INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa • University of Ottawa

Molecular evolutionary conservation of the *Drosophila* α-amylase gene and its regulatory systems

Erin N. Yoshida

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for the
Doctor of Philosophy degree in the
Ottawa-Carleton Institute of Biology

Thèse soumise á
l'École des études supérieures et de la recherche
Université d'Ottawa
en vue de l'obtention de la maîtrise ès sciences à
L'Institut de biologie d'Ottawa-Carleton

1997

ACKNOWLEDGMENTS

I would first like to thank my supervisor, Dr. Donal Hickey, for all of the guidance and support which he has given me, even to the point of potentially suffering permanent hearing loss by attending my drumming shows. Thank you to Dr. Bernie Benkel who not only had the patience to sit on my advisory committee, but also provided many of the starting materials for the project, and shared his lab and technical staff with me for several months. On that note, thank you to Ying Fong for your technical help and entertaining political debates. Thank you to the other member of my advisory committee, Dr. John Vierula, who provided much insight into the biology of yeast which initially, was such a foreign organism for me. A much deserved thank you to Ed Taboada, for rendering the abstract *en français*. I would also like to thank the other members of Dr. Hickey's lab for their help and friendship, especially Kaarina Benkel, Ada Loverre-Chyurlia, Peter Foster and Shaojiu Wang. I will have fond memories of the days that we have all spent together in the subterranean labyrinth.

I am very thankful for all of the encouragement my family has given me over the years. You have always supported my decision to do a graduate degree, for reasons other than having a place to stay when you visit Eastern Canada. Last, Paul Finnie would be uncomfortable if I thanked him in a melodramatic manner for all that he has done for me, so a simple "thanks" will have to suffice.

ABSTRACT

Alpha-amylase is a digestive enzyme involved in the metabolism of starch. It is found in a wide variety of organisms, and its expression is often regulated by carbon source. The *D. melanogaster* α-amylase gene was used here to show the evolutionary conservation of both a structural gene as well as its glucose repression regulatory system. The repression machinery was shown to be functionally conserved between yeast and flies, as the α-amylase promoter was used to express luciferase-based reporter constructs in a glucose repressible manner in *Saccharomyces*. The repression motifs of the α-amylase promoter were mapped to a 126 bp fragment which contains two putative binding sites for MIG1, the yeast transcription factor involved in glucose repression. Further 5' deletions or removal of the 3' MIG1 binding site eliminated glucose repression. Additional evidence for conservation of this machinery was provided by the cloning of *Drosophila* genomic and cDNA copies for *SNF4*, a subunit of the yeast derepression complex. This gene codes for a 684 aa protein which features extended carboxy and amino terminal sequences relative to the known yeast and mammalian homologues. Southern analysis suggested the presence of two gene copies in *Drosophila*, which would agree with the multiple cDNA isoforms seen in mammals. Inferred phylogenies indicated that the short isoform seen in yeasts and most mammalian sequences group separately from the long isoforms seen in *Drosophila*, *C. elegans* and a second human sequence - indicating the possible organization of orthologous gene copies. Aside from low level similarity to a handful of IMP dehydrogenases and IMPDH-like proteins, *SNF4* is not significantly similar to other known proteins. Last, the evolution of the α-amylase gene itself was examined with regards to nucleotide and amino acid biases. Nucleotide bias was observed at all three codon positions, primarily as a pyrimidine preference in the coding strand. Due to a nucleotide bias at nonsynonymous sites, there were significant content differences in GC-rich and AT-rich amino acids depending on the GC content of the gene. Although an amino acid

bias might be expected to affect phylogenetic determination, with α-amylase this bias was shown to not have a strong influence on inferred gene trees.

RÉSUMÉ

L'alpha-amylase est une enzyme digestive utilisée pour la digestion de l'amidon. On la trouve dans une grande variété d'organismes, et la régulation de son niveau d'expression est largement affectée par les sources de carbone. Le gène α-amylase de *Drosophila melanogaster* a été utilisé pour montrer non seulement la conservation évolutionnaire d'un gène structurel, mais aussi, la conservation de son système de régulation, soit une répression du niveau d'expression controlée par le glucose. La conservation fonctionelle de ce mechanisme de répression entre *Saccharomyces* et *Drosophila* a été demontrée quand des ADN chimeriques, contenant la séquence codante de luciferase et le promoteur de l'α-amylase de *D. melanogaster,* ont été introduits dans *S. cerevisiae* et ont exprimé luciferase dans une façon dépendante du glucose. La cartographie des signaux de répression dans le promoteur de l'α-amylase a établi un fragment de 126 paires de bases qui contient deux sites potentiels pour MIG1, le facteur de transcription responsable pour la répression controlée par le glucose dans *S. cerevisiae*. Même si le niveau de base d'expression demeure inchangé, la répression controlée par le glucose est perdue par des délétions de la région 5' du fragment ou par l'enlèvement du site pour MIG1 du côté 3'. Evidence additionelle pour cette conservation de régulation par glucose a été demontrée par le clonage de copies génomiques et de ADN-c pour le gène homologue au gène SNF4 de *Saccharomyces* dans *Drosophila*. Dans *Saccharomyces*, le gène SNF4 code pour une sous-unité du complêxe responsable pour la dérépression. Dans *Drosophila*, ce gène code pour une protéine de 684 acides aminés qui, comparée aux séquences homologues de *Saccharomyces* et de mammifères connues, contient des extensions aux extremités C-terminal et N-terminal. Une analyse par Southern blot montre la présence de deux copies du gène dans le génome de *Drosophila*, ce qui conforme aux multiples copies des different isoformes ADN-c observées dans les mammifères. Une analyse phylogénétique indique que les isoformes longs tels que celui de *Drosophila*, de *C. elegans*, et d'un nouveau isoforme humain se groupent séparement des isoformes courts qui sont observés dans *Saccharomyces* et dans la plupart des

séquences des mammifères. Ceci pourrait indiquer l'organization des copies de gènes orthologues. Sauf pour une faible similarité aux déshydrogenases de IMP, SNF4 n'a pas d'autres similarités à d'autres protéines. Dernièrement, l'évolution du gène de l'α-amylase lui-même fut examinée pour des biais au niveaux des nucléotides et des acides aminés. Des déviations significatives one été observées aux trois positions des codons, surtout en forme d'une préférence pour les bases pyrimidiques dans le brin codant. Des différences ont été observées dans les acides aminés riches en GC et AT à cause du biais aux sites non-synonymes et celles-ci dependent sur le contenu GC de la séquence codante du gène. Pourtant, même si on devrait s'attendre à un effet du biais au niveau des acides aminés sur les reconstructions phylogénétiques, avec l'α-amylase cet effect n'a pas été observé.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABBREVIATIONS USED

aa	=	Amino acid
AMG	=	Anterior midgut
AMP	=	Adenosine-5'-monophosphate
AMPK	=	AMP-activated protein kinase
AT	=	Adenine and thymine
ATP	=	Adenosine-5'-triphosphate
bp	=	Base pair
cDNA	=	Complementary deoxyribonucleic acid
DNA	=	Deoxyribonucleic acid
dNTP	=	A mixture of 2'-deoxy-adenosine-5'-triphosphate, 2'-deoxy-cytosine-5'-triphosphate, 2'-deoxy-guanosine-5'-triphosphate and 2'-deoxy-thymidine-5'- triphosphate
EDTA	=	(Ethylenedinitrilo)tetraacetic acid
EST	=	Expressed sequence tag
GC	=	Guanine and cytosine
GMP	=	Guanosine-5'-monophosphate
IMP	=	Inosine-5'-monophosphate
kb	=	Kilobase pair
MOPS	=	4-Morpholinepropanesulfonic acid
mRNA	=	Messenger ribonucleic acid
PCR	=	Polymerase chain reaction
PEP	=	Phosphoenolpyruvate
pfu	=	Plaque forming unit
PMG	=	Posterior midgut
RACE	=	Rapid amplification of cDNA ends
SDS	=	Sodium dodecylsulfate
SSC	=	0.15 M sodium chloride, 0.015 M sodium citrate
Tris	=	2-Amino-2-(hydroxymethyl)-1,3-propanediol
tRNA	=	Transfer ribonucleic acid
URS	=	Upstream repression sequence
UTR	=	Untranslated region
YPD	=	Rich yeast media (1 % Bacto-Yeast Extract, 2 % Bactopeptone)

CHAPTER 1

INTRODUCTION

1.1 General introduction

Over geological time, evolution has worked its wonders generating vast amounts of genetic and phenotypic variation in its wake. As much as species diverge from one another, they are always anchored together through their descent from a common ancestor. Thus morphologically, taxa as disparate as humans and yeast might hardly seem akin; biochemically however, they are certainly distant cousins. A humbling amount of similarity in many genetic systems has been observed between we mammals and the unicellular eukaryotes. Fundamental molecular mechanisms are often well conserved, but can also show signs of divergence as they may be recruited for different functions in different evolutionary lineages. It is from this perspective that this thesis utilizes the *Drosophila melanogaster* α-amylase gene to illustrate both conservation and divergence in the evolution of a structural protein and the mechanisms which regulate it.

Chapter 1 begins with background information on the *D. melanogaster* α-amylase gene with respect to its organization, evolution, and regulation, including an introduction to glucose repression. This is followed by a general overview of eukaryotic gene regulation - with specific emphasis on the mechanisms which regulate glucose repression in yeast, using the *GAL* genes as a model system. Functional conservation of the regulatory machinery between flies and yeast is shown in Chapter 2, where the *Drosophila* promoter is shown to drive glucose repressible expression in *Saccharomyces*. Using this transgenic expression system, the promoter regions which govern glucose repression are determined for the α-amylase gene. Chapter 3 reports the cloning of *SNF4* from *Drosophila*. Homologues of this gene have been found in several eukaryotes, where it is one subunit in a protein complex that responds to

fluctuations in metabolic energy levels. The function of this complex varies from regulating glucose repression in yeast, to regulating certain macromolecule biosynthesis in mammals. Chapter 4 deals with evolutionary processes in more detail by examining the manner in which α-amylase coding sequences have evolved at both the nucleotide and amino acid levels. The conclusions and summary are presented last, in Chapter 5.

1.2 Description of alpha-amylases

Amylases are a family of enzymes involved in the catabolism of amylose and amylopectin - both of which are components of starch and glycogen. They are subdivided based on their mode of action and substrate specificity. Some of the best characterized members include α-amylase, β-amylase and isoamylase. Alpha-amylases randomly hydrolyze internal α-1,4-D-glucosidic bonds which link the glucose residues in a starch molecule, producing glucose polymers of variable length. In contrast, β-amylases are only active at α-1,4-linkages at the ends of a starch molecule and its only product is maltose. Isoamylases are also known as debranching amylases, as they cleave the 1,6-linkages which are found only at the branch sites (reviewed in Vihinen and Mäntsälä, 1989).

With the cloning of an α-amylase from an archaebacteria (Kobayashi *et al.*, 1994), α-amylases have now been found in all three taxonomic superkingdoms. The sheer number of cloned amylase genes, coupled with the ease in which the protein activity can be assayed, has resulted in α-amylase being one of the first proteins adopted for molecular biological studies (Vihinen and Mäntsälä, 1989). Alignments of eukaryotic, prokaryotic and archeal α-amylase protein sequences reveal five conserved regions (Rogers, 1985; Janacek, 1994). These conserved domains correspond to the $(\alpha/\beta)_8$ barrel protein structure which forms the substrate and calcium binding sites, as determined by X-ray crystallography (Buisson *et al.*, 1987; Boel *et al.*, 1989). Outside these 5 shared domains however, the primary sequence is much more labile - often less than 10 % sequence identity is seen (Nakajima *et al.*, 1986). Phylogenetic

reconstruction using alpha-amylases traditionally clusters the taxa into three groups: the fungi and yeasts; the plants; and last, the bacteria and animals, including *Drosophila melanogaster* (Janacek, 1994).

1.2.1 *The* Drosophila melanogaster *alpha-amylase gene*

For *Drosophila*, the earliest studies on α-amylase characterization date to the late 1960s (reviewed in Milanovic and Andjelkovic, 1993). Doane (1969a) was the first to purify this enzyme from *Drosophila* and also determined that at all stages in the fly life cycle, a single peptide of 54,500 Da is responsible for α-amylase activity. Cytological mapping studies place the α-amy locus between 54B and 55 on chromosome 2 (Kikkawa, 1964; Bahn, 1967; Doane, 1969b). In *Drosophila melanogaster*, this locus shows several polymorphic forms, as eight electrophoretic variants have been detected in both natural and laboratory strains (Doane *et al.*, 1993). Several homozygous strains show co-expression of two electrophoretic forms, which can be explained by the genomic organization of α-amylase as two tightly linked duplicated genes which are divergently transcribed (Boer and Hickey, 1986; Gemmill *et al.*, 1986). The different gene copies are distinguished by the descriptors "proximal" and "distal", reflecting their locations with respect to the centromere (Gemmill *et al.*, 1986).

The gene duplication event is thought to be relatively ancestral, as the genomic organization seen in *D. melanogaster* is conserved within the *melanogaster* subgroup (Payant *et al.*, 1988; Shibata and Yamazaki, 1995). Outside of this subgroup however, copy number is variable. Within the subgenus *Sophophora* for example, *D. pseudoobscura* has 3 amylase gene copies (Brown *et al.*, 1990), and 4 or possibly more have been detected *D. ananassae* (Da Lage *et al.*, 1992). In many other animals as well, amylase is coded for by a multigene family (*e.g.* Gumucio *et al.*, 1985).

The particular genomic arrangement of duplicated genes linked head-to-head seen in *D. melanogaster* has resulted in the non-independent evolution of the two gene copies. Hickey *et al.* (1991) first provided evidence for concerted evolution between the gene copies using an

alignment of the *D. melanogaster* proximal gene with both the proximal and distal copies from its sibling species, *D. erecta*. Within the first 909bp of the coding sequence, 39 nucleotide substitutions were observed between the proximal gene copies of the two species. Of these, 38 of these sites were shared between the proximal and distal copies of *D. erecta*. In total, the level of nucleotide divergence in the coding region was much higher between species (4.25 %) than the level between gene copies within one species (1 %). This was in direct contrast to the untranscribed flanking sequence, where interspecific divergence was measured at ~15 %, and intergenic divergence within each species was closer to 40-60 %. Their observations were supported by the results of Shibata and Yamazaki (1995), who compared proximal and distal α-amylase nucleotide sequences from 8 species within the *melanogaster* subgroup. Most noticeably, in a phylogeny they constructed using both gene copies from all 8 species, aside from one exception, the proximal copy clustered immediately with the distal copy from the same species. At the species level however, the tree topology was as expected, with the *melanogaster*, *yakuba* and *erecta* complexes forming distinct groupings. As the gene duplication event appears to predate the *melanogaster* subgroup speciation event, it is highly unlikely that the topology of the Shibata and Yamazaki tree (1995) is the result of recent gene duplications in each of the different lineages. Rather, it is proposed to be the result of sequence homogenization due to gene conversion. This occurs after a hairpin loop is formed between the two gene copies, bringing the two genes in very close proximity to one another. Such a structure allows gene conversion in transcriptionally active regions. In fact, the flanking sequence evolves completely independently of the other gene copy, as shown by phylogenies that were inferred using only flanking sequences from both the proximal and distal genes. These trees form two very distinct branches - all of the proximal sequences group in one branch, and all of the distal sequences in the other (Shibata and Yamazaki, 1995). Therefore, the sequence flanking the proximal gene is distinct from that surrounding the distal copy, which would not be the case if gene conversion extended into the flanking sequence. This model agrees with the observations that only the coding sequences show extremely low levels

of intergenic sequence divergence whereas the sequence which immediately flanks the transcribed region shows much higher levels of divergence (Hickey *et al.*, 1991).

1.2.2 *Regulation of the* Drosophila melanogaster *alpha-amylase gene*

The α-amylase system in *Drosophila melanogaster* is subject to three different forms of regulation: tissue specific, developmental and dietary (Milanovic and Andjelkovic, 1993). With regards to tissue specificity, several studies have shown that α-amylase production is limited to the anterior (AMG) and posterior (PMG) midgut regions, but not the middle midgut (Doane, 1969a; Abraham and Doane, 1978; Doane *et al.*, 1983). Both the AMG and PMG correspond to the digestive regions of the fly gastrointestinal system, whereas the middle midgut region is non-digestive and highly acidic. Tissue specific expression also changes in some strains with the age of the fly. Doane *et al.* (1983) showed that an $Amy^{2.3}map^{c}$ strain did not express α-amylase in the PMG immediately after eclosure, but expression was seen in this region 14 days later. Both tissue specific and developmental expression are thought to be regulated by a *trans*-acting locus known as *map*, which is linked to *Amy* (Doane *et al.*, 1983; Thompson *et al.*, 1992), although the mechanism has yet to be characterized.

The third form of *D. melanogaster* α-amylase gene regulation, dietary regulation, is the system with which this thesis is concerned. Reminiscent of prokaryotic and lower eukaryotic gene regulation, the *Drosophila* α-amylase gene is subject to glucose repression (reviewed in Hickey and Benkel, 1987). As glucose is a readily metabolized carbohydrate, the presence of dietary glucose results in a downregulation of enzymes required for alternate carbon source catabolism - such as maltase and sucrase, but the effect is most pronounced with α-amylase (Benkel *et al.*, 1985). This response is not seen in all insects, as the flour beetle (*Tribolium castaneum*) and house fly (*Musca domestica*) have amylases which are not responsive to carbohydrate source (Hickey and Benkel, 1987).

In *D. melanogaster*, this change in enzyme levels reflects a decrease in amylase activity due to glucose repression, rather than starch induction (Hickey and Benkel, 1982; Echo and

Doane, 1984; Benkel and Hickey, 1986a), although starch induction has been reported particularly in *D. busckii* and *D. kikkawai* (Inomata *et al.*, 1995). A diet containing as little as 1% glucose has been shown to repress α-amylase levels (Magoulas *et al.*, 1992). Benkel and Hickey (1986a and 1987) confirmed that the decrease in α-amylase activity reflects a decrease in mRNA quantity, rather than a post-translational modification of enzyme activity. At higher glucose concentrations, alpha-amylase mRNA levels are repressed more than 100 fold in the *D. melanogaster* wild type Oregon-R strain, although the degree of repression is dependent on both strain (Benkel and Hickey, 1986b) and species (Inomata *et al.*, 1995). For example, even amongst the *melanogaster* species group, *D. lini* and *D. eugracilis* exhibit very glucose repression even though they are closely related species to *D. melanogaster* (Inomata *et al.*, 1995). Last, the proximal gene copy is more susceptible to glucose repression than the distal copy (Benkel and Hickey, 1986b), so many studies including this one, focus solely on the proximal gene copy.

Although glucose repression of the α-amylase gene is not observed in all species within the *Drosophila* genus, it appears that the biochemical mechanism responsible for this form of gene regulation is conserved. Promoter analyses of the *D. melanogaster* proximal gene (Hawley *et al.*, 1992; Magoulas *et al.*, 1992) have shown that the elements which regulate glucose repression are found within a <500 bp fragment immediately upstream of the coding sequence. This is also the case in *Drosophila virilis*, a distant relative of *D. melanogaster* whose α-amylase gene is glucose repressible. When a 330 bp promoter from *D. virilis* was used to regulate a reporter gene construct in *D. melanogaster*, the reporter construct expressed in a glucose repressible manner, indicating interspecies conservation of the repression signal sequences (Magoulas *et al.*, 1993b). However, interspecies and intergenic alignments of the promoter sequences have not shed any light on conserved regions which may be responsible for glucose repression. Comparisons have been made between the *D. melanogaster* promoter and those of *D. erecta* (Hickey *et al.*, 1991), *D. teissieri* (Okuyama *et al.*, 1997) and 7 other species of the glucose repressible *melanogaster* species subgroup (Shibata and Yamazaki,

1995). As the maximum level of sequence divergence among these sequences is ~15 %, there was not enough sequence divergence to highlight specific regions of the promoter which showed high levels of sequence conservation that might indicate an essential regulatory element. When such alignments are extended beyond the *melanogaster* species group to include *D. pseudoobscura* and *D. virilis* however, sequence conservation is very low (Magoulas *et al.*, 1993b). Only three regions of high conservation could be detected in the promoters of these species. Two were common eukaryotic motifs (TATA and CAAAT boxes), and the third was thought to be a midgut-specific motif due to its presence in other *Drosophila* genes which show midgut-specific expression patterns but are non-glucose repressible. Therefore, although the α-amylase is glucose repressible in many species in the *Drosophila* genus, previous analyses of the structural gene and its promoter have yet to indicate possible mechanisms for such regulation.

1.3 Glucose metabolism in *Saccharomyces*

Gene regulation via glucose repression is very uncommon in higher eukaryotes, but it is a well known phenomenon in prokaryotic and eukaryotic microorganisms. Thus our understanding of the underlying mechanisms as seen in eukaryotes, is largely based on studies conducted in *Saccharomyces*.

In the budding yeast, *Saccharomyces cerevisiae*, glucose metabolism is quickly achieved by fermentation rather than oxidative respiration. This is thought to provide an adaptive advantage to the organism since the resulting ethanol can be further metabolized at a later time through the Krebs cycle - a trait which not all competing organisms possess. Briefly, the pathway for glucose metabolism in yeast is as follows (Figure 1.1). Glucose is transported into the cell where it is phosphorylated by hexokinase PII (*HXK2*). The product of this reaction, glucose-6-phosphate, is then isomerized to fructose-6-phosphate where it enters the glycolytic pathway. The end products of glycolysis are ATP and phosphoenolpyruvate (PEP),

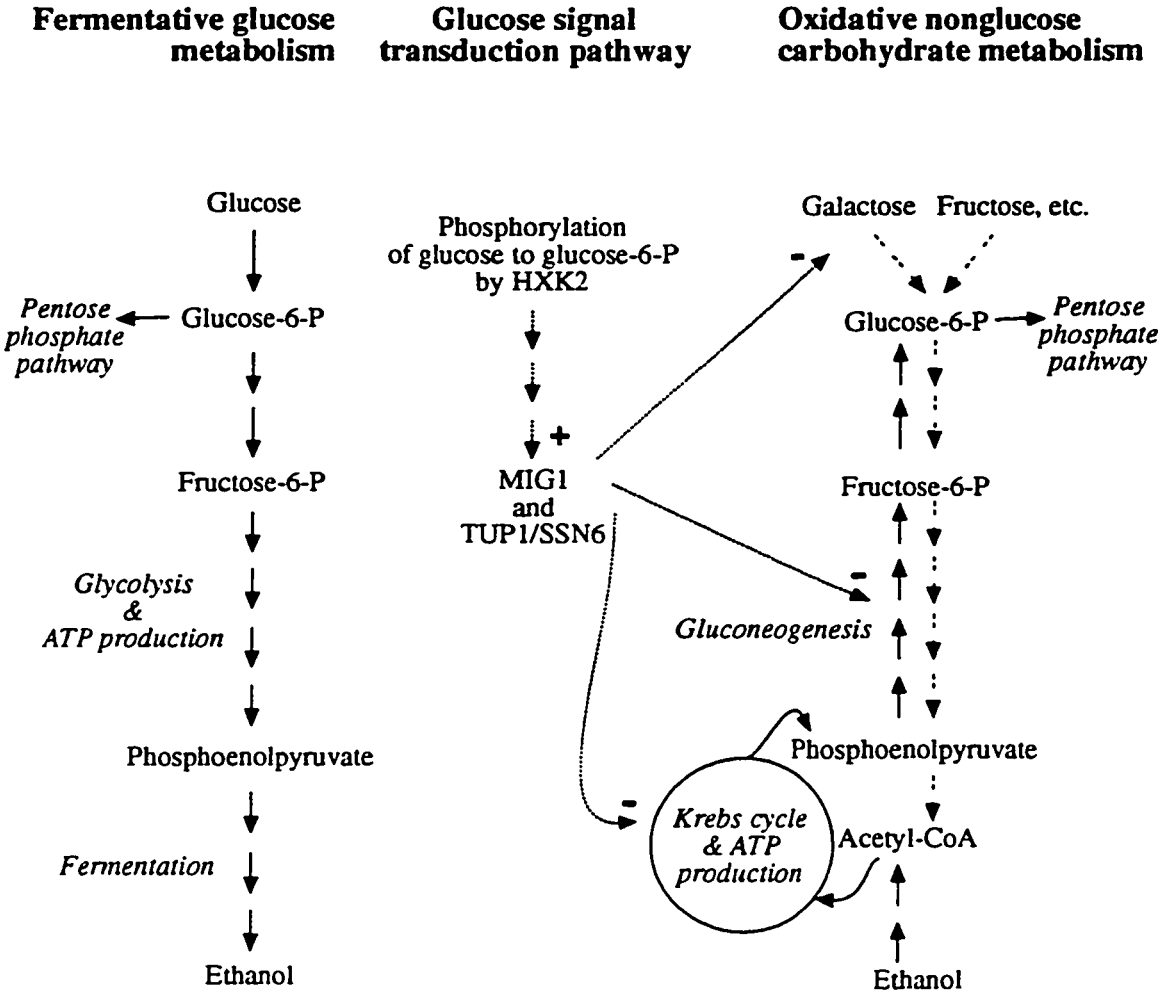


Figure 1.1. Overview of the components of fermentative and oxidative carbon source metabolism in *Saccharomyces cerevisiae* and their regulation by glucose. The fermentation of glucose is coupled to a signal transduction pathway which results in the repression of enzymes involved in alternate carbon source metabolism, gluconeogenesis, and mitochondrial respiration. "+" denotes upregulation and "-" downregulation. (Modified from Thevelein, 1994; Ronne, 1995).

which is subsequently fermented to produce ethanol as the final product (Rawn, 1989). When glucose becomes limiting, carbohydrate metabolism shifts from glucose fermentation to the use of the ethanol which has accumulated in the surrounding media, as well as other alternate carbon sources (such as galactose, maltose, sucrose, etc.) by the more efficient oxidative respiration system (reviewed in Entian and Barnett, 1992; Johnston and Carlson, 1992; Thevelein, 1994; Ronne, 1995). In the case of ethanol catabolism, glucose-6-phosphate must be synthesized for DNA, RNA and protein synthesis by invoking the gluconeogenic pathway, as intermediates from the Krebs cycle are shunted to PEP synthesis. Therefore under conditions of growth in high glucose, three major pathways are targeted for repression (Figure 1.1): i) genes involved in alternate carbon source utilization (*e.g.* the *GAL1* gene coding for galactokinase [Johnston, 1987]), ii) genes involved in gluconeogenesis (*e.g. FBP1* which converts fructose-6-P to glucose-6-P [de la Guerra *et al.*, 1988]), and iii) genes required for the Krebs cycle and the mitochondrial electron transport chain (*e.g. SDH Ip* - the iron-protein subunit of succinate dehydrogenase [Lombardo *et al.*, 1990]).

1.4 Mechanisms for eukaryotic glucose repression and their relation to galactose metabolism in *Saccharomyces*

In total there are dozens of genes which are subject to glucose repression in *Saccharomyces*, and it is known that more than one mechanism is involved (Gancedo and Gancedo, 1986). Generally, there are three recognized ways in which repression occurs in eukaryotes. The simplest mechanism is direct competition between *trans*-acting repressors and antagonistic activators (or basal transcription factors) for overlapping binding sites in a promoter. Although this is a common mechanism in prokaryotes, it is not widely used in eukaryotes (Johnson, 1995). Eukaryotes tend to use more "active" mechanisms for transcriptional repression. These systems rely upon the activity of transcription factors which have separable DNA-binding and peptide activity domains. Such transcription factors exert

their effects either through quenching (also known as titration) whereby the repressor binds and inhibits the activator, or, by interaction with components of the transcriptional complex (reviewed in Cowell, 1994; Johnson, 1995; Hanna-Rose and Hansen, 1996). For the latter class of repressors which function by affecting the transcriptional complex, they function by altering a variety of parameters: the stability of the initiation complex, the affinity of other regulatory elements for the transcriptional complex, the ability of the other regulatory elements to modulate transcription, and the degree of access to promoter elements through changes in chromatin structure (Cowell, 1994). As a result, these active forms of repression are more effective than passive forms since their effects can be localized without affecting flanking genes, they are very sensitive, and similar regulatory pathways can be recycled by transplanting them into and subsequently adapting them for other gene systems (Johnson, 1995).

These different mechanisms are not mutually exclusive, as a single gene system may be coordinately repressed using more than one pathway. Such is the case for the *GAL* genes of *S. cerevisiae*, which are a family of enzymes required for galactose metabolism. There are three main structural genes in this system - *GAL1*, *GAL7*, and *GAL10*, which code for galactokinase, galactose-1-phosphate uridyltransferase, and UDP galactose-4 epimerase, respectively. These three enzymes are involved in the conversion of galactose to glucose-1-phosphate, which is readily metabolized (Fukasawa and Nogi, 1989). At least 8 genes involved in galactose metabolism are known to be subject to glucose repression. Of these, the previously mentioned triad of structural genes are each repressed approximately 1000 fold and have been the subject of much study (reviewed in Fukusawa and Nogi, 1989; Johnston and Carlson, 1992; Lohr *et al.*, 1995). This dramatic decrease in gene expression is the compounded result of 3 mechanisms involving both active repression as well as inhibition of galactose induction (Johnston *et al.*, 1994).

The first mechanism for downregulating the *GAL* gene family involves limiting the amount of functional inducer (Figure 1.2). This is accomplished by direct repression of *GAL2*

and *GAL3*, which are the two genes required for inducer synthesis as they mediate galactose transport into the cell (GAL2), and inducer formation (GAL3). A 10-13 fold decrease in *GAL1* expression is attributed to this first mechanism (Johnston *et al.*, 1994). The resulting decrease in inducer levels forms a signal cascade which overlaps with the second mechanism of *GAL* repression. The promoters of *GAL1*, *GAL7* and *GAL10* all contain binding sites for the transcriptional activator GAL4 (Giniger *et al.*, 1985). GAL4 is possibly one of the best known activators, and its association with its negatively regulating GAL80 peptide is the classic example of protein-protein interactions. When there is no inducer present, GAL4 is unphosphorylated and complexed with GAL80 (reviewed in Johnston, 1987). In this uninduced state, GAL4 binds to the upstream activating sequences of promoters, but GAL80 prevents GAL4-mediated activation of *GAL1* and other genes. In contrast, in the galactose-induced state GAL4 is phosphorylated, and although it is still associated with GAL80, it is no longer subject to GAL80-mediated inactivation (Parthun and Jaehning, 1992). Therefore, the second way of repressing the transcription of the *GAL* genes in *Saccharomyces cerevisiae* is by downregulating GAL4 activity (Figure 1.2). Both direct repression of *GAL4* transcription as well as post-translational inactivation of GAL4 resulting from decreased inducer levels (see mechanism 1 above) account for the 30-50x repression of *GAL1* which is attributed to loss of GAL4 activation (Griggs and Johnston, 1991; Johnston *et al.*, 1994). The third and last mechanism of GAL repression is by direct repression. Out of *GAL1*, *GAL7* and *GAL10*, only *GAL1* is subject to direct repression, which accounts for a 4-5 fold repression of this gene (Flick and Johnston, 1992; Johnston *et al.*, 1994). Although each of these mechanisms account for only a 4-50 fold repression in GAL1 gene expression, the effects are compounded, resulting in the greater than 1000 fold repression seen in wild type cells when all of the mechanisms are working concurrently (Johnston and Carlson, 1992).

Common to two of the above three pathways is MIG1, a zinc finger transcription factor which has been determined to play a key role in glucose repression. It recognizes short GC rich binding sites which have been found upstream of many glucose repressible genes,

including: *GAL1* (Nehlin *et al.*, 1991) , *GAL4* (Nehlin *et al.*, 1991), *MAL63* (Wang *et al.*, 1997), and *FBP1* (Mercado et al., 1991), among others. Therefore, it has been implicated in the direct repression of *GAL1* and *GAL4*, the latter of which in turn, downregulates the GAL4-mediated activation of the *Saccharomyces GAL* genes (Figure 1.2).

The active form of MIG1 binds in a glucose-dependent manner to its target promoters (Wang *et al.*, 1997), where it acts to recruit a general yeast repressor complex composed of TUP1 and SSN6 (Keleher *et al.*, 1992; Johnston and Carlson, 1992; Treitel and Carlson, 1995). Mutants for *ssn6* and *tup1* show a broad range of phenotypes which correspond to a loss of regulation for mating type switching (Mukai *et al.*, 1991), oxygen stress response (Zitomer and Lowry, 1992), DNA damage response (Elledge *et al.*, 1993), and glucose repression (Williams and Trumbly, 1990). One SSN6 subunit and 3 TUP1 subunits are required to form this nonspecific repression complex (Redd *et al.*, 1997), and target specificity is achieved through the use of a small class of transcription factors to direct this repression complex to particular systems. In the case of glucose repression, it is MIG1 which makes this general repression mechanism specific for glucose repressible genes (Keleher *et al.*, 1992; Treitel and Carlson, 1995). Neither SSN6 nor TUP1 have DNA-binding abilities, but it has been determined that SSN6 acts to target the complex to different DNA binding proteins (such as MIG1), while TUP1 provides the repressor activity (Tzamarias and Struhl, 1994 and 1995). SSN6/TUP1-mediated repression occurs through the reorganization of the chromatin by interaction between TUP1 and histones H3 and H4 (Cooper *et al.*, 1994; Edmondson *et al.*, 1996).

1.5 Mechanism of derepression in *Saccharomyces*

The *GAL* system described up to now has only examined the mechanism by which genes are repressed in the presence of glucose. What happens in the presence of other carbon sources? In *Saccharomyces cerevisiae*, when the extracellular glucose supply becomes

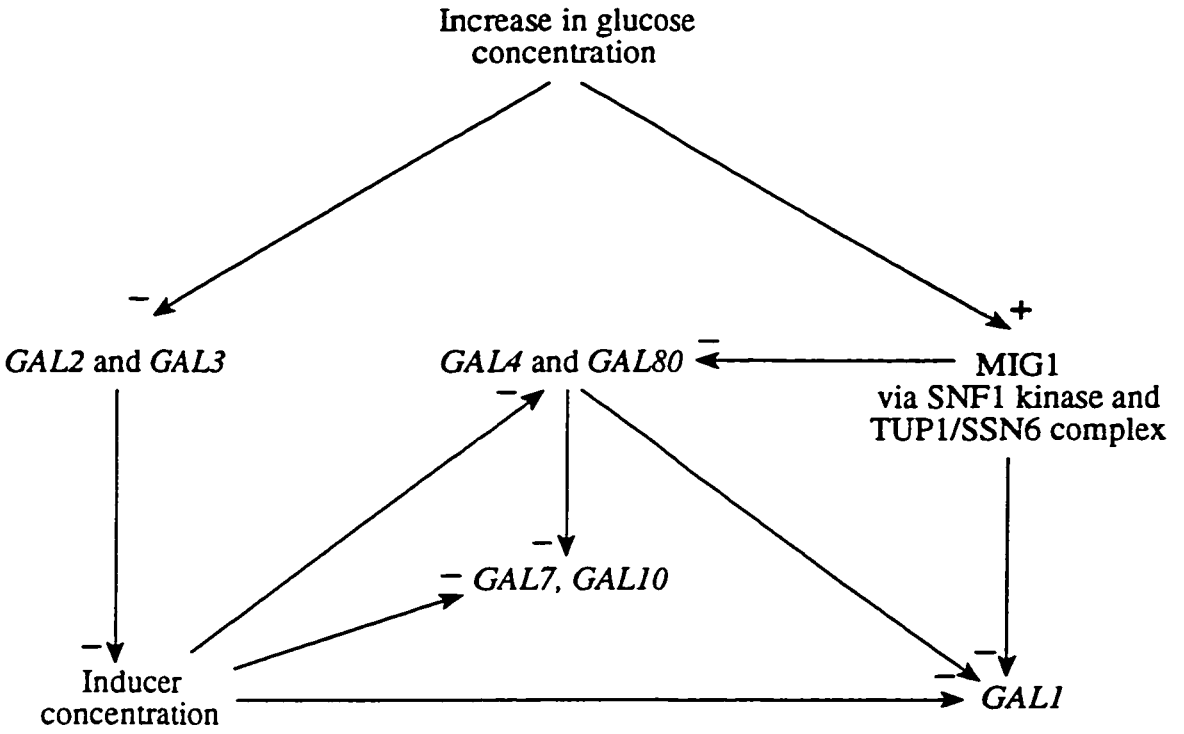


Figure 1.2. The three concurrent mechanisms of glucose repression in the yeast galactose system. One system works by decreasing functional inducer levels through the downregulation of *GAL2* and *GAL3*. The second mechanism involves inactivating the *GAL4* activator, through both direct repression and indirect repression resulting from lower inducer concentrations. The third system involves MIG1-mediated direct repression of several downstream genes. Only two of these systems influence *GAL7* and *GAL10* activity, but all 3 pathways are involved in *GAL1* regulation.

limiting, the shift from fermentation to oxidative respiration requires the upregulation of genes from the three main pathways which are subject to glucose repression (Figure 1.1). Thus, genes which code for electron transport chain components, gluconeogenic enzymes, and alternate carbon source metabolism must be derepressed.

The pathway involved in yeast glucose derepression is less well characterized than that for glucose repression. Approximately 6 loci have been identified as being essential for derepression, but for only 3 of these have the gene products been identified and characterized. Many of the proteins which regulate derepression were initially examined in yeast mutants which could not utilize sucrose (Carlson *et al.*, 1981; Carlson *et al.*, 1984; Gancedo and Gancedo, 1986). Cloning and characterization of one of these mutant genes (Celenza and Carlson, 1986) showed that the gene encodes a protein kinase, which was subsequently named SNF1 (for sucrose nonfermenting, and is also known as CAT1 [Zimmermann *et al.* 1977]). Celenza and Carlson (1989) showed that the catalytic activity of this serine/threonine kinase is essential for the expression of glucose repressed genes when glucose is limiting. This finding was followed by Östling *et al.* (1996), who showed that the kinase negatively regulates MIG1. SNF1-dependent phosphorylation of MIG1 results in the protein conformation changes which prevent MIG1 from associating with the TUP1/SSN6 repression complex.

Classical genetic studies showed a strong relationship between *SNF1* and *SNF4*, another locus identified during the screening for mutants of sucrose metabolism (Carlson *et al.*, 1981). *Snf4* mutants show phenotypes which are very similar to those seen for *snf1*, including the inability of *snf4* mutants to utilize carbon sources that require the expression of glucose-repressible genes (Carlson and Botstein, 1982; Celenza and Carlson, 1984). Celenza and Carlson (1989) performed many biochemical tests on the relationship between SNF1 and SNF4 and showed that the *SNF4* gene product is required for maximal activity of SNF1.

The SNF4 gene was eventually cloned by Schüller and Entian (1988). Although the nucleotide sequence did not shed any light on the possible function of SNF4, subsequent biochemical assays indicated that SNF4 acts as a positive effector of the SNF1 protein kinase

(Celenza and Carlson, 1989). It has been known for several years that SNF1 and SNF4 coprecipitate *in vivo* (Celenza *et al.*, 1989), but the exact mechanism by which these two proteins interact has only been recently deduced. Jiang and Carlson (1996) showed that SNF4 activates SNF1 through competitive binding of SNF4 to the regulatory domain of SNF1. Deletion analysis has shown that the carboxy terminus of SNF1 functions as its regulatory domain, while the kinase activity maps to the amino terminus. Under an environment of high glucose, SNF1 is self-inactivated through the binding of its own regulatory domain to its catalytic domain. SNF1 is then activated under glucose derepressing conditions by an as of yet, unidentified upstream protein kinase. The phosphorylation of SNF1 permits the binding of SNF4 to its regulatory domain, thereby freeing the catalytic domain to interact with its substrates (Jiang and Carlson, 1996). At least one other protein is required for maximal activity of the SNF1 complex. Using the yeast two-hybrid system, Jiang and Carlson (1997) showed that under high levels of glucose SNF4 and SNF1 cannot form a complex without the presence of a third, linker protein. Several different proteins have been shown to act as this bridge protein, including SIP1, SIP2, and GAL83 (Jiang and Carlson, 1997). These three proteins have separate conserved domains for SNF1 and SNF4 binding, indicating that they form a gene family. It is thought that more members of this family exist, and each of these proteins acts as an adapter between the SNF1 derepression complex and a different downstream target (Erickson and Johnston, 1993; Yang *et al.*, 1994).

1.6 Conservation of eukaryotic genes and regulatory mechanisms

It is believed that the yeast MIG1/TUP1/SSN6 and SNF1/SNF4/GAL83 systems are appropriate models for *Drosophila* α-amylase glucose repression and derepression, due to recent transgenic experiments. Initial evidence was provided by Fischer *et al.* (1988) who introduced a copy of the yeast *GAL4* gene into *Drosophila.* Cotransformed along with it was a reporter construct consisting of the *Drosophila hsp*70 promoter region fused to *lacz.* This

reporter gene could only be transcribed if the foreign GAL4 bound to short, 17-mer *GAL4* binding sites that were artificially inserted into the *hsp70* promoter. The yeast GAL4 protein was expressed in *Drosophila*, and it was capable of activating the reporter construct. This early experiment showed that there is evolutionary conservation of the basic transcriptional machinery between *Saccharomyces* and *Drosophila*, to the extent that specialized transcription factors from one species can interact with the transcription complex of another.

More direct evidence comes from Hickey *et al.* (1994). They constructed two yeast shuttle vectors, one of which contained 502 bp of the α-amylase promoter fused to the firefly luciferase gene, and the other 427 bp from the yeast actin promoter linked to the same reporter gene. When these shuttle vectors were transformed into wild-type strains of *Saccharomyces cerevisiae*, they not only expressed the reporter gene, but the expression pattern of the construct which contained the α-amylase promoter was subject to glucose repression, whereas the actin-based construct was not. This indicated that the mechanism for regulating glucose repression was functionally conserved between these two, very divergent eukaryotic taxa. Considering that the homologues of several components identified in yeast glucose regulation - including MIG1, SNF1 and SNF4 have been identified in higher animals (Carling *et al.*, 1994; Mitchelhill *et al.*, 1994; Gao *et al.*, 1996; Woods *et al.*, 1996), it is highly likely that MIG1 and SNF1/SNF4-based pathways are also used in higher taxa. Since the yeast repression and transcription complexes can recognize *Drosophila* glucose repression signals (Hickey *et al.*, 1994), evolutionary conservation must extend from the *cis*-acting promoter sequences to the transcription factors which act in inhibiting mRNA synthesis to the protein complexes which associate with the transcription factors.

What about the evolutionary conservation of the original gene under question - *Drosophila* alpha-amylase? As described earlier (Section 1.2), aside from five functionally conserved domains, the amino acid sequence is highly variable. Due to the differences in α-amylase gene copy number, individual taxa may be subject different evolutionary processes. For instance, the gene conversion seen in *D. melanogaster* is thought to be responsible for a

~90 % G+C content at the third codon position in this gene, due to the observation that DNA mismatch repair mechanisms often favor the incorporation of G-C base pairings (Hickey *et al.*, 1991). Such mechanisms combined with relatively low functional selection, contribute to the wide variety in α-amylase protein sequences characterized to date. As α-amylase sequences are available from such a large diversity of taxa, studies can be conducted which look at the evolution of a single homologous gene which is subject to different evolutionary forces acting at the nucleotide and amino acid levels.

The following chapters develop this theme of the evolution of proteins and their associated regulatory mechanisms. This is accomplished through analyses of the *D. melanogaster* α-amylase coding sequence, its *cis*-acting DNA repression sequences, and protein factors involved in regulating its gene expression by glucose.

CHAPTER 2

CONSERVATION OF THE REPRESSION MACHINERY BETWEEN *DROSOPHILA* AND *SACCHAROMYCES*

2.1 Introduction

Gene expression studies have traditionally been performed by observing the results of reintroduction of a cloned gene back into the original species. More recently however, transgenic experiments are often used to assay gene expression in foreign hosts. This chapter presents the results of a transgenic *Saccharomyces* system which was used to express a reporter gene regulated by the *Drosophila* α-amylase promoter. When >100 bp fragments immediately upstream of the *Drosophila* TATA box were used, glucose repressible expression was detected in *Saccharomyces*. These results support the hypothesis that the mechanism responsible for glucose repression in yeast is conserved in the higher eukaryotes.

2.1.1 *Previous* in vivo *analyses of the α-amylase promoter*

Several studies have examined the regulatory elements in the α-amylase promoter using somatic and germ line *Drosophila* transformation. This is due to the availability of an AMY^{null} strain of *Drosophila melanogaster* (Hickey *et al.*, 1988), which has made it extremely easy to distinguish between the endogenous and exogenous gene products. Using somatic transformation of this null strain, Hawley *et al.* (1992) showed that 446 bp of sequence immediately upstream of the proximal α-amylase gene were sufficient for expression and glucose regulation. This value was later revised by Magoulas *et al.* (1992), who demonstrated by deletion analysis of the promoter that merely 142 bp of the upstream sequence was required for glucose repressible expression. This corresponds to the region extending from -109 to the ATG (which is +33 with respect to the transcriptional start site). In a follow up study

(Magoulas *et al.*, 1993a), characterization of the *cis*-acting sequences in this region was conducted by site-directed mutagenesis. The 142 bp glucose repressible promoter region was divided into 7 sections, and each of these sections was mutagenized in turn. Clones which contained a disrupted TATA or CAAAT box showed no expression, however all of the other constructs were functional and continued to be regulated by glucose. Therefore it was proposed that there is more than one upstream repression sequence (URS) which regulates glucose repression for the *Drosophila melanogaster* α-amylase gene. When one element is mutated or deleted, the other element(s) can compensate for the loss of the single element to maintain glucose repression. When the sequence of this 142 bp promoter region is examined more closely, two GC rich boxes are present, which conform to the yeast MIG1 binding sequence (Lundin *et al.*, 1994).

2.1.2 *Characteristics of* MIG1, *a yeast transcription factor involved in glucose repression*

MIG1 belongs to a family of Cys_2His_2 zinc finger DNA binding proteins. It contains a short region which shares protein sequence similarity with repressors from animals, including the mammalian early growth response factor (EGR1), human Wilms' tumor factor (WT1) and CREA from *Aspergillus nidulans* (Nehlin and Ronne, 1990). CREA, like MIG1, is a global regulatory gene involved in mediating repression due to carbon source (Dowzer and Kelly, 1991). The highly conserved region among these proteins corresponds to the ca. 60 amino acids which span the two zinc fingers of MIG1. Although the primary sequence of the finger region is conserved, the number of fingers is not. Wilms' tumor protein has four such motifs, EGR1 has three, and CREA and MIG1 both contain two (Call *et al.*, 1990; Nehlin and Ronne, 1990, Dowzer and Kelly, 1991).

Nardelli *et al.* (1991) proposed a model of how zinc finger transcription factors bind to DNA. It is known that a single zinc finger recognizes a binding sequence of three base pairs. The specificity of binding is determined by residues 15, 18 and 21 of the zinc finger (Pavletich and Pabo, 1991). It is suggested (Nardelli *et al.*, 1991 and 1992) that the Cys_2His_2 zinc

fingers exemplified by MIG1, CREA, EGR1, and WT1 have zinc fingers which recognize either GGG or GCG. A theory of "wobbly" interactions between the zinc finger and the DNA element has been proposed to explain the permitted variation in binding sites (Lundin *et al.*, 1994). For the proteins which only contain two zinc fingers (MIG1 and CREA), only one of the zinc fingers would need to form bonds with all three nucleotides of its recognition sequence. A stable protein-DNA complex would be formed if the other zinc finger only formed bonds with two of the three nucleotides in its binding site. Therefore, there is some tolerance for different binding motifs, depending on the degree of wobble.

MIG1 binding has been mapped to two regions in the *SUC2* promoter (Nehlin and Ronne, 1990), 1 region in *GAL4* (Nehlin *et al.*, 1991), 2 regions in *GAL1* (Nehlin *et al.*, 1991; Flick and Johnston, 1992), 3 regions in *FBP1* (Mercado *et al.*, 1991), and 2 in *MAL62* (Wang *et al.*, 1997). Although binding is strongly influenced by the surrounding sequence (Lundin *et al.*, 1994), there appears to be a 6 bp consensus binding sequence of 5'-(G/C)(C/T)GG(A/G)G-3', or more simply, SYGGRG (where S = G or C). Recognition sites have also been determined for CREA, WT1 and EGR1, which also adhere to this consensus sequence (Rauscher *et al.*, 1990; Kulmburg *et al.*, 1993; Cubero and Scazzocchio, 1994). Lundin and coworkers (1994) observed that MIG1 binding also requires a short AT rich region immediately 5' to the consensus sequence. Without such a region, even a DNA motif which is identical to the consensus sequence fails to bind the transcription factor. This AT rich region appears to be the site where the DNA is bent to achieve a higher affinity bond with the transcription factor. Therefore, sequence variability in both the recognition site and 5' AT box results in different affinities of MIG1 for different *cis*-acting elements.

2.1.3 *Using transgenic yeast to map the* Drosophila *α-amylase URSs*

In some of the yeast genes for which MIG1 binding sites have been mapped, functional URSs have been shown to be as short as 25 bp (Flick and Johnston, 1992). To achieve a similar degree of resolution in mapping the *Drosophila* α-amylase repression motifs, extremely

short regions of the amylase promoter must be assayed for their ability to confer glucose repression on a non-repressible promoter. One requirement for such a promoter is that it would need to show similar tissue-specific and developmental expression patterns as α-amylase. As transcriptional regulation requires the presence of specific transcription factors, one would need to ensure their presence by coordinating the expression of the exogenous gene with times and areas of natural glucose repression in the fly. Since an endogenous *Drosophila* promoter has not been found which is both non-repressible and has the same expression pattern as α-amylase, another assay system is required. As stated earlier (see Chapter 1), *Drosophila* genes can be expressed in yeast in a glucose repressible manner (Hickey *et al.*, 1994). By using a unicellular assay system such as *Saccharomyces*, the complications associated with tissue specific and developmental regulation can be circumvented. Also, the process of obtaining transgenic yeast, and the time between transformation and assay of gene activity are greatly reduced in comparison to somatic transformation of *Drosophila*. Therefore, the use of transgenic yeast can result in a gene expression system that behaves functionally like the endogenous *Drosophila* system, but removes many of the complications and time consumption associated with a multicellular host.

The transgenic yeast approach used here to analyze the *cis*-acting repression sequences in *Drosophila* is based upon work done by Flick and Johnston (1992) to find the MIG1 regulatory sequences in *GAL1*. They engineered an expression vector which featured a non-repressible *LEU2/HIS3* hybrid promoter fused to a *HIS3* reporter gene. An *Eco*RI site was inserted at the junction between the *LEU2* and *HIS3* sections of this base construct, in order to insert short *GAL1*-derived oligonucleotide linkers to determine their ability to confer glucose repressibility. Using similar methodology, various regions of the *Drosophila* α-amylase promoter were tested. The glucose repression URS could not be resolved as finely as they has been achieved in yeast, but a series of α-amylase-based constructs, the smallest containing 126 bp, were glucose repressible in this transgenic *Saccharomyces* system.

2.2 Materials and methods

2.2.1 *Plasmid construction*

The reporter gene was a gift of Dr. Bernhard Benkel (see description in Hickey *et al.*, 1994). It was constructed by ligating a *Nco*I - *Hin*dIII fragment containing the entire firefly luciferase coding region and a *Hin*dIII - *Bam*HI fragment containing 550 bp of the *Drosophila* α-amylase terminator region into the yeast shuttle vector pBTI-1 (Boehringer Mannheim). The promoters to be assayed were inserted into this reporter construct as *Not*I - *Nco*I fragments. These two restriction sites were incorporated into the PCR generated promoter fragments as part of the primer sequence (Figure 2.1).

In addition to cloning the expression cassette into the episomal pBTI-1 plasmid, integrating forms of two constructs were also made. For these integrating vectors, a *Not*I - *Bam*HI fragment containing the test promoter, luciferase coding sequence, and α-amylase terminator was cloned into the plasmid pRS405 (Stratagene). This plasmid lacked a yeast origin of replication, and would only propagate if integrated into the yeast genome through homologous recombination. It was targeted to integrate within the *HIS3* gene through the addition of 957 bp of *HIS3* coding sequence (+213 to +1169), which was PCR amplified using the following set of primers, containing *Sac*I sites for cloning (underlined):

K1162: 5' AGCTTG<u>GAGCTC</u>GCTAGGAGTCACTGCCAGGTATC 3' (forward)

K1163: 5' CGGTGT<u>GAGCTC</u>ATAAGAACACCTTTGGTGGAGGG 3' (reverse)

All PCR reactions were performed in 25 µl reaction volumes, using 1 U Taq DNA polymerase (Promega), 200 µM of each dNTP (Gibco BRL), 2 mM $MgCl_2$, 10 - 50 ng of template DNA, and 15 pmol of each primer. The template DNAs used were *Saccharomyces* genomic DNA (strain AH22cir$^+$) and a plasmid, pBY59, which was made by Dr. B. Benkel. It contained 502 bp of α-amylase sequence corresponding to -469 to the ATG [+33]. PCR profiles varied depending on the primer combinations used, but generally consisted of a 96 °C

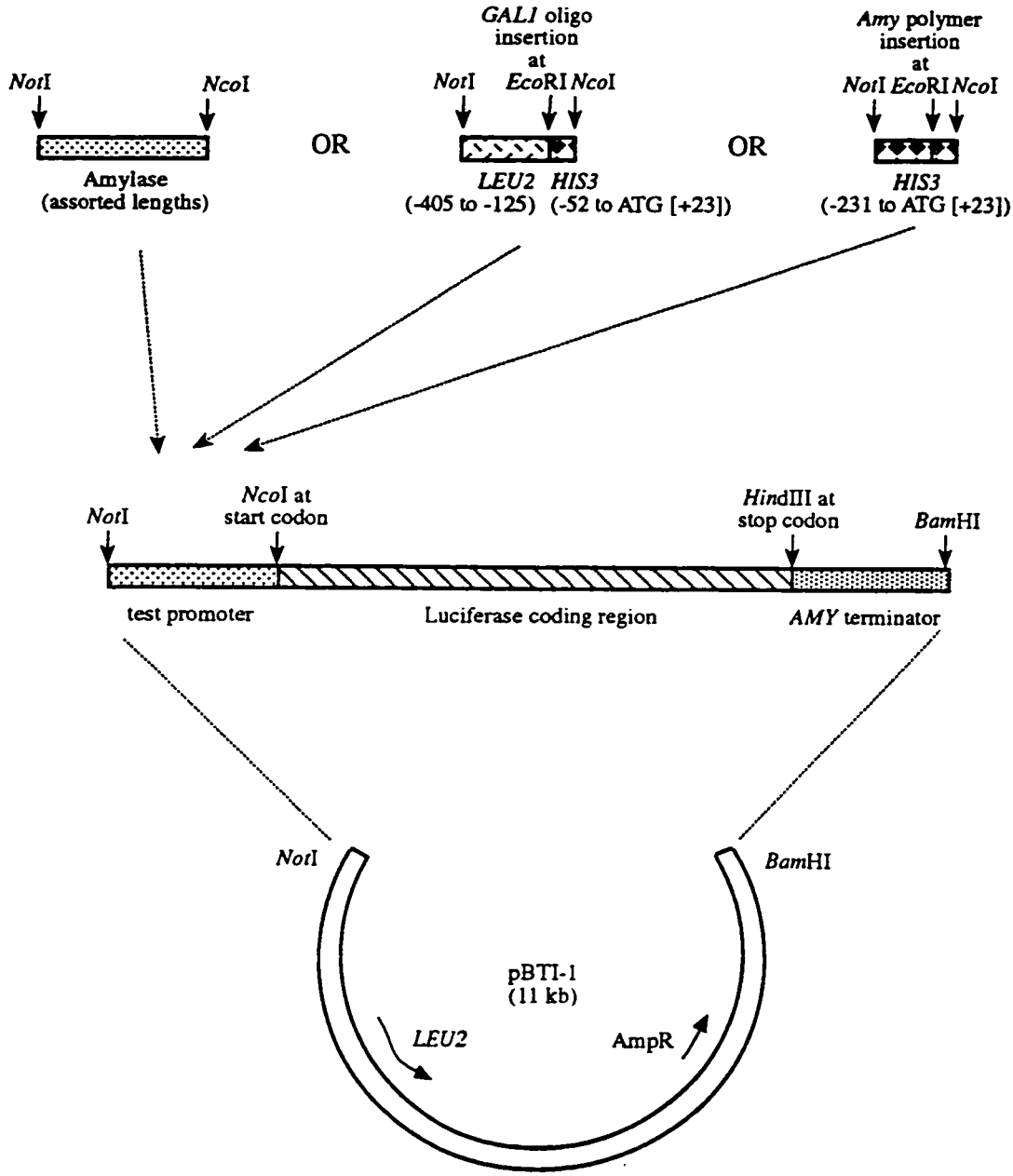


Figure 2.1. Schematic diagram of the yeast expression vectors. Three types of promoters were tested: ones which contained only *Drosophila* α-amylase 5' noncoding sequence, ones which contained putative repression oligomers inserted into a yeast *LEU2/HIS3* fusion promoter, and ones which contained similar oligomers inserted into a yeast *HIS3* promoter. All test promoters were inserted into the luciferase expression casette as *Not*I - *Nco*I fragments. The vector shown is the nonintegrating pBTI, however the same construction strategy was used to insert the *Not*I - *Bam*HI expression casette into the integrating pRS405 plasmid.

denaturation for 30 seconds, a 48 - 60 °C annealing for 20 seconds, and a 0.5 - 1.0 minute extension at 72 °C, for 30 cycles.

The test promoters used fell into three categories (Figure 2.2): ones which contained only *Drosophila* α-amylase promoter sequence; those which recreated the yeast *LEU2/HIS3* Flick and Johnston (1992) promoter; and third, ones which featured α-amylase promoter regions inserted into the yeast *HIS3* promoter. For the first type, there were three constructs assayed whose promoters contained only DNA from the α-amylase promoter (Figure 2.2). The longest, pBY59, is described above. Two shorter versions were made which contained 307 bp (clone EY5) and 201 bp (EY6) of α-amylase upstream sequence. These junction points were chosen because the DNA sequence minimized DNA mismatches between the introduced restriction sites incorporated into the primers and the native sequence, which is shown in Appendix A. The primers used for their synthesis were:

K770: 5' CAGAACCATGGTGATTCCAGATGGAAGTTCAG 3' (reverse at ATG [+33])

K771: 5' CAACTGCGGCCGCTTTCAAAGGAATGCATTTTCCC 3' (forward at -274)

K772: 5' TCAATGCGGCCGCTCGGAATTGTGATTTGACA 3' (forward at -168)

The locations for primer binding are given relative to the α-amylase transcriptional start site. The underlined nucleotides are the *Nco*I (K770) and *Not*I (K771 and K772) sites introduced into the primers as required for cloning into the expression cassette.

For the Flick and Johnston (1992) based promoter, the *Saccharomyces LEU2* portion of the promoter extended from -405 to -125, and the *HIS3* portion from -52 to the ATG [+23], with an *Eco*RI site at the junction (Figure 2.2). These fragments were selected to exactly reproduce the promoter described by Flick and Johnston. The primers used to synthesize this base promoter were:

K817: 5' GAAATATCGCGGCCGCAGTTAACTGTGGGAATACTC 3' (*LEU2* forward)

K821: 5' CTTAAAGAATTCTGTCAACTTCAAGTATTGTG 3' (*LEU2* reverse)

K814: 5' CATAATGAATTCTACATTATATAAAGTAATGT 3' (*HIS3* forward)

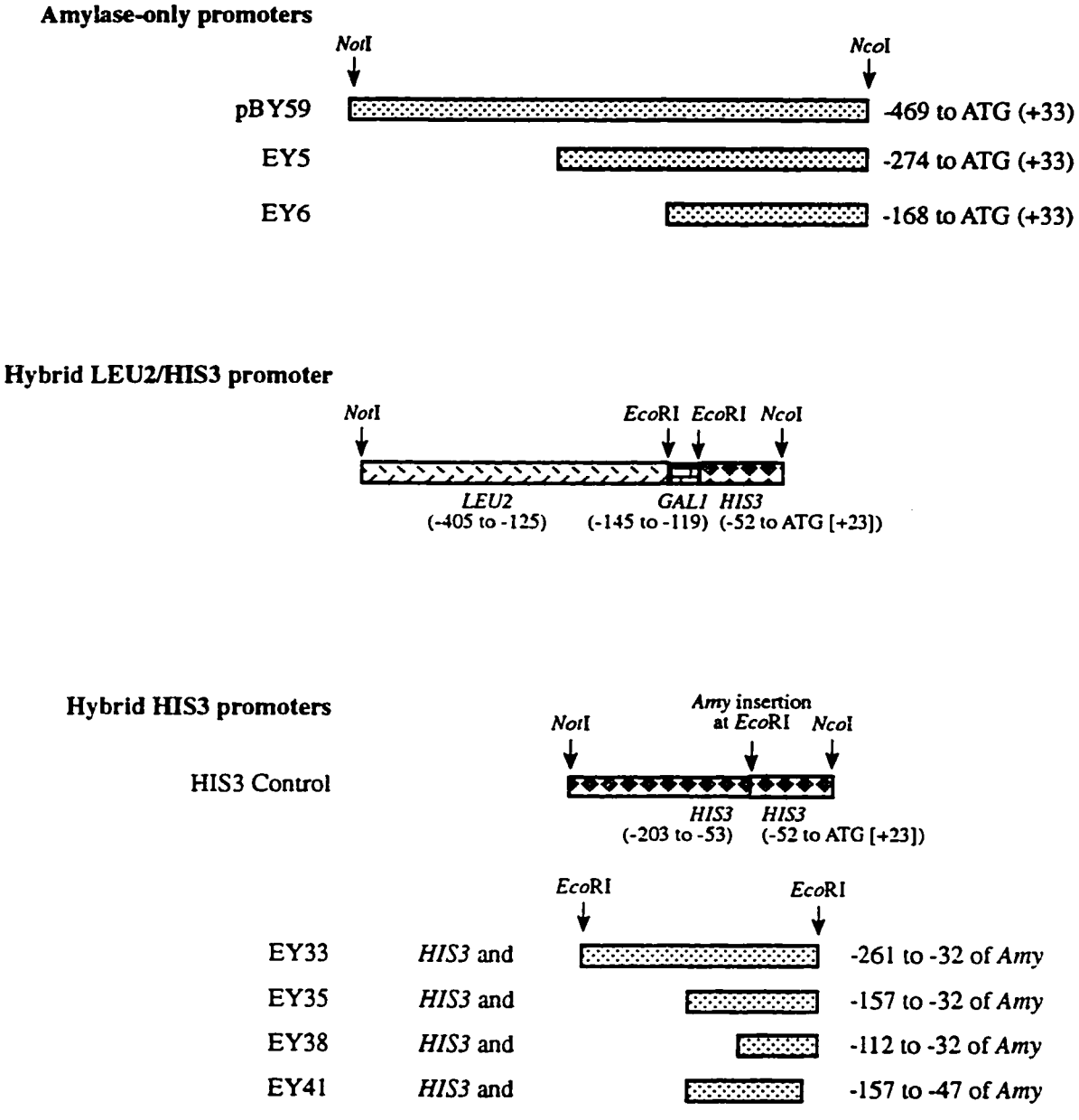


Figure 2.2. Schematic of the promoters used. There were three classes of promoters tested: amylase-only, hybrid *LEU2/HIS3*, and hybrid *HIS3/Amy*. All promoters were inserted into the luciferase expression casette as *Not*I - *Nco*I fragments (see Figure 2.1). Reference points are given relative to the transcriptional start site. Clone names are shown to the left of each construct.

K815: 5' GCTCTGCCATGGTTGCCTTCGTTTATCTTGCC 3' (*HIS3* reverse)

Primer K817 contains an introduced *Not*I site, K821 and K814 both contain *Eco*RI sites, and K815 contains a *Nco*I site, which are shown underlined in the above sequences. Inserted into this base *LEU2/HIS3* vector was the *GAL1* positive control oligonucleotide which was shown to be a MIG1 binding site which regulates glucose repression (Flick and Johnston, 1992). It was synthesized using the following two overlapping primers that were annealed and ligated into the *Eco*RI site of the *LEU2/HIS3* expression cassette (Figure 2.2):

K860: 5' AATTCTTAGCCTTATTTCTGGGGTAATTAATCG 3'
3' GAATCGGAATAAAGACCCCATTAATTAGCTTAA 5': K861

The last series of constructs were based on a 226 bp *HIS3* yeast promoter with an introduced *Eco*RI site at -58 to -53 (Figure 2.2 and Appendix A). This promoter was made by PCR amplifying the region upstream of the *Eco*RI site separately from the region downstream of the site, and ligating the two fragments at the *Eco*RI junction. The previously described K814 and K815 primers were used to synthesize the region downstream of the *Eco*RI site, while the following primers were used to amplify the upstream region with introduced *Not*I (K775) and *Eco*RI (K893) sites:

K775: 5' ATAACGCGGCCGCTTCCCGCAATTTTCTTTTTCTAT 3' (forward)

K893: 5' ATAATGTAGAATTCATTATGTGATAATGCC 3' (reverse)

Putative repression elements in the *Drosophila* α-amylase promoter were tested by amplifying various regions of this promoter with primers containing *Eco*RI sites, and inserting the fragments into the *Eco*RI site of the *HIS3* base promoter (Figure 2.2). The following details the primers used to synthesize the four α-amylase fragments (with the *Eco*RI sites underlined), the clone designations, and the location of the amylase fragment with respect to the transcriptional start site (Appendix A):

EY33 - the *HIS3* base and a 230 bp *AMY* fragment (-261 to -32):

K1012: 5' CAAAGGAATTCATTTTCCCGATGAGTT 3' (forward)

K945: 5' CTCCTGAATTCGTGGACCCCACTG 3' (reverse)

EY35 - the *HIS3* base and a 126 bp *AMY* fragment (-157 to -32), K945 (above) was used with K1013: 5' GCTCGGAATTCTGATTTGACAAACTAAT 3' (forward)

EY38 - the *HIS3* base and a 81 bp *AMY* fragment (-112 to -32), K945 (above) was used with K1042: 5' GCGTGAGAATTCCCCTTAGGGAGCGATAA 3' (forward)

EY41 - the *HIS3* base and a 111 bp *AMY* fragment (-157 to -47), K1013 (above) was used with K318: 5' GTGTCTAGAGAATTCACTTTATCTGAGGGCTTCGCGGGGCGTCATTTG

Due to the presence of an *Eco*RI site within the luciferase coding sequence, all of the constructs containing yeast-based nonrepressible promoters (*i.e.* the *LEU2/HIS3* and *HIS3* based series) were first assembled in pBluescriptKS (Stratagene), and the test promoters transferred to the yeast shuttle vector as *Not*I - *Nco*I fragments. DNA for all constructs was prepared using Qiagen midiprep columns, and the sequences were verified via double stranded plasmid sequencing using the ABI Prism Dye Terminator cycle sequencing kit and an ABI 373 sequencer.

2.2.2. *Transformation and luciferase expression assays*

Two *Saccharomyces* strains were used for gene expression assays. AH22cir$^+$ (*mat* a, *leu2*-3, *leu2*-112, *his4*-519, *can1*, 2 μ plasmid) was most commonly used, but CG378 was also tested (*mat* a, *ade5*, *can1*, *leu2*-3, *leu2*-112, *trp1*-289, *ura3*-52, *gal2*) (Yeast Genetic Stock Center). Yeast were transformed using the lithium acetate method with 0.7 μg plasmid DNA (Schiestl and Gietz, 1989). Transformants were screened by plating on DOBA media lacking leucine (Bio101) and grown at 30 °C for 3 days.

The luciferase assay consisted of first growing single transformant colonies overnight at 30 °C in 5 mls of selective DOB media lacking leucine (Bio101) supplemented with 5 % glucose (wt/vol). This primary culture was then used to inoculate 5 mls of rich media (YPD - 1 % Bacto-Yeast extract [Difco], 2 % Bactopeptone [Difco]) containing either 5 % glucose

(wt/vol) or 5 % glycerol (vol/vol). The glucose-grown cultures grow considerably faster than the glycerol-based cultures, so the starting inoculum for the glucose cultures was generally 100 fold less. After overnight growth at 30 °C, the cell densities were determined by OD_{600} readings. Cells were harvested by centrifugation at 3000 rpm for 10 minutes, the pellets resuspended in 1 x cell lysis buffer (Promega), and the resuspended culture frozen at -70 °C overnight, resulting in cell lysis. To compensate for differences in cell densities between the glycerol and glucose-grown cultures, the volume of cell lysis buffer used to resuspend the cell pellets was adjusted according to the OD_{600} measurement. After thawing the samples for 30 minutes, quantification of luciferase activity was performed with a LKB 1251 luminometer using 0.5 - 5 µl of cell lysate and ≈50 µl of luciferase assay substrate (Promega).

2.3 Results

2.3.1 *The intact* Drosophila *α-amylase promoter drives glucose repressible luciferase expression in* Saccharomyces

Three promoters containing various lengths of α-amylase sequence only (502 bp, 307 bp, and 201 bp, see Figure 2.2) were tested using the luciferase expression assay. All three constructs expressed the reporter gene in a glucose repressible manner (Figure 2.3). The longest, pBY59, showed a mean level of repression of 6.5 fold. In comparison, the two shorter constructs (EY5 and EY6) expressed at a slightly lower level in glycerol-grown cultures and a much lower level in glucose-grown cultures. As a result, the average level of repression for these constructs was 8.5x and 27x for EY5 and EY6, respectively. There was a high level of variation in the degree to which individual clones were repressed, as individual clones varied from 3 fold to more than 125 fold. To test whether the differences in expression levels between the different growth regimes were statistically significant, a paired sample t-test (Freund, 1988) was used. Although this is still a mean-based test, it takes into account the fact that the samples are not independent by using paired sets of data - *i.e.* it compares individual

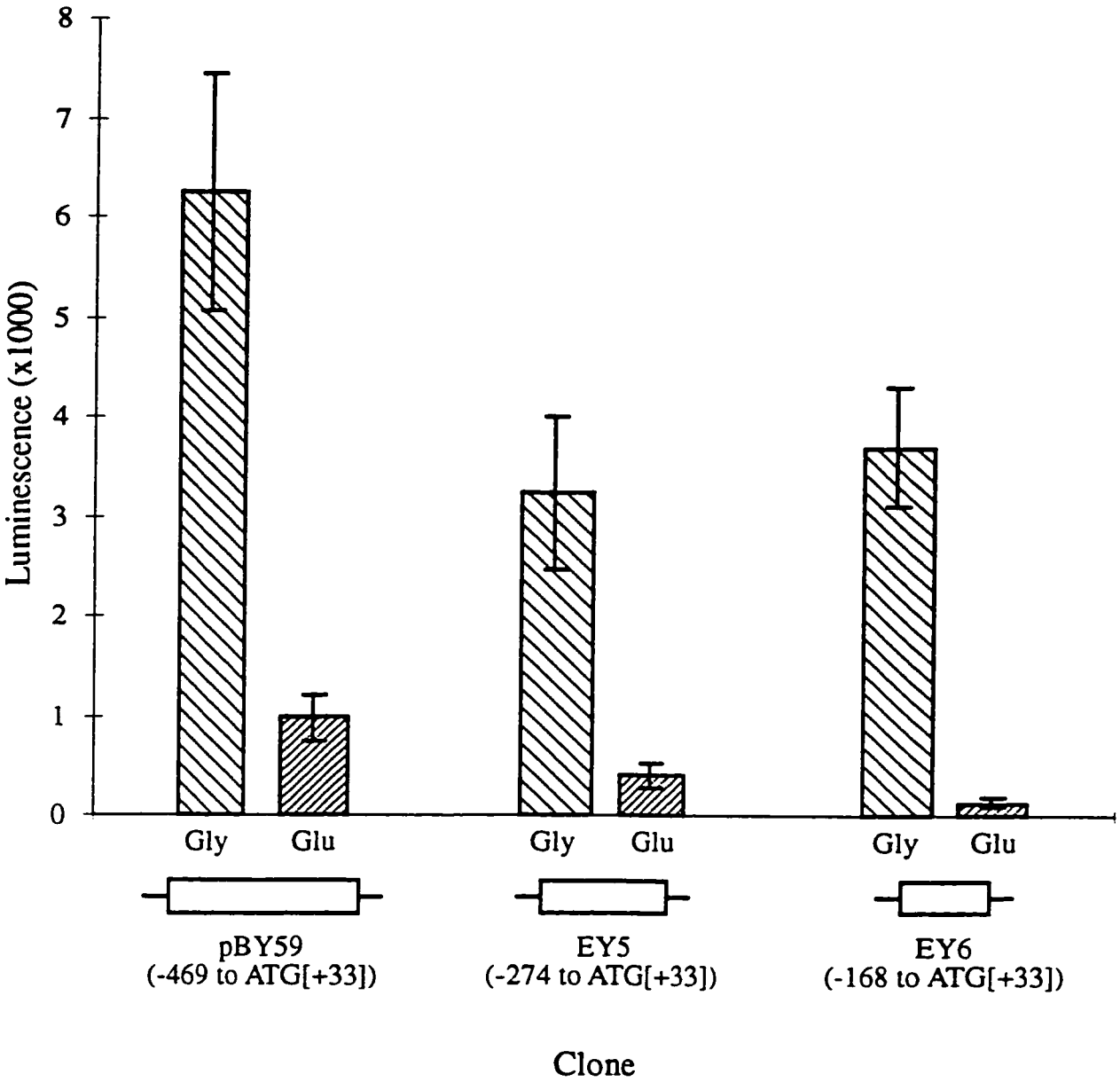


Figure 2.3. Quantification of expression levels for constructs containing only *Drosophila* alpha-amylase promoter sequence. Three different lengths of promoters were assayed, with the specific ranges shown relative to the transcriptional start site. Gly = YPD cultures containing 5% vol/vol glycerol, Glu = YPD cultures containing 5% wt/vol glucose. For all samples, at least 4 replicates from a minimum of 2 independent transformations were used. Data presented are means ± one standard error. All three constructs showed statistically significant differences (p=0.05) in reporter gene expression between the glycerol and glucose treatments.

transformant cultures which were split into the two types of media. When this test was used, all three constructs showed statistically significant differences in reporter gene expression between the glucose and glycerol-grown cultures. These data confirm and expand upon the results of Hickey *et al.* (1994) by showing that constructs containing a minimum region of -168 to the ATG (+33) of the alpha-amylase promoter show glucose repressible expression in yeast.

2.3.2 *Analysis of the* LEU2/HIS3 *hybrid promoter*

Although the previous results showed that a relatively small 201 bp α-amylase promoter was glucose repressible in yeast, from Magoulas *et al.* (1993) it was known that the glucose URSs were located within a 142 bp promoter fragment. However, the same researchers also showed that deletion to a 124 bp promoter nearly abolished expression. Therefore, to further define the *Drosophila* α-amylase promoter region which controls glucose repression, rather than using a deletion-based strategy, the insertion-based strategy of Flick and Johnston (1992) was used.

The *LEU2/HIS3* promoter was recreated to the exact specifications of Flick and Johnston (1992), with the exception of using *Not*I - *Nco*I sites to clone the promoter instead of the *Bam*HI - *Sac*I sites used by them. For a positive control, the 26 bp *GAL1* repression sequence identified in their study was inserted into the base construct. When this *GAL1*-containing hybrid construct was assayed for luciferase expression, glucose repression could not be detected in any of 8 independent assays, as the construct expressed strongly in both types of media. The average level of luciferase expression was 20,000 units for the glucose cultures and 22,300 units for the glycerol ones. This observation was in direct contrast with published results which showed that the *GAL1* oligo caused at least a 10-fold decrease in expression in media containing glucose (Flick and Johnston, 1992).

To address the possibility that differences in host cell genotype are responsible for the unexpected expression patterns seen with the *LEU2/GAL1/HIS3* promoter, the *Saccharomyces*

strain CG378 was transformed with this construct in addition to the AH22cir$^+$ strain (see Section 2.2.2 for genotypes). When the CG378 transformants were assayed for luciferase activity, the glucose grown cultures were quantified at an average of 14,600 units and the glycerol grown cultures, 14,200 units. Thus, although the CG378 strain expressed the heterologous construct at lower levels than the AH22cir$^+$ strain, it too showed nonrepressible expression patterns.

An aliquot of the Flick and Johnston plasmid was subsequently obtained from Dr. Mark Johnston. Upon sequencing the promoter of their plasmid, discrepancies between their sequence and the expected were found at the *Eco*RI junction sites, as shown in Figure 2.4. There is a 10 bp insertion immediately upstream of the *Eco*RI site separating the *GAL1* linker from the *LEU2* region, and another 4 bp one downstream of the site separating the linker from the *HIS3* region. The sequence of these insertions do not correspond to any of the possible *GAL1*, *LEU2* or *HIS3* sequences which might be expected at such sites (Figure 2.4). The presence of a *Bam*HI site (GGATCC) in the upstream insertion would seem to indicate that the insert is part of a linker sequence which was incorporated during the construct assembly. This 10 bp insertion is particularly problematic, since it introduces the sequence 5' CCGGGG 3', which is a MIG1 binding site. Therefore, instead of assaying the activity of the single URS contained in their 26 bp *GAL1* linker, they were potentially assessing the activity of two. Although it is difficult to say with certainty how much of an effect the extra URS copy would have on their data, it is quite conceivable that an additional GC box could bind another molecule of MIG1 and result in cooperative effects in glucose repression. In fact, MIG1 binding is not always necessary for GC boxes to affect glucose repression. In the case of the *GAL1* and *GAL4* promoters, there are GC boxes which are known to not bind MIG1, yet they are involved in glucose repression (Griggs and Johnston, 1991; Flick and Johnston, 1992). Thus, there are known examples of how additional GC boxes adjacent to *bona fide* MIG1 binding sites can have positive effects on glucose repression, perhaps by stabilizing low-affinity MIG1 binding (Lundin *et al.*, 1994). However, as DNA footprinting experiments of

		LEU2	» *Eco*RI «	*GAL1*	» *Eco*RI « *HIS3*	
Expected sequence	5'	ACTTGAAGTTGACA	GAATTC	TTAGCCTTATTTCTGGGGTAATTAAT	GAATTC	TACATTATAT 3'
Flick + Johnston	5'	ACTTGAAGTTCGGATCCGGG	GAATTC	TTAGCCTTATTTCTGGGGTAATTAAT	CAATTC	CTTATACATTATAT 3'
LEU2 genomic	5'	ACTTGAAGTTGACAATATTA	TTTAAG			3'
GAL1 genomic	5'	TAATGCGATT	AGTTTT	TTAGCCTTATTTCTGGGGTAATTAAT	CAGCGA	AGCGAT 3'
HIS3 genomic	5'				TAATGA	ATTATACATTATAT 3'

Figure 2.4. Sequence analysis of the *LEU2-GAL1-HIS3* junction region of Flick and Johnston's (1992) construct. The sequence of their clone is aligned with the expected sequence from their construct specifications. The unexpected nucleotides are single underlined and the motifs which conform to MIG1 binding sites are double underlined. The origin of the two insertions is unknown since they do not match any of the possible constituent sequences as shown in the alignment.

the *GAL1* promoter (Nehlin *et al.*, 1991) show that MIG1 protects the strongly repressing element identified by Flick and Johnston (1992), there is additional evidence to support their determination of the *GAL1* URSs using the *LEU2/HIS3* hybrid promoter.

2.3.3 *Mapping the repression sequences using the α-amylase/*HIS3 *promoter series*

Due to the complications associated with the sequence of the Flick and Johnston *LEU2/HIS3* promoter, it was decided that another base promoter should be used instead. The reason for using both *LEU2* and *HIS3* regions in their promoter was due to historical intermediates in the evolution of their expression cassette. Since there were no obvious reasons for avoiding the use of single gene promoters, and because previous constructs produced in our lab have used the yeast *HIS3* promoter, a *HIS3* promoter was constructed. This promoter differed from the wild type in that it contained an *Eco*RI site introduced at the convenient location used by Flick and Johnston (1992) immediately 5' to the TATA box (see Section 2.2.1). This *Eco*RI site was used to insert four different regions of the α-amylase promoter to produce a series of hybrid constructs (Figure 2.2): EY33 (contains a 230 bp α-amylase fragment from -261 to -32), EY35 (126 bp fragment from -157 to -32), EY38 (81 bp fragment from -112 to -32) and EY41 (111 bp fragment from -157 to -47). Three control plasmids were also used - a *HIS3* promoter with no insert; pBY59, the 502 bp α-amylase promoter construct described earlier (Section 2.2.1); and third, an expression cassette which completely lacked a promoter of any sort. These 7 constructs were tested for their ability to regulate glucose repression, with the results graphed in Figure 2.5.

The one positive and two negative controls all expressed as expected. The *HIS3*-only promoter expressed strongly and was not glucose repressible. Indeed, its expression was slightly higher in glucose-grown cultures compared to the glycerol-grown ones (15,500 units versus 13,600 units), although this was not statistically significantly different. For the construct which contained no promoter, there was virtually no detectable expression in either media type, as luciferase activity was quantified at less than 50 units. The alpha-amylase based

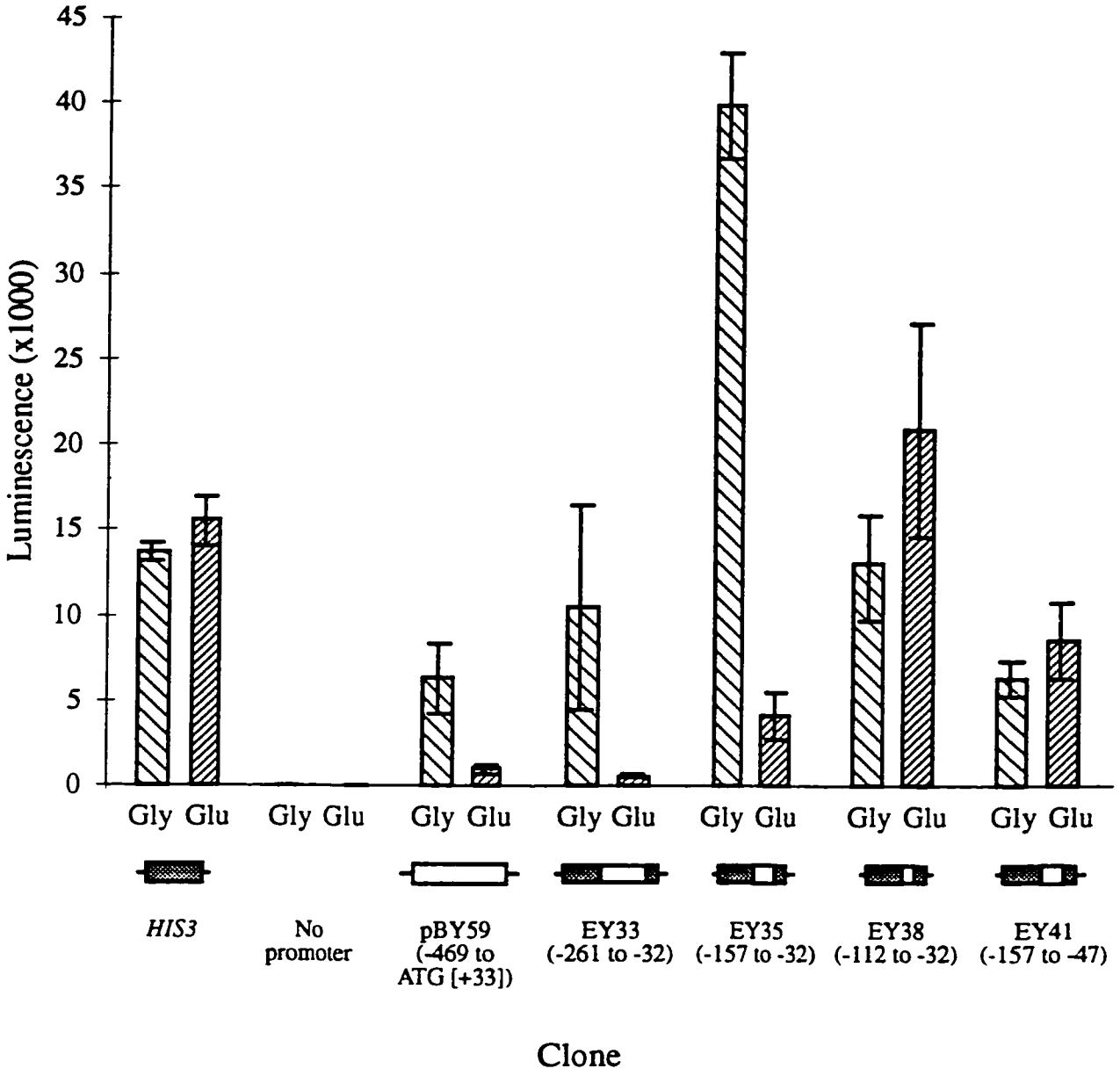


Figure 2.5. Quantification of expression levels for constructs containing hybrid *Drosophila* - yeast promoters. The two negative controls consisted of a yeast *HIS3* promoter with no amylase insert, and an expression construct which contained neither yeast nor *Drosophila* promoter sequence. Several different regions of the alpha-amylase promoter were assayed, with the specific ranges shown relative to the transcriptional start site. Gly = YPD cultures containing 5% vol/vol glycerol, Glu = YPD cultures containing 5% wt/vol glucose. The means of at least 4 clones from 2 independent transformations are shown along with ± one standard error. Only pBY59, EY33 and EY35 showed statistically significant differences (p=0.05) in reporter gene expression between the glycerol and glucose treatments.

positive control, pBY59, had a lower level of expression than the *HIS3* construct, but was subject to glucose repression (Figure 2.5).

Two other constructs in this series showed glucose repressible expression. They were EY33 and EY35, both of which contained hybrid α-amylase/*HIS3* promoters. Relative to the 502 bp α-amylase promoter fragment contained in pBY59, EY33 and EY35 were shorter from both the 5' and 3' ends (Figure 2.2). On the 3' end, both of the amylase/*HIS3* hybrid promoters extended only as far as -32 relative to the transcriptional start site. As this 3' deletion eliminated the sequence from the ATG [+33] up to and including the *Drosophila* TATA box, the only TATA box in the hybrid promoters was from the yeast *HIS3* sequence, located immediately downstream of the α-amylase insertion. This 3' deletion, combined with the shortened 5' ends of the tested α-amylase fragments, resulted in EY33 containing a 230 bp amylase region, and EY35 a 126 bp fragment. When assayed for luciferase expression both of these hybrid promoters showed relatively high levels of expression (Figure 2.5). The expression level for EY33 in glycerol was comparable to the *HIS3* control, however the expression level for EY35 in glycerol medium was highly elevated - approximately 3 times the level seen for the same control plasmid. In glucose-based media, EY33 expressed at a similar level to pBY59, but EY35 showed a 4 fold higher expression level compared to either of these constructs. With regards to the amount of expression considering the different media regimes, the average degree of repression for the longer EY33 construct was 18.2 fold, and 9.9 fold for the shorter EY35. For individual clones, the variation in repression ranged from 48.6x for one EY33 clone to 5.8x for one EY35 clone. When the paired t-test was used, both the EY33 and EY35 constructs, as well as pBY59, showed statistically significant differences between reporter gene expression in glycerol as compared to glucose-grown cultures ($p=0.05$).

These data contrasted with the remaining two hybrid promoters assayed. The first was EY38, which was one of the shortest α-amylase constructs tested. A further 45 bp of 5' amylase promoter sequence was removed, leaving an 81 bp *Drosophila* fragment inserted into the yeast *HIS3* base construct (Figure 2.2). The boundary of this amylase fragment was -112

to -32 with respect to the transcriptional start site. This construct was nearly equivalent to a 3' shortened version of the -109 to ATG [+33] promoter which showed glucose repressible expression in a *Drosophila* somatic transformation assay (Magoulas *et al.*, 1993a). Since the previously assayed EY33 and EY35 also lacked the -31 to ATG [+33] region but were glucose responsive, this 65 bp region does not contain elements which are essential for regulating glucose repression. The EY38 construct in particular showed highly variable expression (Figure 2.5). The average trend was of higher expression in glucose-containing media over glycerol media, although the difference was not statistically significant. Thus, this construct was not glucose repressible in yeast, although a promoter with nearly the same 5' boundary was repressible *in vivo* in *Drosophila*.

The last promoter in this series was EY41, which had the same 5' boundary as the repressible EY35 construct, but featured an extra 15 bp deletion from the 3' end. This plasmid was constructed to test the hypothesis that the α-amylase promoter is regulated by a mechanism homologous to the MIG1-mediated system in yeast. Within the -109 to ATG [+33] region of the α-amylase promoter assayed by Magoulas *et al.* (1993), there are two regions which resemble the MIG1 binding site consensus sequence. One sequence is 5'-GTGGGG-3' (-42 to -37 with respect to the transcriptional start site), and the other is a 5'-CGCCCC-3' sequence found on the lower strand (-70 to -65). As mutagenesis experiments in *Drosophila* (Magoulas *et al.*, 1993a) indicated the presence of more than one glucose repression sequence, by constructing the EY41 clone, an amylase promoter fragment could be tested which contains only one of the two putative URSs. Expression studies of this clone showed that the construct was nonrepressible in glucose (Figure 2.5). Although the sample size for this clone was relatively small (N = 4), none of the independent trials showed any indication of glucose repression. Therefore the *HIS3/Amy* construct series showed that a region extending from -157 to -32 is sufficient for glucose repressible expression in *Saccharomyces*, and 5' or 3' deletions of this fragment maintain expression but abolish glucose repression.

2.3.4. *Comparison of the expression of integrated and nonintegrated plasmids*

Regardless of which plasmid a transformant contained, most cultures grew to an optical density at 600 nm of 0.9 to 1.1 in glucose media and 0.25 to 0.35 in glycerol media. This narrow range in cell density differences cannot explain the large variation in expression levels seen with some clones (*e.g.* EY33 and EY38 of Figure 2.5). The most likely cause of this variation is differences in plasmid copy number between individual transformants. Although the exact copy number was not measured, to test the effect of copy number on expression and glucose repression, two expression cassettes - ones containing the *HIS3* only promoter and the *HIS3*/126 bp *Amy* (EY35) promoter were integrated into the *HIS3* chromosomal region of AH22cir$^+$. To do so, the expression cassette was moved from the autonomous plasmid used up to now to a nonreplicating integration vector which contained a region of homology with the *HIS3* coding sequence to direct the site of integration (see Section 2.2.1). Confirmation of the integration event was performed by PCR amplifying internal sequences specific to the integration vector from genomic DNA preparations of the integrated strains.

Saccharomyces transformants carrying the integrated and nonintegrated forms of these constructs were tested for luciferase expression, along with the nonintegrated pBY59 as a positive control (Figure 2.6). Mean expression of the integrated forms of the plasmids were lower than the nonintegrated forms - especially so with EY35 in glycerol-based media. There is a general trend of higher expression of the integrated *HIS3* construct in glucose media and lower expression of the integrated EY35 in the same media, although statistically, neither difference is significant. Therefore, for clone EY35 in particular, although in its nonintegrated form it expressed in a strongly glucose repressible manner, in its integrated form, the lower expression level in glucose media was not statistically significant. The degree of variation in the data actually increased with the integrated data set, although the sample sizes were smaller than that of the nonintegrated assays (N = 5 vs. N = 8).

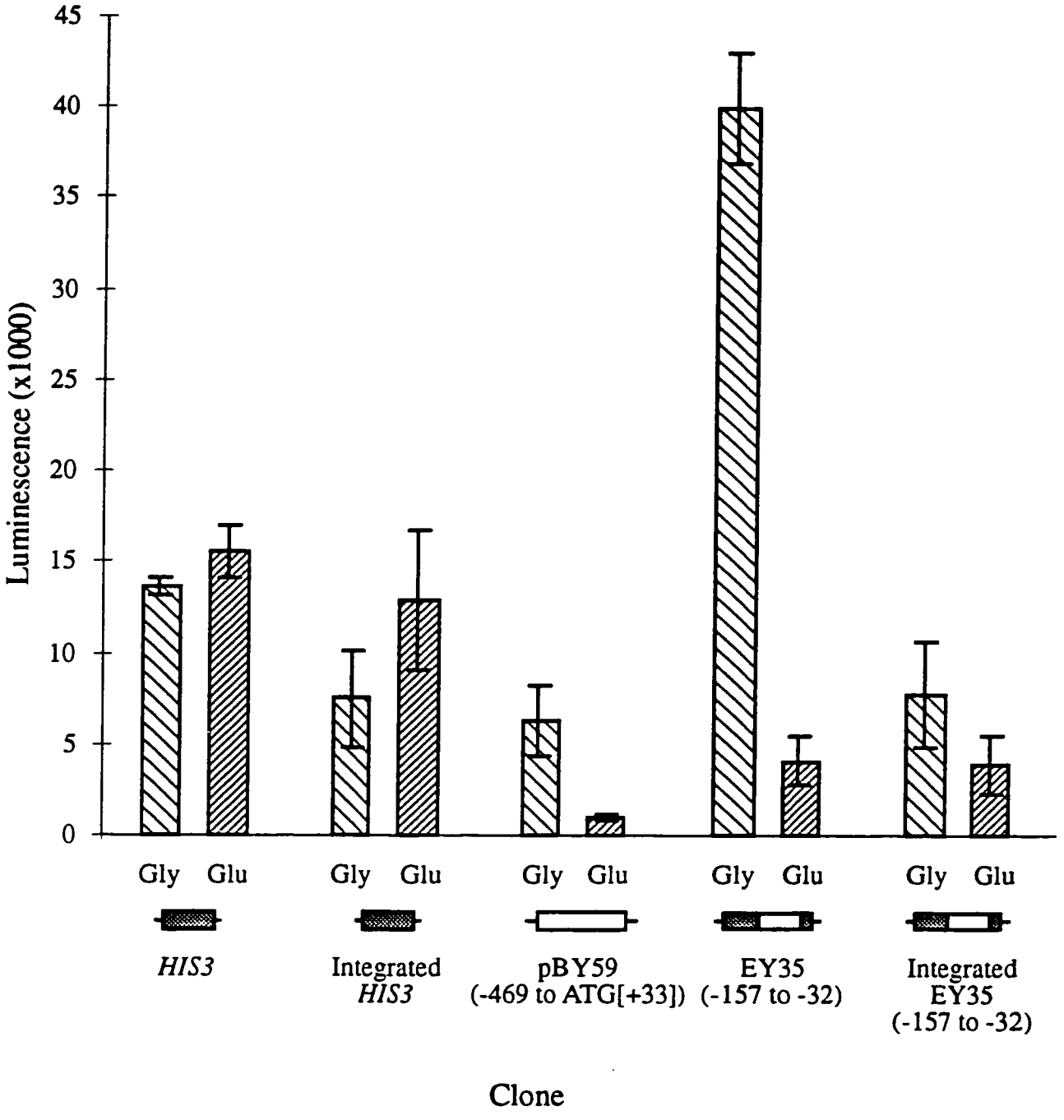


Figure 2.6. Comparison of expression levels between integrated and nonintegrated constructs. The ranges in alpha-amylase promoter sequence shown are all relative to the transcriptional start site. Gly = YPD cultures containing 5% vol/vol glycerol, Glu = YPD cultures containing 5% wt/vol glucose. A minimum of 4 clones from at least 2 independent transformations were used for each mean. Error bars correspond to ± one standard error.

2.4 Discussion

2.4.1 *Comparison of gene expression between the transgenic* Saccharomyces *system and* in vivo Drosophila *assays*

Presented in this chapter are the results from a transgenic *Saccharomyces* system designed to assay *Drosophila* α-amylase promoter activity under different nutritional regimes. Traditionally these types of gene expression assays have been conducted using somatically transformed *Drosophila*, however it was felt that the limit for resolving the specific sequences of the α-amylase regulatory elements was fast being approached with the current system. Given the success with deducing very short (*i.e.* < 30 bp) repression motifs in yeast, the techniques developed in yeast were applied to characterize the *Drosophila* α-amylase glucose repression sequences.

Using yeast transformed with a luciferase-based expression cassette, 5 of 7 nonintegrated constructs driven by promoters containing alpha-amylase promoter sequence expressed in a glucose repressible manner (Figures 2.3 and 2.5). Quantitatively however, the maximum degree of repression seen with the yeast approach was 27 fold, and most constructs showed a 6 to 10 fold average reduction in reporter gene expression in the presence of glucose. This is much less than the greater than 100 fold repression seen naturally for this gene in *Drosophila* (Benkel and Hickey, 1986b). Nonetheless it is the range seen for genes in which direct MIG1-dependent repression has been separated from the other systems of repression and quantified, such as >9 fold for *SUC2* (Nehlin and Ronne, 1990) and 3-4 fold for *GAL1* (Johnston and Carlson, 1992; Johnston *et al.*, 1994).

Such small effects on the expression of a single gene in *Saccharomyces* can result in very dramatic changes in overall gene expression through their amplification in signal transduction pathways and through the presence of multiple pathways of convergent regulation (*e.g.* see the GAL genes of Section 1.4). It is possible that the *Drosophila* α-amylase gene is subject to multiple mechanisms of glucose repression and not all of them can be assayed in yeast, thereby decreasing the levels of repression observed.

An equally plausible explanation for the difference in the degree of repression observed in the transgenic yeast as compared to *Drosophila* is the non-optimization of the fly signal motifs for the yeast transcription system. Although the consensus sequence proposed for MIG1 binding was 5' SYGGRG 3' (Lundin *et al.*, 1994), it was observed on many occasions that not all sequences with this motif bound MIG1 (Flick and Johnston, 1992; Wang *et al.*, 1997). This discrimination between possible MIG1 binding sites has been in part, attributed to the base composition in the 5' flanking region (Lundin *et al.*, 1994). Specifically, the presence of G or C in positions -2 to -6, and especially a C at -3, greatly reduced binding. The corresponding flanking sequence for the upstream putative α-amylase URS is 5' GGGCTT 3' and for the downstream copy, 5' GTAGCA 3'. By this criteria, the upstream element would be expected to be bind MIG1 poorly, resulting in a smaller contribution to overall glucose repression. Such differences can be seen when two repression sequences of *GAL1* are compared. URSA has a flanking sequence of 5' TTATTT 3', but the 5' AGGTTT 3' sequence of URSB is much more GC rich. Repression attributed to the URSA sequence was quantified to be more than 10 fold, but URSB accounted for only 4 fold repression (Flick and Johnston, 1992). Neither WT1 nor EGR1, two human transcription factors in the same family as MIG1 seem to have such flanking sequence requirements (Christy and Nathans, 1989), so it is possible that this is a feature unique to this yeast zinc finger protein. If this is the case, there would be no selective pressure on the α-amylase promoter to maintain such characteristics, thereby accounting for some of the differences in the expression patterns seen when it is expressed in yeast as compared to its native *Drosophila*.

If the presence of the cytosine at position -3 relative to the upstream α-amylase URS results in poor binding of the MIG1 protein, deletion of the more optimal downstream repression sequence should have a negative effect on glucose regulation. This was indeed the case, as shown in Figure 2.7. The different amylase fragments tested are shown relative to their positions in the *AMY* promoter. The expression patterns of clones pBY59, EY5 and EY6 showed that the URSs are located between -168 and the ATG [+33]. The nonintegrated

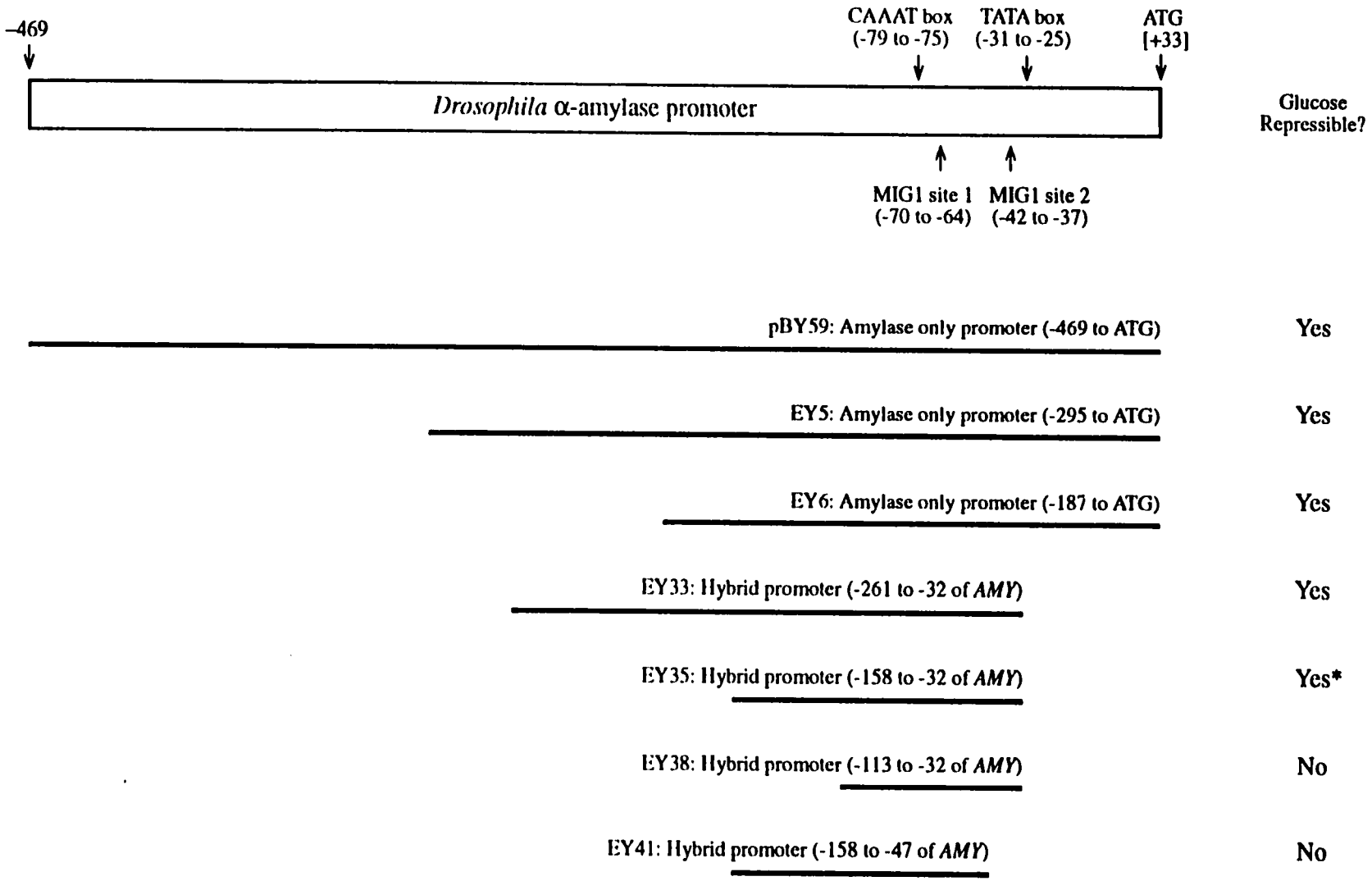


Figure 2.7. Summary of the *Drosophila* α-amylase promoter regions tested (drawn to scale). The reference numbers used are relative to the transcriptional start site of the alpha-amylase promoter. Their ability to regulate glucose repression in a transgenic Saccharomyces expression system is also listed. For clone EY35 which is marked with a "*", statistically significant levels of glucose repression were only observed when the expression cassette was not integrated into the yeast genome.

versions of EY33 and EY35 showed that deletion of the TATA box and the remaining downstream sequence did not affect glucose repression, although this could not be confirmed through the integration of EY35. The elimination of the downstream URS abolished glucose repression in EY41. When this motif was eliminated via mutagenesis in *Drosophila* expression experiments (Magoulas *et al.*, 1993a) glucose repression was not affected, which seems to indicate that the upstream URS is more effective in regulating repression in flies than in yeast. The one unexpected observation was that elimination of the sequence between -158 and -113 also affected glucose repression. This was the case in *Drosophila*. Although there is a 5' ATGGGG 3' motif at -124 to -129 in the lower strand, it is only a 5/6 match to the MIG1 consensus sequence, and unlikely to be a *bona fide* binding site. The difference in expression is most likely due to the obviously different kinetics governing DNA binding by MIG1 and the as of yet to be characterized, *Drosophila* MIG1 homologue.

2.4.2 *Comparison of expression levels between different clones*

In Figure 2.5 one of the most noticeable features is the large range in mean expression levels between the hybrid clones, such as the 3x higher levels of EY35 in glycerol media compared to the *HIS3* control. As in the case of MIG1, most activators and repressors function in a context-dependent manner. When α-amylase fragments of different lengths are inserted into the *Eco*RI site of the base *HIS3* promoter, not only does the sequence at the junction points change, but the distance between the *HIS3* region upstream of the *Eco*RI site and transcriptional start site changes. Many of the resulting effects are unpredictable, but due to recent advances in the development of transcription factor binding site databases, it is possible to test the α-amylase promoter sequence for the presence of putative yeast protein binding sites.

The upstream sequence of the *Drosophila* amylase gene (Hickey *et al.*, 1987) was used to search the TRANSFAC database of DNA binding sites, using the TFSEARCH and

MatInspector[1] (Quandt *et al.*, 1995) programs. Under high stringency search conditions, putative binding sites were identified for a variety of yeast transcription factors, although none of them would be expected to be active under the growth conditions used. These included such proteins as ADR1 (activator of alcohol dehydrogenase 2 which is repressed in glucose grown cells), HSP (cells were not subject to heat shock), and GCR1 (activator of glycolytic enzymes which are repressed in fermentative growth). One hopeful candidate was the yeast ABF1 protein which is known to transactivate the *QCR8* gene - a subunit of the mitochondrial ubiquinol cytochrome c oxidoreductase, under fermentative growth (Mulder *et al.*, 1995). Binding sites for this factor could only be identified under relatively low stringency search conditions however, which greatly decreases the possibility that any "hits" correspond to functional binding sequences. It appears that the enhanced luciferase expression of such clones as EY35 is the result of promoter interactions which could not be predicted beforehand.

2.4.3 *Comparison of expression patterns between integrated and nonintegrated plasmids*

Nonintegrating yeast plasmids are known to occasionally exhibit inconsistent expression patterns (Fukasawa and Nogi, 1989). This is often associated with plasmid loss through differential segregation of the plasmid population in the daughter cells during mitosis. The pBTI-1 plasmid used in these experiments is of the relatively high copy episomal variety (20-50 copies/cell), but since the luciferase assays were conducted in rich media, there was no selection for plasmid maintenance. As well, pBTI-1 contains the autonomous replicon regions from the endogenous 2 µm yeast plasmid. It is known that there are ~600 bp repeat sequences in this region which can cause homologous recombination between plasmids, creating a myriad of products (Lunblad, 1993). With this in mind, the high degree of variance in some of the luciferase assay results may be partially due to the choice of plasmid. By using the nonreplicative and integrative vector pRS405, some of these issues may be eliminated, as no 2 µm-derived sequences are used and the copy number should be conserved at 1 per cell.

[1]TFSEARCH is accessible through the internet at http://www.pdap1.trc.rwcp.or.jp/research/db/TFSEARCH.html MatInspector is found at http://transfac.gbf.de/cgi-bin/matSearch/matsearch.pl

When the expression pattern of the integrated and nonintegrated copies of the *HIS3* control promoter and EY35 were compared (Figure 2.6), it was noted that the integrated forms expressed at lower levels. The simplest explanation is that the lower copy number of the integrated form results in lower expression levels. There are other, more complicated possibilities however. One possibilitity is that the expression cassette has integrated into a transcriptionally inactive region of a chromosome. Since the integration was targetted to the coding region of a key biosynthetic gene (*HIS3*) however, it is very unlikely that the vector is surrounded by heterochromatin. Gene expression studies can be plagued with problems from cryptic or fortuitous promoter sequences in the vector (M. Johnston, pers. comm.). If this is the case, since the vectors used in the integrative and nonintegrative studies were not derived from one another, the observed reporter gene expression may be subject to completely different influences from the vector sequence. Such influences would have to be cryptic for both the pBTI-1 and pRS405 vectors, since the reporter construct was cloned in opposite orientation to the *LEU2* marker in each case, so read-through from this gene should not occur.

With the limited number of comparisons between integrated and nonintegrated plasmids conducted in this study, it is impossible to determine how much of an effect such vector influences had on the expression assay data. With the EY35 construct in particular, the statistically significant levels of glucose expression observed in the nonintegrated state were not repeated when the vector was integrated. When integrated, EY35 did show a trend of higher expression in glycerol-based media compared to glucose-based media (8,200 vs. 4,300 units), but when tested using the paired T-test, the difference in means was not significant. Since a single copy, integrated expression cassette would be presumed to be more similar to the natural genomic organization, can it then be said that the 126 bp α-amylase fragment contained in EY35 has the required components to regulate glucose repression in *Saccharomyces*? This question could have been unequivocably answered had more constructs also been integrated. In hindsight, the EY35 construct might not have been the best choice to test the effects of integration, as it was the only construct which exhibited induction in addition to repression

when unintegrated (Figure 2.5). Both pBY59 and EY33 showed glucose repression but no induction when assayed as episomal plasmids. By observing the behavior of these clones in an integrated state, one would have a better sense of how the *Drosophila* α-amylase promoter expresses when integrated into the yeast genome. This in turn, would have provided a benchmark against which the EY35 results could have been compared to assess their validity.

The integration vector was tested since it was thought that integration would aid in reducing the high levels of variation observed in the data for many of the episomal plasmids. Additional integration experiments were not conducted at the time because the initial data using the EY35 construct did not show any reduction in data variance. It is unknown whether or not vector sequence flanking the expression cassette is responsible for this. The only way to completely eliminate any plasmid-derived effects is to integrate the expression cassette in single copy, with minimal flanking vector sequence, downstream of a strong transcriptional termination signal. This would involve changing the site of integration and designing a intraplasmid recombination system to excise unnecessary vector sequence - a feature which the pRS405 plasmid used here lacked. Indeed, all of these considerations were incorporated into the cloning strategy of Flick and Johnston (1992), which may explain their consistent results in delineating the *GAL1* repression elements to a degree of resolution which was not obtainable with these experiments on the *Drosophila* alpha-amylase gene. The data obtained from these transgenic experiments do indicate however, that there is a high level of evolutionary conservation of the glucose repression mechanism between yeast and a higher eukaryote - to the extent that *Drosophila* URSs are recognized and functional in *Saccharomyces*.

Chapter 3

CONSERVATION OF THE SNF4 GENE IN YEAST, MAMMALS AND *DROSOPHILA*

3.1 Introduction

Up to now, this thesis has only described the process of gene inactivation due to glucose repression. As it is said that with every action there is an equal but opposite reaction, there must be a complementary pathway in which the repressed genes are activated in the absense of glucose. This chapter will introduce some of the protein components which are thought to be involved in the derepression of *Drosophila* genes, as extrapolated from yeast and mammalian systems. Genomic and cDNA copies of one of these components, *SNF4*, were cloned and sequenced, and an analysis of these data shows that it belongs to a small family of related multicopy genes. Different strategies were attempted in the process of isolating this gene - of which not all were successful, therefore a portion of this chapter is dedicated to explaining the early problems encountered in cloning the *Drosophila SNF4* homologue.

3.1.1 *Role of* SNF4 *in yeast glucose derepression*

As explained in the previous chapters, glucose repression in *Saccharomyces* occurs when the transcription factor MIG1 binds to the promoter of the gene to be repressed. The binding of this protein to the DNA results in the subsequent recruitment of a general repression complex composed of TUP1 and SSN6. In the absence of glucose, *i.e.* under glucose derepressing conditions, the TUP1/SSN6 repression complex is inhibited from interacting with MIG1 due to changes in the structural conformation of this transcription factor. The change in conformation is attributed to the phosphorylation of MIG1 by an upstream kinase (Treitel and Carlson, 1995).

The activity of this kinase is dependent on the presence of three subunits. *SNF1* is the locus which codes for the catalytic kinase subunit, the *GAL83* family of loci code for the protein-protein linker subunit and SNF4 has been characterized as the positive effector subunit (see Section 1.5). *SNF4* from *Saccharomyces* was initially cloned by two independent groups. Schüller and Entian (1988) first reported the sequence for a gene involved in glucose derepression which they had named *CAT3*. However, when Celenza *et al.* (1989) published their sequence for *SNF4* a year later, it was revealed that these two genes were in fact, the same locus. *SNF4/CAT3* (hereafter referred to as *SNF4*) codes for a 322 amino acid protein which did not share significant homology to any other genes known at that time, although this has changed since then.

3.1.2 *Comparison of the yeast* SNF1/SNF4/GAL83 *with their mammalian homologues*

The significance of *SNF1* and *SNF4* in metabolic gene regulation underwent a major upheaval when a particular mammalian kinase was cloned and sequenced (Carling *et al.*, 1994; Mitchellhill *et al.*, 1994). This kinase, the AMP-activated protein kinase (AMPK), has long been known to be a key mammalian regulator of biosynthetic pathways. AMPK phosphorylates hydroxymethyl-glutaryl CoA reductase and acetyl CoA carboxylase, which are enzymes that catalyze the rate limiting steps in cholesterol/isoprenoid biosynthesis and fatty acid biosynthesis, respectively (reviewed in Hardie, 1992). When the protein sequences for *AMPK* and *SNF1* were compared, they showed 46% identity - a degree of conservation which would indicate a common evolutionary ancestor.

At first glance, the connection between the regulation of glucose repression in yeast and macromolecule biosynthesis in mammals seems difficult to view, but taken at a more biochemically fundamental level, both of these systems are responsive to cellular energy levels. In *Saccharomyces*, glucose is the most favored carbon source. Depletion of extracellular glucose would represent an energy stress to the organism, and SNF1 is instrumental in derepressing the necessary genes for metabolism of alternative carbon sources to maintain an

inward flux of carbohydrates. Likewise, when mammals experience a stress in their energy levels, it is prudent to downregulate energetically costly pathways such as cholesterol and fatty acid synthesis. Heat shock, oxidative stress and hypoxia are all known to decrease cellular ATP stores, resulting in an increase in AMP. It is an increase in this AMP/ATP ratio which activates AMPK activity in order to preserve ATP for essential processes. Therefore, both SNF1 and AMPK act as metabolite sensing protein kinases, playing key roles in energy utilization. The seemingly large difference in their functions is merely a remnant of the disparate backgrounds which led to their cloning (Hardie, 1994).

There is additional evidence which suggests that AMPK and SNF1 form a kinase family. First, the yeast SNF1 and the mammalian AMPK share similar substrate recognition motifs (Dale *et al.*, 1995), and they have both been shown to phosphorylate acetyl CoA carboxylase (Mitchelhill *et al.*, 1994; Woods *et al.*, 1994). Second, although the yeast kinase which activates SNF1 has not been determined, both SNF1 and AMPK can be activated by AMPKK, the mammalian kinase which activates AMPK (Wilson *et al.*, 1996). There are some differences between the two kinases though - SNF1 is not allosterically regulated by AMP as AMPK is (Carling *et al.*, 1989; Wilson *et al.*, 1996), and the rat AMPK gene did not complement a yeast *snf1* mutant (Woods *et al.*, 1994).

Perhaps one of the most convincing arguments that yeast SNF1 and mammalian AMPK are physiologically and evolutionarily related is the finding that the architecture of the yeast SNF1 complex is maintained in mammals. Mitchelhill and coworkers (1994) showed that AMPK coprecipitates with two other proteins, and when these proteins were cloned and sequenced in rat they were homologous to the yeast *SNF4* and *GAL83* (Gao *et al.*, 1996; Woods *et al.*, 1996). Finally, the mammalian derepression complex assembles in the same manner as seen in yeast, as it has been shown that the mammalian homologues of SNF4 and GAL83 bind to each each other, and this heterodimer interacts with the catalytic subunit for full activation of AMPK (Dyck *et al.*, 1996).

Given this high degree of similarity between the mammalian and yeast SNF1(AMPK) complexes, it was expected that these proteins would also be conserved in *Drosophila.* Indeed, this has turned out to be the case. Our lab has already cloned two gene copies of the *Drosophila SNF1* kinase (E. Taboada, pers. comm.). To complement this work, this thesis reports the cloning and sequencing of the genomic and cDNA copies for *SNF4*, a putative noncatalytic subunit of the derepression complex from *Drosophila melanogaster.*

3.2 Materials and methods

3.2.1 *Degenerate primer design and PCR reaction profiles*

When GenBank was searched for *SNF4* homologues, it contained 4 cDNA sequences and one amino acid sequence. These sequences were derived from rat (X95578), human (U42412), *Saccharomyces cerevisiae* (M30470), *Schizosaccharomyces pombe* (Z69944), and pig (partial amino acid data, Q09138). They were aligned using ClustalV (Higgins *et al.*, 1991) to reveal short regions of local sequence conservation. Degenerate primers were constructed to target several of these regions, with the primer sequence biased towards the mammalian sequences, rather than the yeast ones (Table 3.1). Although the primers were designed to minimize sequence degeneracy, the levels were still quite high, as indicated: K1188F (2048x), K1140F (576x), K1214F (2048x), K1189R (512x), and K1141R (1024x).

PCR reactions were performed in 25 µl reaction volumes, using 1 U Taq DNA polymerase (Promega), 200 mM of each dNTP (Gibco BRL), 1 - 6 mM $MgCl_2$, 50 ng of template DNA, and a range of 72 - 144 pmol of each primer. The template DNAs used were a *Drosophila* larval λgt11 cDNA library (Stratagene) and Oregon-R genomic DNA isolated using the Qiagen genomic DNA extraction kit. PCR profiles varied depending on the primer combinations used, but all reactions used the touchdown strategy (Roux, 1995). A typical touchdown profile consisted of: 98 °C x 1 min, initial annealing temperature x 30 sec, 72 °C x 1 min, 1 cycle => 4 x (95 °C x 30 sec, previous annealing temperature - 2 °C x 30 sec, 72 °C x

Table 3.1. Description of the degenerate primers synthesized to amplify *SNF4*

Primer Name	Nucleotide and Amino Acid Sequence	Known Sequences	Position with respect to rat sequence	Primer melting temperature (°C)
K1188F	5' GA(CT) ACN TCN (TC)TN CA(AG) GTN AA 3'	Rat DTSLQVK	50 - 57	50 - 64
	D T S L Q V K	Human DTSLQVK		
		Sacch DTSLLVK		
		Schiz DVTLFVK		
K1140F	5' ACX AT(TCA) ACX GA(TC) TT(TC) AT(TCA) AA 3'	Rat TITDFIN	80 - 86	46 - 58
	T I T D F I N	Human TITDFIN		
		Sacch TTTDFIN		
		Schiz TMADFVN		
K1214F	5' GCN (TC)TN CCN GTN GTN GA 3'	Rat ALPVVD	226 - 231	48 - 60
	A L P V V D	Human ALPVVD		
		Sacch SVPIID		
		Schiz AVPIVN		
K1189R	3' GGN CAN CAN CT(GA) CT(CT) TT(CT) CC 5'	Rat PVVDEKG	228 - 233	56 - 76
	P V V D E K G	Human PVVDEKG		
		Sacch PIIDENG		
		Schiz PIVNSEG		
K1141R	3' AA(AG) CT(CT) CCN CAN (AG)AN TT(CT) AC 5'	Rat FEGVLKC	272 - 278	50 - 64
	F E G V L K C	Human FEGVLKC		
		Sacch FEGVYTC		
		Schiz FDGVHTC		

"Sacch" refers to *Saccharomyces cerevisiae* and "Schiz" to *Schizosaccharomyces pombe*. Refer to Section 3.2.1 for the accession numbers for the sequences used.

1 min, 5 cycles) => 95 °C x 1 min, final annealing temperature x 30 sec, 72 °C x 1 min, 30 cycles. Optimal conditions were obtained when the touchdown reaction began at 60 °C and was lowered to 53 °C for the majority of the cycles.

3.2.2 *Genomic library screening*

A 684 bp fragment from a human *SNF4* cDNA was PCR amplified for use as a probe using the primers K1188F and K1141R (Table 3.1). The template DNA was an EST clone ordered from the American Type Culture Collection (accession W32636), for which plasmid DNA was prepared using the Qiagen miniprep kit. The reaction conditions to amplify the probe were 30 cycles of: denature at 95 °C for 30 sec, anneal at 50 °C for 15 sec, and extend at 72 °C for 1 min, with an extension time of 5 minutes for the final cycle. The reagent conditions were the same as previously specified (Section 3.2.1), with 72 pmol of each primer being used. The product was gel purified and sequence verified prior to being used as a probe.

The library screened was a Stratagene *Drosophila* genomic library in λFIXII, which had an average insert size of 9-23 kbp. K802 plating cells were prepared (Sambrook *et al.*, 1989) and 40,000 pfu of the library were plated over 2 LB plates containing 10 mM $MgSO_4$ and 0.2 % maltose. After an overnight 37 °C incubation, the plaques were transferred to MAGNA nylon membranes (MSN); denatured in 0.5 M NaOH, 1.5 M NaCl for 5 minutes; neutralized in 1.5 M NaCl, 0.5 M TRIS-HCl (pH 8.0) for 10 min; and the DNA crosslinked to the filters using a Stratagene UV crosslinker. The filters were prehybridized and hybridized overnight in a solution of 5 x SSC, 5 x Denhardt's solution, 50 mM TRIS-HCl (pH 7.5), 0.5 % SDS, 50 % formamide with 0.25 mg/ml sonicated, denatured herring sperm DNA. The probe was radiolabelled using 50 ng of the PCR amplified human probe, 50 μCi of $\alpha^{32}P$ dCTP (Amersham), and the Prime-It® RmT Random Primer Labeling kit (Stratagene). Unincorporated nucleotides were removed using NucTrap columns (Stratagene).

Hybridizations were performed at low stringency at 30 °C overnight, and the filters washed in two steps using 2 x SSC, 0.1 % SDS at room temperature for 20 minutes, followed

by a 30 minute wash at 37 °C in 2 x SSC, 0.1 % SDS. The filters were exposed to BioMax film (Kodak) for 2-3 days at -80 °C.

3.2.3 *Sequencing*

Sequencing primers were made using an ABI 381A DNA synthesizer. All sequencing was done using double stranded plasmid DNA with the Dye Terminator Ready Reaction kit and the 373A automated sequencing system (ABI). Sequencing reactions consisted of 500 ng of plasmid DNA, 10 pmol of primer, 5 µl of halfTERM sequencing reagent (Genpak), and 3 µl of ABI sequencing premix. PCR cycling conditions were as specified (ABI), and the reaction products were cleaned by 95 % ethanol precipitation using 0.3 M sodium acetate, followed by a 70 % ethanol wash.

3.2.4. *Southern and Northern analyses*

Lambda DNA and *D. melanogaster* Oregon-R genomic DNA were prepared using Qiagen's Lambda DNA Extraction kit and Genomic DNA Extraction kit, respectively. For the Southern analysis, 5 µg/lane of the restricted DNA was electrophoresed in a 1% agarose gel, and later capillary transferred overnight in 20 x SSC to a HybondN$^+$ membrane (Amersham). Total RNA for the Northern was prepared by the phenol method (Sambrook *et al.*, 1989), and the polyA$^+$ RNA further purified with the Oligotex mRNA extraction kit (Qiagen). Approximately two micrograms of mRNA per lane were electrophoresed in an ethidium bromide-containing denaturing gel (1.1 % agarose, 2.2 % formamide, 20 mM MOPS, 5 mM sodium acetate, 1 mM EDTA, pH 7.0). The gel was prepared for transfer by soaking in 10 x SSC, 0.05 N NaOH at room temperature for 10 minutes, followed by two room temperature washes in 20 x SSC for 20 minutes each, before transferring the RNA to a nylon support as above. The probe consisted of a 1.1 kb *Drosophila SNF4* cDNA fragment which was PCR amplified from a λgt11 (Stratagene) *Drosophila* larval cDNA library. Radioactive probes and hybridization solution were prepared as per Section 3.2.2. High stringency conditions

consisted of an overnight 45 °C hybridization, followed by washes in 0.1 - 0.5 x SSC, 0.1 % SDS at 65 °C. The low stringency conditions have already been described (Section 3.2.2).

3.2.5. *Sequence alignments and phylogenetic analysis*

Sequences were aligned using ClustalV and phylogenetic trees generated using the parsimony algorithm of Phylip 3.5 (Felsenstein, 1988). The full length sequence of the mouse and second human SNF4 cDNAs were generated as the consensus of several overlapping ESTs. For the mouse sequence, the accession numbers of the clones used were AA198401, AA259329, AA008244, and AA178898. The second human SNF4 was assembled from AA304552, H15390, W39604, HSC2WF121, and H06773.

3.3 Results

3.3.1 *Cloning and analysis of the genomic clone for* Drosophila SNF4

The initial approach involved the use of degenerate primers to PCR amplify fragments spanned by relatively well conserved sequence motifs (Table 3.1). Using both *Drosophila* genomic DNA and a *Drosophila* λgt11 larval cDNA library (Stratagene) as templates, PCR products of the expected size were only amplified using the primer combination K1214F/K1141R (159 bp). Although more than a dozen different clones were sequenced, nucleotide searches of GenBank indicated that the clones were mainly previously characterized *Drosophila* sequences and none showed any homology to yeast or mammalian *SNF4*.

With the lack of success with the degenerate PCR strategy, conventional library screening using a heterologous SNF4 probe was then successfully employed to clone the *Drosophila SNF4* homologue. The genomic library was initially screened using a 37 °C hybridization temperature but no positives were seen after a 3 day exposure. As a result, the same filters were reprobed using fresh solutions and a 30 °C overnight incubation. Under these lower stringency conditions, 21 positive signals were detected after the first screening.

Sixteen of the 21 primary positive clones were subject to a second screening, with 9 of the 16 being positive. Six of these clones were subject to a third round of screening, and all 6 showed that they were purified positives. A final 3 clones were randomly selected and their DNA extracted from plate lysates.

The DNA isolated from these three final clones was cut with several enzymes, and the restriction profiles probed with the human *SNF4* cDNA to identify fragments for subcloning. Two of the clones (isolates #15 and #18) showed nearly identical restriction patterns, and the third (isolate #5) was slightly different. Based on their ability to hybridize with the human sequence, a 3.5 kb *Hin*dIII fragment from isolate #18 and a 2.3 kb *Sac*I fragment from #5 were then subcloned into pBluescriptII KS(+) (Stratagene). The 3.5 kb *Hin*dIII fragment was further subcloned as three *Pst*I fragments with the following lengths: 0.7 kb, 0.9 kb, and 1.9 kb.

The sequencing strategy used for the 3.5 kb *Hin*dIII clone is shown in Figure 3.1. Using a primer walk approach, it was determined that the 3.5kb *Hin*dIII genomic fragment contains the last 4 exons which code for the *Drosophila melanogaster* homologue of *SNF4*. No other putative ORFs were detected. The exons range in size from 178 bp to 779 bp and the introns from 62 bp to larger than 1536 bp (Figure 3.1). The ends of the 2.3 kb *Sac*I fragment from clone #5 were also sequenced. This clone is a shorter version of the 3.5 kb *Hin*dIII fragment which spans from position 366 to the *Sac*I site at 2765, and its restriction profile is in agreement with that of the larger *Hin*dIII fragment (data not shown).

The full sequence of the 3.5 kb fragment and its predicted protein translation are shown in Figure 3.2. Coded within this fragment is a polypeptide of 484 amino acids with high levels of sequence similarity to the yeast and mammalian *SNF4*. It did not contain the translational start codon however. The remainder of the subclone was sequenced, but within the 1,536 bp that were sequenced further upstream of the 5'-most splice site, the junction to the next upstream exon could not be determined.

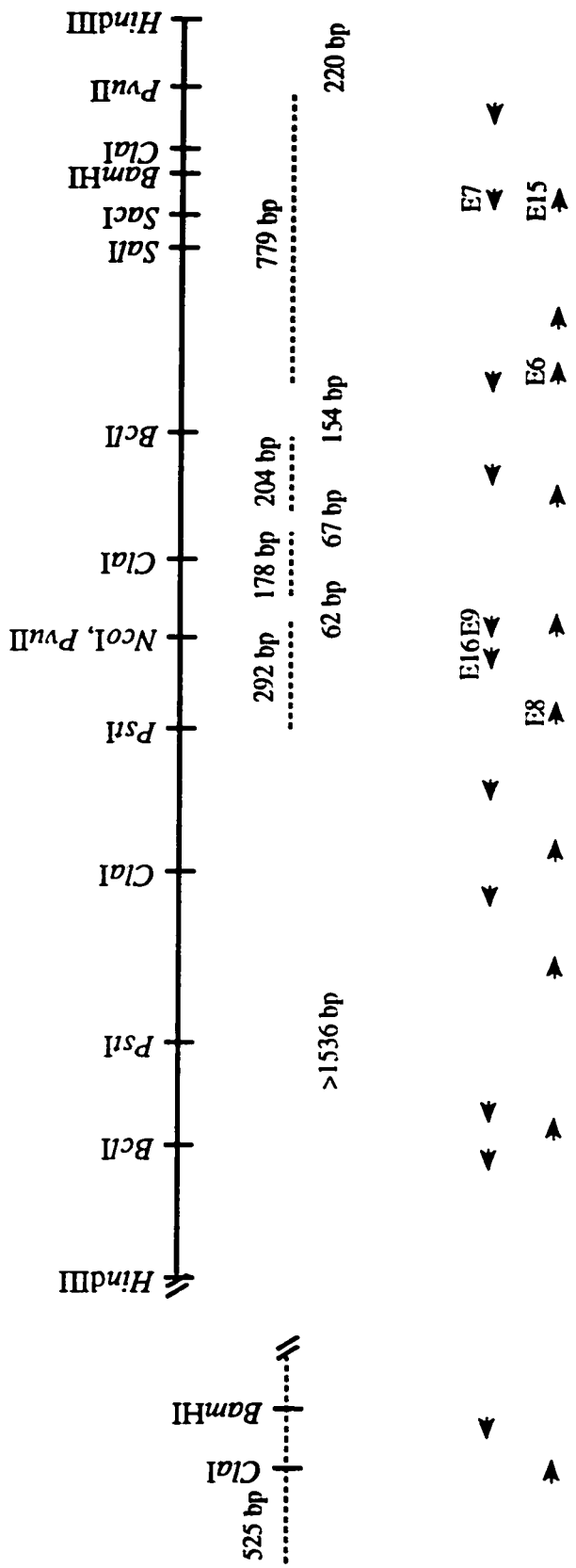


Figure 3.1. Restriction and sequencing map of the 3.5 kb *Hin*dIII genomic fragment and 525 bp 5' cDNA fragment. The lengths of the exons are listed above the dotted line and below the line for the introns. Specific primers which are referred to by name in the text are identified. Contiguous clones containing all exons were not obtaincd, as shown by the double lines at the 3' end of the first exon.

```
SNF4_genomic      1 AAGCTTAGTC GGGAATATCT TTTTGCCGCT TGCATGCATA TAGCTCCAAA TATTTTACAA TTTATTTCGA TGTGTATATT TTCTGGCCGC GGTATCTGGC
SNF4_protein
                                                                                                                          0
                101 CTTCGTGTTT TTTCTATATA TTATATTTGG TCGTTTTTAA ACTTTCTGTT TGTCTGTTGT TTGCTTATAT TGTGGTGGAT ATCATCCGGA TAATTTGTTG
                                                                                                                          0
                201 GAAATACTTT TACCAGTGGT TTTAAAAATT TCCACTAATC TCATACTGGG AGGCGGCAGG TTTTCCTTAA ATTATTTCCG CTCATTAATG CATTCCATTT
                                                                                                                          0
                301 CTATGTCGAT GAATCATAAA TTTACTTTGA TCTGACTTCG TGCCAAATAA AACATCAGAT TGTCTGATCA AAAAATAACC TGCCAATTAG TAAATTGAAC
                                                                                                                          0
                401 GAAAATAATT TTTTGCACTT TAAATAAAGT TATTTATTAT TTCTGAAATA GATTTCCCTA TGGTTTTCGC AAAAGAGGCG CACTAATCCC TACAGACATT
                                                                                                                          0
                501 CGAAATTTCA GTGTTGAGTC TTATGCTGCG GTTAATTAAT TTCTTGGCAC AACTGCCTGT CCCAATATCT GTTATGTGTC GCTATAATGG CTGAATTGGT
                                                                                                                          0
                601 TTGGGTGTAT GTTTGGCTTG CATTTATCAG GCCCCTGACA AGACCGATCT GCAGCCCCTG TGATTTTATA TAGTTTTTGA TTAACACCTT TTGGCCCCGC
                                                                                                                          0
                701 CTCAATATTT ATTTGTTTAC ATTGTGCCAC AGCACAGTAG GCAGTCTACA TTATGCCCCT CTCACTTGGT GGTTTCTCTT TCTCTAAATG GGAATTGAAA
                                                                                                                          0
                801 TAATAATATT TATTGTTGAA CAATTTGGCG AATCTATTCT ATTTCGATGT GTTTTCCTAA GCTGTGGCTG CTTCGGTTGC TTTTGTCGTA CGCTAAATTT
                                                                                                                          0
                901 ATTTTAATTA AAAATGTCTT AATTTAATTC AACGAAACGC GTGCGCCCTC AAATTTATGG GCACCGGTTG TTCTATTTTT GGGTTCGAAT ACGCTGAGCG
                                                                                                                          0
               1001 CGAATAAATA TCGCCTAACA TCCGTAGCTC ACTACAGTGA GAAGCGCGTA GAGCGGCGCA CTCTGTTAGC GGGCTGTTAG CTAGTTGGAC AATCAATAAG
                                                                                                                          0
               1101 GCAAATCAAT TAGCGAAGCA AATCGATTCG CGCAGTTAAT TGCTAGCGCC CACAGTTCAC AGGCAGGAGA CCAGATCAAA TTGGATCGCA TCAATCTAGC
                                                                                                                          0
               1201 GTAATGCGAC ACTGAAAATC GCTTAAAAAC AAAAACCCAA AAATAAAAGA GCGAAGCGTC AAAGTGAAGC ACAGTGGAAA CCCAGAACGC ATGTTACAGT
                                                                                                                          0
               1301 TGCACTTCTA CTCGAATCAT GCGGGTCGTT TGGTACTCGG AGGTGGCAAC CGTCGCCCGA CCGACGTCCT TTGCTATCCG TTGGAGAGCA TCAAAGAGTT
                                                                                                                          0
               1401 CGAGCAGATC TCTGAGGCCT GGAAACGCCT GCTAGCCGAG TTGCTCAAGA AAGGTGGACA TCCCAGATCC CTAAACCTTC AGCCCATATC GTAAATAAAT
                                                                                                                          0
               1501 CCATTATTTA ACCGACTTCT CTTCTTATTT CTGCAGAAGA GGACGACTCA CAGATCTTCG TGAAGTTCTT TCGTTTTCAC AAGTGCTATG ATCTGATACC
                                                                                    E  D  D  S  Q  I  F  V  K  F  F  R  F  H  K  C  Y  D  L  I  P    21
               1601 CACCTCCGCC AAGTTGGTTG TCTTCGACAC CCAGCTTCTT GTAAAGAAGG CCTTCTACGC CCTCGTCTAC AACGGTGTGC GAGCGGCACC GCTCTGGGAT
                     T  S  A  K  L  V  V  F  D  T  Q  L  L  V  K  K  A  F  Y  A  L  V  Y  N  G  V  R  A  A  P  L  W  D    54
               1701 TCGGAGAAGC AACAGTTCGT GGGCATGCTA ACCATCACGG ACTTTATCAA GATCCTGCAA ATGTATTACA AATCGCCAAA TGCGTCCATG GAGCAGCTGG
                     S  E  K  Q  Q  F  V  G  M  L  T  I  T  D  F  I  K  I  L  Q  M  Y  Y  K  S  P  N  A  S  M  E  Q  L  E    88
               1801 AAGAGCACAA ACTGGACACG TGGCGGGAGT AAGTGAAAAC CCACATTTGC ATTGAATGAT GATACTAACT AAATGGTATC CTCTGCTAAG CGTGCTGCAC
                       E  H  K  L  D  T  W  R  E                                                                V  L  H   100
               1901 AACCAGGTGA TGCCGTTGGT CAGCATCGGA CCGGATGCGT CCCTCTACGA TGCCATCAAA ATTCTCATCC ACAGCCGCAT ACATCGCCTG CCCGTCATCG
                     N  Q  V  M  P  L  V  S  I  G  P  D  A  S  L  Y  D  A  I  K  I  L  I  H  S  R  I  H  R  L  P  V  I  D   134
               2001 ATCCGGCGAC CGGCAATGTC CTCTACATCC TGACACATAA ACGCATACTT AGGTTCCTTT TCCTATACGT GAGTTTCTGG AATATTAACT ACTTACCTAA
                       P  A  T  G  N  V  L  Y  I  L  T  H  K  R  I  L  R  F  L  F  L  Y                                  156
               2101 AAGCAAAATA TTTATGTGAT ACTTCCTATG CGCAGATTAA TGAATTACCA AAGCCCGCGT ACATGCAAAA AAGTTTGCGC GAACTGAAGA TTGGCACCTA
                                                                                    I  N  E  L  P  K  P  A  Y  M  Q  K  S  L  R  E  L  K  I  G  T  Y   178
```

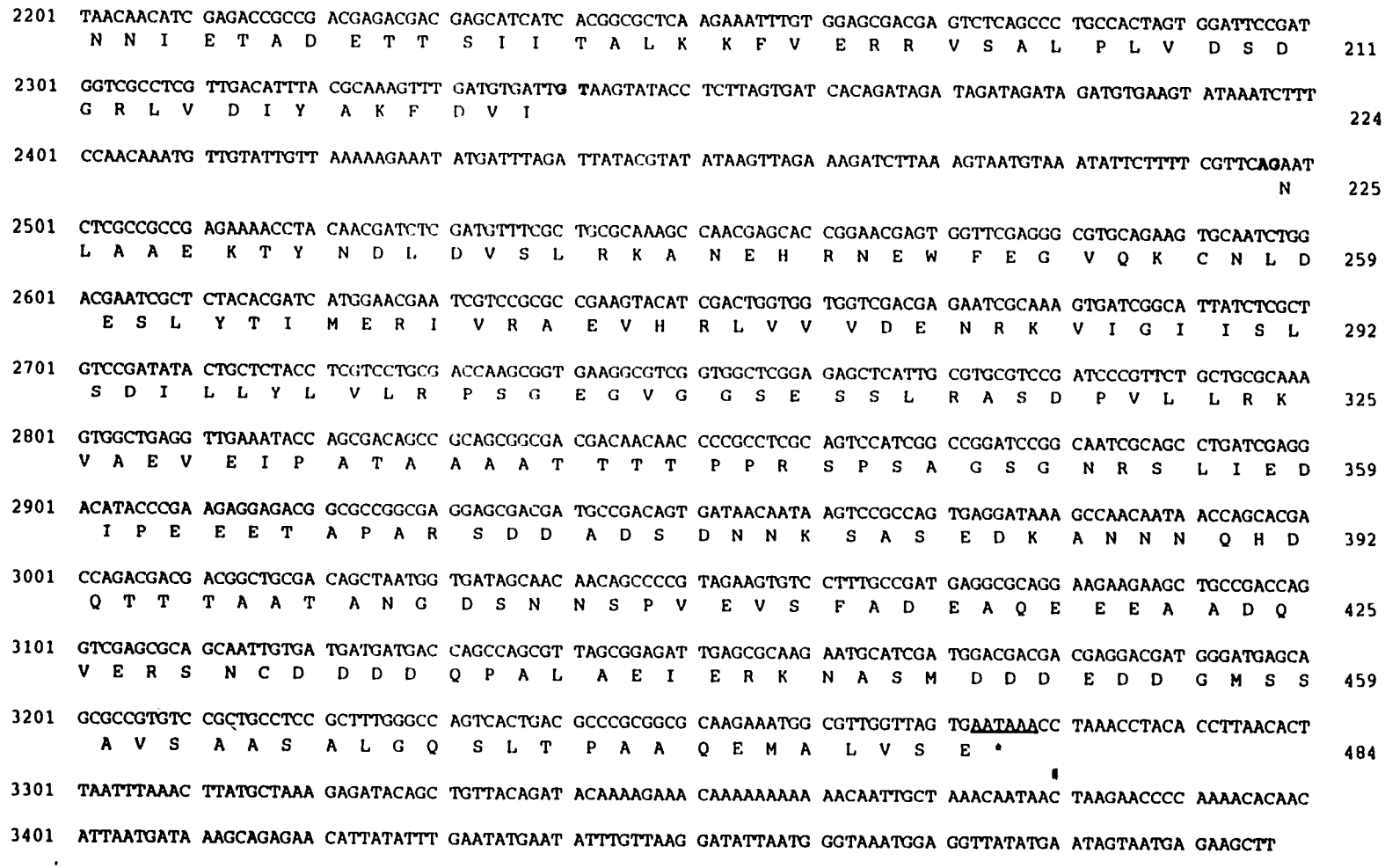


Figure 3.2. Nucleotide and inferred protein sequence of the 3.5 kb *Hin*dIII fragment containing the partial *Drosophila* SNF4 gene. The nucleotide sequence is numbered to the left, and the protein translation to the right. Intron splice sequences are shown in bold and the putative polyA signal sequence, which overlaps with the stop codon is underlined. The site of the polyA tail addition is indicated above the sequence by the "#".

3.3.2 *Cloning and characterization of the cDNA for* Drosophila SNF4

A *Drosophila SNF4* cDNA probe was prepared by PCR amplification of a 1.1 kb fragment using the E8 and E7 primers (Figure 3.1) with a λ gt11 cDNA library template. This probe was used at both high and low stringencies to probe two different cDNA libraries from Stratagene. A total of 1 x 10^5 pfu from each library were screened, but no positive plaques were obtained.

The strategy for obtaining the full length cDNA was then switched to 5' and 3' RACE. For the 5' end, the nested primers E16 and E19 (Figure 3.1) were used with the λ gt11 forward and reverse vector primers (Stratagene). These vector-specific primers were also used with the nested *SNF4* primers E6 and E15 to amplify 3' RACE fragments (Figure 3.1). Using these primer combinations, the cDNA sequence was extended a further 102 bp in the 5' direction and 46 bp in the 3' direction. In the midst of these RACE experiments however, a *Drosophila* embryonic EST was submitted to GenBank which was an identical match to the 5'-most sequence of the *Drosophila* cDNA sequence obtained using the RACE strategy.

This clone (accession LD11267) was ordered from the *Drosophila* EST project (Harvey *et al.*, 1997) and sequenced in its entirety. It contained a 2.1 kb fragment coding for the full length *Drosophila SNF4* cDNA, including 35 bp of 5' UTR and 101 bp of 3' UTR, complete with a polyA tail. The sequence obtained agreed with the gene organization determined from the genomic clone. The nucleotide and amino acid sequences of the 5' end of the cDNA up to the intron junction determined from the *Hin*dIII genomic fragment are shown in Figure 3.3A. When this 525 bp region was used to search the nucleotide database at GenBank, it showed a 43 bp identical match with sequence which flanks the *Mdg3* transposable element (Bayev *et al.*, 1980). This matching sequence was not part of the transposon itself, but ~150 bp downstream of one of the arms of the element. This sequence was mapped to the 93C region of *Drosophila* chromosome 3R using *in situ* hybridization (Bayev *et al.*, 1980). Therefore they must have obtained sequence data from a *Drosophila* strain that contains an insertion of the *Mdg3* transposon into the *SNF4* locus - which must map to 93C.

A.

```
SNF4_cDNA_5'     1  CAGGACAGGG CGCAGCGGGA GGAGCCGGTG GCCTTATGAA CTCCATGAAG GTGGCCATGC AGAACTTTAG CCATCGCCAG CATCCCGCCG TGACGATAAC
SNF4_protein                                          M  N  S  M  K  V  A  M  Q  N  F  S  H  R  Q  H  P  A  V  T  I  T    21

               101  GAGCGCCGAC GGCACGCAAT CAACGGCAAA GAGCAAGTAC AAAGACGGCA GTGCCCATCC GCATCAAGGC AGCGACGCGC AGTATTACCA CACGGTGACG
                     S  A  D  G  T  Q  S  T  A  K  S  K  Y  K  D  G  S  A  H  P  H  Q  G  S  D  A  Q  Y  Y  H  T  V  T    54

               201  GCGGTGCGTC CAAACTCTTC CCAACGGTCG CCGATGACCA AGGTCATGGA TCTGTTCCGG CATCGATCCA GCTCGGTTGT CAGCGAAGCC GACAAACGCA
                     A  V  R  P  N  S  S  Q  R  S  P  M  T  K  V  M  D  L  F  R  H  R  S  S  S  V  V  S  E  A  D  K  R  K    88

               301  AAGCGCGCGC GGCGGCGCAT CAGCAACAGT TGGCTGTGCA AAGTGCTCAC ATGCGTCGTG CCTCCGCGGA TTTGGAGAAA CGTCGTGCAT CAGTTGGTGC
                       A  R  A  A  A  H  Q  Q  Q  L  A  V  Q  S  A  H  M  R  R  A  S  A  D  L  E  K  R  R  A  S  V  G  A   121

               401  CGCAGGTCGA GGACTGCGAG GGGATGGTAC TTTGGATCCA CACCATGCAG CCATCCTCTT CAGAGACTCA CGAGGGTTGC CTGTCGCTGA TCCGTTCCTA
                     A  G  R  G  L  R  G  D  G  T  L  D  P  H  H  A  A  I  L  F  R  D  S  R  G  L  P  V  A  D  P  F  L   154

               501  GAGAAAGTAA ATCTATCAGA TCTGG
                     E  K  V  N  L  S  D  L  E                                                                          163
```


B.

```
SNF4_cDNA_5'   434  GGATCCA CACCATGCAG CCATCCTCTT CAGAGACTCA CGAGGG---- ---------T TGCCTGTCGC  487
Dros_Mdg3_3'     1  ....... ......... ......... ......... ......GTAA GTTTAAAAC. ...AGC.A..   67

               488  TGATCCGTTC CTAGAGAAAG TAAATCTATC AGATCTGG  525
                68  GT..AAA..A GC....G... ..TC..ATAA .TTGTCTC  105
```

Figure 3.3. Sequence of the 5' end of the *Drosophila SNF4* cDNA. (A) The nucleotide and inferred protein sequences. Only shown is the region obtained by sequencing the *Drosophila* EST. The sequence of the full length cDNA continues in Figure 3.2 from position 1537. The DNA sequence is numbered to the left and the amino acid to the right of each line. The region showing 100 % identity with the *Mdg3* transposable element (see Panel B) is underlined. (B) An alignment of the partial *Drosophila* cDNA with the 3' flanking sequence of *Mdg3*, a *Drosophila* transposable element. Only the sequence from position 434 onwards of the *SNF4* clone from panel (A) is shown. The GT intron excision motif is shown in bold. An additional 232bp of *Mdg3* sequence is available in the 3' direction, but is not shown. Sequence identity is shown by ".", gaps are indicated by dashes.

Furthermore, when these two sequences are compared (Figure 3.3B), an alignment can only be made between position 434 and 476 of the SNF4 cDNA sequence. This leaves a 49 bp section of the *SNF4* cDNA sequence (position 477 to 505) which does not align with either the *Hin*dIII genomic clone nor the 295bp of *Mdg3* genomic sequence available extending in the 3' direction. There must be an intron between position 476 and 477 of the cDNA therefore, which is at least 295 bp long. Indeed, the *Mdg3* sequence at this site contains a GT motif (position 44 - 45), which does conform to the upstream splice site consensus sequence (Figure 3.3B). Thus, the *Drosophila SNF4* gene has at least 5 introns, with the possibility of others within the first 433 bases of the cDNA sequence. Due to the fact that only cDNA clones have been isolated from the other metazoans, the genomic organization of the *Drosophila SNF4* gene cannot be compared with other eukaryotes.

The full length 1944 bp cDNA of the fly *SNF4* codes for a protein of 648 amino acids with a predicted molecular weight of 71,600 Da. The GC content of the coding sequence is slightly biased at 55.5 %. Due to the lack of genomic sequence upstream of the first exon, the translational start site has been assigned to the first in frame methionine in the cDNA sequence (Figure 3.3). As most mRNAs begin translation at the first ATG in the transcript (Kozak, 1983), and assuming that the cDNA obtained is of full length, this assignment is not unreasonable. The downstream AATAAA polyadenylation site appears to overlap with the stop codon at position 3275 of the 3.5 kb *Hin*dIII fragment. Within the ~100 bp between this signal and the polyA addition site at position 3380 (Figure 3.2), there are several A_n and CCTCC motifs which are also characteristic of eukaryotic 3' untranslated sequences (Nussinov, 1986).

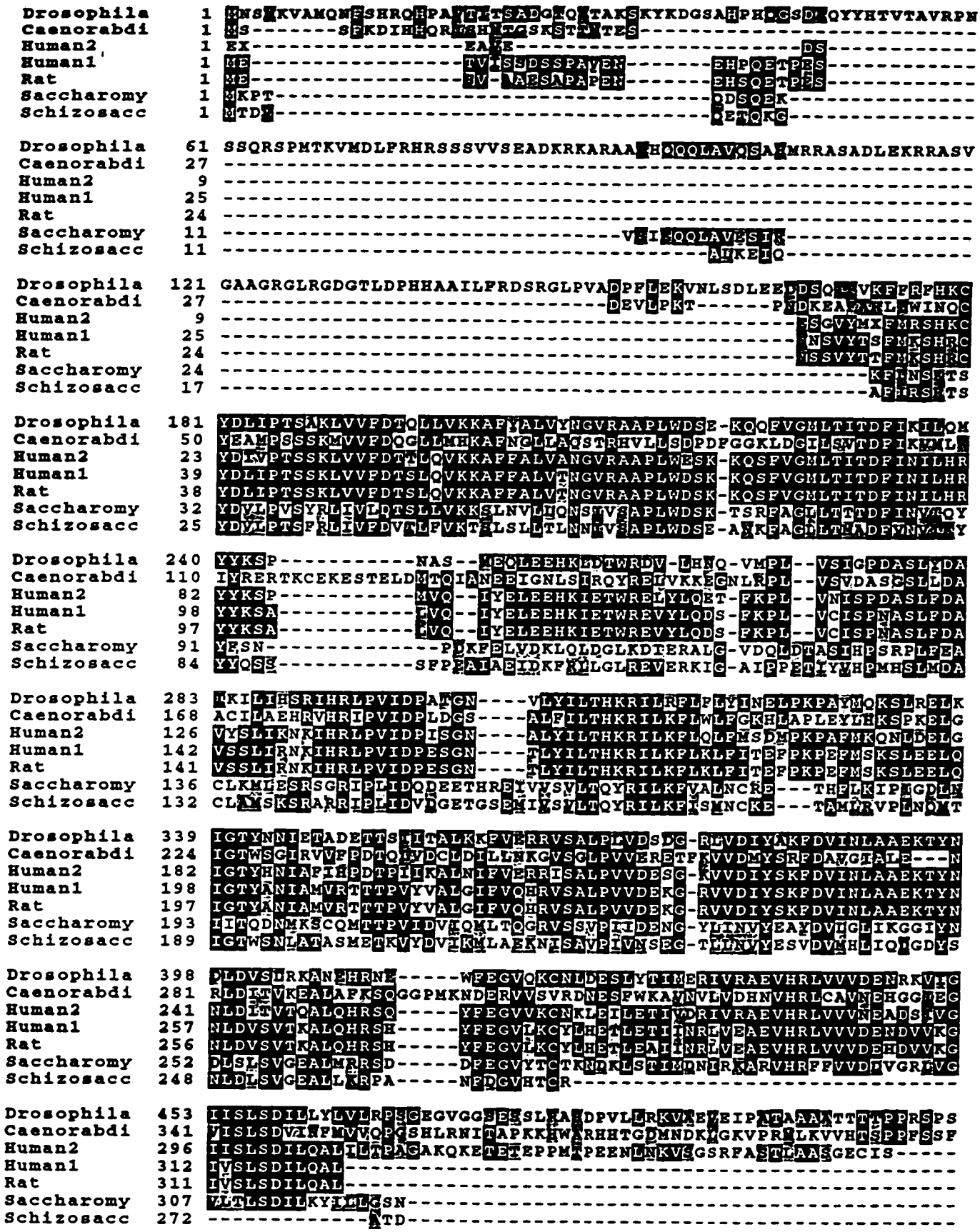
Northern analysis was used to quantify mRNA levels. An initial Northern was prepared from total RNA representing *Drosophila* larvae and adults which had been raised on both glucose and nonglucose media. When this blot was probed under high stringency with the 1.1 kb E7/E8 primed *Drosophila* cDNA, no signal could be detected. A second Northern made from polyA$^+$ RNA rather than total RNA was also probed with this 1.1 kb fragment.

After a 5 day exposure, no signal was detectable for this blot as well. The integrity of the membrane-bound RNA was tested by reprobing the mRNA Northern with a 750 bp *Drosophila* alpha-amylase probe. Using the same high stringency conditions as were used for the *SNF4* probe, a band of the correct size was visible after a 1.5 hour exposure.

3.3.2 *Comparison of* SNF4 *from* Drosophila *with other homologous sequences*

The *SNF4* protein is much longer than any other known homologues for this gene. At 648 amino acids, it is nearly twice the size of the rat (331 aa) and yeast (323 aa) sequences. When the *Drosophila* sequence is aligned against the other eukaryotic homologues, it can be seen that the added length is found at both the carboxy and amino termini (Figure 3.4). *C. elegans* also has a longer carboxy terminus than the mammalian and yeast homologues, but it is still considerably shorter than the *Drosophila* copy. When the complete *Drosophila* amino acid sequence is compared with the other eukaryotic homologues, it is 31 % identical to the rat and human sequences, 22 % to *Caenorabditis*, 18 % to *Saccharomyces*, and 13 % to *Schizosaccharomyces*. However when the long carboxy and amino terminal *Drosophila* stretches are excluded and just 305 aa spanning the inner core sequence is compared, these identity values rise to 64 % for mammals, 25 % for *C. elegans*, 35 % for *Saccharomyces*, and 28% for *S. pombe*.

When the GenBank expressed sequence tag database was searched, the *Drosophila SNF4* cDNA matched dozens of human and mouse ESTs. The human sequences were sorted into different types based on their similarity to the single full length human cDNA which has been published (accession U42412). Thirty four of the human ESTs were identical to the published cDNA. Of the 30 which were not, 27 were overlapping clones which could be aligned to form a second full length consensus sequence which is named Human2. A further 3 clones showed numerous differences with both the published and composite human *SNF4* sequences. They were accession numbers AA434294, AA178898, and M78939. An alignment of the DNA sequences of these 5 human SNF4 isotypes is presented in Figure 3.5.


Drosophila
Caenorabdi
Human2
Human1
Rat
Saccharomy
Schizosacc

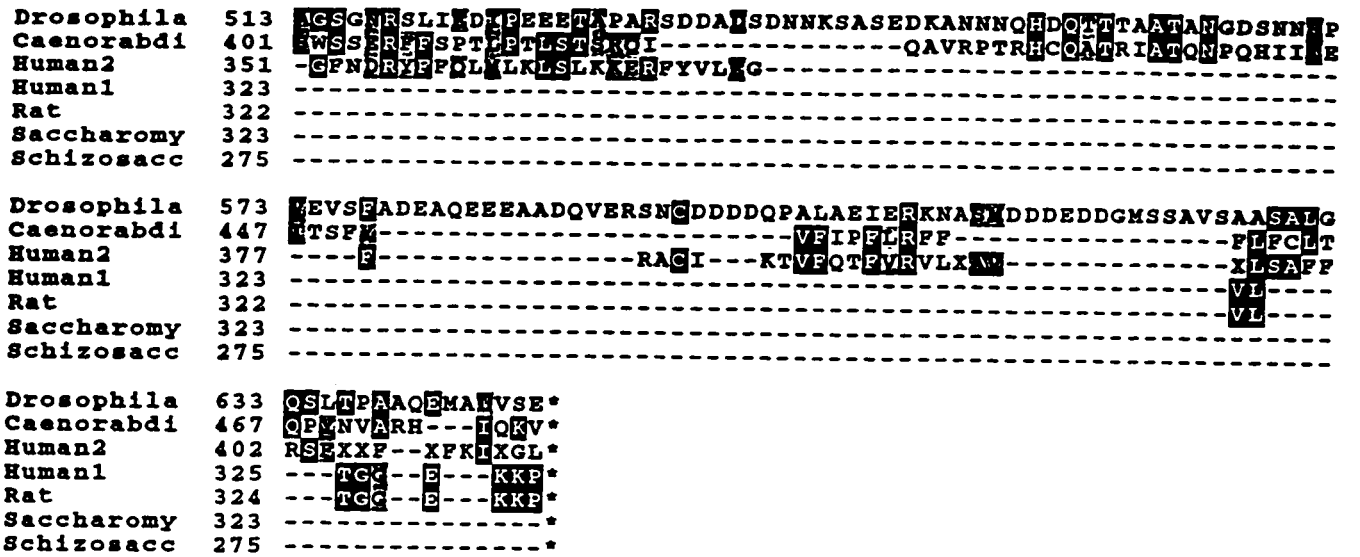
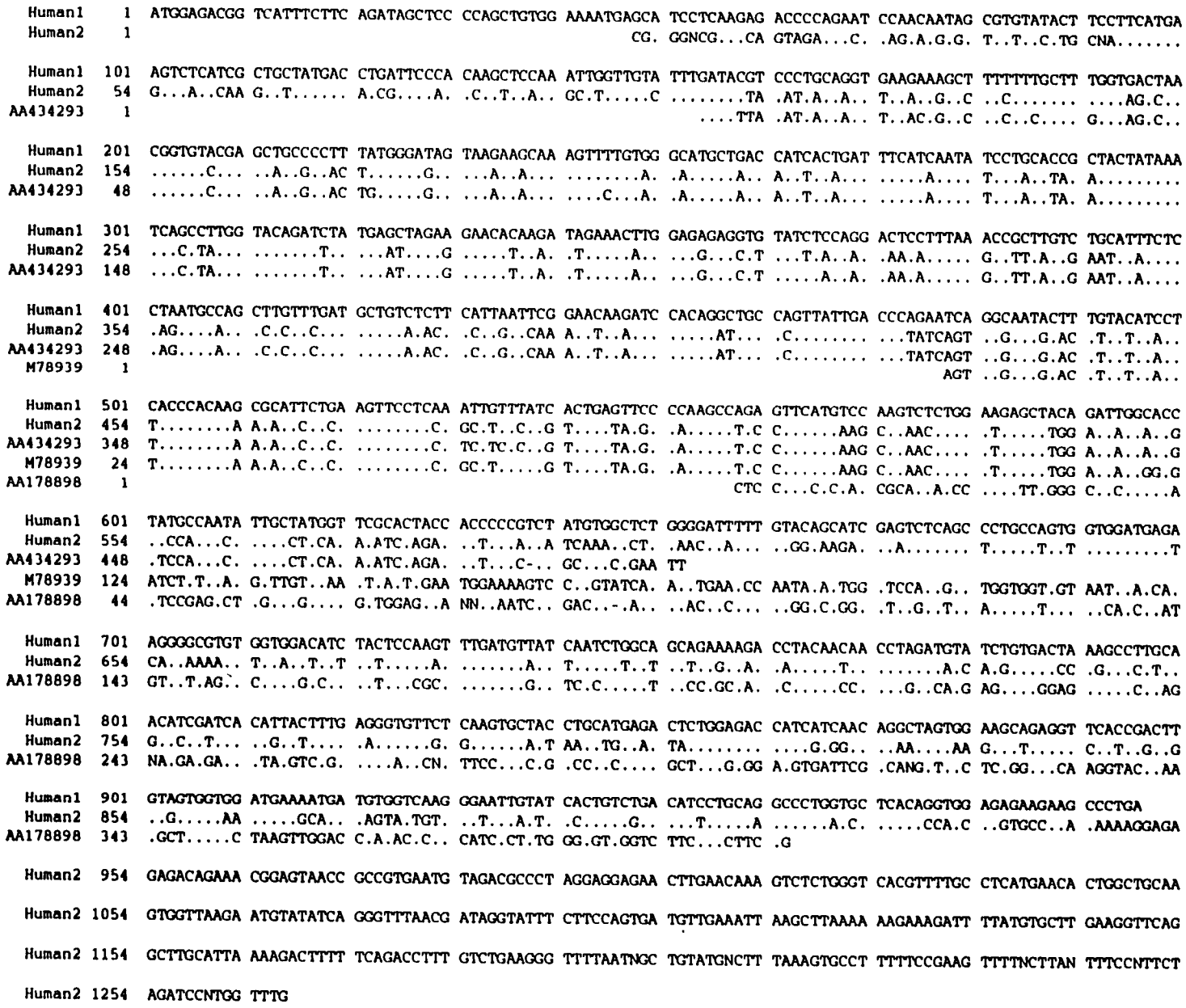


Figure 3.4. Alignment of the *Drosophila SNF4* protein sequence with other animal and yeast homologues. The junction of the sequence derived from the cDNA and genomic clones is at position 184-185 of the *Drosophila* sequence. The Human2 sequence is a composite of several EST sequences in GenBank, but the remaining sequences are published *SNF4* homologues. The accession number for the *C. elegans* homologue is U97550, see sections 3.2.1 and 3.2.5 for the accession numbers of the other sequences used. Amino acid identities are shown boxed in black, and similarities are shown in grey. Gaps in the alignment are shown by the dashes.



```
  Human1     1  ATGGAGACGG TCATTTCTTC AGATAGCTCC CCAGCTGTGG AAAATGAGCA TCCTCAAGAG ACCCCAGAAT CCAACAATAG CGTGTATACT TCCTTCATGA
  Human2     1                                                    CG. GGNCG...CA GTAGA...C. .AG.A.G.G. T..T..C.TG CNA.......

  Human1   101  AGTCTCATCG CTGCTATGAC CTGATTCCCA CAAGCTCCAA ATTGGTTGTA TTTGATACGT CCCTGCAGGT GAAGAAAGCT TTTTTTGCTT TGGTGACTAA
  Human2    54  G...A..CAA G..T..... A.CG....A. .C..T..A.. GC.T.....C ........TA .AT.A..A.. T..A..G..C ..C..... ....AG.C..
AA434293     1                                                        ....TTA .AT.A..A.. T..AC.G..C ..C..C.... G...AG.C..

  Human1   201  CGGTGTACGA GCTGCCCCTT TATGGGATAG TAAGAAGCAA AGTTTTGTGG GCATGCTGAC CATCACTGAT TTCATCAATA TCCTGCACCG CTACTATAAA
  Human2   154  ......C... ..A..G..AC T......G.. ...A..A... ........A. .A.....A.. A..T..A... .....A.... T...A..TA. A.........
AA434293    48  ......C... ..A..G..AC TG.....G.. ...A..A... ....C...A. .A.....A.. A..T..A... .....A.... T...A..TA. A.........

  Human1   301  TCAGCCTTGG TACAGATCTA TGAGCTAGAA GAACACAAGA TAGAAACTTG GAGAGAGGTG TATCTCCAGG ACTCCTTTAA ACCGCTTGTC TGCATTTCTC
  Human2   254  ...C.TA... .......T.. ...AT....G .....T..A. .T.....A.. ...G...C.T ...T.A..A. .AA.A..... G..TT.A..G AAT..A....
AA434293   148  ...C.TA... .......T.. ...AT....G .....T..A. .T.....A.. ...G...C.T .....A..A. .AA.A..... G..TT.A..G AAT..A....

  Human1   401  CTAATGCCAG CTTGTTTGAT GCTGTCTCTT CATTAATTCG GAACAAGATC CACAGGCTGC CAGTTATTGA CCCAGAATCA GGCAATACTT TGTACATCCT
  Human2   354  .AG....A.. .C.C..C... .....A.AC. .C..G..CAA A..T..A... .....AT... .C........ ...TATCAGT ..G...G.AC .T..T..A..
AA434293   248  .AG....A.. .C.C..C... .....A.AC. .C..G..CAA A..T..A... .....AT... .C........ ...TATCAGT ..G...G.AC .T..T..A..
  M78939     1                                                                                      AGT ..G...G.AC .T..T..A..

  Human1   501  CACCCACAAG CGCATTCTGA AGTTCCTCAA ATTGTTTATC ACTGAGTTCC CCAAGCCAGA GTTCATGTCC AAGTCTCTGG AAGAGCTACA GATTGGCACC
  Human2   454  T........A A.A..C..C. ........C. GC.T..C..G T....TA.G. .A.....T.C C......AAG C..AAC.... .T.....TGG A..A..A..G
AA434293   348  T........A A.A..C..C. ........C. TC.TC.C..G T....TA.G. .A.....T.C C......AAG C..AAC.... .T.....TGG A..A..A..G
  M78939    24  T........A A.A..C..C. ........C. GC.T.....G T....TA.G. .A.....T.C C......AAG C..AAC.... .T.....TGG A..A..GG.G
AA178898     1                                                        CTC C...C.C.A. CGCA..A.CC ....TT.GGG C..C.....A

  Human1   601  TATGCCAATA TTGCTATGGT TCGCACTACC ACCCCCGTCT ATGTGGCTCT GGGGATTTTT GTACAGCATC GAGTCTCAGC CCTGCCAGTG GTGGATGAGA
  Human2   554  ..CCA...C. ....CT.CA. A.ATC.AGA. ..T...A..A TCAAA..CT. .AAC..A... ..GG.AAGA. ..A..... T.....T..T .........T
AA434293   448  .TCCA...C. ....CT.CA. A.ATC.AGA. ..T...C-.. GC...C.GAA TT
  M78939   124  ATCT.T..A. G.TTGT..AA .T.A.T.GAA TGGAAAAGTC C..GTATCA. A..TGAA.CC AATA.A.TGG .TCCA..G.. TGGTGGT.GT AAT..A.CA.
AA178898    44  .TCCGAG.CT .G...G.... G.TGGAG..A NN..AATC.. GAC..-.A.. ..AC..C... ..GG.C.GG. .T..G..T.. A.....T... ..CA.C..AT

  Human1   701  AGGGGCGTGT GGTGGACATC TACTCCAAGT TTGATGTTAT CAATCTGGCA GCAGAAAAGA CCTACAACAA CCTAGATGTA TCTGTGACTA AAGCCTTGCA
  Human2   654  CA..AAAA.. T..A..T..T ..T.....A. .......A.. T.....T..T ..T..G..A. .A.....T.. .......A.C A.G.....CC .G...C.T..
AA178898   143  GT..T.AG.. C....G.C.. ..T...CGC. .......G.. TC.C.....T ..CC.GC.A. .C.....CC. ...G..CA.G AG....GGAG .....C..AG

  Human1   801  ACATCGATCA CATTACTTTG AGGGTGTTCT CAAGTGCTAC CTGCATGAGA CTCTGGAGAC CATCATCAAC AGGCTAGTGG AAGCAGAGGT TCACCGACTT
  Human2   754  G..C..T... ..G..T.... .A......G. G......A.T AA..TG..A. TA........ ....G.GG.. ..AA....AA G...T..... C..T..G..G
AA178898   243  NA.GA.GA.. .TA.GTC.G. ....A..CN. TTCC...C.G .CC..C.... GCT...G.GG A.GTGATTCG .CANG.T..C TC.GG...CA AGGTAC..AA

  Human1   901  GTAGTGGTGG ATGAAAATGA TGTGGTCAAG GGAATTGTAT CACTGTCTGA CATCCTGCAG GCCCTGGTGC TCACAGGTGG AGAGAAGAAG CCCTGA
  Human2   854  ..G.....AA .....GCA.. .AGTA.TGT. ..T...A.T. .C.....G.. ...T.....A ......A.C. .....CCA.C ..GTGCC..A .AAAAGGAGA
AA178898   343  .GCT.....C TAAGTTGGAC C.A.AC.C.. CATC.CT.TG GG.GT.GGTC TTC...CTTC .G

  Human2   954  GAGACAGAAA CGGAGTAACC GCCGTGAATG TAGACGCCCT AGGAGGAGAA CTTGAACAAA GTCTCTGGGT CACGTTTTGC CTCATGAACA CTGGCTGCAA

  Human2  1054  GTGGTTAAGA ATGTATATCA GGGTTTAACG ATAGGTATTT CTTCCAGTGA TGTTGAAATT AAGCTTAAAA AAGAAAGATT TTATGTGCTT GAAGGTTCAG

  Human2  1154  GCTTGCATTA AAAGACTTTT TCAGACCTTT GTCTGAAGGG TTTTAATNGC TGTATGNCTT TAAAGTGCCT TTTTCCGAAG TTTTNCTTAN TTTCCNTTCT

  Human2  1254  AGATCCNTGG TTTG
```

Figure 3.5. DNA alignment of the family of human SNF4 sequences. Human 1 is the published cDNA sequence (Gao *et al.*, 1996), Human2 is the assembled full length sequence. AA434293, M78939, and AA178898 are all partial length EST sequences which show homology to SNF4 but are different from the Human1 and Human2 sequences. Sequence identity is shown by ".", gaps by "-".

Most of the mismatches occur at intervals of three nucleotides, which corresponds to the synonymous third codon position. In addition, there are many mismatches which result in amino acid substitutions. As a result, it is highly unlikely that the observed mismatches are due to sequencing errors.

The same technique was used to examine the population of mouse ESTs which showed homology with the *Drosophila SNF4*. A full length published sequence of the mouse EST does not exist, although a rat sequence is available (accession X95578). Of the 19 mouse ESTs, 15 were aligned to form a single consensus sequence. Two others (AA204077 and AA473880) are possible pseudogenes since they were identical to the consensus sequence for the first 18 amino acids but were followed by in frame stop codons. In addition there were a further two clones, AA107296 and AA571379, which featured several substitutions from the mouse consensus sequence which could not be explained by sequencing errors alone.

The presence of multiple isoforms of mammalian *SNF4* cDNAs suggests that *SNF4* is a member of a multigene family. Although only one gene copy has been found in yeast, the population of mammalian ESTs opens the possibility that there may be at least three distinct *SNF4* copies in the mammalian genome. To determine gene copy number in *Drosophila*, the 1.1 kb *Drosophila SNF4* cDNA was used to probe a single genomic Southern under high stringency (Figure 3.6A). Based on the restriction map (Figure 3.1), the expected sizes for the given digests are: 3.5 kb (*Hin*dIII), 1.9 kb (*Hin*dIII/*Pst*I), and 1.1 kb (*Bam*HI/*Pvu*II). The major band for each digest in Figure 3.5A is concordant with the expected size. Additional high intensity bands of 3.1 and ~7 kb were seen in the *Bam*HI/*Pvu*II digest on this 16 hour exposure. These are likely to be the result of incomplete digestion, since the other two digests only showed one major band each. This Southern membrane was stripped and reprobed with the human 684 bp *SNF4* probe, using the same low stringency conditions as were used to isolate the original lambda clones. After a 4 day exposure, multiple bands could be seen for each digest (Figure 3.6B): in the *Hin*dIII digest, a doublet at 3.5 kb; in the *Hin*dIII/*Pst*I digest, bands at 1.9 and 3.8 kb; and for the *Bam*HI/*Pvu*II digest, medium intensity bands at 1.0, 2.3,

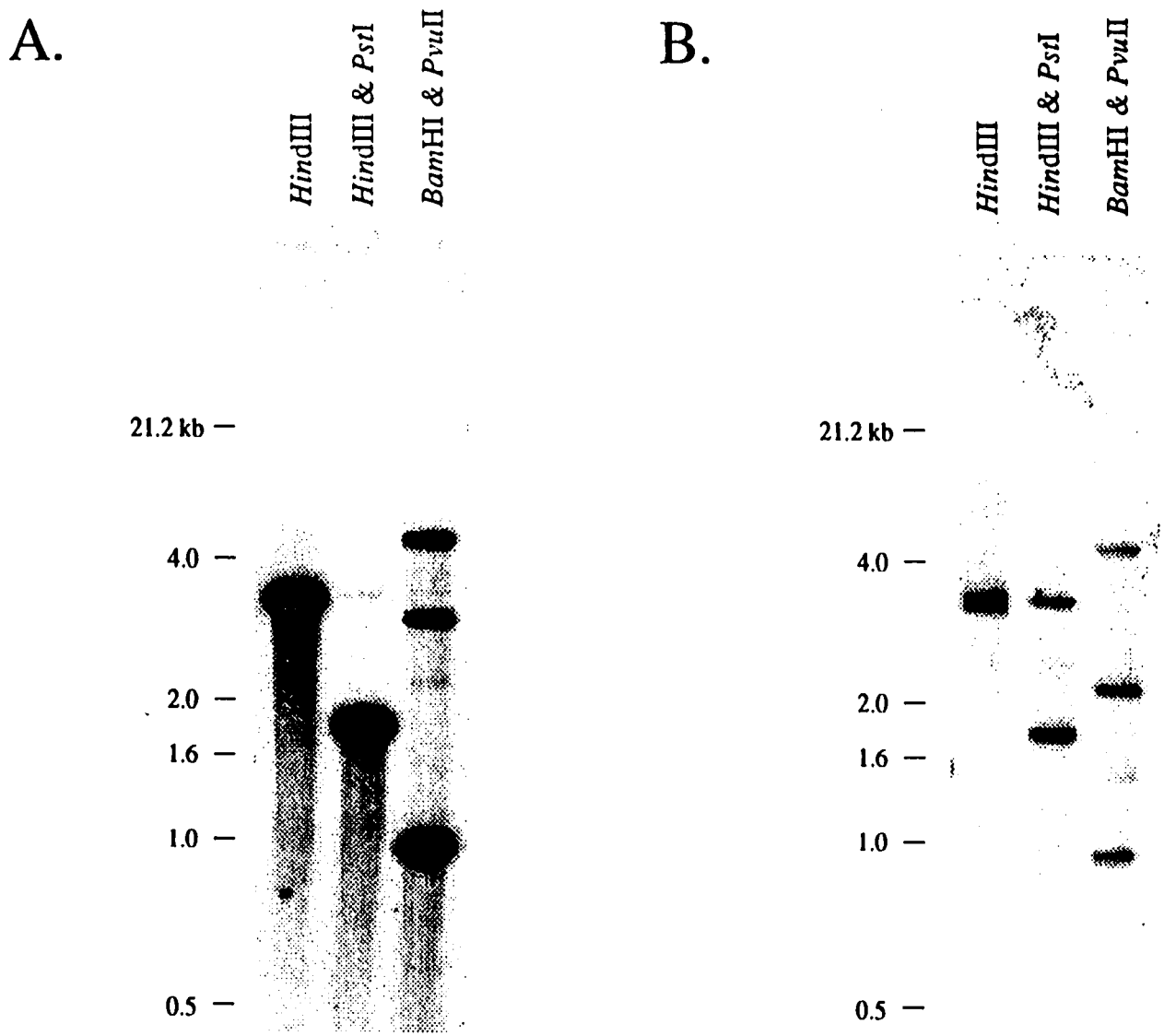


Figure 3.6. Southern analysis of *Drosophila* genomic DNA using *SNF4* probes. (A) Genomic digest probed under high stringency with a 1.1 kb *Drosophila SNF4* cDNA. (B) Genomic digest reprobed under low stringency with a 684 bp human *SNF4* cDNA. See section 3.2.4 for hybridization conditions. Both Southern analyses were performed a single time only.

and 7 kb and low intensity bands at 1.6 and 3.5 kb. The major bands observed in the high stringency Southern (Figure 3.6A) are all seen in the low stringency Southern (Figure 3.6B). Several of the less intense bands are also seen in both Southern profiles - such as the 3.8 kb fragment in the *Hin*dIII/*Pst*I digest, as well as the 7.0 and 2.3 kb bands in the *Hin*dIII/*Pst*I digest. In a shorter exposure of the high stringency Southern however, these fainter bands may not be visible however. The doublet bands observed in the low stringency probing of the *Hin*dIII lane may also be present under high stringency conditions, as the intensity of the single high stringency band precludes such a determination. A shorter exposure for the high stringency probing is unavailable however, as both Southerns represent a single replicate. Due to the presence of bands resulting from incomplete DNA restriction (as seen in the *Bam*HI/*Pvu*II digest of Figure 3.6A), it is possible that the multiple bands observed in the low stringency blot also represent the products of incomplete digestion. However, as there are bands attributed to incomplete digestion which are not seen in both blots (*i.e.* the 3.5 kb fragment in the *Bam*HI/*Pvu*II digest), the multiple bands observed under low stringency cannot all be the result of this problem. A replicate of these Southerns could both resolve the presence of the doublet banding patterns of the *Hin*dIII digest, as well as eliminate these restriction complications definitively. One possible interpretation of the combined of data from the two available Southerns is that *SNF4* is present as two gene copies that are relatively divergent, due to the small degree of probe cross reactivity under high stringency conditions (Figure 3.6A).

An analysis of the different mammalian members of the *SNF4* gene family also agrees with this interpretation. The only taxa for which multiple full length cDNAs are known is humans. Although the published human cDNA codes for a short peptide (331 aa), the Human2 sequence, like the *C. elegans* and *Drosophila* sequences, has a 3' tail that is approximately 300 bp longer than the other human sequence. When the two full length human protein sequences were compared with the *Drosophila* sequence, the longer Human2 protein showed a slightly higher level of identity with the fly sequence (33 % vs. 31 %), although they are more similar to each other (55 %) than either is to the *Drosophila* sequence. The protein

alignment of Human2 with the other published sequences is included in Figure 3.4, and the DNA alignment with the Human1 gene is shown in Figure 3.5.

3.3.3 *Phylogenetic analysis*

The alignment shown in Figure 3.4 was used to generate amino acid-derived phylogenetic trees for *SNF4*, as shown in Figure 3.7. The tree inferred using the parsimony algorithm on full length sequences is presented in Figure 3.7A. The taxa formed three clusters - *Drosophila/Caenorabditis*/Human2, the published mammals, and the yeasts. On a purely quantitative level however, the *Drosophila* sequence is more similar to the mammalian sequences than *C. elegans* (31 % vs. 22 %, see previous section), but both the fly and worm cDNAs had extended 3' sequences. To test the effect on phylogenetic reconstruction of the long 5' and 3' regions of the *Drosophila* gene which are not found in the published mammalian and yeast homologues, a truncated alignment of 305 amino acids (from position 167 to 471 out of the 648 aa in the *Drosophila* sequence) spanning the length of the shorter proteins was used. The topology of this phenogram (Figure 3.7B) differs from that generated from full length sequences by the nesting of the Human2 and *Drosophila* sequences with the short mammalian homologues, rather than grouping with *C. elegans*. However, the Human2 sequence is shown as being clearly distinct from the Human1 sequence, as the Human1 sequence grouped with the Rat sequence, rather than with Human2, the second human gene copy.

The last tree extends the phylogenetic inference beyond the *SNF4* family. When the *Drosophila* amino acid sequence is used to search the GenBank protein database, only a small number of sequences show significant similarity. These include hypothetical proteins identified from the *Methanococcus* genome project (accession numbers G64453 and H64452), a *C. elegans* IMP dehydrogenase-like protein (U67950), and numerous IMP dehydrogenases (IMPDH). When these proteins are aligned (Figure 3.8), they do not show specific areas of high sequence conservation, rather, there is low level similarity across the length of the protein. These non-*SNF4* sequences were added to a subset of the taxa used in the previous gene

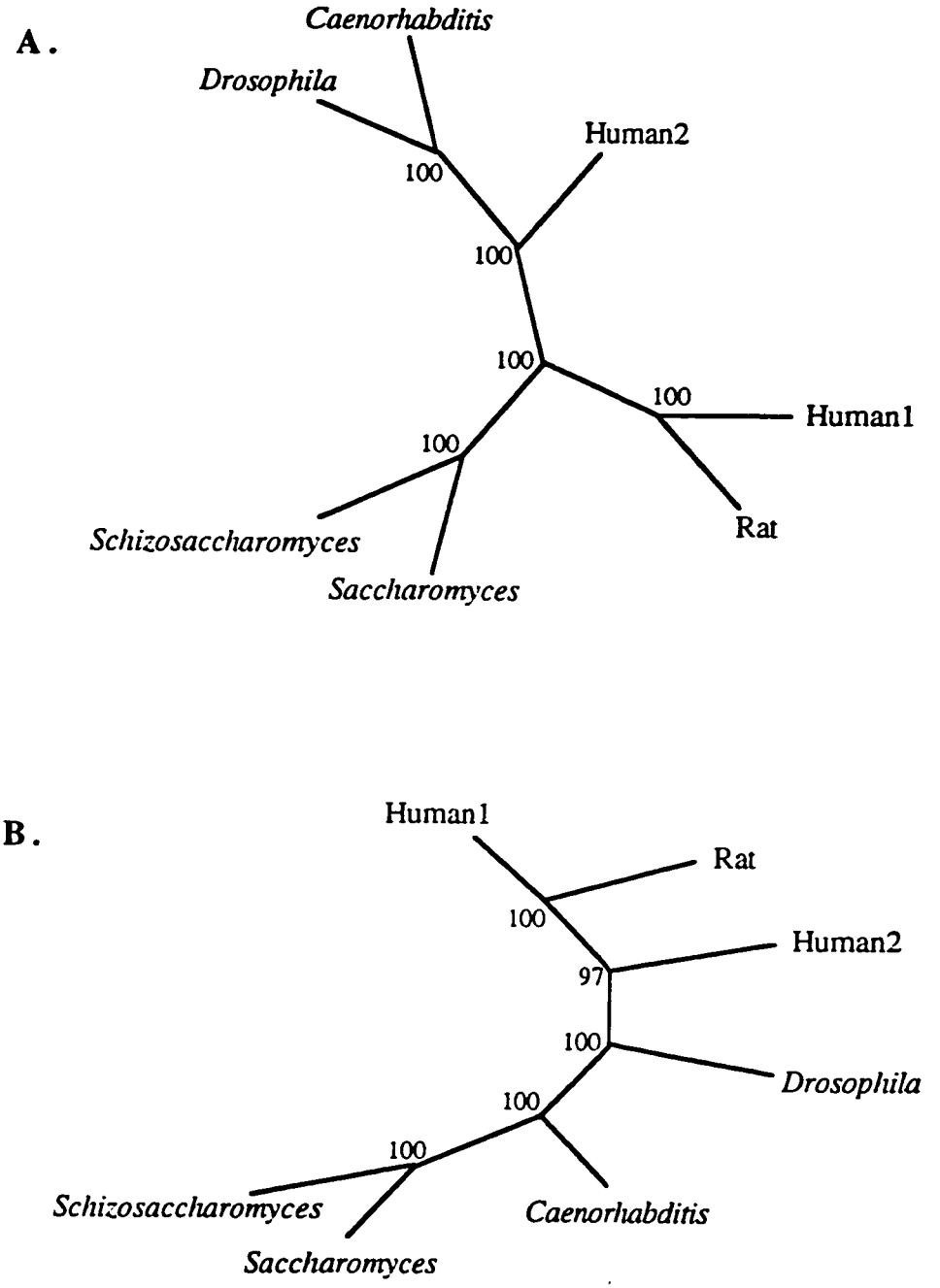


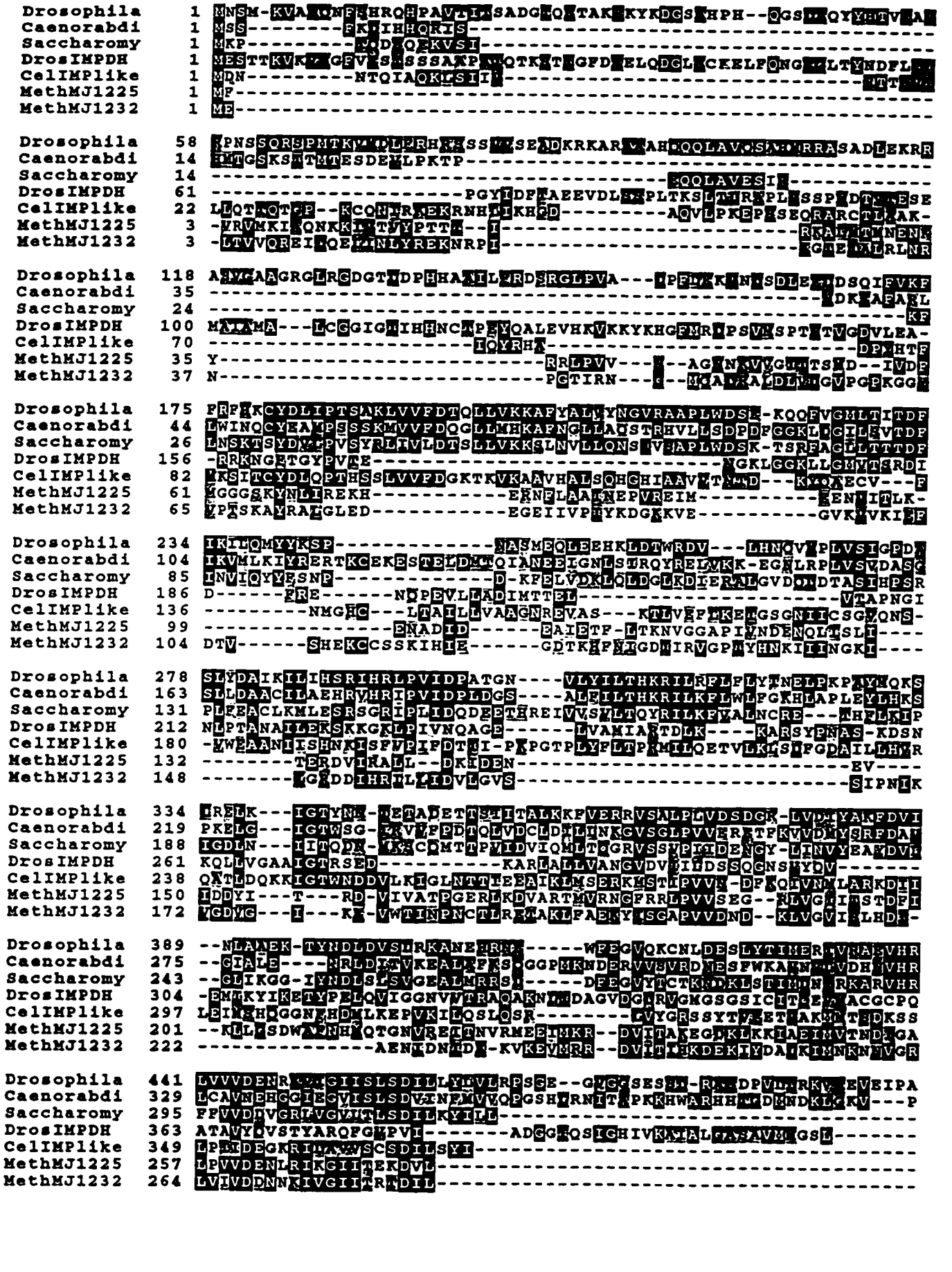
Figure 3.7. Unrooted protein phylogenies of SNF4 generated using parsimony. (A) Inferred phenogram when the full lengths of all sequences are used. (B) Tree generated using only the core 305 aa of sequence conserved between all SNF4 genes. The numbers at the nodes are the bootstrap percentages.

family tree to produce the phylogeny in Figure 3.9. To remove effects associated with the long 5' and 3' ends of the *Drosophila SNF4* clone (see Figure 3.7), the regions not common to all of the sequences were trimmed, resulting in the 305 aa alignment that was used. With the addition of more sequences the bootstrap certainties decrease slightly, but several trends can be observed in this figure. The IMP dehydrogenases form one cluster with the two unknown *Methanococcus* sequences. Therefore, they form a distinct family from the *SNF4* sequences. Nested with the IMPDH group is the *C. elegans* IMPDH-like sequence, which is more similar to the IMPDH sequences than the *SNF4* sequences. Last, the remaining *SNF4* sequences branch together with the *Caenorhabditis SNF4* sequence. The *Drosophila*/Human1 sister group for *SNF4* branches off further away from the *C. elegans* homologue than the *Saccharomyces* sequence, unlike in the previous phylogenies presented (Figure 3.7). When calculated over the short 305 aa alignment, the fly sequence is more similar to the *Saccharomyces* homologue than to the worm one (35 % vs. 25 %). When only these three taxa are included in the analysis therefore, rather than the fly and *C. elegans* sequences pairing, the fly and yeast sequences form a sister group as seen in Figure 3.9. However, the addition of the *S. pombe* sequence changes the topology of the tree. Although the two yeast sequences share many similarities, the *Schizosaccharomyces* sequence is quite divergent from that of *Drosophila*. This causes the yeast group to branch further away from the *Drosophila* sequence. This effect is seen in Figure 3.7, and a subsequent reanalysis of the IMPDH-containing phylogeny confirmed this outcome (data not shown).

3.4 Discussion

3.4.1 *Analysis of the different approaches used to obtain the genomic and cDNA clones*

The *SNF4* proteins for which sequence data were initially available could be easily aligned, even though they were from yeast and mammals. Since short regions of high sequence similarity were identified from this alignment, it was initially thought that the



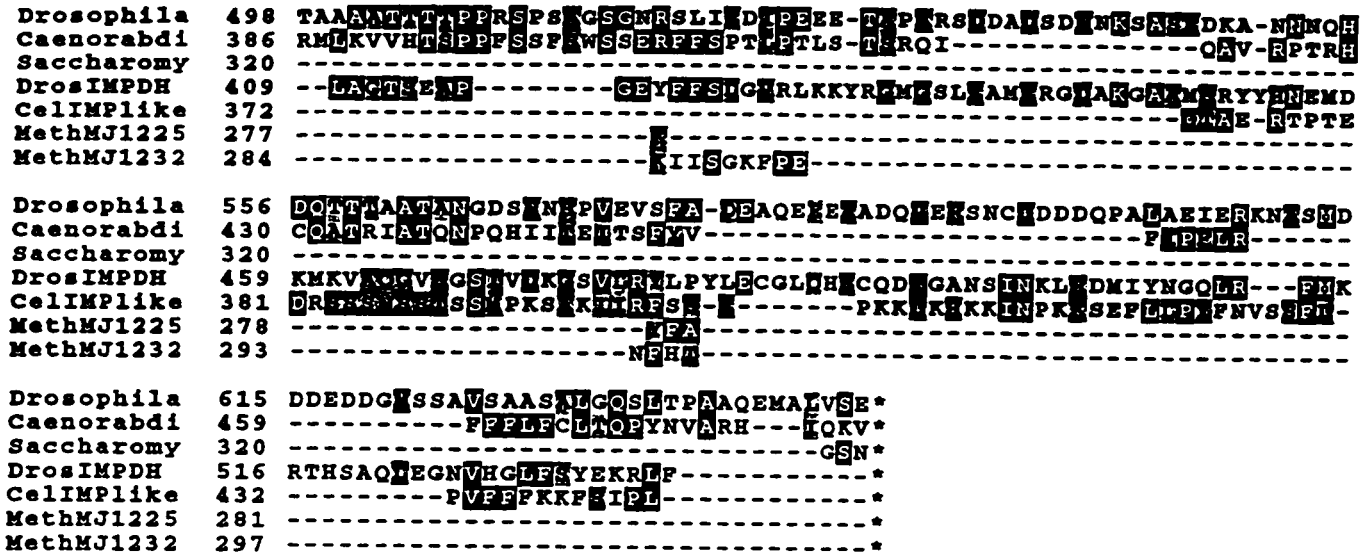


Figure 3.8. Alignment of the *Drosophila SNF4* protein sequence with other related proteins. The first three sequences are the *Drosophila*, *Caenorhabditis* and *Saccharomyces SNF4* homologues, respectively. They are followed by *Drosophila* IMP dehydrogenase, a *C. elegans* IMP dehydrogenase-like protein, and two hypothetical proteins identified in the *Methanococcus* genome identified by their protein numbers. Amino acid identities are shown boxed in black, and similarities are shown in grey. Gaps in the alignment are shown by "-".

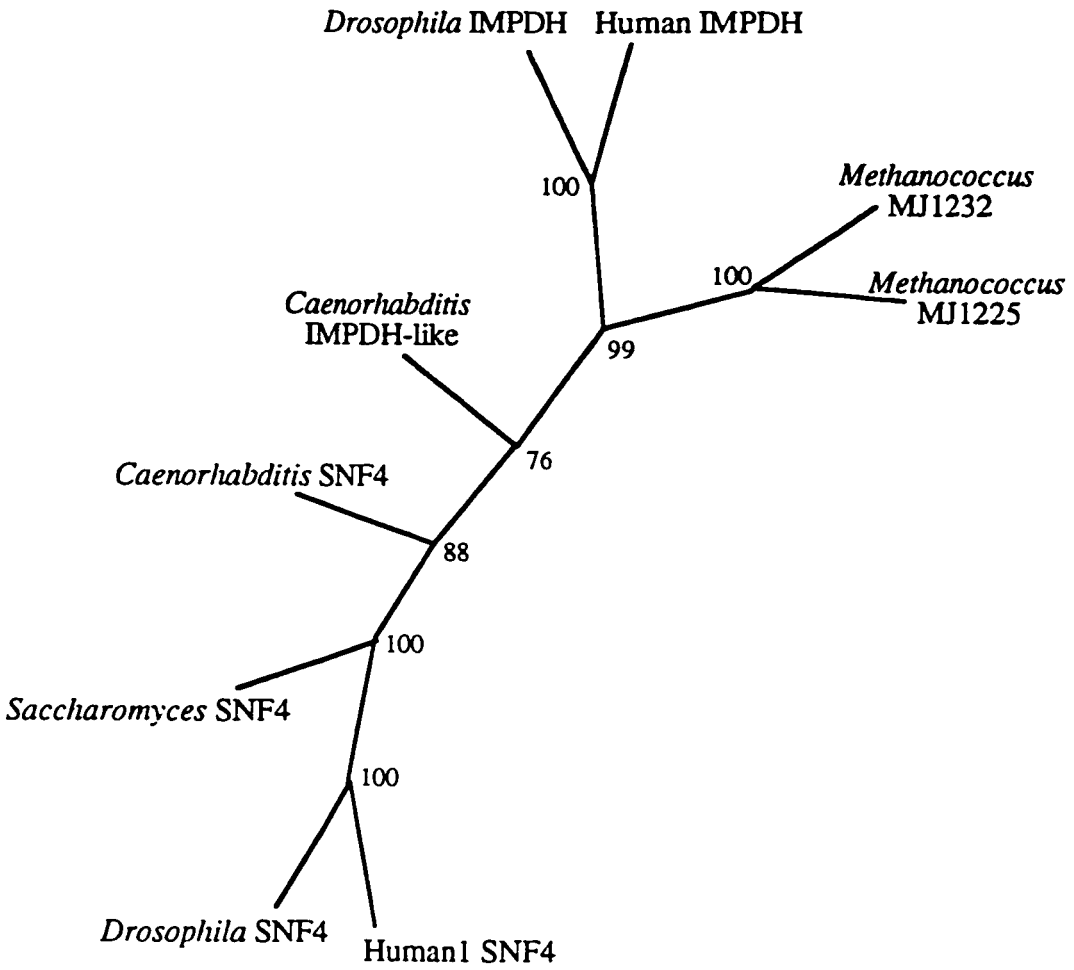


Figure 3.9. Phylogeny of SNF4 and related sequences obtained using parsimony. A truncated alignment of 305 amino acids was used instead of the full length sequences. The numbers at the tree nodes are the bootstrap percentages.

Drosophila homologue could be obtained by PCR using degenerate primers designed for these regions. Exhaustive combinations of primers and amplification conditions were tried using the 5 primers listed in Table 3.1, but the *Drosophila* SNF4 homologue was not successfully amplified.

The sequence for this gene was eventually obtained through a combination of traditional heterologous probing of a genomic library, and the sequencing of a *Drosophila* expressed sequence tag. Using this sequence data, the actual specificity of the degenerate primers can be determined and are shown below:

Table 3.2. Comparison of the degenerate primers synthesized for *SNF4* amplification and the actual sequence of the primer binding site

Primer Name	Amino Acid Sequence of primer	Actual Target Sequence in *Drosophila*	No. of nucleotide mismatches per primer length
K1188F	DTSLQVK	DTQLLVK	4/20
K1140F	TITDFIN	TITDFIK	0/20
K1214F	ALPVVD	ALPLVD	1/17
K1189R	PVVDEKG	PLVDSDG	6/20
K1141R	FEGVLKC	FEGVQKC	1/20

Only one of the degenerate primers, K1140F, would have contained oligonucleotides in its population with exact matches to its target sequence. The remainder had at least one mismatch, often near the 3' end. A single mismatch at this critical end can decrease the product yield by as much as 100 fold (Kwok *et al.*, 1990). Such primer-DNA complex destabilizations would also decrease the specificity of the PCR reactions and increase the probability of spurious primer binding to nontarget sites.

Unrelated problems were also encountered with cloning the putative full length cDNA copy of *SNF4* using a 100% specific probe. When a 1.1 kb probe that was PCR amplified from a *Drosophila* cDNA library was used to screen the same library, no positives were obtained. A total of 200,000 plaques were screened from two different libraries with no success. In a related observation, *SNF4*-specific cDNA fragments could be PCR amplified using the *Drosophila* cDNA library as a template, but only when the reactions were seeded with at least 1 x 10^7 pfu and not the 1 x 10^5 pfu and 5 x 10^5 pfu template also tested. *In vivo* amplification of lambda libraries can result in the overpopulation of certain inserts at the expense of others, so the problems associated with these libraries could be an artifact of the amplification process. The result of the Northern analysis indicates that this is not the case. A gel containing 3 μg of polyA$^+$ mRNA per lane should detect messages whose abundancy is greater than 1/500,000 (Brown, 1993). Due to limited sample availability, the amount of RNA loaded could not be quantified precisely, but 1-2 μg of mRNA was loaded per lane. Even after a several day exposure, no band was visible when the membrane was probed with the 1.1 kb *Drosophila SNF4* cDNA probe. This finding was unexpected, as the cDNA sequence was derived from an EST containing a polyA tail, therefore its mRNA would be present on a polyA$^+$ mRNA Northern. When the Northern blot was probed with *Drosophila* α-amylase as a positive control, the RNA showed slight degradation but a ~1.7 kb band could be clearly detected within a 1.5 hour exposure. Although α-amylase is one of the more abundant messages, the positive control test demonstrated that rare messages should also be visible after a longer exposure. The SNF4 transcript appears to be present in every 1 x 10^6 to 1 x 10^7 transcripts, therefore more than 6 μg of polyA$^+$ mRNA would need to be loaded per well to visualize a band using Northern analysis. A third Northern blot containing a higher concentration of polyA$^+$ mRNA was not performed. Confirmation of this low level abundance of the *Drosophila SNF4* mRNA have yet to be attempted, although this could be achieved by alternate methodology, including semi-quantitative PCR or ribonuclease protection assays.

The relative scarcity of the *Drosophila SNF4* mRNA is quite different from the expression profile seen in mammals. Tissue specific expression has been examined for rat (Woods *et al.*, 1996) and goat *SNF4* (Piosik *et al.*, 1996). The rat gene expressed very strongly in 6 different somatic tissues, and weakly in spleen and testis. The goat sequence, unmentioned up to now due to the overrepresentation of mammalian homologues, expressed strongly in several brain tissues, but in other somatic tissues and testis the expression was weak, although detectable in a total RNA Northern. Expression levels for the *Drosophila SNF1* protein kinase catalytic subunit are also undetectable on the same polyA$^+$ Northern blot (E. Taboada, per. comm.), which raises the possibility that several members of this protein complex may be coordinately subjected to low level expression in *Drosophila.*

3.4.2 *Diversity of* SNF4 *homologues*

One of the first characteristics of the *Drosophila SNF4* cDNA which can be observed even before any sequence comparisons are made, is its larger size relative to all other known *SNF4* homologues. The origin of the additional 5' and 3' terminus sequences is difficult to determine, since they show no homology at the protein and nucleotide levels to any GenBank sequences. The 3' tail codes for a higher level of acidic residues than the rest of the protein (21 % vs 10 %). Such acidic domains are often found in transcription factors where they can influence protein-protein interactions. It is unknown whether the *Drosophila* SNF4 acidic domain affects the interaction with its associated SNF1 and GAL83 homologues in a similar manner.

The *C.elegans* and Human2 sequences also have sequence extensions on the 3' end, but not the 5' end. When the protein translation of these 3' tail sequences are aligned (Figure 3.4), *Drosophila* and *Caenorhabditis* share certain sequence motifs more towards the stop codon, while the Human2 shares other motifs near the opposite end of the *Drosophila* tail. The degree of similarity between the *Drosophila* and Human2 tails certainly seems to indicate homology. This raises the possibility that a gene duplication occurred early in the metazoan

lineage, giving rise to at least two gene copies for *SNF4*. One of the gene copies codes for a short (~330 aa) protein as seen in the yeast and published mammalian sequences, while the other gene copy encodes a longer protein (~410 aa) which is extended in the 3' direction. This second form of the *SNF4* protein is seen in the *Drosophila*, *C. elegans*, and Human2 sequences. When the two human proteins are compared across the 331 amino acids conserved in both, they show 68 % sequence identity and 77 % sequence similarity. Therefore if such a gene duplication does exist, they are unlikely to undergo sequence homogenizing events such as the gene conversion seen in *Drosophila* trypsin and α-amylase genes (Hickey *et al.*, 1991; Hickey *et al.*, 1994b), and would evolve independent of each other. This can be seen in Figure 3.7A, where the full length *SNF4* sequences branch into two orthologous groups at the top of the tree rather than genes from the same species (*i.e.* Human and Human2) clustering together.

No phylogenies were presented for the family of human and mouse ESTs identified as *SNF4* isoforms (see Section 3.3.2). The data contained in the GenBank EST database comes from cloned cDNAs which have only been sequenced from the 5' end. As a result, each entry consists of 100 - 700 bp representing an incomplete cDNA sequence. Full length cDNA strands are only produced in a fraction of the synthesized population, so the EST entries will span a range of regions within a given cDNA. The 6 EST sequences identified as being distinct from known full length sequences do not all align to the same region of the SNF4 protein (Figure 3.5) and do not contain as much phylogenetically informative data as a full length sequence. Therefore they were omitted from a phylogenetic analysis as the true relation of the different ESTs with the published *SNF4* sequences cannot be determined with certainty until sequence data spanning their full length is obtained.

Multiple copies of the *SNF4* gene would allow for differential regulation of the gene copies under different stimuli. As it is, the SNF1/SNF4 protein complex has been implicated in regulating a diverse range of physiological responses in different taxa. These range from glucose derepression in yeast (Celenza *et al.*, 1989), to fatty acid and sterol biosynthesis in rats and humans (Hardie, 1994), to thyroid hormone signal transduction in goats (Piosik *et al.*,

1996). It is currently unknown as to whether or not the SNF1/SNF4 complex regulates more than one of these responses in a single species. There is indirect evidence to support a multifunction role for these proteins, as SNF1/SNF4 activity levels are negatively correlated with acetyl CoA carboxylase activities (*i.e.* fatty acid biosynthesis) in yeast (Mitchelhill *et al.*, 1994) and the mammalian acetyl CoA carboxylase enzyme is regulated by glucose levels via a phosphorylation/dephosphorylation pathway (Louis and Witters, 1992). The presence of multiple gene copies of *SNF4*, which is the activator subunit in the SNF1/SNF4 complex, could produce a regulatory system where each copy of SNF4 is responsive to a different stimulus (*e.g.* glucose, ATP concentration, thyroid hormone, etc.) allowing for the precise gene regulation of very different systems in different tissues.

3.4.3 *Comparison of the* Drosophila SNF4 *with other related proteins*

The *Drosophila* gene does not show significant similarity to many proteins, but one gene which does show low level similarity is IMP dehydrogenase. The similarity between these two proteins is found at low levels over the length of the shorter *SNF4* homologues and is not concentrated in the catalytic region of IMPDH (Sintchak *et al.*, 1996). The observed similarity therefore, is not likely to be based on conserved function.

Inosine monophosphate dehydrogenase catalyzes the conversion of IMP to XMP, a precursor to GMP. As such, it is often a necessary enzyme for DNA synthesis and inhibition of this gene can result in cell cycle arrest (Catapano *et al.*, 1995). Cell cycle arrest is a prudent step if available energy levels are low, as in the biological scenario during which the mammalian SNF1/SNF4 complex is inhibited. The only studies examining SNF1/SNF4 regulation by ATP levels have focused on the effect of these nucleotide triphosphates on the SNF1 catalytic domain. If regulation of this protein complex is accomplished by post translational mechanisms involving the noncatalytic subunits as well, the protein sequence of SNF4 may reflect an ancestral mechanism of regulation common to IMPDH. Therefore, not

only is SNF4 distantly related to nucleotide binding proteins, but it appears to be evolutionarily conserved through the eukaryotic lineage.

Chapter 4

PATTERNS IN THE NUCLEOTIDE AND PROTEIN EVOLUTION OF α-AMYLASE[2]

4.1 Introduction

By focusing on the evolution of protein sequences and associated regulatory mechanisms, the previous chapters have examined the effects of evolution from a macroscopic perspective. Underlying the changes seen at the protein level of course, are mutations at the DNA level which provide the sequence variation required for such evolution. Mutations are usually thought to be selectively neutral, however there is a growing body of evidence which shows that the mechanisms responsible for generating this sequence variation do not see all possible mutations as being equal. As a result of this, biases towards certain types of mutations result in changes in the nucleotide content of genes. In this chapter such patterns are identified using alpha-amylase sequences. A strand specific preference for pyrimidines is shown using more than 40 amylase sequences from a variety of taxa. Correlations between the nucleotide content of a gene and both codon usage and amino acid content are also shown. Biased amino acid content arising from a bias at the DNA level can be particularly problematic in phylogenetic reconstruction, but this effect is minimal as shown in the phylogeny inferred in this study of α-amylase sequence evolution.

4.1.1 *Background information*

The relative content of the 4 nucleotides varies considerably among different organisms (Sueoka, 1959), different organelles (Jukes and Bhushan, 1986), and different regions of a

[2] Adapted from a manuscript prepared by Erin N. Yoshida, Peter G. Foster, and Donal A. Hickey

single genome (Bernardi *et al.*, 1985). In noncoding DNA there are minimal evolutionary constraints, so highly skewed AT and GC contents are possible. Even in protein coding genes where often there are functional properties which must be maintained, wide ranges in overall nucleotide content are not uncommon. This effect is even more pronounced when the nonsynonymous and synonymous positions are examined separately. At synonymous sites, nucleotide changes do not alter the amino acid at that position, which results in a high tolerance for mutation at such sites. In contrast mutations at nonsynonymous sites result in amino acid substitutions, which are subject to the evolutionary pressures that provide inertia to sequence change.

At present, there is a lively debate among molecular evolutionists regarding the forces that affect the nucleotide composition of protein-coding genes. The main focus of interest has been on the third position of codons where nucleotide changes do not usually result in amino acid changes. Instead of observing a frequency of 0.25 for each nucleotide as would be predicted for completely neutral mutation, there is a predominance of only one or two nucleotides at third codon positions. This phenomenon is known as codon bias. Ikemura (1985) first proposed that, in order to optimize the coding sequence with the isoaccepting tRNA population, translational selection results in nonrandom codon use. Several studies support this theory, and show differences in the degree of codon bias based on the expression level of the genes in question (Aota and Ikemura, 1986; Bulmer, 1991; Kurland, 1993; DeBry and Marzluff, 1994). In most higher eukaryotes however, the nucleotide composition of a gene appears to reflect its chromosomal location, rather than its level of expression (Bernardi and Bernardi, 1985; Sueoka, 1992; Collins and Jukes, 1993; Kliman and Hey, 1994). Therefore it is thought that in these cases, non-selected mutational bias is the cause of codon bias.

Mutational biases can generate variation in the GC content of genes, which may be associated with significant differences in the amino acid composition of their proteins (Sueoka, 1961). The different amino acids are affected by nucleotide bias to different degrees however.

The 20 amino acids are often divided into GC-rich or AT-rich categories, depending on the nucleotides in the first two positions of the codon (Grantham *et al.*, 1980). In particular, GC rich sequences contain more Ala, Arg, Gly and Pro; and less Asn, Ile, Lys, Tyr, Met and Phe than GC poor sequences (Sueoka, 1961; D'Onofrio et al., 1981; Collins and Jukes, 1993; Porter, 1995; Foster, 1997).

Previous studies on this topic examined several genes from a small number of species. In contrast, to reduce selective differences for protein function, a single homologous gene, alpha-amylase, is analyzed from many taxa to test predictions about protein evolution arising from nucleotide biases in this study. The results presented in this chapter show that nucleotide content can be an indicator of biases in both codon usage and amino acid composition.

4.2. Materials and Methods

Alpha-amylase sequences from 41 different taxa were retrieved from GenBank and EMBL. The names and accession numbers of the sequences analyzed are: *Aedes atropalpus* U01219, *Aeromonas hydrophila* L19299, *Alteromonas haloplanktis* X58627, *Anopheles gambiae* L04753, *Anopheles merus* U01210, *Aspergillus oryzae* X12725, *Aspergillus shirousamii* D10461, *Bacillus acidocaldarius* X07261, *Bacillus amyloliquefaciens* J01542, *Bacillus licheniformis* X03236, *Bacillus megaterium* X07261, *Bacillus polymyxa* Y00150, *Bacillus stearothermophilus* X02769, *Bacillus subtilis* J01547, *Butyrivibrio fibrisolvens* M62507, *Clostridium thermosulfurogenes* X54654, *Culex tarsalis* U01211, *Dictyoglomus thermophilum* X15948, *Drosophila erecta* M55995, *Drosophila melanogaster* X04569, *Drosophila virilis* U02029, *Homo sapiens* (pancreatic) M18714, 16, 18, 20, 22, 24, 26, 81, 83, 85; *Homo sapiens* (salivary) M18715, 17, 19, 21, 23, 25, 27, 82, 84, 86; *Mus musculus* (liver) V00719, *Mus musculus* (pancreatic) V00718, *Mus musculus* (salivary) V00717, *Natronococcus* strain Ah-36 D26510, *Rattus norvegicus* M24962, *Pseudomonas* strain KO-8940 D01143, *Saccharomycopsis fibuligera* X05791, *Schwanniomyces occidentalis* S77586,

Streptomyces griseus X57568, *Streptomyces hydroscopicus* M15540, *Streptomyces limosus* M18244, *Streptomyces thermoviolaceus* M34957, *Streptomyces venezuelae* M25263, *Sus scrofa* (protein only) A17230, *Thermoactinomyces vulgaris* X69807, *Thermomonospora curvata* X59159, *Tribolium castaneum* X06905, and *Xanthomonas campestris* M85252.

To quantify codon bias, the effective number of codons (Wright, 1990) was calculated using the Microgenie software package. Grantham *et al.*'s (1980) classification scheme was used to divide the 20 amino acids into three categories - GC rich, AT rich, and intermediate richness, to determine if amino acid composition correlates with nucleotide bias. Finally, the sequences were aligned using ClustalV, and a phylogeny inferred using the Phylip v3.5 parsimony algorithm with a subset of 24 taxa, rooted with the β-amylase sequence from *Bacillus polymyxa*.

4.3 Results

The alpha-amylase sequences used in this study include vertebrate, invertebrate, fungal and bacterial amylases, and span a wide range of GC contents. From the data, it is seen that there is no strict correlation between GC content of the gene and phylogenetic grouping of the organism. For instance, a typical example is the insect amylases, where there is a large difference in GC content between *Drosophila* (63%) and *Tribolium* (49%) (Figure 4.1A).

The GC content at the third codon position was plotted against the adenine and thymine contents at the same position. The results (Figure 4.1A) show that, while the frequency of both A and T is negatively correlated with GC content - as expected, the percentage T tends to exceed the percentage A. Likewise, when the percent G and C are plotted separately against the AT content of the third codon position (Figure 4.1B), we see an excess of C over G. This means that, regardless of the GC content of the genes, there is a bias in favor of T and C, and against A and G (Figure 4.1C). This strand-specific bias is hereto referred to as a "pyrimidine

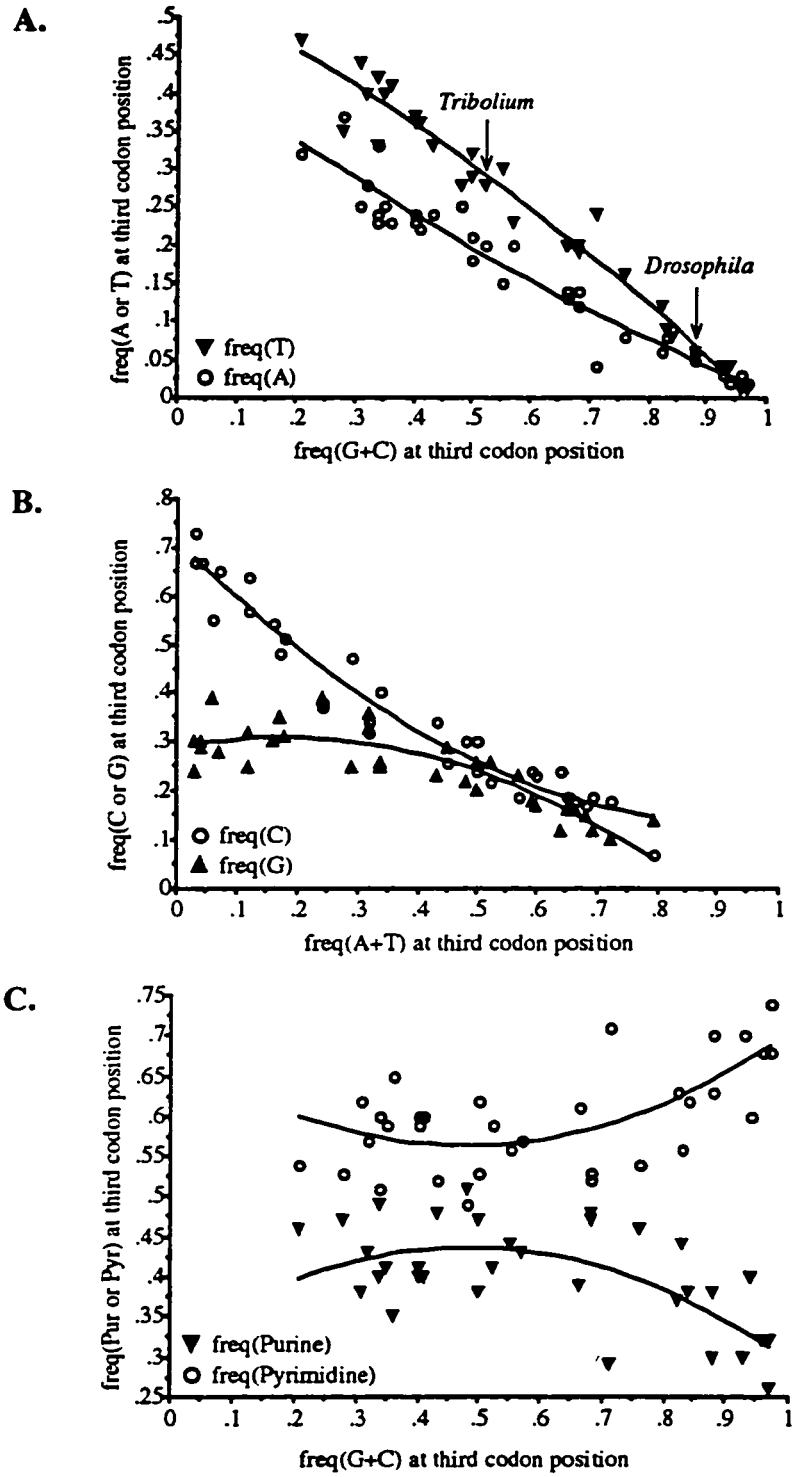


Figure 4.1. Nucleotide biases at the third codon position in alpha-amylases. (A) Frequency of thymine and adenine at the third codon position. For thymine: r = 0.99, P = 0.0001; adenine: r = 0.97, P = 0.0001. (B) Frequency of cytosine and guanine at the third codon position. For cytosine: r = 0.98, P = 0.0001; guanine: r = 0.87, P = 0.0001. (C) Frequency of pyrimidines and purines at the third codon position. For pyrimidine: r = 0.67, P = 0.0001, purine: r = 0.66, P = 0.0001.

preference". Other studies show that this preference is not confined to amylase sequences (Wu and Maeda, 1987; Mrázek and Kypr, 1994).

It is plausible that the extreme nucleotide bias present in some of these sequences would, necessarily, influence the patterns of codon usage. For instance, considering the *Streptomyces* amylases, the percentage GC at the third codon position is over 95%. If one then considers the strand-specific pyrimidine preference, it becomes apparent that the vast majority of codons in these genes must end in C (Figure 4.1B). For the entire data set, the effective number of codons (Wright, 1990) is highly correlated with the GC content of the third codon position (Figure 4.2). Specifically, both positive and negative deviations from an intermediate GC content result in a reduction in the effective number of codons. Again, one should note that organisms that are taxonomically close (such as *Drosophila* and *Tribolium*) may fall in very different parts of the curve.

At this stage, the data show that there is a very strong correlation between the overall nucleotide content of these sequences and the patterns of codon usage. The question to be asked is whether the nucleotide bias is a cause, or an effect, of codon bias. The first indication that the underlying cause may be at the nucleotide level comes from the observation that all codons are biased towards a preponderance of a single nucleotide at the third position (Wright and Bibb, 1992), *e.g.* C in the case of the GC-rich sequences, or T in the case of the AT-rich sequences. Most models of selection for optimal codons do not predict that the same nucleotide should be preferred in all codon groups. An explanation involving nucleotide biases, on the other hand, would predict that all "preferred" codons would share the same nucleotide. It would also predict that the forces causing this bias should affect all codon positions, both synonymous and non-synonymous, equally. It is only because of selective constraints at the non-synonymous positions that the response to the nucleotide pressure will be less evident at these sites. To test this prediction, a comparison of the nucleotide composition at all three codon positions was used (Figure 4.3). There is a highly significant correlation between the overall GC content of the sequences and the GC content of each of the

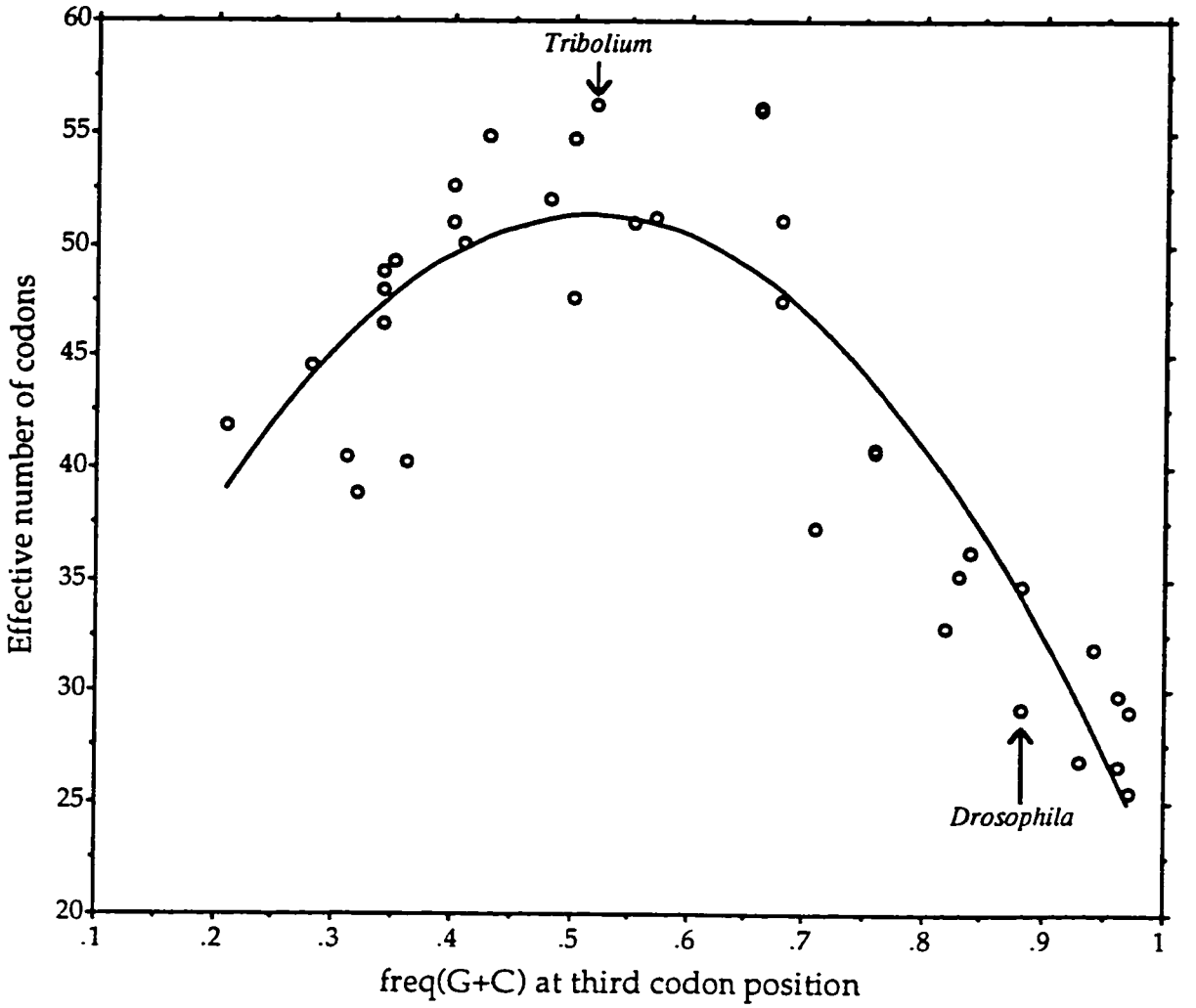


Figure 4.2. Correlation of GC content with codon bias. The effective number of codons (Wright, 1990) was used as the index of codon bias. r = 0.91, P = 0.0001

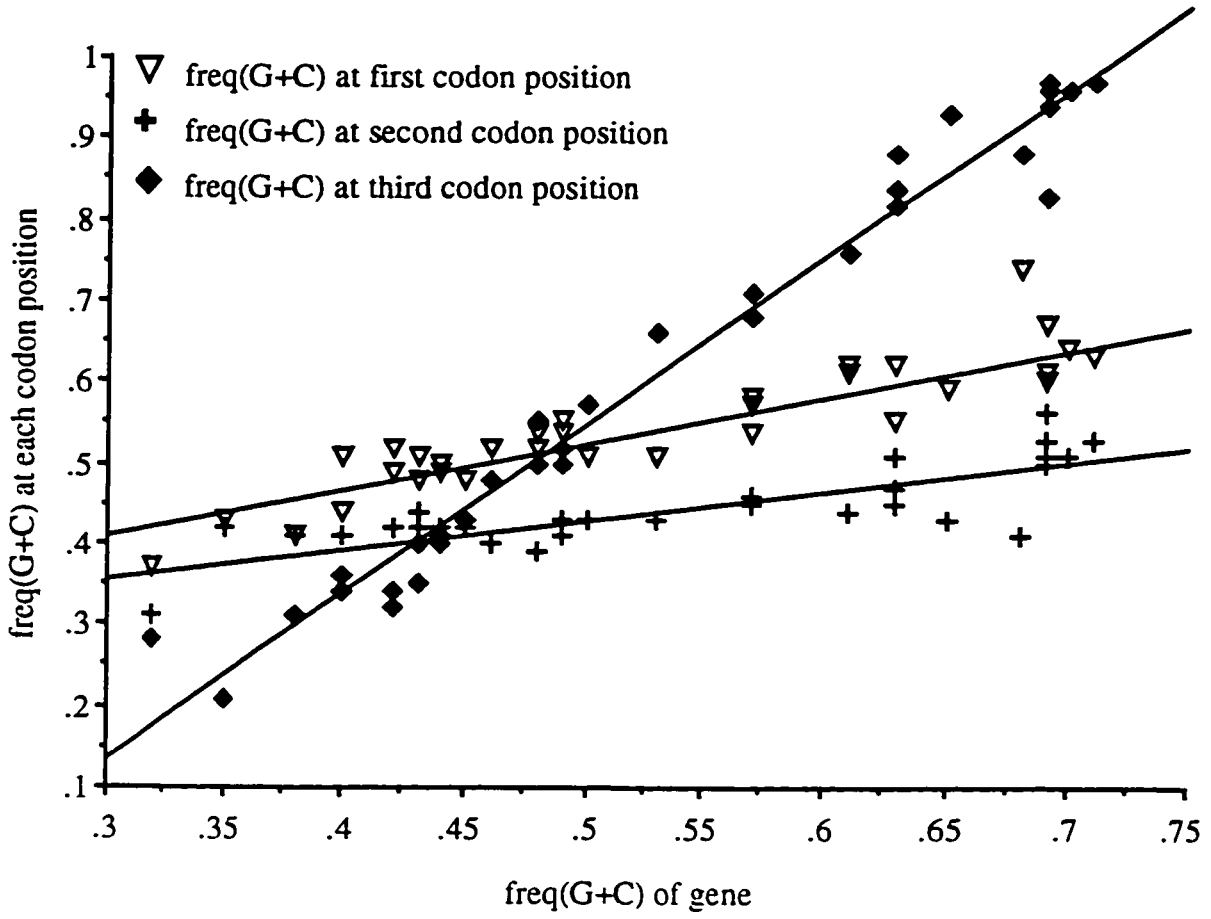


Figure 4.3. Extent of nucleotide bias at each codon position. For the first codon position: m = 0.57, r = 0.9, p = 0.0001; second codon position: m = 0.36, r = 0.81, p = 0.0001; third codon position m = 2.07, r = 0.99, p = 0.0001.

three codon positions, although the magnitude of the changes is much larger for the third position. This difference between the three codon positions is entirely consistent with a model of uniform nucleotide pressure affecting all three positions, but balanced by different levels of selective constraint at the synonymous and non-synonymous sites.

Based on the observation of correlated changes in nucleotide composition at all three codon positions, it is predicted that extreme nucleotide bias would affect, not only the patterns of codon usage, but also the amino acid composition of the encoded proteins. Rather than simply comparing the amino acid composition with the overall GC content, a test of the correlation between GC content at silent sites with changes in the amino acid composition was used. Since there is no *a priori* reason that the amino acid composition of a protein should reflect the nucleotide content of the silent sites, it is reasoned that such a relationship would be a clear indication of a single, underlying cause for both these biases. Using Grantham's classification (Grantham, *et al.*, 1980) of GC-rich and AT-rich amino acids, it was found that there was in fact, a very highly significant correlation between changes in amino acid composition and the nucleotide composition at the silent sites (Figure 4.4A). The earliest investigations of the relationships between nucleotide content and amino acid content (Sueoka, 1961) indicated that some amino acids such as alanine and isoleucine showed a particularly good correlation with nucleotide content. The results for each of these amino acids are shown separately in Figure 4.4B. It should be noted that the relationship between GC content and the frequency of these two amino acids is not linear. It would appear that the initial response to GC or AT pressure is seen only at the silent sites, and that an appreciable change in amino acid composition occurs only when there is extreme nucleotide bias. Because of the selective constraints on amino acid changes, this non-linearity makes intuitive biological sense.

These results have a number of implications for molecular evolutionary studies. The first is that phylogenetic trees constructed from amino acid sequences will not be entirely free of the effects of nucleotide bias. In the most extreme case, one might expect proteins with shared nucleotide biases to group together. A phylogeny based on the sequences used in this

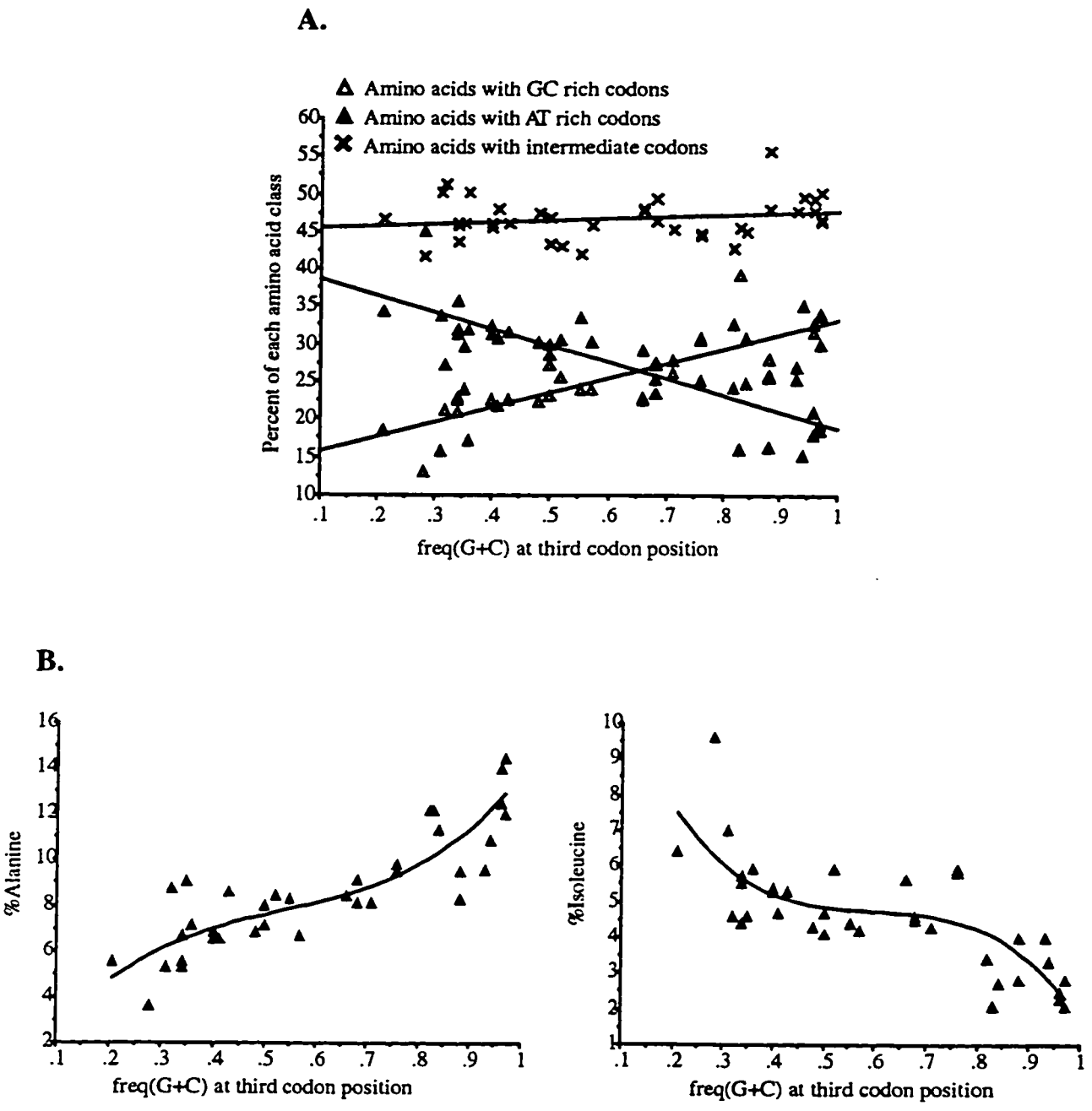


Figure 4.4. Effect of nucleotide bias on amino acid composition. (A) Relation between classes of amino acids and GC content. The twenty amino acids were divided into three classes (Grantham *et al.*, 1980), with leucine and arginine classified as an intermediate and GC rich amino acid, respectively. For the GC rich amino acids: r = 0.80, P = 0.0001; AT rich amino acids: r = 0.81, P = 0.0001; Intermediate amino acids: r = 0.14, P = 0.0001. (B) Correlation of alanine (left) and isoleucine (right) content with nucleotide bias. For alanine: r = 0.89, P = 0.0001; Isoleucine: r = 0.83, P = 0.0001

study shows that this is not true (Figure 4.5). As has been shown previously (Janacek, 1994), vertebrate, invertebrate and bacterial amylases form one clade while fungal amylases form a second clade. Most important, amylases from phylogenetically related organisms cluster together despite large differences in nucleotide composition. This result can be best understood if one considers that, for instance, while *Drosophila* and *Streptomyces* share a relative excess of alanine residues, these "extra" alanines are not necessarily at corresponding positions in the two sequences. More rigorous tests will be needed, however, before one can completely discount the effect of amino acid bias in those phylogenies which attempt to make more subtle phylogenetic distinctions.

Throughout this study, correlations are presented which are highly statistically significant. It can be argued that such relationships would naturally occur, since the common ancestry of the sequences implies that the data points are not independent. Due to this problem, the actual probabilities associated with the regressions may be higher than calculated, but there are several reasons why the correlations are not believed to be artifacts of the data set. First, all of the major lineages show large ranges in GC content, so there is no clustering according to phylogenetic grouping (Figure 4.1). Next, trends observed on a global scale are also observed at a finer resolution in the phylogeny. To illustrate this, the relative amounts of isoleucine - which was particularly responsive to changes in nucleotide content (Figure 4.4B), were examined between 2 pairs of taxa from different regions of the phylogeny (Table 4.1). For both pairs (*Streptomyces/Alteromonas*, and *Drosophila/Tribolium*), the AT rich species has considerably more isoleucine in comparison to the GC rich species, which is in agreement with the results presented in Figure 4.4B. Finally, using mitochondrial sequences from carp (GC rich) and honeybees (AT rich), the observations can be extended to non-nuclear genes (Table 4.1) as an indication of the strength of the correlations.

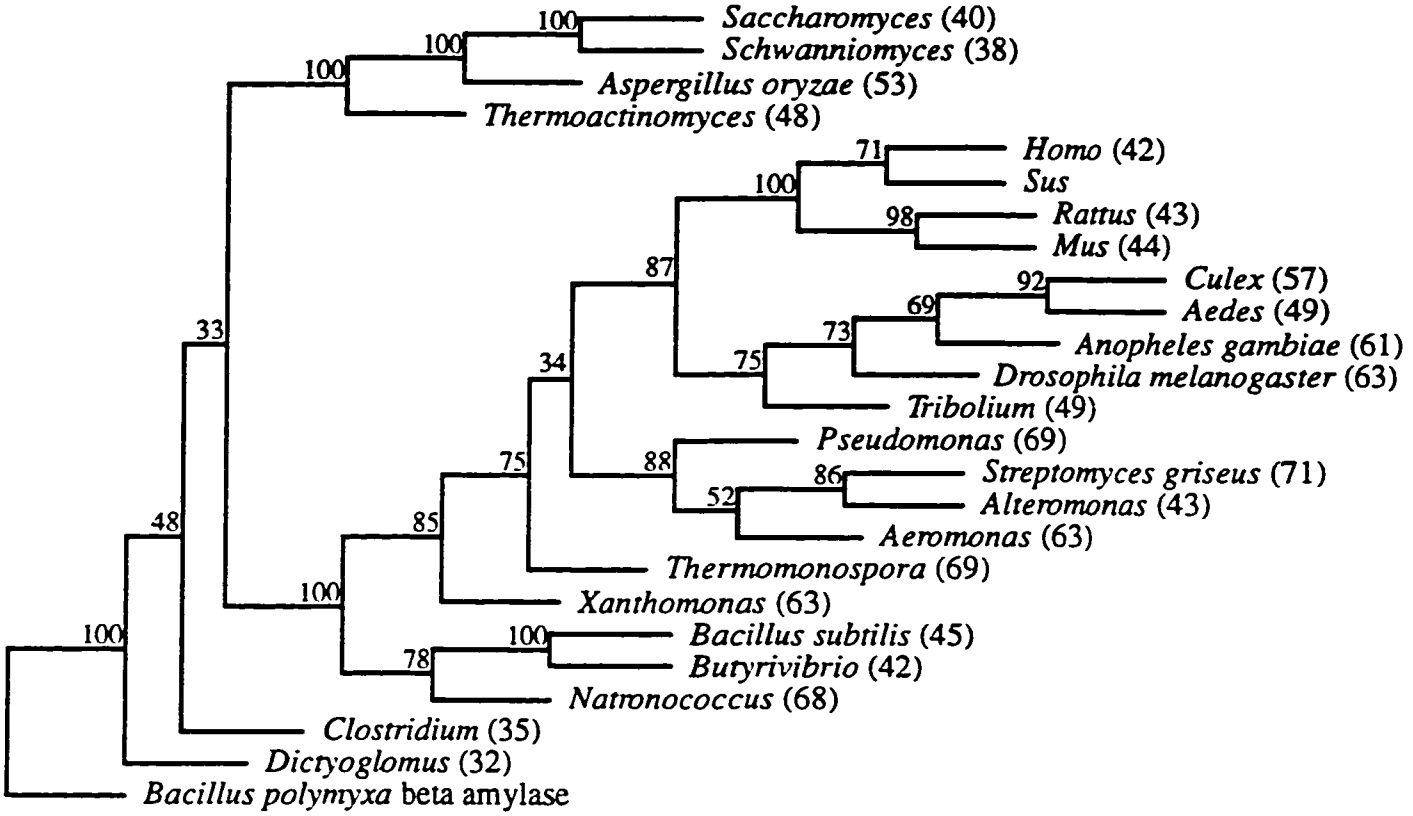


Figure 4.5. Phylogeny derived from α-amylase protein sequences. The parsimony algorithm of Phylip v3.5 was employed, and the tree presented is the consensus of 100 bootstrap replicates. The phylogeny was inferred using twenty-four representative taxa, and rooted with the β-amylase sequence from *Bacillus polymyxa*. For those genera which were represented by more than one species, the genus and species name is used, otherwise just the genus is indicated. The bootstrap values are indicated at the branch nodes, and the values in parentheses are the GC content of the respective sequences.

Table 4.1. Comparison of isoleucine content between AT rich and GC rich taxa

	Number of isoleucine residues	
	AT rich taxon*	GC rich taxon*
Prokaryote amylases	22	12
Insect amylases	29	20
Mitochondrial proteins	500	294

*Pairs of taxa which show large differences in nucleotide content were sampled from different phylogenetic lineages. The species used for the prokaryote amylase comparison were *Alteromonas haloplanktis* and *Streptomyces griseus*; for the insect amylases, *Drosophila melanogaster* and *Tribolium castaneum*; and *Apis mellifera* and *Cyprinus carpio* for the comparison using the 13 protein coding mitochondrial genes.

4.4 Discussion

It is now more than thirty years since Sueoka (1961) first observed a relationship between total nucleotide content and the total amino acid composition of proteins. More recent studies have used molecular sequence data to infer similar relationships within groups of unrelated genes (Bernardi and Bernardi, 1987; D'Onofrio *et al.*, 1991; Collins and Jukes, 1993). What was done in this chapter is to choose a single gene from a wide variety of organisms in order to get a more direct impression of how nucleotide biases might affect the molecular evolution of a particular gene. By using homologous sequences, the data set is less "noisy"; this accounts for the ease with which the various trends can be seen. Ideally, one would like to repeat this analysis using other genes for which sequences are now available from a wide variety of phylogenetically diverse organisms. The recent studies on

mitochondrial genes by Foster *et al.* (1997) and Foster (1997) provide an example of such data and their results are consistent with the ideas presented here. Other examples are provided by the correlated patterns of nucleotide and amino acid change in retroviral, P-450, and mitochondrial cytochrome b genes (Bronson and Anderson, 1994; Jermiin *et al.* 1994; Porter, 1995).

The results presented indicate that nucleotide biases can be an important factor (but most probably not the only factor) in determining both codon usage and amino acid content of proteins. The data do not, however, address the possible causes of the nucleotide biases themselves. Mutational bias is the usual, and most reasonable explanation for species-specific nucleotide bias (Wolfe *et al.*, 1989; Ellsworth *et al.*, 1994). Variations in nucleotide composition between genes within a species can be explained by intragenomic variations in the balance between AT-biased mutations and GC-biased DNA repair (Brown and Jiricny, 1988; Hickey et al., 1991 and 1994). The recent finding that DNA repair and transcription are functionally coupled processes means that transcribed sequences are subject to more DNA repair than adjacent, non-transcribed sequences (Hanawalt, 1994). Not only are transcribed sequences over-repaired, but there is a preferential repair of the transcribed DNA strand. This strand specificity may provide an explanation for the pyrimidine preference seen in this data set. An alternative explanation is that the larger physical size of purines compared to pyrimidines may result in selection against purines at the third codon position. This would reduce possible steric incompatibilities between purine-purine pairings of the codon and anticodon.

Regardless of the mechanism responsible for nucleotide biases, this analysis demonstrates that there are molecular evolutionary forces acting at the nucleotide level that can produce significant, non-random evolutionary changes at the protein level. This aspect of protein evolution does not fit neatly within either of the two main paradigms of evolution: Darwinian selection or genetic drift.

Chapter 5

CONCLUSIONS

5.1 Summary

This thesis dealt with the evolution of the *Drosophila* α-amylase gene and the mechanisms which regulate it. Chapter 1 began with a description of this gene with respect to its function and different modes of regulation, including glucose repression. As the mechanism of glucose repression in *D. melanogaster* is little characterized, the yeast system for regulating galactose metabolism was presented as a paradigm on which to model alpha-amylase gene regulation. Glucose derepression was also introduced and the significance of the SNF1/SNF4 kinase complex was described in yeast gene regulation. It was noted that evolutionary conservation in eukaryotes is seen in α-amylase coding sequences, MIG1 zinc finger proteins, and SNF1/SNF4 subunit proteins.

In Chapter 2, functional conservation of the repression apparatus between yeast and flies was shown. A *Saccharomyces* expression system was developed for the expression of reporter gene constructs driven by regions of the *Drosophila* α-amylase promoter. Both constructs which contained only α-amylase promoter sequence, as well as hybrid *HIS3*-based promoters with shorter α-amylase sequence insertions were tested for expression. Using episomal vectors, promoters which contained two putative repression motifs with a minimum of 126 bp of flanking sequence expressed in a glucose repressible manner. When one of the repression elements was removed, or the promoter shortened to 81 bp, expression was maintained but regulation by glucose repression was not. There was a high level of variance associated with the data, possibly due to differences in plasmid copy number between individual transformants. As a result, integration of the expression cassette into the yeast genome was used for select constructs. Gene expression levels were lower for the integrated

plasmids compared to the nonintegrated plasmids, and an equally high degree of variance was seen for both data sets. One construct which contained 126 bp of *Drosophila* α-amylase sequence was used for both the integrative and nonintegrative assays. Although both vectors showed higher expression in glycerol-grown cultures as compared to glucose-grown ones, the difference was only statistically significant for the nonintegrated construct. These data indicated that the *Drosophila* repression signals are recognizable in a foreign system, and they are likely contained within a 126 bp promoter element. Glucose repression motifs in the yeast *GAL1* gene have been narrowly assigned to 25 bp motifs. Although there is functional conservation in repression machinery between flies and yeast, it is likely that the *Drosophila* regulatory signals are non-optimal in a yeast context, and as a result, they can not be as finely resolved using this transgenic system.

Data from this transgenic yeast expression system did confirm previous observations (Hawley *et al.*, 1992; Magoulas *et al.*, 1993a) that the motifs which regulate glucose repression in the *Drosophila* α-amylase gene are located within a few hundred base pairs 5' to the coding sequence. What was not achieved however, was a fine scale mapping of possible repression motifs which regulate this process. As suggested in Chapter 2, this could be due to the non-optimization of the *Drosophila* motifs for the *Saccharomyces* system. Additionally, this might also be a repercussion of more stringent regulatory requirements placed upon the yeast regulatory system. The combination of tissue-specific, developmental, and glucose repressible regulation of the *Drosophila* α-amylase gene provides several different pathways for maintaining correct expression of this gene. Since yeast do not have these additional pathways it is possible that *Saccharomyces* gene regulation is more straight-forward and precise, with very specific regulatory signals used for different pathways. Christy and Nathans (1989) analyzed the binding motifs recognized by WT1, a MIG1 homologue found in mammals. Although the core 5'-SYGGRG-3' motif is common between both transcription factors, the binding properties of the human homologue were not as context-dependent and well-defineable as the yeast homologue appears to be (Lundin *et al.*, 1994). Even amongst different species of

the genus *Drosophila*, comparisons of promoter regions known to regulate glucose repression fail to highlight specific sequence motifs which may regulate this response (Magoulas *et al.*, 1993b). Therefore, although glucose repression in *Drosophila* may occur via a homologous pathway to what has been characterized in yeast, the actual binding properties of the *trans*-acting factors to the *cis*-acting sequences may be more similar to what is observed in other metazoans than as seen in *Saccharomyces*.

This explanation is of course, based on the hypothesis that glucose regulation in *Drosophila* occurs by a homologous MIG1-dependent pathway. Although the removal of one of the two putative MIG1 binding sites in the *Drosophila* promoter resulted in a loss of glucose repressible expression in a nonintegrative assay, this data in itself does not exclude other possible mechanisms. Certainly one of the keys to drawing stronger parallels between yeast and fly gene regulation is showing that homologous proteins exist in the different eukaryotic lineages, and that the different components have similar functions.

Hence, chapter 3 presented the cloning and characterization of the *Drosophila* homologue of *SNF4*, an essential noncatalytic subunit believed to be involved in the derepression of glucose repressed genes. The cloning of the *Drosophila* copy, along with the two yeast, *C. elegans*, and three mammalian sequences previously identified, supports conservation of this gene throughout the eukaryotic lineage. This gene codes for a 648 amino acid protein which shares high sequence similarity with, but is considerably larger than, the other metazoan and yeast homologues. The expression level was estimated to be 1 copy per 1 x 10^6 to 1 x 10^7 transcripts, based on the inability to obtain a signal from a polyA^+ Northern blot and the requirement of a large amount of a Lambda cDNA library as a template for PCR amplification of the cDNA, although this value has yet to be verified using other methodology. It is suggested that it is present in two copies in the *Drosophila* genome, supporting the observation of multiple cDNA isoforms in mammalian systems. If so, a likely scenario is that one gene codes for a short (ca. 330 aa) protein, and the second copy for a longer protein which has a carboxy terminal extension. When phylogenies are inferred using the full length

the full length proteins, the short *SNF4* proteins and the longer homologues with carboxy terminus tails branch as separate orthologous clades. This is due to a level of sequence similarity between the longer sequences which is lower than that between the longer and shorter homologues, but which extends over a larger number of amino acids. When the amino and carboxy regions that are not conserved in all *SNF4* homologues were removed and phylogenies inferred from a truncated protein alignment, the sequences tended to branch more according to the taxonomic relations of the species from which they were derived. Database searches indicated distant homology with IMP dehydrogenases, a few hypothetical *Methanococcus* proteins, and one *C. elegans* IMP dehydrogenase-like protein, otherwise the *Drosophila SNF4* gene is quite distinct from most proteins.

As SNF4 has been implicated in the regulation of glucose repression in yeast as well as fatty acid and steroid biosynthesis in mammals, there is no doubt that it plays a role in signal transduction in energy utilization pathways. With so many biochemical pathways being regulated by available energy stores, it is quite possible that the SNF1/SNF4/GAL83 complex is a key regulatory step in a wide variety of different pathways. Considering the results of Piosik *et al.* (1996), who showed that *SNF4* mRNA levels are responsive to thyroid hormone, it appears that this peptide also participates in signalling pathways which respond to factors other than energy source. As well, several studies have shown a clear tissue-specific expression pattern for *SNF4* (Gao *et al.*, 1996; Piosik *et al.*, 1996; Woods *et al.*, 1996). Although it has only been inferred and not experimentally demonstrated, if *SNF4* were to be coded for by a gene family, each copy of the gene could evolve specific regulatory sequences which are responsive to different tissues and/or stimuli. Therefore, a kinase cascade which responds quickly and effectively to glucose levels in yeast, may have evolved into a hormone sensitive and/or ATP:AMP sensitive regulatory system in higher eukaryotes.

The cloning of *SNF4* from *Drosophila*, combined with the data from the transgenic yeast experiments, adds more evidence to support the hypothesis that glucose repression in *Drosophila* is mediated by a pathway similar to that deduced in yeast. As *snf4* mutants in yeast

are unable to derepress glucose repressed enzymes, it is possible that *SNF4*^{null} strains of *D. melanogaster* will also be deficient at glucose derepression. Certainly in this case, the Bayev *et al.* (1980) *Drosophila* strain featuring a transposable element insertion into the *SNF4* locus could provide a useful genotype for subsequent analysis. By examining its phenotype one could determine if *Drosophila* uses this kinase complex for the dual function of glucose repression - as seen in Saccharomyces, in addition to biosynthesis regulation, as observed in mammals.

As alpha-amylases too, are found in an extremely wide variety of species, the evolution of the α-amylase coding region itself was analyzed in more detail in Chapter 4. Forty-one α-amylase coding sequences from all three superkingdoms were used to determine evolutionary trends at the nucleotide, codon, and amino acid levels. The sequences spanned a wide range of GC contents, which did not relate to taxonomic grouping. The wide range in GC and amino acid contents observed in the α-amylase sequences used in this analysis underscores the sequence plasticity which has already been attributed to this protein (Nakajima *et al.*, 1986). In certain taxa, such as *Drosophila*, the overrepair of duplicated coding sequences during gene conversion is thought to provide the mechanism for extreme nucleotide bias formation (Hickey *et al.*, 1991). In most cases however, the underlying mechanism cannot be as conveniently explained.

It was this large range in GC content however, which was necessary to illuminate many of the evolutionary patterns observed in this study. To summarize, at the nucleotide level nucleotide bias was shown to manifest itself as a pyrimidine preference in the coding strand of nearly all taxa examined. In turn, this nucleotide bias was correlated to codon bias, as sequences which were very GC or AT rich used only a fraction of the 64 possible codons in protein translation. When individual codon positions were tested for their susceptibility to nucleotide bias, all three sites - *i.e.* both synonymous and nonsynonymous sites, showed some degree of nucleotide bias. As a result, it was predicted and shown that the nucleotide bias

affects amino acid content in these proteins. However this effect was not so substantial as to significantly affect phylogenetic analysis, as shown by a phylogeny of α-amylase sequences.

5.2 Future possibilities

With the cloning of *SNF4* and *SNF1* from *Drosophila melanogaster*, one step has been taken towards elucidating the complexes which regulate glucose repression and derepression in this species. Gaps in our knowledge lie both upstream and downstream of these proteins within the signal transduction pathway. Upstream, it would be interesting to determine the signal molecule which regulates SNF1/SNF4 activity in *Drosophila*. In yeast the signal molecule appears to be a hexose phosphate (Witt *et al.*, 1966), whereas in mammals the ATP:AMP ratio is the key signal (Carling *et al.*, 1989). The question which remains to be answered is whether the regulatory mechanism in *Drosophila* is more similar to yeast or mammals, or is it even more different, involving a cAMP-based signaling mechanism as commonly seen in prokaryotes (Magoulas *et al.*, 1992).

In order to unequivocally map the *Drosophila* URS elements, the transcription factor responsible for direct α-amylase repression needs to be identified. All indications point to a MIG1-like transcription factor being involved. Analyses such as DNA footprinting would provide initial information about the specific promoter sites which are recognized by the *trans*-acting factor. Also, when this gene is cloned, the use of transgenic yeast systems as described here will become a powerful tool for analyzing *Drosophila* repression. By replacing components of the *Saccharomyces* regulatory system with their *Drosophila* homologues, not only would a *Saccharomyces* system be useful in studying protein-protein interactions using techniques like the two-hybrid system, functional expression of the *Drosophila* MIG1 homologue in yeast would provide a means to study *D. melanogaster* protein-DNA interactions in a unicellular system.

This thesis also addressed the effects of nucleotide bias on amino acid content and phylogenetic reconstruction. Gene trees are usually inferred using protein rather than DNA sequences, under the belief that protein data is immune to biases observed at the DNA level. Data presented here and in other accounts (Porter, 1995; Foster *et al.*, 1997) have shown that this is not the case. Algorithms need to be designed which take these factors into account, in order to provide a more robust explanation for the manner in which genes evolve.

LITERATURE CITED

Abraham, I., and Doane, W.W. 1978. Genetic regulation of tissue-specific expression of *Amylase* structural genes in *Drosophila melanogaster. Proc. Natl. Acad. Sci. USA.* **75**: 4446-4450.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Aota, S-I., and Ikemura, T. 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucl. Acids Res.* **14**: 6345 - 6355.

Bahn, E. 1967. Crossing over in the chromosomal region determining amylase isozymes in *Drosophila melanogaster. Hereditas.* **58**: 1-12.

Bayev, A.A.Jr. Krayev, A.S., Lyubomirskaya, N.V., Ilyin, Y.V., Skryabin, K.G., and Georgiev, G.P. 1980. The transposable element *Mdg3* in *Drosophila melanogaster* is flanked with perfect direct and mismatched inverted repeats. *Nucl. Acids Res.* **8**: 3263-3273.

Benkel, B.F., Chow, C., and Hickey, D.A. 1985. Glucose repression of maltase and sucrase activity in *Drosophila melanogaster*. Unpublished.

Benkel, B.F., and Hickey, D.A. 1986a. Glucose repression of amylase gene expression in *Drosophila melanogaster. Genetics* **114**: 137-144.

Benkel, B.F., and Hickey, D.A. 1986b. The interaction of genetic and environmental factors in the control of amylase gene expression in *Drosophila melanogaster. Genetics* **114**: 943-954.

Benkel, B.F., and Hickey, D.A. 1987. A *Drosophila* gene is subject to glucose repression. *Proc. Natl. Acad. Sci. USA.* **84**: 1337-1339.

Bernardi, G., and Bernardi, G. 1986. Compositional constraints and genome evolution. *J. Mol. Evol.* **24**: 1-11.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. 1985. The mosaic genome of vertebrates. *Science* **228**: 953-958.

Boel, E., Brady, L., Brzozowski, A.M., Derewenda, Z., Dodson, G.G., Jensen, V.J., Petersen, S.B., Swift, H., Thim, L., and Woldike, H.F. 1990. Calcium binding in α-amylases: an X-ray diffraction study at 2.1-Å resolution of two enzymes from *Aspergillus. Biochemistry* **29**: 6244-6249.

Boer, P.H., and Hickey, D.A. 1986. The alpha-amylase gene in *Drosophila melanogaster*: nucleotide sequence, gene structure and expression motifs. *Nucl. Acids Res.* **14**: 8399-8411.

Bronson, E.C., and Anderson, J.N. 1994. Nucleotide composition as a driving force in the evolution of retroviruses. *J. Mol. Evol.* **38**: 506-532.

Brown, C.J., Aquadro, C.F., and Anderson, W.W. 1990. DNA sequence evolution of the amylase multigene family in *Drosophila pseudoobscura*. *Genetics* **126**: 131-138.

Brown, T. 1993. "Analysis of RNA by Northern and slot blot hybridization". *in* "Current procols in molecular biology". pp. 4.9.1-4.9.14. Edited by F.M. Ausubel, R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, and K. Struhl. John Wiley & Sons Inc. Cambridge, MA.

Brown, T.C., and Jiricny, J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**: 705-711.

Buisson, G., Duée, E., Haser, R., and Payan, F. 1987. Three dimensional structure of porcine pancreatic α-amylase at 2.9 Å resolution. Role of calcium in structure and activity. *EMBO.* **6**: 3909-3916.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897-907.

Call, K.J., Glase, T., Ito, C.Y., Buckler, A.J., Pelletier, J., Haber, D.A., Rose, E.A., Kral, A., Yeger, H., Lewis, W.H. 1990. Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. *Cell* **60**: 509-520.

Carling, D., Aguan, K., Woods, A., Verhoeven, A.J.M., Beri, R.K., Brennan, C.H., Sidebottom, C., Davison, M.D., and Scott, J. 1994. Mammalian AMP-activated protein kinase is homologous to yeast and plant protein kinases involved in the regulation of carbon metabolism. *J. Biol. Chem.* **269**: 11442-11448.

Carling, D., Clarke, P.R., Zammit, V.A., and Hardie, D.G. 1989. Purification and characterization of the AMP-activated protein kinase. Copurification of acetyl-CoA carboxylase kinase and 3-hydroxy-3-methylglutaryl-CoA reductase kinase activities. *Eur. J. Biochem.* **186**: 129-136.

Carlson, M., and Botstein, D. 1982. Two differentially regulated mRNAs with different 5' ends encode secreted and intracellular forms of yeast invertase. *Cell* **28**: 145-154.

Carlson, M., Osmond, B., and Botstein, D. 1981. Mutants of yeast defective in sucrose utilization. *Genetics* **98**: 25-40.

Carlson, M., Osmond, B.C., Neigeborn, L., and Botstein, D. 1984. A suppressor of *snf1* mutations causes constitutive high-level invertase synthesis in yeast. *Genetics* **107**: 19-32.

Catapano, C., Dayton, J., Mitchell, V., and Fernandes, D. 1995. GTP depletion induced by IMP dehydrogenase inhibitors blocks RNA-primed DNA synthesis. *Mol. Pharm.* **47**: 948-955.

Celenza, J.L., and Carlson, M. 1984. Cloning and genetic mapping of *SNF1*, a gene required for expression of glucose-repressible genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **4**: 54-60.

Celenza, J.L., and Carlson, M. 1986. A yeast gene that is essential for release from glucose repression encodes a protein kinase. *Science* **233**: 1175-1180.

Celenza, J.L., and Carlson, M. 1989. Mutational analysis of the *Saccharomyces cerevisiae* SNF1 protein kinase and evidence for functional interaction with the SNF4 protein. *Mol. Cell. Biol.* **9**: 5034-5044.

Celenza, J.L., Eng, F.J., and Carlson, M. 1989. Molecular Analysis of the SNF4 gene of *Saccharomyces cerevisiae*: Evidence for physical association of the SNF4 protein with the SNF1 protein kinase. *Mol. Cell. Biol.* **9**: 5045-5054.

Christy, B., and Nathans, D. 1989. DNA binding site of the growth-factor-inducible protein Zif268. *Proc. Natl. Acad. Sci. USA.* **86**: 8737-8741.

Collins, D.W., and Jukes, T.H. 1993. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. *J. Mol. Evol.* **36**: 201-213.

Cooper, J.P., Roth, S.Y., and Simpson, R.T. 1994. The global transcriptional regulators, SSN6 and TUP1, play distinct roles in the establishment of a repressive chromatin structure. *Genes Dev.* **8**: 1400-1410.

Cowell, I.G. 1994. Repression versus activation in the control of gene transcription. *Trends Biochem. Sci.* **19**: 38-42.

Cubero, B., and Scazzocchio, C. 1994. Two different, adjacent and divergent zinc finger binding sites are necessary for CREA-mediated carbon catabolite repression in the proline gene cluster of *Aspergilluc nidulans. EMBO.* **13**: 407-415.

Da Lage, J.-L. Lemeunier, F., Cariou, M.-L., and David, J.R. 1992. Multiple amylase genes in *Drosophila ananassae* and related species. *Genet. Res.* **59**: 85-92.

Dale, S., Wilson, W.A., Edelman, A.M., and Hardie, D.G. 1995. Similar substrate recognition motifs for mammalian AMP-activated protein kinase, higher plant HMG-CoA reductase kinase-A, yeast SNF1, and mammalian calmodulin-dependent protein kinase I. *FEBS Lett.* **361**: 191-195.

DeBry, R.W., and Marzluff, W. 1994. Selection on silent sites in the rodent H3 histone gene family. *Genetics* **138**: 191-202.

de la Guerra, R., Valdes-Hevia, M.D., and Gancedo, J.M. 1988. Regulation of yeast fructose-1.6-bisphosphatase in strains containing multicopy plasmids for this enzyme. *FEBS Lett.* **242**: 149-152.

Doane, W.W. 1969a. "Drosophila amylases and problems in cellular differentiation". pp 73-109. *In* "RNA in Development". Edited by E.W. Hanly. University of Utah Press. Salt Lake City, UT.

Doane, W.W. 1969b. Amylase variants in *Drosophila melanogaster*: Linkage studies and characterization of enzyme extracts. *J. Exp. Zool.* **171**: 321-342.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., and Bernardi, G. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* **32**: 504-510.

Dowzer, C.E.A., and Kelly, J.M. 1991. Analysis of the *creA* gene, a regulator of carbon catabolite repression in *Aspergillus nidulans. Mol. Cell. Biol.* **11**: 5701-5709.

Dyck, J.R.B., Gao, G., Widmer, J., Stapleton, D., Fernandez, C.S., Kemp, B.E., and Witters, L.A. 1996. Regulation of 5'-AMP-activated protein kinase activity by the noncatalytic beta and gamma subunits. *J. Biol. Chem.* **271**: 17798-17803.

Echo, B.S., and Doane, W.W. 1984. Effects of diet on genetically regulated alpha-amylase expression in *Drosophila melanogaster. Genetics* **107** (Suppl.): s28.

Edmonson, D.G., Smith, M.M., and Roth, S.Y. 1996. Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes Dev.* **10**: 1247-1259.

Elledge, S.J., Zhou, Z., Allen, J.B., Navas, T.A. 1993. DNA damage and cell cycle regulation of ribonucleotide reductase. *BioEssays* **15**: 333-339.

Ellsworth, D.L., Hewett-Emmett, D., and Li, W-H. 1994. Evolution of base composition in the insulin and insulin-like growth factor genes. *Mol. Biol. Evol.* **11**: 875-885.

Entian, K.D., and Barnett, J.A. 1992. Regulation of sugar utilization by *Saccharomyces cerevisiae. Trends Biochem. Sci.* **17**: 506-510.

Erickson, J.R., and Johnston, M. 1993. Genetic and molecular characterization of *GAL83*: its interaction and similarities with other genes involved in glucose repression in *Saccharomyces cerevisiae. Genetics* **135**: 655-664.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Ann. Rev. Genet.* **22**: 521-565.

Fischer, J.A., Giniger, E., Maniatis, T., and Ptashne, M. 1988. GAL4 activates transcription in *Drosophila. Nature* **332**: 853-856.

Flick, J.S., and Johnston, M. 1992. Analysis of URS_G-mediated glucose repression of the *GAL1* promoter of *Saccharomyces cerevisiae. Genetics* **130**: 295-304.

Foster, P.G. 1997. Phylogenetic implications of the effect of nucleotide bias on amino acid composition. Ph.D. Thesis. University of Ottawa.

Foster, P.G., Jermiin, L.S., and Hickey, D.A. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J. Mol. Evol.* **44**: 282-288.

Freund, J.E. 1988. "Modern elementary statistics". 7th Ed. pp. 315. Prentice-Hall. Englewood Cliffs, NJ.

Fukasawa, T., and Nogi, Y. 1989. "Molecular genetics of galactose metabolism in yeast". pp1-18. In "Yeast genetic engineering". Edited by P.J. Barr, A.J. Brake, and P. Valenzuela. Butterworth Publishers. Stoneham, MA.

Gancedo, J.M., and Gancedo, C. 1986. Catabolite repression mutants of yeast. *FEMS Microbiol. Rev.* **32**: 179-187.

Gao, G., Fernandez, C.S., Stapleton, D., Auster, A.S., Widmer, J., Dyck, J.R.B., Kemp, B.E., and Witters, L.A. 1996. Non-catalytic beta- and gamma-subunit isoforms of the 5'-AMP-activated protein kinase. *J. Biol. Chem.* **271**: 8675-8681.

Gemmill, R.M., Schwartz, P.E., and Doane, W.W. 1986. Structural organization of the *Amy* locus in seven strains of *Drosophila melanogaster. Nucl. Acids Res.* **14**: 5337-5352.

Giniger, E., Varnum, S.M., and Ptashne, M. 1985. Specific DNA binding of GAL4, a positive regulatory protein of yeast. *Cell* **40**: 767-774.

Grantham,R., Gauthier, C., Gouy, M., Mercier, R., and Pave, A. 1980. Codon catalog usage and the genome hypothesis. *Nucl. Acids Res.* **8**: r49-r62.

Griggs, D.W., and Johnston, M. 1991. Regulated expression of the *GAL4* activator gene in yeast provides a sensitive genetic switch for glucose repression. *Proc. Natl. Acad. Sci. USA*. **88**: 8597-8601.

Gumucio, D.L., Wiebauer, K., Dranginis, A., Samuelson, L.C., Treisman, L.O., Caldwell, R.M., Antonucci, T.K., and Meisler, M.H. 1985. Evolution of the amylase multigene family: YBR/Ki mice express a pancreatic amylase gene which is silent in other strains. *J. Biol. Chem.* **260**: 13483-13489.

Hanawalt, P.C. 1994. Transcription-coupled repair and human disease. *Science* **266**: 1957-1958.

Hanna-Rose, W., and Hansen, U. 1996. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* **12**: 229-234.

Hardie, D.G. 1992. Regulation of fatty acid and cholesterol metabolism by the AMP-activated protein kinase. *Biochim. Biophys. Acta.* **1123**: 231-238.

Hardie, D.G. 1994. Ways of coping with stress. *Nature* **370**: 599-600.

Harvey, D., Hong, L., Evans-Holmes, M., Pendleton, J., Su, C., Brokstein, P., Lewis, S., and Rubin, G.M. 1997. BDGP/HHMI *Drosophila* EST Project. Unpublished.

Hawley, S.A., Doane, W.W. and Norman, R.A. 1992. Molecular analysis of *cis*-regulatory sequences at the α-amylase locus in *Drosophila melanogaster. Biochem. Genet.* **30**: 257-277.

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

Hickey, D.A., and Benkel, B.F. 1982. Regulation of amylase activity in *Drosophila melanogaster*: effects of dietary carbohydrate. *Biochem. Genet.* **20**: 1117-1129.

Hickey, D.A., and Benkel, B.F. 1987. Regulation of amylase gene expression: *Drosophila* amylases as a model experimental system. *CRC Crit. Rev. Biotech.* **3**: 229-241.

Hickey, D.A., Benkel, B.F., Abukashawa, S., and Haus, S. 1988. DNA rearrangement causes multiple changes in gene expression at the amylase locus in *Drosophila melanogaster. Biochem. Genet.* **26**: 757-768.

Hickey, D.A. Benkel, K.I., Fong, Y., and Benkel, B.F. 1994. A *Drosophila* gene promoter is subject to glucose repression in yeast cells. *Proc. Natl. Acad. Sci. USA*. **91**: 11109-11112.

Hickey, D.A., Genest, Y., and Benkel, B.F. 1987. Nucleotide sequence upstream of a glucose-repressible *Drosophila* gene. *Nucl. Acids Res.* **15**: 7184.

Hickey, D.A., Wang, S., and Magoulas, C. 1994b. "Gene duplication, gene conversion and codon bias". *In* "Non-neutral evolution". Edited by B. Golding. pp. 199-207. Chapman and Hall. New York, NY.

Higgins, D.G., Bleasby, A.J., and Fuchs, R. 1992. CLUSTAL V: improved software for multiple sequence alignment. *CABIOS.* **8**: 189-191.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**: 13-34.

Inomata, N., Kanda, K., Cariou, M.-L., Tachida, H., and Yamazaki, T. 1995. Evolution of the response patterns to dietary carbohydrates and the developmental differentiation of gene expression of α-amylase in *Drosophila. J. Mol. Evol.* **41**: 1076-1084.

Janacek, A. 1994. Sequence similarities and evolutionary relationships of microbial, plant and animal α-amylases. *Eur. J. Biochem.* **224**: 519-524.

Jermiin, L.S., Graur, D., Lowe, R.M., and Crozier, R.H. 1994. Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. *J. Mol. Evol.* **39**: 160-173.

Jiang, R., and Carlson, M. 1996. Glucose regulates protein interactions within the yeast SNF1 protein kinase complex. *Genes Devel.* **10**: 3105-3115.

Jiang, R., and Carlson, M. 1997. The SNF1 protein kinase and its activating subunit, SNF4, interact with distinct domains of the SIP1/SIP2/GAL83 component in the kinase complex. *Mol Cell. Biol.* **17**: 2099-2106.

Johnson, A.D. 1995. The price of repression. *Cell* **81**: 655-658.

Johnston, M. 1987. A model fungal gene regulatory mechanism: the *GAL* genes of *Saccharomyces cerevisiae. Microbiol. Rev.* **51**: 458-476.

Johnston, M., and Carlson, M. 1992. "Regulation of carbon and phosphate ultilization". pp 194-241. *In* "The molecular and cellular biology of the yeast *Saccharomyces*". Edited by E.W. Jones, J.R. Pringle, and J.R. Broach. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY.

Johnston, M., Flick, J.S., and Pexton, T. 1994. Multiple mechanisms provide rapid and stringent glucose repression of *GAL* gene expression in *Saccharomyces cerevisiae. Mol. Cell. Biol.* **14**: 3834-3841.

Jukes, T.H., and Bhushan, V. 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *J. Mol. Evol.* **24**: 39-44.

Keleher, C.A., Redd, M.J., Schultz, J., Carlson, M., and Johnson, A.D. 1992. Ssn6-Tup1 is a general repressor of transcription in yeast. *Cell* **68**: 709-719.

Kikkawa, H. 1964. An electrophoretic study on amylase in *Drosophila melanogaster. Jpn. J. Genet.* **39**: 401-411.

Kobayashi, T., Kanai, H., Aono, R., Horikoshi, K., and Kudo, T. 1994. Cloning, expression, and nucleotide sequence of the α-amylase gene from the haloalkaliphilic archaeon *Natronococcus* sp. strain Ah-36. *J. Bacteriol.* **176**: 5131-5134.

Kliman, R.M., and Hey, J. 1994. The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* **137**: 1049-1056.

Klumberg, P. Mathieu, M., Dowzer, C., Kelly, J., and Felenbok, B. 1993. Specific binding sites in the *alc*R and *alc*A promoters of the ethanol regulon for the CREA repressor mediating carbon catabolite repression in *Aspergillus nidulans*. *Mol. Microbiol.* **7**: 847-857.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucl. Acids Res.* **12**: 857 - 872.

Kurland, C.G. 1993. Major codon preference: theme and variations. *Biochem. Soc. Trans.* **21**: 841-846.

Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C., and Sninsky, J.J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodificiency virus type 1 model studies. *Nucl. Acids Res.* **18**: 999-1005.

Lohr, D., Venkov, P., and Zlatanova. J. 1995. Transcriptional regulation in the yeast *GAL* gene family: a complex genetic network. *FASEB J.* **9**: 777-787.

Lombardo, A., Carine, K., and Scheffler, I.E. 1990. Cloning and characterization of the iron-sulfur subunit gene of succinate dehydrogenase from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **265**: 10419-10423.

Louis, N.A., and Witters, L.A. 1992. Glucose regulation of acetyl-CoA carboxylase in hepatoma and islet cells. *J. Biol. Chem.* **267**: 2287-2293.

Lunblad, V. 1993. "Yeast cloning vectors and genes". *in* "Current procols in molecular biology". Edited by F.M. Ausubel, R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, and K. Struhl. John Wiley & Sons Inc. Cambridge, MA.

Lundin, M., Nehlin, J.O. Ronne, H. 1994. Importance of a flanking AT-rich region in target site recognition by the GC box-binding zinc finger protein MIG1. *Mol. Cell. Biol.* **14**: 1979-1985.

Magoulas, C., Bally-Cuif, L., Loverre-Chyurlia, A., Benkel, B., and Hickey, D. 1993a. A short 5'-flanking region mediates glucose repression of amylase gene expression in *Drosophila melanogaster*. *Genetics* **134**: 507-515.

Magoulas, C., Loverre-Chyurlia, A., Abukashawa, S., Bally-Cuif, L., and Hickey, D.A. 1993b. Functional conservation of a glucose-repressible amylase gene promoter from *Drosophila virilis* in *Drosophila melanogaster*. *J. Mol. Evol.* **36**: 234-242.

Magoulas, C., Loverre-Chyurlia, A., and Hickey, D.A. 1992. Amylase *cis*-acting sequences mediate the alleviation of glucose repression by cAMP in *Drosophila*. *Biochem. Cell. Biol.* **70**: 751-757.

Mercado, J.J., Vincent, O., and Gancedo, J.M. 1991. Regions in the promoter of the yeast *FBP1* gene implicated in catabolite repression may bind the product of the regulatory gene *MIG1*. *FEBS Lett.* **291**: 97-100.

Milanovic, M., and Andjelkovic, M. 1993. Biochemical and genetic diversity of alpha-amylase in *Drosophila*. *Arch. Biol. Sci. Belgrade* **45**: 63-82.

Mitchellhill, K.I., Stapelton, D.A., Gao, G., House, C., Michell, B., Katsis, F., Witters, L.A., and Kemp, B.E. 1994. Mammalian AMP-activated protein kinase shares structural and functional homology with the catalytic domain of yeast Snf1 protein kinase. *J. Biol. Chem.* **269**: 2361-2364.

Mrázek, J., and Kypr, J. 1994. Biased distribution of adenine and thymine in gene nucleotide sequences. *J. Mol. Evol.* **39**: 439-447.

Mukai, Y., Harashima, S., and Oshima, Y. 1991. AAR1/TUP1 protein, with a structure similar to that of the β subunit of G proteins, is required for a1-α2 and α2 repression in cell type control of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **11**: 3773-3779.

Mulder, W., Scholten, I.H., and Grivell, L.A. 1995. Distinct transcriptional regulation of a gene for a mitochondrial protein in the yeasts *Saccharomyces cerevisiae* and *Kluyveromyces lactis* despite similar promoter structures. *Mol. Microbiol.* **17**: 813-824.

Nakajima, R., Imanaka, T., and Aiba, S. 1986. Comparison of amino acid sequences of eleven different α-amylases. *Appl. Microbiol. Biotechnol.* **23**: 355-360.

Nardelli, J., Gibson, T., and Charnay, P. 1992. Zinc finger-DNA recognition: analysis of base specificity by site-directed mutagenesis. *Nucl. Acids Res.* **20**: 4137-4144.

Nardelli, J., Gibson, T.J., Vesque, C., and Charnay, P. 1991. Base-sequence discrimination by zinc-finger DNA-binding domains. *Nature* **349**: 175-178.

Nehlin, J.O., Carlberg, M., and Ronne, H. 1991. Control of yeast *GAL* genes by MIG1 repressor: a transcriptional cascade in the glucose response. *EMBO.* **11**: 3373-3377.

Nehlin, J.O., and Ronne, H. 1990. Yeast MIG1 repressor is related to the mammalian early growth response and Wilms' tumour finger proteins. *EMBO.* **9**: 2891-2898.

Nussinov, R. 1986. Sequence signals which may be required for efficient formation of mRNA 3' termini. *Nucl. Acids Res.* **14**: 3557-3571.

Okuyama, E., Tachida, H., and Yamazaki, T. 1997. Molecular analysis of the intergenic region of the duplicated *Amy* genes of *Drosophila melanogaster* and *Drosophila teissieri*. *J. Mol. Evol.* **45**: 32-42.

Östling, J., Carlberg, M., and Ronne, H. 1996. Functional domains in the MIG1 repressor. *Mol. Cell. Biol.* **16**: 753-761.

Parthun, M.R., and Jaehning, J.A. 1992. A transcriptionally active form of GAL4 is phosphorylated and associated with GAL80. *Mol. Cell. Biol.* **12**: 4981-4987.

Pavletich, N., and Pabo, C.O. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**: 809-817.

Payant, V., Abukashawa, S., Sasseville, M., Benkel, B.F., Hickey, D.A., and David, J. 1988. Evolutionary conservation of the chromosomal configuration and regulation of amylase genes among eight species of the *Drosophila melanogaster* species subgroup. *Mol. Biol. Evol.* **5**: 560-567.

Perlman, D. and Hopper, J.E. 1979. Constitutive synthesis of GAL4 protein, a galactose pathway regulator in *Saccharomyces cerevisiae. Cell* **16**: 89-95.

Piosik, P.A., van Groenigen, M., Ponne, N.J., Valentijn, L.J., Bolhuis, P.A., and Baas, R. 1996. Caprine homologue of rodent 5'-AMP-activated protein kinase subunit and yeast SNF4/CAT3 is down-regulated by thyroid hormone. *Mol. Brain Res.* **40**: 240-253.

Porter, T.D. 1995. Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome *P*-450 gene superfamily. *Biochim. Biophys. Acta* **1261**: 394-400.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. 1995. MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* **23**: 4878-4884.

Rauscher, F.J.III, Morris, J.F., Tournay, O.E., Cook, D., Curran, T. 1990. Binding of the Wilms' tumor locus zinc finger protein to the EGR-1 consensus sequence. *Science* **250**: 1259-1262.

Rawn, J.D. 1989. Biochemistry. pp 398-399. Neil Patterson Publishers. Burlington, NC.

Redd, M.J., Arnaud, M.B., and Johnson, A.D. 1997. A complex composed of tup1 and ssn6 repressed transcription *in vitro. J. Biol. Chem.* **272**: 11193-11197.

Rogers, J.D. 1985. Conserved amino acid sequence domains in alpha-amylases from plants, mammals, and bacteria. *Biochem. Biophys. Res. Comm.* **128**: 470-476.

Ronne, H. 1995. Glucose repression in fungi. *Trends Genet.* **11**: 12-17.

Roux, K.H. 1995. "Optimizing and troubleshooting PCR". pp 53-62. *In* "PCR primers - A laboratory manual". Edited by C.W. Dieffenbach and G.S. Dveksler. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY.

Sambrook. J., Fritsch, E.F., and Maniatis, T. 1989. "Molecular cloning - A laboratory manual". 2nd ed. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY.

Schiestl, R.H., and Gietz, R.D. 1989. High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. *Curr. Genet.* **16**: 339-346.

Schüller, H-J., and Entian, K-D. 1988. Molecular characterization of yeast regulatory gene *CAT3* necessary for glucose derepression and nuclear localization of its product. *Gene* **67**: 247-257.

Shibata, H., and Yamazaki, T. 1995. Molecular volution of the duplicated *Amy* locus in the *Drosophila melanogaster* species subgroup: Concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation. *Genetics* **141**: 223-236.

Sintchak, M.D., Fleming, M.A., Futer, O., Raybuck, S.A., Chambers, S.P., Caron, P.R., Murcko, M.A., and Wilson, K.P. 1996. Structure and mechanism of inosine

monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic acid. *Cell* **85**: 921-930.

Sueoka, N. 1959. Heterogeneity in deoxyribonucleic acids. II Dependency of the density of deoxyribonucleic acids on guanine-cytosine content. *Nature* **183**: 1429-1431.

Sueoka, N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl. Acad. Sci. USA.* **47**: 1141-1149.

Sueoka, N. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. *J. Mol. Evol.* **34**: 95-114.

Thevelein, J.M. 1994. Signal transduction in yeast. *Yeast* **10**: 1753-1790.

Thompson, D.B., Treat-Clemons, L.G., and Doane, W.W. 1992. Tissue-specific and dietary control of alpha-amylase gene expression in the adult midgut of *Drosophila melanogaster*. *J. Exp. Zool.* **262**: 122-134.

Treitel, M.A., and Carlson, M. 1995. Repression by SSN6-TUP1 is directed by MIG1, a repressor/activator protein. *Proc. Natl. Acad. Sci. USA.* **92**: 3132-3126.

Tzamarias, D., and Struhl, K. 1994. Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. *Nature* **369**: 757-761.

Tzamarias, D., and Struhl, K. 1995. Distinct TPR motifs of Cyc8 are involved in recruiting the Cyc8-Tup1 corepressor complex to differentially regulated promoters. *Genes Dev.* **9**: 821-831.

Vihinen, M., and Mäntsälä, P. 1989. Microbial amylolytic enzymes. *Crit. Rev. Biochem. Mol. Biol.* **24**: 329-418.

Wang, J., Sirenko, O., and Needleman, R. 1997. Genomic footprinting of MIG1p in the *MAL62* promoter. *J. Biol. Chem.* **272**: 4613-1622.

Williams, F.E., and Trumbly, R.J. 1990. Characterization of *TUP1*, a mediator of glucose repression in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **10**: 6500-6511.

Wilson, W.A., Hawley, S.A., and Hardie, D.G. 1996. Glucose repression/derepression in budding yeast: SNF1 protein kinase is activated by phosphorylation under derepressing conditions, and this correlates with a high AMP:ATP ratio. *Curr. Biol.* **6**: 1426-1434.

Witt, I., Kronau, R., and Holzer, H. 1966. Repression von alkoholdehydrogenase, malatdehydrogenase, isocitratlyase und malatsynthase in hefe dürch glucose. *Biochim. Biophys. Acta.* **118**: 522-537.

Wolfe, K.H., Sharpe, P.M., and Li, W-H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283-285.

Woods, A., Cheung, P.C.F., Smith, F.C., Davison, M.D., Scott, J., Beri, R.K., and Carling, D. 1996. Characterization of AMP-activated protein kinase β and γ subunits. Assembly of the heterotrimeric complex *in vitro*. *J. Biol. Chem.* **271**: 10282-10290.

Woods, A., Munday, M.R., Scott, J., Yang, X., Carlson, M., and Carling, D. 1994. Yeast SNF1 is functionally related to mammalian AMP-activated protein kinase and regulates acetyl-CoA carboxylase *in vivo*. *J. Biol. Chem.* **269**: 19509-19515.

Wright, F. 1990. The 'effective number of codons' used in a gene. *Gene* **87**: 23-29.

Wright, F., and Bibb, M.J. 1992. Codon usage in the G+C-rich *Streptomyces* genome. *Gene* **113**: 55-65.

Wu, C-I., and Maeda, N. 1987. Inequality in mutation rates of the two strands of DNA. *Nature* **327**: 169-170.

Yang, X., Jiang, R., and Carlson, M. 1994. A family of proteins containing a conserved domain that mediates interaction with the yeast SNF1 protein kinase complex. *EMBO.* **13**: 5878-5886.

Zimmermann, F.K., Kaufmann, I., Rasenberger, H., and Haussmann, P. 1977. Genetics of carbon catabolite repression in *Saccharomyces cerevisiae*: genes involved in the derepression process. *Mol. Gen. Genet.* **151**: 95-103.

Zitomer, R.S., and Lowry, C.V. 1992. Regulation of gene expression by oxygen in *Saccharomyces cerevisiae*. *Microbiol. Rev.* **56**: 1-11.

APPENDIX A

Nucleotide sequence of the *Drosophila melanogaster* α-amylase promoter region numbered with respect to the transcriptional start site, with clone boundary points indicated:

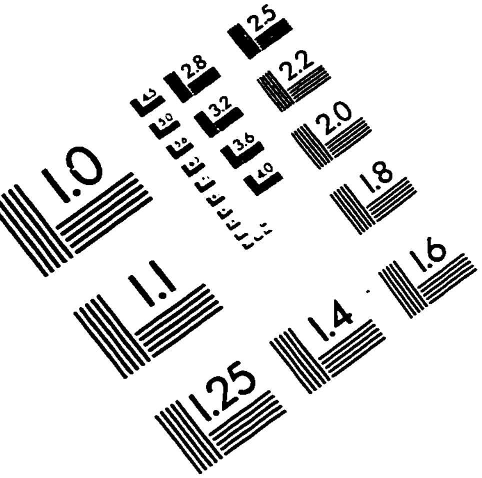
```
-469 (pBY59)
|
CGTATTCCAG AGAACGGCGC AGCCAAAGCT TCAAACCAAA ATCGCTTGCT

ACCTTTATTT TCAACATTTT TAGGCGATAT TGCATGATTT CAATGCTTTC

AAATACGCTA AAAAATCCAA ATAACAATTC ACAGTAAACC CGCTCCTAGG
                                                -274(EY5)
                                                  |
AGCGTGAACG TAATAAATAG TCAATAAATT CCCAACTGAA ACCGATTTCA
      -261(EY33)
        |
AAGGAATGCA TTTTCCCGAT GAGTTATTGA TACAAATATA ACGAAAATAA

GCCGACTCAC TAATCATCAG CGAAAAATTG CGATCTCCAG TCAATACGTC
-168(EY6)   -157(EY35, EY41)
 |            |
TGCTCGGAAT TGTGATTTGA CAAACTAATC GCCAGTCAGA CCCCATGCGT
(EY38)-112  -109(Magoulas et al., 1993a)
       |    |
GAAAAAACCC CTTAGGGAGC GATAAGATCC CATGCAGTCA CAAATCACTC
                        -47(EY41)       -32(EY33, 35, 38)
                          |               |
CCCGCGAAGC CCTCAGATAA AGTAGCAGTG GGGTCCACTA TATAAGGAGC
                    +1
                     |
GGCTCTGAGT AGTTCCGACC AGAGTGAAAC TGAACTTCCA TCTGGAATCA
+34(pBY59, EY5, EY6)
  |
TCATG
```

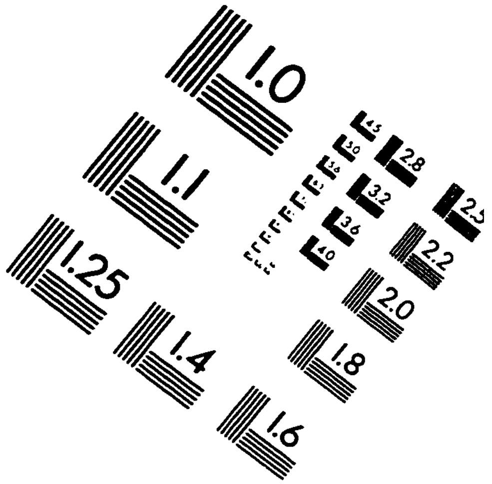
Nucleotide sequence of the *Saccharomyces cerevisiae HIS3* promoter region numbered with respect to the transcriptional start site:

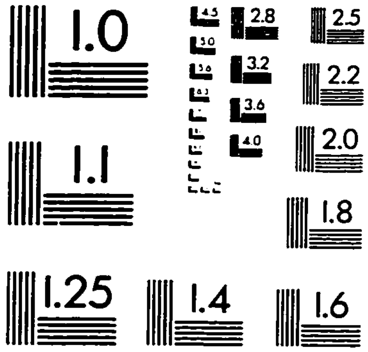
```
-211  -203
|      |
ACAGTCCTTT CCCGCAATTT TCTTTTTCTA TTACTCTTGG CCTCCTCTAG
```

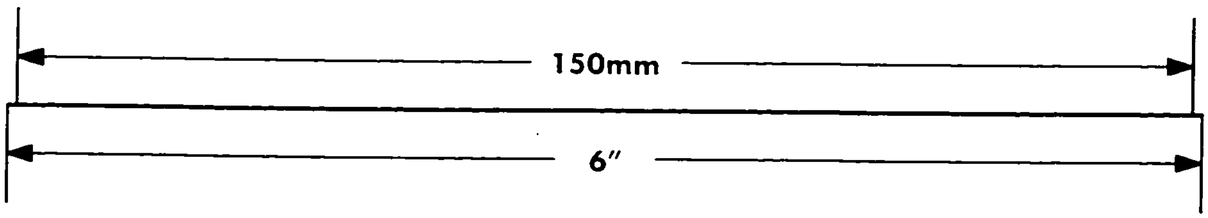
```
TACACTCTAT ATTTTTTTAT GCCTCGGTAA TGATTTTCAT TTTTTTTTTT


CCACCTAGCG GATGACTCTT TTTTTTTCTT AGCGATTGGC ATTATCACAT
 -59    -52
  |      |
AATGAATTAT ACATTATATA AAGTAATGTG ATTTCTTCGA AGAATATACT
       +1                            +23
        |                             |
AAAAAATGAG CAGGCAAGAT AAACGAAGGC AAAGATG
```

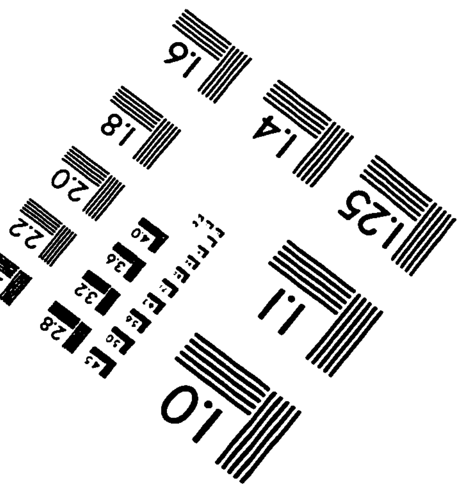
IMAGE EVALUATION TEST TARGET (QA-3)

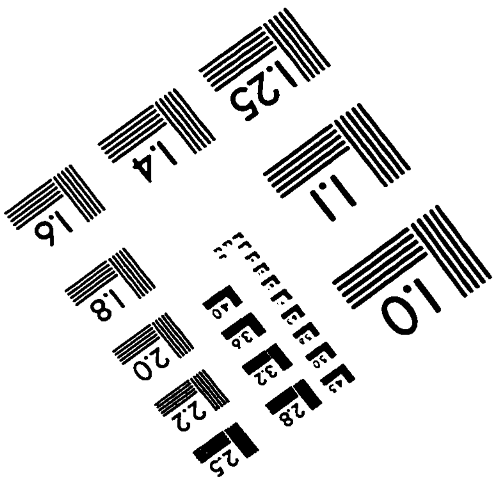












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989