STRESS-INDUCIBLE PROTEIN 1: A BIOINFORMATIC ANALYSIS OF THE HUMAN, MOUSE AND YEAST *STI1* GENE STRUCTURE

A research report submitted in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE
(in Bioinformatics and Computational Molecular Biology)

at
RHODES UNIVERSITY

by

Bronwen Louise Aken

February 2005

ABSTRACT

Stress-inducible protein 1 (Sti1) is a 60 kDa eukaryotic protein that is important under stress and non-stress conditions. Human Sti1 is also known as the Hsp70/Hsp90 organising protein (Hop) that coordinates the functional cooperation of heat shock protein 70 (Hsp70) and heat shock protein 90 (Hsp90) during the folding of various transcription factors and kinases, including certain oncogenic proteins and prion proteins. Limited studies have been conducted on the *STI1* gene structure. Thus, the aim of this study was to develop a comprehensive description of human *STI1* (*hSTI1*), mouse *STI1* (*mSTI1*), and yeast *STI1* (*ySTI1*) genes, using a bioinformatic approach. Genes encoded near the *STI1* loci were identified for the three organisms using National Centre for Biotechnology Information (NCBI) MapViewer and the *Saccharomyces* Genome Database. Exon/intron boundaries were predicted using Hidden Markov model gene prediction software (HMMGene) and Genscan, and by alignment of the mRNA sequence with the genomic DNA sequence. Transcription factor binding sites (TFBS) were predicted by scanning the region 1000 base pairs (bp) upstream of the *STI1* orthologues' transcription start site (TSS) with Alibaba, Transcription element search software (TESS) and Transcription factor search (TFSearch). The promoter region was defined by comparing the number, type and position of TFBS across the orthologous *STI1* genes. Additional putative TFBS were identified for *ySTI1* by searching with software that aligns nucleic acid conserved elements (AlignACE) for over-represented motifs in the region upstream of the TSS of genes thought to be co-regulated with *ySTI1*. This study showed that *hSTI1* and *mSTI1* occur in a region of synteny with a number of genes of related function. Both *hSTI1* and *mSTI1* comprised 14 putative exons, while *ySTI1* was encoded on a single exon. Human and mouse *STI1* shared a perfectly conserved 55 bp region spanning their predicted TSS, although their TATA boxes were not conserved. A putative CpG island was identified in the region from -500 to +100 bp relative to the *hSTI1* and *mSTI1* TSS. This region overlapped with a region of high TFBS density, suggesting that the core promoter region was located in the region approximately 100 to 200 bp upstream of the TSS. Several conserved clusters of TFBS were also identified upstream of this promoter region, including binding sites for stimulatory protein 1 (Sp1), heat shock factor (HSF), nuclear

factor kappa B (NF-kappaB), and the cAMP/enhancer binding protein (C/EBP). Microarray data suggested that *ySTI1* was co-regulated with several heat shock proteins and substrates of the Hsp70/Hsp90 heterocomplex, and several putative regulatory elements were identified in the upstream region of these co-regulated genes, including a motif for HSF binding. The results of this research suggest several avenues of future experimental work, including the confirmation of the proposed core promoter, upstream regulatory elements, and CpG island, and the investigation into the co-regulation of mammalian *STI1* with its surrounding genes. These results could also be used to inform *STI1* gene knockout experiments in mice, to assess the biological importance of mammalian *STI1*.

ACKNOWLEDGEMENTS

I would like to extend my thanks to the following individuals and organisations:

- Prof. Greg Blatch for supervising this work. His enthusiasm and guidance is greatly appreciated.
- Prof. Hugh Patterton, Dr Graeme Bradley, and Prof. Cathal Seioghe for their advice and input.
- Corné Schriek for downloading the required chromsosome sequences, and for his help in writing an algorithm in Python to calculate the base composition of these sequences.
- My family and friends, especially Nick, for their encouragement and support.
- The National Research Foundation and Rhodes University for their generous financial support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3’	-	3 prime (downstream of sense strand)
5’	-	5 prime (upstream of sense strand)
A	-	Adenine
Adr1	-	Alcohol dehydrogenase regulatory gene 1 product
AlignACE	-	Aligns nucleic acid conserved elements
ATP	-	Adenosine triphosphate
BAD	-	BCL2-antagonist of cell death
Bag-1	-	BCL2-associated athanogene 1
bp	-	Base pairs
C	-	Cytosine
C/EBP	-	CCAAT / enhancer-binding proteins
CBF	-	CCAAT-binding factor
Cdc2 kinase	-	Cell division cycle control protein 2 kinase
cDNA	-	Complementary deoxyribonucleic acid
CdxA	-	Caudal-related homeodomain transcription factor
c-Ets	-	Cellular E26 transformation specific sequence
CpG	-	Unmethylated CG dinucleotide
CRE-BP	-	Cyclic adenosine monophosphate response element-binding protein
c-Rel	-	Cellular reticuloendotheliosis proto-oncogene
DNA	-	Deoxyribonucleic acid
DNAJC4	-	DnaJ (Hsp40) homologue, Family C, Member 4
DPE		Downstream promoter element
FKBP2	-	FK506 binding protein 2
FKBP51	-	FK506 binding protein 51
FLRT1	-	Fibronectin leucine rich transmembrane protein 1
FN	-	False negative
FP	-	False positive
G	-	Guanine
GATA-1	-	GATA binding protein 1 (globin transcription factor 1)
Gcn2	-	General control kinase 2

Harc	-	Hsp90-associating relative of Cdc37
Hip	-	Hsp-interacting protein
HMMGene	-	Hidden Markov model gene-predicting software
HSE	-	Heat shock element
HSF	-	Heat shock factor
Hsp	-	Heat shock protein
Inr	-	Initatior region
kb	-	Kilobase pairs
LRP16	-	Low density lipoprotein-related protein 16
MGC11134	-	tRNA splicing 2' phosphotransferase 1
MGC13045	-	Hypothetical protein
mRNA	-	Messenger ribonucleic acid
N	-	Any nucleotide; adenine, cytosine, guanine or thymine
NCBI	-	National Centre for Biotechnology Information
NF-kappa B	-	Nuclear factor kappa B
Nkx-2.5	-	Homeobox protein NK-2 homolog E
NLS	-	Nuclear localisation signal
NNPP	-	Neural network promoter prediction
ORF	-	Open reading frame
OTUB1	-	OTU domain, ubiquitin aldehyde binding 1
Pbx-1	-	Pre-B-cell leukemia transcription factor 1
PLCB3	-	Phospholipase C beta 3
PPP1R14B	-	Protein phosphatase 1 regulatory (inhibitor) subunit 14B
Pu	-	Purine (A or G)
PWM	-	Position weight matrix
Py	-	Pyrimidine (C or T)
SGD	-	*Saccharomyces* Genome Database
Sp1	-	Stimulatory protein 1 (also known as Specific protein 1)
SRF	-	Serum response factor
SRY	-	Sex-determining region Y
STI1	-	Stress-inducible phosphoprotein 1
T	-	Thymine

TBP	-	TATA-binding protein
TESS	-	Transcription Element Search Software
TF	-	Transcription factor
TFBS	-	Transcription factor binding site
TFSearch	-	Transcription factor search
TN	-	True negative
TP	-	True positive
TPR	-	Tetratricopeptide repeat
TSS	-	Transcription start site
Ubx	-	Ultrabithorax
URP2	-	UNC-112 related protein 2
USF	-	Upstream stimulatory factor
UTR	-	Untranslated region
VEGFB	-	Vascular endothelial growth factor B
WT1	-	Wilms tumor suppressor gene zinc finger protein

Amino acid and nucleotide abbreviations are according to the IUPAC standards.

CHAPTER 1: INTRODUCTION

This study focuses on the stress-inducible protein 1 (Sti1), which is a 60 kDa co-chaperone found throughout the eukaryotes. Sti1 is abundant and predominantly cytosolic under non-stress conditions (Longshaw *et al.,* 2004). Within the cytosol, Sti1 co-ordinates the functional co-operation of the heat shock proteins Heat shock protein 70 (Hsp70) and Heat shock protein 90 (Hsp90) in the Hsp70/Hsp90 heterocomplex known to be involved in the folding of various transcription factors and kinases, including certain oncogenic proteins and prion proteins (Odunuga *et al.,* 2004). Details regarding Sti1 and its role within the cell are discussed in the first section of this chapter.

Transcription of the Sti1 gene can be upregulated following environmental stresses that cause an accumulation of denatured proteins, such as heat shock (Nicolet and Craig, 1989). Although the biological importance of Sti1 is not completely understood, it appears that this protein may play a critical role in eukaryotes in times of non-stress (Johnson *et al.,* 1998), and during environmental stress (Nicolet and Craig, 1989) and pathology (Honoré *et al.,* 1992). While some genetic studies have been conducted on the yeast *STI1* (Nicolet and Craig, 1989), the gene boundaries and genomic organisation of the gene encoding Sti1 in mice and humans have not been described.

The aim of this study, outlined in the third section of this chapter, was to add to our understanding of the gene encoding Sti1 by performing a preliminary analysis of the genomic organisation, gene structure and possible genetic elements involved in regulating human, mouse and yeast *STI1* expression at the transcriptional level. This was achieved using the available gene prediction and transcription factor identification programs, as described in the second section of this chapter and in Chapter 2. During the analysis of the data, emphasis was placed on results that correlated between the *STI1* homologues and also with a gene whose protein is known to bind and function with the Sti1 protein, *HSP70*. This approach was based on the premise that a protein known to interact with Sti1 was possibly also regulated by the same transcription factors. The

findings of this study and discussion thereof can be found in Chapter 3 and Chapter 4, respectively.

1.1 Sti1 And The Hsp70/Hsp90-Based Chaperone Machinery

Chapter 1 has been divided into three sections. This first section introduces cell stress and the need for chaperones, and specifically presents knowledge to date regarding the Sti1 protein, its role within the cell and the Hsp70/Hsp90 heterocomplex, and its shuttling between the nucleus and the cytosol.

1.1.1 Cell Stress And Chaperones

Cells respond to heat and other environmental stresses by producing a number of stress signals, and the result is a change in gene expression, protein localisation and post-translational modification (Georgopoulos and Welch, 1993). In general, protein synthesis is inhibited during cell stress. However, heat shock proteins (Hsps) are distinct in that their transcription and translation are rapidly upregulated (Morimoto, 1998): their abundance during cell stress protects cells from damage (Jolly and Morimoto, 2000).

Heat shock proteins are a family of conserved proteins, ubiquitous in both prokaryotes and eukaryotes, which function during stress and non-stress conditions. Certain Hsps are molecular chaperones that are capable of assisting specific non-native proteins in reaching their functional conformational state efficiently by preventing undesirable intra- and intermolecular interactions (Buchner, 1999). Hsps are not included in the active protein complexes that they help to assemble (Ellis, 1997). Probably the best characterised role of Hsps is that of steroid hormone receptor (SHR) maturation by the Hsp70/Hsp90 heterocomplex. A general model of SHR maturation is discussed later in this chapter.

Besides being transcriptionally regulated by cell stress, Hsps are also differentially regulated by cellular events such as cell growth, cell cycle progression and apoptosis. Consequently, if the operation of any of the chaperones or their cofactors is disturbed, disease may occur. Problems experienced by an organism in which chaperone pathways malfunction include poorly regulated physiologic processes, such as cell growth and cell death (Jolly and Morimoto, 2000).

1.1.2 Molecular Chaperones And Co-Chaperones

The process of SHR maturation is complex, requiring the co-ordinated interaction of a number of chaperones and co-chaperones (accessory proteins). This section comprises a brief description of some chaperones and co-chaperones involved in the Hsp70/Hsp90-based chaperone machinery.

Heat Shock Protein 90 (Hsp90)

Hsp90 is a highly conserved, ubiquitous adenosine triphosphate (ATP)-dependent molecular chaperone (Obermann *et al.,* 1998; Grenert *et al.,* 1999; Panaretou *et al.,* 1998), with homologues in both prokaryotes and eukaryotes (Bardwell and Craig, 1988; Buchner, 1999). It is one of the most abundant cellular proteins during non-stress conditions, comprising 1-2% of the total cytosolic protein (Pratt and Toft, 1997).

During non-stress conditions, Hsp90 is involved in the *in vivo* maturation and transport of diverse proteins including certain protein kinases and transcription factors (Pratt and Toft, 1997). Under stress conditions, Hsp90 expression is upregulated; Hsp90 becomes involved in maintaining the protein homoeostasis within the cell and it is also implicated in autoregulation of Hsp transcription (Morimoto, 1998; Zou *et al.,* 1998).

Transcription of mammalian Hsp genes, such as Hsp90, is regulated by the heat shock promoter element (HSE), which consists of multiple adjacent inverted repeats of the motif 5'-nGAAn -3' (Trinklein *et al.,* 2004). The HSE is bound and activated by a transcription factor, heat shock factor 1 (HSF1). During non-stress conditions, the leucine

zipper (Pirkkala *et al.,* 2001) HSF1 is bound by Hsp90 and is thus unable to bind the HSE and promote transcription (Morimoto, 1998). During stress conditions, the proportion of unfolded proteins in the cell increases and these unfolded proteins compete with HSF1 for binding to Hsp90. With fewer Hsp90 molecules available to bind HSF1, the HSF1 trimerises and binds to the HSE, and is thus able to initiate the stress response and upregulate Hsp transcription via re-organisation of chromatin and possibly also by interaction with another transcription factor, the TATA-binding protein (TBP) (Bharadwaj *et al.,* 1999; Mason and Lis, 1997; Morimoto, 1998; Pirkkala *et al.,* 2001; Zou *et al.,* 1998).

Besides HSF1, mammals are known to synthesise two other HSFs, HSF2 and HSF4, whereas organisms such as yeast, *Caenorhabditis elegans* and *Drosophila melanogaster* only manufacture one HSF. Human HSF1, the major stress-induced HSF, is ubiquitous and functionally equivalent to yeast and *Drosophila* HSF (Fernandes *et al.,* 1994; Morimoto, 1998). HSF2 is active during development and during the inhibition of ubiquitin-dependent proteasome. HSF4 is tissue-specific and is able to inhibit stress-induced gene expression. Another HSF exists that is activated with HSF1, HSF3, but it has only been observed in avian specimens. (Morimoto, 1998)

Heat Shock Protein 70 (Hsp70)

As with Hsp90, Hsp70 is also upregulated during cell stress where it becomes the most abundant protein in the cell (Lindquist, 1986; Nicolet and Craig, 1989). Hsp70 is the most conserved of the Hsps, with homologues in eukaryotes and prokaryotes (Georgopoulos and Welch 1993). This ATP-dependent chaperone is responsible for recognising and folding nascent polypeptides (Ellis, 1999) and denatured proteins (Frydman, 2001; Gebauer *et al.,* 1997); it has anti-apoptotic activity (Wei *et al.,* 1995); and is also involved in protein transport (Pratt, 1993) and autoregulation of the heat shock response (Morimoto, 1998). With respect to the latter, Hsp70 increases the rate of HSF1 deactivation during recovery from stress (Bharadwaj *et al.,* 1999) and thus restrains HSF1-mediated transcription.

Hsp70 plays a role in cell cycle progression, and human Sti1 and Hsp70 are reported to activate histone transcription during S-phase (Zheng *et al.,* 2003). Clearly, the proteins involved in maturation of SHRs carry out other functions within the cell besides SHR maturation. This extended functioning is admissible for proteins found at such high concentrations within the cell.

Co-Chaperones

A complete description of the Hsp70/Hsp90-based chaperone machinery has not yet been achieved, yet several functional co-chaperones have been identified. These auxiliary proteins include Hsp40, Hsp-interacting protein (Hip), BCL2-associated athanogene 1 (Bag-1), Sti1 and immunophilins (Smith, 2000).

1.1.3 The Sti1 Co-Chaperone

As the focus of this report will be to characterise the genomic organisation and promoter elements of *ySTI1*, *mSTI1* and *hSTI1*, this next section is devoted entirely to the Sti1 co-chaperone and its role within the cell. The human, mouse and yeast orthologues of Sti1 will be referred to as hSti1, mSti1 and ySti1 respectively. In addition, *STI1* will refer to the gene encoding stress-inducible protein 1, and Sti1 will refer to the protein itself. The uppercase lettering, used to refer to the gene, makes no reference to genetic (allele) dominance but is merely the nomenclature chosen for this report to discriminate between the gene and protein.

The Sti1 Protein

Yeast Sti1 was first described by Nicolet and Craig (1989) as a heat- and canavanine-inducible protein that activates transcription of *lacZ* under the control of the Hsp70 (*SSA4*) promoter. Subsequently, ySti1 homologues were recovered in other eukaryotes such as human (Honoré *et al.,* 1992), mouse (Lässle *et al.,* 1997), rabbit (Gross and Hessefort, 1996), chicken oviduct (Smith *et al.,* 1993) and soybean (Zhang *et al.,* 2003).

A *STI1* pseudogene has also been identified on human chromosome X (Odunuga *et al.,* 2004).

Human Sti1 is also known as the Hsp70/Hsp90 organising protein (Hop), p60, extendin, IEF SSP 3521 and RT-Hsp70. A preliminary analysis has revealed 14 exons for *hSTI1* (Odunuga *et al.,* 2004), although the gene boundaries of *STI1* have not been defined and genetic regulatory elements remain to be identified (Figure 1).

The primary structure of Sti1 has a high level of similarity amongst eukaryotes (Lässle *et al.,* 1997) and contains multiple copies of a loosely conserved sequence, known as the

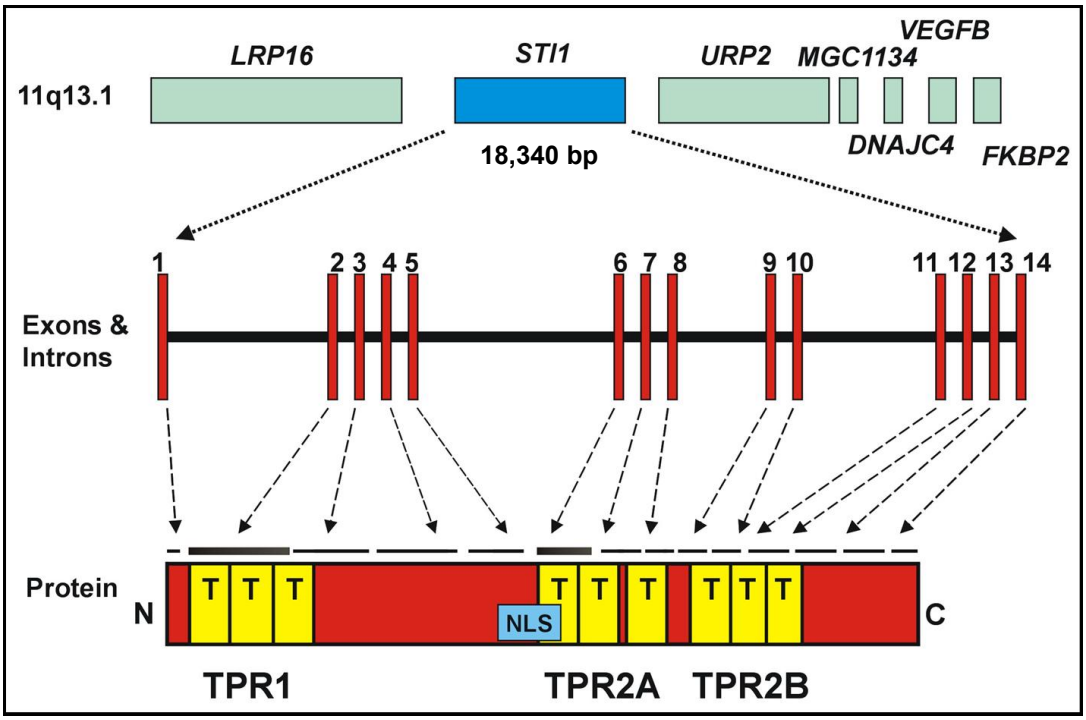

Figure 1: Genomic organization of human *STI1*. The chromosomal region 11q13.1 comprises a number of putative genes, including Low density lipoprotein-related protein 16 (*LRP16*), *STI1*, UNC-112 related protein 2 (*URP2*), tRNA splicing 2' phosphotransferase 1 (MGC1134), DnaJ (*HSP40*) homologue Subfamily C Member 4 (*DNAJC*4), vascular endothelial growth factor B (*VEGFB*), and FK506-binding protein 2 (*FKBP2*). The *hSTI1* transcribed region is 18,340 base pairs (bp) in length and consists of 14 exons (indicated by red bars). The corresponding hSti1 protein has three tetratricopeptide (TPR) domains, with TPR1 near the NH2-terminal (N) of the protein and TPR2A and TPR2B closer to the COOH-terminal (C) of protein. TPR motifs fully encoded entirely on one exon are indicated by a bold line above and it can be seen that none of the three TPR domains is encoded on a single exon. The nuclear localization signal (NLS) has also been indicated. Adapted from Odunuga *et al.,* 2004.

tetratricopeptide repeat (TPR). The helix-turn-helix TPR motif folds into an amphipathic channel or binding groove, allowing protein-protein interactions with the complementary region of target proteins such as Hsp90 and Hsp70 (Blatch and Lässle, 1999; Scheufler *et al.,* 2000). Specifically, structural experiments have shown that the N-terminal TPR1 (Figure 1) domain of Sti1 interacts with the C-terminal PTIEEVD motif of Hsp70, and a C-terminal TPR2A domain of Sti1 interacts with the C-terminal MEEVD of Hsp90. Protein interactions occur via electrostatic and hydrophobic contacts (Scheufler *et al.,* 2000).

Mouse Sti1 is cytosolic under non-stress conditions (Lässle *et al.,* 1997), yet trafficking between the cytosol and nucleus can occur (Longshaw *et al.,* 2004). This shuttling is most probably the result of differential phosphorylation due to cell cycle status and cell cycle kinases (Longshaw *et al.,* 2004): Stress has been shown to change the isoform composition of Sti1 proteins in both human (Honoré *et al.,* 1992) and mouse cells (Lässle *et al.,* 1997), possibly due to differential phosphorylation. Indeed, a putative bipartite nuclear localisation signal (NLS) has been identified at positions 222-239 in mSti1 (Blatch *et al.,* 1997) and putative phosphorylation sites have been identified at positions S189 (casein kinase II, CKII) and T198 (cell division cycle control protein 2 kinase, cdc2 kinase), just upstream of the NLS. Together with the NLS, the phosphorylation sites form a predicted casein kinase II - cdc2 kinase-NLS (CcN) motif at position 180-239.

CKII is important in the transition of cells from G0 phase to G1 phase of cell cycle, and for continuing passage through early G1 phase, and it has been proposed that phosphorylation of the Sti1 casein kinase II (CKII) site promotes relocation of Sti1 to the nucleus. Additionally, phosphorylation of the cdc2 site during the G1/S phase in the cell cycle promotes cytosolic localisation of Sti1 (Figure 2). Thus, it is speculated that the localisation of Sti1 may be cell cycle- dependent and that the NLS and phosphorylation sites could contribute to the shuttling of Sti1 between the cytosol and nucleus (Jans and Jans, 1994; Longshaw *et al.,* 2004).

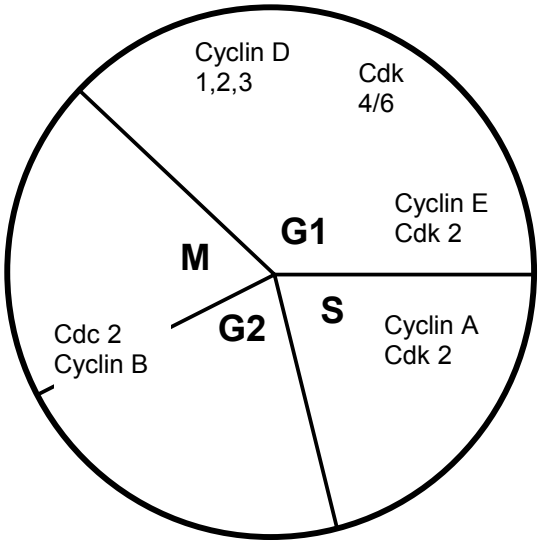

Figure 2: The cell cycle. The cell cycle consists of four main stages: Growth 1 (G1), Stationary (S), Growth 2 (G2), and Mitosis (M). Each stage is characterised by the action of a specific set of proteins, as indicated. These include cyclin dependent kinases (cdk) and cyclins.

The Role Of Sti1 In The Cell

Sti1 is a transient member of the Hsp70/Hsp90-based chaperone machinery with no apparent chaperone function of its own (Scheibel and Buchner, 1998). The role of Sti1 in the Hsp70/Hsp90 heterocomplex is well-studied and it seems that Sti1 optimises the functional co-operation of the Hsp70/Hsp90 complex. The concept of Sti1 functioning in cell proliferation and gene regulation, which requires that Sti1 be located in the nucleus, is less well studied than its functioning in the cytosolic Hsp70/Hsp90-based chaperone heterocomplex.

The Hsp70/Hsp90 heterocomplex is known to be involved in the maturation of a number of proteins including transcription factors such as HSF1 (Nadeau *et al.,* 1993), the glucocorticoid receptor (Sanchez *et al.,* 1985), progesterone receptor (Catelli *et al.,* 1985), estrogen receptor (Catelli *et al.,* 1985; Joab *et al.,* 1984), and E12 (Shue and Kohtz, 1994). Other Hsp90 substrates include the Hepatitis B virus reverse transcriptase (Hu and Seeger, 1996), Hsp90-associating relative of Cdc37 (Harc) (Scholz *et al.,* 2001), and general control kinase 2 (Gcn2) (Donzé and Picard, 1999). Sti1 is also able to interact with proteins independently of Hsp90, as has been noted for Hsp104 (Abbas-

Terki *et al.,* 2001) and the cell division cycle control protein 37 (Cdc37) (Abbas-Terki *et al.,* 2002).

In Vitro Model Of Steroid Hormone Receptor Activation In Vertebrates

The section summarises the maturation of steroid hormone receptors by the Hsp70/Hsp90 heterocomplex. Unless otherwise specified, Sti1 refers to vertebrate Sti1.

Large proteins struggle to achieve their native conformation without the assistance of chaperones, such as Hsp70 and Hsp90, to prevent their aggregation (Nathan *et al.,* 1997; Ellis and Hartl, 1999; and Houry, 2001). The experimental system commonly used to investigate the Hsp70/Hsp90 heterocomplex assembly has involved glucocorticoid (Dittmar *et al.,* 1998) or progesterone receptor maturation in cell-free rabbit reticulocyte lysate (Smith *et al.,* 1990; Nathan *et al.,* 1999). Under these conditions, chaperone-facilitated maturation of the steroid hormone receptors (SHRs) is thought to resemble *in vivo* interactions accurately (Smith *et al.,* 1992; Cheung and Smith, 2000) and to approximate the maturation of other substrate proteins (Nair *et al.,* 1996).

Five proteins are required for *in vitro* formation of the core SHR-Hsp90 heterocomplex: Hsp70, Hsp90 and the three accessory proteins Sti1, Hsp40 and p23 (Kosano *et al.,* 1998; Dittmar *et al.,* 1998). Partner proteins and basic functioning of the Hsp70/Hsp90-based multiprotein chaperone complex seem to be conserved across eukaryotes (Stancato *et al.,* 1996; Buchner, 1999; Smith *et al.,* 1993; Chang and Lindquist, 1994). In addition to these five proteins, several nonessential proteins are associated with *in vitro* heterocomplexes, including Hip, Bag1, and the immunophilins FK506-binding proteins 51 (FKBP51), FKBP52 and cyclophilin 40 (Owens-Grillo *et al.,* 1995; Nair *et al.,* 1997). Heterocomplex assembly and concomitant SHR maturation is a multi-step process driven by (i) ATP/ADP exchange, causing Hsp70 and Hsp90 conformational changes (Cheetham *et al.,* 1994; Grenert *et al.,* 1997), and (ii) competitive binding of TPR proteins to Hsp70 and Hsp90 (Young *et al.,* 1998; Prodromou *et al.,* 1999; reviewed in Kimmins and MacRae, 2000). Refer to Figure 3.

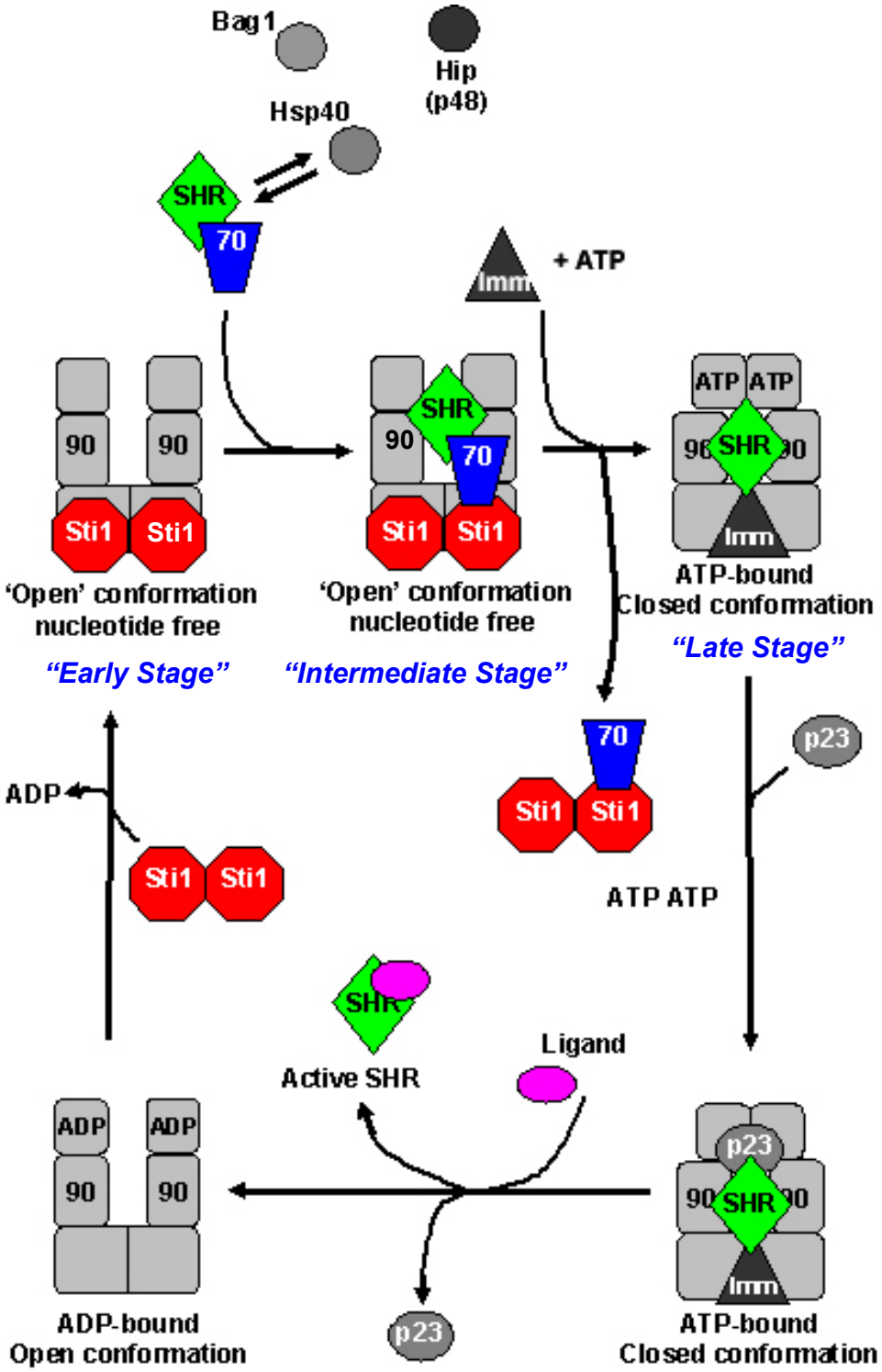

Figure 3: Activation of steroid hormone receptor by the Hsp70/Hsp90 heterocomplex, illustrating the transient binding of Sti1 to the heterocomplex. The Sti1 protein has been drawn as a dimer, although there is evidence that the protein can also act as a monomer (van der Spuy *et al.,* 2001). Abbreviations include: 90 (heat shock protein 90), 70 (heat shock protein 70), Hsp40 (heat shock protein 40), Hip1 (Hsp-interacting protein), SHR (steroid hormone receptor), Imm (immunophilin), Bag 1 (BCL2-associated athanogene 1), p23, ADP (adenosine diphosphate), and ATP (adenosine triphospahte). Adapted from Prodromou and Pearl, 2003.

The sequential cycle of SHR activation can be divided into 3 basic stages (Smith *et al.,* 1995; Dittmar *et al.,* 1996; Smith, 2000). In the early stages of heterocomplex assembly, Hsp70-ATP is linked to an inactive SHR. Peptide and Hsp40 binding stimulate ATP-/K^+-dependent hydrolysis of Hsp70-bound ATP (Flynn *et al.,* 1989; Cyr *et al.,* 1992) and a consequent conformational change of Hsp70 (Cheetham *et al.,* 1994). Hip is recruited to stabilize the complex (Höhfeld *et al.,* 1995).

The intermediate stage incorporates Hsp90 and Sti1 to the Hsp70-SHR complex. *In vivo*, there is evidence that eukaryotic Sti1 and Hsp90 are pre-associated (Chang and Lindquist, 1994; Chang *et al.,* 1997), with *in vitro* experiments showing that Sti1 preferentially binds Hsp90-ADP (Johnson *et al.,* 1998). Binding of Hsp90-Sti1 to the Hsp70-SHR complex constructs a "foldosome" (Hutchison *et al.,* 1994) with Hsp90 contacting the ligand-binding domain of the SHR (Pratt and Toft, 1997). Sti1 mediates Hsp70-Hsp90 interaction *in vitro* by establishing a bridge between Hsp90 and Hsp70, allowing them to communicate indirectly (Smith *et al.,* 1993; Chen *et al.,* 1996; Lässle *et al.,* 1997; Chen and Smith, 1998; Johnson *et al.,* 1998). Thus, it is possible that Sti1 controls certain steps within the SHR-heterocomplex assembly.

The third stage of chaperone-facilitated SHR activation is characterised by the binding of p23 and immunophilins to Hsp90 (Johnson and Toft, 1995). In order to achieve this final state, Hsp90 must exchange ADP for ATP and consequently adopt its ATP-dependent conformation. The steroid-binding cleft of the SHR is opened by Hsp90-ATP (Grenert *et al.,* 1999). Dissociation of Sti1 leaves Hsp90 free to bind p23 (Dittmar *et al.,* 1997) and immunophilins. The binding of Sti1 and p23 is mutually exclusive (Johnson *et al.,* 1998). Depending on the target protein, different immunophilins seem to bind Hsp90; their binding is also prevented while Sti1 is bound to Hsp90 (Ratajczak and Carrello, 1996; Nair *et al.,* 1997; Pratt *et al.,* 1999). Hsp70 may or may not be present in the final chaperone-SHR complex (Smith, 1993; Pratt and Toft, 1997).

It has been suggested that Hsp90, in a complex with immunophilins and the SHR, can act as a "transportosome" to convey the SHR between and within the cytosol and nuclear compartments, *i.e.* retrograde movement (Pratt, 1993; DeFranco, 2000; Pratt and Toft, 2003; Pratt *et al.,* 2004). Although SHRs may be folded within the cytosol, their site of action is the nucleus, where they function as transcription factors (TFs). As long ago as 1989, it was documented that vertebrate Hsp90 holds its substrate proteins in an inactive state until they are activated by ligand binding or until they have arrived at the appropriate intracellular location and are released by the chaperone complex (Smith, 2000).

Hsp90-ATP hydrolysis and ligand binding release p23 and the active SHR (Obermann *et al.,* 1998; Scheibel *et al.,* 1998; Smith *et al.,* 1992; Smith *et al.,* 1998). The role of the chaperone heterocomplex is complete and the constituent proteins are available for reuse.

1.2 Computational Molecular Biology

Following a brief introduction to cell stress and chaperones, emphasising the Sti1 co-chaperone and its role within the cell, the next section of this chapter will focus on an altogether different subject: that of computational molecular biology. The section to follow introduces key concepts regarding transcription in eukaryotes, before discussing *in silico* gene and promoter prediction methods.

1.2.1 Transcription In Eukaryotes

Gene expression in eukaryotes is a complex process that is fine-tuned to suit the spatial, temporal and environmental requirements of the cell. While gene expression may refer to the expression of the final protein product, the focus of this review is on transcription and its regulation.

Transcription refers to the process whereby the double-stranded deoxyribonucleic acid (DNA) blueprint of a gene is converted into a single stranded messenger ribonucleic acid (mRNA) by RNA polymerase II. In eukaryotes, the mRNA is post-transcriptionally

modified and then translated into protein (Figure 4). Gene expression is controlled, *inter alia,* through: (i) dense packaging of chromatin; (ii) transcription initiation via assembly of RNA polymerase II and binding of transcription factors (TFs) at the core promoter; (iii) enhancer elements; (iv) CpG islands; (v) alternative splicing; (vi) polyadenylation; and (vii) translation initiation (Fickett and Hatzigeorgiou, 1997; Johansson *et al.,* 2003; Lareau *et al.,* 2004; Pedersen *et al.,* 1999).

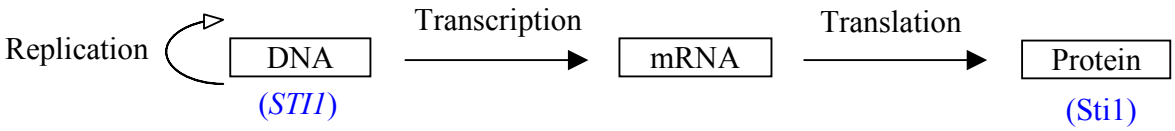

Figure 4: Flow of genetic information. In order to obtain the protein Sti1, the *STI1* gene must be transcribed to mRNA and then this mRNA must be translated into a protein.

In order for transcription to begin, the RNA Polymerase II holoenzyme and a number of accessory proteins, called general TFs, must bind to the core promoter that spans the transcription start site (TSS) and comprises the TATA box, initiation region (Inr) and downstream promoter element (DPE) (Fukue *et al.,* 2004). In addition, regulatory proteins may bind to the proximal promoter, upstream of the core promoter, at the CCAAT box and the GC box (Figure 5). Transcription resulting from the binding of only the minimal components to the core promoter is called basal transcription and is uncommon *in vivo* (Pedersen *et al.* 1999).

In contrast to basal transcription, the activated transcription that occurs *in vivo* is tightly regulated by the binding of additional TFs to additional *cis*-regulatory elements. Bound TFs mediate their regulatory effects by interacting with the basal transcription complex via protein-protein interactions. This is made possible by the DNA strand bending back on itself (Werner *et al.,* 2003). Regulatory elements can be found several kilobases (kb) away up- or downstream from the TSS and include both enhancers, that activate transcription, and silencers, that repress transcription (Pedersen *et al.,* 1999).

The type of TF that binds a DNA sequence will also influence whether transcription will be repressed or stimulated, and to what extent. Some TFs binding upstream of a gene may be nonspecific, whilst others may be specific for the gene (Nikolov and Burley, 1997) and the conditions under which transcription must occur. The binding of TFs to DNA is influenced by the degeneracy of the transcription factor binding site (TFBS) and

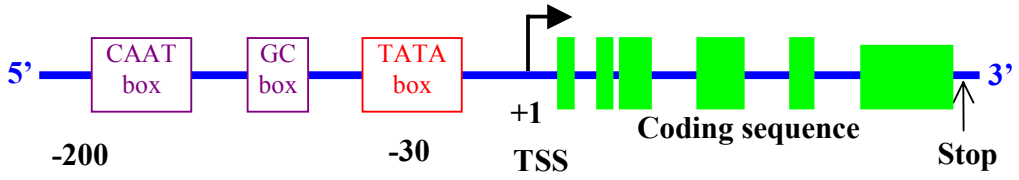

Figure 5: Example of a gene, showing coding sequence and promoter elements. In vertebrates, genes are a mosaic of exons (coding, green blocks) and introns (non-coding). Both introns and exons are transcribed into RNA and the pre-mRNA is spliced to remove the intervening intron sequences and produce mature mRNA. Upstream (5') of the transcriptional start site (TSS) is the promoter region, a DNA sequence which is recognised by RNA polymerase II.

the availability of the TFs. For example, spatial, temporal and environmental conditions may alter the types of TFs that are available to the cell. A specific combination of TFs bound upstream of a gene may thus modulate transcription in a different manner than that of a different combination of TFs bound upstream of the same gene (Gailus-Durner *et al.,* 2001; Pedersen *et al.,* 1999).

1.2.2 Prediction Of Gene Structure And Regulatory Elements

A wide choice of gene and promoter prediction programs is available on the World Wide Web (Mount, 2001). A number of relatively robust gene prediction programs are available (Rogic *et al.,* 2001) yet, unfortunately, none of the available promoter prediction programs has proven to be significantly superior to the others (Fickett and Hatzigeorgiou, 1997).

Gene Prediction Programs

"*Difficulty in deciphering the anatomy of mammalian genes is due to several factors, including large amounts of intervening (non-coding) sequence, the imperfection of gene prediction algorithms, and the incompleteness of cDNA-sequence resources, many of*

which consist of gene tags of variable length and quality. Full-length cDNA sequences are extremely useful for determining the genomic structure of genes, especially when analysed within the context of genomic sequence" (Strausberg *et al.,* 2002).

Aligning a full-length gene sequence with its complementary DNA (cDNA), the product of reverse transcribing mRNA to DNA, allows prediction of exon/intron boundaries and the promoter region. However, cDNA is not always available and in these cases it is necessary to turn to prediction programs in order to determine the exon/intron boundaries of a gene.

The available prediction programs use different algorithms and different training sets and are thus suitable for different query sequences. For example, some programs are optimised for prokaryotic sequences whilst others are optimised for human genes. Yet other programs may be suitable for long multigenic DNA sequences (Scherf *et al.,* 2000) or sequences where only one gene is expected (Hutchinson, 1996). It is preferable to use programs that read the complementary strand because TFBS on the complementary strand may influence transcription on the coding strand (Pedersen *et al.,* 1999).

Gene prediction programs make use of three general methods for promoter prediction. These include (i) signal or site-based searching, (ii) content-based searching, or (iii) homology searching. Signal searching involves searching for transcriptional and translational start and stop sites, exon/intron boundaries, and TFBS, while content-based searching looks at more general trends such as codon usage, word frequencies and periodicities, GC content and CpG islands. Homology searching requires that the query gene has a well-annotated homologous gene from which information can be deduced. Promoter prediction programs generally use a combination of these methods (Table 1).

Table 1: Classification of some web-based prediction programs that can be used to identify promoters and *cis*-elements.

Approach	Method and Output	Discrimination Method	Example	Web address	Reference
Database similarity search	Query DNA is scanned against a transcription factor database for putative transcription factor signals. A list of putative transcription factor binding sites is returned.	Consensus sequence or position weight matrix	TESS	http://www.cbil.upenn.edu/ tess	Schug and Overton (1997a,b)
			TFSearch	http://www.cbrc.jp/research/db/TFSEARCH.html	Heinemeyer *et al.* (1998)
		Known TF binding sites are used to make a temporary matrix	Alibaba	http://www.gene-regulation.com/pub/programs/alibaba2/index.html	Grabe (2002)
Statistical pattern recognition	Query DNA is scanned for promoter patterns. A promoter pattern profile has been constructed by analysis of a training set of promoters. A transcription start site or promoter region is returned.	Linear function	PromoterScan	http://cbs.umn.edu/software/proscan/promoterscan.htm	Prestridge (1995)
		Neural network	NNPP	http://www.fruitfly.org /seq_tools/promoter.html	Waibel *et al.* (1989), Reese and Eeckman (1995), Reese (2000), Reese (2001)
		Nonlinear function	CorePromoter	http://rulai.cshl.org/tools/genefinder/CPROMOTER/index.htm	Zhang (1998)
Combination of above	Query DNA is searched for both transcription factor binding sites and promoter patterns. A promoter region and putative transcription factors are returned.	Linear function	TSSG	http://genomic.sanger.ac.uk/gf/gf.shtml	Solovyev and Salamov (1997)
Search for user-defined promoter	Query DNA is scanned for a user-defined promoter pattern profile.	User-defined	FastM	http://genomatix.gsf.de/cgi-bin/fastm2/fastm.pl	Klingenhoff *et al.* (1999)
Phylogenetic footprinting	Co-regulated or orthologous genes are compared to find over-represented motifs. Putative regulatory motifs are returned.	Gibbs sampling algorithm	AlignACE	http://copan.cifn.unam.mx/Computational_Biology/yeast-tools	Roth *et al.* (1998), Hughes *et al.* (2000)
Alignment with gene message	Query DNA gene structure and transcription start site is determined by alignment of mRNA or ESTs to genomic DNA.	Alignment	GrailEXP	http://compbio.ornl.gov/grailexp/	Hyatt *et al.* (2000a,b), Xu *et al.* (1999)

Promoter And Transcription Factor Binding Site Prediction Programs

Careful evaluation is required when choosing which programs to use for *in silico* promoter predictions. In general, it is best to use more than one program to search for the same elements, in the hope that one program's strengths will compensate for another's weaknesses. Promoter prediction programs can be compared by calculating their sensitivity and specificity using the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) predictions. Sensitivity is defined as TP/(TP+FN) and specificity is defined as TP/(TP+FP) (Mount, 2001). A study conducted by Fickett and Hatzigeorgiou (1997) compared nine promoter prediction programs and found that only 13% to 54% of the true promoters were identified. They also found that, on average, one false positive promoter was predicted per 1000 bp – a specificity of between 11% and 22%. Often sensitivity is sacrificed in order to improve specificity, or *vice versa*.

The lack of correlation between predicted elements and actual promoter elements is largely due to an incomplete understanding of transcription and its regulation. The prediction of biologically functional TFBS in eukaryotes is problematic because they are flexible, relatively common by chance, scattered over large distances, and active in both orientations (Pedersen *et al.,* 1999). The quality of gene prediction also depends on the quality of sequencing and thus sequences with a high number of sequencing errors are prone to prediction errors. Hence, even though the promoter prediction algorithms have been carefully developed, computed results are not considered reliable until proven experimentally (Fickett and Hatzigeorgiou, 1997). Thus, the more consistent the predictions are between programs, the more accurate a prediction can be presumed to be.

Mount (2001) describes six methods of promoter prediction. Promoter prediction methods one to six are characterized by Neural Network Promoter Prediction (NNPP), PromoterScan II, TSSG, CorePromoter, Transcription Element Search Software (TESS), and FastM, respectively. An overview of the main methods behind promoter prediction and *cis*-regulatory element identification follows, using the headings in the

first column of Table 1 as a guide. Details of the specific programs used, can be found in Chapter 2.

Database similarity searches (Table 1), such as performed by Transcription Element Search Software (TESS) and TFSearch, involve searching the query sequence against a database of known promoter elements (TFBS), defined by consensus IUPAC (International Union of Pure and Applied Chemistry) codes or position weight matrices (PWMs). Unlike IUPAC codes, PWMs make use of nucleotide position information by considering the probability of all four nucleotides at each position, thus making it possible to allocate a higher score to a match at a frequently conserved position than to a match at a position that is observed to be conserved less frequently. PWMs are considered more sophisticated (Frech *et al.,* 1997) than consensus sequences because they make use of this additional information (Fickett and Hatzigeorgiou, 1997; Jensen and Liu, 2004; Lavorgna *et al.,* 1998; Quandt *et al.,* 1995) and can therefore provide a score for a match that gives an indication of the probability that a TF will bind to the sequence. This is useful, as most TFs are able to bind to a range of similar sequences that deviate from the consensus. How degenerate the binding site is from the consensus will affect how tightly a particular TF is able to bind to the regulatory element. The TFBS search programs' output is a list of the specific TFBS that are found to match the query DNA sequence. Thus, these programs are useful in identifying TFs that may bind to the query DNA sequence but do not predict the promoter region *per se*.

'Statistical pattern recognition' methods of promoter-finding (Table 1), as the name suggests, rely on assessing the significance of a promoter pattern found in the query sequence with regard to the training set of sequences. Discrimination between promoter and non-promoter regions, on the basis of features such as oligonucleotide frequencies and transcriptional signals, (Fickett and Hatzigeorgiou, 1997; Frith *et al.,* 2001; Hannenhalli and Levy, 2001; Prestridge, 1995), is advantageous because the method does not require knowledge of specific TFBS. Instead, a suitable training set of characterized promoters must be available. Prediction of promoter regions by neural networks has produced competitive results in promoter prediction (Fickett and Hatzigeorgiou, 1997). A second method of statistical promoter prediction is to assign a two-variable score to a DNA window and then employ a linear or non-linear

function to decide whether the DNA falls within a promoter or non-promoter region (Figure 6): the function forms a threshold, pre-determined by the training set, between the promoter and non-promoter regions, and depending on which side of the function line the window is placed it is classified as being a promoter or non-promoter region (Zhang, 1997). Markov models can be also be used to predict promoters but are more commonly employed in gene-finding algorithms (Burge and Karlin, 1997).

User-defined promoter regions (Table 1) can also be searched for, such as for the FastM (Klingenhoff *et al.,* 1999) program. FastM requires input from the user such as which elements (e.g. transcription factors and repeats) to search for, a range of how far apart these elements can be, their sequential order, and their strand orientation. A model is then built that abides by these parameters, and this model is scanned for in query DNA sequences.

Phylogenetic footprinting (Table 1) can be useful in recognising regulatory elements. Phylogenetic footprinting refers to comparing evolutionarily-related sequences and can thus only be employed when such related sequences are known. When the upstream regions of orthologous or co-regulated genes are locally aligned, areas of sequence conservation may be detected (Fessele *et al.,* 2002; Grad *et al.,* 2004). These conserved sequences are most likely to have a biological role such as that of binding specific TFs.

Lastly, aligning a full-length gene sequence with its cDNA, the product of reverse transcribing mRNA to DNA, allows prediction of gene structure and the promoter region. Additionally, CpG islands have been used to identify promoters in vertebrates (Hyatt *et al.,* 2000a, 2000b; Xu *et al.,* 1999) as they are found associated with the 5' end of a gene. A CpG island is a region of greater than 200 bp in length where the GC content is at least 50%, and where the observed frequency of CG dinucleotides / expected frequency of CG dinucleotides is greater than 0.6 (Gardiner-Garden and Frommer, 1987). In general, the observed occurrence of the CG dinucleotides in vertebrate genomes is approximately 0.008, significantly lower than the expected frequency, because the methylated cytosine in the CG dinucleotide is deaminated to be come thymine; thus, the proportion of CG dinucleotides increase and the proportion of TpG dinucleotides increase over time. CpG islands therefore

refer to a section of DNA with CG dinucleotides occurring at a higher frequency than in the rest of the genome (Bird *et al.,* 1987), and these CG dinucleotides are unmethylated (Gardiner-Garden and Frommer, 1987).

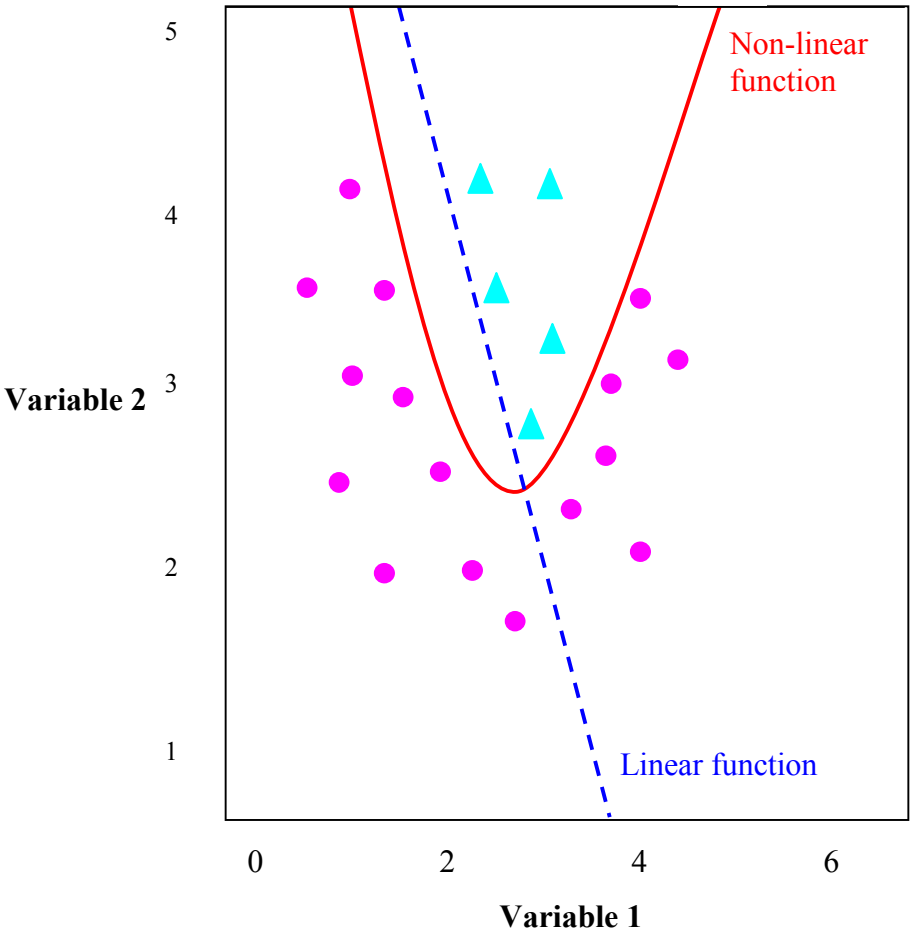

Figure 6: Two-variable example in which a linear function (dotted line) and non-linear function (solid line) separate the promoter (▲) and non-promoter (●) windows of DNA. Adapted from Zhang (1997).

1.3 Problem Statement And Research Objectives

With the preceding introductory sections in mind, the remainder of this chapter outlines the problem statement and associated research objectives for this study.

1.3.1 Problem Statement And Hypotheses

While some work has been done on yeast (*Saccharomyces cerevisiae*) *STI1* (Nicolet and Craig, 1989), neither mouse (*Mus musculus*) nor human (*Homo sapiens*) *STI1* has been characterised in detail with respect to promoters, transcription initiation and termination sites, or regulatory genetic elements. In addition, no work has been done to date with regard to comparing the three abovementioned orthologues at the genetic level. It is possible that the yeast, mouse and human *STI1* orthologues have withstood evolutionary pressure since they diverged from a common ancestor, and that they still have biologically important elements of their gene structure in common.

It is thus hypothesised that:

(i) The basic gene structure of mouse and human *STI1* is conserved.

(ii) Yeast, mouse and human *STI1* orthologues will employ similar transcription factor binding sites and promoter regions, as they are required by the cell under similar conditions.

(iii) Genes co-regulated with *ySTI1* will share common transcription factor binding sites with *ySTI1*.

1.3.2 Research Objectives

The aim of this project was to develop a complete bioinformatic description of the genes encoding Sti1 in humans, mouse and yeast. This was achieved using appropriate prediction programs available on the internet. The following specific objectives were set:

(i) To predict the chromosomal position, extent and direction of the *STI1* orthologues, using National Centre for Biotechnology Information (NCBI) MapViewer and the *Saccharomyces* Genome Database, and also to identify genes

of interest found to occur in close proximity to the *STI1* genes. Additionally, it was intended to gain insight into the *STI1* gene structure (including transcription start and stop sites, introns, exons, and polyadenine tail) using Genscan and the Hidden Markov model gene predictor (HMMGene), and also by aligning the full-length DNA with the corresponding mRNA sequence.

(ii) To predict the promoter region and regulatory elements of the *STI1* orthologues, and also to discover over-represented motifs in genes co-regulated with *ySTI1*. This was to be achieved using a variety of prediction programs:

- CorePromoter, Neural Network Promoter Prediction, PromoterScan and TSSG for promoter prediction;
- Alibaba, TESS and TFSearch for *cis*-regulatory elements; and
- Aligns nucleic acid conserved elements (AlignACE) for prediction of over-represented motifs.

(iii) To collate the data and to map the *STI1* gene structure and conserved regulatory elements

CHAPTER 2: RESEARCH APPROACH AND METHODOLOGY

This chapter outlines the methods used to achieve the objectives stated in the previous chapter, and has been divided into three sections based on the hypotheses presented. A table of the web-based programs used in this report, their website addresses, and the relevant references, can be found in Appendix A.

2.1 Prediction Of *STI1* Gene Structure

The human genomic *STI1* DNA (accession no. NT_033903.7), human *STI1* mRNA (accession no. NM_006819.1), human Sti1 protein (accession no. NP_006810.1), mouse genomic *STI1* DNA (accession no. NT_082892.2), mouse *STI1* mRNA (accession no. NM_016737.1), mouse Sti1 protein (accession no. NP_058017.1) and yeast Sti1 protein (accession no. P15705) sequences were extracted from GenBank at the NCBI Entrez webpage. Yeast genomic *STI1* DNA was accessed from the *Saccharomyces* Genome Database (SGD). These sequences were retrieved on 2 November 2004.

2.1.1 Chromosomal Location And Surrounding Genes

Genes upstream and downstream of human and mouse *STI1* were accessed from NCBI MapViewer and compared. Genes surrounding *ySTI1* were accessed from the SGD.

2.1.2 Exon / Intron Boundaries

The human, mouse and yeast coding regions, including 1 kb up- and downstream of the coding regions, were scanned by Genscan (Burge and Kalin, 1997) and HMMGene (Krogh, 1997) at default settings. Default settings were used as they have been set to optimise gene prediction. Where the user is provided with a choice of which organism the algorithm has been optimised for, Appendix D indicates the optional settings used. With the mean intergenic region in yeast being 536 bp

(Hurowitz and Brown, 2003), a region of 1 kb on either side of each coding region was chosen for analysis, and kept standard for all three test organisms.

Although genes can be predicted by a number of alternate methods such as neural networks (GRAIL II), linear discriminant analysis (FGENEH, HEXON) and quadratic discriminant analysis (MZEF), Genscan and HMMGene were chosen for their robustness (Rogic *et al.,* 2001).

Genscan

Genscan, developed by Burge and Kalin (1997), makes use of a probabilistic model to recognise (i) general transcriptional, translational and splice signals, (ii) length distribution for introns, initial exons, internal exons and terminal exons, and (iii) composition of exons, introns and intergenic regions. Signals such as donor and acceptor splice sites, polyadenylation signals and the TATA-box are defined by weight matrices. Genscan does not make use of homology information. Upon analysis, Genscan searches both strands simultaneously and has been reported to perform steadily across a range of GC contents (Rogic *et al.,* 2001). At default settings, the highest-scoring predictions for promoters, introns, exons and polyadenylation signals are returned as a predicted gene structure. Genscan is almost unique in that it is able to recognize single, multiple or partial genes from the query DNA sequence. As Genscan's statistical model was trained from vertebrate genes, it is best-suited to predicting vertebrate gene structure.

HMMGene

Like Genscan, HMMGene (Krogh, 1997) also predicts gene structure, can predict multiple or partial genes, and performs steadily over a range of GC contents (Rogic *et al.,* 2001). Although HMMGene also makes use of a probabilistic model to predict gene structure, it uses hidden Markov models to compute and assign a probability score that the prediction is accurate. As for Genscan, HMMGene's default parameter is to output only the best gene structure prediction. HMMGene is different to Genscan in that it allows for constraints provided by user-input annotated features,

although this option was not utilized. HMMGene is currently set to predict genes in vertebrates and *C. elegans*.

2.1.3 Transcription And Translation Signals

The transcriptional and translational start and stop sites of all three *STI1* orthologues were compared to consensus sequences. In addition, the transcriptional and translational start sites of *hSTI1* and *mSTI1* were compared with each other as a high degree of conservation suggests that this region is particularly important and thus may be the biologically functional start site. Global pairwise alignment of sequences was performed in BioEdit (Hall, 1999) using a PAM250 similarity matrix (Dayhoff *et al.,* 1978). Multiple sequence alignment was performed in ClustalW version 1.82, using default parameters (Thompson *et al.,* 1994).

Transcription Start And Stop Sites

The mammalian TSS is spanned by the Inr which has a consensus of 5'-PyPyA$_{+1}$N(T/A)PyPy – 3' (Javahery *et al.,* 1994; Smale *et al.,* 1998), where Py represents pyrimidine (cytosine C or thymine T), N represents any nucleotide and A$_{+1}$ represents the transcription start site, adenine. The Inr site is of importance because it is thought to be the most common promoter site (Bajic *et al.,* 2003). Also of interest is the downstream promoter element (DPE), with consensus PuG(A/T)CGTG, where Pu represents a purine (A or guanine G) base (Burke and Kadonaga, 1997), usually found approximately 30 bp downstream of the TSS.

The transcription termination signal is considered to be YGTGTTYY and found at 15 – 30 bp downstream of the polyadenylation signal, AATAAA (McLauchlan *et al.* 1985; Proudfoot and Brownlee, 1976). This pre-mRNA processing signal indicates that a polyadenine tail should be added.

Translation Start And Stop Sites

The translation start signal was assumed to be the Kozak sequence (GCC)GCC(A/G)CCATGG (Kozak, 1987), while the translation stop signal was assumed to be the canonical TGA, TAG or TAA.

2.2 Identification Of Putative Promoter Regions And Upstream Regulatory Elements In *STI1* Orthologues

The second section of this chapter describes the methods used to identify putative promoter and regulatory elements in the *STI1* orthologues. The three main methods used are described in section 2.2.1, section 2.2.2, and section 2.2.3.

2.2.1 Alignment Of Mouse And Human Sequences From -500 To +100

As the core promoter spans the TSS and contains the TATA box (Fukue *et al.,* 2004), the region -500 to +100 with respect to the putative TSS was of interest. Thus, the abovementioned region of *hSTI1* and *mSTI1* were aligned, using a pairwise global alignment in BioEdit (Hall, 1999) with a PAM250 similarity matrix (Dayhoff *et al.,* 1978). The percent identity was determined and the region was scanned manually for sites of local conservation. It was purported that this region immediately surrounding the TSS may be critically involved in gene expression, and that important regulatory signals may have been conserved between the two orthologues. Indeed, highly conserved non-coding regions between human and mouse sequences are more likely to perform an important function, such as comprising regulatory elements to which TFs can bind, than non-conserved non-coding regions (Down and Hubbard, 2004).

The consensus sequence for the TATA box, TATA(A/T)A(A/T)(A/G) (Breathnach and Chambon, 1981; Carey and Smale, 2000; Goldberg, 1979), was used to search for putative biologically functional TATA boxes in the region surrounding the predicted transcriptional start sites.

Because CpG islands are found associated with the 5' end of genes, and thus the promoter region, (Antequera and Bird, 1993), DNAssist (Patterton and Graves, 2000a, 2000b) was used to calculate the GC content for the coding sequences, mRNA, and the region from -500 to +100 for *hSTI1*, *mSTI1*, and *ySTI1*. To calculate the GC content of the *STI1* orthologues' relevant chromosomes, the assembled chromosome sequences were downloaded on 8 May 2005 and an algorithm in Python (Appendix J) used to count their base composition. The sequence for human chromosome 11 and mouse chromosome 19 were accessed from GenBank's FTP site, whilst the sequence for yeast chromosome 15 was accessed at the SGD.

2.2.2 Algorithmic Recognition Of Promoter Regions

A variety of promoter and TFBS prediction programs was used, as suggested in Chapter 1. Although the prediction of TFBS can be used to predict promoter regions, they may also occur, and be functional at, sites distant from the promoter regions. In addition, the methods used to predict promoter regions are different from those that predict TFBS. Thus, this section has two parts; that describing promoter prediction, and that describing TFBS prediction.

Promoter-predicting Programs

Recognition of promoter elements in the *STI1* orthologues was achieved by scanning the genomic DNA 1 kb upstream of the transcriptional start site with Neural Network Promoter Prediction (NNPP), PromoterScan, TSSG and CorePromoter at default settings. These programs, used in preliminary analyses to determine promoter regions, correspond to type one through four of the six promoter prediction methods described by Mount (2001), and are discussed in Table 1 and later on in section 2.2.2. Appendix F indicates the optional settings used.

As a negative control, *STI1* genomic DNA 1 kb upstream of the transcriptional start was run through the Shuffle package (Appendix A) to create a random sequence, and then scanned as for the *STI1* DNA. The Shuffle software generates random DNA sequences from the input sequences and thus nucleic acid content and length of the random sequence remains the same as the test sequence. Although the shuffled DNA

sequences were not analysed as such, they were used as an indicator of how many of each type of TFBS could have occurred randomly, as a result of sequence composition.

In addition to this control, *HSP70* genomic DNA of 1kb upstream of the TSS was scanned in the same way as the corresponding *STI1* DNA for the same organism. This served as a positive control because Hsp70 and Sti1 proteins are functionally related and were thus assumed to be regulated, at least in part, by the same TFs (Down and Hubbard, 2004).

Type 1: NNPP

NNPP (Waibel *et al.,* 1989; Reese and Eeckman, 1995; Reese, 2000; Reese, 2001) makes use of a neural network that has been trained to identify eukaryotic promoter elements such as the TATA-box, CCAAT-box, GC-box and transcription start site (Inr). These elements may occur in any number of combinations and with varying distances between the elements. The trained neural network can be described as 'black box', through which a query DNA sequence is passed, that decides how relevant the combination of elements within the query sequence is based on the training sequences, and then outputs promoter predictions that satisfy a given confidence level. NNPP predicts a 50 bp region from -40 to +10 with respect to the estimated TSS.

Type 2: PromoterScan

PromoterScan (Prestridge, 1995) predicts eukaryotic RNA Pol II promoters, based on a linear function, by recognition of a TATA box, and the density and type of transcription factors in the region. The TATA box is identified using a weight matrix, whilst the significance of the observed TF density is based on comparison of the query sequence, known promoter sequences in the EPD (Périer *et al.,* 1998; Périer *et al.,* 1999; Périer *et al.,* 2000; Praz *et al.,* 2002, Schmid *et al.,* 2004), and known non-promoter sequences in GenBank. PromoterScan outputs a predicted promoter region, promoter score, TATA box position (if present), TSS position (if a TATA box is found), and a list of transcription factors. The Ghosh transcription factor database number is reported for each transcription factor, as well as a strand, position and weight. Unlike most transcription factor predicting programs, the weight reported for

a TF is not an indication of the quality (biological significance) of the signal. Instead, the weight is a measure of how well the TF is able to discriminate between promoter and non-promoter sequences. For example, a TF with the maximum score of 50 is found (to PromoterScan's knowledge) only within promoter sequences and is thus particularly useful in predicting the promoter region of the query DNA sequence.

Type 3: TSSG

TSSG (Solovyev and Salamov, 1997) uses a linear discriminant function to discern promoter from non-promoter sequences. TSSG identifies TF sites in a similar way to PromoterScan (Prestridge, 1995). It searches for a TATA box and TFBS, calculates codon preferences and hexamer frequencies, and also uses oligonucleotide composition as a measure of predicting the transcription start site more accurately. TSSG relies on data from an outdated version of the transcription factor database (Transfac) (Wingender *et al.,* 1996) to search for eukaryotic *cis*-regulatory DNA elements and TFs.

Type 4: CorePromoter

CorePromoter (Zhang, 1998; Ioshikhes and Zhang, 2000; Zhang, 2000) predicts the transcriptional start site (TSS) of a human query DNA sequence using quadratic discrimination analysis. The frequency of pentamers is calculated in 30-bp and 45-bp double-overlapping windows to lessen the background noise. CorePromoter defines the core promoter region as -60 to +40 with respect to the TSS.

2.2.3 Algorithmic Recognition Of Transcription Factor Binding Sites

Recognition of *cis*-acting elements in the *STI1* orthologues was achieved by scanning the genomic DNA 1 kb upstream of the TSS with Alibaba, Transcription Element Search Software (TESS), and TFSearch. Alibaba and TFSearch were used at default settings. Because the default TESS output was very extensive, and therefore probably not stringent enough, TESS results were limited by changing the defaults settings so that the minimum string length was 8 (default 6). Increasing the string length makes the TESS search more stringent because TESS will not search for matches that are

shorter than the minimum string length. The probability of finding any particular 8-bp motif, assuming random distribution of the 4 types of nucleotides, is 1 per 65,536 kb, a far more stringent search than that for a 6-bp motif which can be expected to occur at a rate of 1 per 4,096 kb. The negative and positive controls used were the same as described for section 2.2.2.

While predicting TFBS using TESS and TFSearch may be classified according to Mount (2001) as the fifth method of promoter prediction, Alibaba is unclassified as it was developed after Mount's publication. Mount's method six was not used in this study as it aimed more at searching for a promoter region that corresponds to a set of user-defined parameters, than searching for the promoter region in one unannotated query sequence.

TESS (Shug and Overton, 1997a,b) predicts TFBS by string- and weight matrix-based searching of both strands of the query sequence against the Transfac database (Wingender *et al.,* 1996). Similarly, TFSearch (Heinemeyer *et al.,* 1998) also uses weight matrices to search for transcription factor sites. The method of TFBS detection used by Alibaba (Grabe, 2002) is slightly different: Alibaba searches the query DNA sequence for known transcription factor binding sites, aligning these known sites by pairwise alignment to the query sequence. A score is given to the alignment based on how well the query sequence matches the known TFBS. If the score is above a threshold score, this site is accepted for use in matrix construction. The hierarchical classification in Transfac 4.0 is used to pool all accepted matrices within a subfamily. A temporary matrix is constructed from the set of TFBS in a subfamily and it is these temporary matrices that are the output for Alibaba 2.1.

Analysis Of The Transcription Factor Binding Site Prediction Programs' Output

The output of Alibaba, TESS and TFSearch were compared as follows: For a particular orthologue, each TF predicted by each of the programs was compared to the TFs predicted by the other two programs, for all three orthologues and *HSP70* of the query orthologue, to look for a match. Thus, for *hSTI1*, Alibaba's output would be compared to TESS output and TFSearch output, and then the TESS and TFSearch

output would be compared. Then, the *hSTI1* output would be compared with the output from Alibaba, TESS and TFSearch for *mSTI1*, *ySTI1*, and human *HSP70*. A schematic diagram illustrates the occurrence of predicted TFs that show matches, in both type and position, with two or more predicted TFs amongst the orthologues and positive control. If, for example, transcription factor X was predicted to occur in *ySTI1*, yeast HSP70 and *hSTI1* but not in human HSP70 or *mSTI1*, then the TF was indicated on the human upstream region even though it showed only one match.

It is important to note that the difference in the number of TFBS predicted between Alibaba, TESS and TFSearch is due to the stringency and method by which the TFBS are predicted. TFBS that are predicted by all three programs, especially those that have positional matches with orthologues, are more likely to have functional relevance than the TFBS that are predicted to occur by only one TF search program and not for any orthologues. This hypothesis formed the basis for the method by which the results were compared.

A table was drawn up to summarise the results, showing only the TFs that match with another orthologue or Hsp70 of the same species. For each TF predicted in this table, the number of TFs predicted within a 50 bp region of the query TF were counted and plotted. The number of TFs were counted both up- and downstream of the query TF.

TFs were selected for further investigation if (i) they had a high score, (ii) they occurred often within one program or between programs, (iii) the function of the TFs was well-annotated, (iv) the TFs were known to be involved in cell stress, cell cycle, development or disease, (v) partner proteins were identified, or (vi) the TFs were known to associate with Hsp70 or Hsp90.

2.3 Identification Of Genes Co-Regulated With Yeast *STI1* And Determination Of Over-Represented Regulatory Motifs

It is of interest to determine which genes are co-regulated with *ySTI1* as, because these genes are manufactured by the cell at the same time, it is possible that *ySTI1*

may be involved in the same cellular processes as these cells. Also, these co-regulated genes may have similar upstream regulatory elements that, when compared to *ySTI1*, would enable the genes to be regulated under the same set of conditions. The method described below, of clustering microarray data to find co-expressed genes (section 2.3.1), and then searching the upstream region of these genes for over-represented motifs (section 2.3.2), is a common approach (DeRisi *et al.,* 1997; van Helden *et al.,* 1998; Tavazoie *et al.,* 1999).

2.3.1 Genes Co-Regulated With *ySTI1*

Microarray data hosted by the SGD were searched for experimental conditions where the expression of *STI1* changed by more than one-fold. While choosing experiments where *ySTI1* is upregulated by only one-fold may not be a very stringent approach, it was chosen to allow for the maximum number of experiments to be returned while still requiring the change in *ySTI1* expression to be noticeable. The microarray data for the experiments fitting this one-fold requirement were downloaded and the expression results clustered using Cluster (Eisen *et al.,* 1998). Data for yeast gene expression during diauxic shift were not accessible.

Cluster Analysis

When clustering microarray data, the first decision is whether the data should be clustered using a supervised or unsupervised method. As the name suggests, supervised methods require *a priori* knowledge regarding which genes may be expected to cluster together. There is little knowledge regarding which genes are expected to be co-expressed with *ySTI1*, and therefore an unsupervised method was chosen. Unsupervised clustering techniques are further divided into methods that are hierarchical (agglomerative) or non-hierarchical (divisive). Hierarchical clustering is the most common approach (Tilstone, 2003). It begins with each gene as its own cluster and then iteratively clusters genes together, based on the similarity of their expression data throughout the experimental conditions, until all genes are in one cluster. The settings chosen for cluster analysis were thus hierarchical, correlation centred, average linkage clustering. It is noted that this clustering method assumes that genes with a similar expression pattern to *ySTI1* will be functionally related to

ySTI1, although this is not always the case. Also, this method may group genes that have high expression values at different times in the experiment (Eisen *et al.,* 1998).

The clustering method described produces nested clusters that can be displayed as a two-dimensional hierarchical dendrogram where the length of the branches of the tree are an indication of how similar the expression profile of the genes is (Eisen *et al.,* 1998). Thus, the clustered data were viewed using TreeView (Eisen *et al.,* 1998). Genes clustering with a probability of 0.85 or higher with *STI1* were considered relevant and their upstream sequences were extracted from the SGD in early August 2004. The upstream region of each of the genes co-clustering (and thus perhaps co-regulated) with *ySTI1* was expected to share common sequence motifs with other genes in its cluster, which would explain their similar expression profiles. In this way, each cluster provides information on the cell's response to the experimental condition in question by identifying genes that are most likely to act together.

Extracting The Upstream Region of Genes Clustering With *ySTI1*

Where the upstream (5') gene with respect to the gene of interest occurs on the same strand (i.e. head-to-tail), the DNA sequence retrieved was the sequence occurring from the 3' end of the upstream gene to the 5' end of the gene of interest. Where the upstream gene with respect to the gene of interest occurs on the opposite strand (i.e. head-to-head), the upstream sequence extracted contained the region upstream of the gene of interest, until the TATA box of the gene on the opposite strand. A TATA box was accepted if it contained the sequence TATA (and not necessarily the sequence TATA(A/T)A(A/T)(A/G) (Breathnach and Chambon, 1981; Carey and Smale, 2000; Goldberg, 1979) within the region upstream 60-100 bp. If no TATA box was found, or if the TATA box occurred outside of this region, then the entire region between the start of the sequential genes was extracted. All DNA was extracted from the SGD which returned the Watson (+) strand. Thus, the reverse complement was found for genes occurring on the Crick (-) strand.

2.3.2 Recognition Of Over-Represented Motifs

All relevant sequences from each microarray experiment were assembled alphabetically in FASTA format into a file, and passed through the Gibbs sampling algorithm, AlignACE, at default settings (Roth *et al,.* 1998; Hughes *et al.,* 2000). AlignACE was used to identify the presence of over-represented motifs in the sequences provided. It is of interest to search for over-represented motifs as their presence in more than one sequence, and more than once within a sequence, could be an indication of biological functionality.

AlignACE returns a MAP score as an indication of the statistical significance of the motif alignment when compared to the genomic background and thus shows the specificity of a motif for the sequence. A MAP score of zero means that zero sites have been aligned. The consensus motifs retuned by AlignACE were searched against Transfac database in the SITE table, with the search term being the Sequence (SQ) table field. All output motifs were considered irrespective of the MAP score.

CHAPTER 3: RESULTS

This chapter presents the results of this study, and follows the structure of the previous chapter by being divided into three sections according to the hypotheses presented in Chapter 1.

3.1 *STI1* gene structure

This section is divided into three sections based on different aspects of the broad term, gene structure. Firstly, the context of the *STI1* orthologues is presented in relation to the chromosome and nearby genes. Next, the exon/intron organization of the orthologues is presented, and lastly the transcription and translation signals are discussed.

3.1.1 Chromosomal Location And Surrounding Genes

The *hSTI1* gene is localised to chromosome 11 region q.13, *mSTI1* is localized to chromosome 19, and *ySTI1* is localized to chromosome 15. The genes adjacent to *STI1* correspond between human and mouse (Figure 7 and Appendix C). Indeed, a region including at least three genes (218.6 kb in human and 166.3 kb in mouse) upstream of the 5' end of *STI1*, and a region of at least nine genes (80.1 kb in human

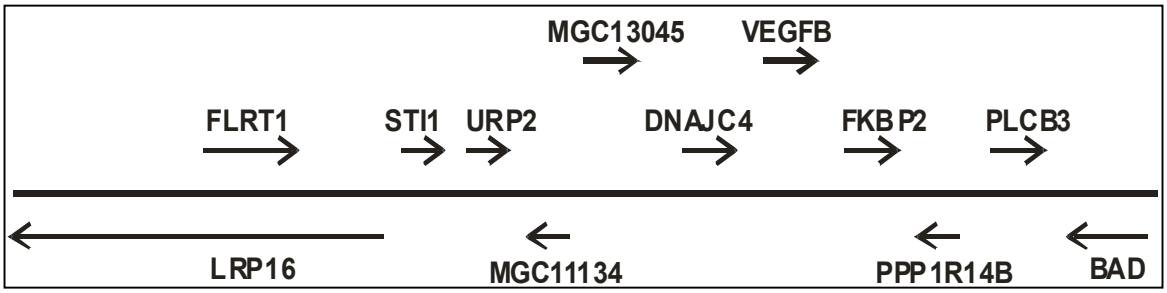

Figure 7: Schematic diagram of genes surrounding *hSTI1*. Human and mouse *STI1* lie in a region of synteny and thus these genes are also found surrounding *mSTI1*. Genes include fibronectin leucine rich transmembrane protein 1 (*FLRT1*), stress-inducible protein 1 (*STI1*), UNC-112 related protein 2 (*URP2*), hypothetical protein MGC13045, DnaJ (*HSP40*) homologue Subfamily C Member 4 (*DNAJC4*), vascular endothelial growth factor B (*VEGFB*), FK506-binding protein 2 (*FKBP2*), phospholipase C beta 3 (*PLCB3*), Low density lipoprotein-related protein 1 (*LRP16*), tRNA splicing 2' phosphotransferase 1 (MGC1134), protein phosphatase 1 regulatory (inhibitor) subunit 14B (*PPP1R14B*), and BCL2-antagonist of cell death (*BAD*). Arrowheads indicate the direction of transcription, with genes above the solid line being genes on the (+) strand and genes below the solid line being genes on the (-) strand. The function of some of these genes has not been identified directly, but has been inferred from homologous genes. This diagram is not to scale.

and 78.9 kb in mouse) downstream of the 3' end of *STI1* correspond with respect to the encoded genes.

Aligning genomic *hSTI1* and genomic *mSTI1* using a PAM250 similarity matrix (Dayhoff *et al.,* 1978) gave a score of -256 for the region -2000 to -1000 (identity 43%), a score of 68 for the region -1000 to -1 (identity 57%), and a score of -94 for the region +1 to +1000 (identity 51%) with respect to the transcriptional start site.

3.1.2 Exon / Intron Boundaries

This section presents the results of exon/intron boundary prediction for the *STI1* orthologues. Where gene prediction programs did not agree, the differences have been identified, and reasons for these differences discussed in Chapter 4. For clarity's sake, the section is divided into three sections (human, mouse and yeast), so that results for the three orthologues are easily distinguished.

Human

Both HMMGene and Genscan predicted 15 exons for *hSTI1* and all predicted exons were identical except for the first two exons, and the 5' end of the sixth predicted exon (Appendix D and Figure 8). Alignment of the genomic *hSTI1* DNA with the corresponding mRNA suggests that HMMGene and Genscan maybe be predicting an extra exon at the 5'-end of the coding region. The first exon predicted by HMMGene overlaps with the second exon predicted by Genscan (Figure 8). The second exon predicted by HMMGene overlaps with the first exon predicted by alignment with mRNA. Neither of the first two exons predicted by Genscan overlaps with the first exon predicted by alignment with mRNA. Based on information from HMMGene, Genscan and alignment with mRNA (cDNA), the putative exon/intron boundaries are given in Table 2.

The putative exons (Table 2) were translated and matched the hSti1 protein sequence exactly. The translational start codon is the first ATG in the putative transcript. Table 2 provides additional information on the exon/intron splice sites in *hSTI1*. The 5'-untranslated region (UTR) and 3'-UTR of *hSTI1* mRNA are 58 and 417 bp in length

Table 2: Putative exon/intron boundaries of *hSTI1*. The boundaries are based on HMMGene and Genscan results, and on alignment of genomic DNA with cDNA. Exon/intron splice-junction sequences of the *hSTI1* gene are also shown.

Exon					Exon/intron boundaries		
No.	**Begin**	**End**	**Length (bp)**	**Length (aa)**	**Exon (n)**	**Intron (bp)**	**Exon (n+1)**
5' UTR[i]	63729036	63729093	58				
0(G)[ii]	(63728125)	(63728331)	(207)				
0(H)[ii]	(63728657)	(63728790)	(134)				
1(G)[iii]	(63728646)	(63728806)	(161)				
1(H)[iii]	(63729015)	(63729102)	(88)				
1	63729094	63729102	9	3	ATGGAGCAGgtgaag M_1 E_2 Q_3	6807	cctcagGTCAAT V_4 N_5
2	63735910	63736119	210	73	GGCAAGgtcagc G_{72} K_{73}	901	tttaagGGCTAT G_{74} Y_{75}
3	63737021	63737162	142	120.3	TTGGCAGgtaggt L_{119} A_{120}	148	ggacagAGAGAAAA E_{121} R_{122} K_{123}
4	63737311	63737452	142	167.7	CTGGGCACgtaagt L_{166} G_{167} T_{168}	1024	tactagGAAACTA K_{169} L_{170}
5	63738477	63738645	169	224	AAGCAGgtcttg K_{223} Q_{224}	1457	atctagGCACTG A_{225} L_{226}
5(H)[iv]	(63738513)	(63738645)	(133)	-		(1061)	
6	63740103	63740229	127	266.3	CAAGCAG gtgagg Q_{265} A_{266}	95	tggcagCGGTATAC A_{267} V_{268} Y_{269}
7	63740325	63740427	103	300.7	ATTGCCAA gtaggc I_{299} A_{300} K_{301}	258	tatcagAGCATAT A_{302} Y_{303}
8	63740686	63740806	121	341	CAGCAGgtgcgt Q_{340} Q_{341}	1965	ttgtagGCAGAG A_{342} E_{343}
9	63742772	63742868	97	373.3	CAGAAAGgtactg Q_{372} K_{373}	132	ccccagGGGACTAT G_{374} D_{375} Y_{376}
10	63743001	63743125	125	415	CTCAAGgtgacg L_{414} K_{415}	2582	ttctagGACTGT D_{416} C_{417}
11	63745708	63745744	37	427.3	ACCTTCA gtaagt T_{426} F_{427}	212	ttgtagTCAAGGGT I_{428} K_{429} G_{430}
12	63745957	63746060	104	462	TGTAAGgtgggg C_{461} K_{462}	221	ctgcagGAGGCG E_{463} A_{464}
13	63746282	63746454	173	519.7	CTCAGCGA gtacgt L_{518} S_{519} E_{520}	431	ctgtagACACTTA H_{521} L_{522}
14	63746886	63746958	73	543	CGGTGA tgactt R_{543} stop	-	-
3' UTR[i]	63746959	63747375	417				

Predicted 5'-untranslated (UTR) and 3'-UTR regions are shown, as well as exons that were incorrectly predicted by Genscan (G) and HMMGene (H). Uppercase nucleotides are translated whilst lowercase nucleotides are not translated. Amino acid type and number are indicated below the codons encoding them.

(i) Region that is transcribed but not translated.

(ii) First exon predicted by gene prediction program Genscan (G) or HMMGene (H). This exon does not correspond to an exon predicted by alignment with mRNA.

(iii) Second exon predicted by gene prediction program. This exon corresponds to, but does not match, the first exon predicted by alignment with mRNA.

(iv) The 5' end of this exon, predicted by HMMGene, differs from that predicted by Genscan and alignment with mRNA.

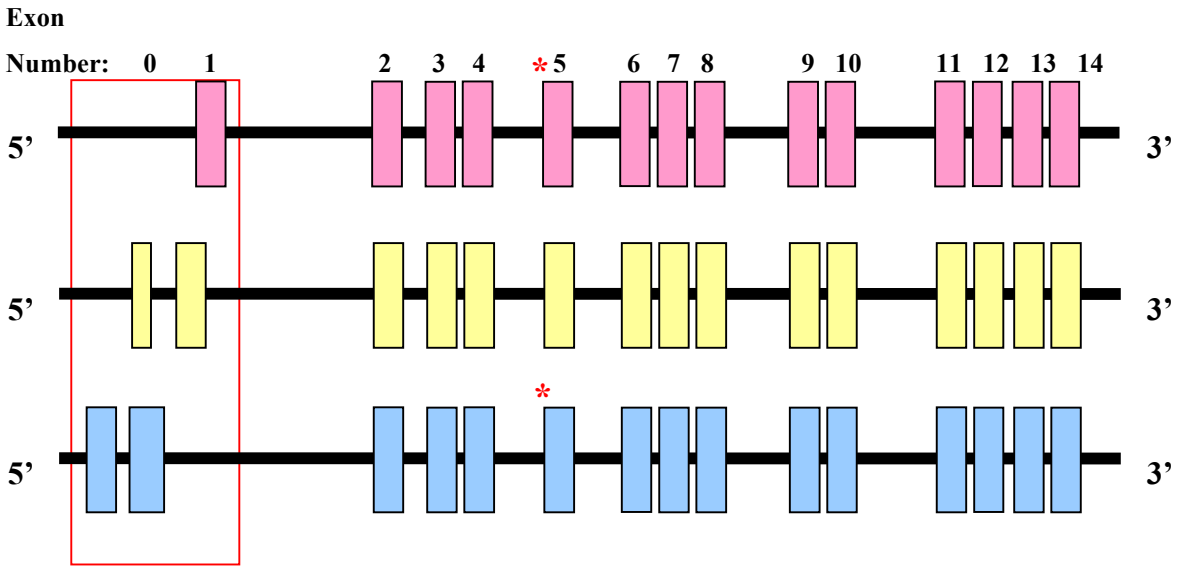

Figure 8: Figure depicting the predicted exons and introns in *hSTI1*. The top (pink) row is the exon/intron structure according to alignment with mRNA. The second (yellow) row and third (blue) row show that HMMGene and Genscan, respectively, do not predict the first exon accurately. The asterix above the top row's 5th exon and third row's 6th exon indicate identical sites, whereas HMMGene does not agree with this splice site. Exons are numbered 0 to 14 because it is most likely that exon 0 is incorrect and that *hSTI1* has only 14 exons, ie. Exons 1 to 14.

while the introns range from 95 to 6807 bp in length. The *hSTI1* transcribed region is 18,340 bp long before splicing (Table 2 and Figure 10).

Mouse

Genscan predicted 15 exons for *mSTI1* whilst HMMGene predicted 14 exons (Figure 9). The first two exons predicted by Genscan do not agree with the first exon predicted by HMMGene, and exon 5 does not match either. Similar to *hSTI1*, alignment of genomic *mSTI1* with the corresponding mRNA suggested that *mSTI1* comprises 14 exons. The putative exon lengths agree perfectly with *hSTI1*.

Table 3 shows the chromosomal position of the exon / intron boundaries. The 5'-UTR and 3'-UTR of *mSTI1* mRNA are 53 and 393 bp in length, and the transcribed region is 19,234 bp before splicing (Table 3 and Figure 10). Similar to *hSTI1*, mouse *STI1* introns range from 103 to 4174 bp in length. Genscan's predicted polyadenylation site is within the 3' UTR shown in Table 3.

Table 3: Putative exon/intron boundaries of *mSTI1*. The boundaries are based on HMMGene and Genscan results, and on alignment of genomic DNA with cDNA. Exon/intron splice-junction sequences of the *mSTI1* gene are also shown.

Exon					Exon/intron boundaries		
No.	**Begin**	**End**	**Length (bp)**	**Length (aa)**	**Exon (n)**	**Intron (bp)**	**Exon (n+1)**
5' UTR[i]	6753370	6753318	53				
0(G)[ii]	(6752806)	(6753176)	(371)				
1(G)[iii]	(6752208)	(6752271)	(64)				
1(H)[iii]	(6749364)	(6749356)	(9)				
1	6753317	6753309	9	3	ATGGAGCAGgtgaag M_1 E_2 Q_3	4174	cctcagGTGAAT V_4 N_5
2	6749134	6748925	210	73	GGCAAGgtaagc G_{72} K_{73}	707	gttcagGGTTAT G_{74} Y_{75}
3	6748217	6748076	142	120.3	TTGGCAGgtgggt L_{119} A_{120}	341	ggacagAGAGGAAA E_{121} R_{122} K_{123}
4	6747734	6747593	142	167.7	CTGGGCACgtgagt L_{166} G_{167} T_{168}	1513	atctag GAAACTA K_{169} L_{170}
5	6746079	6745911	169	224	AAACAGgtcttt K_{223} Q_{224}	3246	atctagGCACTG A_{225} L_{226}
5(H)[iv]	(6744301)	(6744208)	(94)	-		(1609)	
6	6742664	6742538	127	266.3	CAAGCAGgtgagg Q_{265} A_{266}	103	tggcagCTGTGCAC A_{267} V_{268} H_{269}
7	6742434	6742332	103	300.7	ATCGCCAAgtatgc I_{299} A_{300} K_{301}	153	cttcagAGCTTAT A_{302} Y_{303}
8	6742178	6742058	121	341	CAGCAGgtgggt Q_{340} Q_{341}	2239	ttttagGCAGAG A_{342} E_{343}
9	6739818	6739722	97	373.3	CAGAAAGgtacag Q_{372} K_{373}	852	atctagGGGACTAC G_{374} D_{375} Y_{376}
10	6738869	6738745	125	415	CTCAAGgtgagg L_{414} K_{415}	3006	ccctagGACTGT D_{416} C_{417}
11	6735738	6735702	37	427.3	ACCTTCA gtaagt T_{426} F_{427}	141	ttgtagTCAAGGGT I_{428} K_{429} G_{430}
12	6735560	6735457	104	462	TGTAAGgtaagc C_{461} K_{462}	204	atgcagGAAGCA E_{463} A_{464}
13	6735252	6735080	173	519.7	CTGAGCGA gtaagt L_{518} S_{519} E_{520}	478	tcctagACACTTA H_{521} L_{522}
14	6734601	6734529	73	543	CGGTGAtaactt R_{543} stop	-	
3' UTR[i]	6734528	6734136	393				

Predicted 5'-untranslated (UTR) and 3'-UTR regions are shown, as well as exons that were incorrectly predicted by Genscan (G) and HMMGene (H). Uppercase nucleotides are translated whilst lowercase nucleotides are not translated. Amino acid type and number are indicated below the codons encoding them.

(i) Region that is transcribed but not translated.

(ii) First exon predicted by gene prediction program Genscan (G). This exon does not correspond to an exon predicted by HMMGene or alignment with mRNA.

(iii) Second exon predicted by gene prediction program. This exon corresponds to, but does not match, the first exon predicted by alignment with mRNA.

(iv) The 5' end of this exon, predicted by HMMGene, differs from that predicted by Genscan and alignment with mRNA.

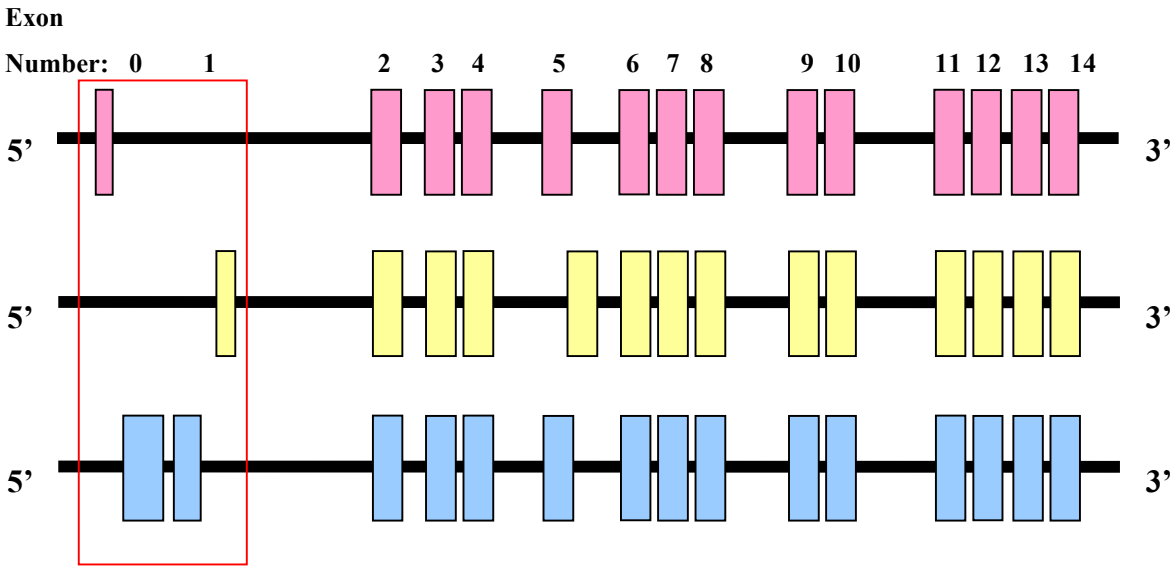

Figure 9: Figure depicting the predicted exons and introns in *mSTI1*. The top (pink) row is the exon/intron structure according to alignment with mRNA. The second (yellow) row and third (blue) row show that HMMGene and Genscan, respectively, do not predict the first exon accurately. Additionally, HMMGene does not predict the Exon 5 accurately, as illustrated by the significant misalignment of this exon when compared to the exon 5 predicted by mRNA and Genscan. Exons are numbered 0 to 14 because it is most likely that exon 0 is incorrect and that *mSTI1* has only 14 exons.

Yeast

HMMGene correctly predicted the *ySTI1* gene to be a single open reading frame (ORF) from 1001 to 2770 when the setting was set as *Homo sapiens*. Setting to *Caenorhabditis elegans* predicted the correct coding region but divided into two exons. Genscan predicted three possible genes in the region; the only full-length predicted gene in the region had the highest score and corresponded with experimental work for *ySTI1*. Appendix D gives the co-ordinates of the *ySTI1* gene. The *ySTI1* 5'-UTR is 64 bp and the 3'-UTR is 374 bp and thus, the transcribed region is 2,208 bp in length (Figure 10). Yeast *STI1* is one of more than 95% of yeast genes that is not interrupted by introns.

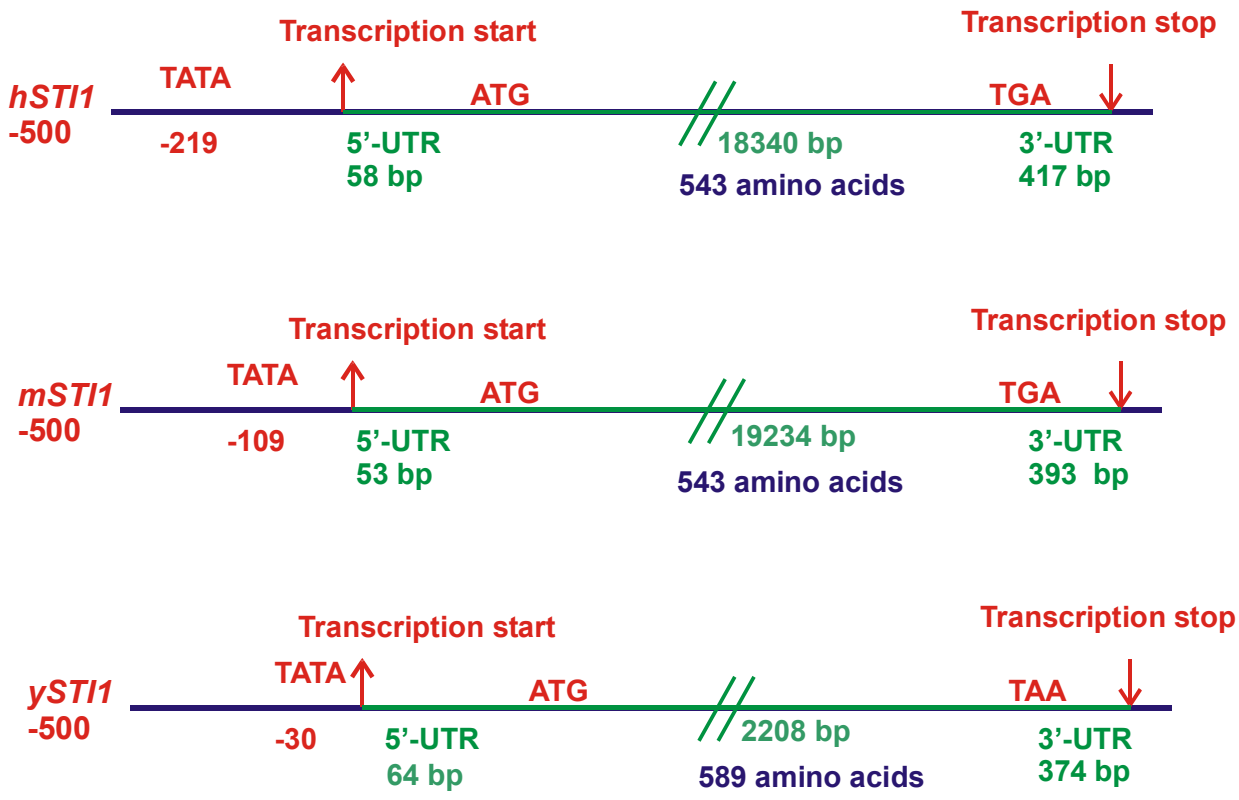

Figure 10: Schematic diagram to show the position of the transcriptional and translational start and stop sites for *hSTI1, mSTI1* and *ySTI1*.

3.1.3 Transcriptional And Translational Signals

The transcriptional start and stop sites are first presented, followed by the translation start and stop sights, and then a summary of putative lengths of the untranslated regions for the three *STI1* orthologues.

Transcriptional Start And Stop Signals

Alignment of human genomic DNA with the corresponding mRNA suggested that the transcriptional start site occurred at aaggcggcgcGTGCGGTTGG where uppercase letters represent transcribed bases. However, the Eukaryotic Promoter Database (EPD) (Périer *et al.,* 1998; Périer *et al.,* 1999; Périer *et al.,* 2000; Praz *et al.,* 2002, Schmid *et al.,* 2004) reports that transcription is initiated most frequently at four bp downstream of this site, as underlined. For the purpose of this report, the EPD transcription initiation site will be used and the first underlined G will be designated the putative +1 transcription start site.

Alignment of the mouse genomic DNA with the corresponding mRNA suggested that the transcriptional start site occurred at ggcggcgcgtGCGGTTGGGA. Alignment of the human and mouse *STI1* genomic DNA over the putative transcription initiation site returned a 55 bp sequence (agcttctagtaggttccagaaggcggcgcgtgcGGTTGGGAACG CGGAGCGGACG) that was perfectly conserved between the two orthologues. For this reason, the putative transcription initiation site of the *mSTI1* will be designated the same as the corresponding human transcription initiation site.

An Inr-like sequence (TTAAGCC) was found –67 to -61 bp upstream of the start codon of *ySTI1*. It conforms to the Inr consensus at all positions except the +3 position, and it is thus quite possible that this sequence may be the biological transcription start site. Thus, this was designated the transcriptional start site for *ySTI1*. No downstream promoter element (DPE) was found in either of the three orthologues. Neither the human nor the *mSTI1* contains a canonical Inr, although this is not uncommon.

In terms of the transcriptional stop sites, the *hSTI1* transcription termination signal (CGTGGTTA) is found at 14 bp downstream of the polyadenylation site and corresponds to the consensus sequence, YGTGTTYY. An identical putative transcription termination signal sequence (CGTGGTTA) is found 13 bp downstream of the AATAAA in *mSTI1*. In addition, 24 bp downstream of the *ySTI1* polyadenylation signal is a putative transcription termination signal, TTCGTTTCTT. This transcriptional termination signal does not match the consensus sequence, YGTGTTYY, exactly.

Translational Start And Stop Signals

In addition to the 55 bp that are conserved across the transcription initiation sites, the region of -3 to +27 of the *hSTI1* predicted translation start site aligns exactly with the *mSTI1* gene. The *hSTI1*, *mSTI1* and *ySTI1* predicted translation start sites do not conform to the optimal eukaryotic translational start site (Kozak, 1987). The translation start site of *hSTI1* is TGCGCTATGG, whilst the translation start for *mSTI1* is CAGGCTATGG and for *ySTI1* is AGAAAGATGT.

The translation stop codons for *hSTI1*, *mSTI1* and *ySTI1* are TGA, TGA, and TAA respectively (Figure 10). Translation of the coding region results in the human, mouse and yeast Sti1 proteins of length 543, 543 and 589 amino acids in length, respectively. An alignment of the *hSTI1* DNA with hSti1 protein can be found in Appendix B. The position of the NLS, casein kinase phosphorylation sites and the proline stretch are indicated.

Polyadenylation Tails And Untranslated Regions

Thus, after splicing, the *hSTI1* mRNA start codon is found at position 59 – 62 (Figure 10) and the stop codon is at position 1688 – 1690. A putative polyadenylation signal (AATAAA) is found 387 bp downstream of the stop codon and 26 bp from the 3'-end of the mRNA. Thus, the 3'-UTR region is 417 bp in length. This information corresponds to previous work (Honoré *et al.,* 1992) except that the transcriptional start site for this study was designated to be 4 bp downstream of the TSS described by Honoré *et al.* (1992). The 5'-UTR of *mSTI1* is 53 bp (Figure 10) in length and a putative polyadenylation signal is found 364 bp downstream of the stop codon and 23 bp upstream from the 3'-end of the mRNA. Thus, the 3'-UTR is 393 bp in length. The 5'-UTR of *ySTI1* is 64 bp in length and very close to the length of the 5'-UTR predicted for *hSTI1* (58 bp) and *mSTI1* (53 bp). A polyadenylation signal (AATAAA) is found 368 bp and 499 bp downstream of the *ySTI1* stop codon. The former signal is most likely to be biologically functional as the length of the 3'-UTR would thus correspond closely to that of *hSTI1* and *mSTI1*. Taking the transcriptional stop site as the position where the mRNA matches the genomic DNA, the transcribed regions of *hSTI1*, *mSTI1* and *ySTI1* are 18340, 19234 and 2208 bp long (Figure 10).

3.2 Putative Promoter Regions And Upstream Regulatory Elements In *STI1* Orthologues

The second section of this chapter presents the results of the promoter and TFBS prediction programs. The first part of this section presents observations from the alignment of *hSTI1* and *mSTI1* in the region –500 to +100 bp with regard to the TSS.

The second part of this section gives the output of the promoter prediction programs, and lastly, the third part of this section shows the results for the TFBS prediction.

3.2.1 Alignment Of Mouse And Human Sequences From -500 To +100

A global alignment of the human and mouse regions from -500 to +100 is shown in Figure 11. Pairwise global alignment of this region using BioEdit (Hall, 1999) gives 70% identity and an alignment score of 226 using a PAM250 similarity matrix (Dayhoff *et al.,* 1978).

Human

The TATA box of *hSTI1* was assigned to the TATA-like sequence (TTTATA) at -224 to -219 from the transcriptional start site. The region of 10 kb upstream of the transcription start site holds 10 potential TATA box-like sequences (Appendix E). Only three of these sequences match the TATA consensus of TATA(A/T)A(A/T) (Fukue *et al.,* 2004). An additional two sequences match the slightly less stringent sequence TATA(A/T)(A/G/T)(G/A). Many promoter prediction programs that search for the TATA box make use of the PWMs described by Bucher (1990).

The human chromosome 11 has a GC% of 42, whilst the GC content for *hSTI1* is 49% and is raised to 60% in the region from -500 to +100 with respect to the transcriptional start site. The region from -500 to +100 from the transcriptional start site of the *hSTI1* gene meets the criteria of a typical CpG island by having a frequency of observed CG dinucleotides / frequency of expected CG dinucleotides ~ 0.8 and GC% of 60 (Table 4).

Lastly, this region of the *hSTI1* gene has 8 repeats of CCAAT sequences and 4 Stimulatory protein 1 (Sp1) sites (3 upstream of the transcriptional initiation site – one of which is the reverse complement - and one downstream).

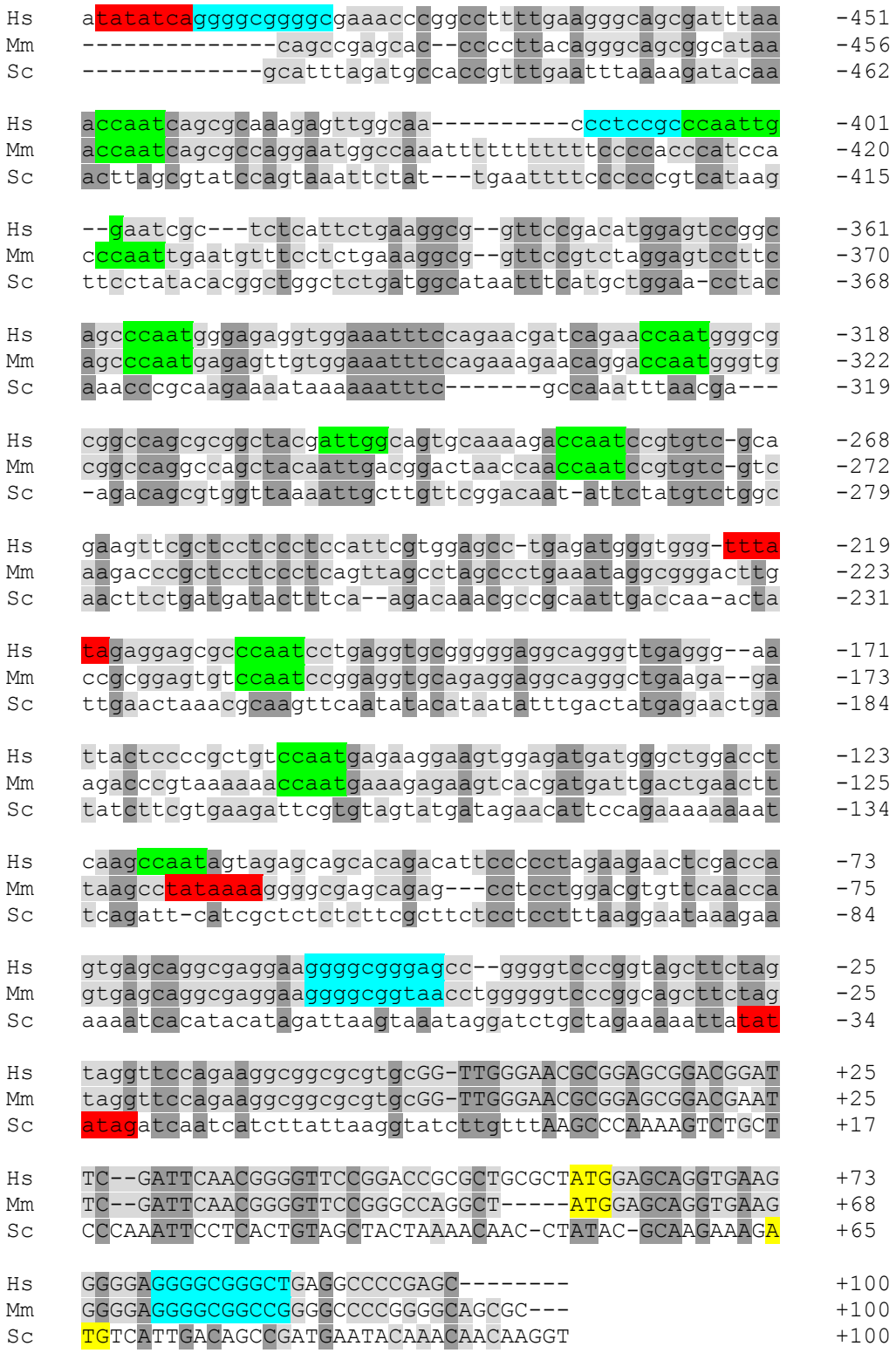

Figure 11: Alignment of human (Hs), murine (Mm), and yeast (Sc) *STI1* DNA from -500 to +100 with respect to the transcription initiation site. The translational start site is coloured yellow, TATA boxes are coloured red, stimulatory protein 1 (Sp1) boxes (GGGCGG) are coloured cyan, and CCAAT boxes are coloured green. Transcription initiation is indicated by the transition from lowercase to uppercase lettering. Nucleotides that are conserved between all three orthologues have been coloured dark grey, whilst those nucleotides conserved between two orthologues are shaded light grey. ClustalW (Thompson *et al.,* 1994) gives the alignment a score of 5098. BioEdit (Hall, 1999) gives human-mouse identity as 70%, human-yeast identity as 43%, and mouse-yeast identity as 44%.

Mouse

The *mSTI1* TATA (TATAAAA) box, occurring at –115 to -108 from the transcriptional start site, is not aligned with the *hSTI1* TATA box which is assumed to occur at –224 to -219. The region from -500 to +100 of the *mSTI1* gene has 7 repeats of the CCAAT sequence and 2 Sp1 sites (one upstream and one downstream of the transcriptional initiation site).

The mouse chromosome 19 has a GC% of 43, which is raised to 46% for the *mSTI1* gene, and further to 58% in the region from -500 to +100. This 600 bp region, similar to *hSTI1*, meets the criteria of a typical CpG island with a frequency of observed CG dinucleotides / frequency of expected CG dinucleotides ~ 0.7 and GC% of 58 (Table 4).

Yeast

Yeast *STI1* has a TATA box (TATATAG) at -36 to -30 with respect to the TSS (-94 to -100 bp upstream of the ATG). This roughly agrees with the yeast orthologues of *HSP70* (TATA at -160 from ATG) and *HSP90* (TATA at -134 from ATG). The region from -500 to +100 of the *ySTI1* gene has no CCAAT or Sp1 sites.

The yeast chromosome 15 has a GC% of 38, and the *ySTI1* gene has a GC content of 42%. Unlike *hSTI1* and *mSTI1*, the GC content in the region –500 to +100 bp is not higher than the *ySTI1* GC% average.

Table 4: Comparison of *STI1* orthologues with respect to composition of putative transcribed region, mRNA, and region surrounding transcriptional start site.

Organism	Sample of Nucleic Acid	Start Position	Stop Position	Length (bp)	G	C	A	T	GC%
Human	Transcribed region[(i)]	63729036	63747375	18340	4769	4286	4384	4905	49
	mRNA	-	-	2113	569	538	605	403	52
	-500 to +100[(ii)]	63728537	63729136	600	213	147	140	100	60
Mouse	Transcribed region[(iii)]	6734137	6753370	19234	4243	4594	5589	4819	46
	mRNA	-	-	2080	550	519	617	394	51
	-500 to +100[(ii)]	6733640	6734239	600	190	156	151	103	58
Yeast	Transcribed region[(iv)]	381052	382821	1770	390	348	629	403	42
	mRNA	Sequence not available							
	-500 to +100[(ii)]	380488	381087	600	91	121	217	171	35

(i) Chromosome 11 (+ strand)
(ii) Region of DNA, on relevant chromosomal strand, with respect to *STI1* transcriptional start site
(iii) Chromosome 19 (+ strand)
(iv) Chromosome 15 (- strand). As yeast *STI1* consists of a single exon, values given for the transcribed region should be essentially equivalent to those for yeast *STI1* mRNA.

3.2.2 Algorithmic Recognition Of Promoter Regions And TFBS

Several promoter prediction algorithms were used to investigate the promoter region of the *STI1* orthologues. Little correlation was found between the programs and, in addition, not every program predicted a promoter region for all sequences entered. A summary of the output of promoter prediction algorithms is shown in Table 5.

In *hSTI1*, the promoter regions predicted by NNPP, PromoterScan and TSSG are over a similar region, as are the promoter regions predicted for *mSTI1* by NNPP and TSSG. Additionally, NNPP predicts promoters in a similar region for both *hSTI1* and human *HSP70,* at -511 to -461 and at -232 to -182 relative to the *hSTI1* TSS. NNPP also predicts promoters for *mSTI1* and mouse *HSP70* at -634 to -584 relative to the *mSTI1* TSS. The promoter regions predicted by NNPP for *ySTI1* and yeast *HSP70*, near -428 to -378 relative to the *ySTI1* TSS, are close to one another but do not overlap as the human and mouse regions do.

Table 5: Promoter prediction output summary. Positions are with respect to the Eukaryotic Promoter Database (EPD) transcriptional start site.

Orthologue		*STI1*		*HSP70*		*Shuffled STI1*	
		Start	**Stop**	**Start**	**Stop**	**Start**	**Stop**
Human	**CorePromoter**(i)	-582 (0.05)		-518 (0.08)		-706 (0.281)	
	NNPP	-511 -232	-461 -182	-468 -304	-418 -254	-916 -897 -833 -825 -730	-866 -847 -783 -775 -680
	PromoterScan	-740	-490	-314	-64	-	-
	TSSG (ii)	-472 (TATA at -500)		-54		-697 (TATA at -721) -395	
Mouse	**CorePromoter**	-98 (0.997)		-198 (1)		-502 (0.037)	
	NNPP	-892 -634 -246 -125	-842 -584 -196 -75	-615 -586 -495	-565 -536 -445	-898	-848
	PromoterScan	-	-	-	-	-	-
	TSSG	-83 (TATA at -115)		-		-	
Yeast	**CorePromoter**	-225 (0.217)		-131 (0.79)		-677 (0.303)	
	NNPP	-428	-378	-744 -490 -228 -176	-694 -440 -178 -126	-951 -426	-901 -376
	PromoterScan	-	-	-	-	-	-
	TSSG	-		-137 (TATA at -167)		-	

(i) CorePromoter, optimised only for human sequences, predicts the position of the transcriptional start site (TSS) and gives it a score between 0 (low probability of being correct) and 1 (high probability of being correct). In this table only the top-scoring TSS position has been included for each *STI1* orthologue, with its score in parentheses. Only the scores for mouse *STI1* and mouse *HSP70* are high enough to consider.

(ii) TSSG also predicts transcription factors found in the putative promoter region.

Table 5 shows that, of the promoter prediction programs used, NNPP seemed to give the most reliable output: NNPP identifies the promoter region as being from -40 to +10 from the predicted TSS and thus although the TSS predicted by NNPP may not have been correct, the regions predicted were still in agreement with the regions of high density (discussed later in section 3.2.2) and conservation of the TFBS (Appendix F). NNPP also had the advantage of predicting promoter sites for all three *STI1* orthologues.

Several promoter prediction programs managed to predict promoters for the shuffled *STI1* DNA. Most predictions did not match the promoters predicted for *STI1* in number or position.

3.2.3 Analysis Of Transcription Factor Binding Sites

Of the TFs predicted for *hSTI1*, 45 are matched to *mSTI1* and 12 are matched to *ySTI1*; of these 7 sites co-occurring in all three orthologues. These seven sites include 5 Sp1 sites at -900, -741, -491, -423 and -189 relative to the *hSTI1* TSS, and sites for Nuclear Factor 1, NF-1 (-739) and HSF (-349). Note that the Sp1 site at -191 in *hSTI1* was predicted by all three prediction programs, and matches were found in all three orthologues and in human *HSP70*. Sp1 is predicted more often than any other TFBS for *hSTI1* and *mSTI1*, whilst HSF was predicted most often for *ySTI1* and followed closely in number by CCAAT/enhancer-binding proteins (C/EBP) and Sp1.

The HSF site occurring at -349 in *hSTI1* is shared with *mSTI1, ySTI1,* human *HSP70* and yeast *HSP70*. A summary of the predicted TFBS for *hSTI1*, *mSTI1* and *ySTI1* are found in Appendix F, and represented schematically in Figure 12.

A number of predicted binding sites are found to overlap or to occur very close to one another, suggesting that competitive binding between the proteins could occur. In *hSTI1*, the list of possible competitive binding pairs includes:

(i) Homeobox protein NK-2 homolog E, Nkx-2 (-957) and Cellular E26 transformation specific sequence, c-Ets (-957),
(ii) Sp1 (-741) and NF-1 (-739),
(iii) rDNA enhancer-binding protein 1, REB1 (-707) and Sp1 (-689),
(iv) Serum response factor, SRF (-504) and Sp1 (-491),
(v) Caudal-related homeodomain transcription factor, CdxA (-455), CCAAT-binding factor, CBF (-451), Sex-determining region Y, SRY (-451), GATA binding protein 1, GATA-1 (-449) and Pre-B-cell leukemia transcription factor 1, Pbx-1 (-449),
(vi) Sp1 (-372), CBF (-366), Enhancer factor I, EFI (-364), C/EBP (-351), Cellular reticuloendotheliosis proto-oncogene, c-Rel (-349), HSF (-349) and Nuclear factor kappa B, NF-kappaB (-349), and
(vii) CBF (-329), Sp1 (-323) and NF-1 (-318).

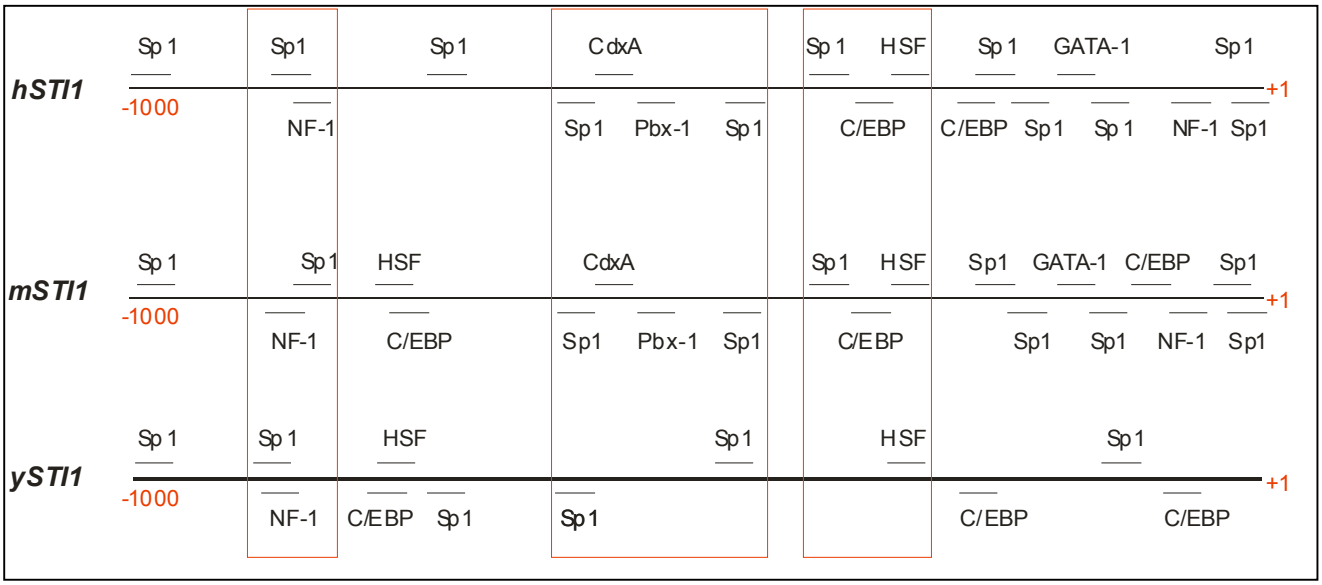

Figure 12: Schematic diagram representing some putative TFBS of interest. Drawing is not to scale. Putative TFBS are relative to one another. Red blocks indicate conserved clusters of TFBS. Numbers indicate position relative to TSS. Stimulatory protein 1 (Sp1), Nuclear factor 1 (NF-1), Caudal-related homeodomain transcription factor (CdxA), Pre-B-cell leukemia transcription factor 1 (Pbx-1), heat shock factor (HSF), CCAAT / enhancer-binding proteins (C/EBP), and GATA binding protein 1 (GATA-1) are shown.

Again, in *mSTI1*, the possible competitive binding pairs include:

(i) Nkx-2 (-969), c-Ets (-947) and GATA-1 (-946),

(ii) Sp1 (-755), NF-1 (-751) and CBF (-748),

(iii) REB1 (-703) and Sp1 (-699),

(iv) SRF (-487) and Sp1 (-481),

(v) CdxA (-469), CBF (-467), SRY (-466), GATA-1 (-464) and Pbx-1 (-464),

(vi) Sp1 (-370), CBF (-367), EFI (-365) and C/EBP (-359), and

(vii) CBF (-330), Sp1 (-324) and NF-1 (-321).

In *ySTI1*, the possible competitive binding pairs include:

(i) HSF (-760), Sp1 (-754), NF-1 (-750) and CBF2 (-743),

(ii) HSF (-510), C/EBP (-501), HSF (-498) and Sp1 (-498), and

(iii) HSF (-441), Sp1 (-435), C/EBP (-425) and Hb (-424).

Figure 13 and Figure 14 show that the density of putative TFBS in *hSTI1* and *mSTI1* is highest near -750, -450, -350, and -200 relative to the TSS. The density of putative TFBS in *ySTI1* roughly follows the same pattern. Additionally, the promoter regions

predicted by NNPP for *hSTI1* (-511 to -461 and -232 to -182), *mSTI1* (-246 to -196), and *ySTI1* (-428 to -378) correlate loosely with these regions.

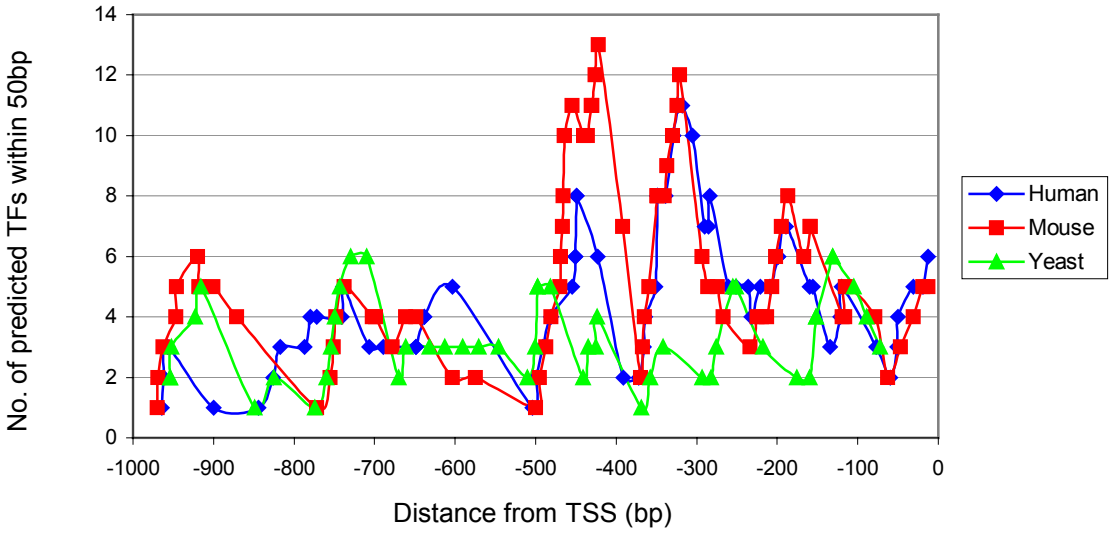

Figure 13: Graph representing the number of putative transcription factor binding sites in the region of 50 bp upstream of each predicted transcription factor binding site.

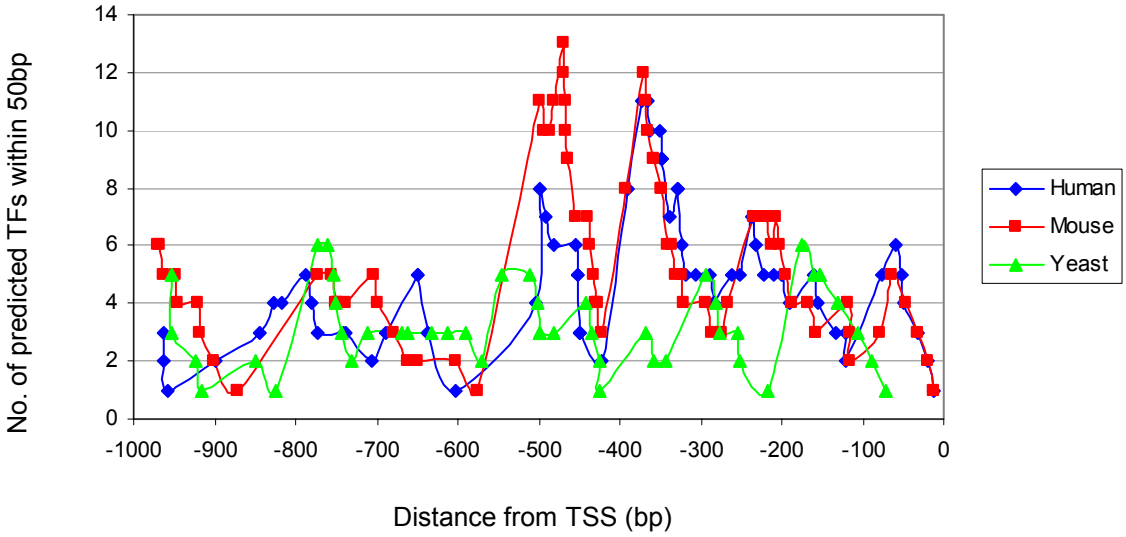

Figure 14: Graph representing the number of putative transcription factor binding sites in the region of 50 bp downstream of each predicted transcription factor binding site.

3.3 Identification Of Genes Co-Regulated With Yeast *STI1* And Determination Of Common Regulatory Motifs

The final section of Chapter 3 deals with the identification of genes that co-cluster with *ySTI1*, and with the identification of over-represented motifs in the co-regulated genes that may be involved in this regulation.

3.3.1 Genes Co-Regulated With *ySTI11*

The SGD hosts data for ten microarray conditions where the expression of *STI1* changes by more than one fold (Table 6). Data from all of these experiments, except expression during the diauxic shift, were analysed.

Table 6: Microarray conditions where the expression of yeast *STI1* changes by more than one-fold.

Data set	Maximum fold increase(i)	Maximum fold decrease(i)	Reference
Expression during sporulation	-1.0	-2.8	Chu *et al.,* 1998
Ploidy regulation of gene expression	1.4	-2.2	Galitski *et al.,*1999
Expression in response to varying zinc levels	2.4	-1.1	Lyons *et al.,* 2000
Expression during the cell cycle	2.5	-2.7	Spellman *et al.,* 1998
Expression during the unfolded protein response	2.5	1.1	Travers *et al.,*2000
Expression regulated by the calcineurin/Crz1 pathway	2.6	-2.2	Yoshimoto *et al.,* 2002
Expression during the diauxic shift	2.9	-1.4	DeRisi *et al.,* 1997
Expression in response to histone depletion	3.4	1.2	Wyrick *et al.* 1999
Expression in response to DNA-damaging agents	6.2	-2.1	Gasch *et al.,* 2000
Expression in response to environmental changes	11.3	-5.9	Gasch *et al.,* 2000

(i) Values are quoted according to the *Saccharomyces* Genome Database, July 2004.

Co-regulated Genes

The number of genes clustering with *STI1* across microarray conditions ranged from 5 to 39 (including *STI1*) with an average of 16 genes. Some of the genes co-regulated with *ySTI1* can be found in Table 7, and a list of all genes co-regulated with *ySTI1* can be found in Appendix G. A number of the genes co-expressed with *ySTI1* belong to the Hsp70 or Hsp90 family of Hsps. Other genes co-expressed with *ySTI1* include Hsp30, Hsp40, Hsp60 and protein kinases (Table 7).

Table 7: Subset of genes thought to be co-regulated with yeast *STI1*.

Locus	Additional information	Microarray experiment in which gene is upregulated with *ySTI1*
YAL005C *SSA1*	• Belongs to Hsp70 family • Role in protein folding	Calcineurin
YLL024C *SSA2*	• Belongs to Hsp70 family • Role in protein folding	Calcineurin, Zinc
YER103W *SSA4*	• Belongs to Hsp70 family • Role in protein folding	Cell cycle
YBR101C *FES1*	• Hsp70 (Ssa1p) nucleotide exchange factor	DNA damaging agents, Histone depletion, Zinc
YPL240C *HSP90*	• Belongs to Hsp90 family • Role in protein folding • Involved in negative regulation of Hsf1p • Interacts with co-chaperones Cpr6p, Sti1	Cell cycle, DNA damaging agents, Environmental stress, Histone depletion
YMR186W *HSC86*	• Belongs to Hsp90 family • Role in protein folding	DNA damaging agents, Environmental stress, Histone depletion
YDR214W *AHA1*	• Activator of Hsp90 ATPase • Role in protein folding	Environmental stress
YLR216C *CPR6*	• Binds Hsp82p • Role in protein folding	Environmental stress
YNL281W *HCH1*	• Hsp90p activator activity • Role in protein folding	Environmental stress
YGR249W *MGA1*	• Shows similarity to heat shock transcription factor	Zinc
YCR021C *HSP30*	• Role in response to stress	Zinc
YNL007C *SIS1*	• Belongs to Hsp40 family • Role in unfolded protein binding	Cell cycle, DNA damaging agents, Histone depletion, Zinc
YLR259C *HSP60*	• Role in protein folding: prevents aggregation	Environmental stress
YBR082C *UBC4*	• Mediates degradation of short-lived and abnormal proteins	Unfolded protein response
YMR104C *YPK2*	• Protein kinase	Zinc
YOL016C *CMK2*	• Calmodulin-dependent protein kinase • Role in signal transduction	Zinc

3.3.2 Recognition Of Over-Represented Motifs

Of the motifs returned by AlignACE, the program used to find sequence motifs that may be over-represented, several matches were found with currently known TFBS in Transfac (Appendix I). Of interest were the HSF, C/EBP, Antennapedia and YY1 motifs that were also recognised by the TF prediction programs for *ySTI1* (Appendix F).

CHAPTER 4: DISCUSSION AND CONCLUSION

4.1 Genomic Organisation Of The *STI1* Locus

This study showed that the genes surrounding human and mouse *STI1* were orthologous and thus *mSTI1* joins the 96% of mouse genes that can be found in a corresponding conserved human syntenic interval (Boguski, 2002). Furthermore, it was shown that this interval holds an above-average density of genes, with an average of 1 gene every 24 kb. The human genome, approximately 3,200 Mb in length and holding approximately 30,000 genes, has an average gene density of approximately 1 gene per 100 kb (International Human Genome Sequencing Consortium, 2001). The above–average gene density at the *STI1* locus suggests that *cis*-acting sequences may be shared between surrounding gene loci. Moreover, this may result in active chromatin hub formation, which is a recent concept and has been described for successive β-globulin genes (de Laat and Grosveld, 2003). In this model, inactive genes loop out in three-dimensional space, leaving the active genes in the active chromatin hub to bind the necessary combination of *cis*-regulatory elements for transcription to occur. Which genes loop out and which genes are active depends on the type and concentration of available TFs, on the specificity of their DNA binding, and how they interact with one another. The concept of genes encoded in close proximity to the *STI1* gene being co-regulated with *STI1* is of particular interest since several of these nearby genes are cancer-related (Odunuga *et al.,* 2004). Of the genes shown in this study to occur on the syntenic interval in humans and mice, three genes of particular interest have been identified for discussion in section 4.1, with reference to their possibly being regulated under the same conditions as *hSTI1* in the active chromatin hub.

The first gene that could be expected to be co-regulated with *hSTI1* is *DNAJC4.* This gene encodes an Hsp40 homologue (subfamily C, member 4) (Silins *et al.,* 1998) that may interact with Hsp70, a partner protein of Sti1. The second gene of interest is *FKBP2*, encoding the 13 kDa FK506 binding protein 2. This protein is a peptidyl-prolyl *cis*-trans isomerase (Hendrickson *et al.,* 1993) that, like *hSTI1*, assists the

folding of proteins. Even more noteworthy is that cluster analysis in this study shows that *ySTI1* is co-regulated with *CPR6* (YLR216C), also a peptidyl-prolyl *cis*-trans isomerase, under environmental stress conditions. This supports the argument that *FKBP2* could be regulated in tandem with *hSTI1*.

The third candidate for co-regulation with *hSTI1* is *PLCB3,* which encodes phospholipase C beta 3, a calmodulin-binding protein (Hempel and DeFranco, 1991). Support for the co-regulation of *hSTI1* and *PLCB3* is provided by microarray experiments on yeast, where this study recognises that, under stressful conditions, the calcineurin pathway regulates the expression of *ySTI1*, along with several Hsps and a calmodulin-dependent protein kinase (Yoshimoto *et al.,* 2002). It is thus proposed that a similar situation exists in humans and mice, whereby stress stimulates the Hsp90-dependent maturation of calcineurin, a Ca^{2+}/calmodulin-dependent protein phosphatase. Consequently, the cellular levels of Sti1 and calmodulin-binding proteins such as PLCB3, would be increased. An additional factor to consider is that calcineurin can be inhibited by immunosuppressant drugs such as FK506 (Yoshimoto *et al.,* 2002). This is of interest because it has been proposed that an FK506-binding protein, FKBP2, is also co-regulated with *hSTI1*. Lastly, it is reasonable to assume that an Hsp90 substrate such as calcineurin could regulate *STI1* expression, since another Hsp90 substrate, HSF, regulates the expression of heat shock genes (Bharadwaj *et al.,* 1999).

4.2 Intron-Exon Organisation

While the gene prediction programs used in this study predicted *ySTI1* to be a single exon, as is common of most yeast genes, *hSTI1* was predicted to comprise 15 exons and *mSTI1* was predicted to comprise either 14 exons (HMMGene) or 15 exons (Genscan). Upon translation of these exons, it appeared that the mammalian *STI1* gene has 14 exons and that the additional 15th exon, occurring 5' to the *STI1* gene, was a false prediction. Although the predicted position of mammalian *STI1* exons 2 to 14 appears to be accurate, there is evidence to suggest that the position of the first exon is less certain.

The first indication for this is that HMMGene, Genscan and *hSTI1* mRNA do not agree with regard to the position of the first exon. While HMMGene and Genscan may not be specifically designed to identify the first exon of a gene correctly, both programs identified a first exon to be upstream of the currently-accepted position. Additionally, it is possible that the mRNA sequence currently available for *hSTI1*, and used to determine the TSS, may be shorter than the *in vivo* mRNA sequence. If as few as 5 or 10 bp are omitted from the 5'-end of the mRNA, and the first intron is large – a common occurrence amongst higher eukaryotes – it may result in the first exon being completely absent from the mRNA sequence. As the EPD (Périer *et al.,* 1998; Périer *et al.,* 1999; Périer *et al.,* 2000; Praz *et al.,* 2002, Schmid *et al.,* 2004) predicts several TSS sites for *hSTI1* within close proximity to one another, this could indicate truncation of the 5' end of the mRNA. The second indication that the predicted first exon may be incorrect is that the alignment of the -500 to +100 bp region surrounding the TSS of *hSTI1* and *mSTI1* shows that their TATA boxes have not been conserved with regard to sequence or position (Figure 15). This observation is particularly noteworthy given the importance of the TATA box in transcription initiation (Wang *et al.,* 1996), and that the analysis of regulatory elements shows a high degree of conserved TFBS in the 1000 bp region upstream of the TSS (Figure 15). Lastly, it has been shown that neither *hSTI1* nor *mSTI1* has a consensus Inr region spanning the TSS, a consensus DPE, or a consensus translational start site. While these factors are not critical, it is reasonable to expect that at least some of the important regulatory signals for transcription and translation initiation should follow the consensus pattern. That they are not observed suggests that the TSS and the first exon may not have been correctly predicted.

Despite evidence suggesting that the position of the currently-acknowledged position of the first exon may need to be revised, at least three factors support the current positions of the *hSTI1* and *mSTI1* first exon. Firstly, alignment of the region -500 to +100 relative to the TSS shows a perfect conservation between *hSTI1* and *mSTI1* across the putative TSS and translation start sites, indicating that these sites are most probably biologically functional. Secondly, it may be possible for transcription to initiate at a number of sites close to one another (Jacquet *et al.,* 1989) and thus the mRNA sequence may not actually be 5'-truncated, as was previously suggested in the previous paragraph. Thirdly, this study shows an increase in GC content over the

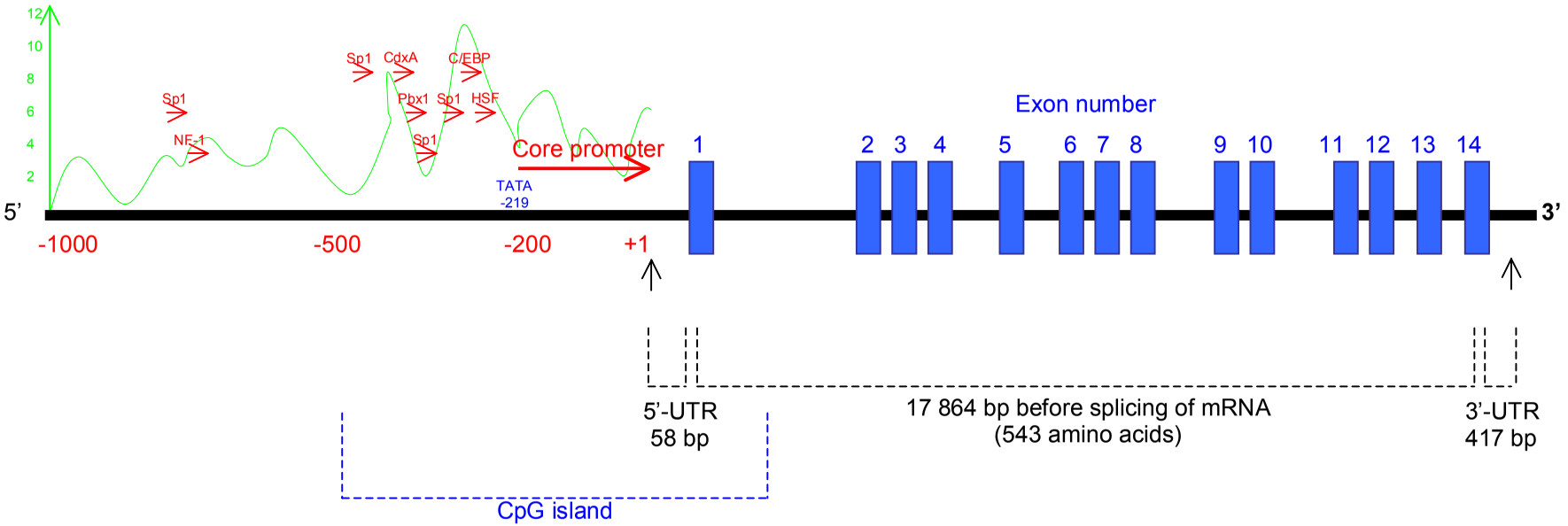

Figure 15: Schematic diagram summarizing the *hSTI1* gene features. The transcription start and stop sites are indicated, as well as the number and position of the exons, and the length of the transcribed region. The putative promoter containing region 1000 bp upstream of the transcription start site has been included. The green curve indicates the density of putative transcription factor binding sites determined using a 50 bp sliding window. Transcription factors found to be conserved in terms of type and relative position in *hSTI1*, *mSTI1* and *ySTI1* are indicated (→ red arrows). Based on the position of the TATA box, taking into account the density and position of transcription factors, and assuming a proximity of approximately 200 bp upstream of the transcription start site, the predicted core promoter region is given. This drawing is not to scale.

region -500 to +100 bp for *hSTI1* and *mSTI1*, leading to the suggestion that these regions may be CpG islands. Approximately 50% of human and mouse genes are associated with CpG islands, which is why CpG islands have been used to identify vertebrate promoters (Hyatt *et al.,* 2000a, 2000b; Xu *et al.,* 1999). The *in vivo* mRNA sequence, the TSS, and the TATA boxes, are best confirmed experimentally.

4.3 Promoter And Regulatory Elements

The promoter region was estimated by manual consideration of the type, position and density of transcription factors in the region upstream of the putative TSS of all three orthologues. This method, known as phylogenetic footprinting, can help to determine which TFBS are functionally relevant because functionally relevant DNA sequences tend to accumulate mutations at a slower rate than neutral sequences. Because yeast diverged from the ancestors of metazoans much earlier than mice diverged from humans, yeast has had more time to collect mutations and thus it is not surprising that there are fewer corresponding TFBS between yeast and the metazoans than between *hSTI1* and *mSTI1*.

The region of approximately 200 bp upstream of the *hSTI1* TSS was identified as the core promoter region because it includes the TATA box and has a high TFBS density, including a number of Sp1 and CCAAT sites responsible for basal transcription. For the purpose of uniformity, this region was also defined as the core promoter region for *mSTI1* and *ySTI1*.

While the core promoter may hold the necessary elements for basal transcription, analysis of TFBS density and conservation between orthologues has identified several regions upstream of the core promoter which are proposed to regulate *STI1* gene expression. These putative regulatory regions are conserved clusters of TFBS and regions of high TFBS density (Figure 15), where TFs in close proximity could interact, either synergistically or competitively (mutually exclusively), as a method of controlling and fine-tuning gene expression. Of particular interest is the conserved cluster containing the HSF binding site, at -349 in *hSTI1*, as HSF is known to regulate Hsp gene expression.

Although analysis of DNA sequences by promoter prediction software is time-efficient and easy, more valuable information is generated using TFBS predicition software and phylogenetic footprinting. However, predicting biologically important TFBS using TFBS prediction software is labour-intensive because most programs, at default settings, have compromised specificity for sensitivity and thus the number of putative TFBS predicted is enormous. Besides the time inefficiency of this method, three other important problems are associated with this method. Firstly, different TFBS programs make use of different TF databases. For example, TFSearch may make use of Transfac and TSSG may make use of the Gosch database. The accession numbering is different in the different databases and makes cross-examination of the program output difficult. Secondly, some of the programs were written a number of years ago and make use of outdated versions of the TF databases. Thus, TFs that have recently been discovered will not be recognised by the search engine. Lastly, some programs seem to output the name of a TF protein whilst others output the name of the DNA element to which the TF protein binds. These do not necessarily have the same name. This last point includes the observation that some programs will output the name of a multi-component TF for a particular site, whilst others will predict a subunit of the same complex to bind to the site. An example of this is the prediction of NF-kappaB as opposed to p65 or RelA. Also, different TFs with different roles in the cell may be able to bind competitively for the same TFBS.

Problems were also encountered in this study when searching for putative TFBS by phylogenetic footprinting. Because some of the yeast gene clusters used in AlignACE contained few genes, the AlignACE MAP scores were low despite the possibility that the motifs could be biologically significant. It was for this reason that all returned motifs were considered regardless of the MAP score. The reason that few motifs were matched by Transfac could be due to several reasons: the motifs may not be present in Transfac in any form and are thus not recognised as being functional TFBS; a PWM may have been a better way to represent the motifs returned by AlignACE than a consensus sequence, as a mismatch in any one position in a motif would result in Transfac not recognizing the motif; and the length of the motifs may differ from those present in Transfac and thus, for example, if the AlignACE motif is 12 nucleotides long but the motif in Transfac is only 10 nucleotides long, a match may not occur.

The success of phylogenetic footprinting as a method of TFBS identification is dependent on finding genes, co-expressed with *ySTI1*, whose products are required to function alongside *ySTI1* – and then searching for over-represented motifs in the upstream region of these genes. In this study, many of the genes identified as being co-expressed with *ySTI1* are likely to contribute toward over-represented motifs that are biologically important. These genes include the Hsp70 family proteins, Hsp90 family proteins, Hsp-associated genes, and Hsp90 substrates (eg. protein kinases) which are involved in protein folding pathways (discussed in section 1.1.3).

To demonstrate the value of phylogenetic footprinting, four of the over-represented motifs that were identified in Transfac, were found to have been predicted by TFBS prediction programs: HSF, C/EBP, YY1 and possibly Antennapedia (as indicated by Hb and Ftz, the homeobox transcription factors, in Appendix F). Thus, phylogenetic footprinting could be a useful tool for recognizing putative TFBS in the upstream region of co-regulated genes.

4.4. Conclusion And Future Work

The aims of this project were to predict the gene structure of the *STI1* orthologues, to predict the regulatory elements upstream of the *STI1* orthologues and upstream of genes co-regulated with *ySTI1*, and finally to collate the information and draw a simple schematic diagram depicting the main results of the research. With regard to the gene structure of the *STI1* orthologues, it has been shown that *hSTI1* and *mSTI1* share a number of gene features, including common genes surrounding the loci that may be regulated in tandem with mammalian *STI1*. The promoter regions of the *STI1* orthologues have been predicted, and in addition several conserved clusters of TFBS have been identified upstream of the *STI1* promoters as being putative regulatory regions. Yeast *STI1* shares several gene features with mammalian *STI1*, although less conservation is observed. Genes thought to be co-regulated with *ySTI1* consist of several Hsp genes, Hsp-associated genes, and Hsp90 substrates. Several over-represented motifs, indicating possible TFBS, were identified in the upstream regions of the genes co-expressed with *ySTI1*. The main findings of this research are represented in Figure 15.

Thus, this research has provided a comprehensive bioinformatics analysis of the genetic structure and putative transcription factors involved in the regulation of human, mouse and yeast *STI1*. The results generated by this *in silico* research provide a foundation for future experimental work, including transgenic and knockout studies to investigate the biological importance of the human and mouse Sti1 proteins. Five main areas of future work are suggested:

i. The biologically functional TSS and first exon should be verified. Included in this area of research could be the determination of the TATA box, and also the degree of methylation of the CG repeats, by means of restriction enzyme analysis, to determine whether the region surrounding the TSS is indeed a CpG island.

ii. The second area of research could be to analyse the type, position and density of TFs binding downstream of the TSS. Further *in silico* investigation concerning the TF density is also recommended. It is noted that, whilst an alignment of the region -500 to +100 of *hSTI1* and *mSTI1* showed several conserved CCAAT boxes and Sp1 sites, these were not all included in Figure 15. The reason for this was that the TFBS prediction algorithms could not always identify a CCAAT box in these regions (because they identify sites longer than the simple 5-bp CCAAT) or because they could have identified different CCAAT-binding proteins with different names. If these TFBS were included in the TF density profile calculations, it could change the profile such that the region immediately upstream of the TSS showed a higher density of TFs.

iii. Linked to the abovementioned *in silico* investigations, the *in vitro* and *in vivo* role of a range of different TFs could be investigated, to determine the general conditions that usually regulate *STI1*, and the TFs required to do so. The TFBS with the most potential seems to be the HSF site occurring at -349 in *hSTI1,* which was shared with *mSTI1, ySTI1,* human *HSP70* and yeast *HSP70*. This high degree of conservation suggested that this specific HSF-binding site was likely to be biologically functional.

Besides the HSF binding site, the role of Sp1, CCAAT, NF-kappaB, and cyclic adenosine monophosphate response element-binding protein (CRE-BP) binding sites are suggested as a starting point for investigating the functional TFBS upstream of the *STI1* orthologues, as certain of these proteins are known to interact with one another. For example, Sp1 is known to bind NF-kappaB, of which several sites have been predicted in *hSTI1* and *mSTI1*. NF-kappaB is the main TF in response to inflammatory cytokines and plays a role in apoptosis. Reports reveal that pathways leading to apoptosis and stress response are linked and it would be of interest to investigate this further. NF-kappaB also interacts with the CRE-BP, a component of basal transcription machinery, of which several sites have also been predicted in *hSTI1* and *mSTI1*. Additionally, NF-kappaB, C/EBP and CRE-BP could possibly be inhibited by glucocorticoids. Glucocorticoid receptors (GR) are substrates of the Hsp70/Hsp90 chaperone complex in which Sti1 is involved. Thus, while GR binding sites do not seem to be common or conserved between the *STI1* orthologues, it is possible that this Hsp90 substrate is still involved in regulating *STI1* gene expression.

In addition to the more obvious TFBS, a number of TFBS predicted by the prediction programs were not shown to be conserved between the *STI1* orthologues, but could potentially be very interesting. While they may not be responsible for the standard means of *STI1* expression, they may be able to influence the expression under more special or disease circumstances. For example, in *hSTI1* a binding site for Wilms tumor suppressor protein 1 (WT1) was predicted. *WT1* has been mapped to the human chromosome 11q13, near that of *hSTI1*. WT1 induces G1 phase arrest, has a role in cell differentiation, and may cause Wilms tumor (nephroblastoma). Also of interest are Ste11 (an Hsp90 substrate), the upstream regulatory factor (USF) which interacts with TFIID, and alcohol dehydrogenase regulatory protein 1 (Adr1) in *ySTI1* because of the link between the cell stress response and the requirement for alcohol catabolism.

iv. From the available literature on the role of ySti1 protein in the cell, the change in *ySTI1* expression in response to conditions such as the cell cycle, unfolded protein response and environmental changes came as no surprise as *STI1* is known to be involved in G1/S arrest and unfolded proteins can induce the heat

shock response in eukaryotes (Ananthan *et al.,* 1986). The role of *STI1* in response to other experimental conditions is less clear and should be investigated. The calcineurin experiment is a particularly interesting experiment, as already discussed.

v. Last but not least, a fifth area of potential research could be that of the regulation of the genes surrounding mammalian *STI1*, related to the spatial and temporal regulation of *STI1*.

REFERENCES

Abbas-Terki,T., Briand,P-A., Donzé,O. and Picard,D. (2002) The Hsp90 Co-Chaperones Cdc37 and Sti1 Interact Physically and Genetically. *Biol.Chem.,* **383,** 1335-1342.

Abbas-Terki,T., Donzé,O. Briand,P-A., and Picard,D. (2001) Hsp104 Interacts with Hsp90 Cochaperones in Respiring Yeast. *Mol. Cell. Biol.,* **21,** 7569-7575.

Ananthan,J., Goldberg,A.L. and Voellmy,R. (1986) Abnormal proteins serve as eukaryotic stress signals and trigger the activation of heat shock genes. *Science,* **232,** 522-524.

Antequera,F. and Bird,A. (1993) Number of CpG Islands and Genes in Human and Mouse. *Proc. Natl. Acad. Sci. USA,* **90,** 11995-11999.

Bajic,V.B., Choudhary,V. and Hock,C.K. (2003) Content analysis of the core promoter region of human genes. *In Silico Biology*, **4,** 0011 (electronic version).

Bardwell,J.C.A. and Craig,E.A. (1988) Ancient heat shock gene is dispensable. *J. Bacteriol.,* **170,** 2977-2983.

Bharadwaj,S., Ali,A. and Ovsenek,N. (1999) Multiple components of the Hsp90 chaperone complex function in regulation of Heat Shock Factor 1 *in vivo*. *Mol. Cell. Biol.,* 19, 8033-8041.

Bird,A.P., Taggart,M.H., Nicholls,R.D. and Higgs,D.R. (1987) Non-methylated CpG-rich islands at the the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO J.,* **6,** 999-1004.

Blatch,G.L. and Lässle,M. (1999) The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays*, 21, 932-939.

Blatch,G.L., Lässle,M., Zetter,B.R. and Kundra,V. (1997) Isolation of a mouse cDNA encoding *mSTI1*, a stress-inducible protein containing the TPR motif. *Gene,* **194,** 277-282.

Boguski,M.S. (2002) Comparative genomics: The mouse that roared. *Nature,* **420,** 515 – 516.

Breathnach,R. and Chambon,P. (1981) Organization and expression of eucaryotic split genes coding for proteins. *Annu. Rev. Biochem.,* **50,** 349-383.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.,* **212**, 563-578.

Buchner, J. (1999) Hsp90 & co. – a holding for folding. *Trends Biochem. Sci.,* **24,** 136-141.

Burge,C. and Karlin,S. (1997) Prediction of complete genome structures in human genomic DNA. *J. Mol. Biol.,* **268,** 78-94.

Burke,T.W. and Kadonaga,J.T. (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev.,* **11,** 3020–3031.

Carey,M. and Smale,S.T. (2000) Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Catelli,M.G., Binart,N., Jung-Testas,I., Renoir,J.M., Baulieu,E.E., Feramisco,J.R. and Welch,W.J. (1985) The common 90-kd protein component of non-transformed '8S' steroid receptors is a heat-shock protein. *EMBO J.,* **4,** 3131-3135.

Chang,H.J. and Lindquist,S. (1994) Conservation of Hsp90 macromolecular complexes in *Saccharomyces cerevisiae*. *J. Biol. Chem.,* **269,** 24983-24988.

Chang,H.J., Nathan,D.F. and Lindquist,S. (1997) *In vivo* analysis of the Hsp90 cochaperone Sti1 (p60). *Mol. Cell. Biol.,* **17,** 318-325.

Cheetham,M.E., Jackson,A.P. and Anderton,B.H. (1994) Regulation of 70-kDa heat-shock-protein ATPase activity and substrate binding by human DnaJ-like proteins, HSJ1a and HSJ1b. *Eur. J. Biochem.,* **226,** 99-107.

Chen,S. and Smith,D.F. (1998) Hop as a adaptor in the heat shock protein 70 (Hsp70) and Hsp90 chaperone machinery. *J. Biol. Chem.,* **272,** 35194-35200.

Chen,S., Prapapanich,V., Rimerman,R.A., Honoré,B. and Smith,D.F. (1996) Interactions of p60, a mediator of progesterone assembly, with heat shock proteins Hsp90 and Hsp70. *Mol. Endocrinol.,* **10,** 682-693.

Cheung,J. and Smith,D.F. (2000) Molecular chaperone interactions with steroid receptors: an update. *Mol. Endocrinol.,* **14,** 939-946.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science,* **282,** 699-705.

Cyr,D.M., Lu,X. and Douglas,M.G. (1992) Regulation of Hsp70 function by a eukaryotic DnaJ homolog. *J. Biol. Chem.,* 267, 20927-20931.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* **5,** suppl. 3, 345-352.

de Laat,W. and Grosveld,F. (2003) Spatial organization of gene expression: The active chromatin hub. *Chromosome Res.* **11,** 447-459.

DeFranco,D.B. (2000) Role of molecular chaperones in subnuclear trafficking of glucocorticoid receptors. *Kidney Internat*., **57,** 1241-1249.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* **278,** 680-686.

Dittmar,K.D., Banach,M., Galigniana,M.D. and Pratt,W.B. (1998) The role of DnaJ-like proteins in glucocorticoid receptor-Hsp90 heterocomplex assembly by the reconstituted Hsp90-p60-Hsp70 foldosome complex. *J. Biol. Chem.,* **273,** 7358-7366.

Dittmar,K.D., Demady,D.R., Stancato,L.F., Krishna,P. and Pratt,W.B. (1997) Folding of the glucocorticoid receptor by the heat shock protein (Hsp) 90-based chaperone machinery. *J. Biol. Chem.,* **272,** 21213-21220.

Dittmar,K.D., Hutchison,K.A., Owens-Grillo,J.K. and Pratt,W.B. (1996) Reconstitution of the steroid receptor-Hsp90 heterocomplex assembly system of rabbit reticulocyte lysate. *J. Biol. Chem.,* **271,** 12833-12839.

Donzé,O. and Picard,D. (1999) Hsp90 Binds and Regulates the Ligand-Inducible Subunit of Eukaryotic Translation Initiation Factor Kinase Gcn2. *Mol. Cell. Biol.,* **19,** 8422-8432.

Down,T.A. and Hubbard,T.J.P. (2004) What can we learn from noncoding regions of similarity between genomes. *BMC Bioinformatics,* **5,** 131-137.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA,* **95,** 14863-14868.

Ellis,R.J. (1997) Do molecular chaperones have to be proteins? *Biochem. Biophys. Res. Commun.,* **238,** 687-692.

Ellis,R.J. (1999) Molecular chaperones: pathways and networks. *Curr. Biol.,* **9,** R137-139.

Ellis,R.J. and Hartl,F.U. (1999) Principles of protein folding in the cellular environment. *Curr. Opin. Struct. Biol.,* **9,** 102-10.

Fernandes,M., O'Brian,T. and Lis,J.T. (1994). Structure and regulation of heat shock gene promoters. In The biology of heat shock proteins and molecular chaperones (ed. R.I. Morimoto, A. Tissieres and C. Georgopoulos), pp. 375-393. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Fessele,S., Majer,H., Zischek,C., Nelson,P.J. and Werner,T. (2002) Regulatory context is a crucial part of gene function. *Trends Genet.,* **18,** 60-63.

Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter prediction. *Genome Res.,* **7,** 861-878.

Flynn,G.C., Chappell,T.G. and Rothman,J.E. (1989) Peptide binding and release by proteins implicated as catalysts of protein assembly. *Science,* **245,** 385-390.

Frech,K., Quandt,K. and Werner,T. (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.,* **22,** 103-104.

Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics,* **17,** 878-889.

Frydman,J. (2001) Folding of newly translated proteins *in vivo*: the role of molecular chaperones. *Annu. Rev. Biochem.,* **70,** 603-647.

Fukue,Y., Sumida,N., Nishikawa,J. and Ohyama,T. (2004) Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. *Nucleic Acids Res.,* **32,** 5834-5840.

Gailus-Durner,V., Scherf,M. and Werner,T. (2001) Experimental data of a single promoter can be used for *in silico* detection of genes with related regulation in the absence of sequence similarity. *Mamm. Genome,* **12,** 67-72.

Galitski,T., Saldanha,A.J., Styles,C.A., Lander,E.S., and Fink. (1999) Ploidy regulation of gene expression. *Science,* **285,** 251-254.

Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertegrate genomes. *J. Mol. Biol.,* **196,** 261-282.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.,* **12,** 4241-42.

Gebauer,M., Zeiner,M. and Gehring,U. (1997) Proteins interacting with the molecular chaperones Hsp70/hsc70: physical associations and effects on refolding activity. *FEBS,* **417,** 109-113.

Georgopoulos,C. and Welch,W.J. (1993) Role of the major heat shock proteins as molecular chaperones. *Annu. Rev. Cell. Biol.,* **9,** 601-34.

Goldberg,M.L. (1979) Sequence analysis of *Drosophila* histone genes. Ph.D. Dissertation, Stanford University.

Grabe,N. (2002) AliBaba2: Context specific identification of transcription factor binding sites. *In Silico Biol.*, **2,** S1-1.

Grad,Y.H., Roth,F.P., Halfon,M.S. and Church,G.M. (2004) Prediction of similarly-acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *D.melanogaster* and *D.pseudoobscura*. *Bioinformatics,* **20,** 2738-2750.

Grenert,J.P., Johnson,B.D. and Toft,D.O. (1999) The importance of ATP binding and hydrolysis by Hsp90 in formation and function of protein heterocomplexes. J. *Biol. Chem.,* **274,** 17525-17533.

Grenert,J.P., Sullivan,W.P., Fadden,P., Haystead,T.A.J., Clark,J., Mimnaugh,E., Krutzsch,H., Ochel,H., Schulte,T.W., Sausville,E., Neckers,L.M. and Toft,D.O. (1997) The amino-terminal domain of heat shock protein 90 (Hsp90) that binds geldanamycin is an ATP/ADP switch domain that regulates Hsp90 conformation. *J. Biol. Chem.,* **272,** 23843-23850.

Gross,M. and Hessefort,S. (1996) Purification and characterization of a 66-kDa protein from rabbit reticulocyte lysate which promotes the recycling of Hsp70. *J. Biol. Chem.,* **271,** 16833-16841.

Hall,T.A. (1999) BioEdit: a user-friendly biological sequence alignment *Nucl. Acids. Symp. Ser.,* **41,** 95-98.

Hannenhalli,S. and Levy,S. (2001) Promoter prediction in the human genome. *Bioinformatics,* **17,** S90-S96.

Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) Databases on transcriptional regulation: Transfac, TRRD, and COMPEL. *Nucleic Acids Res,*. **26,** 364-370.

Hempel,W.M. and DeFranco,A.L. (1991) Expression of phospholipase C isozymes by murine B lymphocytes. *J. Immunol.,* **146,** 3713-3720.

Hendrickson,B.A., Zhang,W., Craig,R.J., Jin,Y.J., Bierer,B.E., Burakoff,S. and DiLella,A.G. (1993) Structural organization of the genes encoding human and murine FK506-binding protein (FKBP) 13 and comparison to FKBP1. *Gene,* **134,** 271-275.

Höhfeld,J., Minami,Y. and Hartl,F.U. (1995) Hip, a novel chaperone involved in the eukaryotic Hsc70/Hsp40 reaction cycle. *Cell,* **83,** 589-598.

Honoré, B., Leffers,H., Madsen,P., Rasmussen,H.H., Vandekerckhove,J.V. and Celis,J.E. (1992) Molecular cloning and expression of a transformation-sensitive human protein containing the TPR motif and sharing identity to the stress-inducible yeast protein *STI1*. *J. Biol. Chem.*, **267,** 8485-8491.

Houry,W.A. (2001) Chaperone-assisted protein folding in the cell cytoplasm. *Curr. Protein Pept. Sci.,* **2,** 227-244

Hu,J. and Seeger,C. (1996) Hsp90 is required for the activity of a hepatitis B virus reverse transcriptase. *Proc. Natl. Acad. Sci. USA,* **93,** 1060-1064.

Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae. J. Mol. Biol.,* **296,** 1205-1214.

Hurowitz,E.H. and Brown,P.O. (2003) Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae. Genome Biology,* **5,** R2

Hutchinson,G. (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput. Applic. Biosci.,* **12,** 391-398.

Hutchison,K.A., Stancato,L.F., Owens-Grillo,J.K., Johnson,J.L., Krishna,P., Toft,D.O. and Pratt,W.B. (1995) The 23-kDa acidic protein in reticulocyte lysate is the weakly bound component of the Hsp foldosome that is required for assembly of the glucocorticoid receptor into a functional heterocomplex with Hsp90. *J. Biol. Chem.,* **270,** 18841-18847.

Hyatt,D., Snoddy,J., Schmoyer,D., Chen,G., Fischer,K., Parang,M., Vokler,I., Petrov,S., Locascio,P., Olman,V., Land,M., Shah,M. and Uberbacher,E. (2000a) Improved analysis and annotation tools for whole-genome computational annotation and analysis: GRAIL-EXP. *Genome Analysis Toolkit and Related Analysis Tools, Genome Sequencing & Biology Meeting*.

Hyatt,D., Snoddy,J., Schmoyer,D., Chen,G., Fischer,K., Parang,M., Vokler,I., Petrov,S., Locascio,P., Olman,V., Land,M., Shah,M. and Uberbacher,E. (2000b) GRAIL-EXP and the Genome Analysis Toolkit. *The 13th Annual Cold Spring Harbor Meeting on Genome Sequencing & Biology.*

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome, *Nature,* **409,** 860 – 921.

Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genetics,* **26**, 61-63.

Jacquet,M.A., Ehrlich,R. and Reiss,C. (1989) *In vivo* gene expression directed by synthetic promoter constructions restricted to the -10 and -35 consensus hexamers of *E. coli. Nucleic Acids Res.,* **17,** 2933-2945.

Jans,D.A. and Jans,P. (1994) Negative charge at the casein II kinase site flanking the nuclear localization signal of the SV40 large T-antigen is mechanistically important for enhanced nuclear import. *Oncogene*, **9,** 2961-2968.

Javahery,R., Khachi,A., Lo,K., Zenzie-Gregory,B. and Smale,S.T. (1994) DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.,* **14,** 116-127.

Jensen,S.T. and Liu,J.S. (2004) BioOptimizer: A Bayesian scoring function approach to motif discovery. *Bioinformatics,* **20,** 1557-1564.

Joab,I., Radanyi,C., Renoir.,M., Buchou,T., Catelli,M.G., Binart.N., Mester,J. and Baulieu,E.E. (1984) Common non-hormone binding component in non-transformed chick oviduct receptors of four steroid hormones. *Nature,* **308,** 850-853.

Johansson,O., Alkema,W., Wasserman,W.W. and Lagergren,J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics,* **19,** i169-i176.

Johnson,B.D., Schumacher,R.J., Ross,E.D. and Toft,D.O. (1998) Hop modulates Hsp70/Hsp90 interactions in protein folding. *J. Biol. Chem.*, **273,** 3679-3686.

Johnson,J.L. and Toft,D.O. (1995) Binding of p23 and Hsp90 during assembly with progesterone receptor. *Mol. Endocrinol.,* **9,** 670-678.

Jolly,C. and Morimoto,R.I. (2000) Role of the heat shock response and molecular chaperones in oncogenesis and cell death. *J. Natl. Cancer Inst.,* **92,** 1564-1572.

Kimmins,S. and MacRae,T.H. (2000) Maturation of steroid receptors: an example of functional cooperation among molecular chaperones and their associated proteins. *Cell Stress Chaperones,* **5,** 76-86.

Klingenhoff,A., Frech,K., Quandt,K. and Werner,T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics,* **15,** 180-186.

Knudsen,S. (1999) Promoter2.0: for the recognition of Pol III promoter sequences. *Bioinformatics,* **15,** 356-361.

Kosano,H., Stensgard,B., Charlesworth,M.C., McMahon,N. and Toft,W. (1998) The assembly of progesterone receptor-Hsp90 complexes using purified proteins. *J. Biol. Chem.,* **273,** 32973-32979.

Kozak,M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.,* **15,** 8125-8148.

Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. In *Proc. of Fifth Int. Conf. on Intelligent Systems for Molecular Biology*, ed. Gaasterland,T. *et al.,* Menlo Park, CA: AAAI Press, 179-186.

Lareau,L.F., Green,R.E., Bhatnagar,R.S. and Brenner,S.E. (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.,* **14,** 273-282.

Lässle,M., Blatch,G.L., Kundra,V., Takatori,T. and Zetter,B.R. (1997) Stress-inducible, murine protein *mSTI1*. *J. Biol. Chem.*, **272,** 1876-1884.

Lavorgna,G., Boncinelli,E., Wagner,A. and Werner,T. (1998) Detection of potential target genes *in silico*. *Trends Genet.,* **14,** 375-376.

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

Longshaw,V.M., Chapple,J.P., Balda,M.S., Cheetham,M.E. and Blatch,G.L. (2004) Nuclear translocation of the Hsp70/Hsp90 organizing protein *mSTI1* is regulated by cell cycle kinases. *J. Cell. Sci.,* **117,** 701-710.

Lyons,T.J., Gasch,A.P., Gaither,L.A., Botstein,D., Brown,P.O. and Eide,D.J. (2000) Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc. Natl. Acad. Sci. USA,* **97,** 7957-7962.

Mason,P.B. and Lis,J.T. (1997). Cooperative and competitive protein interactions at the Hsp70 promoter. *J. Biol. Chem.,* **272,** 33227-33233.

McLauchlan,J., Gaffney,D., Whitton,J.L. and Clements,J.B. (1985) The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. *Nucleic Acids Res.,* **13,** 1347-1368.

Morimoto,R.I. (1998) Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. *Genes Dev.,* **12,** 3788-3796.

Mount,D.W. (2001) Bioinformatics: Sequence and genome analysis. Cold Spring Harbor Laboratory Press. p337-379.

Nadeau,K., Das,A. and Walsh,C.T. (1993) Hsp90 chaperonins possess ATPase activity and bind heat shock transcription factors and peptidyl prolyl isomerases. *J. Biol. Chem.,* **268,** 1479-1487.

Nair,S.C., Rimerman,R.A., Toran,E.J., Chen,S., Prapapanich,V., Butts,R.N. and Smith,D.F. (1997) Molecular cloning of human FKBP51 and comparisons of immunophilin interactions with Hsp90 and progesterone receptor. *Mol. Cell. Biol.,* **17,** 594-603.

Nair,S.C., Toran,E.J., Rimerman,R.A., Hjermstad,S., Smithgall,T.E. and Smith,D.F. (1996) A pathway of multi-chaperone interactions common to diverse regulatory proteins: estrogen receptor, Fes tyrosine kinase, heat shock transcription factor Hsf1, and the aryl hydrocarbon receptor. *Cell Stress Chaperones,* **1,** 237-250.

Nathan,D.F., Vos,M.H. and Lindquist,S. (1997) *In vivo* functions of the *Saccharomyces cerevisiae* Hsp90 chaperone. *Proc. Natl. Acad. Sci. USA,* **94,** 12949-12956.

Nathan,D.F., Vos,M.H. and Lindquist,S. (1999) Identification of *SSF1, CNS1*, and *HCH1* as multicopy suppressors of a *Saccharomyces cerevisiae* Hsp90 loss-of-function mutation. *Proc. Natl. Acad. Sci. USA,* **96,** 1409-1414.

Nicolet,C.M. and Craig,E.A. (1989) Isolation and characterization of *STI1*, a stress-inducible gene from *Saccharomyces cerevisiae*. *Mol. Cell. Biol.,* **9,** 3638-3646.

Nikolov,D.B. and Burley,S.K. (1997) RNA polymerase II transcription initiation: a structural view. *Proc. Natl. Acad. Sci. USA,* **94,** 15-22.

Obermann,W.M.J., Sondermann,H., Russo,A.A., Pavletich,N.P. and Hartl,F.U. (1998) *In vivo* function of Hsp90 is dependent on ATP binding and ATP hydrolysis. *J. Cell Biol.,* **143,** 901-910.

Odunuga,O.O., Longshaw,V.M. and Blatch, GL. (2004) Hop: more than an Hsp70/Hsp90 adaptor protein. *BioEssays,* **26,** 1058-1068.

Owens-Grillo,J.K., Hoffmann,K. and Hutchison,K.A. (1995) The cyclosporin A-binding immunophilins CyP40 and the FK506-binding immunophilins Hsp56 bind to a common site on Hsp90 and exist in independent cytosolic heterocomplexes with the untransformed glucocorticoid receptor. *J. Biol. Chem.,* **270,** 20479-20484.

Panaretou,B., Prodromou,C., Roe,S.M., O'Brien,R., Ladbury,J.E., Piper,P.W. and Pearl,L.H. (1998) ATP binding and hydrolysis are essential to the function of the Hsp90 molecular chaperone *in vivo*. *EMBO J.,* **17,** 4829-4836.

Patterton,H-G. and Graves,S. (2000a) DNAssist, a C++ program for editing and analysis of nucleic acid and protein sequences on PC-compatible computers running Windows 95, 98, NT4.0 or 2000. *Biotechniques,* **28,** 1192-1197.

Patterton,H-G. and Graves,S. (2000b) DNAssist: the integrated editing and analysis of molecular biology sequences in Windows. *Bioinformatics,* **16,** 652-653.

Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1999) The biology of eukaryotic promoter prediction – a review. *Comput. Chem.,* **23,** 191-207.

Périer, RC., Junier, T., Bonnard, C. and Bucher, P. (1999) The Eukaryotic Promoter Database EPD: Recent Developments. *Nucleic Acids Res.,* **27,** 307-309.

Périer,R.C., Junier,T. and Bucher, P. (1998) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.,* **26,** 353-357.

Périer,RC., PrazV., Junier,T., Bonnard,C. and Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.,* **28,** 302-303.

Pirkkala,L., Nykänen,P. and Sistonen,L. (2001) Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. *FASEB J.,* **15,** 1118-1131.

Pratt,W.B. (1993) The role of heat shock proteins in regulating the function, folding, and trafficking of the glucocorticoid receptor. *J. Biol. Chem.,* **268,** 21455-21458.

Pratt,W.B. and Toft,D.O. (1997) Steroid receptor interactions with heat shock protein and immunophilins chaperones. *Endocr. Rev.,* **18,** 306-360.

Pratt,W.B. and Toft,D.O. (2003) Regulation of signaling protein function and trafficking by the Hsp90/Hsp70-based chaperone machinery. *Exp. Biol. Med.*, **228,** 111-133.

Pratt,W.B., Galigniana,M.D., Harrell,J.M. and DeFranco,D.B. (2004) Role of Hsp90 and Hsp90-binding immunophilins in signaling protein movement. *Cell. Signal.,* **16,** 857-872.

Pratt,W.B., Silverstein,A.M. and Galigniana,M.D. (1999) A model for the cytoplasmic trafficking of signaling proteins involving the Hsp90-binding immunophilins and $p50^{cdc37}$. *Cell. Signal.,* **11,** 839-851.

Praz,V., Périer,RC., Bonnard,C. and Bucher,P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.,* **30,** 322-324.

Prestridge,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.,* **249,** 923-932.

Prodromou,C. and Pearl,L.H. (2003) Structure and functional relationships of Hsp90. *Curr. Cancer Drug Targets,* **3,** 301-323.

Prodromou,C., Siligardi,G., O'Brien,R., Woolfson,D.N., Regan,L., Panaretou,B., Ladbury,J.E., Piper,P.W. and Pearl,L.H. (1999) Regulation of Hsp90 ATPase activity by tetratricopeptide repeat (TPR)-domain co-chaperones. *EMBO J.,* **18,** 754-762.

Proudfoot,.N.J. and Brownlee,G.G. (1976) 3' non-coding region sequences in eukaryotic messenger RNA. *Nature,* **263,** 211-214.

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. *Nucleic Acids Res.,* **23,** 4878-4484.

Ratajczak,T. and Carrello,A. (1996) Cyclophilin 40 (CyP-40), mapping of its Hsp90 binding domain and evidence that FKBP52 competes with CyP-40 for Hsp90 binding. *J. Biol. Chem.,* **271,** 2961-2965.

Reese,M.G. (2000) Computational prediction of gene structure and regulation in the genome of *Drosophila melanogaster*. PhD Thesis, California Berkeley/University Hohenheim.

Reese,M.G. (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.,* **26**, 51-56.

Reese,M.G. and Eeckman,F.H. (1995) Novel neural network algorithms for improved eukaryotic promoter site recognition. *The Seventh International Genome Sequencing And Analysis Conference*, South Carolina.

Rogic,S., Mackworth,A.K. and Ouellette,F.B.F. (2001) Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Res.,* **11,** 817-832.

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, **16,** 939-945.

Sanchez,E.R., Toft,D.O., Schlesinger.M.J. and Pratt,W.B. (1985) Evidence that the 90-kDa phosphoprotein associated with the untransformed L-cell glucocorticoid receptor is a murine heat shock protein. *J. Biol. Chem.,* **260,** 12398-12401.

Scheibel,T. and Buchner,J. (1998) The Hsp90 complex – a super-chaperone machine as a novel drug target. *Biochem. Pharmacol.,* **56,** 675-682.

Scheibel,T., Weikl,T. and Buchner,J. (1998) Two chaperone sites in Hsp90 differing in substrate specificity and ATP dependence. *Proc. Natl. Acad. Sci. USA,* **95,** 1495-1499.

Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.,* **297,** 599-606.

Scheufler,C., Brinker,A., Bourenkov,G., Pegorano,S., Moroder,K., Bartunik,H., Hartl,F.U. and Moarefi,I. (2000) Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell,* **101,** 199-210.

Schmid,C.D., Praz,V., Delorenzi,M., Périer,R. and Bucher,P. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res.,* **32,** D82-85.

Scholz,G.M., Cartledge,K. and Hall,N.E. (2001) Identification and characterization of Harc: A novel Hsp90 associating relative of Cdc37. *J. Biol. Chem.,* **276,** 30971-30979.

Schug,J. and Overton,G.C. (1997a) Technical Report CBIL-TR-1997-1001-v0.0. Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania.

Schug,J. and Overton,G.C. (1997b) TESS:Transcription Element Search Software on the WWW. Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, Philadelphia.

Shue,G. and Kohtz,D.S. (1994) Structural and functional aspects of basic helix-loop-helix protein folding by heat-shock protein 90. *J. Biol. Chem.,* **269,** 2707-2711.

Silins,G.,Grimmond,S. and Hayward,N. (1998) Characterisation of a new human and murine member of the DnaJ family of proteins. *Biochem. Biophys. Res. Commun.,* **243,** 273-276.

Smale,S.T., Jain,A., Kaufmann,J., Emami,K.H., Lo,K. and Garraway,I.P. (1998) The initiator element: a paradigm for core promoter heterogeneity within metazoan protein-coding genes. *Cold Spring Harbour. Symp. Quant. Biol.,* **63,** 21–31.

Smale,S.T., Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem*., **72**, 449-479.

Smith,D.F. (1993) Dynamics of heat shock protein 90-progesterone receptor binding and the disactivation loop model for steroid receptor complexes. *Mol. Endocrinol.,* **7,** 1418-1429.

Smith,D.F. (2000) Chaperones in progesterone receptor complexes. *Semin. Cell. Dev. Biol.,* **11,** 45-52.

Smith,D.F., Schowalter,D.B., Kost,S.L. and Toft,D.O. (1990) Reconstitution of progesterone receptor with heat shock proteins. *Mol. Endocrinol.,* **4,** 1704-1711.

Smith,D.F., Stensgard,B.A., Welch,W.A. and Toft,D.O. (1992) Assembly of progesterone receptor with heat shock proteins and receptor activation are ATP mediated events. *J. Biol. Chem.,* **267,** 1350-1356.

Smith,D.F., Sullivan,W.P., Marion,T.N., Zaitsu,K., Madden,B., McCormick,D.J. and Toft,D.O. (1993) Identification of a 60-kilodalton stress-related protein, p60, which interacts with Hsp90 and Hsp70. *Mol. Cell. Biol.,* **13,** 869-876.

Smith,D.F., Whitesell,L. and Katsanis,E. (1998) Molecular chaperones: biology and prospects for pharmacological intervention. *Pharmacol. Rev.,* 50, 493-513.

Smith,D.F., Whitesell,L., Nair,S.C., Chen,S., Prapapanich,V. and Rimerman,R.A. (1995) Progesterone receptor structure and function altered by geldanamycin, an Hsp90-binding agent. *Mol. Cell. Biol.,* **15,** 6804-6812.

Solovyev,V. and Salamov,A. (1997) The gene-finder computer tools for analysis of human and model organisms genome sequences. *Proceedings of the fifth international conference on Intelligent Systems for Molecular Biology* (AAAI Press), 294-302.

Spellman,T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.,* **9,** 3273-32.

Stancato,L.F., Hutchison,K.A., Krishna,P. and Pratt,W.B. (1996) Animal and pant cell lysates share a conserved chaperone system that assembles the glucocorticoid receptor into a functional heterocomplex with Hsp90. *Biochemistry,* **35,** 554-561.

Stothard,P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques,* **28**, 1102-1104.

Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F., *et al*. (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA,* **99,** 16899-16903.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genetics.,* **22**, 281-285.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.,* **22,** 4673-4680.

Tilstone,C. (2003) Vital statistics. *Nature,* **424,** 610-612.

Travers,K.J., Patil,C.K., Wodicka,L., Lockhart,D.J., Weissman,J.S. and Walter,P. (2000) Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell,* **101,** 249-258.

Trinklein,N.D., Murray,J.I., Hatmean,S.J., Botstein,D. and Myers,R.M (2004) The role of heat shock transcription factor 1 in the genome-wide regulation of the mammalian heat shock response. *Mol. Biol. Cell.,* **15,** 1254-1261.

van der Spuy,J., Cheetham,M.E., Dirr,H.W. and Blatch,G.L. (2001) The cochaperone murine stress-inducible protein 1: overexpression, purification, and characterization. *Protein Expr. Purif.,* **21,** 462-469.

van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.,* **281,** 827-842.

Waibel,A.H., Hanazawa,T., Hinton,G.E., Shikano,K. and Lang,K.J. (1989) Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustic, Speech, and Signal Processing*., **37**, 328-339.

Wang,Y., Jensen,R.C. and Stumph,W.E. (1996) Role of TATA box sequence and orientation in determining RNA polymerase II/III transcription specificity. *Nucleic Acids Res.,* **24,** 3100-3106.

Wei,Y.Q., Zhao,X., Kariya,Y., Teshigarawa,K. and Uchida,A. (1995) Inhibition of proliferation and induction of apoptotsis by abrogation of heat-shock protein (Hsp) 70 expression in tumor cells. *Cancer Immunol. Immunother*., **40,** 73-8.

Werner,T., Fessele,S., Maier,H. and Nelson,P.J. (2003) Computer modeling of promoter organisation as a tool to study transcriptional coregulation. *FASEB J.,* **17,** 1228-1237.

Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) Transfac: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.,* **24**, 238-241.

Wyrick,J.J., Holstege,F.C., Jennings,E.G., Causton,H.C., Shore,D., Grunstein.M., Lander,E.S. amd Young,R.A. (1999) Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature,* **402,** 418-421.

Xu,Y., Shah,M., Hyatt,D., Mural,R. and Uberbacher,E.C. (1999) GRAIL-EXP: Multiple gene modeling using pattern recognition and homology. *The Seventh Department of Energy Contractor and Grantee Workshop.*

Yoshimoto,H., Saltsman,K., Gasch,A.P., Li,H.X., Ogawa,N., Botstein,D., Brown,P.O. and Cyert,M.S. (2002) Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces* cerevisiae. *J. Biol. Chem.,* **277,** 31079-31088.

Young,J.C., Obermann,W.M.J., Hartl,F.U. (1998) Specific binding of tetratricopeptide repeat proteins to the C-terminal 12-kDa domain of Hsp90. *J. Biol. Chem.,* **273,** 18007-18010.

Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discrimination analysis. *Proc. Natl. Acad. Sci. USA,* **94,** 565-568.

Zhang,M.Q. (1998) Identification of human gene core promoters *in silico*. *Genome Res.,* **8,** 319-326.

Zhang,M.Q. (2000) Discriminant analysis and its application in DNA sequence motif recognition. *Briefings in Bioinformatics*, **1,** 331-342.

Zhang,Z., Quick,M.K., Kanelakis,K.C., Gijzen,M. and Krishna,P. (2003) Characterisation of a plant homolog of Hop, a cochaperone of Hsp90. *Plant Physiol.,* **131,** 525-535.

Zheng,L., Roeder,R.G. and Luo,Y. (2003) S phase activation of the histone H2B promoter by OCA-S, a coactivator complex, that contains GAPDH as a key component. *Cell,* **114,** 255-266.

Zou,J., Guo,Y., Guettouche,T., Smith,D.F. and Voellmy,R. (1998) Repression of heat shock transcription factor HSF1 activation by Hsp90 (Hsp90 complex) that forms a stress-sensitive complex with HSF1. *Cell,* **94,** 471-480.

APPENDIX A – Web-based programs used

Table A.1: Web-based software used in this study.

Program	Website	Reference
Alibaba	http://www.gene-regulation.com/pub/programs/alibaba2/index.html	Grabe (2002)
AlignACE	http://copan.cifn.unam.mx/Computational_Biology/yeast-tools http://atlas.med.harvard.edu/cgi-bin/alignace.pl	Roth *et al.* (1998), Hughes *et al.* (2000)
BioEdit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html	Hall (1999)
ClustalW	http://www.ebi.ac.uk/clustalw/	Thompson *et al.* (1994)
Cluster	http://www.rana.lbl.goc/EisenSoftWare.htm	Eisen *et al.* (1998)
CorePromoter	http://rulai.cshl.org/ tools/genefinder/CPROMOTER/index.htm	Zhang (1998), Ioshikhes and Zhang (2000), Zhang (2000)
DNAssist	http://www.dnassist.org/dnassist.htm	Patterton and Graves (2000a, 2000b)
EPD	http://www.epd.isb-sib.ch/	Périer *et al.* (1998); Périer *et al.* (1999); Périer *et al.* (2000); Praz *et al.* (2002), Schmid *et al.* (2004)
FastM	http://genomatix.gsf.de/cgi-bin/fastm2/fastm.pl	Klingenhoff *et al.* (1999)
Genscan	http://genes.mit.edu/GENSCAN.html	Burge and Kalin (1997)
HMMGene	http://www.cbs.dtu.dk/services/HMMgene/	Krogh (1997)
NCBI MapViewer	http://www.ncbi.nlm.nih.gov/mapview/	
NNPP	http://www.fruitfly.org/seq_tools/promoter.html	Waibel *et al.* (1989), Reese and Eeckman (1995), Reese (2000), Reese (2001)
Promoter 2.0	http://www.cbs.dtu.dk/service/promoter/	Knudsen (1999)
PromoterScan	http://cbs.umn.edu/ software/proscan/ promoterscan.htm	Prestridge (1995)
Reverse Complement	http://www.cbio.psu.edu/sms/rev_comp.html	Stothard (2000)
SGD	http://www.yeastgenome.org	
Shuffle	http://www.cbio.psu.edu/sms/shuffle_dna.html	Stothard (2000)
TESS	http://www.cbil.upenn.edu/tess	Schug and Overton (1997a,b)
TFSearch	http://www.cbrc.jp/research/db/TFSEARCH.html	Heinemeyer *et al.* (1998)
Treeview	http://rana.Stanford.EDU/software/	Eisen *et al.* (1998)
TSSG	http://www.softberry.com	Solovyev and Salamov (1997)

APPENDIX B – Human *STI1* cDNA and protein sequence

Protein sequence alignment of Sti1 orthologues in human, mouse and yeast showed a strong similarity between species. Pairwise global alignment, using the PAM250 similarity matrix in BioEdit (Hall, 1999), was employed. The human and mouse proteins were 97% identical (alignment score 2601), the human and yeast proteins were 39% identical (alignment score 998), and the mouse and yeast proteins were 39% identical (alignment score 1000).

```
1       GGTTGGGAACGCGGAGCGGACGGATTCGATTCAACGGGGTTCCGGACCGCGCTGCGCTATGGAGCAGGTCAATGAGCTGAAGGAGAAAGGCAACAAGGCC
1                                                                   M  E  Q  V  N  E  L  K  E  K  G  N  K  A

101     CTGAGCGTGGGTAACATCGATGATGCCTTACAGTGCTACTCCGAAGCTATTAAGCTGGATCCCCACAACCACGTGCTGTACAGCAACCGTTCTGCTGCCT
15       L  S  V  G  N  I  D  D  A  L  Q  C  Y  S  E  A  I  K  L  D  P  H  N  H  V  L  Y  S  N  R  S  A  A

201     ATGCCAAGAAAGGAGACTACCAGAAGGCTTATGAGGATGGCTGCAAGACTGTCGACCTAAAGCCTGACTGGGGCAAGGGCTATTCACGAAAAGCAGCAGC
48      Y  A  K  K  G  D  Y  Q  K  A  Y  E  D  G  C  K  T  V  D  L  K  P  D  W  G  K  G  Y  S  R  K  A  A  A

301     TCTAGAGTTCTTAAACCGCTTTGAAGAAGCCAAGCGAACCTATGAGGAGGGCTTAAAACACGAGGCAAATAACCCTCAACTGAAAGAGGGTTTACAGAAT
82       L  E  F  L  N  R  F  E  E  A  K  R  T  Y  E  E  G  L  K  H  E  A  N  N  P  Q  L  K  E  G  L  Q  N

401     ATGGAGGCCAGGTTGGCAGAGAGAAAATTCATGAACCCTTTCAACATGCCTAATCTGTATCAGAAGTTGGAGAGTGATCCCAGGACAAGGACACTACTCA
115      M  E  A  R  L  A  E  R  K  F  M  N  P  F  N  M  P  N  L  Y  Q  K  L  E  S  D  P  R  T  R  T  L  L

501     GTGATCCTACCTACCGGGAGCTGATAGAGCAGCTACGAAACAAGCCTTCTGACCTGGGCACGAAACTACAAGATCCCCGGATCATGACCACTCTCAGCGT
148     S  D  P  T  Y  R  E  L  I  E  Q  L  R  N  K  P  S  D  L  G  T  K  L  Q  D  P  R  I  M  T  T  L  S  V

601     CCTCCTTGGGGTCGATCTGGGCAGTATGGATGAGGAGGAAGAGATTGCAACACCTCCACCACCACCCCCTCCCAAAAAGGAGACCAAGCCAGAGCCAATG
182      L  L  G  V  D  L  G  S  M  D  E  E  E  E  I  A  T  P  P  P  P  P  P  P  K  K  E  T  K  P  E  P  M

701     GAAGAAGATCTTCCAGAGAATAAGAAGCAGGCACTGAAAGAAAAAGAGCTGGGGAACGATGCCTACAAGAAGAAAGACTTTGACACAGCCTTGAAGCATT
215      E  E  D  L  P  E  N  K  K  Q  A  L  K  E  K  E  L  G  N  D  A  Y  K  K  K  D  F  D  T  A  L  K  H

801     ACGACAAAGCCAAGGAGCTGGACCCCACTAACATGACTTACATTACCAATCAAGCAGCGGTATACTTTGAAAAGGGCGACTACAATAAGTGCCGGGAGCT
248     Y  D  K  A  K  E  L  D  P  T  N  M  T  Y  I  T  N  Q  A  A  V  Y  F  E  K  G  D  Y  N  K  C  R  E  L

901     TTGTGAGAAGGCCATTGAAGTGGGGAGAGAAAACCGAGAAGACTATCGACAGATTGCCAAAGCATATGCTCGAATTGGCAACTCCTACTTCAAAGAAGAA
282      C  E  K  A  I  E  V  G  R  E  N  R  E  D  Y  R  Q  I  A  K  A  Y  A  R  I  G  N  S  Y  F  K  E  E

1001    AAGTACAAGGATGCCATCCATTTCTATAACAAGTCTCTGGCAGAGCACCGAACCCCAGATGTGCTCAAGAAATGCCAGCAGGCAGAGAAAATCCTGAAGG
315      K  Y  K  D  A  I  H  F  Y  N  K  S  L  A  E  H  R  T  P  D  V  L  K  K  C  Q  Q  A  E  K  I  L  K

1101    AGCAAGAGCGGCTGGCCTACATAAACCCCGACCTGGCTTTGGAGGAGAAGAACAAAGGCAACGAGTGTTTTCAGAAAGGGGACTATCCCCAGGCCATGAA
348     E  Q  E  R  L  A  Y  I  N  P  D  L  A  L  E  E  K  N  K  G  N  E  C  F  Q  K  G  D  Y  P  Q  A  M  K

1201    GCATTATACAGAAGCCATCAAAAGGAACCCGAAAGATGCCAAATTATACAGCAATCGAGCTGCCTGCTACACCAAACTCCTGGAGTTCCAGCTGGCACTC
382      H  Y  T  E  A  I  K  R  N  P  K  D  A  K  L  Y  S  N  R  A  A  C  Y  T  K  L  L  E  F  Q  L  A  L

1301    AAGGACTGTGAGGAATGTATCCAGCTGGAGCCGACCTTCATCAAGGGTTATACACGGAAAGCCGCTGCGCTGGAAGCGATGAAGGACTACACCAAAGCCA
415      K  D  C  E  E  C  I  Q  L  E  P  T  F  I  K  G  Y  T  R  K  A  A  A  L  E  A  M  K  D  Y  T  K  A

1401    TGGATGTGTACCAGAAGGCGCTAGACCTGGACTCCAGCTGTAAGGAGGCGGCAGACGGCTACCAGCGCTGTATGATGGCGCAGTACAACCGGCACGACAG
448     M  D  V  Y  Q  K  A  L  D  L  D  S  S  C  K  E  A  A  D  G  Y  Q  R  C  M  M  A  Q  Y  N  R  H  D  S

1501    CCCCGAAGATGTGAAGCGACGAGCCATGGCCGACCCTGAGGTGCAGCAGATCATGAGTGACCCAGCCATGCGCCTTATCCTGGAACAGATGCAGAAGGAC
482      P  E  D  V  K  R  R  A  M  A  D  P  E  V  Q  Q  I  M  S  D  P  A  M  R  L  I  L  E  Q  M  Q  K  D

1601    CCCCAGGCACTCAGCGAACACTTAAAGAATCCTGTAATAGCACAGAAGATCCAGAAGCTGATGGATGTGGGTCTGATTGCAATTCGGTGATGACTTGTTC
515      P  Q  A  L  S  E  H  L  K  N  P  V  I  A  Q  K  I  Q  K  L  M  D  V  G  L  I  A  I  R  *  *

1701    ATCCCCCCTTCCCTTCGCCCTCATGTGGAAAGAGGAGCTGGGACCGCGGCGAGCAGCACGGAGCGGAAGGGAGAGCAGGGGAGAGAAGGCCTCATCTCTC
1801    TATATTTATACATAACCCCGGGGAAGACACAGAGACTCGTACCTGCGCTGTTTGTGCCGCCGCTGCCTCTGGGCCCTCCCAGCACACGCATGGTCTCTTC
1901    ACCGCTGCCCTCGAGTTCCATGTCTCTTTCCCCTGCCCCTAGTTGCTGTCTCGGCTGCTCTCCCATAGTTGGTTTTTTTTTTATTTGGGGCAGTGGGCAT
2001    GTTATGGGGAGGGGAGGGGGTTCTTCCAGCCTCAGGTCCCAGCTGTCTCACGTTGTTTATTCTGCGTCCCCTTCTCCAATAAAACAAGCCAGTTGGGCGT
2101    GGTTATAAC
```

Figure B.1: Human *STI1* cDNA and protein sequence, including untranslated nucleotides up- and downstream of the coding region. The TPR1, TPR2A, and TPR2B domains are shaded cyan, yellow and green respectively. Within each domain, TPR repeats are indicated by a break in shading. Boxed sequences indicate potential casein kinase II phosphorylation sites. Thin underline indicates highly conserved amino acids. Double underline indicates potential NLS sites. Grey shaded areas indicate the polyproline stretch, DPEV, and DPAM sequences respectively. Adapted from Odunuga *et al.,* 2004.

APPENDIX C – Genes surrounding *hSTI1* and *mSTI1*

Table C.1: Genes surrounding *hSTI1* on chromosome 11.

Gene name	Start position	Stop position	Length of gene	Strand
OTUB1	63510461	63522463	12003	+
LRP16	63541390	63708893	167504	-
FLRT1	63646601	63662005	15405	+
STIP1	63729036	63747375	18340	+
URP2	63749566	63766723	17158	+
MGC11134	63766631	63768982	2352	-
MGC13045	63769122	63772848	3727	+
DNAJC4	63777368	63777016	3330	+
VEGFB	63777626	63781619	3994	+
FKBP2	63783773	63787046	3274	+
PPP1R14B	63787312	63792098	4787	-
PLCB3	63794419	63810388	15970	+
BAD	63812662	63827524	14863	-

i. BAD - BCL2-antagonist of cell death
ii. DNAJC4 - DnaJ (Hsp40) homologue, Family C, Member 4
iii. FKBP2 - FK506 binding protein 2
iv. FLRT1 - Fibronectin leucine rich transmembrane protein 1
v. LRP16 - Low density lipoprotein-related protein 1
vi. MGC11134 - tRNA splicing 2' phosphotransferase 1
vii. MGC13045 - Hypothetical protein
viii. OTUB1 - OTU domain, ubiquitin aldehyde binding 1
ix. PLCB3 - Phospholipase C beta 3
x. PPP1R14B - Protein phosphatase 1 regulatory (inhibitor) subunit 14B
xi. *STI1* - Stress-inducible phosphoprotein 1
xii. URP2 - UNC-112 related protein 2
xiii. VEGFB - Vascular endothelial growth factor B

Table C.2: Genes surrounding *mSTI1* on chromosome 19.

Gene name	Start position	Stop position	Length of gene	Strand
OTUB1	6911633	6919711	8079	-
FLRT1	6808411	6819108	10698	-
LRP16	6770237	6911484	141248	+
STIP1	6734137	6753370	19234	-
URP2	6712408	6732865	20458	-
MGC11134	6709763	6712476	2714	+
MGC13045	6706449	6709467	3019	-
DNAJC4	6701341	6705702	4362	-
VEGFB	6695902	6701033	5132	-
FKBP2	6691171	6693860	2690	-
PPP1R14B	6688416	6690752	2337	+
PLCB3	6667144	6683182	16039	-
BAD	6655285	6665323	10039	+

i. BAD - BCL2-antagonist of cell death
ii. DNAJC4 - DnaJ (Hsp40) homologue, Family C, Member 4
iii. FKBP2 - FK506 binding protein 2
iv. FLRT1 - Fibronectin leucine rich transmembrane protein 1
v. LRP16 - Low density lipoprotein-related protein 1
vi. MGC11134 - tRNA splicing 2' phosphotransferase 1
vii. MGC13045 - Hypothetical protein
viii. OTUB1 - OTU domain, ubiquitin aldehyde binding 1
ix. PLCB3 - Phospholipase C beta 3
x. PPP1R14B - Protein phosphatase 1 regulatory (inhibitor) subunit 14B
xi. *STI1* - Stress-inducible phosphoprotein 1
xii. URP2 - UNC-112 related protein 2
xiii. VEGFB - Vascular endothelial growth factor B

APPENDIX D – Genscan and HMMGene results

Table D.1: Settings available for the gene prediction programs used.

Software	Settings used		
	Human	**Mouse**	**Yeast**
Genscan	Vertebrate	Vertebrate	Vertebrate, (*Arabidopsis*, maize)
HMMGene	Human (and other vertebrates)	Human (and other vertebrates)	Vertebrate (*C. elegans*)

Table D.2: HMMGene and Genscan's predicted exon / intron structure for *hSTI1*(i).

	Exon number	DNA Strand	Begin	End	Length	Score
HMMGene	0	+	**63728657**	**63728790**	134	0.623
	1	+	**63729015**	**63729102**	88	0.389
	2	+	63735910	63736119	210	0.996
	3	+	63737021	63737162	142	0.989
	4	+	63737311	63737452	142	0.99
	5	+	**63738513**	**63738645**	133	0.672
	6	+	63740103	63740229	127	1
	7	+	63740325	63740427	103	1
	8	+	63740686	63740806	121	0.765
	9	+	63742772	63742868	97	0.982
	10	+	63743001	63743125	125	0.97
	11	+	63745708	63745744	37	0.916
	12	+	63745957	63746060	104	0.831
	13	+	63746282	63746454	173	0.956
	14	+	63746886	63746958	73	0.792
Genscan	0	+	**63728125**	**63728331**	207	0.637
	1	+	**63728646**	**63728806**	161	0.812
	2	+	63735910	63736119	210	0.999
	3	+	63737021	63737162	142	0.998
	4	+	63737311	63737452	142	0.999
	5	+	**63738477**	**63738645**	169	0.997
	6	+	63740103	63740229	127	0.999
	7	+	63740325	63740427	103	0.999
	8	+	63740686	63740806	121	0.992
	9	+	63742772	63742868	97	0.92
	10	+	63743001	63743125	125	0.774
	11	+	63745708	63745744	37	0.767
	12	+	63745957	63746060	104	0.633
	13	+	63746282	63746454	173	0.733
	14	+	63746886	63746958	73	0.93
	PolyA	+	63747346	63747351	6	

(i) Shaded regions indicate sites that deviate between gene prediction programs.

Table D.3: HMMGene and Genscan's predicted exon / intron structure for *mSTI1*(i).

	Exon number	DNA strand	Begin	End	Length	Score
HMMGene	1	-	**6749364**	**6749356**	9	0.298
	2	-	6749134	6748925	210	0.985
	3	-	6748217	6748076	142	0.991
	4	-	6747734	6747593	142	0.986
	5	-	**6744301**	**674420**	94	0.564
	6	-	6742664	6742538	127	1.000
	7	-	6742434	6742332	103	1.000
	8	-	6742178	6742058	121	0.998
	9	-	6739818	6739722	97	0.783
	10	-	6738869	6738745	125	0.923
	11	-	6735738	6735702	37	0.742
	12	-	6735560	6735457	104	0.999
	13	-	6735252	6735080	173	0.984
	14	-	6734601	6734529	73	0.547
Genscan	Promoter	-	6753488	6753449	40	
	0	-	**6753176**	**6752806**	371	0.514
	1	-	**6752271**	**6752208**	64	0.457
	2	-	6749134	6748925	210	0.999
	3	-	6748217	6748076	142	0.999
	4	-	6747734	6747593	142	0.995
	5	-	**6746079**	**6745911**	169	0.875
	6	-	6742664	6742538	127	0.999
	7	-	6742434	6742332	103	0.999
	8	-	6742178	6742058	121	0.999
	9	-	6739818	6739722	97	0.993
	10	-	6738869	6738745	125	0.973
	11	-	6735738	6735702	37	0.800
	12	-	6735560	6735457	104	0.998
	13	-	6735252	6735080	173	0.989
	14	-	6734601	6734529	73	0.922
	PolyA	-	1029	1024	6	

(i) Shaded regions indicate sites that deviate between gene prediction programs.

Table D.4: Gene prediction results for *ySTI1*(i).

Exon number	Begin	End	Exon length (bp)	No. amino acids	Score
Gene 1 (H)	380245	380805			0.784 (H)
Promoter (G)	380923	380962			
5' UTR	380988	381051	64		
Gene 2 (G)(H)	381052	382821	1770	589	1.001 (H) 0.000 (G)
3' UTR	382822	383195	374		
PolyA (G)	383190	383195			

(i) Genscan results (G) and HMMGene results (H) are indicated.

APPENDIX E - TATA boxes

Table E.1: Possible TATA boxes within 10 kb upstream of the predicted TSS for *hSTI1*.

Location With respect to transcription start site	Sequence	Comply with consensus TATA[AT]A[AT]	Comply with consensus TATA[AT][AGT][GA]
-219	GTTTATAGAG		
-494	TATATCA		
-2880	TATAAAC		
-4138	TATATTG		✓
-5056	ATTTATA	✓	
-5489	TATATAT	✓	
-6236	GTTTATATTA		✓
-6523	TTTTATA	✓	✓
-7981	TATAATT		
-9042	TATAATC		

Where a TATA box has been identified, a tick has been placed in the appropriate column. Putative TATA boxes at positions -219, -5056, -6236, and -6523, occur on the reverse complement strand.

APPENDIX F – Promoter prediction results

Table F.1: Putative TFBS predicted for *hSTI1* by three TFBS prediction programs: Alibaba, TESS and TFSearch.

TF	Position	Alibaba	TESS	TFSearch	Hsp70	Mouse	Yeast
Ftz	-964	✓	✓				
Nkx-2	-962			✓		✓(-969)	
C-Ets	-957		✓	✓		✓ (-947)	
Sp1	-900	✓				✓ (-901)	✓ (-916)
Sp1	-844	✓	✓				
MZF1	-826			✓	✓ (-812)		
Sp1	-817	✓	✓		✓ (-810)		
USF	-787	✓	✓	✓			
Elk-1	-780	✓		✓			
HSF	-772			✓			✓ (-774)
Sp1	-741	✓	✓			✓ (-738)	✓ (-754)
NF-1	-739	✓				✓ (-751)	✓ (-750)
REB1	-707		✓			✓(-703)	
Sp1	-689	✓				✓ (-699)	
Sp1	-649	✓	✓				
WT1	-649	✓	✓				
AP	-638	✓			✓TESS		
Sp1	-638	✓	✓				✓ (-632)
Sp1	-603	✓				✓	
SRF	-504	✓				✓ (-487)	
GATA-1	-499	✓		✓			
Sp1	-491	✓	✓	✓	✓ (-494)	✓ (-495)	✓ (-498)
Sp1	-481	✓				✓	
CdxA	-455			✓	✓ (-458)	✓ (-469)	
CBF	-451		✓			✓ (-467)	
SRY	-451			✓		✓ (-466)	
GATA-1	-449			✓		✓ (-464)	
Pbx-1	-449			✓	✓	✓ (-464)	
Sp1	-423	✓	✓		✓ (-432, -413)	✓ (-422)	✓ (-435)
Sp1	-391	✓				✓ (-392)	
Sp1	-372	✓				✓ (-370)	
CBF	-366		✓			✓ (-367)	
EFI	-364	✓				✓(-365)	
C/EBPalpha	-351	✓				✓ (-359beta)	
c-Rel	-349			✓		✓	
HSF	-349	✓	✓	✓		✓	✓ (-342)
NF-kappaB	-349	✓	✓	✓		✓	
HSF	-339			✓			✓ (-342)
C/EBPalpha	-329	✓			✓ (-337, -332)		
CBF	-329		✓			✓ (-330)	
Sp1	-323	✓	✓			✓ (-324)	
NF-1	-318	✓				✓ (-321)	
CP1	-305		✓			✓ (-293)	
C/EBPalpha	-290	✓	✓				
CBF	-285		✓			✓ (-286)	
C/EBPalpha	-284	✓					✓ (-276)
Sp1	-261	✓	✓			✓ (-267)	
SRF	-251	✓	✓				

Table F.1 continued

GATA-1	-236			✓			✓ (-251)
Sp1	-232	✓	✓			✓ (-234)	
Sp1	-221	✓			✓ (-215)	✓ (-219)	
GATA-1	-209			✓	✓ (-200)	✓ (-207)	
Egr-1	-199	✓					
Sp1	-199	✓	✓			✓ (-195)	
Sp1	-189			✓	✓ (-193)	✓ (-187)	✓ (-176)
CBF	-160		✓			✓ (-159)	
Oct-1	-156	✓			✓ (-147)		
Sp1	-134	✓	✓				
NF-1	-122	✓				✓ (-119)	
C/EBPalpha	-120	✓					✓ (-131)
CBF	-77		✓			✓ (-79)	
Sp1	-60	✓			✓ (-67)	✓ (-63)	
AP	-51	✓				✓ (-47)	
Sp1	-50	✓			✓ (-44)		
HSF	-31		✓			✓	
Sp1	-19	✓				✓	
Sp1	-13	✓				✓	

Where a particular TFBS prediction program has predicted a TFBS of the type indicated in the TF column, within 10 bp of the position indicated, it has been indicated by a tick (✓) in the relevant column. Where a match of TF, of both type and position (within 20 bp), was found for either human *HSP70* or for one of the *STI1* orthologues, it has been indicated by a tick (✓) in the appropriate column. Shaded rows indicate TFs shown in Figure 12.

i. AP-1 - Activator protein 1
ii. C/EBP - CCAAT / enhancer-binding proteins
iii. CBF - CCAAT-binding factor
iv. CdxA - Caudal-related homeodomain transcription factor
v. c-Ets - Cellular E26 transformation specific sequence
vi. CP2 - CCAAT-binding protein
vii. c-Rel - Cellular reticuloendotheliosis proto-oncogene
viii. EFI - Enhancer factor I (CCAAT-binding)
ix. Egr – Early growth response
x. Elk - ETS-like transcription factor-1
xi. Ftz - Fushi tarazu transcription factor
xii. GATA-1 - GATA binding protein 1 (globin transcription factor 1)
xiii. HSF – Heat shock transcription factor
xiv. MZF1 – Myeloid zinc finger protein 1
xv. NF-kappaB – Nuclear factor kappa B
xvi. Nkx-2.5 - Homeobox protein NK-2 homolog E
xvii. Oct-1 - Octamer-binding transcription factor 1
xviii. Pbx-1 - Pre-B-cell leukemia transcription factor 1
xix. REB1 - rDNA enhancer-binding protein 1
xx. Sp1 - Stimulatory protein 1 (also known as Specific protein 1)
xxi. SRF - Serum response factor
xxii. SRY - Sex-determining
xxiii. USF - Upstream stimulatory factor

Table F.2: Putative TFBS predicted for *mSTI1* by three TFBS prediction programs: Alibaba, TESS and TFSearch.

TF	**Position**	**Sequence**	**Alibaba**	**TESS**	**TFSearch**	**Hsp70**	**Human**	**Yeast**
CdxA	-970				✓	✓ (-957)		
Nkx-2	-969				✓		✓ (-962)	
GATA-1	-963		✓	✓	✓			
c-Ets	-947				✓		✓ (-957)	
GATA-1	-946				✓			✓ (-954)
NF-1	-920		✓			✓ (-918		

Table F.2 continued

C/EBPalpha	-918		✓					✓ (-923)
Sp1	-901		✓			✓ (-894)	✓ (-900)	✓ (-916)
Sp1	-871		✓			✓ (-881)		
C/EBPalpha	-772		✓			✓ (-783)		
NF-1	-751		✓				✓ (-739)	✓ (-750)
CBF	-748			✓				✓ (-743 CBF2)
Sp1	-738		✓			✓ (-745)	✓ (-741)	✓ (-754)
REB1	-703			✓			✓ (-707)	
Sp1	-699		✓				✓ (-689)	
HSF	-678		✓		✓			✓ (-661)
C/EBPalpha	-661		✓			✓ (-670)		✓ (-670)
C/EBPalpha	-649		✓			✓ (-632)		
Sp1	-603		✓				✓	
GR	-575		✓	✓		✓		
RAP1	-500		✓	✓				
Sp1	-495		✓				✓ (-491)	✓ (-498)
SRF	-487			✓			✓ (-504)	
Sp1	-481		✓				✓	
CP1/2	-470		✓		✓			
CdxA	-469				✓		✓ (-455)	
CBF	-467			✓			✓ (-451)	
SRY	-466				✓		✓ (-451)	
GATA-1	-464				✓		✓ (-449)	
Pbx-1	-464				✓		✓ (-449)	
SRF	-455		✓	✓				
Hb	-440		✓	✓				✓ (-424)
C/EBPalpha	-436		✓					✓ (-425)
NF-kappaB	-431		✓			✓ (-439)		
RAP1	-426		✓	✓				
Sp1	-422		✓				✓ (-423)	✓ (-435)
Sp1	-392		✓				✓ (-392)	
Sp1	-370		✓			✓	✓ (-372)	
CBF	-367			✓			✓ (-366)	
EFI	-365		✓				✓ (-364)	
C/EBPbeta	-359		✓			✓ (-349)	✓ (-351alpha)	
c-Rel	-349				✓		✓	
HSF	-349		✓	✓	✓		✓	✓ (-342)
NF-kappaB	-349		✓	✓	✓		✓	

Table F.2 continued

CdxA	-340				✓	✓ (-325)		
SRY	-337				✓	✓ (-351)		
CBF	-330			✓			✓ (-329)	
Sp1	-324		✓	✓			✓ (-323)	
NF-1	-321		✓				✓ (-318)	
CP1/2	-293				✓		✓ (-305)	
CBF	-286			✓			✓ (-285)	
CRE-BP	-274		✓		✓			
Sp1	-267		✓	✓		✓ (-259)	✓ (-261)	
Sp1	-234		✓	✓			✓ (-232)	
NF-kappaB	-225			✓	✓			
Sp1	-219		✓				✓ (-221)	
C/EBPbeta	-213			✓		✓ (-219 alpha)		
GATA-1	-207				✓		✓ (-209)	
c-Ets	-202				✓		Egr01 (-199)	
Sp1	-195		✓				✓ (-199)	
Sp1	-187			✓	✓		✓ (-189)	✓ (-176)
Hb	-167		✓	✓				
C/EBPalpha	-159		✓			✓ (-154)		✓ (-152)
CBF	-159			✓			✓ (-160)	
NF-1	-119		✓			✓ (-108)	✓ (-122)	
SRF	-116		✓	✓				
TBP	-115		✓					
CBF	-79			✓			✓ (-77)	
Sp1	-63		✓				✓ (-60)	
AP	-47		✓				✓ (-51)	
HSF	-31			✓			✓	
Sp1	-19		✓			✓ (-20)	✓	
Sp1	-13		✓				✓	

Where a particular TFBS prediction program has predicted a TFBS of the type indicated in the TF column, within 10 bp of the position indicated, it has been indicated by a tick (✓) in the relevant column. Where a match of TF, of both type and position (within 20 bp), was found for either mouse *HSP70* or for one of the *STI1* orthologues, it has been indicated by a tick (✓) in the appropriate column. Shaded rows indicate TFs shown in Figure 12.

i. C/EBP - CCAAT / enhancer-binding proteins
ii. CBF - CCAAT-binding factor
iii. CdxA - Caudal-related homeodomain transcription factor
iv. c-Ets - Cellular E26 transformation specific sequence
v. CP2 - CCAAT-binding protein
vi. CRE-BP - cAMP response element binding protein
vii. c-Rel - Cellular reticuloendotheliosis proto-oncogene
viii. EFI - Enhancer factor I (CCAAT-binding)
ix. GATA-1 - GATA binding protein 1 (globin transcription factor 1)
x. GR – glucocorticoid receptor
xi. Hb - Homeobox transcription factor
xii. HSF – Heat shock transcription factor
xiii. NF-kappaB – Nuclear factor kappa B
xiv. Nkx-2.5 - Homeobox protein NK-2 homolog E
xv. Pbx-1 - Pre-B-cell leukemia transcription factor 1
xvi. RAP1 - Repressor activator protein 1
xvii. REB1 - rDNA enhancer-binding protein 1
xviii. Sp1 - Stimulatory protein 1 (also known as Specific protein 1)
xix. SRF - Serum response factor

xx. SRY - Sex-determining
xxi. TBP – TATA-binding protein

Table F.3: Putative TFBS predicted for *ySTI1* by three TFBS prediction programs: Alibaba, TESS and TFSearch.

TF	Position	Sequence	Alibaba	TESS	TFSearch	Hsp70	Human	Mouse
GATA-1	-954		✓					✓ (-963)
Oct-1	-954		✓	✓				
YY1	-952		✓	✓				
C/EBPalpha	-923		✓					✓ (-918)
Sp1	-916		✓				✓ (-900)	✓ (-901)
HSF	-849				✓	✓ (-868, -864)		
HSF	-825				✓	✓ (-831)		
HSF	-774				✓		✓ (-772)	
HSF	-760				✓	✓		
Sp1	-754		✓			✓ (-759)	✓ (741)	✓ (-755)
NF-1	-750		✓				✓ (-739)	✓ (-751)
CBF2	-743			✓				✓ (-748)
GATA-1	-730		✓			✓ (-744)		
HSF	-710				✓	✓ (-699)		
C/EBPalpha	-670			✓				✓ (-661)
HSF	-661				✓	✓ (-655, -666, -670)		✓ (-678)
Sp1	-632		✓			✓ (-613)	✓ (-638)	
C/EBPalpha	-613		✓			✓ (-618)		
C/EBPalpha	-591		✓			✓ (-596)		
Oct-1	-571		✓	✓				
Oct-1	-546		✓	✓		✓ (-543)		
HSF	-510		✓		✓	✓ (500, -513)		
C/EBPalpha	-501			✓		✓ (-499)		
HSF	-498				✓	✓ (-490)		
Sp1	-498		✓				✓ (-491)	✓ (-495)
HSF	-482				✓	✓ (-470)		
HSF	-441				✓	✓ (-449, -455)		
Sp1	-435		✓			✓	✓ (-423)	✓ (-430)
C/EBPalpha	-425		✓					✓ (-436)
Hb	-424			✓				✓ (-440)
HSF	-369				✓	✓ (-360, -362, -387, -365)		

Table F.3 continued

HSF	-358				✓	✓ (-364, -355, -360, -354)		
HSF	-342				✓	✓ (-346, -349, -335)	✓ (-349)	✓ (-349)
HSF	-293				✓	✓ (-306)		
HSF	-282				✓	✓ (-271, -287)		
C/EBPalpha	-276		✓			✓ (-278, -292)	✓ (-284)	
HSF	-255				✓	✓ (-257, -265, -259, -268, -252, -262)		
GATA-1	-251		✓				✓ (-236)	
HSF	-218		✓	✓	✓	✓ (-214, -229)		
Sp1	-176		✓			✓ (-161)	✓ (-189)	✓ (-187)
TBP	-160			✓		✓ (-168)		
C/EBPalpha	-152		✓			✓ (-168beta, -156)		✓ (-159)
HSF	-152				✓	✓ (-137, -142)		
C/EBPalpha	-131		✓	✓			✓ (-120)	
Oct-1	-131		✓			✓ (-280)		
TBP	-105		✓	✓				
Oct-1	-89		✓			✓ (-90)		
C/EBPbeta	-72			✓		✓ (-59)		

Where a particular TFBS prediction program has predicted a TFBS of the type indicated in the TF column, within 10 bp of the position indicated, it has been indicated by a tick (✓) in the relevant column. Where a match of TF, of both type and position (within 20 bp), was found for either yeast *HSP70* or for one of the *STI1* orthologues, it has been indicated by a tick (✓) in the appropriate column. Shaded rows indicate TFs shown in Figure 12.

i. C/EBP - CCAAT / enhancer-binding proteins
ii. CBF - CCAAT-binding factor
iii. GATA-1 - GATA binding protein 1 (globin transcription factor 1)
iv. Hb - Homeobox transcription factor
v. HSF – Heat shock transcription factor
vi. NF-1 – Nuclear factor 1
vii. Oct-1 - Octamer-binding transcription factor 1
viii. Sp1 - Stimulatory protein 1 (also known as Specific protein 1)
ix. TBP – TATA-binding protein
x. YY1 – Yin-yang 1

Table F.4: Settings available for the promoter and TFBS prediction programs used.

Software	**Settings used**		
	Human	**Mouse**	**Yeast**
Alibaba	No options		
CorePromoter	Human	Human	Human
NNPP	Eukaryote	Eukaryote	Eukaryote
PromoterScan	No options		
TESS	No options		
TFSearch	Vertebrate	Vertebrate	Yeast
TSSG	Human	Human	Human

APPENDIX G – Genes clustering with *ySTI1*

Table G.1: Genes clustering with *ySTI1* using Cluster (Eisen *et al.,* 1998). Cluster scores are indicated.

	Calcineurin	Cell Cycle	DNA-Damaging Agents	Environmental Stress	Histone Depletion	Ploidy	Sporulation	Unfolded Protein	Zinc
Cluster	**0.869**	**0.853**	**0.861**	**0.856**	**0.860**	**0.883**	**0.810**	**0.861**	**0.863**
1	YAL001C	YER103W	YBR101C	YDR214W	YBR101C	YDR159W	YHR007C	CAT5	YBR101C
2	YAL005C	YNL007C	YGR142W	YLR216C	YGR142W	YOR027W	YOR027W	COX14	YCR021C
3	YGL122C	YOR027W	YMR186W	YLR217W	YMR186W	YPL077C		*STI1*	YCR103C
4	YGL222C	YPL240C	YNL006W	YMR186W	YNL006W			UBC4	YDR151C
5	YHR115C		YNL007C	YNL281W	YNL007C				YER067W
6	YLL024C		YOR027W	YOR010C	YOR027W				YER100W
7	YLR438W		YPL240C	YOR027W	YPL240C				YGL045W
8	YMR147C		YPR158W	YPL240C	YPR158W				YGL046W
9	YOL066C			YLR259C					YGR142W
10	YOR027W								YGR249W
11	YOR298W								YJL144W
12	YPL258C								YKL223W
13									YKR075C
14									YLL024C
15									YLR327C
16									YLR446W
17									YMR020W
18									YMR022W
19									YMR104C
20									YMR316W
21									YNL006W
22									YNL007C
23									YOL016C
24									YOR027W
25									YPL014W
26									YPR013C
27									YPR066W
8									YPR154W
29									YPR158W

APPENDIX H – Motifs returned by AlignACE

Table H.1: Putative regulatory motifs, returned as output by AlignACE.

	Motif Number	Motif	MAP Score	Number of occurrences
Calcineurin (12 genes)	1	ACNCNNNNNATNAAAAA	41.9334	32
	2	AAAAGTGAAA	41.3794	35
	3	ATAAAATTTC	17.104	20
	4	ANGNTNNTNGNGGAAA	11.8293	22
	5	ATCTNGNAAGANA	9.83017	15
	6	GNNGNNAANNNNANATAAA	6.58225	27
	7	CGAGNNNTTGACG	4.64654	14
	8	TANANGAGGNGNAG	2.64851	14
	9	AGNNANNNAGAAAAA	2.0425	15
	10	GNAAGGGNTGAA	1.8788	10
	11	ANGTTCTNNAAAG	0.819452	10
Cell cycle (4 genes)	1	CNNGAAAANNNANANNNANA	25.1154	21
	2	AATAANAAGNAA	14.3441	20
	3	ANTNTNNNAGAACTT	9.90791	11
	4	CNAGAANNNNNNNNAANANAA	8.70882	17
	5	TNACNTTCCNGAG	7.97982	9
	6	AGAAAAATTC	4.36303	13
	7	ANNCNTGATAGAA	4.21397	15
	8	GAAGGGATNNGC	3.02069	9
	9	ANGNANNANNNNNANNNGNGNGNTA	2.00468	12
	10	CNNNNANANANAANNNNAANCNNNC	1.49549	18
DNA-damaging agents (8 genes)	1	AGACNNNNNNANAANATA	36.4665	26
	2	GGAAGAAAAA	31.9315	20
	3	TNNNNNNNNNTTCNAGAANANNA	27.3498	18
	4	TANANAANAGNNNNAANA	18.5892	25
	5	AAGAAAGNGCNT	18.0328	25
	6	GGNNNNAAAANNCGNNNNNGG	17.9432	12
	7	TANAANCANNAGNNCA	12.4467	32
	8	AAAAGNANAANNNAA	8.1873	21
	9	ATANNACCATNCG	8.16474	16
	10	ANNANNNTNNANAANANNNNTTA	6.13776	9
	11	TNAGNNGGTNANAGA	5.78313	12
	12	TNGAGNTNNAGNANNNAT	0.451682	20
Environmental stress (9 genes)	1	AAAAAAAGANA	46.3991	34
	2	TTCNAGAAANCNNA	20.7409	14
	3	TTNNAGGNNGNGNNNGAA	9.76909	15
	4	ACNNANGNNCAAANAA	4.01432	39
	5	AGNNTNNGNAGGNAANG	3.14356	13
	6	AAANNNGNNGCCNNGGC	2.52841	11
	7	GAAGNNATNAGNNGC	0.663105	20
	8	CAAANACGCAA	0.506064	12
	9	AAACAAAAGC	0.367363	17
	10	GCNNNTCNTTGCCNNNNG	0.147904	12
Histone (8 genes)	1	AGAANNNNNNANTANATA	34.6948	24
	2	GGAAGAAAAA	31.9315	20
	3	GNNNNAANAANNCGNNNATG	27.5759	10
	4	GCNNTCTAGAAA	25.2027	14
	5	TTCNNGAANNNNAANANA	21.8161	16

Table H.1 continued

	6	AAAAAGNANAANT	17.6707	21
	7	ANNCNNNAGAGGANNNANNA	16.3669	13
	8	GNNTGTNNCANNNTNTNNNANC	14.6948	11
	9	ANGCGNATNNTNTTA	10.2505	13
	10	ANNGAANNTTCNNGAA	6.08498	6
	11	ANAAACAAANNNNNANA	5.45794	23
	12	GANNGNNCCAGAANNNT	4.70777	14
	13	ANNACAAANANANANNNA	4.65899	16
	14	ANANNNNGAANNNTTCAA	3.79046	13
	15	GGTTGGTNNTAA	2.47731	18
	16	TGCGTGTNTGT	1.76875	6
	17	GANAAGNNGTTNGC	1.7007	8
	18	AGAGNTCAAGA	0.277679	11
Ploidy (3 genes)	1	AGGNAAATNNNANA	18.7962	15
	2	ANGATNANTNANNNANNNAT	4.67535	11
	3	AANNAAAAANNAAA	0.901623	8
Sporulation (2 genes)	1	GAANNGAGANCGA	18.2556	15
	2	ANNGANANNGAANAAA	15.499	11
	3	ANGCAAGAAAG	13.8113	21
	4	ANACGCAANAAA	9.89551	11
	5	TAANGAAAAAA	6.15287	12
	6	GCCTNCTGCAA	4.3528	7
	7	GNANNNNNNAANANAANNNANNCG	1.82162	10
	8	ANAANANNNNANANGNNNNANNGA	0.763953	9
	9	AGAAAAAAAA	0.467177	7
	10	AAAAAAAATT	0.169044	5
Unfolded Protein Response (4 genes)	1	GAANAAAGANNNNANNNA	21.8295	19
	2	AGNAAGAAANNCA	20.8779	20
	3	GNAGANAANNAAANA	17.8457	20
	4	AAANAAGGAAA	17.2431	20
	5	ANAANAAANNNNNCNTNNGNA	12.7123	14
	6	TTCCNGAAAAT	10.3098	20
	7	GGAAAAANANNNANNNA	6.96802	16
	8	GGCAANNNNNAGANNANNNG	3.45366	18
Zinc Depletion (29 genes)	1	AAAATAGAAA	34.6482	34
	2	CACAANNNANNANAGA	28.6433	32
	3	GNGNNTACNNGNNAAANA	18.8487	16
	4	GNAACAGANNCNGNNA	14.4841	25
	5	GNTANTTCNNGAAG	14.0804	20
	6	GTNNTNGAANANNCCG	10.0923	16
	7	TGGATNGNTNTNCA	8.41713	15
	8	GNAANAANAANAAA	7.84771	13
	9	GNCTNNGNNNGAAAAA	6.67008	27
	10	AANCCTNGNANNNGCA	6.58552	19
	11	AANAANGAAGGC	5.73562	13
	12	GNCCGGAAAGG	4.67284	9
	13	AGGAAATTNTG	4.55577	14
	14	GCNNATANNANNGNNANNNTG	2.78771	15
	15	TACAGAACTA	2.51207	11
	16	CNGNGAAANGGAT	2.02354	17
	17	CAGTGGTNGAA	1.69826	8
	18	GCGNATCNGTTT	0.869176	12
	19	GGGTGACNCGT	0.82289	10
	20	GCGGGNAGNNCGNC	0.407902	8

APPENDIX I – AlignACE motifs recognized by Transfac

Table I.1: AlignACE motifs from Appendix H that were recognized by Transfac as being putative regulatory elements. Regulatory elements of interest have been highlighted.

Microarray experiment	Motif	AlignACE score	Regulatory element	Gene regulated by this element	Reference
Calcineurin	ACNCNNNNNATNAAAAA	41.9334	Myocyte-specific enhancer-binding factor (MEF-2)	Cardiac alpha-myosin heavy chain (*Mus musculus*)	Adolph *et al.*, 1993
	ATAAAATTTC	17.104	(A+T)-stretch binding protein (ATBP)	Lectin (*Sarcophaga peregrina*)	Nakanishi-Matsui *et al.*, 1995
DNA damaging agents	AAAAGNANAANNNAA	8.1873	(1) Protein-binding site identified by Methidiumpropyl-EDTA (MPE) footprinting (2) Ascorbate oxidase binding protein 1 (AOBP1)	(1) Actin (*Oryza sativa*) (2) Ascorbate oxidase (*Cucurbita maxima*)	(1) Wang *et al.*, 1992 (2) Kisu *et al.*, 1997; Kisu *et al.*, 1998
Environmental stress	AAAAAAAGANA	46.3991	(1) MPE (2) Agamous-like MADS-box protein 3 (AGL3)	(1) Actin (*O. sativa*) (2) Artifical sequence	(1) Wang *et al.*, 1992 (2) Huang *et al.*, 1995
	AAACAAAAGC	0.367363	Photoreceptor conserved element I (PCEI)	Arrestin (*M. musculus*)	Kikuchi *et al.*, 1993
Histone	ANNGAANNTTCNNGAA	6.08498	Heat Shock Factor (HSF)	Heat shock protein 70 (*Drosophila melanogaster*)	Wiederrecht *et al.*, 1987
Ploidy	AGGNAAATNNNANA	18.7962	Dorsal 1(D1)	Snail (*D. melanogaster*)	Ip *et al.*, 1992
	AANNAAAAANNAAA	0.901623	(1) MPE (2) ATBP (3) AOBP1	(1) Actin (*O. sativa*) (2) Lectin (*S. peregrina*) (3) Ascorbate oxidase (*C. maxima*)	(1) Wang *et al.*, 1992 (2) Nakanishi-Matsui *et al.*, 1995 (3) Kisu *et al.*, 1997; Kisu *et al.*, 1998
Unfolded protein response	GNAGANAANNAAANA	17.8457	TGTTT-binding (TGT3)	Apolipoprotein B (*Homo sapiens*)	Kardassi *et al.*, 1990; Paulweber *et al.*, 1991
	AAANAAGGAAA	17.2431	(1) AGL3 (2) Agamous-like protein 1 (AG1)	(1) Artifical sequence (2) Artifical sequence	(1) Huang *et al.*, 1995 (2) Schmidt *et al.*, 1993
Zinc	AAAATAGAAA	34.6482	Yin Yang 1 (YY1)	Beta-casein (*Rattus norvegicus*)	Raught *et al.*, 1994
	GNAANAANAANAAA	7.84771	(1) Antennapedia (2) CAAT/enhancer-binding protein (C/EBPalpha)	(1) Ultrabithorax (*D. melanogaster*) (2) Alcohol dehydrogenase 1 (*H. sapiens*)	(1) Winslow et al., 1989 (2) Stewart et al., 1991

References for Appendix I

Adolph,E.A., Subramaniam,A., Cserjesi,P., Olson,E.N. and Robbins J. (1993) Role of myocyte-specific enhancer-binding factor (MEF-2) in transcriptional regulation of the alpha-cardiac myosin heavy chain gene. *J. Biol. Chem.*, **268**, 5349-5352.

Huang,H., Tudor,M., Weiss,C.A., Hu,Y. and Ma,H. (1995) The *Arabidopsis* MADS-box gene AGL3 is widely expressed and encodes a sequence-specific DNA-binding protein. *Plant Mol. Biol.*, **28**, 549-567.

Ip,Y.T., Park,R.E., Kosman,D., Yazdanbakhsh,K. and Levine,M. (1992) Dorsal-twist interactions establish snail expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev.*, **6**, 1518-1530.

Kardassis,D., Hadzopoulou-Cladaras,M., Ramji,D.P., Cortese R., Zannis V.I. and Cladaras,C. (1990) Characterization of the promoter elements required for hepatic and intestinal transcription of the human apoB gene: Definition of the DNA-binding site of a tissue-specific transcriptional factor. *Mol. Cell. Biol.*, **10,** 2653-2659.

Kikuchi,T., Raju,K., Breitman,M.L. and Shinohara,T. (1993) The proximal promoter of the mouse arrestin gene directs gene expression in photoreceptor cells and contains an evolutionarily conserved retinal factor-binding site. *Mol. Cell. Biol.*, **13,** 4400- 4408.

Kisu,Y., Harada,Y., Goto,M. and Esaka,M. (1997) Cloning of the pumpkin ascorbate oxidase gene and analysis of a *cis*-acting region involved in induction by auxin. *Plant Cell Physiol.*, **39,** 631-637.

Kisu,Y., Ono,T., Shimofurutani,N., Suzuki,M. and Esaka,M. (1998) Characterization and expression of a new class of zinc finger protein that binds to silencer region of ascorbate oxidase gene. *Plant Cell Physiol.*, **39**, 1054-1064.

Nakanishi-Matsui,M., Kubo,T. and Natori,S. (1995) Molecular cloning and nuclear localization of ATBP, a novel (A+T)-stretch-binding protein of *Sacophaga peregrina* (flesh fly). *Eur. J. Biochem.*, **230**, 396-400.

Paulweber,B., Onasch,M.A., Nagy,B.P. and Levy-Wilson B. (1991) Similarities and differences in the function of regulatory elements at the 5' end of the human apolipoprotein B gene in cultured hepatoma (HepG2) and colon carcinoma (CaCo-2) cells. *J. Biol. Chem.*, **266**, 24149-24160.

Raught,B., Khursheed,B., Kazansky,A. and Rosen,J. (1994) YY1 represses beta-casein gene expression by preventing the formation of a lactation-associated complex. *Mol. Cell. Biol.*, **14**, 1752-1763.

Schmidt,R.J., Veit,B., Mandel,M.A., Mena,M., Hake,S., Yanofsky,M.F. (1993) Identification and molecular characterization of ZAG1, the maize homolog of the *Arabidopsis* floral homeotic gene AGAMOUS. *Plant Cell*, **5**, 729-737.

Stewart,M.J., Shean,M.L., Paeper,B.W. and Duester,G. (1991) The role of CCAAT/enhancer-binding protein in the differential regulation of a family of human liver alcohol dehydrogenase genes. *J. Biol. Chem.*, **266**, 11594-11603.

Wang,Y., Zhang,W., Cao,J., McElroy,D. and Wu,R. (1992) Characterization of *cis*-acting elements regulating transcription from the promoter of a constitutively active rice actin gene. *Mol. Cell. Biol.*, **12**, 3399-3406.

Wiederrecht,G., Shuey,D.J., Kibbe,W.A. and Parker,C.S. (1987) The *Saccharomyces* and *Drosophila* heat shock transcription factors are identical in size and DNA binding properties. *Cell*, **48**, 507-515.

Winslow,G.M., Hayashi,S., Krasnow,M., Hogness,D.S. and Scott,M.P. (1989) Functional activation by the Antennapedia and fushi tarazu proteins in cultured *Drosophila* cells. *Cell*, **57**, 1017-1030.

APPENDIX J – Python algorithm used to calculate base composition of DNA

```
##############################################
# Authors: Bronwen Aken and Corné Schriek
# Description: Counts  A,C,G,T,N's. Calculates GC%, ignoring N's.
# Date: May 2005
# For support e-mail : bronwen.aken@gmail.com or caschriek@gmail.com
##############################################

import sys

try:
   infile = open(sys.argv[1])
   outfile = open(sys.argv[2],"w")
except:
   print "Invalid or no filenames given "
   print "Usage: countBases inputfilename outputfilename"
   sys.exit()

a=0
c=0
g=0
t=0
n=0
total=0
for line in infile:
   if not line.startswith(">"):
     line = line.strip()
     for x in line:
       total +=1
       if x == "A" or x == "a":
            a+=1
       elif x == "C" or x == "c":
            c+=1
       elif x == "G" or x == "g":
            g+=1
       elif x == "T" or x == "t":
            t+=1
       elif x == "N" or x == "n":
            n+=1

outfile.write("A    : "+str(a)+"\n")
outfile.write("C    : "+str(c)+"\n")
outfile.write("G    : "+str(g)+"\n")
outfile.write("T    : "+str(t)+"\n")
outfile.write("N    : "+str(n)+"\n")
outfile.write("GC/ACGT  %: "+str(float(c+g)/(a+c+g+t) * 100)+"\n")
outfile.close()
```

DECLARATION

I certify that this research report has not been submitted for a degree in any other university and that it is my original work.

Signature: ……. Date: ……