988), and is used to describe this ubiquitous family of proteins which are found to perform the same functions across numerous species. In addition to their conservation in function, hsp60 homologs are very similar in terms of their structure. When visualized by electron microscopy, the quaternary structure of GroEL consists of two stacked layers of seven 60 kDa monomer subunits, each layer organized into a ring, leaving a hollow interior cavity. Mitochondrial hsp60 proteins of eukaryotes, and the Rubisco large subunit binding protein

of plant chloroplasts also possess this toroidal structure (see reviews by Lindquist, 1986; Lindquist and Craig, 1988; Rothman, 1989; Buchner, 1996).

In eukaryotic species, a homolog for the GroEL protein was first identified in the protozoan *Tetrahymena*. This protein was induced upon heat shock, but was also present under normal cellular conditions (McMullin and Hallberg, 1987). Antibodies were raised against this protein and used to investigate other eukaryotic species for the presence of GroEL homologs. These homologs were found in every eukaryotic organism studied, mainly within the mitochondria, and designated hsp60 or cpn60 (Lindquist and Craig, 1988). Experiments conducted in temperature sensitive yeast, by Cheng *et al.*, (1989), confirmed the essential nature of hsp60 proteins for cellular growth, even under normal conditions. The mutant yeast were found to contain an alteration in the nuclear gene mif4, later identified as hsp60. One of the originally characterized molecular chaperones, the Rubisco subunit binding protein from plant chloroplasts, has also been identified as a GroEL homolog, demonstrating amino acid sequence identity of 46% plus additional conservative substitutions when compared to GroEL (Hemmingsen *et al.*, 1988). In eukaryotes, hsp60 proteins are found mainly within mitochondria; however, the protein Tcp-1, which is distantly related to hsp60 and appears to carry out a similar function, is present within the cytoplasm (Gupta, 1990).

In prokaryotes, GroEL homologs are organized in the genome as part of an operon with a second gene encoding a 10 kDa protein, designated GroES. Induction of both GroEL and GroES occurs efficiently from this operon upon heat shock or the addition of other cellular stresses (Yamamori and Yura, 1982). The presence of more than one GroEL

homolog has been found for a number of different bacteria, including representatives from cyanobacteria, Gram-positive and Gram-negative bacteria (Mazodier *et al.*, 1991; Fischer *et al.*, 1993). Interestingly, the hsp60 protein from a number of pathogenic bacteria, such as *Mycobacterium tuberculosis, Mycobacterium leprae,* and *Coxiella burnetii,* has been shown to be a major antigenic protein of these organisms. This has implicated the involvement of immune responses to hsp60 in a number of autoimmune diseases (Young, 1990; Cohen, 1991).

The crystal structure of GroEL has recently been solved to 2.8 Å, confirming the existence of a ring shaped organization for this protein as shown in Fig. 1 (Braig *et al.*, 1994). The amino acid residues which line the central cavity and the opening to the central cavity are hydrophobic, which are believed to be the sites of interaction of the GroEL complex with target proteins whose hydrophobic residues would be exposed (Fenton *et al.*, 1994). The central cavity is rather large, estimated to hold proteins of up to ~90 kDa (Christensen and Pain, 1991). This region, also referred to as the apical domain or peptide binding domain, is also the region for GroES binding. The equatorial domain is located closer to the N-terminus of the amino acid sequence, but is more centrally located in the ring structure. This has been designated as the ATP-binding site. These two regions are connected by an intermediate domain (Braig *et al.*, 1994). The majority of hsp60 proteins have been found to contain a conserved C-terminal amino acid repeat motif of GGM or GGGM, but the functional or structural significance of this repeat is not known (Gupta, 1996; McLennan *et al.*, 1993). Eukaryotic hsp60 sequences also contain a typical conserved leader sequence at their N-

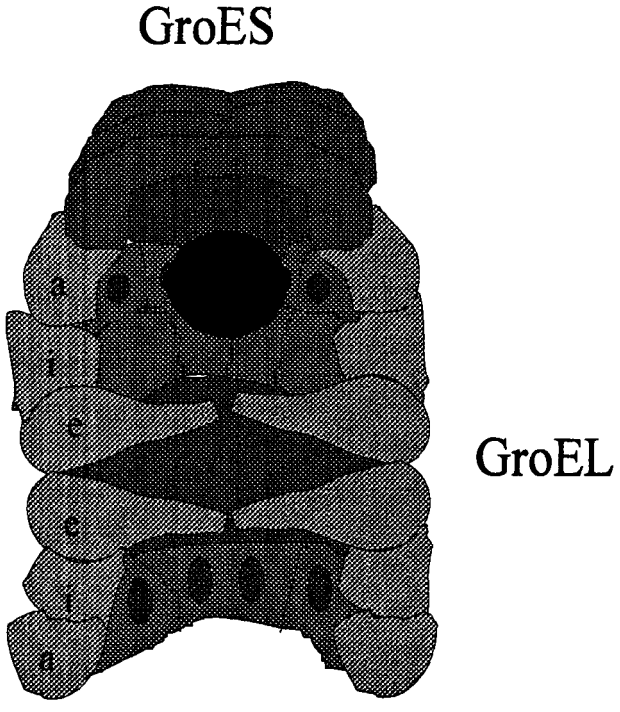

Figure 1 **Schematic representation of the GroEL/GroES complex from their crystal structures**

Domains of the GroEL complex are labelled as a, apical domains, i, intermediate domains, and e, equitorial domains. Polypeptide bound within the complex is represented as a black circle. Black patches in the apical domains represent patches of hydrophobic residues. Both complexes are shown as cross-sections cut through the center.
Figure was adapted from Mayhew and Hartl, 1996.

termini for import into mitochondria (Schatz, 1987; Yoshida *et al.*, 1995). The ATPase function of the hsp60 proteins is an important requirement, along with the cofactor GroES or hsp10, for their function in protein folding (Lubben *et al.*, 1990; Becker and Craig, 1994).

The cDNA clones for hsp60 have been isolated from a number of different eukaryotic species, including, Chinese hamster ovary (CHO), rat, mouse, and human, (Picketts *et al.*, 1989; Venner and Gupta, 1990a; Venner and Gupta, 1990b; Jindal *et al.*, 1989), as well as the genomic clone from yeast (Yoshida et al., 1995). However, at the onset of this project, the genomic clone for hsp60 had not been isolated from any mammalian species. Because of hsp60's abundance in the cell, its role in the cellular stress response, its function in protein folding and assembly, and its suggested involvement in autoimmune diseases, the isolation of a functional gene for hsp60 became of great interest. Venner *et al.*, (1990), attempted to isolate genomic hsp60 from both human and CHO cells. Southern analysis of human and CHO genomic DNA revealed the existence of multiple copies of hsp60, with an initial estimation of 8-12 copies present for mammalian genomes. Further investigation of clones isolated from genomic library screens, using cDNA probes, revealed the presence of numerous nonfunctional hsp60 related genes for both CHO and human. These were classified as pseudogenes for hsp60 based on the presence of various mutations, including deletions, additions and base pair substitutions, rendering them unable to encode functional protein.

Currently it is estimated that there are well over 30 pseudogenes for hsp60 in mammalian genomes, creating a large multigene family of pseudogenes for hsp60 (Venner *et al.*, 1990; Pochon and Mach, 1996; Gupta, unpublished). Pseudogenes appear to be

restricted to mammalian genomes, with the most common type lacking introns, known as the processed pseudogene (see review by Vanin, 1985). Despite numerous mutations, the pseudogene sequence maintains a high degree of similarity towards the cDNA of its functional gene (Vanin, 1985; Venner *et al.*, 1990; Pochon and Mach, 1996). This creates difficulties when attempting to isolate the functional gene through genomic library screens using the cDNA as probe. Binding of the cDNA probe may occur equally well to both the functional gene and any pseudogene sequences present. Therefore, based on the ratio of pseudogenes to functional gene and the lack of introns for most pseudogenes, the probability of the probe binding to pseudogenes becomes much higher.

1.4 GroES/hsp10 Chaperonin Family

The GroEL protein from *E.coli* had initially been identified as being essential for growth of various bacteriophages. During these same studies by Georgopoulos *et al.*, (1973), and Zweig and Cummings, (1973), a second protein, designated GroES, was also found to be necessary for bacteriophage replication functioning as a cofactor for GroEL. The GroES protein was smaller than GroEL with a molecular weight of 15 kDa. Similar to GroEL's dual designation, GroES is both a heat shock protein and a molecular chaperonin. Functioning as a cofactor for GroEL, it carries out an important role for cell survival under stressful conditions, as well as in protein folding and assembly under normal conditions. GroES homologs have also been identified in eukaryotes, commonly referred to as hsp10 or cpn10, based on their molecular weight of approximately 10 kDa and their role as co-chaperonins for

the eukaryotic GroEL homologs, hsp60. In many different eukaryotes, including yeast, the hsp10 protein is localized within mitochondria which is consistent with the localization of hsp60. Hsp10 homologs have also been identified in higher plants within both the chloroplasts and mitochondria. Interestingly, the chloroplast homolog has a mature molecular weight of approximately 24 kDa, and it consists of two bacterial GroES-like proteins organized as a single protein, clearly distinguishable from plant mitochondrial hsp10 (Bertsch *et al.*, 1992). Unlike hsp60, which has a distantly related cytoplasmic homolog Tcp-1, hsp10 homologs have not been identified within the eukaryotic cytosol. The hsp10 family of proteins are not as highly conserved as the hsp60 proteins. Comparison of hsp10 sequences from different species by Gupta, (1995), revealed a minimum identity of 30% with additional 15-20% residues as conservative substitutions. As demonstrated by Hartman *et al.*, (1992), this conservation is also extended towards the function of these chaperonins. Isolated rat hsp10 protein was able to substitute for bacterial GroES and participate in the *in vitro* folding of rat OTC along with GroEL (see reviews by Lindquist, 1986; Lindquist and Craig, 1988; Rothman, 1989; Gupta, 1995).

Isolation of the hsp10 functional gene from the yeast, *S. cerevisiae*, by Hohfeld and Hartl, (1994), allowed for study of the role of this protein as a co-chaperonin. Deletion of the hsp10 gene produced non-viable cells providing evidence for the essential function of hsp10, specifically for folding of mitochondrial proteins, even under normal cellular conditions. Isolated mitochondria, from a temperature sensitive yeast defective in hsp10, were used to study the effect of folding and assembly of mitochondrial matrix proteins. The

α-matrix processing protease and rat ornithine transcarbamylase proteins were found to be incorrectly folded and nonfunctional. This was due to decreased interaction of the mutant hsp10 with hsp60 within the mitochondria. However, one protein tested, dihydrofolate reductase, DHFR, targeted to mitochondria as a fusion protein, was still correctly folded in the mutant cells. This finding supported previous *in vitro* observations that hsp10 may not be necessary as a cofactor for hsp60 in the folding of every protein (Viitanen *et al.*, 1991; Hohfeld and Hartl, 1994).

The first mammalian hsp10 homolog isolated was from rat hepatoma cultured cells by Hartman *et al.*, (1992). This protein was shown to be hsp10 related based on its high degree of amino acid sequence identity, at least 45% when compared to various prokaryotic sequences, and is essential for the *in vitro* folding of rat ornithine transcarbamylase, OTC, in conjunction with bacterial GroEL and ATP. At the genetic level, cDNA sequences for mammalian hsp10 have been isolated from various sources including, rat, CHO, bovine and human (Ryan *et al.*, 1994; Gupta, unpublished, Pilkington and Walker, 1993; Chen *et al.*, 1994). Comparison of amino acid sequence identity demonstrates the high degree of conservation for this protein in mammals. Human hsp10 and bovine hsp10 are exactly the same, while they differ only by one amino acid compared with rat hsp10. At the nucleotide level, the sequence identity is very extensive with human and bovine of ~93% and human and rat of ~90% (Chen *et al.*, 1994). Unlike hsp60 from mammalian cells, pseudogenes have not been identified and reported for hsp10 to date. However, based on its close association with hsp60 in terms of function and its shared characteristics as part of the chaperonin family, it

is probable that pseudogenes are present in mammalian genomes, such that hsp10 would also be part of a large multigene family (Gupta, unpublished).

Crystallographic studies utilizing the GroES proteins from *M. leprae* and *E.coli* have resulted in the solving of the structure of GroES from both sources, at 3.5 Å and 2.8 Å, respectively, as illustrated in Fig. 1 (Mande *et al.*, 1996; Hunt *et al.*, 1996). Similar to GroEL, GroES forms a dome-like ring structure from seven monomer proteins through hydrophobic interactions. The inner side of the dome contains hydrophobic residues, while the opening at the center of the dome is lined with negatively charged residues. A flexible region, consisting of ~16 amino acid residues, is present near the N-terminus of the primary amino acid sequence. This is known as the mobile loop and is thought to be important for GroES interactions with GroEL (Mande *et al.*, 1996; Hunt *et al.*, 1996; Mayhew and Hartl, 1996).

Studies on purified GroES from *E.coli* have demonstrated the importance of both the C-terminal and N-terminal regions of the protein. Digestion of GroES with carboxypeptidase Y to remove seven C-terminal amino acids resulted in the inability of the monomer proteins to correctly assemble into the necessary dome-like ring structure (Seale and Horowitz, 1995). Similarly, proteolysis of the first 20 amino acids from the N-terminus of GroES removes the ability of the monomers to assemble, as well as disrupting the ability of GroES to interact with GroEL (Llorca *et al.*, 1997).

In eukaryotes, the N-terminal region of hsp10 does not contain a mitochondrial target sequence like other mitochondrial imported proteins. Instead, the initiator methionine residue

of the hsp10 protein is removed and the N-terminus acetylated (Ryan *et al.*, 1994). Experiments by Ryan *et al.*, (1995), demonstrated that lack of N-terminal acetylation does not affect import of the protein into mitochondria, its ability to assemble into its heptameric structure or its activity as a molecular chaperone. However, hsp10 protein lacking this N-terminal acetylation was degraded very quickly, by an unidentified extramitochondrial protease source. Thus, acetylation may not affect function or structural organization for hsp10 present in mitochondria, but it does become important for the persistence of hsp10 outside of mitochondria. Recently hsp10 has been shown to correspond to the early pregnancy factor, EPF. This is a growth factor which is important for normal fetal development through exertion of its immunosuppressive properties (Cavanagh and Morton, 1994). The acetylation of the N-terminus of hsp10 would prolong its life when acting as the EPF which is secreted from cells (Ryan *et al.*, 1995).

1.5 Mechanism of Protein Folding and Assembly for GroEL/GroES

The mechanism by which the hsp60 family of proteins and their cofactors, the hsp10 family, carry out protein folding and assembly has mainly been studied using the *E.coli* proteins, GroEL and GroES. The acquisition of the 3-dimensional structures of both these proteins by X-ray crystallography, in conjunction with electron microscopy and biochemical studies has lead to the development of a detailed model for their protein folding cycle as outlined in Fig. 2 (see reviews by Becker and Craig, 1994; Hartl *et al.*, 1994; Buchner, 1996). Proteins that are misfolded or denatured may still contain some secondary or weak tertiary

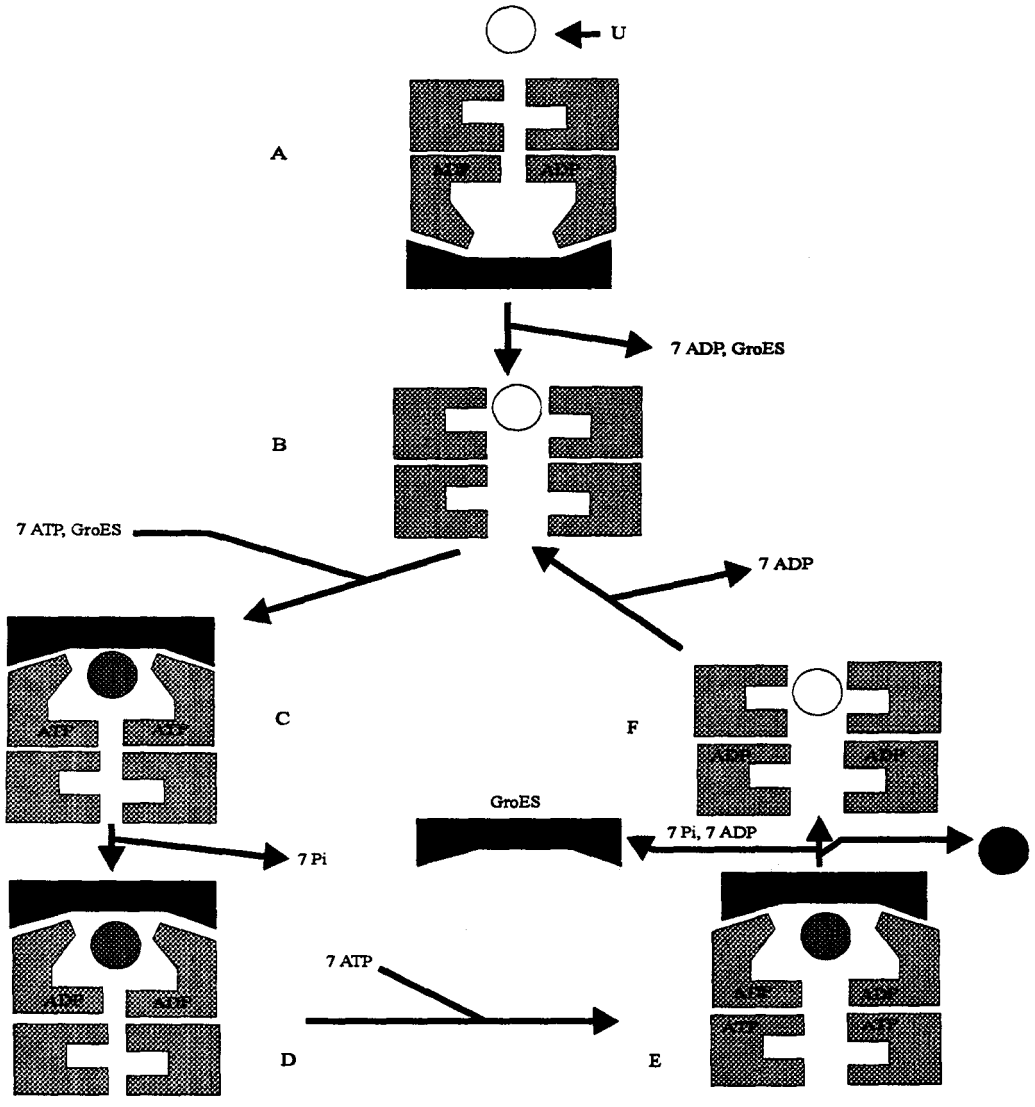

Figure 2 **Schematic representation of the protein folding mechanism of GroEL and GroES**

Shaded structures represent the GroEL complex visualized as a side view cut through the center, and black structures the GroES complex. White circles correspond to unfolded or denatured polypeptide, labelled as U in the first stage of the cycle. Shaded circles correspond to a mixture of folded and unfolded proteins for a population of GroEL complexes, the black circle represents folded protein. Conformational changes to the GroEL complex correspond to changes in GroES binding and ATP hydrolysis. Stages of the reaction cycle are labelled A-F. Polypeptide which has not reached the native state at the end of the cycle is re-bound by other GroEL complexes to repeat the cycle.
Figure was adapted from Mayhew *et al.*, 1996.

structure and have been referred to as in a "molten-globule"-like state by Martin *et al.*, (1991). Proteins in this state generally will have exposed hydrophobic regions with which they can interact with the hydrophobic regions near the top of the GroEL ring structure (Martin *et al.*, 1991; Christensen and Pain, 1991). Up to two GroES heptamers may bind to GroEL for protein folding, one at either end of GroEL, also requiring Mg^{2+}-ATP or Mg^{2+}-ADP (Llorca *et al.*, 1994; Azem *et al.*, 1994). The binding of non-native protein to GroEL usually occurs at the opposite end to which GroES is bound, causing GroES and ADP to be released (Martin *et al.*, 1993). GroES rebinds to the end of the GroEL stacked structure, mainly through interactions involving the mobile loop region with the hydrophobic ring exterior of GroEL, (Mande *et al.*, 1996), to begin the folding process. If binding of GroES occurs on the same side as the polypeptide, the result is a capping effect that forces the substrate protein into the interior cavity of GroEL. Binding of GroES to the opposite end, may result in premature release of the polypeptide. The binding of GroES also results in conformational changes to GroEL that may be translated to the protein within the cavity to promote its folding (Mayhew *et al.*, 1996). GroES also coordinates the ATPase activity of GroEL and thus can increase the rate of folding (Hartl *et al.*, 1994). Once a polypeptide has bound to GroEL, and ADP and GroES dissociate, the nucleotide binding sites on the GroEL monomers are then available for ATP binding. Seven ATP molecules bind, followed by the re-binding of GroES. Hydrolysis of these ATP molecules affects the GroEL conformation, returning it to the ADP bound state which has a lower affinity for polypeptide binding (Todd *et al.*, 1994; Mayhew *et al.*, 1996). Subsequently, GroES dissociates from GroEL along with

ADP and the substrate protein. Binding and hydrolysis of ATP molecules in the opposite ring can also aid in polypeptide release. If the protein has not yet reached its native state or has not attained a structure committed to fold, than it may be rebound by GroEL to repeat the cycle (Mayhew *et al.*, 1996).

The different molecular chaperones involved in protein folding do not function in isolation, but rather they are part of protein folding pathways. For example, hsp70 within mammalian mitochondria binds proteins as they are translocated into the organelle, preventing their aggregation so that they may bind to the hsp60/hsp10 complexes for productive folding to their native state within the matrix (Langer *et al.*, 1992). Although the majority of studies concerning the chaperonin protein folding mechanism have been carried out on GroEL and GroES, the same basic mechanism can be applied for the eukaryotic homologs as well (Staniforth *et al.*, 1994; Tian *et al.*, 1995). However, many questions remain to be answered in order to achieve a more in depth understanding of this complex protein folding mechanism.

1.6 Pseudogenes and DNA Repetitive Elements

Studies on *Xenopus laevis* twenty years ago by Jacq *et al.*, (1977) resulted in the isolation of a sequence for 5S RNA which was very similar to the functional gene, but contained numerous base pair changes. This 5S RNA-related sequence was termed a pseudogene, suggesting that this gene was very close to the functional sequence, but was mutated so that it could not encode a functional protein. Many other pseudogenes have since been isolated for numerous genes including, ribosomal proteins, structural proteins,

immunoglobulin proteins, metabolic enzymes and of particular interest to this project, heat shock proteins. The extensive isolation of pseudogenes has provided a large database for use in further characterizing these genomic elements in attempts to understand their possible function and origin in the genome. To date, the occurrence of pseudogenes has mainly been restricted to mammalian genomes. Wagner, (1986), presents two possible explanations for this phenomenon; simply that the mammalian genome is larger than the genomes of other organisms which tolerates the presence of non-coding excess DNA and that perhaps mammals have an early developmental stage that can be easily subjected to retroviral infection resulting in the production of pseudogenes. The majority of genes that have pseudogenes associated with them belong to the designation of "housekeeping" genes, or genes that encode proteins essential for cellular function. The majority of genes expressed in germ cells are those essential for cellular function, in other words, housekeeping genes. Therefore, genes that are transcriptionally active in germ cells have the ability to gain pseudogenes. Additionally, the number of pseudogenes appears to vary for each gene family. Some genes are found to be associated with two or three pseudogenes, while other families are rather large. For example, human β-actin and rat α-tubulin, both have at least 20 pseudogenes, (Weiner *et al.*, 1986; Leavitt *et al.*, 1984), mouse and rat cytochrome c have at least 30, (Weiner *et al.*, 1986), and CHO hsp60 has over 30 pseudogenes (Gupta, unpublished).

Comparison of numerous pseudogenes has resulted in identification of shared structural characteristics. The most abundant type of pseudogene is known as a processed pseudogene. Processed pseudogenes are very similar to the mRNA copy of the functional

gene in terms of structure; they lack introns, have polyadenylation signals at their 3' ends followed by a poly(A) tail, are flanked by direct repeats, and are almost always full length copies. In addition, they usually contain numerous mutations, such as single base pair changes, deletions, additions, and frameshift mutations, all of which result in premature stop codons rendering them unable to encode a functional protein. The presence of flanking direct repeats, usually between 7-17 bp in length, mark the end of homology between the functional gene and the pseudogene. Direct repeats are also found associated with other genomic elements such as, LINEs and the Alu family, which led to the suggestion by Vanin (1985) that pseudogenes may arise by a similar mechanism as for these genomic elements.

A second category of pseudogenes are also present in the genome which differs from processed pseudogenes by their retention of one or more introns that are found within the functional gene. These are referred to as semi-processed pseudogenes, and most probably arise via gene duplication rather than via a retroinsertion mechanism (Dyer *et al.*, 1989; Ueyama *et al.*, 1993). These pseudogenes are not abundant, for example, the CHO hsp60 gene was found by Venner *et al.*, (1990), to have only one related semi-processed pseudogene, which contained an 87 bp intron near its 3' end, however, as mentioned, it is estimated that there are over 30 processed pseudogenes.

The exact mechanism for the production of pseudogenes is not known, but two models have been put forth based on the information gathered thus far. Both of these mechanisms involve an RNA intermediate, but differ by the mechanism of insertion into the genome. The first model suggested by Moos and Gallwitz (1983), requires the presence of

a reverse transcriptase, RT, which could be provided by a retroviral infection or endogenous RT, resulting in reverse transcription of the housekeeping mRNA, producing a cDNA copy which inserts randomly into A-rich regions on a different chromosome from the functional gene. Alternatively, Vanin (1985), suggests that a direct insertion of mRNA in an A-rich region in the genome occurs followed by copying to DNA during replication by DNA polymerase or by a RT. Semi-processed pseudogenes are usually closely linked to the functional gene on the same chromosome, suggesting gene duplication as their main mechanism of propagation. Pseudogenes usually become inactive at the time of random insertion, since the probability that they would insert behind a functional RNA pol II promotor is low. It was suggested by Fotaki and Iatrou (1993) and others, that pseudogenes are not just non-functional DNA elements present in the genome but, in some cases may actively participate in the evolution of genes through involvement in gene conversion and duplication, providing a potential diversity. (Jackers *et al.*, 1996; Varga-Madrazo *et al.*, 1995). For example, the immunoglobulin proteins in humans contain about 30% of their sequences as pseudogenes. It is suggested that the presence of these related sequences to the functional genes provides more diversity to these proteins, allowing for greater variety in immune responses (Varga-Madrazo *et al.*, 1995). Although the exact mechanism for pseudogene formation is not known and their potential purpose in the genome can only be speculated, it is reasonable to assume that they may have a role in increasing the fluidity of the genome by absorbing mutations and providing genetic diversity (see reviews by Vanin, 1985; Weiner *et al.*, 1986; Wagner, 1986).

Interestingly, several examples have been reported for the association of repetitive DNA elements, such as LINEs with pseudogenes (Vilotte *et al.*, 1993; Ardehali *et al.*, 1995; Foord *et al.*, 1996; Gupta, unpublished). Long interspersed repetitive DNA elements, or LINEs, are a very abundant family of retropseudogenes, with over 100,000 copies throughout the mammalian genome, within introns, flanking genes and in centromeric DNA. Similar elements are found in other organisms, which arose from a common ancestral LINE, designated L1. A full length LINE is approximately 6-7 kb, but the majority of these elements are truncated to varying degrees at their 5' ends as well as containing deletions, inversions, and single base pair changes. Two open reading frames are present in a full length LINE, separated by an inter-ORF region which contains numerous stop codons. The first ORF is capable of encoding a protein of 40 kDa, which currently has not been identified. The second ORF sequence contains a high degree of homology to reverse transcriptase proteins, and has been shown in some cases to encode an active reverse transcriptase (Singer *et al.*, 1993). The 3' end usually contains a poly(A) signal, followed by an A-rich region, while the 5' end upstream region contains an RNA polymerase II promotor. All LINEs, regardless of length, are flanked by direct repeats, suggesting a retroposition mechanism for their insertion and propagation in the genome.

The existence of such a high number of LINEs within the genome suggests that they must have some advantage. Originally, it was thought that LINEs were a part of the large amount of repetitive DNA present in the eukaryotic genome which does not encode any specific proteins. It is now suggested by Singer and Skowronski (1985), Charlesworth *et al.*,

(1994) and others, that LINEs may have a functional role in the genome, providing it with fluidity in terms of novel combinations of DNA sequences, as well as providing a means for the movement of DNA sequences utilizing their reverse transcriptase. The association of LINEs with genes within introns or flanking regions, has been shown to result in mutations, via insertions or chromosomal rearrangements, as well as affecting the transcription of a gene. Additionally, LINEs have been found associated with many different pseudogenes, (Packer *et al.*, 1993; Dombroski *et al.*, 1994), suggesting that they may be one important mechanism for generating and propagating pseudogenes within the mammalian genome (see reviews by Singer and Skowronski, 1985; Weiner *et al.*, 1986; Martin, 1991).

Another abundant DNA repetitive element is known as a SINE, or short interspersed repetitive element. These elements are similar to LINEs, containing 5' and 3' end direct repeats of 7-21 bp, and a 3' end poly(A) repeat, but are much smaller in size, ranging from ~75-500 bp. They also contain internal RNA polymerase III promotors so that they can remain active after insertion. SINEs can be divided into two basic families, all of which originated from RNA polymerase III genes. The tRNA related SINEs are structurally related to the tRNA molecules containing the stem-loop configuration, but they have accumulated a number of mutations so that they lose the ability to fold into this structure. They are usually around 65% identical to the parental tRNA (Daniels and Deininger, 1985). The more commonly studied SINEs arose from the 7SL RNA gene, which is a part of the signal recognition particle involved in protein synthesis. The most common SINE of this type is the Alu family, originally identified by its distinct cleavage site for the AluI restriction enzyme.

These are small, ~300 bp, dimeric elements usually found adjacent to genes or within introns. The dimer consists of a right and left monomer 7SL RNA related sequence, both missing 155 bp from the original 7SL RNA gene, with the right monomer also 31 nucleotides longer than the left. These monomers are separated by a short A-rich region. The left monomer contains box A and box B sequences which are important for transcription of the Alu element. Because of the direct repeats that flank SINEs, and the presence of the poly(A) tract, these elements are considered to move through the genome via retroposition by a similar mechanism as LINEs. However, these elements are small and therefore do not encode any proteins themselves. Thus, it is thought that their propagation may also be mediated by LINEs, which can provide a source of reverse transcriptase (Mathias *et al.*, 1991).

Similar to LINEs, these elements were originally considered part of the vast amount of junk DNA in the genome which did not encode protein or carry out any specific function. Recent studies have shown that these elements can have a dramatic effect on the genome organization, function and evolution. Alu elements have been shown to affect transcription by functioning as enhancers or silencers. Alu elements present within introns or exons of a gene can promote chromosomal rearrangements, differential splicing, or translation of the Alu element resulting in production of new genes or mutations to existing genes. Therefore, both LINE and SINE elements are present in the genome for a purpose, they actually play an active role in affecting its evolution (see reviews by Weiner *et al.*, 1986; Novick *et al.*, 1996).

1.7 Goals of This Project

Our laboratory has carried out extensive studies on the drug resistance mechanisms of CHO cells. We have developed a number of original CHO mutant cell lines, specifically utilizing antimitotic drugs, such as podophyllotoxin (Gupta, 1981; Gupta *et al.*, 1982, 1985). One particular CHO mutant cell line, Pod^{RII}, was found to contain an altered form of an abundant cellular protein, P1, later identified as hsp60 (Gupta *et al.*, 1982; Jindal *et al.*, 1989). This lead to the question as to how a mitochondrial matrix protein could be involved in resistance to a drug which acts on a cytoplasmic target. Studies on the localization of the hsp60 protein were pursued, resulting in discovery of its presence at various extramitochondrial sites, possibly acting in a chaperonin capacity (Jones *et al.*, 1994; Soltys and Gupta, 1996). Subsequent isolation of both the human and CHO cDNAs for hsp60 were carried out and these sequences used as probes for attempted isolation of both human and CHO hsp60 functional genes (Jindal *et al.*, 1989; Picketts *et al.*, 1989). However, screening of both human and CHO genomic libraries resulted in isolation of numerous pseudogenes, estimated at over 30 pseudogenes for CHO hsp60 alone. These clones closely resembled the cDNA, but contained various mutations so that they would be unable to encode functional protein (Venner *et al.*, 1990).

In addition to the association of hsp60 with a large number of pseudogenes, this protein is also interesting in terms of its involvement in autoimmune diseases, its dual role as a molecular chaperone and heat shock protein, its presence and possible functions at extramitochondrial sites, as well as its involvement in drug resistance. Therefore, one of the objectives of this project was to isolate the full length functional hsp60 gene, including

flanking regions, specifically from CHO cells. Cloning of genomic hsp60 would allow for studies of the regulation of this important gene, as well as provide insight into why pseudogenes are so abundant for this particular heat shock protein.

As a cofactor for hsp60, studies involving the hsp10 protein are also of interest in this laboratory. At the onset of this project only the cDNA sequences from higher eukaryotes had been isolated for hsp10 (Pilkington and Walker, 1993; Ryan *et al.*, 1994; Chen *et al.*, 1994; Gupta unpublished). Possessing many of the same characteristics as hsp60, hsp10 is both a molecular chaperone and a heat shock protein (Lindquist and Craig, 1988; Rothman, 1989). Thus, it was of particular interest to our laboratory and to this study to isolate the functional gene from CHO cells, including flanking regions. Studies could then follow involving the regulation of hsp10 expression, as well as comparisons to the regulatory mechanisms of hsp60. Additionally, pseudogenes for hsp10 had not previously been reported so that cloning of the functional gene, as well as pseudogenes, would allow for characterization of a possible multigene family for hsp10.

Both hsp60 and hsp10 are important proteins because of their role in protecting the cell against environmental stresses, as well as playing a role in normal cellular function. These proteins also appear to have diversified functions, including roles in autoimmune diseases, localization to extramitochondrial sites, and for hsp60, a role in drug resistance. Our laboratory has been extensively involved in studies on these proteins, more recently in terms of their cDNA sequences and discovery of pseudogene association for hsp60. Thus, the objectives of this study involving the isolation of the functional genes for both hsp60 and

hsp10 is of importance in order to gain further insight into these heat shock proteins.

2.0 MATERIALS and METHODS

2.1 Materials

2.1.1 Reagents and Chemicals

Dulbecco's modified Eagle's medium, fetal calf serum, 1 kb DNA ladder, and buffered phenol were purchased from GibcoBRL Life Technologies, Burlington, Ont. BSA, (Albumin, Fraction V), was purchased from Boehringer Mannheim, Laval, Que. Bactotryptone, bactoyeast extract and agar were purchased from Becton Dickinson and Co., Cockeysville, MD. Ampicillin and salmon sperm DNA were purchased from Sigma-Aldrich Canada Ltd., Oakville, Ont. X-gal and IPTG were purchased from Vector Biosystems, Toronto, Ont. All dNTPs used in PCR reactions were purchased from Pharmacia Biotech, Baie D'Urfe, Que. All procedures utilized autoclaved ddH_20.

2.1.2 Molecular Biology Kits and Enzymes

TA cloning kit, containing pCRII vector was purchased from Invitrogen, Calsbad, CA. Purification of DNA from agarose gels utilized the Gene Clean Kit from Bio-Can Scientific, Mississauga, Ont. Random primer DNA labelling system was purchased from GibcoBRL/Life Technologies. T_7 sequencing kit was purchased from Pharmacia Biotech, containing the T_7 DNA polymerase. Isolation of lambda DNA utilized the Lambda Maxi Kit purchased from Qiagen, Chatsworth, CA.

Calf intestinal phosphatase, T_4 ligase, *E. coli* DNA polymerase large fragment (Klenow), and restriction enzyme, XbaI, were purchased from GibcoBRL/Life Technologies. Proteinase K, RNase A, and DNase were purchased from Boehringer Mannheim, trypsin protease was purchased from Difco Laboratories, Detroit, MI. *Taq* DNA polymerase was purchased from Sangon Ltd. Restriction enzymes, BamHI, ClaI, EcoRI, HindIII, KpnI and XhoI were purchased from MBI Fermentas Inc., Flamborough, Ont., with AluI and Vent DNA polymerase purchased from New England Biolabs Inc., Mississauga, Ont. All enzyme reaction buffers were supplied by their respective manufacturers.

2.1.3 Radiochemicals and DNA Oligonucleotide Primers

[α-^{32}P]dATP (10 μCi/μl, 3000 Ci/mmol) and [α-^{35}S]dATP (10 μCi/μl, 2500 Ci/mmol) were purchased from Dupont NEN Life Science Products, Guelph, Ont. Oligonucleotide primers, used in sequencing and PCR reactions, listed below were produced by the Central Facility of the Institute for Molecular Biology and Biotechnology, McMaster University, Hamilton, Ont. The primers were designed using Primer Designer Version 2.0 computer program by Scientific and Educational Software.

Forward primers for hsp10:

degenerate primer P3796 5'-GGATCCATATGGCT(A)GGACAA(G)GCT(A)(G)TTTAG-3', P6942 5'-TCTACCGCTCTTTGACAG-3', P10400 5'-AGACGTGCTACCACCAAG-3', P10152 5'-GAGATGCTGACATTCTGG-3'

reverse primers for hsp10:

degenerate primer P3797 5'-GGATCCTCAGTCG(T)ACA(G)TAC(T)TTT(C)CC-3',

P6943 5'-CCAGAATGTCAGCATCTC-3', P10153 5'-GTCAAAGAGCGGTAGAAACT-3',

P10401 5'-CACAACCTTGGTCTCCTC-3', P10784 5'-GCTACTGAGTGTGCATCC-3',

P10996 5'-GGACACTCACCATGACTC-3'

Forward primers for hsp60:

P5814 5'-GATGGAGTAGCAGTGTTG-3', P6178 5'-ATTTTGTGAACATGGTGG-3',

P6766 5'-AGCTAAATGAGTCAGC-3', P7832 5'-ATCGTAAGCCCTTGGTC-3',

P8331 5'-GGCCAAAGGTACCAGTAT-3', P9914 5'-GCATGTCATTGACCCTAC-3',

Forward primers previously designed for hsp60 cDNA sequencing:

P730 5'-GCACTACCACTGCTACTG-3', P725 5'-AAGTTGGAGAGGTCATTG-3',

P731 5'-ATCATCTCTGATGCAATG-3', P729 5'-GAGCCTTAATGCTTCAAG-3'

Reverse primers for hsp60:

P5815 5'-AGTTAGCAAGGAGGCCAC-3', P6179 5'-TCACTACAGCTTCTGCTG-3',

P6767 5'-AGCAGTTCTTACAACC-3', P8332 5'-CATCACACCTAGTTCAAC-3',

P6961 5'-CCATAGAACAAGCCATCC-3', P9915 5'-TGCAAGCTCTCATGTCAG-3'

Reverse primer previously designed for hsp60 cDNA sequencing:

P674 5'-AATTACAGCATCAACAGC-3', P513 5'-GTGATTTCTAGCTGCTCAG-3'

Primers supplied by T_7 polymerase sequencing kit:

Universal forward primer: 5'-GTAAAACGACGGCCAGT-3'

M13 1254 reverse primer: 5'-CAGGAAACAGCTATGAC-3'

2.1.4 Cell Lines, Bacterial Strains, and Cloning Vectors

The wild type Chinese hamster ovary cell line, CHO-K1 pro$^-$, was used as the sole CHO-WT genomic DNA source, purchased from American Type Culture Collection (Rockville, MD). Two *E. coli* strain derivatives were used, JM109, purchased from Promega Corporation, Madison, WI and LE392, purchased from Clontech Laboratories, Palo Alto, CA.

Subcloning of PCR amplified fragments was carried out in plasmid pCRII, unique restriction enzyme digested DNA fragments were subcloned into plasmid pGEM-7Zf(+), or plasmid pBluescript SK (+). The general structures of these plasmids are shown in Fig. 3.

Lambda genomic CHO bacteriophage library, in EMBL3 vector, with an average insert size of 15 kb, was purchased from Clontech Laboratories, catalogue number, JL1003j.

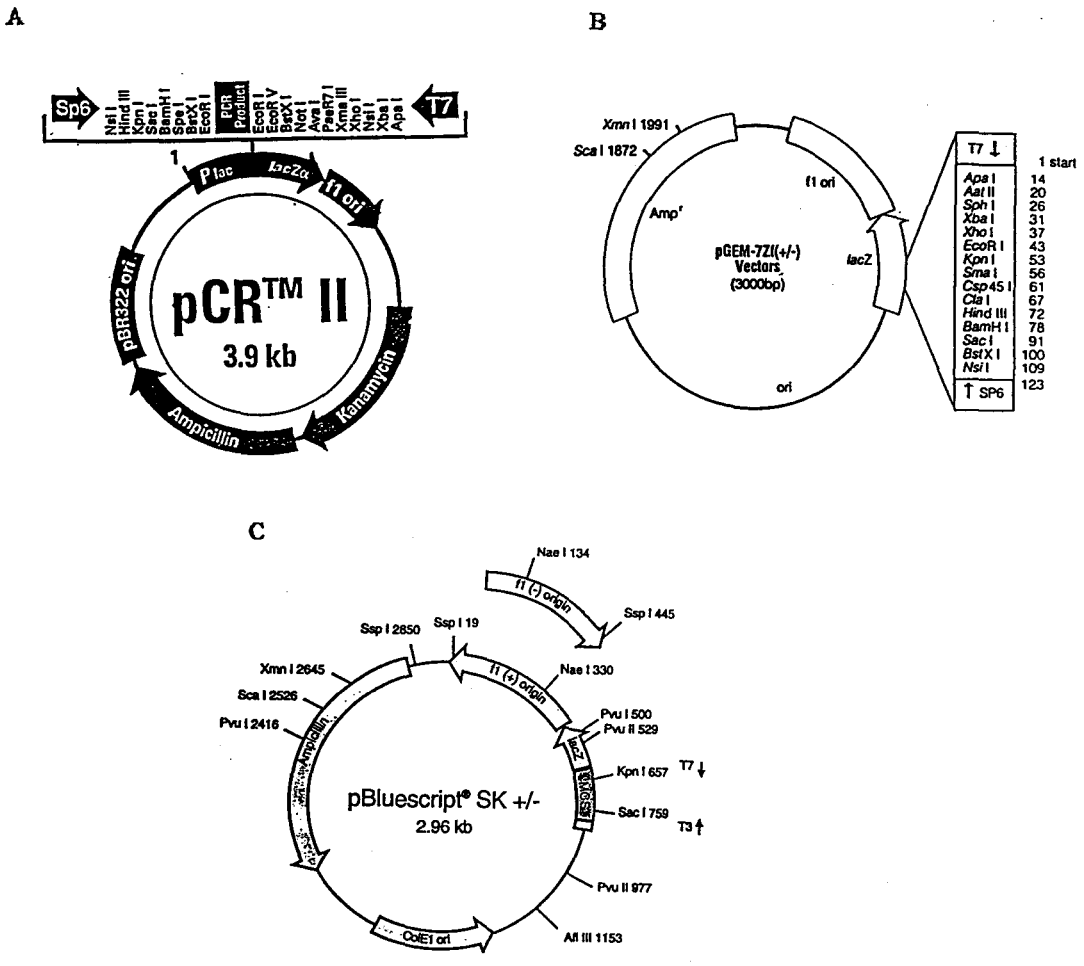

Figure 3 **DNA cloning vectors used in these studies**

A: pCRII vector (Invitrogen); B: pGEM-7Zf(+) vector (Promega Corp.); C: pBluescript SK (+/-) vector (Stratagene)

2.2 Methods

2.2.1 Cell Culture and Isolation of Mammalian Genomic DNA

Chinese Hamster Ovary wild type cells, (CHO-WT), were grown as a monolayer using Dulbecco's modified Eagle's medium containing 5% fetal calf serum at 37°C in a 5% CO_2 humidified incubator, following standard procedures (Gupta and Gupta, 1984). Isolation of genomic DNA utilized a modification of methods by Blin and Stafford, (1976), as follows. Ten confluent plates of wild type CHO cells were collected and treated with Proteinase K digestion buffer mixture (100 mM Tris-HCl pH 8.0, 25 mM EDTA, 100 mM NaCl, 0.5% SDS) to give 0.1 mg/ml/plate. The cell suspension was incubated with agitation overnight at 37°C. Genomic DNA was isolated from cellular debris using phenol:chloroform:isoamyl alcohol (25:24:1) extraction, precipitated, dried by dessication, then quantified using UV absorbance spectroscopy at 260/280 nm.

2.2.2 Bacteria Culture and Plasmid DNA Preparation

E.coli JM109 strain was used in all subcloning procedures, except for bacteriophage growth, which required the LE392 *E.coli* strain (see 2.2.8). Plasmid DNA was isolated from transformed bacteria grown overnight in 2 ml LB (containing per litre, 10 g Bactotyrptone, 5 g Bactoyeast extract, 10 g NaCl, pH to 7.2 with 5 M NaOH) containing 100 μg/ml ampicillin, after inoculation from a single colony, following the basic procedure by Birnboim, (1983). As a variation of this protocol, 2.5 M ammonium acetate was used in place of potassium acetate and DNA precipitation was performed using 0.7 vol isopropanol rather

than 95 % ethanol.

2.2.3 Enzymatic Digestion and Agarose Gel Electrophoresis

Plasmid DNA, lambda DNA and mammalian genomic DNA were digested with the appropriate enzymes as a means of analysis and fragment isolation. Digestion reactions were performed according to the instructions supplied with each enzyme by the manufacturer. Results of digestions were analyzed on agarose gels of either 0.8 % or 1 % in 1X TAE (containing per litre, 242 g Tris-Base, 57.1 ml glacial acetic acid, 100 ml 0.5 M EDTA pH 8.0) with 0.45 μg/ml ethidium bromide to allow for visualization by UV light. If required, DNA fragments were isolated from agarose gel slices by using the Gene Clean Kit, resulting in yields of 50-80 %.

2.2.4 Subcloning Restriction Enzyme DNA Fragments

Subcloning of DNA fragments into unique restriction enzyme sites within pGEM-7Zf(+) or pBluescript-SK(+) vectors first required dephosphorylation. This was carried out as described by Tabor, (1987), using calf intestinal phosphatase (CIP). DNA fragments amplified by PCR were ligated directly into pCRII vector from the TA cloning kit, overnight at 14°C. Both ligation reactions were carried out with 0.5 units of T_4 ligase, 1X T_4 ligation buffer, 50-100 ng vector and 100-200 ng fragment, to a final volume of 5 μl. Recombinant plasmids were transformed into competent JM109 E.coli cells by the method described by Hanahan, (1983). Cells were made competent by the $CaCl_2$ method by Dagert and Ehrlich,

(1974). As a modification of this protocol, the original cell pellet was first suspended in a 1X Pipes buffer solution made from 2X Pipes buffer stock (containing per 100 ml, 1.375 g Pipes, 240 mg $RbCl_2$, pH to 6.8 with 50% HCl), before $CaCl_2$ treatment. The identification of cells containing recombinant plasmids was carried out using LB/agar plates containing 100 µg/ml ampicillin, 150 µg/ml Xgal and 100 µmol/ml IPTG for blue/white selection.

2.2.5 Polymerase Chain Reaction (PCR) and Probe Reactions

PCR reactions were carried out in final volumes ranging from 10 µl, for experimental reactions, to 50 µl, for preparative or probe reactions. Reaction mixtures contained 100-200 ng/µl DNA template, 1.5 mM $MgCl_2$, 1X reaction buffer, 0.2 mM dNTP's, and 5 units of *Taq* DNA polymerase or if Vent DNA polymerase was used, 10 mM $MgSO_4$ and 2.5 units of Vent DNA polymerase. In addition, the appropriate forward and reverse synthetic oligonucleotide primers were used at concentration range of 5-10 pmol/µl, adjusted to achieve the best results. Reaction mixtures were overlayed with an equal volume of light white mineral oil, and placed for 1 min at 96°C in a pre-heated PCR machine. PCR was then carried out for 35 cycles, each consisting of 1 min at 96°C for denaturing of DNA strands, 30 sec at 53°C for annealing of primers, and 2.5 min at 72.5°C for extension of DNA strands. An additional extension program, with 95°C for 30 sec, 55°C for 30 sec and 73.5°C for 10 min, occurred after the 35 cycles to fill in any gaps in the ends of the new DNA fragments. A modification of this PCR protocol was used to make [^{32}P]dATP labelled DNA fragment for use as probe. The dNTP mixture was modified to contain all dNTP's except dATP in equal concentrations

to each other and to the amount of [^{32}P]dATP to be added (~10 μM). Cycle temperatures and times remained the same, but the program was reduced to 30 cycles and the 10 min extension time was removed. DNA fragments were also labelled with [^{32}P]dATP for probe using the Random primer DNA labelling system according to the protocol provided by the manufacturer. All DNA probe fragments were quantified for the amount of radioactivity incorporated in counts per minute, (cpm), using a scintillation counter by comparison of a washed aliquot and an original aliquot. Only probe with counts of at least 1.0×10^5 cpm or greater were used for further analysis.

2.2.6 Southern Blot Analysis

The transfer of DNA fragments to nylon membrane by the technique of Southern Blotting was carried out following the basic protocol by Southern, (1975). DNA samples were also analyzed using the manual Dot Blot method, as described by Katatos *et al.*, (1979), allowing for direct addition of DNA to positively charged nylon membrane. Agar plates containing numerous colonies from ligations and transformations using blue-white selection were analyzed using the colony hybridization method of Hanahan and Meselson, (1983), by transfer to Hybond C+ nitrocellulose membranes in duplicate. DNA on filters was denatured in 0.5 N NaOH/ 1.5 M NaCl solution, followed by 0.5 M Tris-HCl pH 8.0/1.5 M NaCl to neutralize, and 2X SSC (20X SSC per litre; 175.3 g NaCl, 88.2 g sodium citrate, pH to 7.0 with 5 M NaOH), to fix the DNA to the filters prior to hybridization.

Hybridizations of all membranes, except phage library screening, were carried out

following the basic protocol by Brown, (1993), in aqueous pre-hybridization solution (APH), (containing 5X SSC, 1% SDS, 5X Denhardt's solution; containing per 500 ml, 5 g BSA, 5 g Ficoll, 5 g polyvinylpyrroldine), with hybridization temperatures of 65°C. In all hybridizations, ~0.1 mg/ml of salmon sperm DNA was included to minimize background binding of probe. DNA blots were usually washed twice with a low stringency buffer (2X SSC, 0.1% SDS) at room temperature for 10 min each time, then depending on the gieger counter readings, 1-2 high stringency washes with high stringency buffer (0.2X SSC, 0.1% SDS) were carried out in the hybridization oven at 65°C, for short durations of 1-5 min.

2.2.7 DNA Sequencing

In preparation for sequencing, 2-5 μg of recombinant plasmid DNA was treated with 1X RNase A (50 μg/ml), denatured at 42°C for 30 min after addition of denaturing solution (0.2 M NaOH, 0.2 mM EDTA), then precipitated following a modification of protocol provided with the T_7 sequencing kit. Sequencing reactions were carried out using the T_7 sequencing kit according to the protocol provided by the manufacturer. Sequencing reaction samples were run on a denaturing urea-acrylamide gel containing 57.6 g urea, 6.84 g acrylamide, 0.36 g bisacrylamide, 24 ml of 5X TBE (per litre, 54 g Tris-base, 27.5 g boric acid, 20 ml 0.5 M EDTA pH 8.0) to a final volume of 120 ml. DNA samples were incubated for 2.5 min at 85°C then quickly cooled on ice before loading onto a pre-warmed gel. DNA sequences read from sequencing gel autoradiograms were entered into the PC/Gene computer program by IntelliGenetics Inc., for analysis. Multiple sequence alignments were performed

using CLUSTAL, and alignment of pairs of sequences with NALIGN of the PC/Gene program.

2.2.8 Bacteriophage

(a) bacterial growth and phage titer

All lambda EMBL3 vectors were grown on the host LE392 *E.coli* bacterial strain. The LB media was supplemented with 10 mM $MgSO_4$ and 0.2% maltose, as described by the manufacturer. Before growing the EMBL3 CHO genomic library for screening, the correct dilution for plating was calculated by titration of the library. A general phage library stock was made at a 1:50 dilution of the original followed by several, (~7), stepwise dilutions of phage in SM media (containing, 100 mM NaCl, 10 mM $MgSO_4$, 50 mM Tris-HCl pH 7.5). The phage and bacteria were then grown as described by the manufacturer. Plates containing 50-100 plaques were counted and used to calculate the average titer of the library in pfu/ml.

(b) screening of lambda genomic library

Lambda phage were grown at the correct calculated concentration to yield 1.0×10^5 pfu/plate onto 15 plates. The resulting plaques were transferred to Hybond C+ nitrocellulose membrane discs in duplicate. DNA was denatured and fixed to the filters before hybridization as described in section 2.2.6.

Pre-hybridization and hybridization of library filters were carried out as described in 2.2.6 with the following modifications. The pre-hybrization/hybridization solution (containing per 100 ml., 1 g BSA, 1 mM EDTA pH 8.0, 7% SDS, 0.5 M Na_2HPO_4) heated to 65°C, was

used in place of the APH solution. Low stringency wash solution (containing per 500 ml, 1 mM EDTA pH 8.0, 0.04 M Na_2HPO_4, 5% crude SDS, 2.5 g BSA) was used at room temperature for 10 min for two consecutive washes. High stringency washes, when necessary, were performed in a 55°C water bath for 3-5 min in high stringency solution (containing per 500 ml, 1 mM EDTA pH 8.0, 0.04 M Na_2HPO_4, 1% crude SDS). Plaques which bound the radioactive probe in the same positions on the original and duplicate filters were judged to be positives. These positive plaques were removed from the agar plates and stored in 1 ml of SM solution containing a drop of chloroform at 4°C. The first screening of the library usually resulted in the presence of contaminating non-positives plaques so that second and third screenings were usually necessary in order to isolate a single positive clone.

(c) isolating DNA from lambda clones

Recombinant phage containing positive clone DNA were grown to ~1.0 x 10^5 pfu/plate and isolated by two subsequent additions of 2 ml of SM solution for 2 hrs each time. This lysate was collected and used as stock for future isolation of DNA. Five plates were grown to confluence using this purified plaque stock and the new lysate isolated as previously described. DNA was purified using the protocol provided with the Lambda Maxi kit, except for column purification, which was replaced by phenol/chloroform extraction. The isolated DNA was then quantified by absorbance readings at 260/280 nm and loaded onto an agarose gel (see 2.2.3) to check for any contaminating bacterial DNA, degraded phage DNA or the presence of RNA.

3.0 RESULTS

3.1 Isolation and Analysis of hsp10 Functional Gene and Pseudogenes

3.1.1 PCR, Southern Analysis and Identification of Pseudogenes

Isolation of the functional hsp10 gene from CHO-WT genomic DNA was initially pursued utilizing PCR, and the sequence of the hsp10 CHO cDNA, previously cloned in our laboratory. Oligonucleotide primers of 18 base pairs were created near the 5' and 3' ends of the hsp10 CHO cDNA sequence, consisting of one forward and one reverse primer, to amplify ~270 bp of the central sequence (Fig. 4). Primers P3796/P3797 were utilized for a preliminary PCR with *Taq* DNA polymerase, on CHO-WT genomic DNA. This resulted in amplification of ~5 discrete bands, visualized on a 0.8 % agarose gel, of ~300 bp, ~350 bp, ~500 bp, ~600 bp and ~800 bp. In order to remove any non-specific hsp10 bands, a second PCR was carried out using a 1:10 dilution of the first reaction and a nested pair of oligomers, P6942/P6943, for forward and reverse primers, respectively (Fig. 4). This resulted in amplification of one prominent band of ~300 bp, as was expected according to the size of hsp10 CHO cDNA region between these primers (data not shown). Consequently, this fragment was purified from agarose gel and subcloned into the pCRII vector. From this transformation, 8 white colonies were picked and the insert sequenced. The inserts in all these clones were very similar to the hsp10 CHO cDNA, but contained various mutations, including deletions and single base pair changes such that they could not encode functional

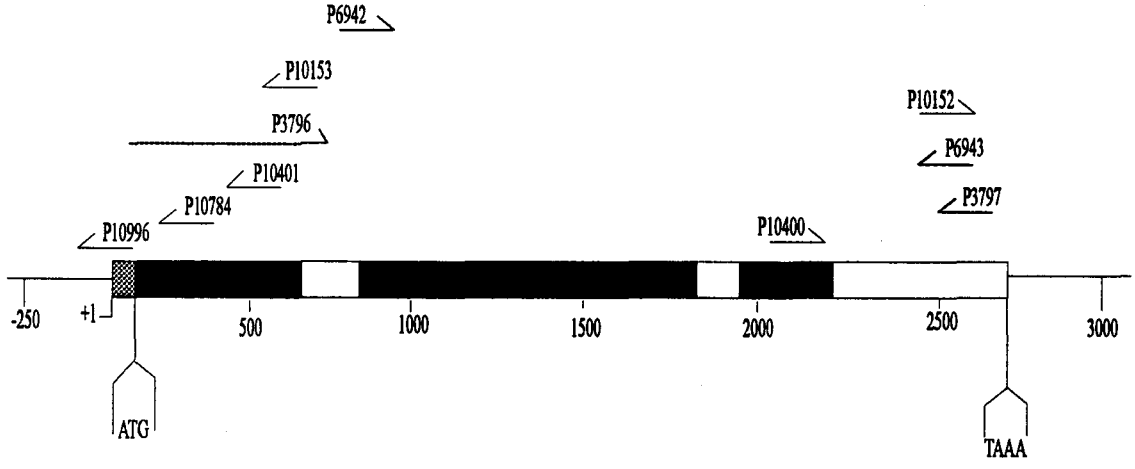

Figure 4 **Location of primers used in PCR and sequencing of the cpn10 3Aii and 4Bi clones**

The sequence from the putative transcription initiation site, (+1), to the translation initiation codon, ATG, is shown as a shaded square. Introns are represented by black rectangles, and the exons as open rectangles. The 5' and 3' end flanking regions are drawn as single lines. Arrows denote the direction of the primers. P3796, overlaps the first exon/intron boundary and was useful for sequencing of the 4Bi pseudogene, shown as an extended arrow.

hsp10 protein, with none of the clones containing any introns. Of the eight clones, seven were identical containing 24 single bp changes and a deletion of 2 bp near the 3' end, which resulted in a reading frameshift. One of these clones was designated hsp10-1A, showing ~91 % identity to the hsp10 CHO cDNA sequence. The other type of clone present in the original eight, also contained numerous single bp changes, plus 3 separate deletions and a single bp addition near the 3' end. This clone designated hsp10-2A, showed ~90 % identity to the hsp10 CHO cDNA sequence and ~96 % identity to hsp10-1A. Because of their high degree of similarity to the hsp10 CHO cDNA sequence, but the presence of numerous mutations, these clones were considered partial pseudogenes for hsp10 which previously had not been demonstrated.

From two independent PCR reactions, using CHO-WT genomic DNA and the above primers, 14 additional clones were sequenced, 13 of which were found to correspond to the previously identified hsp10-1A pseudogene. The other clone was found to contain unique single bp changes and deletions, not seen in any previous clones, and was designated hsp10-3A. This pseudogene had only ~85 % identity to the hsp10 CHO cDNA, the lowest identity observed for any of the pseudogenes, as well as only ~84 % identity to hsp10-1A.

In order to confirm that the mutations were actually present in the genomic DNA and not a result of PCR errors, the PCR amplifications were repeated utilizing Vent DNA polymerase. Of the eight clones sequenced, 6 were identical to the hsp10-1A pseudogene and 2 clones were unique, Vent2B and Vent15, which were also classified as pseudogenes. These 2 clones were ~89 % and ~96 % identical to the hsp10 CHO cDNA, and ~89 % and ~94%

identical to hsp10-1A, respectively. An alignment of hsp10-1A, along with the other four unique pseudogenes is shown compared to the hsp10 CHO cDNA in Fig. 5. By comparing all the pseudogenes to each other, only 5 common changes in the region between primers P6942/P6943 are observed. As well, the majority of the changes are concentrated from the middle of the sequence towards the 3' end, including the deletions, or insertions which would lead to frameshift mutations. The hsp10-3A pseudogene contains the lowest identity to the hsp10 CHO cDNA, (~ 85%), and consequently the largest number of bp changes.

To further investigate the existence of pseudogenes in the CHO genome for hsp10, a Southern blot analysis was performed utilizing ~10 µg of CHO-WT genomic DNA digested with BamHI, EcoRI, HindIII, KpnI and XbaI. These enzymes were chosen based on the fact that they did not cut within the hsp10 cDNA sequence, except HindIII which cuts only at bp 53. This decreases the possibility that single hsp10 sequences would be cut into multiple fragments, leading to a false observation of multiple hsp10 sequences in the CHO genome. One of the partial pseudogene clones, hsp10-1A, was used as probe, since despite its numerous mutations, it still maintains ~91 % similarity to the hsp10 CHO cDNA, and an average of ~91 % similarity to the other pseudogenes, making it useful for identifying other hsp10 sequences in the CHO genome (Fig. 5). The resulting blot revealed the presence of multiple bands in each restriction enzyme lane, confirming the existence of numerous hsp10 sequences in the CHO genome (Fig. 6).

3.1.2 Screening of the EMBL3 CHO Genomic Library

```
                                                                P6942 →
hsp10cDNA    69 TCTACCGCTCTTTGACAGAGTATTGGTTGAAAGGAGTGCCGCCGAAACTG
Vent15          ------------------A--------------------A--A------
hsp10-1A        ------------------A--------------------A--A------
hsp10-2A        ------------------A--------------------A--A------
Vent2B          ------------------A--------------------A--A------
hsp10-3A        ---C---------------A------------T***-A-TA-T-------

hsp10cDNA   119 TGACCAAAGGTGGCATTATGC TTCCAGAAAAGTCTCAAGGAAAAGTATT
Vent15          -------------------- ----------------A----------
hsp10-1A        -----------------C--- -----------C---------------
hsp10-2A        -----------------C--- -----------C---------------
Vent2B          -----------------C--- -----------C---------------
hsp10-3A        -A---------------C---C-----A---CA---G---------C---

hsp10cDNA   168 GCAATCAACAGTAGTGGCTGTGGGAGCAGGCTCGAAAGGAAAGGGTGGAG
Vent15          ----G-------------------------------A-----------
hsp10-1A        ----G---------CA---A-----TT----G-A--G------T----G-
hsp10-2A        ----G---------CA---A-----TT----G-A--G------T----G-
Vent2B          ----G-----------------T-T-----C-A---------------
hsp10-3A        -T-----T*----AGT---------T-----A-A----A----------

hsp10cDNA   218 AGATTCAGCCGGTCAGTGTGAAAGTTGGAGACAAAGTTCTTCTCCC AGA
Vent15          ----------A-----------------------------------C---
hsp10-1A        ----------AA-----------------T---------***--C--G
hsp10-2A        ----------AA-----------------T---------***--C--G
Vent2B          ---------T----T---CTG--**********------****--C---
hsp10-3A        -------A--A----------G--CC--------------****--C*--

hsp10cDNA   267 ATATGGAGGCACCAAAGTAGTTCTAATCGACAAGGATTATTTCTTATTTA
Vent15          ------------------------GA-A--------------------
hsp10-1A        ------------------------GA-A---*A------------A--
hsp10-2A        ----------------GGAT****GA-AT*G*-AC-AT--A--AC--A-
Vent2B          ---C-------T-------A-----GAT-----ATA-------------
hsp10-3A        G----------------------*-GA-A-------C----C---C--G-

                          ← P6943
hsp10cDNA   317 GAGATGCTGA   CATTCT GG
Vent15          ----------TGA------ --
hsp10-1A        ---------*  A------ --
hsp10-2A        ---------*  A------ --
Vent2B          ------G-**  A------T--
hsp10-3A        ----------  *------GA-
```

Figure 5 **Alignment of the partial nucleotide sequence of the CHO hsp10 cDNA and five hsp10 pseudogenes amplified from CHO genomic DNA**

Primers used in the PCR amplification are indicated in bold italics. The nucleotide sequence is numbered according to the position of this partial sequence within the CHO hsp10 cDNA sequence. Identical nucleotides are indicated by (----), deletions in the pseudogene sequences are indicated by (*), gaps in the cDNA or pseudogene sequences represent insertions in one or more of the sequences. Mutations in the pseudogene sequences are concentrated towards the 3' end, with the most divergent pseudogene being hsp10-3A.

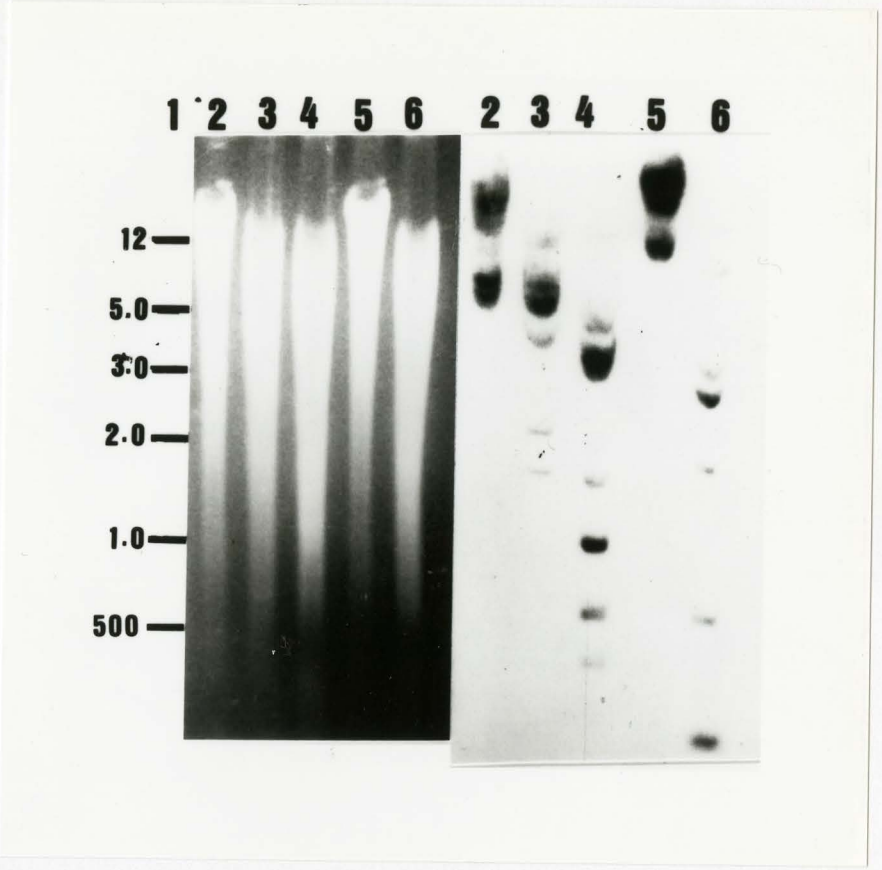

Figure 6 Southern blot analysis of CHO genomic DNA

Approximately 10 µg of genomic DNA was digested to completion with the enzymes listed below for Southern analysis using the hsp10 cDNA probe as described in the text. Only a single high stringency wash was used for 10 min at 65° C in 0.2x SSC/0.1% SDS and the blot exposed overnight.
Lanes: 1= 1kb DNA ladder, 2= BamHI, 3= EcoRI, 4= HindIII, 5= KpnI, 6= XbaI

In order to isolate the functional gene for hsp10, as well as full length pseudogenes, screening of a lambda EMBL3 CHO genomic library was performed. Because of the small size of the cDNA sequence for hsp10, ~475 bp, it was expected that even with the presence of introns, the complete functional hsp10 gene would be found within a single lambda clone. The CHO cDNA clone for hsp10 and the internal primers, P6942/P6943, were used to carry out PCR to amplify an ~270 bp fragment of the hsp10 CHO cDNA. This was then used as template for a second PCR to make the radiolabeled probe for the library screen. Initial screening of the CHO genomic libary resulted in isolation of 5 positive clones, four of these, 3Aii, 4Bi, 5Ciii, and 9Bii, were purified by re-plating and screening of the plaques.

Based on the previous findings of hsp10 pseudogenes within the CHO genome, the possibility existed that these 4 clones could correspond to pseudogenes. Therefore, DNA was isolated from the lambda clones for use as template in PCR. It was expected that the functional hsp10 gene may contain introns and any pseudogenes would most likely be of the processed type and lack introns. Therefore, if PCR was performed utilizing primers P6942/P6943, amplification of pseudogenes would result in a fragment of ~300 bp, corresponding to the distance between these primers in the hsp10 cDNA sequence. However, the functional gene, expected to contain introns, would result in amplification of a much larger fragment. The resulting PCR reactions were subjected to Southern analysis using the hsp10 cDNA fragment as probe. From this analysis a prominent band in the lane for 3Aii of ~2.1 kb bound to the probe, while the other clones, 4Bi, 5Ciii, and 9Bii, bound a smaller band of ~300 bp (data not shown). This suggested that the 3Aii clone possibly corresponded to

the functional hsp10 gene, and the other clones corresponded to hsp10 pseudogenes.

3.1.3 Southern Analysis and Subcloning of Pseudogene and Functional Gene Fragments

In order to confirm the identities of the hybridizing bands from the above Southern analysis, a larger scale PCR reaction of 50 µl was carried out for the 3Aii and 4Bi clones. The ~2.1 kb fragment for 3Aii and the smaller ~300 bp fragment for 4Bi were subcloned into the pCRII vector and sequenced.

The sequence for the 3Aii clone corresponded to the hsp10 CHO cDNA sequence in the region between P6942/P6943 with the addition of two introns. The intron closer to the 3' end of the clone, of ~200 bp, was completely sequenced, while the other intron, closer to the 5' end was larger, estimated at ~1.6 kb. Only ~200 bp of the 5' end and ~100 bp of the 3' end of this intron were sequenced. All boundaries of the introns contained the consensus, 5' GT and 3' AG sequences at the splice sites (Mount, 1982). Based on this preliminary sequence comparison, it strongly suggested that the 3Aii clone corresponded to at least part of the CHO hsp10 functional gene.

In order to isolate the full length hsp10 CHO functional gene, complete with 5' and 3' end flanking regions, DNA from the 3Aii lambda clone was digested with various restriction enzymes, BamHI, EcoRI, HindIII, ClaI, and XhoI, and Southern blot analysis was carried out using the hsp10 CHO cDNA probe. A band in the BamHI lane of ~5.5 kb, hybridized strongly to the cDNA probe, (Fig. 7), and was subsequently isolated from the

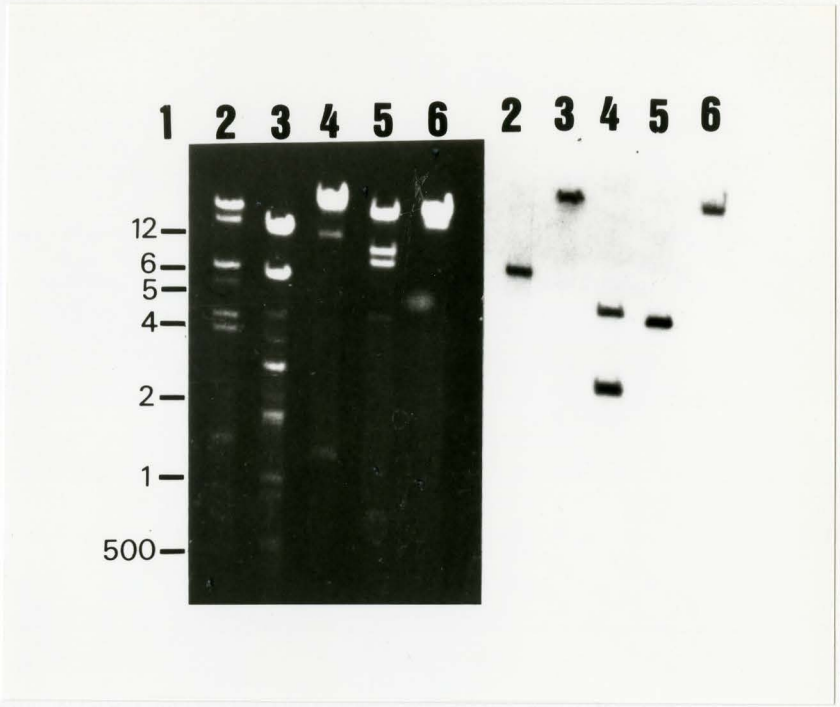

Figure 7 **Southern blot analysis of 3Aii clone DNA from CHO genomic library screen**

Approximately 5 μg of 3Aii lambda DNA was digested for 3 hrs at 37°C with the enzymes listed below, for Southern analysis with the hsp10 cDNA probe as described in the text. A single high stingency wash for 2 min heated to 65°C in 0.2x SSC/0.1% SDS but added to the blot at room temperature was required. Exposure to X-ray film was ~16 hrs.
Lanes: 1= 1 kb DNA ladder, 2= BamHI, 3= ClaI, 4= EcoRI, 5= HindIII, 6= XhoI

agarose gel and subcloned for sequencing.

Sequencing of the ~300 bp fragment of the 4Bi clone revealed it contained a high degree of similarity, ~93%, to this region of the hsp10 CHO cDNA. This fragment lacked introns and contained 13 bp changes, resulting in 10 amino acid changes, as well as 2 single bp deletions leading to 2 frameshift mutations. This sequence information strongly suggested that this clone corresponded to a hsp10 pseudogene. This 4Bi pseudogene differed from the previously identified pseudogenes, with identities of ~ 89%, ~ 87%, ~ 85%, ~ 87% and ~ 93%, for hsp10-1A, hsp10-2A, hsp10-3A, Vent2B and Vent15, respectively.

Similar to the isolation of a larger fragment for the 3Aii clone, Southern analysis of the 4Bi lambda DNA resulted in isolation and subcloning of an ~1.5 kb fragment which hybridized to the hsp10 cDNA probe. DNA sequencing of the resulting clone, in the region between primers P6942/P6943, revealed they were identical to the previously sequenced ~300 bp PCR fragment for 4Bi DNA, containing the same bp changes, and deletions, as well as lacking introns. Assuming that the expected size of a hsp10 processed pseudogene would be similar to the cDNA of ~475 bp, the ~1.5 kb subcloned fragment was expected to contain the full length 4Bi pseudogene including flanking regions.

3.1.4 Sequencing and Analysis of the 3Aii Functional Gene and 4Bi Pseudogene

The sequencing of the 3Aii functional gene was carried out by extending outwards, both in the 5' and 3' directions, to read beyond the ends of the hsp10 cDNA sequence. Various primers were used in the sequencing of this clone as illustrated in Fig. 4. Initially,

overlapping primers were designed for the P6942/P6943 primers, but in opposite directions, to begin sequencing upstream and downstream. This resulted in identification of a third, large intron, ~600 bp, at the 5' end of the clone. Additional synthetic primers were created as necessary once the previous primers revealed new sequence information (Fig. 4).

Analysis of the upstream region of the 3Aii functional gene revealed characteristic regulatory sequences. The 5' UTR contained a consensus CAAT box located at -39 position, two Sp1 consensus elements, located at -68 and -10 from the +1 start site, and a putative HSE at -225 bp upstream. This region is ~64% GC rich over ~270 bp, however, the functional gene appears to lack a conventional TATA consensus sequence or Inr sequence. The 3' end region contains a consensus polyadenylation signal, ATAAA about 65 bp after the termination signal.

The sequencing of the 4Bi pseudogene was carried out using the primers, P10153 and P10152, which allowed for determination of ~230 bp downstream of the 3' end of the hsp10 CHO cDNA and ~300 bp upstream of the 5' end of the cDNA. An alignment of the 3Aii full length functional gene and the 4Bi full length pseudogene is shown in Fig. 8. Comparing the 4Bi pseudogene sequence to the 3Aii functional gene, excluding introns, reveals ~90 % similarity. The upstream region of the two genes is completely divergent and shows limited homology from the transcription initiation site to the ATG initiation codon. In the region downstream of the position of the poly(A) tail in the cDNA, they retain significant homology, until a large stretch of A's is reached in the pseudogene sequence. The upstream region of the 4Bi pseudogene shows identity to the ORF2 reverse transcriptase consensus amino acid

```
4Bi    -303  TGTCTGAATGGGCTAAATGATAACATCAATGACAATCTATGCTGGAAAGGATGTGGAGGA
                                    HSE
3Aii   -244  TGTTCCTGGTCTCTCCGTTCTAGAACTTCCGGAAGGCTGCGCCTCCCGTGGGTCAAGGTC
4Bi   (-243) AAGAGGAAC-----CT-CATTGCTGGTGGGA-TGGAAACTTTTAC-A-TACTT-GG-AAG-

3Aii   -184  AGTGGGACTCATTTCCCGGGTGTGGTACAAGGGCTCGTTCTTTCAACCTCAGCGGGCCGC
4Bi   (-183) ---ATT-TGGT--CTGTA--AAAATGGG--TC-G--TAC--CAAG-TC-AGCAATC--T-
                                                                    Sp1
3Aii   -124  CTGAAAAGCCTAGAAAGTACGTACGCCCCTTTTCTTCGTCCGGTCTCCGCCTTCGGCGGG
4Bi   (-123) TCTTGGCATA--CCC-AAGAA-G-A-A-TCA-A-AATAAGGATATCTGTT-AA-TAT-TT
                                                                    Sp1
3Aii   -64   CCTGGCGGCGGAGTCCTCGCCACCAATCGGCGGCGGTACGTACGGCTGCCGTGGGCCGGC
4Bi   (-63)  -A-A--A--ATTA-TTGTAAT-T-TGGAAC-TAGAAGCAACT-TC-----A-*-------
                 ⌈+1                          ⇓ intron 1
3Aii    -4   CCGAGTGCAGAGTCCGAGCGGCGGCGCCG GAGTCATGgtgagtgtccctgggcaggcct
4Bi    (-4)  ---------A- -------T--A---AT-C------A-

3Aii    56   gcctgcgggtcggggtcccccgcgacccacggtgtccccccctgtgtcggcgtttagcag

3Aii   116   ccggcagggcaggggccacgcagtgttcggcgcccgccgcccgggttccgcaggagcga

3Aii   176   ggcctgtccgatgcggttctccaccacggcctctttgcacgcaacgtgtgctgcaggagc

3Aii   236   ccggggtaggggcgggaaccggatgcacactcagtagcccacctgccctgccagggcttt

3Aii   296   catggacgtatactgtgttggagaatatatgacccagctcaaaaattgatactctctccc

3Aii   356   cggcgttactagaattaaaaaaagaaatctaaagaagggaaaacgtttgaccttggagga

3Aii   416   gaccaaggttgtggttaaaggtagtgtgggagctgatgggcaatgaagacacccccccca

3Aii   476   cccacccacacgcactaaaatatacagggtatgtgaaacatccaagaatattttctaaag

3Aii   536   gcacacctagatttttcttgtcagtttggacgtttgacccgttataggttaaggcatgca

3Aii   596   gcggtattaactgttttcgtggtctgacgctgcgaacgcaatgacagtccgtttgaacac
                          ⇓
3Aii   656   gtttctctcgcagGCTGGACAAGCTTTCAAAAAGTTTCTACCGCTCTTTGACAGAGTATT
4Bi    (33)                ----------------*------------------------**---

3Aii   716   GGTTGAAAGGAGTGCCG CCGAAACTGTGACCAAAGGTGGCATTATGCTTCCAGAAAAGT
4Bi    (47)  ------------A-----G-------------------------------------------
                                                                        ⇓
3Aii   775   CTCAAGGAAAAGTATTGCAATCAACAGTAGTGGCTGTGGGAGCAGGCTCGAAAGGAAAGg
4Bi   (107)  -------------------------T--------------------A-C--------G
              intron 2
3Aii   835   taactgggagctactacagcagtactgagtcttatatgtgatatggaaggaaaggagtag

3Aii   895   aaaatcctagattaccaagtggcttttaaaatagtagatgctttttttgaaataatttaa

3Aii   955   cttcacaaagaaccgcgcatgaaggaaaaccattgtggagtttgtgtttatttttctaag
                                ~1.3 kb
3Aii  1015   gaatggagaggcattctatnnnnnnnnnnnngagctcctgtgtaactgtacgaaagaaag
                                                                        ⇓
3Aii  2315   tgttcattgtgtgtcttgtgaacaagatgccttctaattaacgcaactatttttttaagG
                                                                        -
```

```
3Aii 2375   GTGGAGAGATTCAGCCGGTCAGTGTGAAAGTTGGAGACAAAGTTCTTCTCCCAGAATATG
4Bi (167)   ------C--------A-----------------------------------------A----
                                          ⇓intron 3
3Aii 2435   GAGGCACCAAAGTAGTTCTAATCGACAAGgtagataaagctatgatttatttaaatcaag
4Bi (227)   ----------------------------

3Aii 2595   acagggtttctctgtgtcactctggctgtcctggaacctctctgtagaccaggctggctt

3Aii 2555   tgaactcacaaagatttgcctgcctttgcctcccaagtgctggaattaaagacgtgctac

3Aii 2615   caccaagcaataaagctaaaatgagttgtgtgtttgcttttactgtcttaactggtggtt
                      ⇓
3Aii 2675   tctttattcacagGATTATTTCTTATTTAGAGATGCTGACATTCT GGAAAATATACGGA
4Bi (256)                 ---------------------------------T----------T-CT

3Aii 2734   CTGAAATCACTGTTGAAATGGTATCACAGTGAATCTGCCCATTCCGCC  G A C
4Bi (303)   ---------------T--------------A-------------T---AA-G-A-CCTGA

3Aii 2785         ATTCTGAACTATTCATCATGTAAATAATTTCTGTGCCTCCCTTTTATAATAAA
4Bi (363)   CCAAACT----------------------------C----------*------------

3Aii 2838   CTGATGATGCCTAAACTGATGACATCCATTGCTCTGAAGGTTAGTATCAATGTACTGTTA
4Bi (422)   ------------T-----------G-----------------T----------------

3Aii 2898   TAAACAT   CGAATAAA TAA GTAAATGAGTGACGGTGTTTAGTAATTTAAGGTGGGT
4Bi (482)   -------TTC--------A---T--------AAA-AA-GAAG--TGGCAG--CACA--CA

3Aii 2953   TTCTCCTTTAGCTGCAGAGTAAGAGTCTGAGAAGAGGGTCTTAATAATAATTTGAGGATG
4Bi (542)   GG-CTGGCCGCTA-AGT----GCT-AAA-CTC-C-TCA-GA-CC-CGAGC-CCCTAT-G-

3Aii 3013   ATTGCTCTCAAAAGAGGG
4Bi (602)   TGCT-AAGTCGT-TTA--CCAGTGAGGCTAGAGTCGACCAGACC
```

Figure 8 Alignment of the full length nucleotide sequence of CHO hsp10 clone 3Aii and pseudogene 4Bi isolated from CHO genomic library

The putative transcription initiation site is labelled as +1, initiation codon and termination codons are underlined. Putative promotor elements, including a heat shock element,(HSE), are shown in bold. The polyadenylation signal and the poly(A) tail from the cDNA position are in bold italics. Intron positions are indicated by (⇓), with the sequence in lowercase. Only ~300 bp of intron#2 was sequenced of an estimated ~1.6 kb, the unknown sequence is represented by (nnn). Gaps represent insertions in the pseudogene, (*) deletions. The intron regions are left blank in the pseudogene which lacks these introns. The two clones show significant homology within the coding region, but are completely divergent in flanking regions. The upstream region of the 4Bi pseudogene shows homology to the ORF2 of a human LINE sequence.

sequence of ~53 %. This sequence is commonly found as part of the 3' end of the DNA repetitive element, known as a LINE (Martin, 1991; Singer *et al.*, 1993), for which this region of the 4Bi pseudogene shows ~47 % similarity to the amino acid sequence of a human LINE.

3.2 Isolation Attempts of Genomic hsp60 and Partial Pseudogenes

3.2.1 PCR and Southern Analysis of CHO-WT Genomic DNA

An hsp60 CHO pseudogene, GC1, had previously been isolated from CHO-WT genomic DNA in our laboratory. Sequencing of this pseudogene revealed the presence of a small ~87 base pair intron near its 3' end, as well as numerous single bp changes, additions and deletions when aligned with the hsp60 CHO cDNA (Venner *et al.*, 1990). Generally, the majority of pseudogenes are found to be of the processed type, or those that lack introns, (see review by Vanin, 1985), therefore in order to isolate the functional gene for CHO hsp60 screening of a CHO genomic library was undertaken, using a probe fragment that contained the small ~87 bp intron of the GC1 pseudogene. This was expected to result in isolation of the functional gene, which may contain introns, and possibly additional pseudogenes, but only those of the less common semi-processed type. Utilizing PCR with primers P6178/P6179, (Fig. 9), a ~200 bp fragment of the GC1 pseudogene containing the ~87 bp intron was amplified for use as radiolabeled probe and called PSC-4. In Southern blot analysis of CHO-WT genomic DNA, digested with AluI, BamHI, EcoRI, HindIII, KpnI, and XbaI, this probe recognized numerous bands (data not shown). Most of these enzymes, with the exception of EcoRI and AluI, do not cut within the hsp60 CHO cDNA sequence. Thus, the presence of multiple hybridizing bands on the Southern analysis would most probably correspond to the recognition of both functional and pseudogene sequences in the CHO genome. Attempts were then made to make the probe fragment more specific for the ~ 87 bp intron sequence. The 18 bp oligonucleotide primer pair, P6766/P6767, was created to more closely flank the

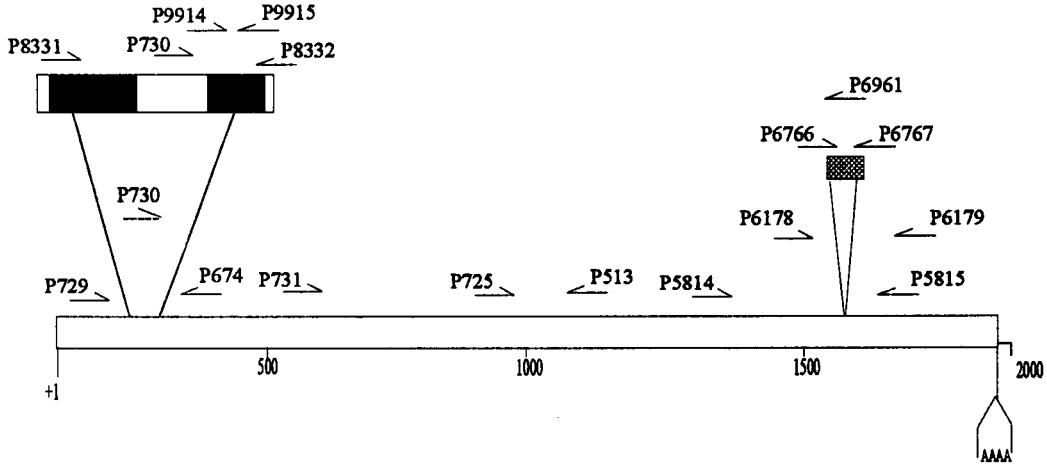

Figure 9 **Location of primers used in PCR and sequencing for CHO hsp60 related sequences**

The CHO hsp60 cDNA is represented as the large open rectangle. The translation initation site is shown as +1, the position of the poly(A) tail as (AAA). The position of the small ~87 bp intron of the GC1 pseudogene is shown as a grey inset. The ~2.5 kb fragment of the CHO hsp60 functional gene isolated is shown as an inset at the 5' end. Intron#2 and intron#3 contained within this fragment are shown as dark rectangles, exon#3 as an open rectangle. Arrows denote the direction of the primer.

~87 bp intron in the GC-1 pseudogene, leaving only 17 bp of exon in a ~104 bp fragment (Fig. 9). A probe based on this fragment, PSC-4#2, bound to fewer bands in each restriction enzyme lane on a similar Southern analysis than the previous probe fragment, PSC-4, (data not shown), and therefore, this probe was used to screen the lambda EMBL3 CHO genomic library.

3.2.2 Screening of EMBL3 CHO Genomic Library and Identification of Pseudogenes

The PSC-4#2 probe fragment, used to screen the EMBL3 library, resulted in identification of 50 positive clones. Two of the positive clones, designated λ4 and λ5, were plaque purified by additional rounds of screening, and their DNA isolated for further analysis. A reverse 18 bp primer, P6961, was designed within the middle of the ~87 bp GC1 intron, so that only clones corresponding to the hsp60 functional gene or semi-processed pseudogenes, like GC1, would bind it for sequencing or PCR amplification (Fig. 9). In conjunction with a previously designed primer, P731, used in the sequencing of the hsp60 CHO cDNA sequence, (Fig. 9), P6961 was used in PCR with the λ4 and λ5 DNA. A fragment of ~1.0 kb was expected if these clones corresponded to pseudogenes lacking introns, based on the distance between these primers on the GC1 pseudogene sequence. A larger fragment was expected if the clones corresponded to the functional gene which may contain additional introns. Amplification of λ4 and λ5 DNA resulted in a prominent fragment of ~1.0 kb for both clones (data not shown). The amplified fragments were isolated from agarose gel and subcloned into the pCRII vector. Sequencing of the ends of these clones,

using the universal and reverse primers, revealed that they showed homology to the hsp60 CHO cDNA. Both clones contained the 5' end sequence of the ~87 bp intron, from the position of the P6961 primer, as well as numerous mutations, confirming their identity as pseudogenes for CHO hsp60. Additional sequencing with various other primers, confirmed the above finding. The sequencing also revealed that the λ4 clone lacked the ability to bind to the primer used for sequencing of the central area of the clones and thus lacks ~230 bp in this region. An alignment of the partial sequence of the hsp60 CHO cDNA, the GC1 pseudogene, λ4 and λ5 is shown in Fig. 10. The partial sequence of λ4 shows ~81 single bp changes, 2 large deletions of 12 bp and 21 bp, both also found within the GC1 clone, a large 24 bp addition near the 3' end, as well as single bp deletions and additions when compared to the hsp60 CHO cDNA. Considering the ~230 bp missing from its central region, λ4 shows ~ 67% similarity to the hsp60 CHO cDNA sequence, and ~ 74% and ~72% similarity to the GC1 and λ5 pseudogenes, respectively. The complete ~1.0 kb sequence of the λ5 clone shows ~100 single bp changes, numerous single and multiple bp deletions and several additions. As well, this clone contains a large 12 bp deletion that is common to both the GC1 and λ4 pseudogenes. Comparison of the hsp60 CHO cDNA sequence and the GC1 pseudogene to the λ5 clone, reveals ~ 90% and ~ 94% similarity, respectively.

3.2.3 Isolation of CHO hsp60 Fragment and Sequence Analysis

The initial screeing of the CHO genomic library resulted in isolation of two semi-processed pseudogenes, therefore attempts were made to isolate a more specific probe

```
                  P731→
cDNA 556   CATCATCTCTGATGCAATGAAAAAGGTTGGAAGG AAAGGCGTCATCACAGTGAAGGATG
GC1        ------A---------------C-T------*--G----A----------C--------A
λ5         ----------------------C-T---------G---------------C--------A
λ4         ----------------------C-T------*--G---------------C---------

cDNA 605   GAAAAACCCTGAATGATGAGTTAGAAATTATTGAAGGCATGAAGTTTGACAGAGGATATA
GC1        ----T--------------A------------G---------------A---A---G--
λ5         ----T--------------A-----------C-----------------A---A---G--
λ4         ----T--------------A------------G---------------A---A---G--

cDNA 655   TTTCCCCGTATTTTATTAACACATCAAAAGGTCAAAAATGTGAATTCCAAGATGCCTATG
GC1        ----T-A-***********-------T-A-------------------------------
λ5         ----T-A-***********-------T-A-------C--G----------------G--
λ4         ----T-A-***********-------T-A-------C--G----------------G--

cDNA 705   TTCTGTTGAGTGAAAAGAAAATTTCTAGTGTCCAGTCCATTGTACCTGCTCTTGAAATTG
GC1        ---A----------------------T----G--C-----AC****************
λ5         C------------------------A-T----------G--AC--*-----A---****-
λ4         --------------------------C--GT---------AC****************
                       P7832→
cDNA 755   CTAATGCT**CATCGTAAGCCCTTGGTC**ATTATTGCTGAAGATGTTGATGGAGAAGCTCTAA
GC1        ****-------A-G--A-AAA----G--AT-------------------------------
λ5         ---*-------A-G--A-AAA----G--AT-*-----*-----------------------
λ4         ****-------A-G--A-AAA----G--AT---------------------

cDNA 855   GCACACTGGTTTTGAACAGGCTAAAAGTTGGTCTTCAGGTGGTAGCAGTCAAAGCTCCAG
GC1        --------------------------C----------A-C--------T----------
λ5         ------------------------G-C----------C-----T--T----------
λ4

cDNA 905   GTTTT GGGGACAATAGGAAGAATCAGCTTAAAGATATGGCTATTGCTACTGGTGGTGCG
GC1        -----T-----------------A-----------C-------------A-----A-
λ5         -----**----------------A---C--------C-- -----A-------------
λ4

cDNA 954   GTGTTTGGAGAAGAGGGGCTAAATCTAAATCTTGAAGATGTTCAAGCTCATGACTTGGGA
GC1        CAA---------C----------A------A----T----G-----T------T--A---
λ5         CAA----A----C----------A-------C---T----------G---A--T--A---
λ4

cDNA 1004  AAAGTTGGAGAGGTCATTGTCACCAAAGATGACGCCATGCTTTTGAAAGGAAAGGGTGAG
GC1        ----T----------------C----------T-------------------CT-A---C
λ5         --GTG----------------C----------T-------------------CT-A---C
λ4                                                  ------------CT-A---C

cDNA 1054  AAAGCTCAGATTGAGAAACGAATTCAAGAAATCACTGAG CAGCTAGA
GC1        T--A----A----------C---------------A-- --------
λ5         T--A----A----------AC--------------A--G--------
λ4         T--A----A----------C---------------A-- -------TGAATAAATGTAA

cDNA 1141                AATCACAACTAGTGAATATGAAAAGGAAAAACTAAACGAAAGACTCGC
GC1                      C----------------------------G----------C---T--
λ5                       C----------------------------G--G---*---C---T--
λ4         GAAATCACTAAGC----------------------------GCTA----A---C---T--
```

```
cDNA 1179 TAAACTTTCAGATGG AGTAGCAGTGTTGAAGGTTGGAGGAACAAGTGATGTTGAA GTG
GC1       --------------- T-C---T-----T----------------G--------- ---
λ5        --------------- A-----T-----T----A----------G-G---------A---
λ4        C---------------GT-C--CT-----T----T*----------G--------- ---

cDNA 1228 AATGAGAAGAAAGACAGAGTTACAGATGCTCTCAATGCTACAAGAGCAGCTGTTGAAGAA
GC1       --G--------------------A-------------------------------A-G
λ5        --G-----------------------------------------------------A--
λ4        --G-----------------------------------------------------A-G

cDNA 1277 GGCATTGTTTTAGGAGGGGGCTGT GCTCTGCTTCGATGCATCCCAGCCTTGGATTCCTT
GC1       ----------------TTA----- ------C--TG----------------------
λ5        ----------------TTA----- ------C-----------------------------
λ4        ---------------CTTA-----T------C--C------------G----A-------

cDNA 1327 AAAGCCTTCTAA TGAAGATCAGAAAATAGGTATAGAAATTATTAAAAGAGCACTCAAAA
GC1       ---------A-- ----------------------------------------------
λ5        ---------A-- ----------G-------------C----------------------
λ4        ---------A--A--------------------------------------------

cDNA 1425 TTCCTGCAATGACAATTGCTAAGAATGCAGGTGTTGAAGGATCTTTGATAGTTGAGAAAA
GC1       --------G----------G-----------------TA--------G-----------
λ5        -------CG----------G-----------------TA--------G-----------
λ4        --------G----------G-----------------TA--------G-----------

cDNA 1475 TTCTGCAGAGTTCCTCGGAAATTGGTTATGATGCTATGCTCGGAGATTTTGTGAACATGG
GC1       --------G------A--G-G--------T-G-------C-C-A---------------
λ5        -----A---------A--G-G----------G---------A-----------------
λ4        --------G------A--G-G--------T-C-------C-A-----------------

cDNA 1525 TGGAAAAGGGAATCATTGATCCAACAAAG     ~87 bp intron 5'end
GC1       -------A---------A----------gt aaatgagtcagcaattctta aacattg
λ5        -A-----------------------------t--------------------------
λ4        -A-----A---------A------------ ----------------------t---

cDNA (1574)       ← P6961
GC1       gatggcttgttctatgg
λ5        -----------------
λ4        -----------------
```

Figure 10 **Alignment of the partial nucleotide sequence of the CHO hsp60 cDNA and three partial hsp60 pseudogenes**

Primers used in the PCR amplification of the partial λ4 and λ5 sequences from the lambda clone are shown in italics. The nucleotide sequence is numbered according to the position of this sequence in the hsp60 cDNA. The P7832 designed based on this alignment is shown in bold. Identical nucleotides are indicated by (----), deletions in the pseudogene sequences are indicated by (*), gaps in the cDNA or pseudogene sequences represent insertions. The gap in the λ4 sequence of ~230 bp corresponds to the region unsequenced in this clone. The ~87 bp intron sequence is shown in lowercase. The GC1 pseudogene sequence was previously reported by Venner *et al.*, 1990.

towards the hsp60 CHO functional gene. Using the alignment in Fig. 10, an 18 bp primer, P7832, was designed across a sequence corresponding to the cDNA which contained 7 bp changes within the 18 bp region. However, in preliminary testing by PCR, this primer amplified fragments of ~800 bp for some of the library positive clones tested, suggesting that this region may not be conserved in all pseudogenes and was not further pursued (data not shown).

At this stage of our work, Pochon and Mach, (1996), described the isolation of a functional genomic clone for human hsp60. Using the information, primers were created at the exon#2/intron#2, P8331, and the intron#3/exon#4, P8332, boundaries for amplification of an actual fragment of the human hsp60 for a cross-species hybridization (Fig. 9). However, testing of the human fragment on either Hela genomic DNA or CHO genomic DNA, digested with four different enzymes in a Southern analysis, resulted in non-specific binding of the probe (data not shown).

Shortly after this time, we discovered that the sequences of the rat hsp60 and hsp10 had been isolated by Ryan *et al.*, (1997). It had been shown that the position of introns within a functional gene are conserved across species (Ala-Kokko *et al.*, 1995). By comparing the positions of the first seven introns at the 5' end of the human hsp60 functional gene, with the rat hsp60 intron positions, it was found that they were exactly conserved. Based on the fact that rat and CHO are more evolutionarily closely related than rat and human, it was assumed that the positions of these introns would be the same in the CHO hsp60 functional gene. Comparing the exon#2/intron#2 boundary sequences for both rat and human hsp60, for which

the primer P8331 had previously been designed, they were found to be identical, so it was assumed that the CHO boundary sequence would also be similar. Using this primer and a reverse primer, P674, located ~12 bp downstream of the position of intron#3, (Fig. 9), PCR amplification was carried out using CHO-WT genomic DNA. These two primers resulted in amplification of an ~2.5 kb fragment which hybridized strongly to a probe made for ~0.7 kb of the 5' end region of the CHO hsp60 cDNA (data not shown). The ~2.5 kb fragment was then subcloned into the pCRII vector for sequencing using the universal primer, reverse primer and an internal primer, P730, within exon#3 (see Fig. 9). This revealed that the subcloned CHO fragment was identical to the hsp60 CHO cDNA sequence in this region, containing the two introns in the same positions as in the rat and human hsp60 functional genes. Therefore, this ~2.5 kb fragment was considered a small portion of the CHO hsp60 functional gene.

Because this fragment corresponded to an actual piece of the CHO hsp60 functional gene, it was decided that it would be used as probe for re-screening of the lambda EMBL3 CHO genomic library. In order to decrease the possibility that this probe fragment would bind to pseudogenes, primers were created within intron#3, P9914/P9915, to amplify by PCR a fragment containing only intron sequence (Fig. 9). PCR amplification with these primers resulted in a fragment of ~600 bp (data not shown), which was subcloned into the pCRII vector and ~500 bp of it sequenced to ensure it was the same as the previously sequenced intron#3.

3.2.4 Re-screening of the EMBL3 CHO Genomic Library

The screening of the CHO genomic library with the CHO hsp60 intron#3 probe yielded 12 positive clones. After repeated plating and screening with the intron#3 probe, three purified plaques were chosen from the plate corresponding to clone#3; 3A, 3B and 3C. The DNA for all three clones was isolated for Southern analysis using the hsp60 inton#3 probe. The probe hybrized to all three clones, suggesting they corresponded to the hsp60 CHO functional gene (Fig. 11).

The DNA from clones 3B and 3C was digested with several restriction enzymes, BamHI, ClaI, EcoRI, HindIII, and KpnI, and Southern analysis with intron#3 probe was carried out. The probe hybridized strongly to a fragment in the HindIII lane of ~5.5 kb for clone 3C, which was large enough to contain a significant size fragment of the functional gene and was an appropriate size for subcloning (Fig. 12). Subcloning of this fragment was carried out in pBluescript vector after a second Southern analysis of the gene cleaned fragment confirmed this fragment was correct (data not shown). A large number of blue and white colonies were present on the agar plate so that in order to identify a positive clone colony hybridization was carried out using the intron#3 probe fragment. Positives in the same positions on both filters were picked, their DNA isolated and digested with HindIII for another Southern analysis. Three of the clones that hybridized to the intron#3 probe fragment were then sequenced using various primers, including P9914, used to amplify the intron#3 probe fragment, P729/P730, present within exon sequences, and P8331/P8332, present at intron/exon boundaries (Fig. 9). However, these failed to bind to the inserts in these three

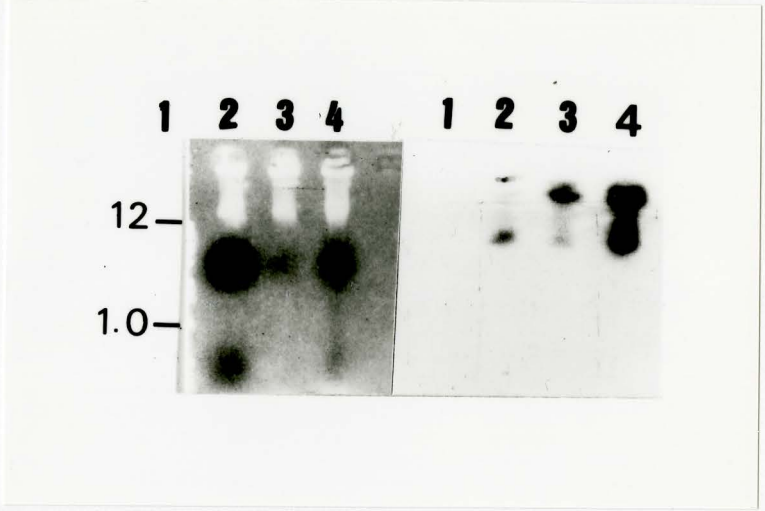

Figure 11 **Southern blot analysis on lambda DNA from 3A, 3B, and 3C CHO hsp60 positive clones.**

For each clone, ~6 µg of undigested lambda DNA was hybridized to the intron#3 probe fragment. A single high stringency wash for 5 min at 65°C in 0.2x SSC/0.1% SDS was necessary. The blot was exposed to X-ray film for ~4 hrs.
Lanes: 1= 1 kb DNA ladder, 2= 3A DNA, 3= 3B DNA, 4= 3C DNA

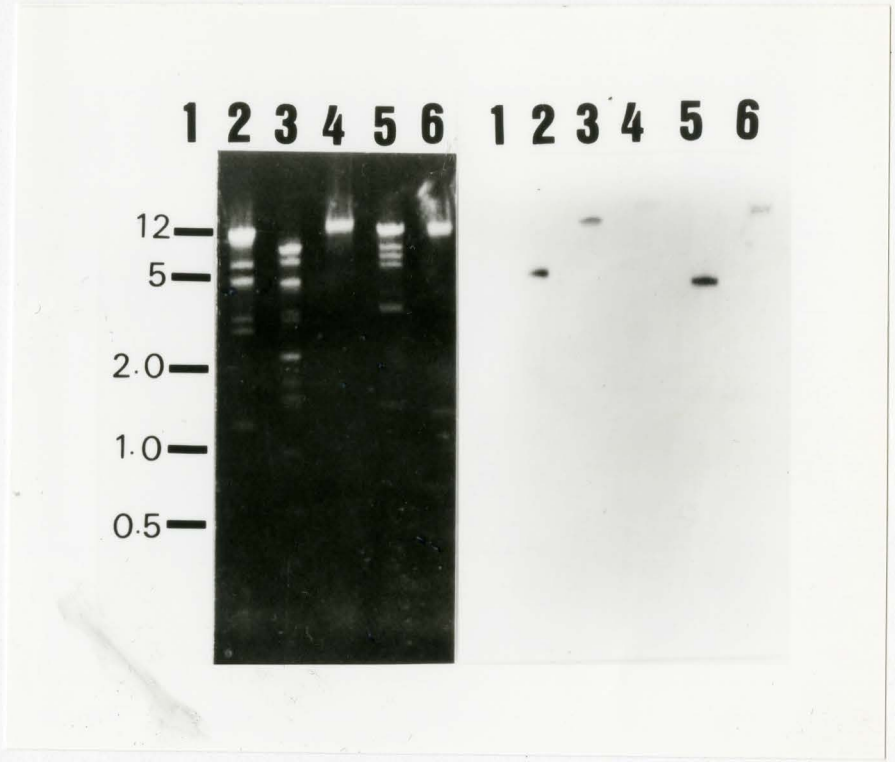

Figure 12 **Southern blot analysis of restriction enzyme digested lambda DNA from the 3C CHO hsp60 positive clone.**

Approximately 2 μg of lambda DNA was digested with the enzymes shown below for 2 hrs, then hybridized with the CHO hsp60 intron#3 probe fragment. One high stingency wash for 5 min at 65°C in 0.2x SSC/0.1% SDS was carried out and the blot exposed to X-ray film for ~7 hrs.

Lanes: 1= 1 kb DNA ladder, 2= BamHI, 3= ClaI, 4= EcoRI, 5= HindIII, 6= KpnI.

clones suggesting they did not correspond to the CHO hsp60 gene, but were positive for a contaminating fragment that may have been present with the hsp60 CHO intron#3 probe fragment. Based on time constraints at this point in the project, further investigation of this problem was not pursued. However, using the partial ~2.5 kb CHO hsp60 gene fragment isolated during this project, re-screening of the EMBL3 CHO genomic library should result in isolation of a full length clone for the CHO hsp60 functional gene.

4.0 DISCUSSION

4.1 Summary of Results

The main objectives of this project were to isolate the functional genes for both the hsp10 and hsp60 proteins from CHO-WT genomic DNA for characterization of structural features, such as upstream regulatory regions and exon/intron boundaries. Any pseudogene sequences isolated would be sequenced for comparison to the CHO hsp10 or hsp60 cDNA sequences, including flanking regions.

For CHO hsp10, screening of the lambda EMBL3 genomic library resulted in isolation of the putative full length functional gene, including ~240 bp upstream and ~200 bp downstream of the hsp10 cDNA sequence. Sequencing of the upstream region allowed for identification of regulatory elements, including a CAAT box, two Sp1 consensus sequences and a single heat shock element, HSE, however, a consensus TATA box or an initiator element, Inr, were not found. Screening of the CHO genomic library also resulted in isolation of a full length hsp10 pseudogene, 4Bi, which was sequenced including, ~300 bp upstream and ~200 bp downstream of the hsp10 cDNA sequence. Analysis of the upstream region of 4Bi revealed the presence of the 3' end of a human LINE sequence.

Amplification of CHO-WT genomic DNA with a nested set of primers designed based on the CHO hsp10 cDNA sequence, also resulted in isolation of five partial pseudogenes for hsp10. Southern analysis on CHO-WT genomic DNA digested with various enzymes, which

did not cut within the cDNA sequence, revealed numerous hybridizing fragments in most enzyme lanes. This further confirmed the existence of pseudogenes in the CHO genome for hsp10. In the past, no pseudogenes for hsp10 had been identified and this constitutes the first report on the presence of a large family of hsp10 pseudogenes in mammalian cells. Thus, as found for hsp60, hsp10 can be considered as a member of a multigene family, consisting of a functional gene and numerous non-functional pseudogenes.

Isolation of the full length CHO hsp60 functional gene was first attempted using information provided by the previous isolation of a semi-processed pseudogene, designated GC1 (Venner *et al.*, 1990). A probe was designed, which contained the ~87 bp intron from the GC1 pseudogene, for screening of the lambda EMBL3 CHO genomic library. Screening of the library resulted in isolation of two additional partial semi-processed pseudogenes, λ4 and λ5, which were similar to the GC1 pseudogene, containing the small ~87 bp intron and numerous unique mutations when compared to the hsp60 CHO cDNA sequence. Further attempts to isolate the CHO hsp60 functional gene were pursued utilizing information provided from the cloning of human hsp60 by Pochon and Mach, (1996), and rat hsp60 by Ryan *et al.*, (1997), which revealed conserved positions of introns in these sequences. Using a primer designed at an intron/exon boundary and a primer in the cDNA sequence, an ~2.5 kb fragment of the CHO hsp60 functional gene, containing one exon and two introns, was isolated. A smaller fragment, containing only intron sequence was then used to re-screen the lambda EMBL3 CHO genomic library. Three purified positive clones were isolated from the library, however, they did not bind to a number of hsp60 specific primers when sequenced.

The full length hsp60 functional gene remains to be isolated from the CHO genome, which can be carried out using the CHO hsp60 ~2.5 kb fragment cloned during this project.

4.2 Isolation of the CHO hsp10 Functional Gene and Identification of Pseudogenes

Southern blot analysis was performed on CHO-WT genomic DNA digested with six different restriction enzymes that did not cut within the hsp10 CHO cDNA sequence, except HindIII, which cut 53 bp into the 5' end. Based on the numerous bands present in each restriction enzyme lane, it can be estimated that at least 6-7 hsp10-related genes are found within the CHO genome. Considering that there is a least one functional copy of the gene present in the CHO genome, then the other fragments most likely correspond to non-functional pseudogene sequences. Analysis of the five partial pseudogenes amplified by PCR, revealed they contained many common mutations, which may lead to similar restriction enzyme patterns This would result in overlapping of fragments on a Southern analysis. Therefore, the actual number of hsp10-related sequences may be much higher than the 6-7 originally proposed. This appeared to be the case when analysis of hsp60 sequences from the human genome was carried out. Initially, the number of copies was estimated at 10-12 based on results from a Southern analysis, with restriction enzyme digested human DNA. However, this estimate was substantially increased to between 40-60 copies when dot blot analysis was performed, with at least 12 of these clones having been sequenced. Similarly, an estimate of at least 30 hsp60 related sequences for the CHO genome has been proposed, with one semi-processed clone, GC1, completely sequenced (Venner *et al.*, 1990; Gupta, unpublished).

Acting as a cofactor for hsp60, hsp10 shares many of the same characteristics, specifically the ability to act as both a heat shock protein and a molecular chaperone (see reviews by Lindquist and Craig, 1988; Rothman, 1989). This similarity may now be extended to the genomic level in terms of the association of pseudogenes.

Five pseudogenes were isolated from the CHO genome by utilizing PCR amplification with a nested set of primers on CHO-WT genomic DNA. The isolated clones are only partial pseudogenes for hsp10, since they corresponded to only ~270 bp of the central portion of the hsp10 CHO cDNA sequence. These clones were classified as processed pseudogenes for hsp10 based on their high degree of similarity towards the CHO hsp10 cDNA of ~90.1 % on average, their lack of introns, and the presence of numerous mutations which would prevent them from encoding functional protein. An alignment of all five pseudogenes with the hsp10 CHO cDNA sequence allowed for further analysis of these clones (see Fig. 5).

The Vent15 clone contains the highest similarity to the hsp10 cDNA sequence, of ~96 %, with mutations mainly due to single base pair changes. Analysis of the other four pseudogenes, reveals a hierarchy of identity, with hsp10-1A exhibiting ~91 % identity to the hsp10 CHO cDNA, followed by hsp10-2A, Vent2B and finally, hsp10-3A, the most divergent with only ~85 % similarity. Comparing these pseudogenes to each other, excluding Vent15, shows they share many of the same base pair changes, with at least 9 identical changes, as well as a common 3 bp deletion. In terms of similarity, they have an average of ~90.6 % similarity towards each other, this is including the Vent15 clone. This appears to be the same when they are compared to the cDNA, resulting in an average of ~ 90.1 % similarity. It is

known that the older a pseudogene is, the more time it has had to acquire mutations (Lewis and Cowan, 1986). Therefore, the high degree of identity for these pseudogenes, specifically Vent15, towards the CHO hsp10 cDNA suggests that their divergence occurred relatively recently.

These pseudogenes have maintained a high degree of homology towards each other, as well as sharing many common mutational characteristics, suggesting they may be part of a subfamily of pseudogenes. The existence of pseudogene subfamilies has previously been observed for human hsp60 pseudogenes, with 7 out of 9 clones sequenced containing many common mutations as well as sharing common flanking regions, creating the R subfamily of pseudogenes (Gupta, unpublished). However, the existence of a subfamily of pseudogenes for hsp10 is speculative since only partial sequence information of these clones is known. Isolation of additional full length clones, including flanking regions would aid in further characterizing the hsp10 multigene family, specifically in terms of the possible pseudogene subfamilies.

For all the PCR amplifications carried out on CHO-WT genomic DNA, the hsp10-1A pseudogene was the major isolated clone. The apparent abundance of this pseudogene may result from association with a DNA repetitive element, such as Alu or a LINE sequence. The association of pseudogenes with a DNA repetitive element, such as Alu or a LINE sequence, has been shown to be a relatively common occurrence. Some examples include, the human hexokinase II pseudogene, bovine α-lactalbumin pseudogene, human glucocerebrosidase pseudogene, as well as two human hsp60 pseudogenes (Horowitz *et al.*, 1989; Violette *et al.*,

1993; Ardehali *et al.*, 1995; and Gupta, unpublished). This association would allow for an increase in its rate of movement in the genome, by a retrotransposition mechanism which is one of the main methods suggested for pseudogene propagation (Vanin, 1985). However, because only the central sequence of the pseudogene was isolated, such that characteristics like a poly(A) tract, and flanking direct repeats, common for processed pseudogenes, are not known, this is only one possibility. Alternatively, many of the clones classified as hsp10-1A may actually correspond to unique clones. Because only ~270 bp of sequence for these clones is known, there is a possibility that this region is one of high conservation amoung certain pseudogenes. Large segments of conserved sequence has been previously observed for several hsp60 pseudogenes, with numerous unique mutations occurring outside these regions (Venner *et al.*, 1990). Therefore, if additional regions of these clones were isolated and sequenced, they may reveal several different mutations indicating they do not all correspond to the same pseudogene.

Screening of the lambda EMBL3 CHO genomic library also resulted in isolation of a full length pseudogene, designated 4Bi. This clone lacked introns, and contained numerous single base pair changes, several deletions and insertions, maintaining ~90 % similarity to the hsp10 CHO cDNA. When aligned with the full length functional hsp10 sequence, a relatively high degree of similarity is maintained between the two genes in the region downstream of the polyadenylation signal in the functional gene. The similarity is lost around bp position 512 when a stretch of ~8 A residues is reached in the 4Bi sequence. Upstream, the 4Bi and hsp10 functional genes are completely divergent past the -19 bp position. This is similar to

what was observed for the GC1 pseudogene from CHO DNA for hsp60. The full length GC1 pseudogene was more similar to the cDNA at the 3' end and very divergent at the 5' end (Venner *et al.*, 1990).

The region upstream of the putative start position in the 4Bi sequence showed homology to the DNA repetitive element, L1, a human LINE sequence. Specifically, the homology is towards the ORF2 consensus sequence which is found at the 3' end of many LINE sequences. This ORF2 codes for a protein with homology to reverse transcriptase sequences (Martin, 1991). In some cases functional LINEs have been observed that encode an active reverse transcriptase (Singer *et al.*, 1993). The association of a pseudogene with such an element could provide a means for its own propagation through the genome, since transcription of the upstream LINE element may lead to transcription of the pseudogene associated with it. Additionally, the presence of a full length functional LINE that can provide a source of active reverse transcriptase could result in production of a cDNA copy of the mRNA transcript for the functional gene, and therefore result in the production of new pseudogenes.

A comparison of the central ~270 bp of the 4Bi pseudogene with the five previously isolated PCR amplified pseudogenes showed a close similarity amoung all the clones, with the highest similarity to the Vent15 clone of ~93%. As suggested for the partial pseudogenes, the full length 4Bi clone may be a member of the same subfamily of pseudogenes. However, isolation of additional full length pseudogenes for hsp10, including flanking regions, is necessary in order to identify possible pseudogene subfamilies.

Only recently, the hsp10 functional gene was isolated from rat genomic DNA by Ryan *et al.*, (1997), being the first reported for hsp10 from any higher eukaryotic source. During this project, isolation of the hsp10 CHO functional gene resulted from screening of the CHO genomic library. Approximately 3.3 kb of the gene was sequenced which included the full coding region as well as ~240 bp upstream and ~200 bp downstream. Within this ~3.3 kb gene, 3 introns were found with intron#1 of ~600 bp, intron#2 of ~1.6 kb and intron#3 of ~200 bp, being separated by four small exons. The exon/intron boundaries were marked by typical consensus splice site sequences of 5' GT and 3' AG (Mount, 1982). Analysis of the upstream region of the hsp10 functional gene revealed the presence of typical regulatory sequences, including two Sp1 consensus sequences, a CAAT box and an HSE sequence (Gidoni *et al.*, 1984; Nassinnov *et al.*, 1986; Bienz, 1986). The HSE is commonly found upstream or within the first intron of heat shock protein genes to provide efficient expression during heat shock (Amin *et al.*, 1988; Rebbe et al., 1989). However, analysis of the first intron of the gene did not show any additional HSE sequences. Interestingly, a TATA box consensus sequence or an initiator element Inr, was not located upstream in the region that was sequenced. A number of promoters that lack these common consensus elements have been identified, mainly of the housekeeping classification, and have been demonstrated for other heat shock proteins, such as *Xenopus* hsp70 and human hsp60 genes (Bienz and Pelham, 1986; Quon, *et al.*, 1994; Pochon and Mach, 1996). These types of promoters also tend to to be GC rich, containing one or more Sp1 factor binding motifs, (Dynan, 1986), as found in the upstream region of the CHO hsp10 functional gene. Comparing the exon regions

of the functional gene with the cDNA sequence reveals they are identical, which provides further evidence that this sequence does correspond to the actual hsp10 functional gene.

4.3 Attempted Isolation of CHO hsp60 Functional Gene and Pseudogenes

Previously, approximately 40-60 pseudogenes for hsp60 had been estimated as present in the human genome, and possibly as many as 30 pseudogenes for hsp60 in the CHO genome. (Venner *et al.*, 1990; Gupta, unpublished). Isolation of a single semi-processed pseudogene, GC1, from the CHO genome was also found prior to this project (Venner *et al.*, 1990). The initial screening of the lambda EMBL3 CHO genomic library for hsp60, in this report, resulted in isolation of two additional partial semi-processed pseudogenes, λ4 and λ5. Both the λ4 and λ5 clones are very similar to the GC1 pseudogene, sharing a majority of the single bp changes, as well as the 5' end of the ~87 bp intron and a 12 bp deletion. However, these two clones can be considered individual semi-processed pseudogenes since they contain numerous bp changes unique to their own sequences. Additionally, λ5 lacks a 21 bp deletion found in both λ4 and GC1, while λ4 contains a 24 bp addition not found in either λ5 or GC1. However, these pseudogenes do share many common changes and have a high degree of homology to each other, for example λ5 and GC1 are ~95 % similar, suggesting that they belong to a subfamily of semi-processed pseudogenes which may have arisen from a common ancestral pseudogene. As previously mentioned, the existence of pseudogene subfamilies was first identified during investigations on pseudogenes present for the human hsp60 gene. Out of 9 pseudogene clones sequenced, 7 were found to correspond to pseudogenes with very

similar mutations as well as similar flanking regions, and were suggested to belong to a subfamily, having arisen from a common ancestral pseudogene (Gupta, unpublished). In the case of the λ4 and λ5 clones, only partial sequence information was found, such that flanking regions can not be compared. Thus, it can only be suggested that these pseudogenes are part of a smaller subfamily of hsp60 pseudogenes, along with GC1.

Generally, semi-processed pseudogenes arise from a functional gene by gene duplication events that allow for the retention of introns (Dyer *et al.*, 1989; Ueyama *et al.*, 1993). The semi-processed pseudogenes, λ4 and λ5 as well as GC1, have retained only one intron from the hsp60 functional gene of their origin. However, based on sequence information available from the cloning of the human hsp60 by Pochon and Mach, (1996), and of rat hsp60, Ryan *et al.*, (1997), 7 introns were found within the first ~7.0 kb of the 5' end of the functional gene. Therefore, the presence of only a single intron in these semi-processed pseudogenes is puzzling. It is possible that these pseudogenes may have arisen from an ancestral pseudogene that contained only this single intron. This ancestral pseudogene in turn may have arisen from an incompletely spliced mRNA transcript of the hsp60 functional gene, which was copied to cDNA and inserted randomly into the genome. Subsequent gene duplication events would give rise to progeny pseudogenes, creating a pseudogene subfamily that contained only this one intron.

Subsequent studies, utililizing the information available from the rat and human hsp60 genes, allowed for PCR primers to be designed resulting in amplification a ~2.5 kb fragment of the CHO functional gene from CHO-WT genomic DNA. This fragment was found to

contain two introns in identical positions as in rat and human hsp60 sequences and was identical to the CHO hsp60 cDNA in the partial exon region that was sequenced. Using only a ~500 bp fragment of intron#3 as probe, the CHO EMBL3 genomic library was re-screened. Because of the greater specificity of this probe, fewer positive clones were isolated than in the previous library screen. Unfortunately, the 3 purified positive clones isolated from this library screen, failed to bind to various hsp60 primers for sequencing and were considered false positives. However, because an actual ~2.5 kb fragment of the CHO hsp60 functional gene has been isolated, it will be easier to re-screen a CHO genomic library, to isolate the full length CHO hsp60 functional gene.

4.4 Proposed Future Research

The full length pseudogene, 4Bi, is the first identified for hsp10. This pseudogene can be used as probe to further isolate full length hsp10 pseudogenes. Sequencing of additional pseudogenes, including flanking regions will allow for further characterization of their relatedness in terms of proposed pseudogene subfamilies. Isolation of additional pseudogenes will also allow for a more accurate estimation as to the size of this multigene family. It is possible that the hsp10 family is quite large, similar to that for hsp60. Information obtained from flanking regions of additional pseudogenes may provide insight into why these pseudogene families are so large. The upstream and downstream sequences may also demonstrate association of DNA repetitive elements, such as observed for 4Bi. This may offer one reason why these sequences are spreading throughout the genome.

The identification of the full length functional gene for CHO hsp10, provided information concerning its upstream regulatory region. Through the techniques of S1 nuclease mapping or primer extension, the transcription initiation site can be more accurately identified. Subcloning of a large fragment of this region into a construct containing a reporter gene would allow for studies on the regulation of hsp10 expression, such as constitutive levels and inducibility upon heat shock. Identification of the full length CHO hsp60 functional gene carried out using the ~2.5 kb CHO gene fragment cloned in this study would allow for similar regulation studies. Particularly, a comparison of the two genes in terms of sequence and promoter element conservation and levels of expression would be carried out. Because hsp60 and hsp10 function together, it would be interesting to see if they are also co-ordinately regulated by a similar mechanism. Similarly, *in situ* hybridization experiments with both hsp10 and hsp60 would map their chromosomal positions in the genome, perhaps demonstrating localization in close proximity to each other on the same chromosome.

The association of numerous pseudogenes with the hsp60 protein has previously been reported by Venner *et al.*, (1990), for both CHO and human genomes. The identification of the association of pseudogenes with the CHO hsp10 here is the first reported for this protein. Many have suggested that pseudogenes are present in the genome, not as genetic lesions, but because they provide a source of genetic diversity, the raw material for evolution (Wagner, 1986; Fotaki and Iatrou, 1993; Varga-Madrazo *et al.*, 1995). The hsp60 protein has diversified roles, including involvement in protein folding both constitutively, and upon heat shock, localization at various extramitochondrial sites, roles in various autoimmune diseases

and even association with drug resistance mechanisms. Acting as a cofactor for hsp60 and recently identified as the early pregnancy factor, hsp10 is also an important protein. Perhaps the assocation of numerous pseudogenes with these heat shock proteins aids in their evolution in the constantly changing mammalian genome.

5.0 REFERENCES

Ala-Kokko, L., Kvist, A., Metsaranta, M., Kivirikko, K.I., Crombrugghe, B. de, Prockop, D.J., and Vuorio, E. (1995) Conservation of the sizes of 53 introns and over 100 intronic sequences for the binding of common transcription factors in the human and mouse genes for type II procollagen (COL2A1) Biochem. J. 308: 923-929.

Amin, J., Anathan, J., Voellmy, R. (1988) Key features of heat shock regulatory elements. Mol. Cell Biol. 8: 3761-3769.

Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. Science 181: 223-230.

Ardehali, H., Printz, R.L., Koch, S., Phillips III, J.A., and Granner, D.K. (1995) Isolation, characterization, and chromosomal localization of a human pseudogene for hexokinase II. Gene 164: 357-361.

Azem, A., Kessel, M., and Goloubinoff, P. (1994) Characterization of a functional $GroEL_{14}(GroES_7)_2$ chaperonin hetero-oligomer. Science 265: 653-655.

Bardwell, J.C.A., and Craig, E.A. (1987) Eukaryotic M_r 83,000 heat shock protein has a homologue in *Escherichia coli.* PNAS USA 84: 5177-5181.

Barraclough, R., and Ellis, J.R., (1980) Protein synthesis in choloroplasts IX, assembly of newly-synthesized large subunits into ribulose bisphosphate carboxylase in isolated intact pea chloroplasts. Biochim. Biophys. Acta 608: 19-31.

Becker, J., and Craig, E.A., (1994) Heat shock proteins as molecular chaperones. Eur.J. Biochem. 219: 11-23.

Bertsch, U., Soll, J., Seetharam, R., and Viitanen, P.V. (1992) Identification, characterization, and DNA sequence of a functional "double" groES-like chaperonin from chloroplasts of higher plants. PNAS USA 89: 8696-8700.

Bienz, M. (1985) Transient and developmental activation of heat-shock genes. TIBS 10: 157-161.

Bienz, M. (1986) A CCAAT box confers cell-type-specific regulation on the *Xenopus* hsp70

gene in oocytes. Cell 46: 1037-1042.

Bienz, M., and Pelham, H.R.B. (1986) Heat shock regulatory elements function as an inducible enhancer in the *Xenopus* hsp70 gene and when linked to a heterologous promoter. Cell 45: 753-760.

Birnboim, H.C. (1983) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 1.6.1.

Blin, N., and Stafford, D.W. (1976) In Molecular Cloning: A laboratory manual, second edition. Sambrook, J., Fritsch, E.F., and Maniatis, T. eds. Cold Spring Harbor Laboratory Press, pp. 9.16.

Braig, K., Otwinowski, Z., Hedge, R., Boisvert, D.J., Joachimiak, A., Horwich, A.L., and Sigler, P.B. (1994) The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. Nature 371: 578-586.

Brown, T. (1993) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 2.10.2.

Buchner, J. (1996) Supervising the fold: functional principles of molecular chaperones. FASEB 10: 10-19.

Cavanagh, A.C., and Morton, H. (1994) The purification of early-pregnancy factor to homogeneity from human platelets and identification as chaperonin 10. Eur.J. Biochem. 222: 551-560.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371: 215-220.

Chen, J.J., McNealy, D.J., Dalal, S., and Androphy, E.J. (1994) Isolation, sequence analysis and characterization of a cDNA encoding human chaperonin 10. Biochim. Biophys. Acta 1219: 189-190.

Cheng, M.Y., Hartl, F-U., Martin, J., Pollock, R.A., Kalousek, F., Neupert, W., Hallberg, E.M., Hallberg, R.L., and Horwich, A.L. (1989) Mitochondrial heat-shock protein hsp60 is essential for assembly of proteins imported into yeast mitochondria. Nature 337: 620-625.

Christensen, H., and Pain, R.H. (1991) Molten globule intermediates and protein folding. Eur.

Biophys. J. 19: 221-229.

Cohen, I.R. (1991) Autoimmunity to chaperonins in the pathogenesis of arthritis and diabetes. Ann. Rev. Immunol. 9: 567-589.

Cowing, D.W., Bardwell, J.C.A., Craig, E.A., Woolford, C., Hendrix, R.W., and Gross, C.A. (1985) Consensus sequence for *Escherichia coli* heat shock gene promotors. PNAS USA 82: 2679-2683.

Craig, E.A., and Gross, C.A. (1991) Is hsp70 the cellular thermometer? TIBS 16: 135-139.

Dagert, M., and Ehrlich, S.D. (1974) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 1.8.1.

Daniels, G.R., and Deininger, P.L. (1985) Repeat sequence families derived from mammalian tRNA genes. Nature 317: 819-822.

Dombroski, B.A., Feng, Q., Mathias, S.L., Sassaman, D.M., Scott, A.F., Kazazian Jr., H.H., and Boeke, J.D. (1994) An *in vivo* assay for the reverse transcriptase of human retrotransposon L1 in *Saccharomyces cerevisiae.* Mol. Cell Bio. 14: 4485-4492.

Dyer, M.R., Gay, N.J., and Walker, J.E. (1989) DNA sequences of a bovine gene and of two related pseudogenes for the proteolipid subunit of mitochondrial ATP synthase. Biochem. J. 260: 249-258.

Dynan, W.S. (1986) Promoters for housekeeping genes. Trends Genet. 2: 196-197.

Ellis, J. (1987) Proteins as molecular chaperones. Nature 328: 378-379.

Fenton, W.A., Kashi, Y., Furtak, K., and Horwich, A.L. (1994) Residues in chaperonin GroEL required for polypeptide binding and release. Nature 371: 614-619.

Fischer, H.M., Babst, M., Kaspar, T., Acuna, G., Arigoni, F., and Hennecke, H. (1993) One member of a groESL-like chaperonin multigene family in *Bradyrhizobium japonicum* is co-regulated with symbiotic nitrogen fixation genes. EMBO J 12: 2901-2912.

Foord, S.M., Marks, B., Stolz, M., Bufflier, E., Fraser, N.J., and Lee, M.G. (1996) The structure of the prostaglandin EP4 receptor gene and related pseudogenes. Genomics 35: 182-188.

Fotaki, M.E., and Iatrou, K. (1993) Silk moth chorion pseudogenes: Hallmarks of genomic evolution by sequence duplication and gene conversion. J. Mol. Evol. 37: 211-220.

Georgopoulos, C.P., Hendrix, R.W., Casjens, S.R., and Kaiser, A.D. (1973) Host participation in bacteriophage lambda head assembly. J. Mol. Biol. 76: 45-60.

Gidoni, D., Dynan, W.S., and Tijan, R. (1984) Multiple specific contacts between a mammalian transcription factor and its cognate promoter. Nature 312: 409-413.

Gupta, R.S. (1981) Resistance to the microtubule inhibitor podophyllotoxin: selection and partial characterization of mutants in CHO cells. Som. Cell Genet. 7: 59-71.

Gupta, R.S., Ho, T.K.W., Moffat, M.R.K., and Gupta, R. (1982) Podophyllotoxin-resistant mutants of Chinese hamster ovary cells. JBC 257: 1071-1078.

Gupta, R.S., and Gupta, R. (1984) Mutants of Chinese hamster ovary cells affected in two different microtubule-associated proteins. JBC 259: 1882-1890.

Gupta, R.S., Venner, T.J., and Chopra, A. (1985) Genetic and biochemical studies with mutants of mammalian cells affected in microtubule-related proteins other than tubulin: mitochondrial localization of microtubule-related protein. Can. J. Biochem. Cell Bio. 63: 489-502.

Gupta, R.S. (1990) Sequence and structural homology between a mouse T-complex protein TCP-1 and the 'chaperonin' family of bacterial (GroEL, 60-65 kDa heat shock antigen) and eukaryotic proteins. Biochem. Internat. 20: 833-841.

Gupta, R.S., and Golding, G.B. (1993) Evolution of HSP70 gene and its implications regarding relationships between Archaebacteria, Eubacteria and Eukaryotes. J. Mol. Evol. 37: 573-582.

Gupta, R.S. (1995) Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol. Microbio. 15: 1-11.

Gupta, R.S. (1996) Evolutionary relationships of chaperonins. In The chaperonins. Ellis, R.J. ed. Academic Press, pp. 27-57.

Hanahan, D. (1983) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 1.8.2.

Hanahan, D., and Meselson, M. (1983) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 6.2.2.

Hartl, F-U., Hlodan, R., and Langer, T. (1994) Molecular chaperones in protein folding: the art of avoiding sticky situations. TIBS 19: 20-25.

Hartman, D.J., Hoogenraad, N.J., Condron, R., and Hoj, P.B., (1992) Identification of a mammalian 10-kDa heat shock protein, a mitochondrial chaperonin 10 homologue essential for assisted folding of trimeric ornithine transcarbamolylase *in vitro*. PNAS USA 89: 3394-3398.

Heinrich, P. (1993) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 7.3.7.

Hemmingsen, S.M., Woolford, C., van der Vies, S.M., Tilly, K., Dennis, D.T., Georgopoulos, C.P., Hendrix, R.W., and Ellis, R.J. (1988) Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature 333:330-334.

Hohfeld, J., and Hartl, F-U. (1994) Role of the chaperonin cofactor hsp10 in protein folding and sorting in yeast mitochondria. J. Cell Bio. 126: 305-315.

Horowitz, M., Wilder, S., Horowitz, Z., Reiner, O., Gelbart, T., and Beutler, E. (1989) The human glucocerebrosidase gene and pseudogene: Structure and evolution. Genomics 4: 87-96.

Hunt, J.F., Weaver, A.J., Landry, S.J., Gierasch, L., and Deisenhofer, J. (1996) The crystal structure of the GroES co-chaperonin at 2.8 Å resolution. Nature 379: 37-45.

Jackers, P., Clausse, N., Fernandez, M-T., Berti, A., Princen, F., Wewer, U., Sobel, M.E., and Castronovo, V. (1996) Seventeen copies of the human 37 kDa receptor presursor/p40 ribosome-associated protein gene are processed pseudogenes arisen from retropositional events. Biochim. Biophys. Acta 1305: 98-104.

Jacq, C., Miller, J.R., and Brownlee, G.G. (1977) A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell 12: 109-120.

Jindal, S., Dudani, A.K., Singh, B., Harley, C.B., and Gupta, R.S. (1989) Primary structure of a human mitochondrial protein homologous to the bacterial and plant chaperonins and to the 65-kilodalton mycobacterial antigen. Mol. Cell Bio. 9: 2279-2283.

Jones, M., Gupta, R.S., and Englesberg, E. (1994) Enhancement in amount of P1 (hsp60) in mutants of Chinese hamster ovary (CHO-K1) cells exhibiting increases in the A system of amino acid transport. PNAS USA 91: 858-862.

Katatos, F.C., Jones, C.W., and Efstratiadis, A. (1979) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 2.9.18.

Langer, T., Lu, C., Echols, H., Flanagan, J., Hayer, M.K., and Hartl, F-U. (1992) Successive action of DnaK, DnaJ, and GroEL along the pathway of chaperone-mediated protein folding. Nature 356: 683-689.

Laskey, R.A., Honda, B.M., Mills, A.D., and Finch, J.T. (1978) Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. Nature 275: 416-420.

Laskey, R.A., and Earnshaw, W.C. (1980) Nucleosome assembly. Nature 286: 763-767.

Leavitt, J., Gunning, P., Porreca, P., Ng, S.-Y., Lin, C.-S., and Kedes, L. (1984) Molecular cloning and characterization of mutant and wild-type human beta-actin genes. Mol. Cell Biol. 4: 1961-1969.

Lewis, S.A., and Cowan, N.J. (1986) Tubulin pseudogenes as markers for hominoid divergence. J. Mol. Biol. 187: 623-626.

Lindquist-McKenzie, S., Henikoff, S., and Meselson, M. (1975) Localization of RNA from heat-induced polysomes at puff sites in *Drosophila melongaster*. PNAS USA 72: 1117-1121.

Lindquist, S. (1981) Regulation of protein synthesis during heat shock. Nature 293: 311-314.

Lindquist, S. (1986) The heat-shock response. Ann. Rev. Biochem. 55: 1151-1191.

Lindquist, S., and Craig, E.A. (1988) The heat-shock proteins. Ann. Rev. Biochem. 22: 631-677.

Llorca, O., Marco, S., Carrascosa, J.L., and Valpuesta, J.M. (1994) The formation of symmetrical GroEL-GroES complexes in the presence of ATP. FEBS Letts. 345: 181-186.

Llorca, O., Schneider, K., Carrascosa, J.L., Mendez, E., and Valpuesta, J.M. (1997) Role of the amino terminal domain in GroES oligomerization. Biochim. Biophys. Acta 1337: 47-56.

Lubben, T.H., Gatenby, A., Donaldson, G.K., Lorimer, G.H., and Viitanan, P.V. (1990) Identification of a groES-like chaperonin in mitochondria that facilitates protein folding. PNAS USA 87: 7683-7687.

Mande, S.C., Mehra, V., Bloom, B.R., and Hol, W.G.J. (1996) Structure of the heat shcok protein chaperonin-10 of *Mycobacterium leprae*. Science 271: 203-206.

Martin, J., Langer, T., Boteva, R., Schramel, A., Horwich, A.L., and Hartl, FU. (1991) Chaperonin-mediated protein folding at the surface of groEL through a 'molten globule'-like intermediate. Nature 352: 36-42.

Martin, S.L. (1991) LINEs Curr. Opin. Genet. Develop. 1: 505-508.

Martin, J., Mayhew, M., Langer, T., and Hartl, F-U. (1993) The reaction cycle of GroEL and GroES in chaperonin-assisted protein folding. Nature 366: 228-233.

Mathias, S.L., Scott, A.F., Kazazian, H.H., Boeke, J.D., and Gabriel, A. (1991) Reverse transcriptase encoded by a human transposable element. Science 254: 1808-1810.

Mayhew, M. and Hartl, F.U. (1996) Lord of the Rings: GroES structure. Science 271: 161-162.

Mayhew, M., da Silva, A.C.R., Martin, J., Erdjument-Bromage, H., Tempst, P., and Hartl, FU. (1996) Protein folding in the central cavity of the GroEL-GroES chaperonin complex. Nature 379: 420-426.

Mazodier, P., Guglielmi, G., Davies, J., and Thompson, C.J. (1991) Characterization of the groEL-like genes in *Streptomyces albus*. J. Bacteriol. 173: 7382-7386.

McLennan, N.F., Girshovich, A.S., Lissin, N.M., Charters, Y., and Masters, M. (1993) The strongly conserved carboxy-terminus glycine-methionine motif of the *Escherichia coli* GroEL chaperonin is dispensible. Mol. Microbiol. 7: 49-58.

McMullin, T.W., and Hallberg, R.L. (1987) A normal mitochondrial protein is selectively synthesized and accumulated during heat shock in *Tetrahymena thermophila*. Mol. Cell Biol. 7: 4414-4423.

Moos, M., and Gallwitz, D. (1983) Structure of two human β-actin-related processed genes, one of which is located next to a simple repetitive sequence. EMBO J. 2: 757-761.

Mount, S. M. (1982) A catalogue of splice junction sequences. Nuc. Acid Res. 10: 459-472.

Musgrove, J.E., Johnson, R.A., and Ellis, J.R. (1987) Dissociation of the ribulose bisphosphate-carboxylase large-subunit binding protein into dissimilar subunits. Eur. J. Biochem. 163: 529-534.

Nakai, A., Tanabe, M., Kawazoe, Y., Inazawa, J., Morimoto, R.I., and Nagata, K. (1997) HSF4, a new member of the human heat shock factor family which lacks properties of a transcriptional activator. Mol. Cell Bio. 17: 469-481.

Narberhaus, F., and Bahl, H. (1992) Cloning, sequencing and molecular analysis of the *groESL* operon of *Clostridium acetobutylicum.* J. Bacteriol. 174: 3282-3289.

Novick, G.E., Batzer, M.A., Dininger, P.L., and Herrera, R.J. (1996) The mobile genetic element *Alu* in the human genome. Bioscience 46: 32-41.

Nussinov, R., Owens, J., and Maizel, J.V. Jr. (1986) Sequence signals in eukaryotic upstream regions. Biochim. Biophys. Acta 866: 109-119.

Packer, A.I., Manova, K., and Backvarova, R.F. (1993) A discrete LINE-1 transcript in mouse blastocysts. Develop. Bio. 157: 281-283.

Picketts, D.J., Mayanil, C.S.K., and Gupta, R.S. (1989) Molecular cloning of a Chinese hamster mitochondrial protein related to the "chaperonin" family of bacterial and plant proteins. JBC 264: 12001-12008.

Pilkington, S.J., and Walker, J.E. (1993) Complementary DNA sequence of bovine cpn10 (Hsp10), a chaperone protein from mitochondria. DNA Seq. 3: 291-295.

Pochon, N.A., and Mach, B. (1996) Genetic complexity of the human hsp 60 gene. Int. Immun. 8:221-230.

Quon, D.V., Delgadillo, M.G., Khachi, A., Smale, S.T., and Johnson, P.J. (1994) Similarity between a ubiquitous promoter element in an ancient eukaryote and mammalian initiator elements. PNAS USA 91: 4579-4583.

Rebbe, N.F., Hickman, W. S., Ley, T.J., Stafford, D.W., and Hickman, S. (1989) Nucleotide sequence and regulation of a human 90-kDa heat shock protein gene. JBC 264: 15006-15011.

Rothman, J.E. (1989) Polypeptide chain binding proteins: Catalysts of protein folding and related processes in cells. Cell 59: 591-601.

Ryan, M.T., Hoogenraad, N.J. and Hoj, P.B. (1994) Isolation of a cDNA clone specifying rat

chaperonin 10, a stress-inducible mitochondrial matrix protein synthesized without a cleavable presequence. FEBS Letts. 337: 152-156.

Ryan, M.T., Naylor, D.J., Hoogenraad, N.J., and Hoj, P.B. (1995) Affinity purification, overexpression, and characterization of chaperonin 10 homologues synthesized with and without N-terminal acetylation. JBC 270: 22037- 22043.

Ryan, M.T., Herd, S.M., Sberna, G., Samuel, M.M., Hoogenraad, N.J., and Hoj, P.B. (1997) The genes encoding mammalian chaperonin 60 and chaperonin 10 are linked head-to-head and share a bidirectional promoter. Gene 196: 9-17.

Schatz, G. (1987) Signals guide proteins to their correct locations in mitochondria. Eur. J. Biochem. 165: 1-6.

Scheutz, T.J., Gallo, G.J., Sheldon, L., Tempst, P., and Kingston, R.E. (1991) Isolation of a cDNA for HSF2: Evidence for two heat shock factor genes in humans. PNAS USA 88: 6911-6915.

Seale, J.W., and Horowitz, P.M. (1995) The C-terminal sequence of the chaperonin GroES is required for oligomerization. JBC 270: 30268-30270.

Seidman, C.E. (1989) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 1.8.1-1.8.2.

Singer, M.F., and Skowronski, J. (1985) Making sense out of LINES: Long interspersed repeat sequences in mammalian genomes. TIBS 10: 119-121.

Singer, M.F., Krek, V., McMillan, J.P., Swergold, G.D., and Thayer, R.E. (1993) LINE-1: a human transposable element. Gene 135: 183-188.

Soltys, B.J., and Gupta, R.S. (1996) Immunoelectron microscopic localization of the 60-kDa heat shock chaperonin protein (Hsp60) in mammalian cells. Exp. Cell Res. 222: 16-27.

Sorger, P.K., and Pelham, H.R.B. (1988) Yeast heat shock factor is an essential DNA-binding protein that exhibits temperature-dependent phosphorylation. Cell 54: 855-864.

Southern, E.M. (1975) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 2.9.3 - 2.9.7.

Staniforth, R.A., Burston, S.G., Atkinson, T., and Clarke, A.R. (1994) Affinity of chaperonin-60 for a protein substrate and its modulation by nucleotides and chaperonin-10. Biochem. J. 300: 651-658.

Storti, R.V., Scott, M.P., Rich, A., and Pardue, M.L. (1980) Translational control of protein synthesis in response to heat shock in *D. melanogaster* cells. Cell 22: 825-834.

Straus, D.B., Walter, W.A., and Gross, C.A. (1987) The heat shock response of *E.coli* is regulated by changes in the concentration of σ^{32}. Nature 329: 348-356.

Tabor, S. (1989) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 3.10.2.

Tian, G., Vainberg, I.E., Tap W.D., Lewis, S.A., and Cowan, N.J. (1995) Specificity in chaperonin-mediated protein folding. Nature 375: 250-253.

Tilly, K., Erickson, J., Sharma, S., and Georgopoulos, C. (1986) Heat shock regulatory gene rpoH mRNA level increases after heat shcok in *Escherichia coli*. J. Bacteriol. 168: 1155-1158.

Tissiéres, A., Mitchell, H.K., and Tracy, U.M. (1974) Protein synthesis in salivary glands of *Drosophila melongaster*: Relation to chromosome puffs. J. Mol. Bio. 84: 389-398.

Todd, M.J., Viitanen, P.V., and Lorimer, G.H. (1994) Dynamics of the chaperonin ATPase cycle: Implications for facilitated protein folding. Science 265: 659-666.

Ueyama, H., Han-Xiang, D., and Ohkubo, I. (1993) Molecular cloning and chromosomal assignment of the gene for human Zn-α_2-glycoprotein. Biochem. 32: 12968-12976.

Vanin, E.F. (1985) Processed pseudogenes: Characteristics and evolution. Ann. Rev. Genet. 19: 253-272.

Vargas-Madrazo, E., Almagro, J.C., and Lara-Ochoa, F. (1995) Structural repertoire in V_H pseudogenes of immunoglobulins: Comparison with human germline genes and human amino acid sequences. J. Mol. Biol. 246: 74-81.

Venner, T.J., and Gupta, R.S. (1990a) Nucleotide sequence of rat hsp60 (chaperonin, GroEL homolog) cDNA. Nuc. Acid Res. 18: 5309.

Venner, T.J., and Gupta, R.S. (1990b) Nucleotide sequence of mouse HSP60 (chaperonin,

GroEL homolog) cDNA. Biochim. Biophys. Acta 1087: 336-338.

Venner, T.J., Singh, B., and Gupta, R.S. (1990) Nucleotide sequences and novel structural features of human and Chinese hamster hsp60 (Chaperonin) gene families. DNA Cell Bio. 9: 545-552.

Viitanen, P.V., Donaldson, G.K., Lorimer, G.H., Lubben, T.H., and Gatenby, A.A. (1991) Complex interactions between the chaperonin 60 molecular chaperone and dihydrofolate reductase. Biochem. 30: 9716-9723.

Vilotte, J.L., Soulier, S., and Mercier, J.C. (1993) Complete sequence of a bovine α-lactalbumin pseudogene: the region homologous to the gene is flanked by two directly repeated LINE sequences. Genomics 16: 529-532.

Wagner, M. (1986) A consideration of the origin of processed pseudogenes. Trends Genet. 2: 134-137.

Weiner, A.M., Deininger, P.L., and Efstratiadis, A. (1986) Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Ann. Rev. Biochem. 55: 631-661.

Weis, J.H. (1987) In Current Protocols in Molecular Biology, (1994), Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Sedman, J.G., Smith, J.A., and Struhl, K., eds., Greene Publishing Associates and Wiley-Interscience, pp. 6.2.1.

Wu, C. (1980) The 5'ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNaseI. Nature 286: 854-860.

Wu, C., Wilson, S., Walker, B., Dawid, I., Paisley, T., Zimarino, V., and Ueda, H. (1987) Purification and properties of *Drosophila* heat shock activator protein. Science 238: 1247-1253.

Xiao, H., and Lis, J.T. (1988) Transformation used to define key features of heat-shock response elements. Science 239: 1139-1142.

Yamamori, T., and Yura, T. (1982) Genetic control of heat-shock protein synthesis and its bearing on growth and thermal resistance in *Escherichia coli* K-12. PNAS USA 79: 860-864.

Yoshida, H., Yanagi, H., and Yura, T. (1995) Cloning and characterization of the mitochondrial HSP60-encoding gene of *Schizosaccharomyces pombe*. Gene 167: 163-166.

Young, D.B. (1990) Chaperonins and the immune response. Sem. Cell Biol. 1: 27-35.

Zweig, M., and Cummings, D.J. (1973) Cleavage of head and tail proteins during bacteriophage T5 assembly: Selective host involvement in the cleavage of a tail protein. J. Mol. Biol. 80: 505-518.