Functional Analysis and Recombinant Expression of a Sea Urchin G-string Binding Factor

Johann Riedemann

Thesis Presented in Fulfilment of the Requirements for the Degree of Masters of Science at the University of Stellenbosch

Academic Supervisor : Prof. J.P. Hapgood

December 2001

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Summary

The sea urchin G-string binding factor 1 (suGF1) has previously been shown to bind with high affinity and selectivity to stretches of contiguous deoxyguanosine residues, a DNA motif found in the upstream regions of many unrelated genes from several organisms. It has been proposed that suGF1 plays a role in transcriptional regulation. Homopurine.homopyrimidine stretches have been shown to form unusual DNA structures, *in vitro*. To investigate the potential of the suGF1 binding site to form unusual structures under certain conditions, synthetic oligodeoxyribonucleotides containing the suGF1 poly(dG).(dC) binding site were subjected to circular dichroism (CD) analyses. The CD results indicate that the suGF1 binding site forms a mixture of unusual DNA structures, as deduced by comparison with the spectra obtained for B-DNA, triplex and quadruplex conformations. These results are consistent with the hypothesis that suGF1 specifically recognises G-strings that exhibit unusual structures.

Exhaustive database searches showed that suGF1 has no significant homology with any previously identified proteins or cDNAs from any species. Given the relevance of mammalian models to medical science, and since no sea urchin cell lines are currently available, the identification of a mammalian functional homologue would facilitate determination of the *in vivo* function of such a potentially important, putative, novel DNA-binding protein in mammalian cell lines. In this study sequence analysis tools were used to identify hORFX, a putative human functional homologue of suGF1. Similarities in the domain organisation of the two proteins, prompted an investigation into the DNA-binding properties of hORFX, as well as a more detailed structure prediction analysis, with a view to determining whether hORFX is a functional homologue of suGF1. hORFX was successfully expressed *in vitro*, but lacked the ability to specifically bind G-strings.

Theoretical predictions suggest that suGF1 has a DNA-binding domain belonging to a different family to that predicted for hORFX, consistent with differences in their respective DNA-binding specificities. suGF1 and hORFX were predicted to have helix-turn-helix and helix-loop-helix DNA-binding domains, respectively. Taken together the results do not support the hypothesis that hORFX is a suGF1 homologue.

To date, no direct evidence for the *in vivo* function of suGF1 has been obtained. With a view to performing transactivation assays in the future, the expression of suGF1 in yeast was investigated in this project. An suGF1 expression construct was engineered and transformed into a protease-deficient yeast strain. Nuclear extracts were prepared and subjected to SDS-PAGE and electrophoretic mobility shift assays (EMSAs). suGF1 was shown to be successfully expressed in yeast cells and exhibited similar G-string-binding properties to that of native and *in vitro* transcribed and translated (IVT) suGF1. The suGF1 cDNA was also subjected to *in silico* expression, which together with the SDS-PAGE results of yeast nuclear extracts and IVT suGF1, indicated that the protein might be expressed as multiple truncated products, due to the utilisation of multiple AUG translation start sites. These *in vitro* results are crucial for the ultimate outcome and correct interpretation of future transactivation experiments and lay the foundation for further investigation into the possible role of suGF1 in transcriptional regulation.

Opsomming

In die verlede is bewys dat die seepampoentjie G-string-bindende faktor (suGF1) hoë affiniteit en spesifisiteit vir aaneenlopende volgordes van deoksiguanosien residue besit. Hierdie DNA motief kom algemeen voor in die stroom-op gebiede van verskeie gene in verskillende organismes. Daar is 'n veronderstelling dat suGF1 betrokke is by die regulering van geenuitdrukking.

Vroeër is bewys dat homopurien.homopirimidien-ryke areas die vermoë besit om *in vitro* ongewone DNA-strukture te vorm. Die potentiaal van die suGF1-bindingsetel om ongewone DNA-strukture te vorm is gevolglik deur sirkulêre dikroïsme (SD) analise ondersoek. Vergelyking van die spektra vir B-DNA-, tripleks- en kwadrupleks-strukture met dié van die suGF1-bindingsetel, toon duidelik dat laasgenoemde 'n mengsel van ongewone DNA konformasies, onder die spesifieke eksperimentele omstandigehede, aanneem.

Deeglike inspeksie van die beskikbare geen- en proteïendatabasisse vir alle spesies het aangetoon dat suGF1 geen merkbare kDNA- of proteïenhomoloë besit nie. As gevolg van die belang van soogdiermodelsisteme in die mediese wetenskappe, asook die onbeskikbaarheid van seepampoentjie-sellyne, is 'n soektog na 'n funktionele suGF1 homoloog in soogdiere geloods. Die ontdekking van só 'n homoloog sal dit moontlik maak om die rol van hierdie potensiaal belangrike en unieke DNA-bindingsproteïen te ondersoek. Tydens hierdie soektog is spesiale analise-programme gebruik en 'n potensiële menshomoloog van suGF1, hORFX, is geïdentifiseer. Die mees prominente ooreenkoms tussen die twee proteïene is die soortgelyke rangskikking van funksionele motiewe. Gevolglik is die DNA-bindings eienskappe van die hORFX-proteïen ondersoek,

insluitende 'n detaileerde struktuur-funksie-voorspelling ten einde vas te stel of dit wél 'n homoloog van suGF1 is. hORFX is suksesvol uitgedruk *in vitro*, maar besit nie die vermoë om dieselfde G-string waaraan suGF1 spesifiek bind te herken nie. Teoretiese analise het voorspel dat suGF1 en hORFX aan verskillende DNA-bindings proteïen-families behoort, aangesien suGF1 'n heliks-draai-heliks en hORFX 'n heliks-lus-heliks motief bevat. Hierdie inligting, tesame met die eksperimentele resultate, dui aan dat hORFX nie 'n homoloog van suGF1 is nie.

Tot op hede is daar niks bekend aangaande suGF1 se funksie *in vivo* nie. Met die oog op transaktiveringseksperimente in die toekoms, is die ekspressie van suGF1 in gisselle tydens hierdie navorsingsprojek ondersoek. 'n suGF1 ekspressievektor is berei en gebruik om 'n protease-negatiewe gissellyn te transformeer. Kernekstrakte is ondersoek deur SDS-PAGE en elektroforetiese mobiliteitsessais. Daar is gevind dat suGF1 suksesvol uitgedruk is in die gisselle. Die rekombinante suGF1 besit G-volgorde bindingsaktiwiteite soortgelyk aan dié van suGF1 in kernekstrakte van seepampoentjies, asook *in vitro* getranskribeerde-en getransleerde suGF1. Die kDNA vir suGF1 is ook *in silico* uitgedruk. Tesame met die SDS-PAGE-resultate het laasgenoemde aangetoon dat die suGF1-kDNA veelvuldige AUG-kodons bevat vir die inisiasie van proteïentranslasie. Dit lei moontlik tot die translasie van 'n reeks proteïenprodukte wat verkort is aan die N-terminale kant, afgesien van die volledige suGF1-proteïen. Die *in vitro* resultate in geheel is essensieel vir die toekomstige uitvoering en interpretasie van transaktiveringseksperimente. Hierdie projek lê gevolglik die fondasie vir 'n verdere ondersoek na die rol van suGF1 in die regulering van geenuitdrukking.

Part of the work presented in this thesis has been published:

Hapgood JP, Riedemann J and Scherer SD, 2001. **REGULATION OF GENE EXPRESSION BY GC-RICH DNA *CIS*-ELEMENTS.** *Cell Biology International* **25**: 17 - 31.

Acknowledgements

I wish to thank

My supervisor, Professor Janet Hapgood, for her guidance, support and encouragement during the course of this project.

Professor Hugh Patterton (University of Cape Town) for supervising the recombinant expression of suGF1 in yeast.

Doctor Marina Rautenbach for expert advice on circular dichroism and protein-structure.

Professor Theo Mevissen (German National Research Center for Information Technology) for expert advice on tertiary structure prediction.

My parents, sisters, grandmother, friends and Tommie for their continual support, motivation and enthusiasm in my work.

My fiancée, Trix Lubbe, for her endless support, motivation and enthusiasm in my work.

The National Research Foundation and the Harry Crossley Foundation for financial assistance.

Abbreviations

Δε	Change in Ellipticity
3-D	Three Dimensional
ATP	Adenosine Triphosphate
BGP1	β-Globin Binding Protein 1
bp	Base Pairs
BSA	Bovine Serum Albumin
-C	Negative Control
CAT	Chloramphenicol Acetyltransferase
CD	Circular Dichroism
cpm	Counts Per Minute
dATP	Deoxyadenosine 5'-Triphosphate
DB	Dialysis Buffer
DBD	DNA-Binding Domain
dCTP	Deoxycytosine Triphosphate
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
dpm	Disintegrations Per Minute
DSM	Discrete State-Space Model
DTT	Dithiothreitol
EDTA	Ethylenediaminetetra-acetate
EMSA	Electrophoretic Mobility Shift Assay
Gal	Galactose
H1 – H4	Histone 1 – Histone 4
HPLC	High Pressure Liquid Chromatography
IVT	*In vitro* Transcription and Translation
kDa	Kilo Dalton
kPa	Kilo Pascal
LB	Luria Bertani
Mr	Molecular Weight
mRNA	Messenger RNA
NE	Nuclear Extracts

NS-Oligo	Non-Specific Oligodeoxyribonucleotide
OD	Optical Density
PAGE	Polyacrylamide Gel Electrophoresis
PDB	Protein Database
PEG	Polyethylene Glycol
PMSF	Phenylmethyl Sulfonyl Fluoride
PSA	Protein Sequence Analysis
R	Ribonucleic Acid
RNase A	Ribonuclease A
rpm	Revolutions Per Minute
SDS	Sodium Dodecyl Sulphate
S-Oligo	Specific Oligodeoxyribonucleotide
SpGCF1	Strongylocentrotus Purpuratus GC-Binding Factor 1
suGF1	Sea Urchin G-string Binding Factor 1
TAE	Tris Acetate EDTA
TE	Tris-EDTA
TGE	Tris Glycine EDTA
TSB	Transformation and Storage Buffer
Ura	Uracil

Contents

CHAPTER 4 - RESULTS

CHAPTER 5 - RESULTS

CHAPTER 6

Chapter 1

Introduction

Central to transcriptional regulation and deciphering structural or regulatory information encoded in genomes, is the ability of sequence-specific proteins to recognise and bind regulatory gene sequences (Johnson and McKnight, 1989; Mitchel and Tijian, 1989; Saltzman and Weinmann, 1989). The biochemical interaction between message-carrying proteins and regulatory DNA sequences constitute the crux of gene regulation, a research field of intensive activity.

Sea urchin embryos have been extensively utilised as model systems to study eukaryotic developmental processes such as gene regulation. A natural advantage of sea urchin embryos for the molecular analysis of gene regulation is the relatively large amount of biological material available. The embryos can be grown, manipulated and studied with relative ease, and is therefore an ideal candidate system for investigating the molecular mechanics of eukaryotic developmental gene regulation (Calzone *et al.*, 1991).

1.1 Sea Urchin Development

Sea urchin embryos develop in a relatively uncomplicated fashion, in which different lineages descendant from a uniform set of cleavage-stage founder cells, express specific sets of genes to mediate exact formation of prospective cell territories (Lee and Calzone, 1986). The cell lineages construct five territories that are defined in terms of patterns of macromolecular expression and the ultimate cell fate. The prospective aboral ectoderm, oral ectoderm, skeletogenic mesenchyme, vegetal plate and eight small micromeres

constitute these five territories of differential gene expression, by means of an invariant pattern of complete cleavage (Cameron and Davidson, 1991). The rapid division of the fertilised egg into these polyclonal territories eventually gives rise to specific cell types and structures (Type I embryogenesis). During this process certain sets of predestined genes start to be expressed in a highly ordered and regulated manner. These genes code for proteins that mediate specific and controlled cellular differentiation and specialisation. High stringency control of temporal, spatial and quantitative gene expression comprises a delicate framework of combinations of sequence-specific factors occupying various specific target sites. These transcription factors bring chemical messages to their target genes, by means of interactions with one another, ancillary proteins and other components of the basal transcription machinery (Kirchhamer *et al.*, 1996). Several parameters are therefore involved in the precise mechanism by which a target gene in the developing embryo is controlled as to specify the events by which the blastomeric cell mass diverge into the distinct territorial identities.

1.2 Regulation of Gene Expression During Sea Urchin Development

In sea urchins many genes and their respective *cis*-regulatory control machinery have been researched and documented in the literature. Two of these genes, *Endo*16 and *CyIIIa*, are expressed in different embryonic territories and are subjected to strict spatial, temporal and quantitative control. The regulatory regions of these genes contain numerous G.C-rich target sites and are able to bind nuclear proteins (Zeller *et al.*, 1995). A *Parenchinus Angulosis* sea urchin G-string binding factor 1 (suGF1) and its specie homologue *Strongylocentrotus Purpuratus* GC-binding factor 1 (SpGCF1) have been shown to bind preferentially to G.C-rich elements within the promoter control regions of the *Endo16* and *CyIIIa* genes, *in vitro* (Zeller *et al.,* 1995). Their function might therefore be

essential for the correct expression patterns of these two genes. The suGF1 / SpGCF1 binding sites, as well as various other diverse transcription factor binding sites are arranged into discrete modules, each module responsible for some particular sub-element of the overall expression pattern of the gene (Chiou-Hwa *et al.*, 1998; Kirchhamer *et al.*, 1996; Roush, 1996; Wray, 1998).

The *Endo*16 gene codes for a cell surface glycoprotein, the expression of which is restricted to the vegetal plate of the blastula stage embryo and continues through the archenteron (to which the vegetal plate gives rise) in gastrulation (Kirchhamer *et al.*, 1996; Yuh *et al.*, 1994). Transcription is eventually shut down in all other regions except the midgut, where it is increased. The positive regulatory functions of the proximal, central and distal promoter modules are curbed by the negative interactions, which prevent incorrect expression in the adjacent skeletogenic and ectodermal territories. Diverse target sites, including suGF1 / SpGCF1 sites, are distributed throughout the *cis*-regulatory domain of this gene (Fig. 1.1). 23 suGF1 / SpGCF1 sites (present in different modules) are present in the *Endo*16 promoter region and could therefore function in intermodule communication (Kirchhamer *et al.*, 1996; Yuh and Davidson, 1996; Zeller *et al.*, 1995a).

The *CyIIIa* aboral ectoderm-specific actin gene is activated in the late cleavage stage embryo (Franks *et al.*, 1990; Kirchhamer *et al.*, 1996; Roush, 1996; Zeller *et al.*, 1995b). The gene is initially expressed in eleven clones of the original blastomeres, which ultimately give rise to the aboral ectoderm of the embryo. The *CyIIIa* gene contains clusters of suGF1 / SpGCF1 sites in the distal regulatory domain in a variety of patterns and is often in close proximity to other DNA-binding sites (Fig. 1.2) (Kirchhamer *et al.*, 1996; Zeller *et al.*, 1995b). The modules comprising the *cis*-regulatory region (more than 2.3 kb) have separate functions but are quantitatively dependent on each other, and are all

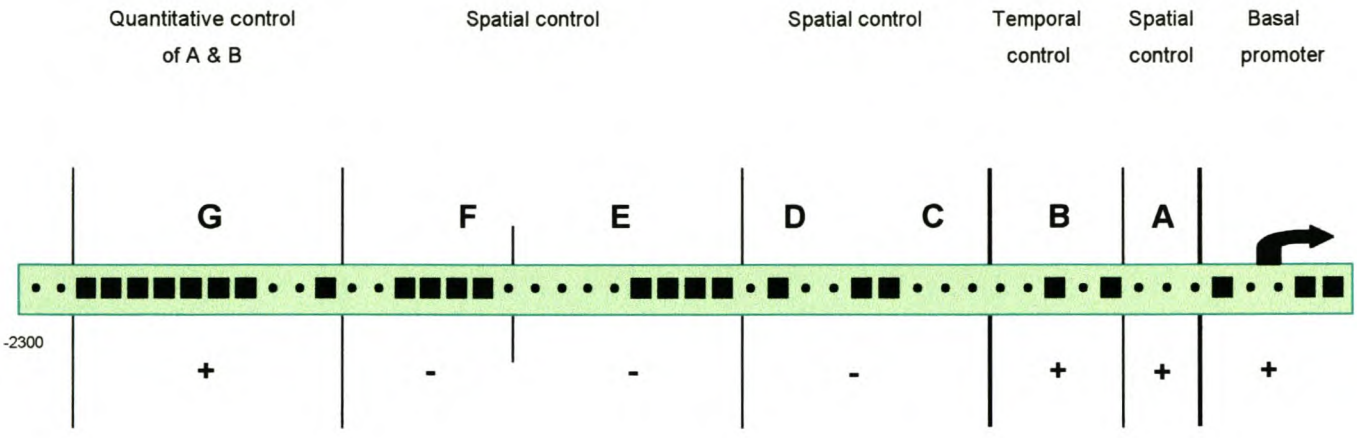


Fig. 1.1 **Diagram depicting the modular arrangement of the 2300-bp 5'-upstream regulatory control region of the *Endo16* gene.**

The 5'-upstream *cis*-regulatory control region of the *Endo16* gene is functionally organised in discrete modules, which function interdependently of each other to produce the global expression pattern of the gene. suGF1 / SpGCF1 sites are indicated by squares and the sites for other factors by the dots. suGF1 / SpGCF1 binding sites are present in all modules, emphasizing the putative role of this protein in regulation of the gene. Module A restricts expression to the vegetal plate during early embryogenesis, while module B promotes expression in the midgut later during postgastrulation. Modules C – F shut off expression at the boundaries of the vegetal plate, ensuring expression only in the vegetal half. Module G controls the level of expression of modules A and B. Positive and negative effects are indicated by + and – signs. The arrow indicates the transcription start site (Yuh et al., 1998) (Diagram compiled by J.Riedemann).

required for normal embryonic expression of the *CyIIIa* gene (Franks *et al.*, 1990; Hough-Evans *et al.*, 1990), possibly by intercommunication between their regulatory sites (Kirchhamer *et al.*, 1996).

The *cis*-regulatory systems of these two genes therefore seem to display a modular (or regional) functional organisation, as opposed to a dispersed or interspersed arrangement. Specific modules (or sub-elements) of the regulatory DNA perform specific and highly controlled developmental subfunctions, which is separable from the basal promoter of the gene. The obvious importance of suGF1 / SpGCF1 in regulation of these and many other developmentally significant genes during sea urchin embryogenesis, spurred an investigation into the molecular mechanism by which this protein acts.

1.3 The Sea Urchin G-string Binding Factor 1 (suGF1)

GC-rich *cis*-regulatory elements and their cognate binding proteins have been strongly implicated in developmental gene regulation and provide a good example of the general principles governing eukaryotic transcriptional regulation. These sequences are the most ubiquitous regulatory elements and are present in upstream (Denver *et al.*, 1999; Kohwi and Kohwi-Shigematsu, 1991; Li and Seetharam, 1998; Redell and Tempel, 1998) and downstream (Baumann *et al.*, 1999; Lisowsky *et al.*, 1999; Oda *et al.*, 1998) regions of several unrelated eukaryotic genes in many organisms. Several functions have been ascribed to these regions, in most cases involving positive or negative regulation of transcription. Several GC-box binding factors from a variety of tissues and organisms are able to associate with these sequences (Denver *et al.*, 1999; Li and Seetharam, 1998; Lisowsky *et al.*, 1999; Redell and Tempel, 1998), making this mechanism of gene regulation a very relevant and interesting topic.

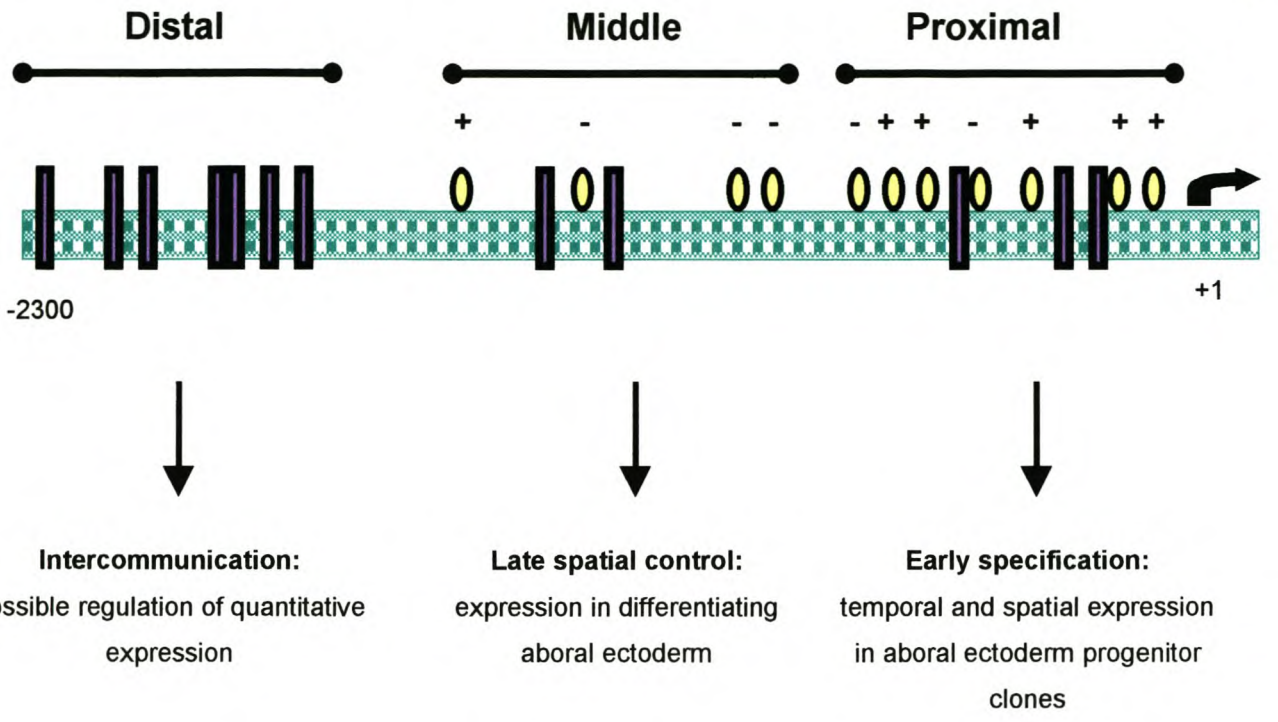


Fig. 1.2 Diagram depicting the modular organisation of transcriptional machinery within the 2300 bp *cis*-regulatory domain of the *CyIIIA* gene.

The 5'-upstream regulatory region of the *CyIIIA* gene consist of an estimated twenty or more sites to which at least 9 different transcription factors bind. This domain consists of three discrete modules, which have been shown to be sufficient for spatial, temporal and quantitative expression of the gene. Rectangular bars indicate specific suGF1 / SpGCF1 binding sites and the oval-shaped circles the binding sites for other transcription factors. Positive and negative effects on gene expression via binding of these factors (all except suGF1 / SpGCF1) are indicated by + and – symbols. The transcription start site is indicated by the arrow at the +1 position (Kirchhamer *et al.*, 1996) (Diagram compiled by J.Riedemann).

In sea urchins suGF1 / SpGCF1 appears to be the major embryonic factor binding to GC-rich DNA (Patterton and Hapgood, 1994; Zeller *et al.*, 1995b). suGF1 / SpGCF1 has high binding affinity and specificity for GC-rich DNA sequences *in vitro* (Hapgood and Patterton, 1994; Zeller *et al.*,1995). The consensus recognition sequence for suGF1 / SpGCF1 is 5'-GGGNGGG-3' or 5'-GGGGGGC-3' (Hapgood and Patterton, 1994), but further degeneracy within the binding site is possible. Although suGF1 / SpGCF1 does not form an obligate homodimer when binding to DNA target sequences (Scherer, 1997), it is nevertheless able to form sequence-specific multimers via suGF1-suGF1 interactions (Patterton and Hapgood, 1994; Zeller *et al.*, 1995a; Zeller *et al.*, 1995b) *in vitro*. This indicates that suGF1 / SpGCF1 could be involved in looping DNA *in vivo*, thereby bringing distant regulatory regions into close proximity to each other (Zeller *et al.*, 1995a).

suGF1 / SpGCF1 contains a highly basic DNA-binding domain (Scherer, 1997; Zeller *et al.*, 1995b), a feature that is common to a diverse set of transcription factors. The DNA-binding domain is located centrally in the protein and is closely associated with a potential heptad of repeats of hydrophobic amino acids (Seid *et al.*, 1996), which are also found in other regions of the protein and are reminiscent of a putative dimerisation domain. Furthermore, a proline-rich putative transactivation domain occurs in the N-terminus of suGF1 / SpGCF1, consistent with its role as a transcription factor that interacts with other proteins.

The cDNAs for suGF1 (Scherer, 1997) and SpGCF1 (Zeller *et al.*, 1995b) show that these proteins do not contain zinc fingers and are structurally unrelated to Sp1 (the classic, ubiquitous mammalian GC-box binding factor). However, suGF1 may be functionally and / or structurally related to BGP1 (β-globin binding protein 1), a chicken transcription factor involved in regulation of the β-globin genes. Both factors bind to G-strings in their potential

target genes *in vitro*, which occur within regions that interact with a nucleosome. Alterations in chromatin structure occur for both genes as a function of transcriptional activation *in vivo* (Patterton and Hapgood, 1994; Patterton and Hapgood, 1996). suGF1 is able to bind with similar affinity to both the G_{11}-string in the H1-H4 intergenic region and the β-globin G-string (Hapgood and Patterton, 1994), *in vitro*. Indeed, suGF1 and BGP1 produce identical footprints *in vitro* on the β-globin gene promoter (Hapgood and Patterton, 1994). However, neither suGF1 (Patterton and Hapgood, 1996) nor BGP1 (Clark *et al.*, 1990) can displace a nucleosome *in vitro.* suGF1 / SpGCF1 could also be related to the mammalian IF-1 factor which binds to the α1 and α2 collagen gene promoters (Hapgood and Patterton, 1994; Karsenty and DeCrombrugghe, 1991). Whether suGF1 is related to BGP1 or IF-1 will become apparent if the cDNAs for these factors are eventually cloned.

Alterations in chromatin structure of the sea urchin early histone gene battery *in vivo* correlate with the temporal expression pattern of these genes. After blastulation, sea urchins express the late histone genes and the early genes are switched off never to be expressed again (Chiou-Hwa *et al.,* 1998; Palla *et al.*, 1999). The early histone genes have nuclease hypersensitive intergenic spacers, whereas the shutdown of the genes correlates with the presence of a positioned nucleosome (Patterton and Hapgood, 1994; Patterton and Hapgood, 1996). The control mechanism governing the gene switch from early to late histone genes has not yet been elucidated. However, there is strong indirect evidence that suGF1 may be involved in regulation of the sea urchin histone gene battery via alterations in chromatin structure within the GC-rich DNA region (Patterton and Hapgood, 1994; Patterton and Hapgood, 1996). suGF1 binds *in vitro* to a region comprising eleven G residues in the H1-H4 spacer (Hapgood and Patterton, 1994; Patterton and Hapgood, 1996). This region has been shown *in vitro* to contain a strong nucleosome-positioning site (Patterton and Hapgood, 1996; Patterton and Von Holt, 1993). The G_{11}-string lies close to

the dyad of the positioned nucleosome core, and the nucleosome positioning signal lies over the sequence $(GA)_{16}(G)_{11}$. This G.C stretch has the ability to form an unusual triple helical DNA structure under conditions of negative superhelical stress and low pH, *in vitro* (Patterton and Von Holt, 1993; Stokorva *et al.*, 1989), consistent with a link between the occurrence of unusual DNA structures and alterations in chromatin structure.

It appears that suGF1 / SpGCF1 may function as a general transcriptional regulator of several unrelated genes in sea urchin development. Various promoter deletion experiments (Flytzanis *et al.*, 1987) and *in vivo* target site competition studies (Franks *et al.*, 1990) in sea urchin embryos and zygotes validate this hypothesis. However, direct evidence that this protein is a transcription factor has yet to be obtained.

1.4 An Overview of Gene Regulation by GC-box Binding Proteins

When searching the literature, it was interesting to discover that many different species contain proteins that recognise GC-rich DNA sequences (Tables 1.1 and 1.2). Several different transcription factors interact with GC-rich DNA sequences i.e. sequences which contain predominantly Gs and Cs on both strands, where a particular strand can contain only Gs or only Cs, or both Gs and Cs. For the purposes of this introduction, GC-rich *cis*-elements will be defined as those that contain at least 78% GC content. These would include the so-called GC boxes (Izmailova *et al.*, 1999; Nielsen *et al.*, 1998), as well as *cis* elements that contain runs of only Gs on one strand and runs of only Cs on the other strand (so called G-strings). A summary of these transcription factors from different species, as well as their DNA-binding sites, is given in Tables 1.1 and 1.2. These tables will be referred to extensively when discussing these *cis*-elements and GC-box binding proteins (section 1.4.1 to 1.4.4).

GC box-binding protein	DNA binding site	Genes regulated	Tissue distribution	Effect on transcription	Structural features of DNA-binding and/or multimerisation domains	Possible *in vivo* function	Reference
Sp1	GGGGCGGGGC and variants of this sequence	Almost all genes containing GC-boxes	Ubiquitous	+	3 C2H2 Zn-fingers.	Ubiquitous regulation of transcription Essential for early embryonic development and maintenance of differentiating cells	Al-Asadi *et al*., 1995 Birnbaum *et al*., 1995 Marin *et al*., 1997 Nielsen *et al*., 1998 Philipsen and Suske, 1999
Sp3	Sp1-like binding sites	Almost all genes regulated by Sp1 and many others	Ubiquitous	+/-	3 C2H2 Zn-fingers	Repress Sp1-mediated transcriptional activation	Nielsen *et al*., 1998 Philipsen and Suske, 1999
Sp4	Sp1-like binding sites	ADH5 gene gERE promoters	Brain, testes, developing teeth, epithelial tissue	+/-	3 C2H2 Zn-fingers	Repress Sp1-mediated transcriptional activation. Highly expressed in developing central nervous system	Kwon *et al*., 1999 Philipsen and Suske, 1999
BTEB-1	Sp1-like binding sites	CYP7A αI(I) collagen Uteroferrin	mRNA ubiquitous	+/-	C2H2 Zn-finger	Neuronal process formation, Developmental function	Philipsen and Suske, 1999 Rosalia *et al*., 1999
BTEB-2	Sp1-like binding sites	?	Intestine and placenta	+	C2H2 Zn-finger	Cell proliferation in intestine Also expressed in placenta	Philipsen and Suske, 1999
GZP1	GC box-like sequence	TAS regions in mitochondrial DNA	?	+	C2H2 Zn-finger	Cell proliferation	Lisowsky *et al*., 1999
ZBP-89	GCCCCTCCCCC	Promoters of rat gastrin, human ht β T-cell receptor, mouse BFCOL1, type I collagen, human β enolase, ODC genes	Ubiquitous	+/-	4 C2H2 Zn-fingers	Inhibition of cell proliferation Often represses Sp1-mediated activation	Law *et al*., 1998
Zif268 (NGFI-A, Krox-24 or Egr1)	CGCCCCCGC	Membrane type matrix metalloproteinase	Ubiquitous	+	Zn-finger	Mitogenesis and cell differentiation	Patletich and Pabo, 1991
NGFI-C (Egr-4)	GCGGGGGCG	Early response genes	Neuronal cells	+	Zn-finger	Rapid response to certain cellular signals (e.g. NGF)	Crosby *et al*., 1991 Philipsen and Suske, 1999
Zf9/CPBP	GC-box-like sequence	TGF-β1 Type 1 and 2 TGF-β receptor Minimal collagen α1(I) promoter	Ubiquitous	+	C2H2 Zn-finger,	Possible role in up-regulation of genes involved in tissue repair. ZF9 upregulated during early hepatic fibrosis in rats	Kim *et al*., 1998
IF-1	GGGGGGG	α1(I) and α2(I) Collagen gene promoters	?	-	?	Developmental regulation of collagen genes	Karsenty and DeCrombrugghe, 1991
GCF	NG/CCGG/CG/CG/CCN	EGFR, β-actin and Ca^{2+}-dependent protease promoters	Widely expressed; not ubiquitous	-	N-terminal DNA-binding domain. Dimerisation via 2 leucine zipper	Down-regulation of unrelated genes	Takimoto, 1999
TIEG	GGGGCGGGGC (Sp1)	?	Osteoblasts and other tissue	-	?	Cell growth	Cook *et al*., 1998
G10BP-1	GGGGGGGCGGG and variants of sequence	Rat fibronectin	?	-	Putative basic-zipper structure capable of forming homodimers	Sp1 negative regulator	Li and Seetharam, 1998 Oda *et al*., 1998

Table 1.1. A summary of the essential features of mammalian GC box-binding factors. Note the abundance of factors containing either zinc-fingers or leucine zippers (table compiled by J.Riedemann).

Abbreviations: TAS - termination associated sequence; TSS - transcription start site; ODC - ornithine decarboxylase; TGF - transforming growth factor; NGF - nerve growth factor; EGFR - epidermal growth factor

GC box-binding protein	Species	DNA binding site	Genes regulated	Tissue distribution	Effect on transcription	Structural features of DNA-binding and/or multimerisation domains	Possible *in vivo* function	Reference
suGF1	Sea urchin (*P.angulosis*)	5' GGGNGGG 3' and variations on this	*Endo16, CyIIIA, SM30,* early histone genes	Various embryonic tissue types	+	Basic DNA binding domain	spatial, temporal and quantitative regulation	Hapgood and Patterton, 1994 Patterton and Hapgood, 1994 Patterton and Hapgood, 1996
SpGCF1	Sea urchin (*S.purpuratus*)	5' GGGNGGG 3' and variations on this	*Endo16, CyIIIA, SM30,* early histone genes	Various embryonic tissue types	+	Basic DNA binding domain	spatial, temporal and quantitative regulation	Zeller *et al.*, 1995
D-Sp1	Fruitfly	Possibly Sp1-like sites	?	Head-specific	?	3 C2H2 Zn-fingers	sensory organ development	Schock F, 1999. D-Sp1 is involved in sensory organ development. *Unpublished.*
BHD	Fruitfly	Possibly Sp1-like sites	?	?	?	3 C2H2 Zn-fingers	head segmentation	WImmer *et al.*, 1993
BGP1	Chicken	Various G-strings and GC-boxes e.g. suGF1 and Sp1 consensus sites	Chicken β_A-globin	Erythroid cells, possibly other tissues	?	Requirement for zinc; possible Zn-finger	erythrocyte development	Clark *et al.*, 1990

Table 1.2. Summary of the essential features of known non-mammalian GC box-binding factors (table compiled by J.Riedemann).

Abbreviations: BHD - buttonhead; BGP1 - beta globin protein 1.

1.4.1 GC-Rich DNA *cis* elements

GC-rich *cis* regulatory sequences do not appear to be confined to a particular class of genes. They have, for example, been identified in the regulatory domains of several housekeeping (Redell and Tempel, 1998), tissue-specific (Clare *et al.*, 1997) and viral (Birnbaum *et al.*, 1995) genes, including genes for enzymes (Al-Asadi *et al.*, 1995; Arcott and Deininger, 1992; Lee *et al.*, 1999; Taketani *et al.*, 1999), receptors (Augustin *et al.*, 1995; Lacy *et al.*, 1994; Maouche *et al.*, 1995), ion-channels (Chu *et al.*, 1999; Redell and Tempel, 1998), cytokines (Masuda *et al.*, 1994), structural proteins (Oda *et al.*, 1998) and DNA-binding proteins (Marin *et al.*, 1997). Interestingly, they do appear to occur frequently in the upstream regions of genes that do not contain TATA boxes, initiator elements or CCAAT boxes (Asundi *et al.*, 1998; Blake *et al.*, 1990; Koritschoner *et al.*, 1997; Li and Seetharam, 1998; Redell and Tempel, 1998). This implies that factors binding to these elements may be essential for basal transcription from these promoters. For example, the 5' proximal minimal promoter of the *KCNJ2* potassium channel gene contains three GC box consensus elements, but lacks TATA- and CCAAT-box elements (Redell and Tempel, 1998). The rat proteoglycan *GPC1* gene is also devoid of classic TATA- and CCAAT-box motifs, but contains multiple GC boxes (Asundi *et al.*, 1998). GC-rich sequences have been detected in upstream regions of genes from many species including insects, sea urchins, amphibia and mammals (Table 1.1 and 1.2 and references therein). The location of GC-rich sequences also varies between genes, being found in promoters (Chen *et al.*, 1997), enhancers (Masuda *et al.*, 1994) and locus control regions (Pruzina *et al.*, 1994). GC-rich DNA is found in genes that are under different modes of control, such as cell cycle regulation (Birnbaum *et al.*, 1995), hormonal activation (Rosalia *et al.*, 1997) and developmental patterning (Kwon *et al.*, 1999; Philipsen and Suske, 1999). In addition GC boxes are involved in regulation of mitochondrial replication (Lisowsky *et al.*, 1999). Thus,

it appears that the function of GC-rich DNA is not linked to a particular cellular process or mechanism of regulation.

1.4.2 Role of GC boxes in Development

A role in development has been shown for many of the studied genes containing GC-rich upstream regions (Table 1.1 and 1.2). There is strong evidence that GC boxes, the binding site for the ubiquitous Sp1 mammalian transcription factor, play an essential role in early mammalian embryonic development. Marin *et al.* (1997) observed that $Sp1^{-/-}$ mutant mouse embryos are retarded in development, exhibit a broad range of growth abnormalities and ultimately die around day eleven of gestation, although the embryonic stem (ES) cells deficient in Sp1 showed normal growth and differentiation. The results indicate that once ES cells are differentiated, Sp1 is necessary for maintenance of the differentiated cells, most likely via regulation of genes like *MeCP2* (methyl-CpG-binding protein). The MeCP2 protein binds to methylated DNA and is thought to repress transcription *in vivo* via interaction with specific histone deacetylases (Mastrelangelo *et al.*, 1991; Wade *et al.*, 1998). Although the expression levels of most of the Sp1-regulated genes were unchanged, the levels of MeCP2 were greatly reduced. Thus the authors concluded that Sp1 is essential for early embryonic development, but not for growth and differentiation of primitive cells (Marin *et al.*, 1997). In addition to Sp1, other mammalian GC factors have also been implicated in developmental gene regulation. The gastrin EGF response element (gERE), which is thought to function in the developing and neoplastic stomach, is a GC-rich element to which many factors (including Sp1) bind (Merchant *et al.*, 1995). Binding of the Sp1-like factor BTEB-1 to GC- boxes in the uteroferrin gene is thought to play a role in the pregnancy-associated growth and development of endometrial epithelial tissue (Rosalia *et al.*, 1999). Interaction of the factor IF-1 with a

poly(dG).poly(dC) stretch in the αI(1) and αI(2) collagen gene promoters plays a role in developmental regulation of the collagen genes (Hasegawa *et al.*, 1996; Karsenty and DeCrombrugghe, 1991).

1.4.3 The GC-box Binding Protein Family

Mammalian Factors:

Sp1 is by far the most well studied example of a transcription factor, which binds to GC-rich *cis* elements (Kadonaga, 1987; Mitsuhiro, 1998). Sp1 belongs to a family of transcription factors characterised by a highly conserved DNA-binding domain consisting of three C_2H_2 zinc-fingers. The GC-rich target sites for Sp1 are variations of the sequence 5' –GGGGCGGG- 3' and this factor binds to these sites with high affinity. However, in addition, Sp1 also recognises GT or CACCC boxes, although with slightly lower affinity (Hagen *et al.*, 1992; Philipsen and Suske, 1999). Sp1 binding sites often appear in clusters in promoter regions, allowing the Sp1 protein to act synergistically through adjacent binding sites (Al-Asadi *et al.*, 1995). Sp1 interacts with itself to form multimeric complexes (Mastrelangelo *et al.*, 1991; Nielsen *et al.*, 1998; Pascal and Tjian, 1991) resulting in looping out of the intervening DNA, reminiscent of the ability of suGF1 / SpGCF1 to do the same (Section 1.3). This suggests a mechanism whereby distant DNA elements are brought into close proximity of each other as a result of stabilising protein-protein interactions (Mastrelangelo *et al.*, 1991; Pascal and Tjian, 1991; Philipsen and Suske, 1999). Sp1 has been shown to interact and co-operate with a variety of proteins involved in transcription, which include regulatory factors such as NF-κB (Fuminori *et al.*, 1998; Mastrelangelo *et al.*, 1991), E2F, p53, RB, STAT-1, GATA-1 (Merika and Orkin, 1995) as well as transcription factors like TATA-box associated factors and even TATA-box binding protein itself (Izmailova *et al.*, 1999; Philipsen and Suske, 1999). Since Sp1

regulates many different genes, the specificity of control is ensured via several mechanisms. These include post-translational modifications such as phosphorylation and O-linked glycosylation (Jackson and Tijian, 1988; Philipsen and Suske, 1999), regulation of Sp1 affinity for its target site (Hagen *et al.*, 1992; Sogawa *et al.*, 1993), alteration of its *trans*-activation potential (Kim *et al.*, 1992) and regulation of Sp1 concentration relative to other proteins (Courey and Tjian, 1992; Nehls *et al.*, 1992).

More recently, several other mammalian three C_2H_2 zinc-finger factors that bind to GC boxes have been identified (Table 1.1), illustrating the diversity and flexibility of eukaryotic transcriptional regulation by zinc-finger GC box-binding factors. Members of the family include Sp1, Sp2, Sp3, Sp4 (Izmailova *et al.*, 1999; Marin *et al.*, 1997; Philipsen and Suske, 1999), BTEB1, TIEG1, TIEG2 and the Krüppel-like factors BTEB2, ZF9 (Kim *et al.*, 1998), ZNF741, AP-2rep (Philipsen and Suske, 1999). Sp1, Sp3 and Sp4 can bind the Sp1 consensus sequence and recent evidence suggests that Sp3 and Sp4 can repress Sp1-mediated transcriptional activation by competing with Sp1 for binding to the core *cis* elements (Fuminori *et al.*, 1998; Nielsen *et al.*, 1998). The relative ratios of Sp1, Sp3 and Sp4 in different cells may therefore be a critical parameter of gene regulation. While not members of the three C_2H_2 zinc-finger GC box-binding factor family, other mammalian zinc-finger GC box-binding proteins have also been detected, which may function in a similar manner to Sp1 (Suzuki *et al.*, 1998). For example, the transcription factor ZBP-89 contains four C_2H_2 zinc fingers and is ubiquitously expressed (Law *et al.*, 1998). Interestingly, unlike Sp1, ZBP-89 is predominantly a repressor of transcription (Law *et al.*, 1998; Lee *et al.*, 1999). A GC-box in the proximal promoter of the ornithine decarboxylase gene is required for basal and induced transcriptional activity, with Sp1 and ZBP-89 binding to this region in a mutually exclusive manner (Law *et al.*, 1998). ZBP-89 has also

been shown to act as repressor for basal and inducible expression of the human gastrin gene (Law *et al.*, 1998; Nielsen *et al.*, 1998).

Not all the mammalian transcription factors binding to GC-rich DNA, for which the structures are known, contain zinc fingers. Examples of these are GC-binding factor (GCF) (Takimoto, 1999) and G_{10}-binding protein 1 (G10BP-1) (Oda *et al.*, 1998). GCF is characterised by a leucine zipper-like motif that might function as a dimerisation domain. Sp1 can also recognise its binding site (5' –GCGGGGC- 3' and variations thereof). GCF acts as a sequence-specific repressor, either by competing with various activators for DNA-binding sites or by interaction with other proteins to achieve repression (Kageyama and Pastan, 1989b). G-rich sequence binding factor, G10BP-1, recognises a G_{10}-string present in the rat fibronectin promoter region and is responsible for repression of Sp1-mediated transcriptional activation by excluding the binding of Sp1 to this site. G10BP-1 forms homodimers through its basic-zipper structure (Oda *et al.*, 1998).

Several other mammalian factors, which bind to GC-rich DNA, but for which the structures are not known, have also been identified e.g. IF-1 (Karsenty and DeCrombrugghe, 1991), ETF (Kageyama and Pastan, 1989a) and H4TF1 (Daily *et al.*, 1986) (see Table 1.1).

Non-Mammalian Factors:

Relatively few non-mammalian proteins that bind to GC-rich DNA *cis* elements have been identified and characterised. Sp1 homologues have been detected in *Drosophila melanogaster* (Wimmer *et al.*, 1993), indicating that Sp1-like proteins are conserved through evolution. The *buttonhead* and *D-Sp1* genes of *Drosophila melanogaster* are involved in head-specific segmentation and sensory organ development, respectively (Wimmer *et al.*, 1993). The chicken erythrocyte factor, BGP1, (see section 1.3) is

implicated in β-globin gene regulation via binding to GC-rich DNA (Clark *et al.*, 1990). Since the BGP1 cDNA has not been cloned nor has the protein been sequenced, the structural relationship between BGP1 and other mammalian and non-mammalian GC-rich factors is not known. An Sp1 homologue does not appear to exist in sea urchins. Instead, in sea urchins, suGF1 / SpCGF1 appears to be the major embryonic factor binding to GC-rich DNA (Patterton and Hapgood, 1994; Zeller *et al.*, 1995b).

1.4.4 DNA-Binding Specificity

If one examines the DNA-binding sites for proteins that recognise GC-rich DNA (see Tables 1.1 and 1.2), a striking feature is the similarity between these sequences. In addition, many of the factors have also been shown to tolerate considerable degeneracy in their high-affinity binding sites *in vitro*. For example, Sp3, Sp4, BTEB1, TIEG2 (Cook *et al.*, 1998; Hagen *et al.*, 1992) and Egr-1 (Haas *et al.*, 1999) all recognise classic Sp1-binding elements (Chiou-Hwa *et al.*, 1998; Pascal and Tijan, 1991). A K_D of 4.6×10^{-10} to 3.1×10^{-9} has been determined for Sp1 binding to the GGGCGGG motif and other classic Sp1-binding sites (Hagen *et al.*, 1992; Sogawa *et al.*, 1993). Sp1-like factors are also capable of binding to GT or CACCC boxes, albeit with slightly lower affinity. Additionally, Sp1 has been shown to bind to NF-κB DNA binding sites. NF-κB can, however, not recognise Sp1-binding sites, which occur frequently in the promoter or enhancer regions of NF-κB-regulated genes e.g. HIV-1 (Jones *et al.*, 1994), c-Rel (Viswanathan *et al.*, 1996) and the δ opoid receptor (Augustin *et al.*, 1995) genes. suGF1 can also recognise the 5'-GGGCGGG-3' Sp1 site with high affinity (Hapgood and Patterton, 1994).

Binding of these GC box-binding factors to their respective target sites therefore seems to be quite promiscuous. This might increase the capability of these factors to regulate

genes with different patterns of deoxyguanosine and deoxycytosine distribution, ensuring maximum flexibility with regard to promoter activity. It is also apparent that, despite having very similar DNA-binding specificities, some of these factors have very different structures in their DNA binding domains (DBDs). Although many of the factors contain zinc fingers in the DBDs, some contain the basic leucine zipper structure e.g. G10BP-1 (Oda *et al.*, 1998), while others do not conform to any previously characterised DBD (e.g. suGF1 / SpGCF1).

1.5 Aims of this MSc Project

The unique features of suGF1, as well as its potential role in the regulation of gene expression, possibly via binding to unusual DNA structures, prompted further investigation towards establishing the *in vivo* function of this protein. The specific aims of this research project were as follows:

- *In vitro* transcription and translation of suGF1 to provide a ready available source of protein. This would be essential for future investigations into the DNA-binding properties of suGF1 and would provide controls for the presence of a functional protein.

- Identification of a mammalian functional homologue of suGF1 to provide clues as to the *in vivo* function of this protein. The function of such a mammalian protein could then in future be investigated in mammalian cell lines, since no sea urchin cell lines are as yet available.

- *In vitro* transcription and translation of the putative mammalian homologue of suGF1, to provide a ready available source of the protein. This would be useful for investigations into the DNA-binding properties of the mammalian protein.

- Investigation into the DNA-binding properties of native suGF1, *in vitro* transcribed and translated suGF1, as well as the putative mammalian homologue by electrophoretic mobility shift assays, to compare their respective G-string-binding properties.

- Theoretical sequence analysis and structure-function predictions of suGF1 and the putative mammalian homologue, to investigate the possible structural similarities and differences between these two proteins.

- Investigation into the structural properties of the suGF1 poly(dG).d(C) binding site by circular dichroism. Homopurine.homopyrimidine stretches have been shown to form unusual DNA structures, *in vitro*, and it might therefore be possible that the G-strings to which suGF1 binds specifically, have the potential to exhibit such structures.

- Preparation of an suGF1 expression construct that could be transformed into yeast cells to produce recombinant suGF1.

- Expression of suGF1 in yeast and the preparation of nuclear extracts to investigate whether the protein was successfully expressed.

- Investigation into the DNA-binding properties of recombinant suGF1. This would be essential for the ultimate outcome and correct interpretation of future transactivation

experiments and lay the foundation for further investigation into the possible role of suGF1 in transcriptional regulation.

Chapter 2

Materials and Methods

2.1 Materials

All solvents and chemical reagents were analytical grade unless otherwise stated. The sources of solvents and chemical are mentioned once and remain the preferred and utilised agents throughout the methods section, unless otherwise stated. All solutions, glassware and plastics were sterilised by autoclaving (at 120 °C and 10 kPa for 60 minutes) or sterile filtering. Distilled and / or analytical quality (prepared via a Milli-Q filter system) water was used throughout.

2.2 Plasmid Propagation and Isolation

2.2.1 Plasmids

A list of the plasmids used within the scope of this research project is given in Table 2.1.

2.2.2 Competent Cells

E.coli strains, JM109, HB101 and DH5α (Pharmacia) were streaked out on Luria Bertani (LB) agar plates (1% (w/v) tryptone (Merck), 0.2 M NaCl (Saarchem; chemically pure), 0.5% (w/v) yeast extract powder (Amersham), 1.5% (w/v) agar (Amersham)) and incubated overnight at 37°C. Competent bacterial cells were prepared and transformed, by either the method described by Chung *et al.* (1992) or by a standard protocol in which

Plasmid	Description	Vector Size	Insert Size	Antibiotic resistance	Source	Reference
pHP2	Part of the H1-H4 intergenic region of the major early histone gene battery of *P.miliaris* inserted at *HindIII-AflIII* site of the pHP5 vector	1800 bp	201 bp	Ampicillin	Prof. H-G Patterton, University of Cape Town	Patterton and von Holt, 1993
pBluescript SK^+ - hORFX	Human cDNA clone HA1331, gene name KIAA0043. Inserted at *EcoRV-NotI* site of the pSK^+ vector	3000 bp	3028 bp	Ampicillin	Constructed by Prof. T Nagase from Kazusa DNA Research Institute pBluescript SK^+ from Stratagene	Beck *et al.*, 1992
pcDNAI/Amp-suGF1	suGF1 cDNA inserted at *XhoI–NotI* site of the pcDNAI/Amp vector	4801 bp	2000 bp	Ampicillin	Constructed by Dr S.D. Scherer pcDNAI/Amp from Invitrogen	Scherer, 1997
pGEMT-suGF1	suGF1 cDNA inserted at *SalI - SacI* site of the pGEMT vector	3003 bp	2000 bp	Ampicillin	Constructed by Dr S.D. Scherer. pGEMT from Promega	Scherer, 1997
pYES2-suGF1	suGF1 cDNA inserted at *HindIII – XbaI* site of the pYES2 vector	5900 bp	2000 bp	Ampicillin	pYES2 from Invitrogen	http://www.invitrogen.com/content/vectors/pyes2.pdf

Table 2.1 A summary of the plasmids used within the scope of this research project.

the cells are made electrocompetent (Sambrook *et al.,* 1989, Chang *et al.,* 1992). Both methods will be described.

2.2.2.1 Preparation of Competent *E.coli* Cells using the DMSO Chemical Method

A single colony was picked from a fresh LB agar plate and inoculated in 10 ml LB broth (1% (w/v) tryptone, 0.2 M NaCl, 0.5% (w/v) yeast extract powder), overnight. This starter culture was incubated at 37°C overnight on a shaker platform (220 rpm). The next morning 1% (v/v) of the starter culture was inoculated in 100 ml fresh LB medium and incubated at 37°C on a shaker platform (220 rpm). The culture was grown to early log-phase ($OD_{600\ nm}$ = 0.3 – 0.6). The cells were pelleted by centrifugation (3020 x g for 10 minutes at 4°C) and subsequently resuspended in $1/12.5^{th}$ volume of transformation and storage buffer (TSB: LB broth (pH6.1) containing 10% (w/v) polyethylene glycol (PEG) (Merck, Mw = 4000 $g.mol^{-1}$), 5% (v/v) DMSO (Merck, For synthesis), 10 mM $MgCl_2$ (Saarchem) and 10 mM $MgSO_4 \cdot 7H_2O$ (Saarchem)) at 4°C. The resuspended cells were then incubated at 4°C for 10 minutes, and either stored at -80°C or used immediately for transformation.

2.2.2.2 Transformation of DMSO Competent *E.coli* Cells

100 µl competent cells were transformed with 100 - 200 ng supercoiled plasmid DNA, mixed well and incubated on ice for 60 minutes. The cells were supplemented with 900 µl TSB containing 20 mM D-glucose (Synthon) and incubated on a shaker platform at 37°C with vigorous shaking (250 rpm) to allow expression of the antibiotic resistance gene. Colonies containing the plasmid of interest were selected by plating the cells on LB agar plates containing 50 µg/ml of the appropriate antibiotic (ampicillin (Sigma) unless otherwise stated). The transformation efficiencies were calculated as the number of

colonies per μg DNA. Typically transformation efficiencies of 10^5 to 10^6 transformants per μg DNA were obtained.

2.2.2.3 Preparation of Electrocompetent *E.coli* Cells

A single colony was picked from a fresh LB plate (described in 2.2.1), inoculated in 50 ml SOB medium (2% (w/v) Bactotryptone (Biolab), 0.5% (w/v) Bacto Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$ and 10 mM $MgSO_4 \cdot 7H_2O$) and incubated overnight at 37°C on a shaker platform. 500 ml SOB was then inoculated with 5 ml of the overnight culture, in a two-liter flask, and incubated at 37°C with vigorous shaking on a shaker platform, until an optical density of 0.8 at 595 nm was obtained. The culture flask was chilled on ice, together with two 250 ml centrifuge bottles. Subsequent steps were all performed at 4°C. The culture was transferred into the two pre-chilled centrifuge bottles and the cells were harvested by centrifugation (2500 x g at 4°C for 15 minutes). The supernatant was decanted and the cells were washed and resuspended in 200 ml (100 ml for each bottle) ice-cold, sterile, distilled water. Cells were pelleted by centrifugation (2500 x g at 4°C for 15 minutes) after which the washing step was repeated and the cells were harvested by centrifugation (2500 x g at 4°C for 15 minutes). The cells were washed in ice-cold sterile 10% (v/v) glycerol (HolPro Analyticals), while care was taken not to disturb the pellet when decanting the supernatant. An extra, small volume of ice-cold sterile 10% (v/v) glycerol was added to the cell pellet to resuspend the cells to a density of 100–200 OD units as measured at 595 nm. Aliquots of the cell suspension were dispensed into pre-chilled microcentrifuge tubes, snap-frozen in liquid nitrogen and stored at -80°C.

2.2.2.4 Electroporation of Competent *E.coli* Cells

The Savant Gene Transformer™ was set up and tested as outlined in the instructions manual. Typical bacterial electroporations were performed in cuvettes with gap size 1 mm and the voltage setting at 1800 V, which induced a field strength of 18 kV/cm for 5 milliseconds. For each electroporation, one sterile microcentrifuge tube and one electroporation cuvette was chilled on ice. The required amount of electrocompetent cells was removed from the -80°C freezer and thawed on ice. The DNA was dispensed into pre-chilled microcentrifuge tubes at concentrations of 0.1 to 1 ng/μl (maximum volume of DNA was 10 μl). The DNA used for electroporation was always dissolved in sterile analytical water or low ionic strength buffer to avoid arcing. 40–80 μl of electrocompetent bacteria was added to the DNA, pipetted up and down to mix, and then quickly transferred into the gap of the pre-chilled electroporation cuvette (supplied with the apparatus). The cuvette was tapped against the side, allowing the sample to settle in the bottom of the slot and to minimize the introduction of air bubbles that might cause arcing. The cuvette was wiped with tissue paper and placed into the cuvette chamber after which a single pulse was delivered. The cuvette was quickly removed and 1 ml SOC medium (SOB containing 20 mM D-glucose) was added to remove the cells from the cuvette gap. The cells were transferred to a culture tube and incubated for 1 hour at 37°C in a shaking incubator (225 rpm). Dilutions of the grown cultures were prepared and 50–100 μl aliquots were plated on LB-agar plates containing 50 μg/ml of the appropriate antibiotic (ampicillin, unless otherwise stated). The plates were incubated overnight at 37°C. Typical transformation efficiencies ranged from 10^6 to 10^7 transformants per μg DNA, which was slightly higher than the DMSO method described in section 2.2.1.2.

2.2.3 Plasmid DNA Mini-Preparation by the Alkaline Lysis Method

Small-scale plasmid DNA preparations were performed according to the established method of Birnboim and Doly (1979) as described by Sambrook *et al.* (1995). A single, transformed bacterial colony was picked from a plate containing the appropriate antibiotic and transferred to 10 ml LB medium containing 50 μg/ml of the appropriate antibiotic, in a sterile, loosely capped 50-ml tube. The cells were grown overnight at 37°C with vigorous shaking. The cells were harvested by centrifugation (2000 x g at 4°C for 10 minutes). The supernatant was carefully decanted and the remaining medium was removed with a pipette, leaving the bacterial pellet as dry as possible, while avoiding manipulation of the cells. The pellet was dissolved in 200 μl ice-cold Solution 1 (50 mM D-glucose, 25 mM Tris-HCl (Merck) pH 8.0 and 10 mM EDTA (Saarchem) pH8.0) by vigorous vortexing. 400 μl of a freshly prepared Solution 2 (freshly prepared 0.2 N NaOH (Saarchem), 1% (w/v) SDS (BDH)) was added. The tube was closed tightly and rapidly inverted five times, ensuring that the entire surface of the tube came into contact with Solution 2. The tubes were stored on ice. 300 μl ice-cold solution 3 was added and the tube was gently vortexed in an inverted position for at least 10 seconds. The tubes were incubated on ice for 5 minutes after which the viscous bacterial lysate was aliquoted into pre-chilled microcentrifuge tubes. The lysate was centrifuged (12000 x g at 4°C for 5 minutes) in a microfuge. The supernatant was transferred to a fresh pre-chilled microcentrifuge tube. The DNA was extracted once by the addition of an equal volume phenol:chloroform (1:1 v/v) (Merck) followed by precipitation with 600 μl isopropanol (Merck) at room temperature. The DNA was recovered by centrifugation (12000 x g at 4°C for 5 minutes), washed with 1 ml of 70% (v/v) ethanol (Merck) at 4°C, evaporated to dryness and resuspended in 50 μl TE (10 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0 containing 100 μg/ml RNase A (Boehringer Mannheim)). The sample was incubated at 37°C for one hour to allow optimal

digestion of the RNA molecules by RNase A. All plasmid DNA samples were stored at -20°C. Typical plasmid mini-preparations yielded 5 to 10 μg supercoiled DNA per milliliter original cell culture. Often DNA minipreps had to be re-digested with RNase A, due to the presence of undigested RNA. The integrity of the DNA samples was subsequently analysed by agarose gel electrophoresis.

2.2.4 Large Scale Plasmid Isolation

Plasmids were propagated in *E.coli* strains JM109, HB101 or DH5α. A 10–100 ml culture of transformed cells were grown overnight at 37°C on a shaker platform in LB medium containing 50 μg/ml of the appropriate antibiotic. Plasmids were then isolated by using Wizard®Midipreps DNA Purification System (Promega) according to the supplier's recommendations.

Bacterial cultures were pelleted by centrifugation (10000 x g at 4°C for 10 minutes, JA14 rotor (Beckman)). The cell pellet was resuspended in 3 ml Cell Resuspension Solution (10 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, and 100 μg/ml RNase A) after which lysis was achieved by addition of 3 ml Cell Lysis Solution (0.2 N NaOH, 1% (w/v) SDS). The plasmids were released from the lysed cells by gently inverting the tubes 4-6 times, after which the mixture was supplemented with 3 ml Neutralisation Solution (1.32 M potassium acetate, pH 4.8). Chromosomal material and cell debris were pelleted by centrifugation (14000 x g at 4°C for 15 minutes, JA20 rotor (Beckman)). The supernatant was collected and filtered through a Whatman® #1 filter. The Wizard® Purification Resin (7 M Guanidine HCl) was resuspended and 10 ml was added to the supernatant and passed directly over a Wizard® MidiPrep MidiColumn. Solvents were eluted from the column by applying a vacuum to the column, which was then washed twice with 15 ml Column Wash Solution

(8.3 mM Tris-HCl pH 7.5, 83 mM NaCl, 2 mM EDTA, 58% (v/v) ethanol). The column was dried and the DNA was eluted with 300 µl pre-heated (65°C) TE (pH 8.0) by spinning for 20 seconds in a microcentrifuge at 10 000 x g. DNA samples were aliquoted and stored in microcentrifuge tubes at -20°C or -80°C. The concentration of DNA samples was determined spectrophotometrically from the absorbance value measured at 260 nm. The ratio of the absorbance values at 260 and 280 nm respectively was used to assess the purity of the DNA. The integrity of the DNA samples was analysed by agarose gel electrophoresis.

2.3 Enzymatic Manipulation of DNA

2.3.1 Restriction Enzyme Digests

Restriction enzyme digest reactions were performed as described by Sambrook *et al.* (1989). Typical reactions contained the appropriate amount of DNA (0.2 µg to 10µg), 1 x reaction buffer (supplied with the enzyme) and 4–6 units of each restriction enzyme (Amersham or Roche) per µg DNA. The final volume was adjusted with nuclease free water (Promega) so that the volume of enzyme present is always equal to or less than 10% (v/v) of the final reaction volume. The digest reaction was incubated at 37°C overnight and terminated by the addition of EDTA to a final concentration of 0.1 M, followed by addition of 1/6th volume of gel loading buffer (0.25% (w/v) bromophenol blue (BDH), 0.25% (w/v) xylene cyanol (Sigma), 30% (v/v) glycerol). The integrity of the digest products was analysed by agarose gel electrophoresis.

2.3.2 Ligation Reactions

Ligation reactions were performed as described by Sambrook *et al.* (1989). Typical reactions contained predetermined amounts (see equation 1) of insert and plasmid DNA in

Equation 1:

$$\frac{\text{ng of vector x kb size of insert}}{\text{kb size of vector}} \times \text{molar ratio of } \frac{\text{insert}}{\text{vector}} = \text{ng of insert}$$

a 5:1 to 10:1 ratio (insert : plasmid), 1 x ligation buffer (30 mM Tris-HCl pH7.8, 10 mM $MgCl_2$, 10 mM DTT and 1 mM ATP; supplied with the enzyme) and 2 – 4 units T4 DNA Ligase (Promega) per μg total DNA. The final volume was adjusted with nuclease-free water so that the volume enzyme present is always equal to or less than 10% (v/v) of the final reaction volume. Sticky-end ligation reactions were incubated at room temperature for 12 to 16 hours.

2.4 Isolation and Purification of DNA from Preparative gels

2.4.1 Isolation and Purification of DNA fragments from Polyacrylamide Gels

DNA fragments were isolated and purified from polyacrylamide gels by a technique originally described by Maxam and Gilbert (1977). The fragments obtained from enzymatic digestion of the pHP2 plasmid (containing part of the H1-H4 intergenic region of the major early histone gene battery of *P.miliaris* as a 201 bp *HindIII / AflIII* insert) (Table 2.1) with *EcoRI* and *HindIII,* were resolved on a 4% (w/v) polyacrylamide gel in 1 x TBE buffer (0.089 M Tris-borate, 0.089 M boric acid, 0.002M EDTA) at room temperature. Pre-electrophoresis was for 2 h at 100 V. Electrophoresis was for 5 hrs at 130 V (1-8 V/cm).

The 335 bp *EcoRI-HindIII* fragment was excised from the gel (Appendix 2) after visualisation by UV-shadowing on a thin-layer chromatography plate (Merck; silica gel F254). DNA was eluted from the chopped-up gel slices by addition of 2-3 volumes freshly prepared elution buffer (0.5 M ammonium acetate (Saarchem), 1 mM EDTA (pH 8.0)) and shaking overnight at 37°C. The acrylamide was pelleted by centrifugation at 10 000 x g in a microcentrifuge. The supernatant was removed. The acrylamide-containing pellet was washed with two volumes elution buffer and centrifuged at 10 000 x g for 1 minute at 4°C in a microcentrifuge. The supernatants were combined and passed through a siliconised glass wool plug. The DNA was precipitated with 1/10th volume of 3 M sodium acetate (pH 5.2) and two volumes ice-cold absolute ethanol (30 minutes, 4°C) and recovered by centrifugation at 12000 rpm for 10 minutes (4°C). The DNA pellet was dissolved in 200 µl TE (pH 7.5) by heating for 5 minutes at 56°C. The DNA was re-precipitated with 1/10th volume of 3 M sodium acetate (pH 5.2) and 2 volumes ice-cold absolute ethanol, washed with 70% (v/v) ethanol, dried, resuspended in TE (pH 7.5) and stored in aliquots. The DNA concentration was determined spectrophotometrically or by ethidium bromide spotting (Ausubel *et al.,* 1987).

2.4.2 Isolation and Purification of DNA Fragments from Agarose Gels

The pcDNAI/Amp construct containing an suGF1 cDNA insert (Table 2.1) was subjected to restriction enzyme digestion as described in section 2.3.1. The insert was isolated and purified from preparative agarose gels by using the Nucleospin™ Extract 2 in 1 kit (Macherey and Nagel) according to the supplier's recommendations. The fragments were resolved on 1% (w/v) agarose gels in TAE (0.04 M Tris acetate (Saarchem), 0,002 M EDTA) containing 0.5 µg/ml ethidium bromide. The relevant bands were visualised by a hand-held ultraviolet lamp and excised from the gel with a sterile scalpel. 300 µl buffer

NT1 (supplied with the kit) was added to each 100 mg agarose gel and incubated for 10 minutes at 50°C with mild vortexing every 2 minutes. The sample was loaded onto a Nucleospin™ Column tube, placed into a 2 ml centrifuge tube and centrifuged at 6000 x g for 60 seconds in a microcentrifuge at room temperature. The flowthrough was discarded, the Nucleospin™ Column Tube was replaced into a 2 ml centrifuge tube (supplied with the kit) and re-centrifuged after addition of 700 µl buffer NT3 (supplied with the kit). The washing step with buffer NT3 was repeated once. The flowthrough was discarded and the Nucleospin™ Column Tube was centrifuged for 60 seconds at maximum speed in order to remove residual ethanol. The Nucleospin Column Tube was placed in a sterile 1.5 ml microcentrifuge tube after which 50 µl elution buffer NE (pre-heated to 70°C) was added. The DNA was eluted from the column by centrifugation for 60 seconds at 20 000 x g in a microcentrifuge at room temperature. The DNA concentration was determined spectrophotometrically or by ethidium bromide spotting (Sambrook *et al.,* 1989). DNA samples were stored at -20°C or -80°C.

2.5 Oligodeoxyribonucleotides

2.5.1 Sequences

A list of the oligodeoxyribonucleotides and their sequences is given in Table 2.2.

2.5.2 Synthesis and Annealing

Oligodeoxyribonucleotides were synthesised on a Beckman Oligo 1000M DNA synthesiser (Beckman Instruments Inc.) by Ms Pei-Yin Ma (University of Cape Town). The S-Oligo (specific oligodeoxyribonucleotide) contained a sequence from the H1-H4 intergenic region

Oligodeoxyribonucleotide	Sequence	Reference
S-Oligo	5′AGAGAGAGAGGGGGGGGGGAGGGAGAATTGC 3′ 3′TCTCTCTCTCCCCCCCCCCCTCCCTCTTAACG 5′	Patterton and Hapgood, 1994.
NS-Oligo	5′GATCTTCTGCACTCTCACCGGTACTGGACTGA 3′ 3′CTAGAAGACGTGAGAGTGGCCATGACCTGACT 5′	Patterton and Hapgood, 1994.
B-DNA	5'GAAGAGAGG 3' 3'CTTCTCTCC 5'	Klump and Chauhan, 2000. Personal Communications (University of Cape Town)
Triplex	5'CTTCTCTCC 3' 3'GAAGAGAGGCCTTGGAGAGAAG 5'	Klump and Chauhan, 2000. Personal Communications (University of Cape Town)
Quadruplex	5'GGGGTTTTGGGG 3'	Klump and Chauhan, 2000. Personal Communications (University of Cape Town)

Table 2.2 A summary of the oligodeoxyribonucleotides used within the scope of this research project.

of the early histone gene battery of *P.miliaris*, to which suGF1 binds specifically (Scherer, 1997). The NS-oligo (non-specific oligodeoxyribonucleotide) contained a random, heterologous sequence to which suGF1 did not bind (Hapgood and Patterton, 1994). The molar extinction coefficient of each oligodeoxyribonucleotide was determined from the extinction coefficients of the individual bases (Ausubel, *et al.,* 1987). Complementary strands (500 ng/μl per strand; in water) were annealed at equimolar ratios (1:1), by incubating at 88°C for 2 min, 65°C for 10 min, 37°C for 25°C for 5 min and placing the sample on ice (Sambrook *et al.,* 1989). Annealed oligodeoxyribonucleotides were resolved on a 2% (w/v) agarose gel in TAE containing 0.5 mg/ml ethidium bromide to assess the annealing efficiency. Electrophoresis was for 1 h at 90 V. The DNA was visualised on an ultraviolet box or by using a hand-held ultraviolet lamp. The annealed oligodeoxyribonucleotides were stored at -20°C. The B-DNA and triplex oligodeoxyribonucleotides were annealed by Ms M.Chauhan (University of Cape Town, laboratory of Prof. H.Klump) in a similar fashion. The annealed triplex and B-DNA oligodeoxyribonucleotides as well as the single-stranded quadruplex oligodeoxyribonucleotide were kindly provided by Ms M.Chauhan for CD analysis.

2.6 Radioactive Labeling of DNA

The *EcoRI – HindIII* (E/H) fragment and oligodeoxyribonucleotides (Appendix 2) were end-labeled by a Klenow (Roche) fill-in reaction according to an established protocol (Sambrook *et al.,* 1995). 100 ng of an *EcoRI - HindIII* fragment was labeled by filling in the *HindIII* site using 10 μCi/μl [α-^{32}P]dCTP (Amersham) as radionucleotide. Oligodeoxyribonucleotides were labeled with the same concentration [α-^{32}P]dCTP. The reaction (final volume 20 μl) was incubated for 1 h at 37°C after which 1 μl EDTA (0.5 M) and 79 μl TE was added. The labeled DNA was separated from unincorporated

nucleotides on a Sephadex G50 (Pharmacia) spin column equilibrated with TE buffer, as described by Sambrook *et al.* (1989). The E/H fragment presented with specific activities of 2 x 10^7 to 5 x 10^7 dpm/μg.

The synthetic oligodeoxyribonucleotides (Table 2.2) were radioactively labeled in the presence of [γ-^{32}P]ATP (Amersham) and Polynucleotide Kinase (from *E.Coli*; Boehringer Mannheim). 200 ng double stranded oligodeoxyribonucleotides (equivalent to 20 pmol 5'-hydroxy termini) were incubated with 5 μl [γ-^{32}P]ATP , 1 X ligation buffer (30 mM Tris-HCl pH 7.8, 10 mM $MgCl_2$, 10 mM DTT and 1 mM ATP) and 10 units of Polynucleotide Kinase (10 U/μl). The final reaction volume (20 μl) was adjusted with nuclease-free water so that the volume enzyme present in the reaction mixture, is always equal to or less than 10% (v/v) of the final volume. The samples were incubated at 37°C for 30 minutes. The reaction was stopped by cooling in an ice bath for 5 minutes. The labeled DNA was separated from unincorporated nucleotides on a Sephadex G50 (Pharmacia) spin column equilibrated with TE buffer, as described by Sambrook *et al.* (1989). The specific activities of the radiolabeled oligodeoxyribo-nucleotides ranged between 1 x 10^7 and 1.5 x 10^7 dpm/μg. All radiolabeled DNA samples were stored at -20°C.

2.7 *In vitro* Transcription and Translation of Expression Constructs

In vitro transcribed and translated (IVT) suGF1 was expressed from the full-length suGF1 cDNA cloned into the *SalI - SacII* site of the pGEM-T vector from Promega (Table 2.1), using the TNT® T7 Quick Coupled Transcription-Translation procedure (Promega). The same kit was used to express the hORFX protein from the full-length cDNA cloned into the *NotI - EcoRV* sites of the pBluescript SK^+ vector (constructed and provided by T. Nagase from the Kazusa DNA research institute) (Table 2.1). The kit was used according to the

supplier's recommendations. All kit components were stored at -70°C. The lysate was stored in aliquots and was never subjected to more than two freeze-thaw cycles. All reactions were performed in a designated RNase-free hood, using RNase-free chemicals, plastics and glassware. In some cases the suGF1 and hORFX proteins were radioactively labeled by the addition of [^{35}S]Methionine (10.5 μCi/μl) to the IVT reaction. For a standard reaction 40 μl TNT rabbit reticulocyte Mastermix, 1 μg DNA, 5 μl RNase-free water, 1 μl RNasin and 1 μl [^{35}S] labeled or unlabeled methionine were mixed, the final reaction volume was adjusted to 50 μl with nuclease-free water (supplied with the kit) and the reaction mixture was incubated for 2 hours at 30 °C. Aliquots of the *in vitro* expressed protein products were then stored at -80°C. The radioactively labeled IVT products were analysed by SDS polyacrylamide gel electrophoresis (SDS-PAGE) as described in section 2.9. The unlabeled IVT reaction products were used for investigating protein-DNA interactions in electrophoretic mobility gel shift assays.

2.8 Electrophoretic Mobility Shift Assays

Electrophoretic mobility shift assays (EMSA) were performed, as previously described (Hapgood and Patterton, 1994), using proteins from different sources. Sea urchin nuclear extracts were prepared and kindly provided by Dr S.Scherer (Ph.D. thesis, 1997). IVT suGF1 and hORFX were prepared as described in section 2.7. Recombinant suGF1 was heterologously expressed in a yeast expression system, from which nuclear extracts were prepared (Section 2.10).

Radiolabeled DNA (5 ng) (10000 cpm/lane) was dissolved in EMSA incubation buffer (16 mM Tris-HCl (pH 8.0), 175 mM KCl, 16% (v/v) glycerol, 1.6 mM $MgCl_2$, 0.8 mM DTT (Merck), 0.4 mM PMSF (Merck), 1 mM EDTA, 0.5 to 1.0 μg pdIdC (Roche) and 10 μg of

BSA (Molecular Biology Grade, Roche)) in a total volume of 20 μl, prior to incubation with the protein source. Sea urchin embryo nuclear extracts (5-10 μl) (0.54 μg/μl total protein); 5, 10 or 15 μl IVT protein products; or 5 μl recombinantly expressed suGF1 (0.5 μg/μl total protein) was then added to the cocktail containing radiolabeled DNA, to start the reaction. In some experiments the amount of DNA and protein was increased to a total volume of either 45 μl or 65 μl (incubation/buffering conditions were changed proportionally). In the case of competitor EMSAs, 100-fold molar excess of either the S-Oligo (2 μg/μl) or NS-Oligo (2 μg/μl) (Table 2.2) was added to the DNA cocktails (before incubation with the relevant protein), as appropriate. Four percent (w/v) (29:1 acrylamide (Amersham) to bisacrylamide (Merck)) non-denaturing polyacrylamide gels (22 cm x 18.5 cm x 0.15 cm) were prepared in 1 x TGE buffer (50 mM Tris-HCl (pH 8.4), 380 mM electrophoresis grade glycine (Merck), 2 mM EDTA) as described by Sambrook *et al.* (1989). Pre-electrophoresis was for 2 h at 4°C at a constant voltage of 100 V. 1 X TGE was used as running buffer. Before loading the incubation mixtures into the wells of the gel, fresh buffer was added. Electrophoresis was for twelve to fourteen hours at 80 to 90 V (4°C). Gels were dried and exposed to X-ray film (Hyperfilm™ MP Amersham) with an intensifying screen at -70°C.

To test for the possible requirement of divalent cations for binding of hORFX to the suGF1 binding site, some reaction cocktails containing the *in vitro* expressed hORFX were titrated against a concentration range (0 μM – 500 μM) of a freshly prepared $ZnCl_2$ (Saarchem) solution.

2.9 Sodium-Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

A 12.5% (w/v) (30:0.5 acrylamide (Amersham):bisacrylamide (Merck)) polyacrylamide gel (7 cm x 6 cm x 0.1cm) was prepared using a BioRad Protean® II xi gel apparatus, as described in the literature (Sambrook *et al.,* 1989). The resolving gels were prepared in 1.5 M Tris (pH 8.8) and 10% (w/v) SDS solution, while stacking gels were prepared in 1.0 M Tris (pH 6.8) and 10% SDS (w/v) solution. Samples were boiled for 3 minutes in 1 x SDS sample application buffer (0.0625 M Tris-HCl (pH6.8), 2% (w/v) SDS (Sigma), 10% (v/v) glycerol, 5% (v/v) β-mercaptoethanol (BDH), 0.001% (w/v) bromophenolblue) and loaded directly onto the gel. Electrophoresis was performed at room temperature for 1 hr at a constant voltage of 120 V using a 1 X SDS solution (10% (w/v) SDS in H_2O; pH 7.2) as running buffer.

For electrophoresis of the radiolabeled IVT proteins, gels were stained for 2 hours in Coomassie staining solution (50 % (v/v) methanol (BDH), 10% (v/v) acetic acid (BDH), 0.1% (w/v) Coomassie R250 (Merck)). Gels were destained for a minimum of 1 h in destaining solution (25% (v/v) ethanol, 10% (v/v) acetic acid). Gels were subsequently dried and exposed to preflashed X-ray film with an intensifying screen at -70°C.

For the SDS-PAGE of yeast whole cell and nuclear extracts, gels were stained with silver salts, as described by Sambrook *et al.* (1995). The proteins were fixed in the gel matrix by incubating the gel for 12 hours at room temperature with gentle shaking, in at least 5 gel volumes of an ethanol:glacial acetic acid:water (30:10:60) solution. The fixing solution was discarded and the gel was incubated in at least 5 gel volumes, freshly prepared 30% (v/v) ethanol solution for 30 minutes at room temperature with gentle shaking. This step was repeated, after which the ethanol was discarded and the gel incubated in 10 gel volumes

of deionised water for 10 minutes at room temperature with gentle shaking. This rehydration step was repeated twice, after which the water was discarded and the gel was incubated in 5 gel volumes of a freshly prepared 0.1% (w/v) $AgNO_3$ (Merck; High Purity) solution for 30 minutes at room temperature with gentle shaking. The $AgNO_3$ solution was discarded and both sides of the gel were washed with water. The gel was soaked in 5 volumes of a freshly prepared 2.5% (w/v) sodium carbonate (Saarchem), 0.02% (v/v) formaldehyde (Merck) solution with gentle agitation, until the desired band contrast was obtained. The reaction was stopped by washing the gel in 1% (v/v) acetic acid for 5 minutes and soaking the gel several times in water for at least 10 minutes.

2.10 Recombinant Protein Expression

2.10.1 Growth and Preparation of Competent *S.cerevisiae* Cells using the Lithium Acetate Protocol

2.10.1.1 Competent Cells

For the preparation of yeast whole cell extracts, the lithium acetate transformation protocol was used to prepare competent FY23 yeast cells (Beggs, 1978).

Forty-eight hours before the transformation procedure, 5 ml YPD (1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) D-glucose) was inoculated with a single colony from the *S.cerevisiae* strain FY23 (MATa ura3-52 trp1Δ63 leu2Δ1 GAL2+) (Winston *et al.*, 1995). The culture was grown overnight to saturation at 30°C on a shaker platform. A 1-liter flask containing 300 ml YPAD (YPD medium supplemented with 30 mg/liter adenine hemisulfate (Sigma)) was subsequently inoculated with 2.5 ml of the saturated overnight culture, and

grown at 30°C to a cell density of 1 x 10^7 cells/ml ($OD_{600\ nm} \approx 0.3 - 0.5$). The cells were harvested by centrifugation (4000 X g at room temperature for 5 minutes) and resuspended in 10 ml autoclaved analytical water. The cells were transferred to a smaller tube and pelleted by centrifugation (6000 X g at room temperature for 5 minutes). The pellet was resuspended in 1.5 ml of a freshly prepared, buffered lithium solution (1 volume 10 x TE buffer pH 7.5, 1 volume 10 X lithium acetate (Sigma) stock solution (1 M lithium acetate, filter sterilised, pH 7.5) and 8 volumes sterile water) and incubated for 30 minutes at 30°C. The cells were either used immediately for transformation or stored at 4°C for a maximum of two weeks

2.10.1.2 Transformation using Lithium Acetate

Salmon sperm carrier DNA (200 μg) (Roche) was mixed with 5 μg supercoiled plasmid DNA (pYES2 or pYES2-suGF1) (Table 2.1), in a sterile microcentrifuge tube. Competent yeast cells (200 μl) were added to the DNA mixture. 1.2 ml freshly prepared PEG solution (8 volumes 50% (w/v) PEG (BDH), 1 volume 10 X TE buffer pH 7.5 and 1 volume 10 X lithium acetate stock solution (1 M, pH 7.5) was added to the tube and incubated for 30 minutes at 30°C. The cells were heat-shocked for exactly 15 minutes at 42°C and centrifuged for 5 seconds at room temperature in a microfuge. The transformed yeast cells were resuspended in 500 μl 1 X TE buffer (pH7.5), plated onto YPD CM –Ura (Bio101) selective dropout agar plates and incubated at 30°C for 3 days, until transformants were visible as single off-white colonies.

2.10.1.3 Preparation of Yeast FY23 Whole Cell Extracts

The Y-PER™ Yeast Protein Extraction Reagent kit (Pierce) was used to prepare whole cell extracts from the FY23 strain transformed with pYES2 or pYES2-suGF1. 500 ml YPG

(1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) D-galactose (Saarchem)) medium was inoculated with one colony of the transformed FY23 cells and grown for 48 hours to induce expression of the GAL1 promoter. Cells were pelleted by centrifugation (3000 x g at 4°C for 5 minutes). The pellet was resuspended in an appropriate volume (5 ml per 1g cell pellet) Y-PER™ solution (containing 10 μg/ml leupeptin (Roche) and 500 μg/ml PMSF) by gentle vortexing and up-and-down pipetting. The mixture was agitated on a shaker platform for 20 minutes at room temperature. After the lysis step the cell debris was collected by centrifugation (13 000 x g at 4°C for 10 minutes). The supernatant was carefully removed and dialysed for 12 hours against 100 volumes dialysis buffer (20 mM Tris-HCl pH 8, 2 mM $MgCl_2$, 0.2 mM EDTA, 20% (v/v) glycerol, 1 mM DTT and 0.5 mM PMSF). The dialysate was centrifuged (13 000 x g for 10 minutes at 4°C) and the supernatant was collected, aliquoted and quick-frozen in a methanol bath at -80°C. Aliquots of whole cell extracts were stored at -80°C and never thawed more than twice for usage in experimental procedures.

2.10.2 Growth and Maintenance of Y294 Yeast Cultures

2.10.2.1 Competent Cells

Competent yeast cells were prepared using a standard electroporation protocol as previously described (Chang *et al.,* 1991). The protease deficient *S.cerevisiae* strain Y294 (Matα leu2-3 leu2-112 ura3-52 his3Δtrp1 GAL+ [cir+]) (Lai *et al.*, 1997) was grown on standard YPD plates (1% (w/v) yeast extract, 2% (w/v) peptone 2% (w/v) D-glucose and 2% (w/v) Bacto-agar) at 30°C overnight. A single colony was inoculated in YPD (1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) D-glucose) grown at 30°C overnight with light shaking, until an optical density of 0.7 (≈ 1 x 10^8 cells / ml) was reached. Growth was

stopped by chilling the culture in an ice bath for 15 minutes. The culture was filtered through a Nalgene™ disposable filter without drying the cells. The cells were washed three times with two volumes ice-cold 1 M sorbitol (Sigma). They were pelleted by centrifugation (5000 x g for 5 minutes at 4°C). The supernatant was discarded and the cells were resuspended in the drops of media remaining on the side of the centrifuge tube. The competent cells were aliquoted into sterile micro-centrifuge tubes and were immediately used for electroporation.

2.10.2.2 Electroporation of Competent Y293 Yeast Cells

Electroporation of competent yeast cells was performed as previously described by Chang *et al.* (1989). Approximately 0.1–1 μg supercoiled plasmid DNA (pYES2 (Invitrogen) or pYES2-suGF1) (Table 2.1) was used for standard electroporation procedures. The DNA samples were pipetted into pre-chilled micro-centrifuge tubes and kept on ice. Competent Y294 cells (40 μl) were gently mixed with the DNA sample (final volume of 50 μl or less, depending on the amount of DNA added) and incubated on ice for 5 minutes. The mixture was transferred to a 0.2 cm electroporation cuvette and quickly placed into the cuvette chamber of a BioRad Micropulser which was set at Sc2. After pressing the pulse button once the cell suspension was removed and resuspended in 400 μl ice-cold 1 M sorbitol. Pulse parameters were recorded and were in the range of 1.0–1.5. The transformed cells were plated on YPD CM –Ura selective dropout agar plates containing 1 M sorbitol, and grown at 30°C for 48 to 72 hours.

2.10.2.3 Preparation of the Y294 nuclear extracts

Yeast nuclear extracts were prepared by the classic spheroplast-lysis method as described by Ausubel *et al.* (1995).

A 100 ml overnight culture of electroporated Y294 cells were grown to mid-log phase, with vigorous shaking (OD ≈ 1.8–2.2) in YPG (1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) D-galactose) to induce expression of the GAL1 promoter. The cells were harvested by centrifugation (1500 x g for 5 minutes at 4°C) in pre-weighed bottles. The wet weight of the cells was determined and taken as the packed cell volume in millimeters, which was considered to be equal to 1 volume. The cells were resuspended in two volumes ice-cold sterile water and pelleted by centrifugation (1500 X g for 5 minutes at 4°C). The supernatant was discarded, the pellet was resuspended in one volume zymolyase buffer 1 (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 M sorbitol and 30 mM DTT) and incubated for 15 minutes at room temperature. The cells were pelleted by centrifugation (1500 x g for 5 minutes at 4°C) and resuspended in three volumes zymolyase buffer 2 (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 M sorbitol and 1 mM DTT). 200 U zymolyase 100T (Seikagaku Corporation) per milliliter original packed cell volume were subsequently added to resuspended cells and incubated for two hours on a shaker platform at 50 rpm. The enzymatic conversion of the cells to spheroplasts was visually monitored by microscopy every 30 minutes. After full conversion the spheroplasts were pelleted by centrifugation (1500 X g for 5 minutes at 4°C) and the pellet was resuspended in YPD containing 1 M sorbitol. The spheroplasts were allowed to undergo metabolic recovery by incubation at 30°C for one hour. From this point onward all procedures were performed at 4°C. The spheroplasts were pelleted by centrifugation (1500 X g for 5 minutes at 4°C) after which the pellet was washed in two volumes ice-cold zymolyase buffer 2. The spheroplasts were harvested by centrifugation (1500 X g for 5 minutes at 4°C). This washing step was repeated three times. The pellet was resuspended in two

volumes ice-cold lysis buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgSO_4$, 1 mM EDTA, 10 mM potassium acetate, 1 mM DTT and 1 mM PMSF) by gently swirling the tube 10 to 20 times. The spheroplasts were pelleted by centrifugation (1500 X g for 10 minutes at 4°C) and resuspended in one volume ice-cold lysis buffer with a sterile glass rod. Extensive manipulation of the spheroplast pellet was avoided as this may cause premature osmotic lysis. The spheroplasts were lysed with 15 to 20 strokes of a sterile, Teflon-fitted (1 to 3 μm clearance) Dounce homogeniser. Ultracentrifuge tubes were half-filled with lysate and an equal volume of extraction buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgSO_4$, 1 mM EDTA, 10 mM potassium acetate, 1 mM DTT, 1 mM PMSF, 0.8 M ammonium sulfate (Saarchem) and 20% (v/v) glycerol) was added. The tubes were closed and gently inverted on a rotating wheel for 30 minutes at 4 °C. The samples were ultracentrifuged for 90 minutes at 100 000 X g (4°C). The supernatant was carefully removed and dialysed for 4 hours against 100 volumes dialysis buffer (20 mM Tris-HCl pH 8, 2 mM $MgCl_2$, 0.2 mM EDTA, 20% (v/v) glycerol, 1 mM DTT and 0.5 mM PMSF). The dialysis bag was transferred to 100 volumes fresh dialysis buffer and dialysed overnight. The dialysate was centrifuged (10 000 x g for 10 minutes at 4°C) and the supernatant collected, snap-frozen in small aliquots and stored at -80°C. The total protein concentration of the nuclear extracts ranged between 0.2 and 0.6 μg/μl. The nuclear extracts derived from Y294 cells were never thawed more than twice for experimental use.

2.11 Protein Determination

Protein concentrations were determined by the Bradford protein assay essentially as described by Bradford (1976) and modified by Zor and Selinger (1996). A dilution series of the unknown sample was prepared. 5 μl of each sample of this series was loaded in duplicate in wells of a microplate. 250 μl Bradford reagent (0.01% (w/v) Coomassie

Brilliant Blue G-250 (Roche), 4.7% (v/v) ethanol and 8.5% (w/v) phosphoric acid) was added to the samples. The microplate was incubated at room temperature for at least 2 minutes. The absorbance of each well were determined in a TiterTek™ microplate reader at 620 nm, within 1 hour after addition of the Bradford reagent. The protein concentration of the unknown samples was determined from a standard curve (0–2 mg protein/ml) obtained from a dilution series of BSA standards with pre-determined concentrations.

2.12 Circular Dichroism of Oligodeoxyribonucleotides

2.12.1 Instrumentation and Measurement

Oligodeoxyribonucleotides were analysed by circular dichroism (CD) spectroscopy, on a JASCO model J-810 spectropolarimeter (Department of Biochemistry, University of Cape Town). The special optical contributions of the cuvette and buffer were subtracted from original readings. Samples were subjected to circular polarised light at wavelengths from 220 nm to 320 nm. The absorption spectra for each sample was obtained as outlined in the instruction manual for the CD spectropolarimeter.

2.12.2 Sample Preparation

Single stranded oligodeoxyribonucleotides (Table 2.2 and Appendix 2) were synthesised and annealed as described in section 2.5. Gel-purified double stranded oligodeoxyribonucleotides were dissolved in sterile injection water (pH 7.0) at concentrations that were in the range of 1.5–2 μg/μl. Before measurement of the absorption values, each sample was filtered through a Watman™ 0.02 μm filter. The samples were loaded into a cylindrical quartz cuvette (1 mm pathlength) at a concentration

of 7.5 μM, which is the optimal amount required for absorption measurement (Hashizume and Imahori, 1967). Control samples (kindly provided by Ms. M.Chauhan, University of Cape Town) were also dissolved in sterile water (pH 7.0) and subjected to the same spectrophotometric analysis. The absorption values were analysed using the CDFIT program. Graphs were drawn using the Microsoft Office Excel program.

2.13 Sequence Analysis and Structure Prediction

The full-length suGF1 and hORFX amino acid sequences, as well as different fragments of these sequences (Appendix 3), were subjected to various methods of sequence analysis and structure prediction. The one-letter amino acid sequences were submitted to these programs.

A list of the tools that are commercially available on the Internet and were utilised for sequence analysis during this research project is given below:

123-D threading - http://www-lmmb.ncifcrf.gov/~nicka/123D.html

Align - http://vega.igh.cnrs.fr/bin/align-guess.cgi

Blast - http://www.ncbi.nlm.nih.gov/BLAST/

Blocks - http://www.blocks.fhcrc.org/

ClustalW - http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA

Entrez - http://www3.ncbi.nlm.nih.gov/htbin-post/Entrez/

Fasta - http://www.ebi.ac.uk/fasta/

FindMod - http://www.expasy.ch/tools/findmod/

Genbank - http://www.ncbi.nlm.nih.gov/Genbank

Local Alignments - http://vega.igh.cnrs.fr/bin/lalign-guess.cgi

Localisation Sites - http://psort.nibb.ac.jp/

Motif Searches - http://www.motif.genome.ad.jp/

OWL database - http://www.bioinf.man.ac.uk/dbbrowser/OWL/

Patterns - http://www.isrec.isb-sib.ch/software/PATFND_form.html

PEST - http://www.icnet.uk/LRITu/projects/pest/

pI, Mw etc - http://www.expasy.ch/tools/pi_tool.html

PredictProtein- http://www.embl-heidelberg.de/predictprotein/submit_adv.html

ProDom Blast - http://prodes.toulouse.inra.fr/prodom/doc/blast_form.html

Prosite Search - http://vega.igh.cnrs.fr/prosite/prosite-query.html

Protein Data Bank - http://www.rcsb.org/pdb/

Protein Sequence Viewer Swiss - http://www.pdb.bnl.gov/expasy/spdbv/mainpage.html

PSA Secondary Structure Prediction - http://bmerc-www.bu.edu/psa/index.html

Rasmol - http://www.umass.edu/microbio/rasmol/

SignalPeptide - http://www.cbs.dtu.dk/services/SignalP/

Threader - http://www.hgmp.mrc.ac.uk/Registered/Option/threader.html

Translate - http://www.expasy.ch/tools/dna.html

Weblab Viewer - http://www.accelrys.com/include/processdata.php

Yeast Protein Database - http://www.proteome.com/databases/index.html

Chapter 3 - Results

DNA-Binding Properties of Native and *in vitro* Transcribed-Translated suGF1

3.1 Rationale

The purification of suGF1 from sea urchin embryos is an extremely laborious process. Another source of the protein, which would be more readily available, was needed to further investigate the DNA-binding properties of suGF1 as well as to provide controls for the presence of a functional protein in future transactivation assays. Thus it was decided to investigate whether the *in vitro* expression of suGF1 in a mammalian reticulocyte lysate system would generate an suGF1 protein that exhibit the same DNA-binding properties as the native protein present in sea urchin embryo nuclear extracts.

3.2 *In vitro* Transcription and Translation of suGF1

The suGF1 protein was recombinantly expressed from the full-length suGF1 cDNA (pGEMT-suGF1) (Table 2.1), using a rabbit reticulocyte lysate transcription-translation kit. Transcriptional initiation was effected from the T7 promoter. A luciferase control plasmid, which was provided with the kit, was also expressed. To verify the presence and integrity of the IVT products, [^{35}S]-Methionine was added to the reaction mixture (incorporated during the translation of the protein). The end products were subsequently analysed by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and autoradiography (Fig. 3.1).

Fig. 3.1 shows the autoradiograph obtained from the SDS-PAGE analysis of the suGF1 IVT products. A negative control reaction, containing no plasmid DNA and thus producing no protein product (lane 2) and a positive control reaction generating a band of 61 kDa (relative to the marker bands), representing the full-length luciferase protein (lane 3), are shown.

To verify the estimated sizes of protein products obtained from the SDS-PAGE analysis, a graphical approach was taken. Figure 3.2 shows the graph of the natural logarithm of molecular weight versus the Rf-values of the respective standard molecular weight marker proteins. From this graph the exact molecular weights of the respective IVT protein products were determined. Table 3.1 summarises the results obtained for this experiment and gives the calculated molecular weight for each suGF1 protein product, as well as that for the positive control protein luciferase.

The SDS-PAGE analysis clearly verified the presence of a 58 kDa band representing the full-length suGF1 protein (lane 1). This is approximately the molecular weight predicted from the cDNA, if translation commenced from the first methionine in the open reading frame (Table 3.2). Several smaller bands exhibiting increased electrophoretic mobility, relative to the full-length protein, can be seen in the same lane (sizes indicated in the margin), suggesting that the processed suGF1 mRNA transcript might be translated from multiple AUG start sites (Table 3.1). To investigate whether the sizes of the suGF1 protein products obtained by *in vitro* transcription and translation of the cDNA are consistent with the utilisation of multiple AUG translation start sites, the full-length suGF1 cDNA sequence was subjected to *in silico* transcription and translation (theoretical strategy). Figure 3.3 presents the full-length cDNA sequence for the suGF1 protein, as well as the amino acid sequence. The suGF1 cDNA was transcribed and translated *in silico* generating a range of

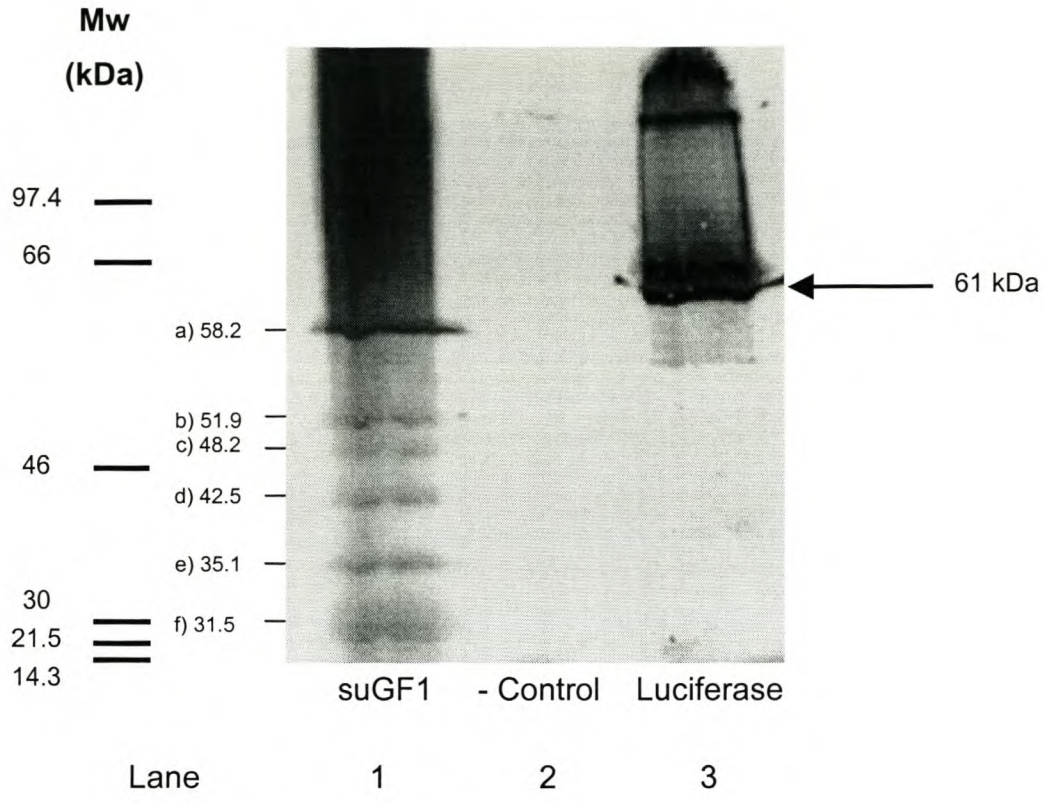


Fig 3.1 *In vitro* transcription and translation of the suGF1 and luciferase proteins.

A representative autoradiograph of the SDS-PAGE analysis of the suGF1 IVT products is shown. Lane 3 shows the results using a luciferase control expression construct as a positive control for protein expression. The negative control reaction (-C) performed in the absence of any plasmid DNA is shown in lane 2. A 58.2 kDa band representing the full-length suGF1 protein (band marked (a) in margin) is shown in lane 1. Five other bands constituting minor suGF1 protein products of 51.9 (b), 48.2 (c), 42.5 (d), 35.1 (e) and 31.5 (f) kDa respectively are shown in the same lane. An unlabeled, standard protein molecular weight (Mw) marker was also subjected to electrophoreses and was understandably not detected on the autoradiographic film. The position of the respective marker proteins was visualised directly on the gel and is indicated in the margin. This result was reproducible.

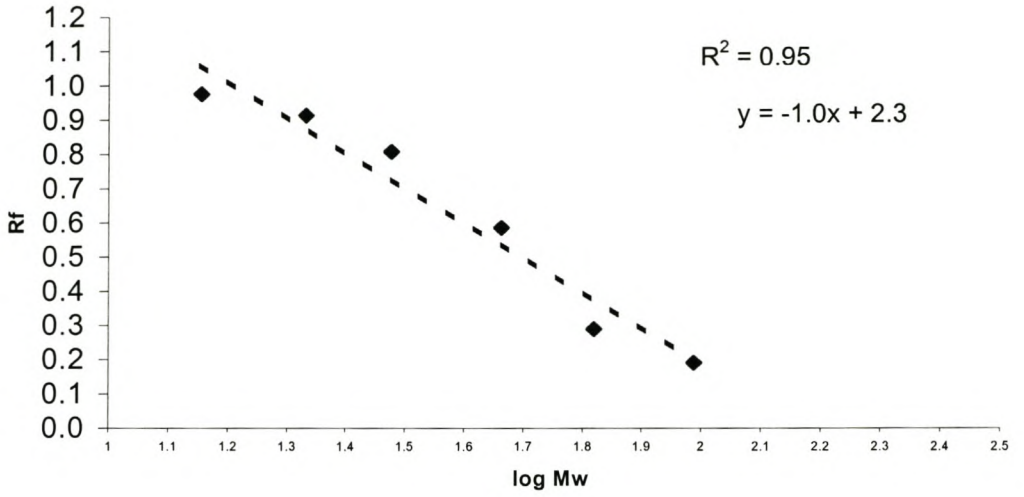


Fig. 3.2 Graph of the natural logarithms of molecular weight versus the Rf-values for the proteins present in the standard molecular weight marker mixture.

The natural logarithms of the molecular weights (log Mw) of the proteins in the rainbow marker are plotted against their corresponding Rf-values (obtained from the actual gel). The molecular weights of the IVT suGF1 protein products as well as the luciferase protein were determined from their respective Rf-values. A trendline (dashed line) (regression coefficient R^2 indicated) is shown as a mean of the polynomial function generated from the experimental data. The equation for the trendline (indicated) was used to calculate the log Mw (x-axis) of the respective suGF1 protein products.

Marker protein	Mw (kDa)	Log Mw	Rf
Phosphorylase b	98	2.0	0.2
BSA	66	1.8	0.3
Ovalbumine	46	1.7	0.6
Carbonic anhydrase	30	1.5	0.8
Trypsin inhibitor	22	1.3	0.9
Lysozyme	14	1.2	1.0
***In vitro* expressed proteins**	**Mw (kDa)**	**Log Mw**	**Rf**
Luciferase	61	1.8	0.4
suGF1 (a)	58	1.8	0.4
suGF1 (b)	52	1.7	0.5
suGF1 (c)	48	1.6	0.6
suGF1 (d)	42	1.5	0.7
suGF1 (e)	35	1.4	0.8
suGF1 (f)	32	1.4	0.9

Table 3.1 Summary of the molecular masses of the *in vitro* transcribed and translated suGF1 protein products as determined using a graphical approach.

The table summarises the results for the graphical determination of molecular mass for the IVT protein products exhibited on the autoradiograph after SDS-PAGE analysis (Fig. 3.1). The masses for the positive control reaction (luciferase) and the standard marker proteins are also given as determined from the graph in Fig. 3.2.

```
      59 kDa
      suGF1(a)➟
suGF1cDNA : ATGTCCACTCTGCCCCAGCCCCTGTCCCATTGCCTGCTGAACCAGGTGAA :  350
             M  S  T  L  P  Q  P  L  S  H  C  L  L  N  Q  V  N

suGF1cDNA : CACTGCAGCCATCAACCTACCACATCAACAACCTGGACTCATCACAGACA :  400
             T  A  A  I  N  L  P  H  Q  Q  P  G  L  I  T  D
             53 kDa
             suGF1(b)➟
suGF1cDNA : TCAAACCAATGATTAGTAACAAACCCCCTCCTACTCAGGAGGTCAAACCA :  450
             I  K  P  M  I  S  N  K  P  P  P  T  Q  E  V  K  P

suGF1cDNA : AACATCTTAGCTGCGGCTGCTGCTGGCTTGACCTACCCTCCACTCAACGT :  500
             N  I  L  A  A  A  A  A  G  L  T  Y  P  P  L  N  V
                                    48 kDa
                                    suGF1(c)➟
suGF1cDNA : GCCTAGCCTACCTGCAATGCCCAACGTGTCGATGCCTAATGTGTCATTGC :  550
             P  S  L  P  A  M  P  N  V  S  M  P  N  V  S  L

suGF1cDNA : CCAACGTGTCAATGCCTAATGTGTCTATGCCCAATGTGTCTATGCCAACC :  600
             P  N  V  S  M  P  N  V  S  M  P  N  V  S  M  P  T

suGF1cDNA : AGCGTTTCAATGCCGAGTGTGTCCATGCCCAGCGTTTCTATGCCGAGTGC :  650
             S  V  S  M  P  S  V  S  M  P  S  V  S  M  P  S  A

suGF1cDNA : GTCCATGCCAAGTGTTACTCTTCACAACCAACAGGGAAACAATAGCCAAC :  700
             S  M  P  S  V  T  L  H  N  Q  Q  G  N  N  S  Q
                                    42 kDa
                                    suGF1(d)➟
suGF1cDNA : TGAGCAACAGTAATTCTCAACGGCTGTCCCAAATGAAGAAATGCCCCAAT :  750
             L  S  N  S  N  S  Q  R  L  S  Q  M  K  K  C  P  N

suGF1cDNA : GAGTTTCTCCATCAGAATCCCCAAAGTGAGCGTCAGCTATTCTACAATGA :  800
             E  F  L  H  Q  N  P  Q  S  E  R  Q  N  L  F  Y  N
             39 kDa
             suGF1(e)➟
suGF1cDNA : TGTAGCCATGCAGCTGTATAACAGTGACTTCAACAAGTTTGCTTCCAAGA :  850
             V  A  M  Q  L  Y  N  S  D  F  N  K  F  A  S  K

suGF1cDNA : AGGAATTTCATGGCTACCTGTTAGAGCAGCAGAAGTGGAGATGGGATACC :  900
             K  E  F  H  G  Y  L  L  E  Q  Q  K  W  R  W  D  T

suGF1cDNA : CACAGCTACATAGGTAACCTGGAGACCAGAGTCCATAACTTGCTCATCAA :  950
             H  S  Y  I  G  N  L  E  T  R  V  H  N  L  L  I  N

suGF1cDNA : TCCAAACAGTGGGGTTGCCCAAAACGTTGCTCGATATCGCAGCGTCCCAA : 1000
             P  N  S  G  V  A  Q  N  V  A  R  Y  R  S  V  P

suGF1cDNA : TCAAATGTAAAAGCGAAGATGTGAAGCGATGTGAAGCCACGTCAAAGGAG : 1050
             I  K  C  K  S  E  D  V  K  R  C  E  A  T  S  K  E
                29 kDa
                suGF1(f)➟
suGF1cDNA : CTGGAGAATATGGCAACGCGTATTGCCAGTGTACGACAGCAGCTGCTGCA : 1100
             L  E  N  M  A  T  R  I  A  S  V  R  Q  Q  L  L  H
```

```
suGF1cDNA : CAAAAAGGGCACCTTGCTAACATCCAGCGATAATAGTGTCATAGTGTGGC : 1150
              K  K  G  T  L  L  T  S  S  D  N  S  V  I  V  W

suGF1cDNA : AGAATGAGCTAGCCTACATAGAACAGCTATTTGACAGAACGGATCAGATG : 1200
              Q  N  E  L  A  Y  I  E  Q  L  F  D  R  T  D  Q  M

suGF1cDNA : TACAACGAGGTCTTGTCCACACTTGCAAGTGTTAACCAAACTTTCTCCCA : 1250
              Y  N  E  V  L  S  T  L  A  S  V  N  Q  T  F  S  H

suGF1cDNA : CCTTCAGACTAGTTTCACTGCCGAAGCTGCAGAGCTGGCCGATCGGAGAC : 1300
              L  Q  T  S  F  T  A  E  A  A  E  L  A  D  R  R

suGF1cDNA : GCCTTTGGAGGCGGCGGAAGGAGAACAACCGAAAGAGACGGAAGCGCATG : 1350
              R  L  W  R  R  R  K  E  N  N  R  K  R  R  K  R  M

suGF1cDNA : GAGAAACAACTTGAAAAAATTGAGCAGCGATCTTGCGAGCTTCTCTTTCA : 1400
              E  K  Q  L  E  K  I  E  Q  R  S  C  E  L  L  F  H

suGF1cDNA : CATCACATCACGGGGGGCGTACGACAGGGTGCGTTCCCACCCTGAGATGC : 1450
              I  T  S  R  G  A  Y  D  R  V  R  S  H  P  E  M

suGF1cDNA : CTCGCATCGGACCCAGCGAGGTGAACACAGACATGTTAAATGGGATTAAA : 1500
              P  R  I  G  P  S  E  V  N  T  D  M  L  N  G  I  K

suGF1cDNA : TCCAAATCAGAAGTGAGGCCTCTAATGCATCTACTGAGTAAAGGTTACAT : 1550
              S  K  S  E  V  R  P  L  M  H  L  L  S  K  G  Y  M

suGF1cDNA : GACTCCAGGTGCGATGGAAATGGTCTCGCAAAAGATTCAGAAACTAGAAT : 1600
              T  P  G  A  M  E  M  V  S  Q  K  I  Q  K  L  E

suGF1cDNA : GTGGTATTAAGACTGAAGCTCACCAACAGGCAACCCAGGTCGGTATCAAC : 1650
              C  G  I  K  T  E  A  H  Q  Q  A  T  Q  V  G  I  N

suGF1cDNA : TCTCTGGCCATCAACAAAATGCCAGTTCCTGCTTCCAGAATTAAATCCAT : 1700
              S  L  A  I  N  K  M  P  V  P  A  S  R  N  K  S  I

suGF1cDNA : ACTGCCTCCTGCTCCTCCTCCAGTCACTGGCGTTGCCTCATCCACTATGA : 1750
              L  P  P  A  P  P  P  V  T  G  V  A  S  S  T  M

suGF1cDNA : TCTCATCAACCATGGTGTCGTCAGTAAACTCTGCTGCCCCTGTTACACAG : 1800
              I  S  S  T  M  V  S  S  V  N  S  A  A  P  V  T  Q

suGF1cDNA : CAATCAGTGCCCACCGTTAATCTCAATACTCAGCTAGCAAAG         : 1842
              Q  S  V  P  T  V  N  L  N  T  Q  L  A  K
```

Fig. 3.3 *In silico* translation of suGF1 versus IVT suGF1 (previous two pages).

The full-length suGF1 cDNA sequence (base numbers indicated in the margin are relative to the transcription start site), as well as the theoretically predicted amino acid sequence is shown. The ATG codons are underlined, whereas the ATG translational start sites predicted to initiate suGF1 expression are indicated as bold and underlined. The expected protein products, suGF1 (a) to suGF1 (f), with molecular weights of 59, 53, 48, 42, 39 and 29 kDa respectively are indicated, whereas the arrows denote the directional orientation of translation.

suGF1 Protein	Length (amino acids)	Molecular Mass from SDS-PAGE (kDa)	Theoretical Molecular Mass (kDa)
suGF1 (a) (full length)	514	58	59
suGF1 (b)	478	52	53
suGF1 (c)	437	48	48
suGF1 (d)	370	42	42
suGF1 (e)	345	35	39
suGF1 (f)	261	32	29

Table 3.2 A comparison between the molecular weights of the suGF1 IVT products determined after SDS-PAGE analysis and the *in silico* translation.

The table compares the molecular weights of the suGF1 IVT products obtained during SDS-PAGE analysis, to the theoretically determined values obtained from the sequence alone (*in silico* translation). Sequence analysis and calculations were performed using the Gene Tool / Pep Tool Lite (provided by DoubleTwist Inc. 2000) and Genedoc Version 2.5.000 (multiple sequence alignment editor and shading tool), molecular biology analysis programs.

protein products due to the presence of multiple AUG translation start sites (bold and underlined in the DNA sequence). The protein products generated due to utilisation of these sites are indicated as suGF1 (a) to suGF1 (f). suGF1 (a) represents the full-length protein (595 amino acids - 59 kDa), while suGF1 (b) to (f) represent truncated protein products. The results for this theoretical, molecular weight prediction strategy are summarised in Table 3.2. Results show that the molecular weights determined from SDS-PAGE analysis (graphical analysis) are similar to those predicted from theoretical analysis. This result is consistent with data obtained from the literature in which the authors found that the species homologue of suGF1 is expressed as five nested variants encoded from a single mRNA (Zeller *et al.*, 1995).

3.3 IVT suGF1 Produces Similar Protein-DNA Complexes to that of Native suGF1

To compare the DNA-binding properties of the IVT suGF1 to that of native suGF1 (present in sea urchin embryo nuclear extracts), the two different protein sources were subjected to electrophoretic mobility shift assays (EMSAs). Incubation of either source of suGF1 with a radiolabeled synthetic oligodeoxyribonucleotide (Fig. 3.4), containing a central G_{11}-string (part of the wild-type H1-H4 intergenic sequence), or a radiolabeled *EcoRI – HindIII* restriction digest fragment (Fig. 3.5) containing a $(GA)_{16}G_{11}$ sequence (part of the H1-H4 early histone gene battery of *P.miliaris*), resulted in multiple protein-DNA complexes, suggesting that both IVT suGF1 and native suGF1 can bind G-strings.

Figure 3.4 shows the autoradiograph for the EMSA analysis of IVT suGF1 (lanes 2 and 3) versus native suGF1 present in sea urchin embryo nuclear extracts (lane 4). When incubated with a radiolabeled, double-stranded oligodeoxyribonucleotide containing a central G-string, the IVT suGF1 exhibited multiple protein-DNA complexes (B1 – B5 in lane

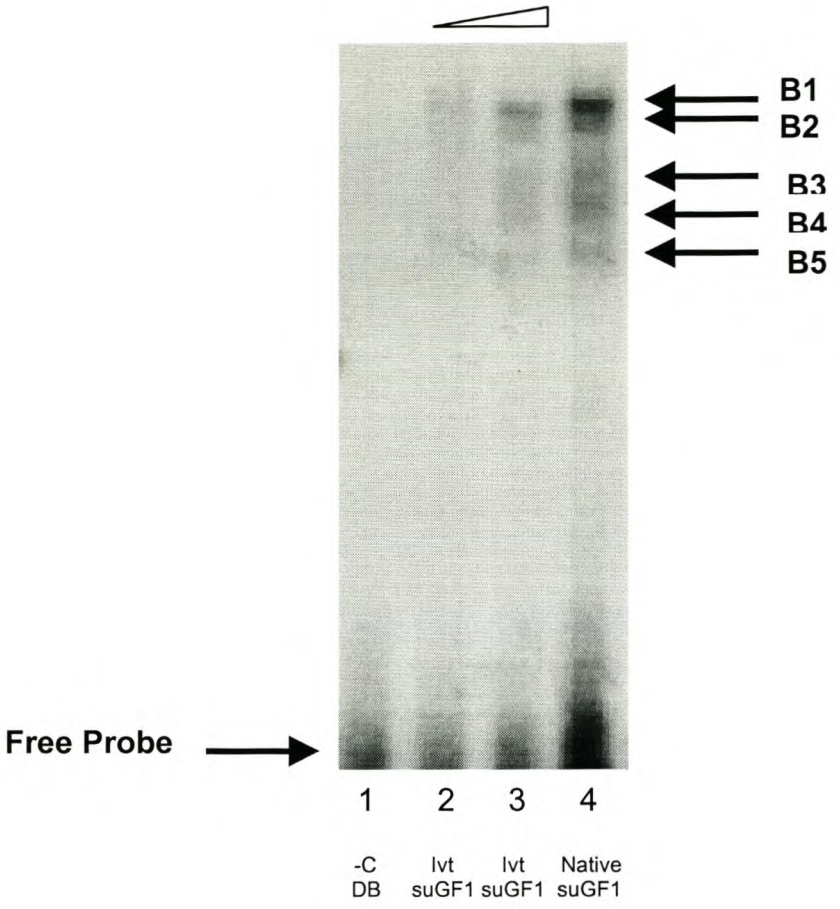


Fig 3.4 IVT and native suGF1 exhibit similar DNA-binding properties when incubated with a synthetic oligodeoxyribonucleotide containing a central G_{11}-string.

An autoradiograph is shown for the EMSA analysis of suGF1 when incubated with the S-Oligo as probe. Increasing amounts of IVT suGF1 (5 μl in lane 2 and 5 μl in lane 3) and native suGF1 (lane 4, sea urchin embryo nuclear extracts, 2.7 μg total protein) were incubated with the ^{32}P-labeled synthetic oligodeoxyribonucleotide and subjected to EMSA. The negative control (-C) reaction containing 5 μl of dialysis buffer (DB) is indicated in lane 1. The final reaction volumes (20 μl) remained unchanged when increasing amounts of IVT suGF1 was added. Arrows indicate free labeled DNA and the specific protein-DNA complexes (B1 to B5).

2 and 3) of reduced electrophoretic mobility relative to the free probe (indicated). Native suGF1, incubated with the same radiolabeled probe, exhibit a range of bands of reduced electrophoretic mobility representing multiple protein-DNA complexes (B1 – B5 in lane 4) of reduced electrophoretic mobility relative to the free probe. The negative control reaction (lane 1) containing the same DNA probe incubated with dialysis buffer produced no visible bands on the autoradiograph.

Figure 3.5 shows the result for a competition EMSA, using as radiolabeled probe, a 330 bp DNA fragment, in the absence and presence of unlabeled specific- or non-specific competitor DNA. The 330 bp radiolabeled *EcoRI – HindIII* fragment containing a $(GA)_{16}G_{11}$ sequence (part of the H1-H4 early histone gene battery of *P.miliaris*) has previously been shown to bind suGF1 specifically (Hapgood *et al.*, 1994). suGF1 present in sea urchin nuclear extracts (lane 2) and IVT suGF1 (lane 6) produced the same DNA-binding patterns. Both reaction cocktails produced multiple protein-DNA complexes (Fig 3.5, B1 – B4) of reduced electrophoretic mobility, relative to the free probe (indicated). The negative control reactions for this experiment, dialysis buffer (lane 1) and rabbit reticulocyte lysate (lane 5 and 9), incubated with the same probe, produced no specific complexes of decreased electrophoretic mobility. These results suggest that it is suGF1 binding to the radiolabeled probe and that both IVT suGF1 and native suGF1 have identical DNA-binding properties. To investigate the specificity of these protein-DNA interactions, a 100 fold molar excess of unlabeled, synthetic specific-oligo (S-Oligo) or non-specific-oligo (NS-Oligo) (Fig. 2.2) were added to the reaction mixtures. The native suGF1-DNA complexes present in lane 2 are completely competed away for by the addition of 100 fold molar excess unlabeled S-Oligo (lane 3), whereas addition of the same amount unlabeled NS-Oligo (lane 4) showed no competition. Similarly the IVT suGF1-DNA complexes present in lane 6, were also competed away for by the addition of 100 fold molar excess, unlabeled

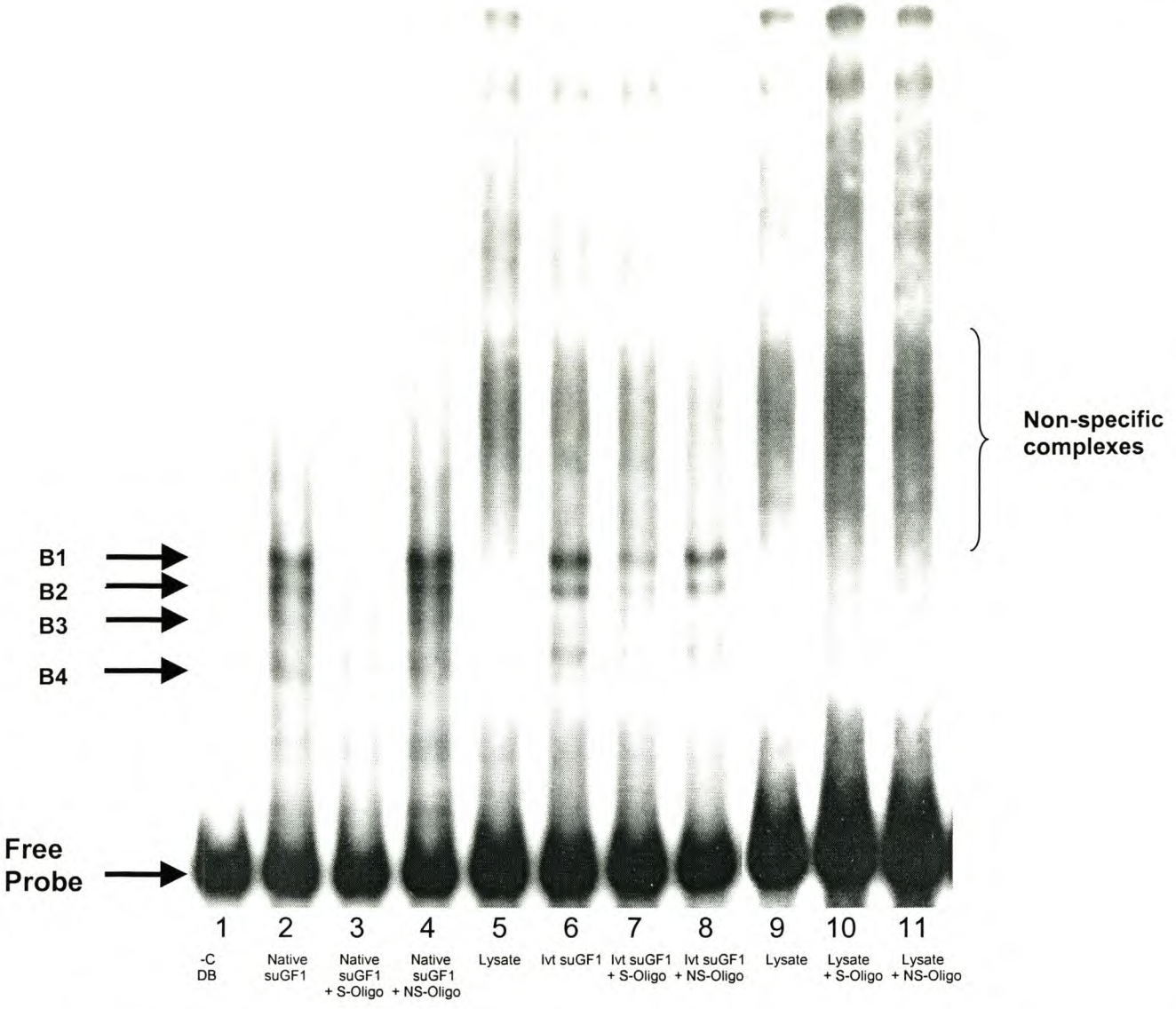


Fig 3.5 Competition electrophoretic mobility shift assays (EMSA) showed that IVT and native suGF1 bind specifically to GC-rich DNA.

An autoradiograph of the EMSA analysis of native and IVT suGF1, using a 330 bp DNA fragment as probe, is shown. EMSAs were performed with IVT suGF1 (10 µl in lanes 6 – 8) or native suGF1 present in sea urchin nuclear extracts (lanes 2 – 4; 2.7 µg total protein), in the presence of a ^{32}P-labeled *EcoRI – HindIII* restriction digest fragment. Complexes B1 to B4 in lanes 2 and 6 (native suGF1 and IVT suGF1 respectively) are competed away for by the addition of 100 fold molar excess unlabeled S-Oligo (lanes 3 and 7), while the addition of 100 fold molar excess non-specific competitor DNA caused no significant competition (lanes 4 and 8). Lanes 1 and 5 indicate negative control (-C) reactions containing 5 µl dialysis buffer (DB) and 10 µl reticulocyte lysate, respectively, plus the radiolabeled probe. Lanes 9 to 11 represent the same amount of reticulocyte lysate plus S-Oligo (lane 10) and NS-Oligo (lane 11) and are also negative control reactions. The final reaction volumes for all the incubations were 20 µl. Arrows indicate free-labeled DNA, non-specific complexes, as well as the specific suGF1-DNA complexes (B1 – B4).

S-Oligo (lane 7), whereas addition of the same amount, unlabeled NS-Oligo (lane 8) showed no competition for binding.

3.4 Circular Dichroism Analysis of Oligodeoxyribonucleotides

Homopurine.homopyrimidine stretches, including poly(dG).poly(dC) stretches (also referred to as G.C stretches or G-strings), have been shown to form unusual DNA structures, such as triple helices, *in vitro* (Kinniburgh *et al.*, 1994; Kohwi-Shigematsu and Kohwi, 1985; Maueler *et al.*, 1998; Musso *et al.*, 1998; Patterton and Von Holt, 1993). To investigate whether suGF1 binds to DNA exhibiting unusual conformations, the synthetic oligodeoxyribonucleotides (Table 2.2) were subjected to circular dichroism (CD).

The CD profile of a molecule is a characteristic that reflects asymmetric features of the physical, molecular structure. CD spectroscopy is a very useful technique for rapidly generating structural data, when only small amounts of material are available. CD spectra allow characterisation of, in particular, the secondary structure of proteins e.g. α-helices, β-sheets, β-turns etc., as well as obtaining information on the geometry of nucleic acids e.g. A-form, B-form, right- or left-handed helices etc. (Campbell and Dwek, 1984). In conjunction with the increased understanding of the three-dimensional structures of biomolecules, CD is a useful and informative technique.

A molecule that physically interacts differently with left- and right-circularly polarised light is said to be optically active (Johnson, 1988; Tinoco and Bustamante, 1980). Optical activity can be detected either as the differential change in velocity of two beams through a sample (optical rotatory dispersion), or as the differential absorption of each beam (CD). The latter generates a higher resolution and is therefore more commonly used. CD spectra

are characterised by ΔA (the differential absorption of the two beams) or θ_m (molar elipticity):

$$\Delta A = \Delta \varepsilon cL = A_L - A_R$$

where A_L and A_R represent the absorbance of left-and-right circularly polarised light beams respectively. $\Delta\varepsilon$ is the difference between the two extinction coefficients, *c* is the sample concentration in mol/L and L is the pathlength in centimeters. When passing the circular polarised light source through the sample, the two beams initially have equal amplitudes, generating a plane-polarised wave as the resultant beam. However when this light source passes through optically active material, the amplitude of the two circularly polarised beams differ drastically, generating a resultant beam that is elliptically polarised. This change in elipticity ($\Delta\varepsilon$ measured in mdeg) is a direct reflection on the asymmetric positioning of purine and pyrimidine bases in a DNA string. The sugar moieties and phosphodiester linkages also present in the DNA do not absorb light at these wavelengths (180 – 300 nm). The CD spectra therefore solely portray the mode of base stacking within the DNA molecule. Differences in base stacking result from many factors, including the sequence of the nucleotides, the geometry (A-form, B-form etc.), type of nucleic acid (RNA or DNA) and the number of nucleotide strands (i.e. single, double, triple helices etc.) (Campbell and Dwek, 1984). Usually a specific structural conformation corresponds to a unique CD spectrum, e.g. when DNA is in a right-handed conformation the CD spectrum exhibits a positive peak at longer wavelengths and a negative peak at shorter wavelength values (the reverse is true for left-handed DNA) (Fig. 3.6).

For the CD analysis of the double-stranded oligodeoxyribonucleotides (S-Oligo and NS-Oligo) used as probe for *in vitro* protein-DNA interaction analysis (EMSAs), all samples

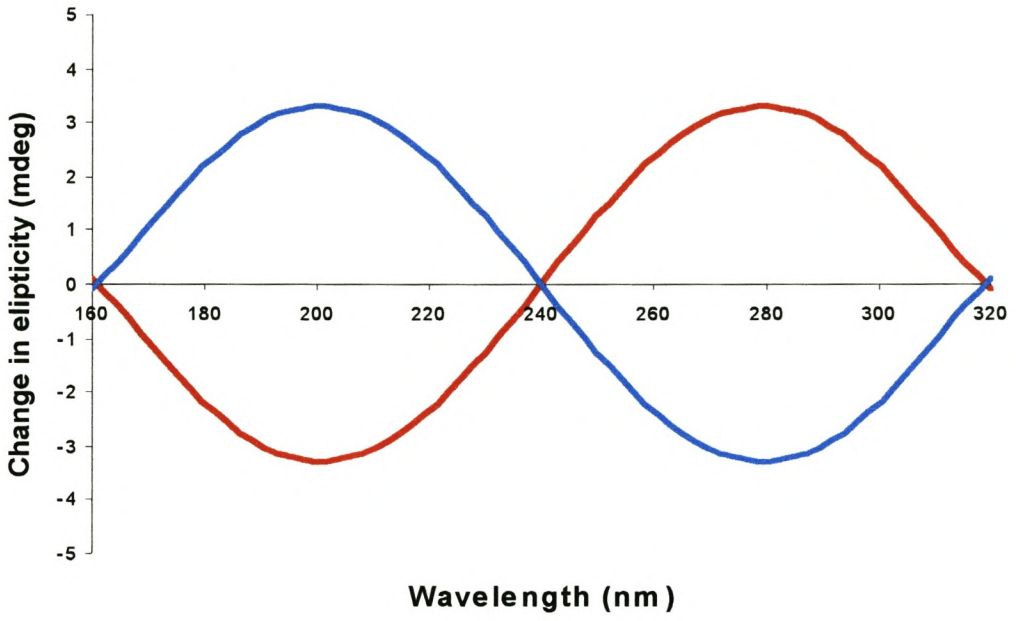


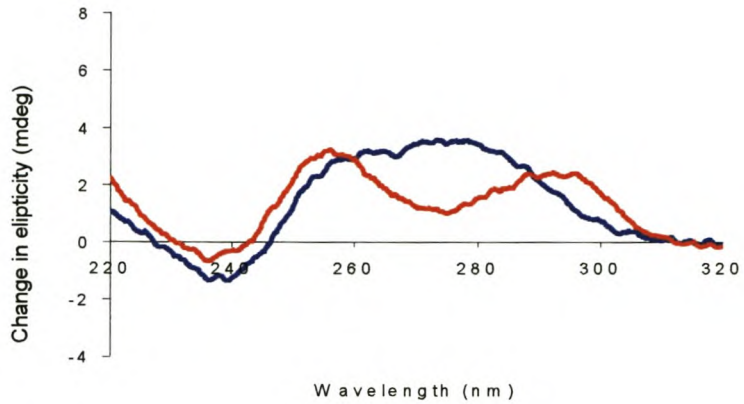
Figure 3.6 Left-handed and right-handed supercoiled DNA exhibit opposite changes in elipticity when subjected to circular polarised light.

The red line shows the CD spectrum of a typical right-handed nucleic acid and the blue line that of a typical left-handed nucleic acid (Zacharius *et al.*, 1982).

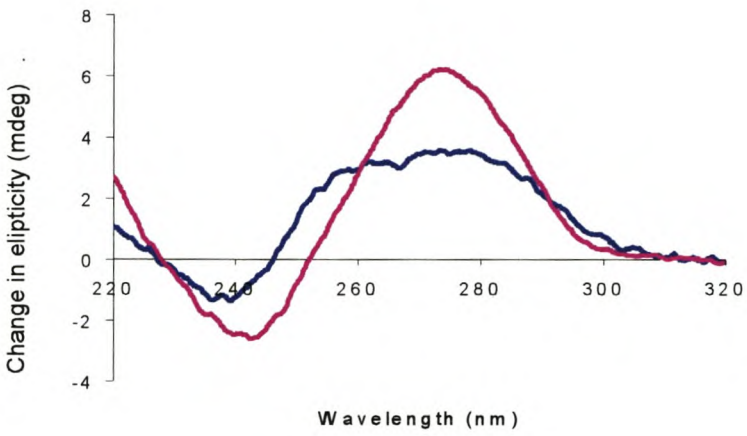
were dissolved in analytical water, because the DNA molecules analysed in the control studies were also dissolved in water. Unfortunately, samples could not be analysed in the presence of the buffer used in EMSAs, since it contains components (e.g. 175 mM KCl) that interfere with the movement of the polarised light through the medium. The DNA samples were not dissolved in physiological buffer either, due to the same reason.

The graphs for the CD analysis of the S-Oligo and NS-Oligo are shown in Fig's 3.7 and 3.8 respectively, whereas Table 3.3 summarises these results. The blue line in Fig. 3.7 represent the CD spectrum for the S-Oligo, whereas the red, purple and orange lines represent the CD spectra for the quadruplex, triplex and classical B-DNA control samples respectively. The spectrum for the NS-Oligo is indicated in Fig. 3.8 as a blue line and is compared to the same control samples. The spectrum for the S-Oligo exhibit a broad Cotton effect (maximum $\Delta\varepsilon \approx 3.56$ mdeg) at higher wavelengths (245 nm to 310 nm) and a smaller negative Cotton effect (maximum $\Delta\varepsilon \approx -1.38$ mdeg) at lower wavelength values (228 nm to 245 nm) (Neuberger and Van Deenen, 1985). The NS-Oligo CD spectrum exhibited a much smaller Cotton effect (maximum $\Delta\varepsilon \approx 3.1$ at 278 nm) from 260 nm to 300 nm, compared to that obtained for the S-Oligo and exhibited a broad, large negative Cotton effect (maximum $\Delta\varepsilon \approx -1.51$) at lower wavelength values of 230 nm to 260 nm. The quadruplex DNA sequence presented with a similar CD spectrum, however instead of having a single positive Cotton effect, this spectrum was characterised by two positive peaks (maximum $\Delta\varepsilon \approx 3.23$ mdeg) from 242 nm 310 nm, and a much smaller negative peak (maximum $\Delta\varepsilon \approx -0.68$ mdeg) from 232 nm to 242 nm (Fig. 3.7 (a)). A classical spectrum for the triplex control sample (Fig. 3.7 (b)) was obtained i.e. a large, sharp positive Cotton effect (maximum $\Delta\varepsilon \approx 6.2$) at higher wavelengths of 252 nm to 310 nm and a negative Cotton effect (maximum $\Delta\varepsilon \approx -2.64$ mdeg) from 228 nm to 253 nm, constituting approximately half the total area of the positive peak. The B-DNA control sample also

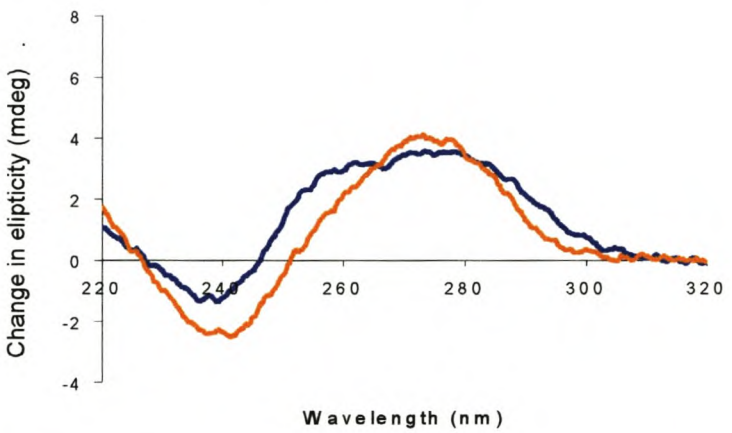
a)



b)



c)



S-Oligo sequence: 5′ **AGAGAGAGAGGGGGGGGGG**AGGGAGAATTGC 3′
3′ TCTCTCTCTCCCCCCCCCCTCCCTCTTAACG 5′

Quadruplex sequence: 5′ GGGGTTTTGGGG 3′

Triplex sequence: 5′ CCTCTCTTCCCTTCTTCTCTCC 3′
3′ GGAGAGAAGGGAAGAAGAGAGG 5′

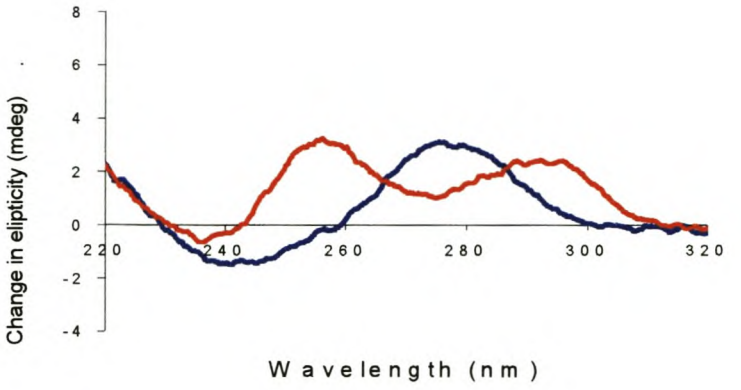
B-DNA sequence: 5′ GAAGAGAGG 3′
3′ CTTCTCTCC 5′

Figure 3.7 The CD spectrum of the double-stranded specific oligodeoxyribonucleotide (S-Oligo).

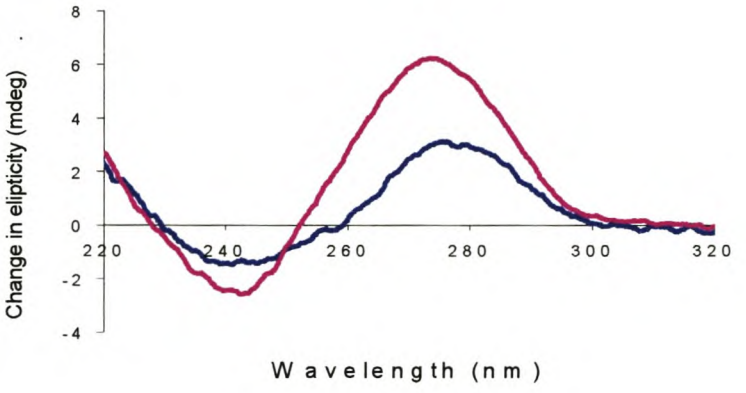
The CD spectrum of the S-Oligo (blue line in a, b and c) is interpreted relative to the CD spectra of the control samples that has been experimentally shown to exhibit quadruplex (red line in (a)), triplex (purple line in (b)) or classical B-DNA (orange line in (c)) conformations, respectively. The sequences of the double-stranded DNA samples subjected to CD analysis are given. This is the result of one experiment.

exhibited a profile reminiscent of previous spectra generated for B-DNA (Moore and Wagner, 1974). A broad, large positive Cotton effect (maximum $\Delta\varepsilon \approx 4.1$ mdeg) at higher wavelengths (252 nm to 305 nm), and a relatively large negative Cotton effect (maximum $\Delta\varepsilon \approx -2.53$ mdeg) at lower wavelengths (227 nm 252 nm), are the most prominent features of this spectrum. The results for the CD spectra are summarised in Table 3.3 and compare the total areas of the respective peaks, as well as maximum and minimum values for changes in ellipticity. All the samples seemed to exhibit a right-handed conformation, as positive peaks were observed for all samples at higher wavelength values and negative peaks at lower wavelength values. This is consistent with data from the literature. Note that when the reverse is observed the conformation has been experimentally proven to be left-handed (Zacharius et al., 1982).

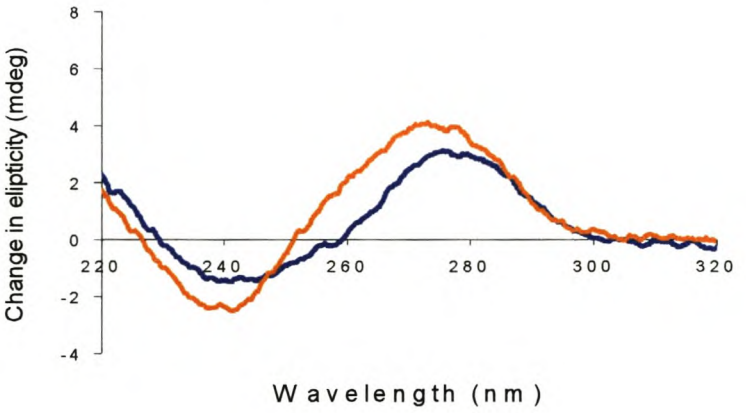
a)



b)



c)



NS-Oligo sequence:	5′ GATCTTCTGCACTCTCACCGGTACTGGACTGA 3′ 3′ CTAGAAGACGTGAGAGTGGCCATGACCTGACT 5′
Quadruplex sequence:	5′ GGGGTTTTGGGG 3′
Triplex sequence:	5′ CCTCTCTTCCCTTCTTCTCTCC 3′ 3′ GGAGAGAAGGGAAGAAGAGAGG 5′
B-DNA sequence:	5′ GAAGAGAGG 3′ 3′ CTTCTCTCC 5′

Figure 3.8 The CD spectrum of the double-stranded non-specific oligodeoxyribonucleotide (NS-Oligo).

The CD spectrum of the NS-Oligo (blue line in a, b and c) is shown relative to the CD spectra of the control samples that have been experimentally shown to exhibit quadruplex (red line in (a)), triplex (purple line in (b)) or classic B-DNA (orange line in (c)) conformations, respectively. The sequences of the DNA samples subjected to CD analysis are given. This is the result of one experiment.

Total Area of Peaks	**S-Oligo**	**NS-Oligo**	**Quadruplex**	**Triplex**	**B-DNA**
Positive Peak	1379.7	705.2	1178.1	1655.7	1099.2
Negative Peak	149.4	272.5	41.1	386.6	391.5
Ratio of Peak Areas	9.2	2.6	28.7	4.3	2.8
Maximum / Minimum	**S-Oligo**	**NS-Oligo**	**Quadruplex**	**Triplex**	**B-DNA**
Positive Peak (mdeg)	3.5	3.1	3.2	6.2	4.1
Negative Peak (mdeg)	-1.4	-1.5	-0.7	-2.6	-2.5
X-axis Intercepts	**S-Oligo**	**NS-Oligo**	**Quadruplex**	**Triplex**	**B-DNA**
Positive Peak (nm)	246; 310	259; 301	243; 312	252; 310	251; 303
Negative Peak (nm)	227; 246	229.5; 259	230; 243	228; 252	227; 251

Table 3.3 A summary of the results obtained for the CD analysis of the S-Oligo and NS-Oligo.

The relative areas of the negative and positive peaks constituting the CD spectra for the various samples that were subjected to CD analysis, the specific maximum and minimum peak values as well as the x-axis intercept values of the spectra, are shown relative to the control samples. The maximum and minimum peak values are given in mdeg, while the x-axis intercept values are given in nanometers.

Chapter 4 - Results

Searching for a Functional Homologue to suGF1

4.1 Introduction

An unprecedented wealth of data is being generated by genome sequencing projects and other experimental efforts to determine the structure and function of biological molecules. The demands and opportunities for interpreting these data are expanding more than ever. Rapid advances in technology and the ubiquity of the Internet offer unprecedented opportunities for scientists to gain access to, share, and analyse critical data and information stored in these databases. These vast stores of information have rich potential to accelerate scientific discovery and prevent costly duplication of experiments.

The currently exploding field of Bioinformatics therefore furnishes modern scientists with powerful computer-driven search and analysis tools. The following list summarises some of the tools and techniques which are by now commonly utilised in laboratories around the world to study DNA and protein sequences, as well as to facilitate the design of experimental strategies:

- Sequence analysis
  - database searches (e.g. Blast, Smith-Waterman)
  - alignments (e.g. CLUSTALW and ALIGN)
  - pattern and profile searches (e.g. ScanProsite and SMART)
  - motif searches and comparisons (e.g. MOTIF and DSMP)

- post-translational modification predictions - proteomics tool (e.g. NetPhos and NetOGlyc)
- exon / intron boundary estimations – genomics tool (e.g. Genquest)
- transcription factor binding site predictions - genomics tool (e.g. Genquest, TRANSFAC)
- Primer design and restriction enzyme mapping (Primer, Genekraal and DNAssist)

- Structure prediction
  - protein primary structure (e.g. REP, ProteinTranslator, ProtPlot)
  - protein secondary structure (e.g. PSA and nnPredict)
  - protein tertiary structure (e.g. SWISS-MODEL, 3D-PSSM, TopAlign and 123-D)
  - DNA structure (e.g. LOOP and CURVE)
- Special tools
  - identification and characterisation (e.g. FindPept and TagIdent)
  - DNA to protein translations and vice versa (e.g. Translate and MBS)
  - physicochemical properties (e.g. ProtParam and Compute pI /Mw)
  - transmembrane region detection - proteomics (e.g. DAS and TopPred)
- Phylogenetics (e.g. PAUP and PHYLIP)

The mere knowledge of a protein's sequence, or primary structure, does not allow a detailed understanding of its cellular purpose and relevance, even though the physiochemical properties of the macromolecule are a function of its monomers, the amino acids. However, when considering only the primary structure of a protein, the hydropathy patterns, iso-electrical points, determination of consensus domains and post-translational modification sites, are some of the useful parameters that can be predicted with reasonable accuracy by using Bioinformatics programs.

The way amino acids interact with their neighbours gives a protein its secondary structure.

The order and interactions between residues ultimately determine whether a specific amino acid sequence will assume a specific secondary structure e.g. α-helix, β-sheet or coiled region for example. The prediction of protein secondary structure is based purely on the sequence of the amino acid residues and the data generated will reflect the statistical probability that a specific residue will be in a specific structural state. Useful information regarding protein domains belonging to a certain class of protein structures can therefore be obtained e.g. whether a specific region conforms to a helix-turn-helix motif. In this case the secondary structure of the molecule would suggest a potential role in transcriptional regulation as many DNA-binding proteins involved in this process have this motif which ultimately establishes a specific functional purpose. Secondary structure can therefore conveniently be utilised as a yardstick for possible structural homologies between proteins, which can lead to a greater understanding of the function of an unknown protein.

The unique, well-defined, three-dimensional (3-D) structure of a protein dictates the way in which it performs its biological function. Knowing the 3-D structure of a protein allows researchers to gain insight into the active site of the protein or into the way it interacts with small molecules and other proteins. The generation of 3-D structures is therefore critical for a detailed understanding of biological processes at the molecular level. Although the determination of the complete genome sequences of various organisms is now customary practice, the experimental determination of the 3-D structures of the proteins encoded in these genomes is currently a very laborious process. With the aid of strong computing power, quantum mechanics and statistical expertise, protein models and simulations of their function can now be obtained albeit with a degree of uncertainty (especially as the

sequence length increases). The basic assumption is that the information the protein needs in order to fold into its unique 3-D structure lies entirely in its amino acid sequence. It is widely accepted that, for most proteins, the native 3-D structure of a protein has the lowest free energy possible for its combination of amino acids. Thus, in principle, predicting the unique 3-D structure of a protein given its amino acid sequence alone, may in future be an achievable goal.

4.2 Database Searches for an suGF1 Homologue

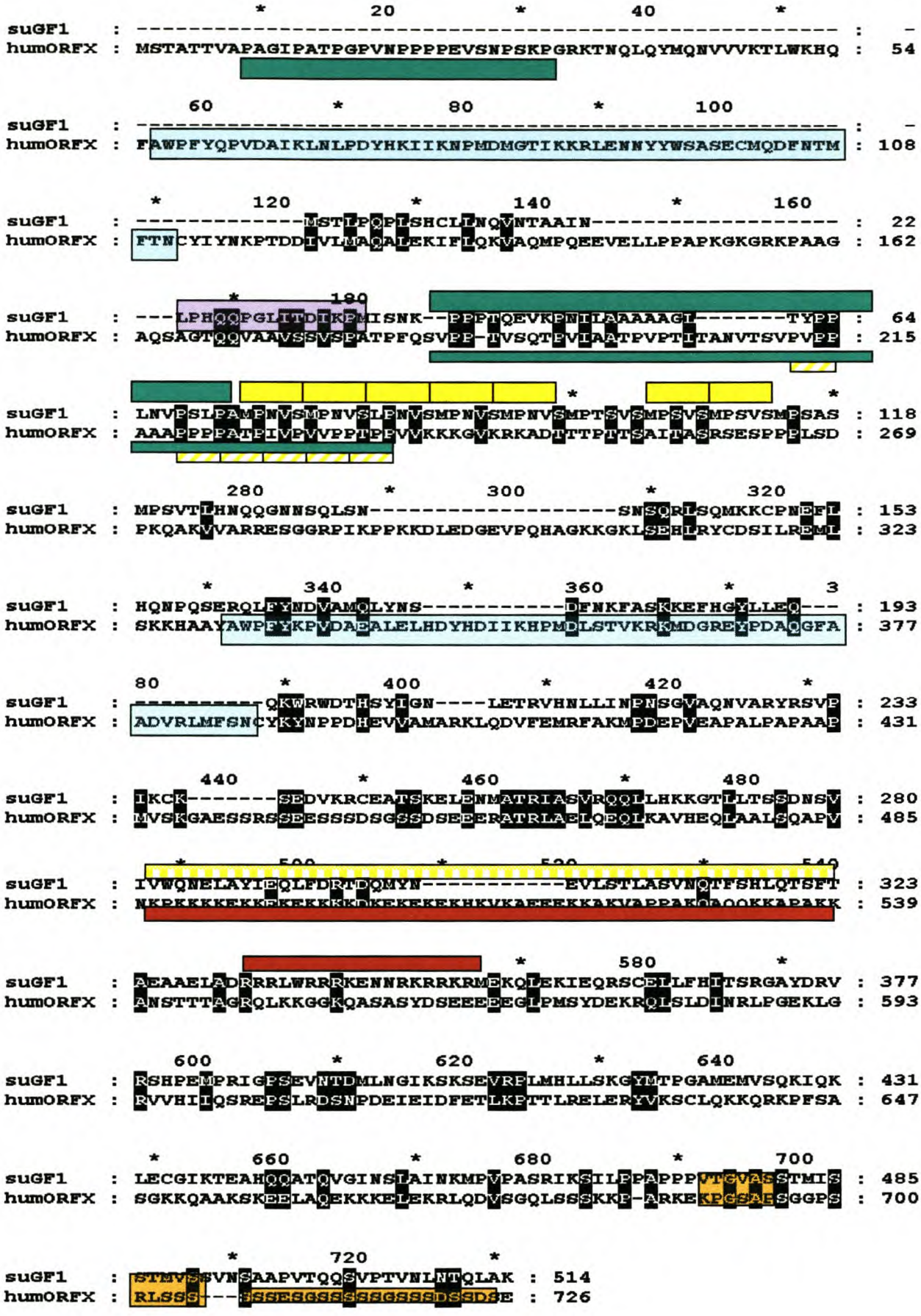
The search for an suGF1 homologue was complicated by the fact that many of the databases that are essential for a thorough sequence search, have not been completed or are currently still unavailable for commercial use. The full-length suGF1 cDNA and amino acid sequences were nonetheless submitted to exhaustive database searches as to identify a functional sequence homologue. The results obtained using different search and retrieval engines (e.g. Blast and Smith-Waterman) produced no significant sequence homology between suGF1 and any of the database entries. Using these search engines, databases for *S.cerevisiae, C.elegans*, *D.melanogaster*, *Mammalia* as well other general databases e.g. those for insects, birds, reptiles and fish were thoroughly searched. As these searches produced no positive results, various permutations and combinations of fragments (domains) of the suGF1 amino acid sequence were subjected to sequence search and analysis. This strategy facilitated the search for a more general, functional homologue to suGF1, as the homology would not necessarily be a function of the amino acid sequences alone, but might reside within the presence and distribution of specific functional domains. Indeed within the scope of this project sequence analysis tools (ProDom, Blast, Smith-Waterman and Genquest) were utilised to identify a putative, functional, human homologue to suGF1 called hORFX (Accession number D26362).

Prior to the present study, no information was available regarding the biochemical properties of the hORFX protein. Alignment of the two full-length amino acid sequences (suGF1 vs. hORFX) showed that they share only 15.9% global homology (Fig. 4.1). However, both proteins contain similar domain features i.e. an N-terminal proline-rich domain (a putative transactivation domain), hydrophobic amino acid repeats (putative dimerisation domains), a highly basic region (a putative DNA-binding domain) and a serine-rich C-terminus. Moreover, these domains are orientated in exactly the same order within the sequence (Fig. 4.1). This prompted an investigation into the DNA-binding properties of hORFX as well as a more detailed structure prediction analysis, with a view to determining whether hORFX is a functional homologue of suGF1.

4.3 *In vitro* Transcription and Translation of hORFX Produced Multiple Products

The hORFX protein was transcribed and translated *in vitro* from the full-length hORFX cDNA (pBluescript SK^+-hORFX) (Table 2.1), using the rabbit reticulocyte lysate transcription-translation system (Promega) as described in Section 2.7 of the materials and methods. The autoradiograph generated from SDS-PAGE analysis verified the existence of multiple protein products, of which the 73 and 80 kDa species appear to be the dominant proteins (Fig. 4.2, lane 2). This is consistent with the theoretical translation (*in silico*) of the hORFX cDNA that also produced multiple protein products, of which the 73 and 81 kDa transcripts are the species of highest molecular mass. The negative control reaction (lane 1) containing the IVT reaction cocktail (without any plasmid DNA) produced no bands on the autoradiograph. The bands in lane 2 are therefore specific for the reaction cocktail in which the hORFX cDNA was present. A rainbow marker is indicated in the margin.

a)



b)

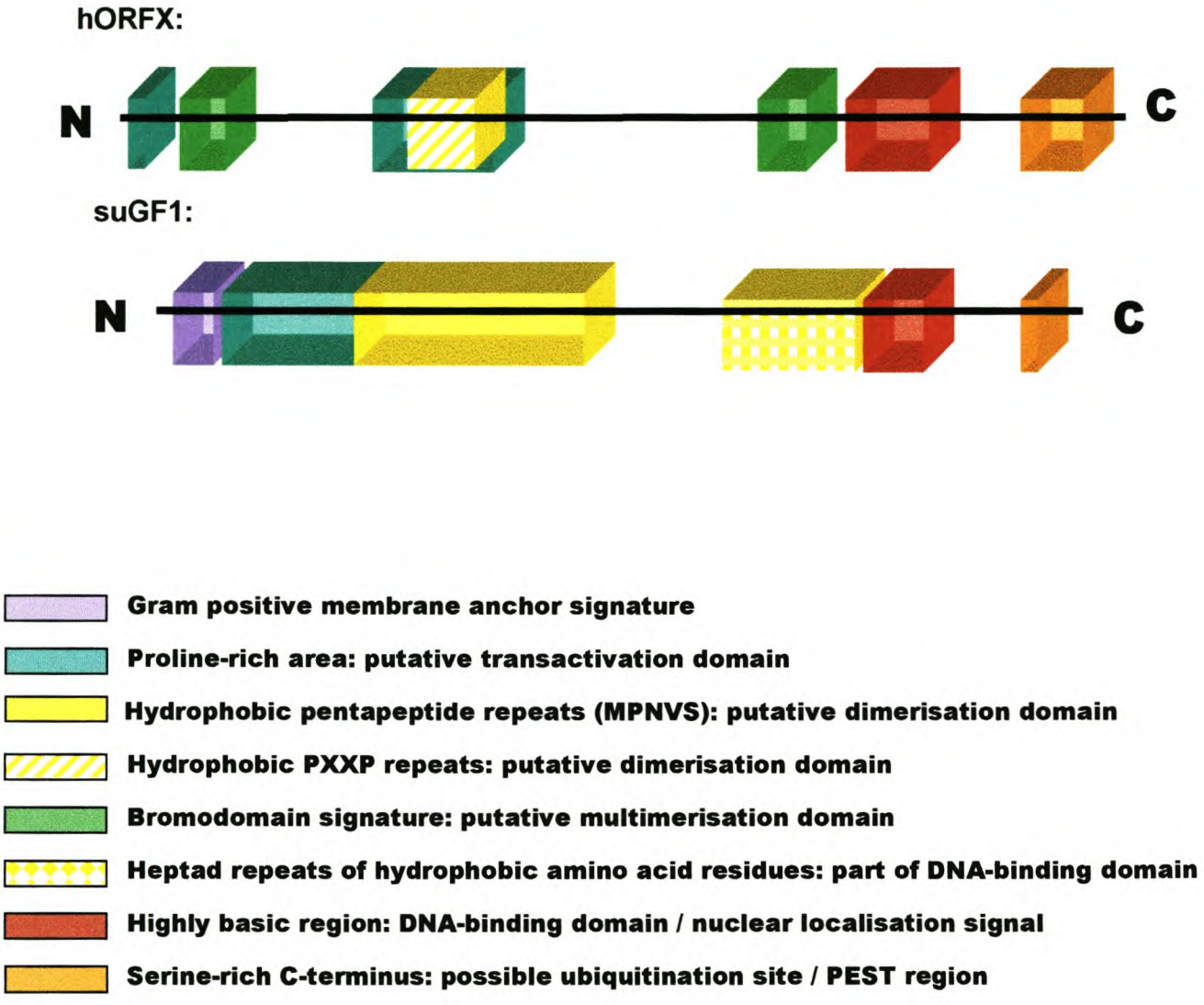


Fig. 4.1 **Sequence alignment of the suGF1 and hORFX amino acid sequences.**

The sequence alignment of the full-length suGF1 and hORFX amino acid sequences is shown in (a). The coloured boxes indicate specific domains whereas the aligned residues are indicated in black. The schematic diagram in (b) depicts the relative distribution of the domains present in the full-length suGF1 and hORFX amino acid sequences. The diagram is drawn approximately to scale. The names of the various domains as indicated in (a) and (b) are given with respect to a specific box colour.

4.4 IVT hORFX Does Not Exhibit Similar DNA-Binding Properties to suGF1

To investigate whether hORFX can recognise and interact with G-strings, EMSAs were performed using IVT hORFX as protein source.

The autoradiograph shown in Fig. 4.3 clearly indicates that hORFX (lanes 5 (5 μl), 6 (10 μl) and 7 (15 μl)) does not bind the specific, synthetic oligodeoxyribonucleotide (containing a G_{11}-string), as no bands of decreased electrophoretic mobility, relative to the free probe, can be observed. Increasing the amount of hORFX that is added to the reaction cocktail (lanes 5 – 7) also did not result in the formation of a specific protein-DNA complex. The positive control reactions, containing IVT suGF1 (lane 2 (5 μl) and 3 (10 μl)) and native suGF1 (present in sea urchin nuclear extracts) (lane 4), incubated with the same probe, produced multiple bands (indicated with the arrow) representing the characteristic suGF1-DNA complexes of reduced electrophoretic mobility. These complexes have previously been shown to be specific (Fig. 3.5). The negative control reaction (dialysis buffer incubated with the probe) produced no bands of reduced electrophoretic mobility (lane 1), relative to the free probe.

The result presented in Fig. 4.4 is consistent with the results shown in Fig. 4.3, suggesting that hORFX does not specifically recognise the G-string sequence *in vitro*. Incubation of IVT hORFX with the radiolabeled *EcoRI* - *HindIII* fragment, produced one band of reduced electrophoretic mobility (lane 3). This band was however not substantially competed away for by the addition of 100-fold molar excess of specific (S-Oligo) competitor DNA (lane 4). However, this protein-DNA complex is significantly competed away for by the addition of 100-fold molar excess of non-specific competitor DNA (NS-Oligo) (lane 5), showing that the complex is not specific for G-strings. The positive control reaction (lane 1), containing

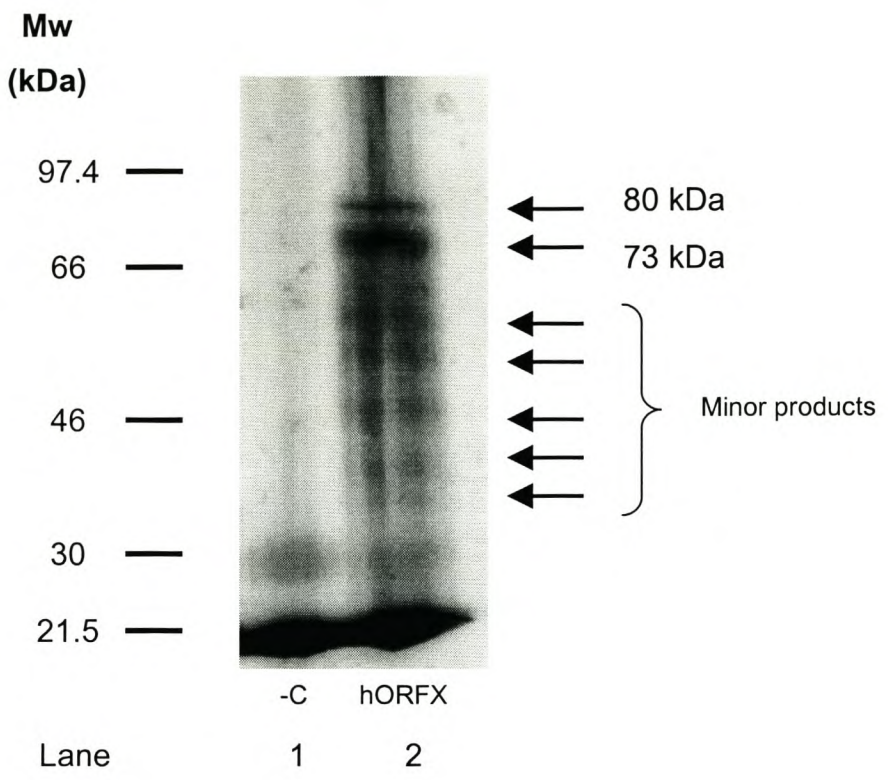


Fig 4.2 *In vitro* transcription and translation of an 80 kDa hORFX protein.

An autoradiograph of the SDS-PAGE analysis of the hORFX IVT products is shown. A negative control (-C) reaction was performed in the absence of any plasmid DNA (lane 1). Two main bands (lane 2), representing the full-length hORFX proteins of 80 and 73 kDa respectively, are indicated with arrows. Five other smaller protein products can be observed in the same lane and are indicated with arrows (sizes not given). The sizes of the proteins present in the rainbow marker are indicated in the margin.

native suGF1 (present in sea urchin embryo nuclear extracts) incubated with the same probe, produced the same classical suGF1-DNA complexes as before (Fig. 3.4 and 3.5). The negative control reaction (lane 2), containing only lysate from the rabbit reticulocyte lysate system, incubated with the same probe, produced no slower migrating bands.

To investigate whether the presence of divalent cations is a pre-requisite for the binding of hORFX to the suGF1-binding element, increasing concentrations of a $ZnCl_2$ (commonly required for binding of proteins to DNA (Bossone *et al.,* 1992) solution was added to the reaction cocktails and subjected to EMSA. Fig. 4.5 shows an autoradiograph of this EMSA, which clearly shows that the addition of Zn^{2+}-ions to the incubation mixture has no effect on the ability of hORFX to bind the probe to which suGF1 binds specifically in the absence of divalent cations. The positive control reaction containing native suGF1 produced the characteristic, specific suGF1-DNA complexes of reduced electrophoretic mobility, while lane 2 (the negative control reaction constituting lysate from the rabbit reticulocyte lysate system) produced no bands. Lanes 3 to 8 contain 5 μl of the hORFX *in vitro* transcription-translation products, while lanes 9 to 14 contain double this amount. The reactions were incubated in the presence of increasing concentrations (0 – 500 μM) of $ZnCl_2$, as indicated. Similar results to that obtained in Fig. 4.4 were observed i.e. no specific bands of reduced electrophoretic mobility, suggesting that no protein-DNA complexes were formed. Non-specific complexes at double the amount of recombinant protein were again observed (lanes 9 to 14), but these have previously been shown to be non-specific (Fig. 4.4). The relative amount of free probe, remaining after electrophoresis of the protein-DNA incubation mixture, is an indication of occupied probe. It is therefore interesting to observe that at the 5 μl amount protein added in the presence of $ZnCl_2$ (lanes 4 – 8), the amount of free probe remaining is relatively low relative to the control

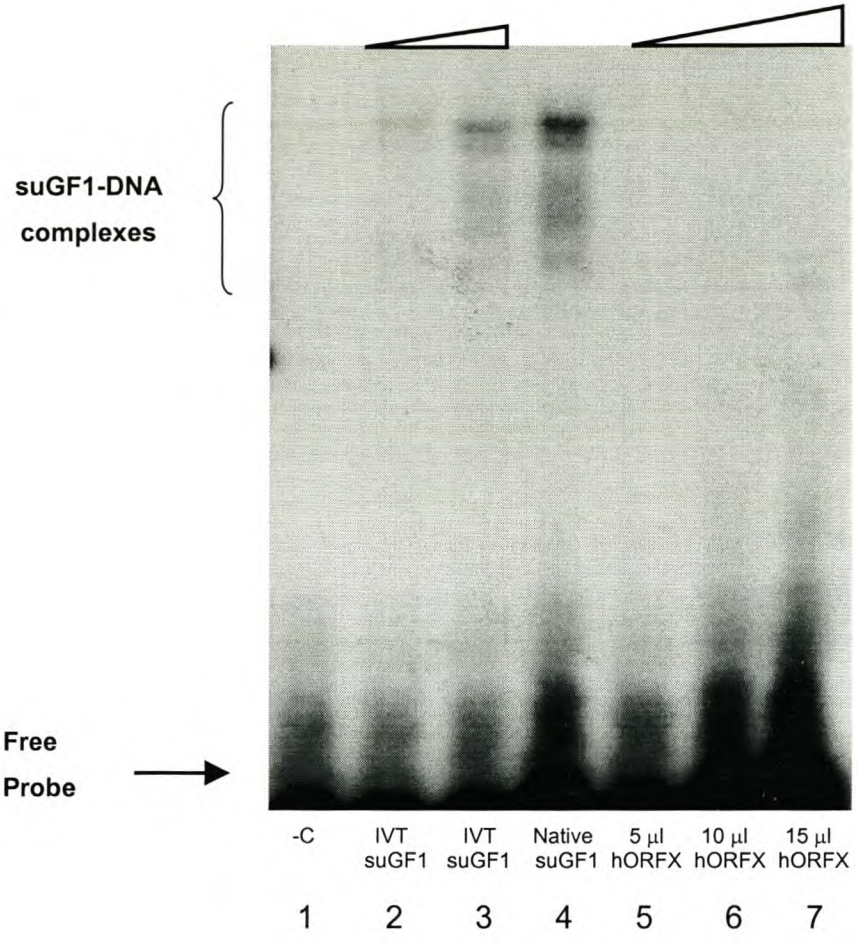


Fig. 4.3 EMSAs showed that IVT hORFX does not exhibit similar DNA-binding properties to suGF1.

An autoradiograph of the EMSA analysis of IVT hORFX when incubated with the radiolabeled S-Oligo. The negative control reaction (-C), containing radiolabeled probe incubated with dialysis buffer, produced no bands of reduced electrophoretic mobility and is shown in lane 1. The positive control reactions containing IVT suGF1 (lanes 2 (5 µl) and 3 (10 µl)), as well as native suGF1 (present in sea urchin embryo nuclear extracts – 2.7 µg total protein) (lane 4) produced characteristic bands of reduced electrophoretic mobility as indicated with the bracket. When IVT hORFX was incubated with the same probe (lanes 5 (5 µl), 6 (10µl) and 7 (15 µl)) no bands of reduced electrophoretic mobility (relative to the free probe) can be observed on the autoradiograph. The final reaction volume (containing different amounts of IVT hORFX products) for all reactions was 30 µl. The free, labelled DNA probe is indicated with an arrow.

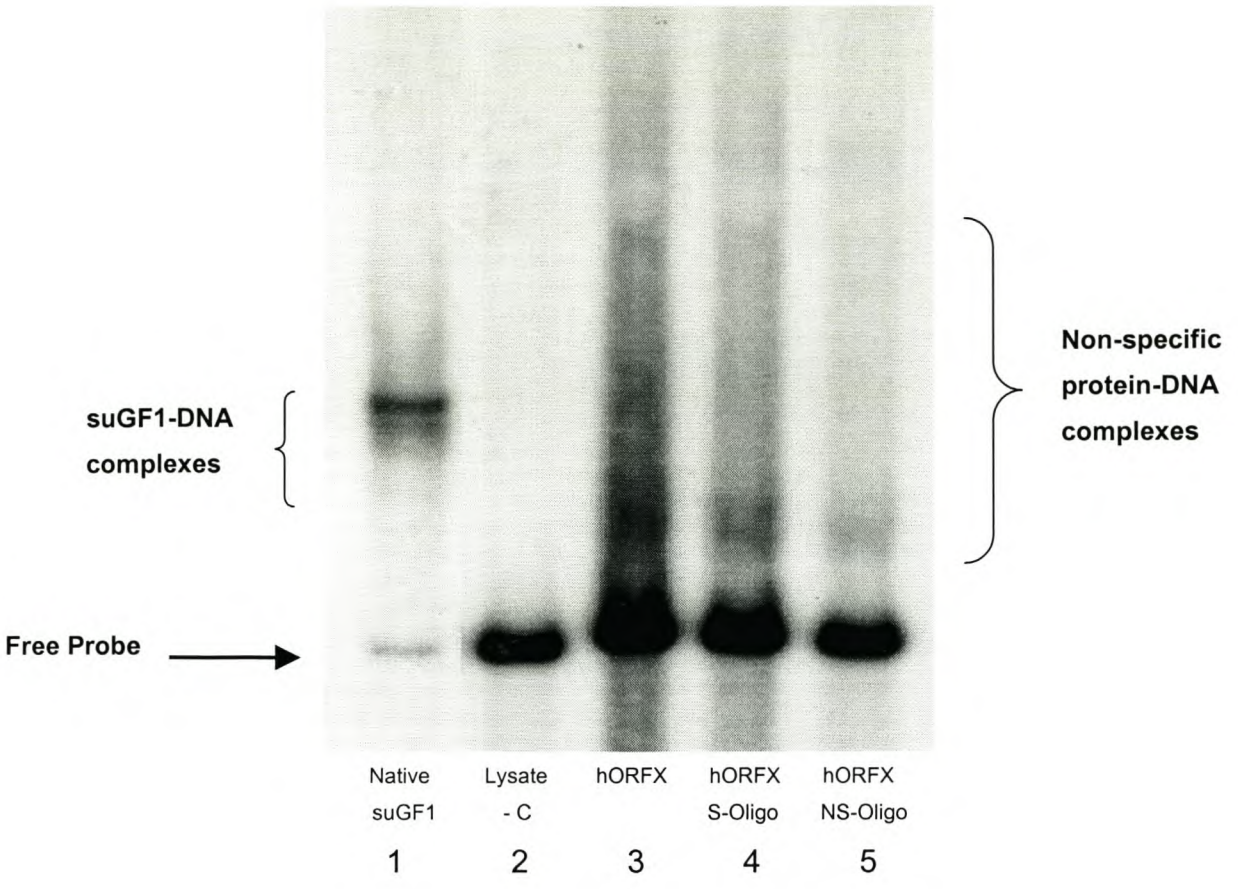


Fig. 4.4 Competitive EMSAs verified the inability of hORFX to recognise G-strings.

An autoradiograph of the EMSA analysis of IVT hORFX when incubated with a 330 bp radiolabeled *EcoRI* - *HindIII* fragment is shown. Lane 1 shows the positive control reaction, which exhibit multiple bands representing suGF1-DNA complexes (indicated with a bracket). The negative control reaction (-C), containing lysate from the reticulocyte lysate system produced no bands of reduced electrophoretic mobility (lane 2). Slower migrating complexes (indicated with a bracket) are observed when IVT hORFX was added to the reaction cocktail (lane 3), suggesting the formation of a protein-DNA-complex. This complex exhibited in lane 3 is however not competed away for by the addition of 100 fold molar excess specific oligo (S-Oligo) (lane 4) or non-specific oligo (NS-Oligo) (lane 5). The free, radiolabeled probe is indicated with an arrow.

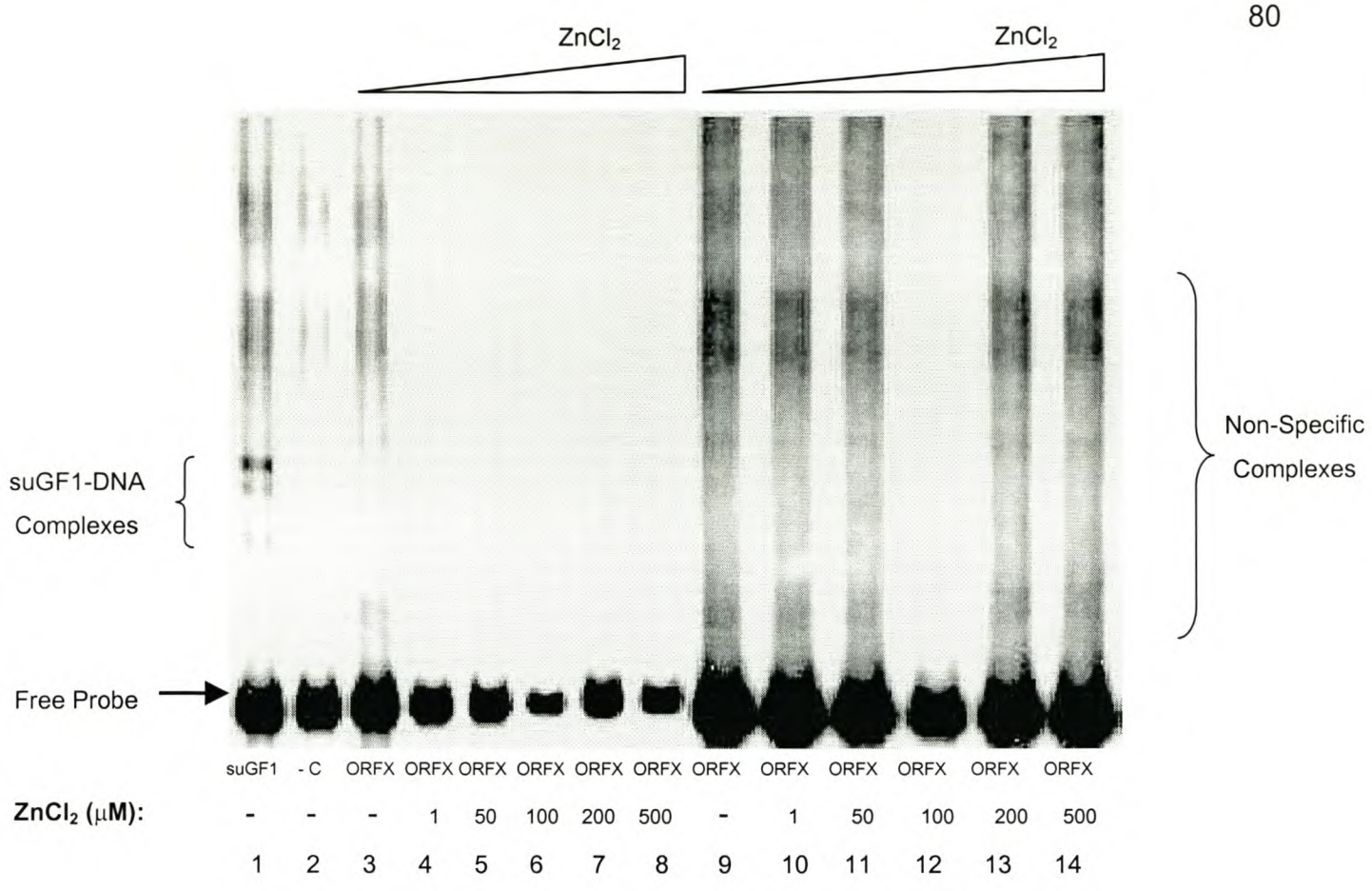


Fig. 4.5 hORFX does not bind to G-strings in the presence of $ZnCl_2$.

An autoradiograph of the EMSA analysis of IVT hORFX (ORFX) when incubated with radiolabeled *EcoRI-HindIII* fragment, in the presence of increasing concentrations of $ZnCl_2$ (0 – 500 μM as indicated), is shown. Lanes 3 to 8 represent reaction mixtures containing 5 μl hORFX IVT products, while lanes 9 to 14 contain double this amount (10 μl). Protein-DNA complexes at 10 μl hORFX are observed in lanes 9 to 14, but these have previously been shown to be non-specific (Fig.4.4). The positive control reaction, i.e. the reaction cocktail in the presence of native suGF1, produced multiple retarded bands that are indicated with a bracket. The negative control reaction (-C) containing lysate from the reticulocyte lysate system produced no bands of reduced electrophoretic mobility (lane 2). The free probe is indicated with an arrow.

reaction (lanes 1 and 2) and the reactions containing 10 μl of IVT hORFX products (lanes 9 – 14). This observation will be discussed in the conclusion to this thesis.

4.5 Primary Structure Analysis

The experimental data presented suggest that hORFX is not a functional homologue of suGF1 because it does not recognise and bind to the same DNA probe *in vitro*. The highly basic domains present in both proteins initially suggested that the putative DNA-binding domains of suGF1 and hORFX might be similar. The results from EMSAs (see Fig. 4.3 and 4.4), however, suggest significant differences between these domains despite the similarity in sequence composition and physicochemical character of their respective basic domains.

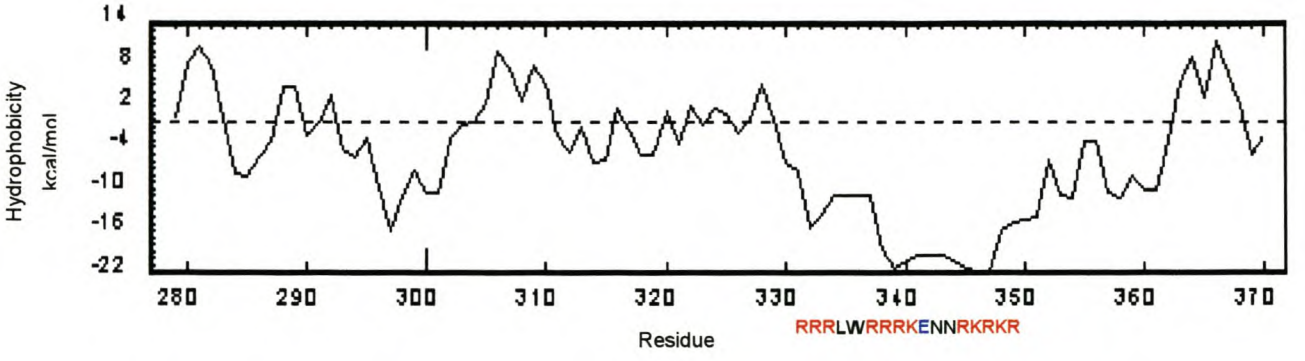
suGF1 basic domain : **RRRLWRRRKENNRKRKR**

hORFX basic domain : **NKPKKKKEKKEKEKKKKDKEKEKEKHKVK**

At first glance the apparent abundance of positive amino acid residues in both protein sequences, suggests a shared ability of both proteins to bind DNA, especially via electrostatic interactions between these positive residues and the negatively charged phosphate backbone of the DNA.

Given the negative results in the EMSAs, the physicochemical and structural properties of these two basic regions were compared using sequence analysis and prediction tools, in order to understand why the two proteins exhibit different DNA-binding specificities. Using the Genetool (DoubleTwist) hydropathy-plot program, the hydrophobicity plots for these two regions were generated and are graphically presented in Figure 4.6. The

a) suGF1 DNA-binding domain (amino acids 330 – 350):



b) Putative humORFX DNA-binding domain (amino acids 488 – 517):

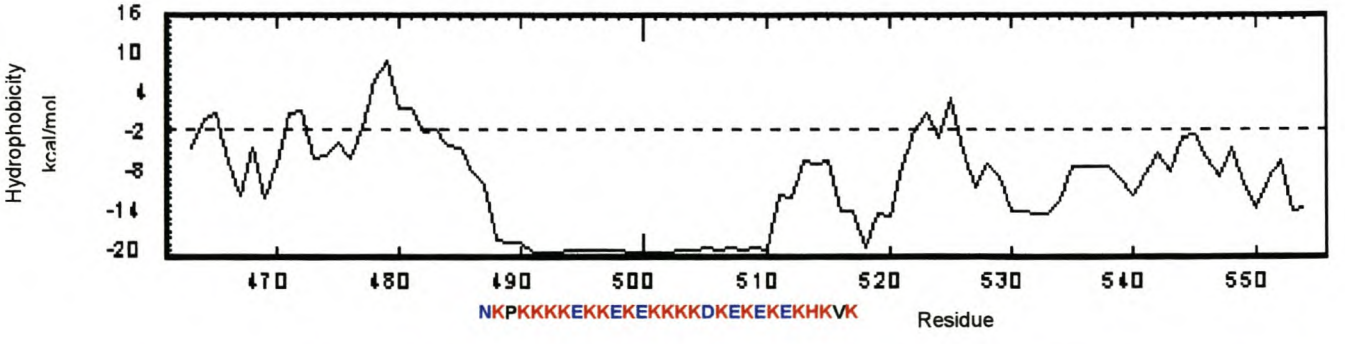


Fig. 4.6 <u>Hydrophobicity plots for the basic regions of the suGF1 and humORFX proteins respectively.</u>

The hydrophobicity plots for the suGF1 (a) and hORFX (b) are shown. A hydrophobicity algorithm (Wisshart *et al.*, 1994) was used to calculate the smoothed hydrophobicity of the given amino acid sequences. The hydrophobicity values are given on the y-axis in kcal/mol and the residue numbers are given on the x-axis. The amino acid sequences for the respective suGF1 and hORFX basic regions are given, where negatively charged residues are indicated by blue one letter abbreviations, uncharged residues in black and positively charged residues in red.

hydrophobicity plots display the hydrophobicity of any given amino acid sequence by using a specific algorithm that creates a smoothed hydropathy pattern which can be used to generate physicochemical information regarding the molecule. Figure 4.6 illustrates that both proteins contain regions of extremely low hydrophobicity. The DNA-binding domain of suGF1 (Fig. 4.6 (a)) is unique in the sense that more than 70% (13 / 18) of this region is occupied by positively charged amino acid residues (Arg and Lys; indicated in red) and contains only one negatively charged residue (Glu; indicated in blue). The hydrophobicity value reached a minimum of approximately –22 kcal/mol, which is extremely low, compared to that of the rest of the sequence. Similarly, the basic region of the hORFX protein (Fig. 4.6 (b)) contains almost 65% positively charged amino acid residues (Arg and Lys; indicated in red). However, scattered through the sequence are multiple, negatively charged residues (Glu and Asp; indicated in blue) that also participate in the overall hydrophobicity profile. This basic region extends for more than 20 amino acid residues and also has a very low minimum hydrophobicity value of approximately –20 kcal/mol, similar to that for suGF1. Both proteins are therefore predicted to have comparable physicochemically features, especially within their respective basic regions.

4.6 Secondary Structure Analysis

The question therefore arose — if these domains are so similar at first glance, why can suGF1 specifically recognise and bind G-strings, whereas hORFX can not? Both basic domains contain an unusual amount of positively charged amino acid residues, producing very similar hydropathy plots. This however is where the resemblance ends. Secondary and tertiary structure prediction and analysis, as well as careful inspection of the respective basic region sequences, predicted significant differences between these two domains.

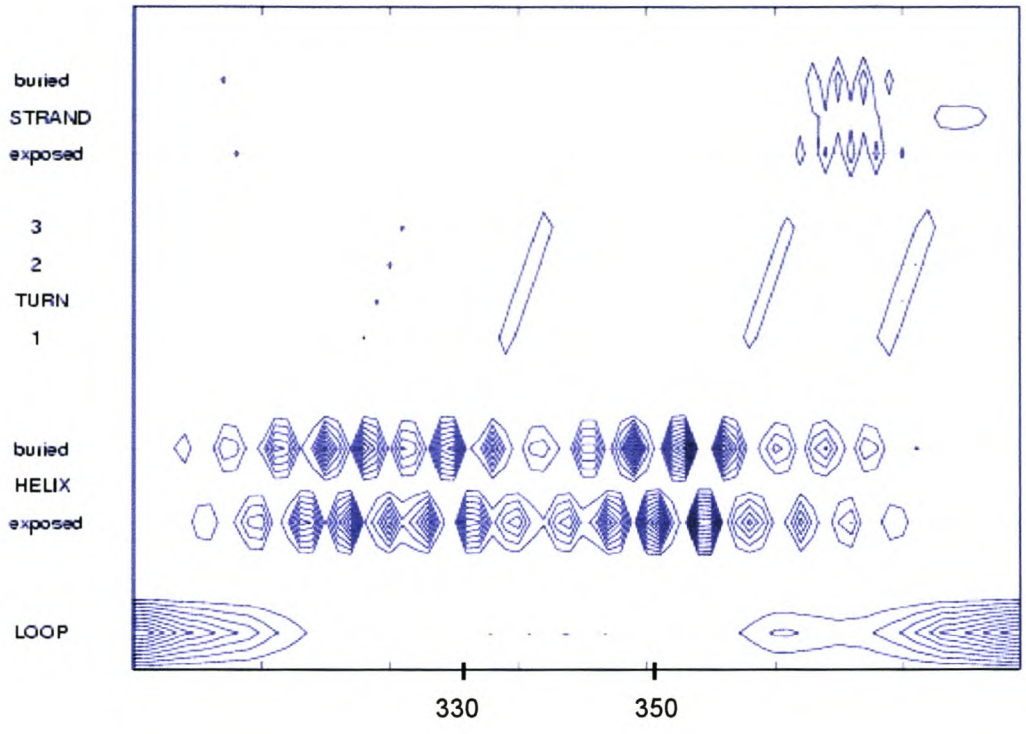
Exhaustive database searches showed that suGF1 and hORFX share no sequence homology to any of the entries contained in the searched databases for which structural information (3-D models, folding patterns etc.) is available. Therefore the amino acid sequences for the basic regions of suGF1 and hORFX were submitted to the PSA server for a Type-1 structure-prediction analysis, to learn more about possible structural differences.

The Protein Sequence Analysis (PSA) server at the BioMolecular Engineering Research Center (BMERC) of Boston University, is a program developed to predict secondary structures and folding classes for a given amino acid sequence. This prediction is based on sequence only and can be used for amino acid sequences for proteins of unknown structure. The amino acid sequence is submitted to the server and the user indicates the sequences to be analysed in one of three ways: Type-1, type-2 or WD-repeat (four or more copies of a Trp-Asp repeat) analysis. These discrete state-space models can be used to predict characteristic patterns of alpha helices, strands, tight turns and loops in specific structural classes. Table 4.1 summarises the essential features of these three different analysis models. For the analysis of the suGF1 and hORFX basic regions, the type 1 analysis was used, specifically because this model is more representative for smaller, single domain sequences that might fall into several distinct structural classes. Furthermore, the presence of the highly basic domains in both proteins suggest that they might be water-soluble, a property which is a pre-requisite for Type-1 analyses (Table 4.1). Because the sequence of the basic domains alone was used for PSA analysis, one can assume a single-domain status for the given amino acid sequence, another mandatory characteristic for Type-1 analysis.

Model	Sequence Properties	Sequence Length
Type-1	Several recognised structural classes for complete sequences including monomeric, single-domain, globular, water-soluble proteins	40 – 350 residues
Type-2	Partial or complete sequences for multimeric or multi-domain proteins, which are not globular or soluble	< 1000 residues
WD-repeats	Only for WD-repeat proteins	< 1000 residues

Table 4.1 Discrete state-space models to predict secondary structure from the PSA server.

Figure 4.7 is a graphical depiction of the results obtained for the structure prediction analysis of the suGF1 basic region. Areas surrounded by a dense mass of lines represent regions of high probability, compared to areas outside the contours which represent probabilities of lower than 0.1. The relative abundance of contours preceding the suGF1 basic region (residues 320–330), in the buried and exposed helical state, suggest this area to be an alpha helix. Immediately following this alpha helix, the contours reposition and are more abundant within the turn-state, suggesting this region to be a turn-like structure, after which the contours again accumulate solely in the helix state, predicting the last section of the basic region to be in a helical conformation again. Residues 332–350, constituting the suGF1 DNA-binding domain (suGF1 DBD) are therefore predicted to have a helix-(irregular turn / β-turn)$_2$-helix structure. Due to the electrostatic repulsion of positive charges this sequence is most likely exposed and capable of interacting with DNA. The abundance of the positive charges might induce the formation of the irregular turn / β-turn, which protrudes from the rest of the molecule and exposes the positive charges to the



Amino acid residues 332 – 350:

RRRLWRRRKENNRKRRKR

Fig. 4.7 A contour graph depicting the Type-1 secondary-structure probabilities of the suGF1 DNA-binding domain (DBD) as predicted by the PSA server.

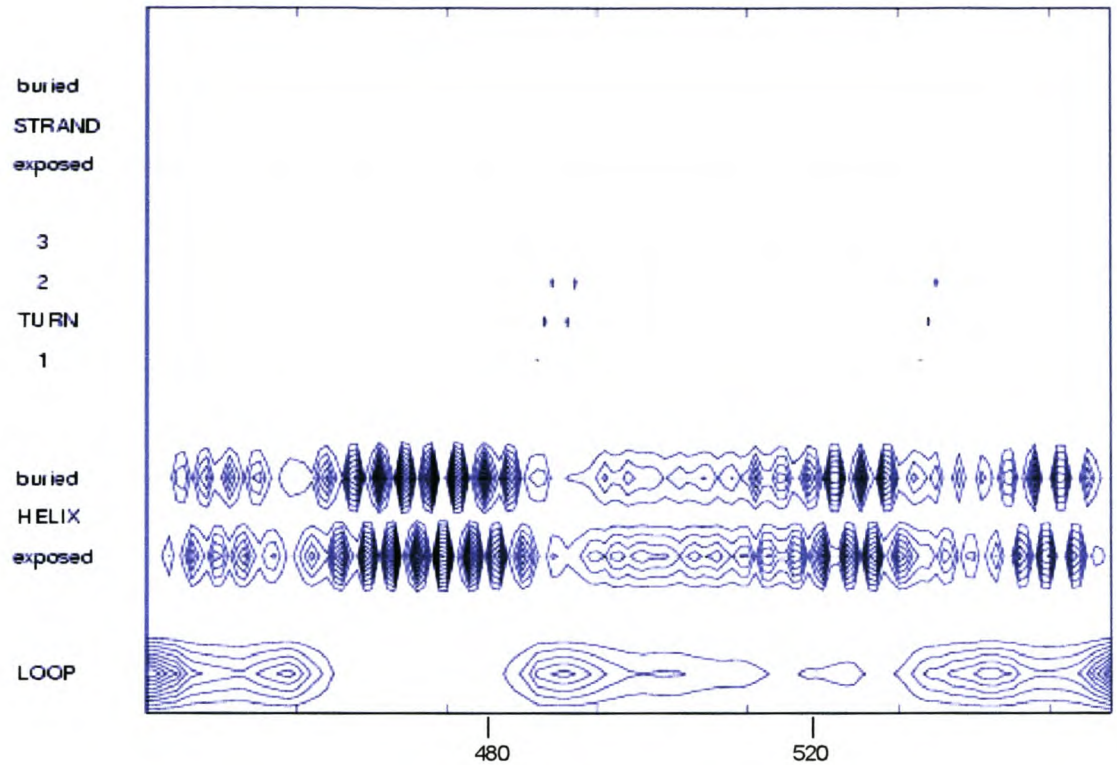
The x-axis (columns) represents the position of a specific residue, while each row on the y-axis correspond to a different secondary structural state. The probability of a specific residue being in each of the different structural states is depicted using contour lines of constant probability in increments of 0.1. The sequence of the suGF1 basic region is given. Positively charged residues are shown in red, whereas uncharged and negatively charged residues are indicated in green and purple respectively.

exterior. This hypothesis is consistent with the fact that suGF1 is a DNA-binding protein *in vitro* and substantiates the predicted secondary structure for the DBD as illustrated in Fig. 4.7.

The structure prediction for the hORFX basic region (amino acids 488 - 517) (Fig. 4.8) by the PSA server clearly displayed significant structural differences in comparison to that of the suGF1 basic region. The relative abundance of contours, preceding the start to the hORFX basic region, in the buried and exposed helical states, suggests this area to be an alpha helix. Immediately following this alpha helix, the contours reposition (amino acids 488 - 517) and are more abundant within the loop state, suggesting this region to be a loop-like structure, after which the contours again accumulate solely in the helix state, suggesting that the C-terminal sequence flanking the basic domain is again an alpha helix. Taken together, the PSA structure prediction indicates that the basic region of the hORFX protein is structured as a helix-loop-helix structure and is different to that of the helix-(irregular turn / β-turn)$_2$-helix predicted for the suGF1 basic region. The scattered presence of negatively charged residues in between the positive residues might induce a closed and buried loop-like structure within the tertiary assembly of the molecule, concealing and also diminishing the net positive charge of the region, due to the electrostatic attraction between positive and negative residues. This conformation would most likely be incapable of binding DNA via electrostatic interaction.

4.7 Tertiary Structure Analysis

The primary and secondary structure analysis of suGF1 and hORFX implied certain structural similarities and differences in the basic regions of the respective proteins. This



Amino acids 488 – 517:

NKPKKKKEKKEKEKKKKDKEKEKEKHKVK

Fig. 4.8 **A contour graph depicting the Type-1 secondary-structure probabilities of the hORFX basic region as predicted by the PSA server.**

The x-axis (columns) represents the position of a specific residue, while each row on the y-axis correspond to a different secondary structural state. The probability of a specific residue being in each of the different structural states is depicted using contour lines of constant probability in increments of 0.1. Therefore, areas surrounded by a dense mass of lines represent regions of high probability, compared to areas outside the contours which represent probabilities of lower than 0.1. Positively charged residues are shown in red, whereas uncharged and negatively charged residues are indicated in green and purple respectively.

information, however, revealed nothing regarding the 3-D structures of the two proteins, which would significantly expand the overall picture generated from the prediction results. In the absence of experimental data, model building on the basis of the known 3-D structure of a homologous protein is at present the only reliable method to obtain structural information. Comparisons of the tertiary structures of homologous proteins have shown that three-dimensional structures have been better conserved during evolution than protein primary structures, and massive analysis of databases holding results of these 3-D comparison methods, as well as a large number of well studied examples indicate the feasibility of model-building by homology (White, 1994). Due to the fact that the 3-D structures of neither these two proteins have been solved by experimental methods, the amino acid sequences of the respective basic regions were subjected to a novel method of tertiary structure prediction, threading, which is based on classical homology modeling principles (Hartl, 1994). Both methods are actually based on sequence homology and similarity. Homology modeling uses structural, sequence homology between amino acid regions and superimposes the unknown sequence on the known sequence. Threading (fingerprinting) is a subset of homology modeling, however a library containing different fold types is searched for sequences similar to the query sequence, after which different folds can be combined to create the full protein. When a query sequence therefore exhibits low global alignment and therefore no significant sequence homology with database entries, homology modeling would be impossible. Fragmenting the query sequence and performing local alignments, however, might generate regions of significant similarity, which can ultimately be used to predict a specific fold (if the fold is present in the library). In this case threading is a much better option and will generate structural data with much higher confidence than with homology modeling.

The highly basic region of suGF1 (putative DNA-binding domain) clearly showed no global sequence homology to any of the entries present in the databases up to date. Significant, however, is the fact that the basic region of suGF1 did exhibit good local alignments with ten database entries (TopLign 123D used for searches), suggesting a possible similarity at the domain level (Fig. 4.9 and 4.11).

The two most significant suGF1 alignments happened to be those for transcription factors, validating the potential role of suGF1 as a transcription factor *in vivo*. The first potential candidate identified, was a murine helix-turn-helix protein, Ets-1 (PDB classification: 1etc), a member of the Ets transcription factor family (Donaldson *et al.,* 1996). The basic region of suGF1 showed 61% mapped (only the aligned sequences) amino acid similarity (over 28 residues) and 18% homology to the helix-turn-helix (HTH) region of the Ets-1 transcription factor (Fig. 4.9). The 3-D structure for this region clearly shows two protruding, anti-parallel β-sheets (indicated in yellow and turquoise) connected to two right-hand twisted, α-helical bundles on both sides of each turn (end of the β-sheet) (Fig. 4.10). The winged HTH motif is a classic feature of the Ets-family of transcription factors which belong to the Winged HTH superfamily of DNA-binding proteins.

The second potential candidate is a transcription factor present in the Phage Mu. The structure of the DNA-binding of domain this Phage Transposase DNA-binding protein has been solved by NMR and was found to belong to the homeodomain-like superfamily of transcription factors. The suGF1 basic region shared 28% identity and 82% similarity to this domain when performing a 50 amino acid mapped alignment (Fig. 4.11). The 3-D structure for this region (Fig. 4.12) shows a similar appearance to that obtained for the murine Ets-1 protein, however, four protruding turns / loops (turquoise, yellow, orange and green) are present with two right-hand twisted, α-helical bundles on both sides of each

```
1etc              0000:......GSGPIQLWQFLLELLTDKSCQSFISWTGDGWEFKLSDPDEVARRWGKRKNKPKM
Sequence          0000:SFTAEA.................................AELADRRRLWRRRKENNRK
score             0000:.....................................|| ||  R W |RK|| ||
No. of '1etc'     0000:       ........10........20........30........40........50....
No. of 'Sequence' 0000:......                                    ..10........20.....
SECSTR '1etc'     0000:                    hhhhhhhh  hhh            eeee     hhhhhh

1etc              0060:NYEKLSRGLRYYYDKNIIHKTAGKRYVYRFVCDLQSLLGYTPEELHAMLDVKPDAD....
Sequence          0060:RRKRMEKQL.............................................EKIE
score             0060:    || | L.....................................................
No. of 'Sequence' 0060:....60........70........80........90.......100.......110
No. of 'Sequence' 0060:...30....                                                ....
SECSTR '1etc'     0060: hhhhhhhhhh                              hhhhhh

1etc              0120:.....................
Sequence          0120:QRSCELLFHITSRGAYDRVRSH
score             0120:.....................
No. of '1etc'     0120:
No. of 'Sequence' 0120:40........50........60
SECSTR '1etc'     0120:

Alignment                    value  Alignment    Prof-1    Prof-2    Mapped
Alignment length       =        28                110        60        28
Alignment value        =     50.20       1.79      0.46      0.84

Alignment ids          =         5     17.86 %    4.55 %    8.33 %   17.86 %
Alignment homs         =        17     60.71 %   15.45 %   28.33 %   60.71 %
```

Fig. 4.9 Local alignment of the suGF1 basic region with the helix-turn-helix domain from murine Ets-1 (PDB classification: 1etc).

The alignment identities (ids) and homology (homs) are given as percentage of the overall alignment. When the alignments are mapped (i.e. just the aligned regions, indicated in bold) the two sequences share 17.86% identity and 60.71% homology (similarity) on the amino acid level. The secondary structure for 1etc (**SECSTR)** is also given (h = helix, e = strand, c = coil).

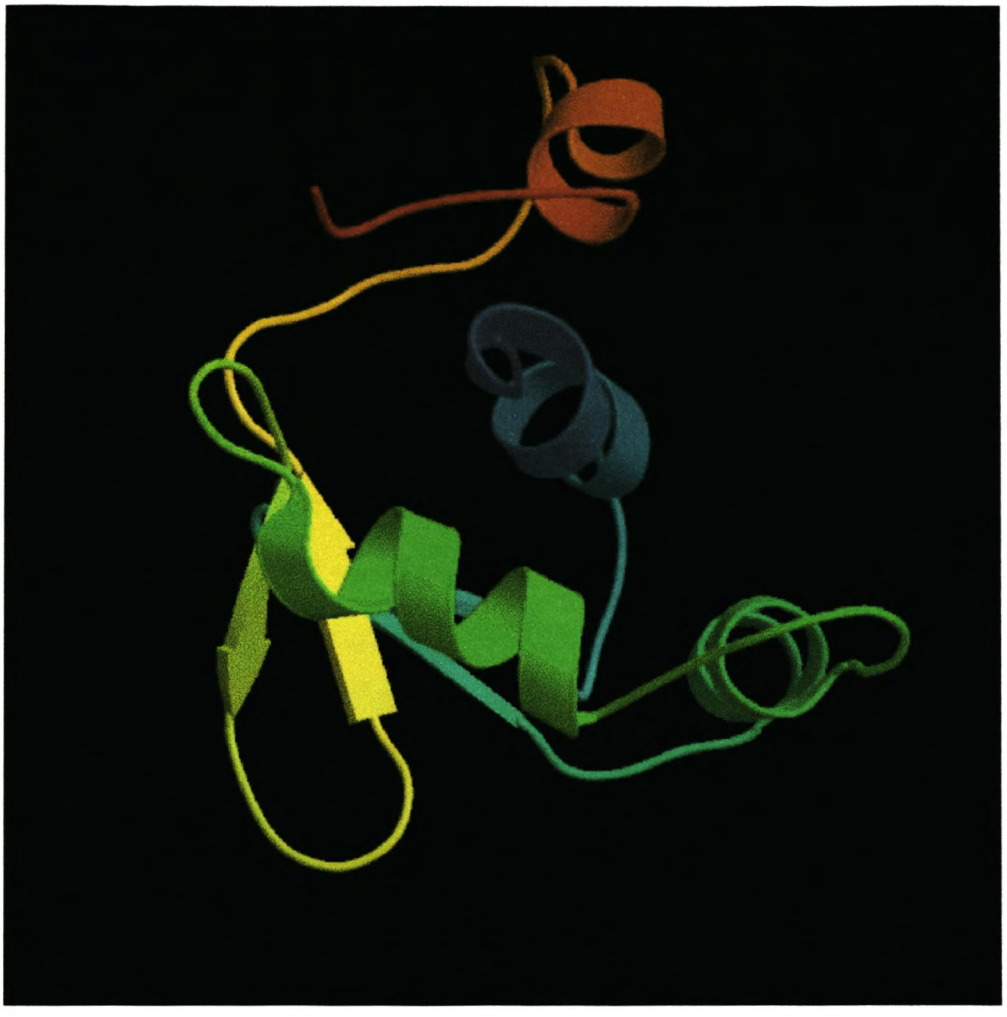


Fig. 4.10 **Structure of the suGF1 DNA-binding domain as predicted by fold recognition.**

The ribbon-structure of the suGF1 DNA-binding domain as predicted by fold recognition (threading), is shown. The modelling was based on the crystal-structure of a murine ETS transcription factor, which has a winged helix-turn-helix DNA-binding motif.

```
2ezl_0_scop           0000:MIARPTLEAHDYDREALWSKWDNASDSQRRLAEKWLPAVQAADEMLNQGISTKTAFATVA
suGF1                 0000:------------DRRRLWRRRKENNRKRRKRMEKQLEKIEQRSCELLFHITSRGAYDRVR
consensus             0000:MIARPTLEAHDYDRRRLWRRWKDDNRKRRRMMEKWLPKIQQRDCMLLFHITTRGAYDRVR
score                 0000:            DR||LW||||||||||R|||EK|L|||| ||  L   I |||A|||V|
SECSTR '2ezl_0_scop'  0000:             hhhhhhhhh   hhhhhhhhhhhhhhhhhhhhhh    hhhhhhhhh

2ezl_0_scop           0060:GHYQVSASTLRDKYYQVQKFAKPDWAAALVDGRGASRRN
suGF1                 0060:SH-----------------------------------
consensus             0060:GHYQVSASTLRDKYYQVQKFAKPDWAAALVDGRGASRRD
score                 0060:|H
SECSTR '2ezl_0_scop'  0060:hh    hhhhhhhhhhhh      hhhh
```

Alignment		value	Alignment	Prof-1	Prof-2	Mapped
Alignment length	=	99		99	50	50
Alignment value	=	724.84	7.32	7.32	14.50	
Alignment ids	=	13	13.13 %	13.13 %	26.00 %	26.00 %
Alignment homs	=	41	41.41 %	41.41 %	82.00 %	82.00 %

Fig. 4.11 Local alignment of the suGF1 basic region with the DNA-binding domain of Phage Mu Transposase (PDB classification: 2ezl).

The alignment identities (ids) and homology (homs) are given as percentage of the overall alignment. When the alignments are mapped (i.e. just the aligned regions, indicated in bold) the two sequences share 26% identity and 82% homology (similarity) on the amino acid level. The secondary structure for 2ezl (**SECSTR)** is also given (h = helix, e = strand, c = coil).

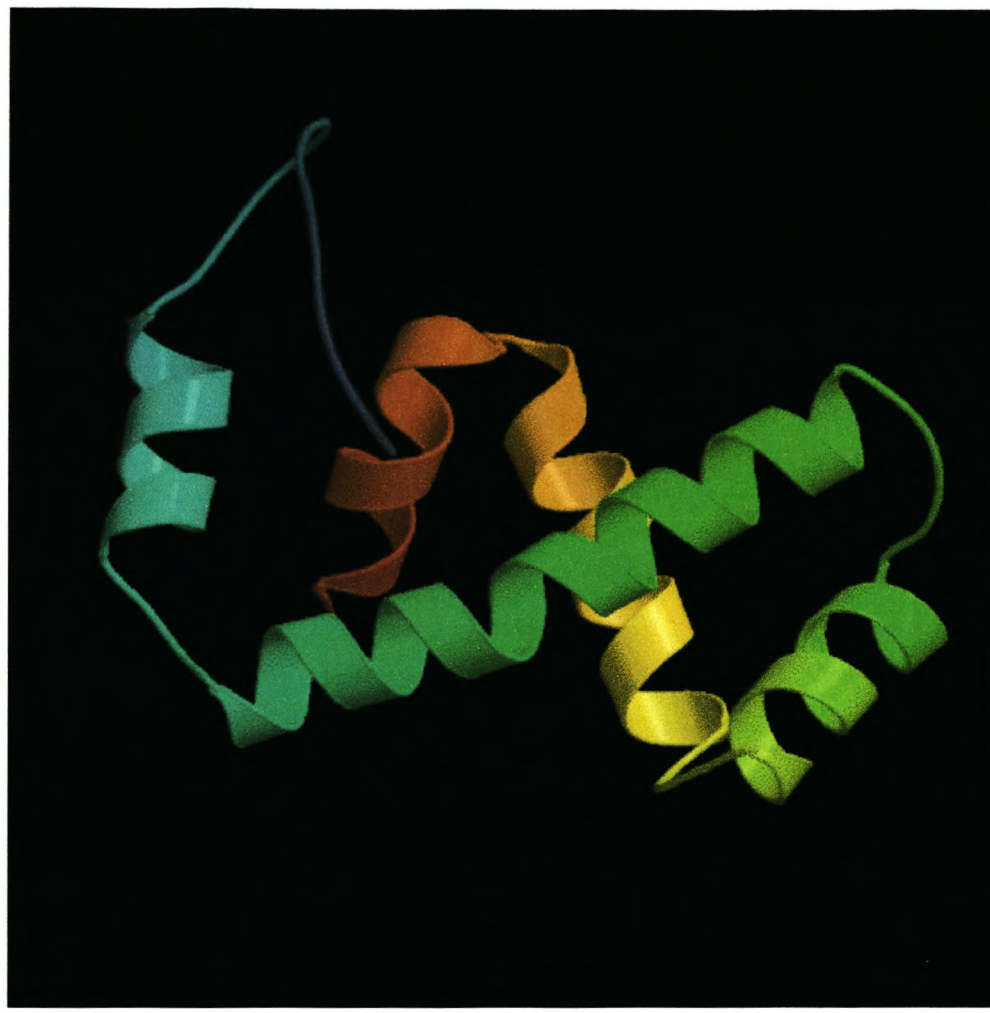


Fig. 4.12 **Structure of the suGF1 DNA-binding domain as predicted by fold recognition.**

The ribbon-structure of the suGF1 DNA-binding domain, as predicted by fold recognition (threading), is shown. The modelling was based on the structure for a bacteriophage Mu DNA-binding protein, which has a helix-turn-helix DNA-binding motif.

turn. The Phage Mu Transposase protein exhibit slightly altered domain-protrusions in that they seem to loop out of the alpha helix, rather than going over in an anti-parallel β-sheet, as is the case for the murine Ets-1 factor. Significant however is the fact that both structures resemble common, comparable DNA-binding motifs that share significant similarity to the suGF1 basic region.

When the hORFX protein was subjected to the same threading process to predict the tertiary structure, the amino acid sequence of this proteins basic region showed high similarity to one entry present in the fold databases. Although the two sequences only share 7% amino acid homology, a 93% similarity value (Fig. 4.13) indicates that they are chemically highly comparable, and that their respective 3-D structures might therefore also exhibit comparable features. The 3-D structure for this region, present in the Max DNA-binding protein, exhibits one small coil-like protrusion with two long left-hand twisted, α-helices on both sides (Fig. 4.14). The Max protein belongs to the helix-loop-helix superfamily of transcription factors, suggesting that hORFX might afterall be involved in gene regulation *via* binding to DNA (although not G-strings). Table 4.2 summarises the structural similarities and differences between suGF1 and hORFX, based on the results obtained for threading.

```
1hloA_0_scop          0000:----NDDIEVE-------SDADKRAHHNALERKRRDHIKDSFHSLRDSVPSLQGEKASRA
humORFX               0000:LKAVHEQLAALSQAPVNKPKKKKEKKEKEKKKKDKEKEKEKHKVKAEEEKKAKVAPP---
consensus             0000:LKAVHDELEVLSQAPVDKPKKKKRKHHKEMKRKRRDHIKDKFHVMRDEVPKLKGEPPSRA
score                 0000:     |||||||        ||||K|||||||||K|||||K||||||||||||||||||
SECSTR '1hloA_0_scop' 0000:      hhhhh          hhhhhhhhhhhhhhhhhhhhhhhhhhhhhh hhh      hh
SECSTR 'humORFX'      0000:                       cchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccc

1hloA_0_scop          0060:QILDKATEYIQYMRRKNHTHQQDIDD---LKRQN
humORFX               0060:--------AKQAQQKKAPAKKANSTTTAGRQLKK
consensus             0060:QILDKATEYKQYMRRKDHTHKQDIDDTAGMKMKK
score                 0060:           Q||||K|||||||       |||||
SECSTR '1hloA_0_scop' 0060:hhhhhhhhhhhhhhhhhhhhhhhhhh    hhhh
SECSTR 'humORFX'      0060:         chhhhhhccccccccccccccchhcc

Alignment                                        Prof-1    Prof-2    Mapped
Alignment length          =       94               80        83        69
Alignment value           =  1080.93     11.50     13.51     13.02
Alignment ids             =        5      5.32 %    6.25 %    6.02 %    7.25 %
Alignment homs            =       64     68.09 %   80.00 %   77.11 %   92.75 %
Alignment mapped          =       69     73.40 %   86.25 %   83.13 %  100.00 %
```

Fig. 4.13 Local alignment of the hORFX basic region with the helix-loop-helix domain from the human Max DNA-binding domain.

The alignment identities (ids) and homology (homs) are given as percentage of the overall alignment. When the alignments are mapped (i.e. just the aligned regions, indicated in bold) the two sequences share 7.25% identity and 92.75% homology (similarity) on the amino acid level. The secondary structure for 1etc (SECSTR) is also given (h = helix, e = strand, c = coil).

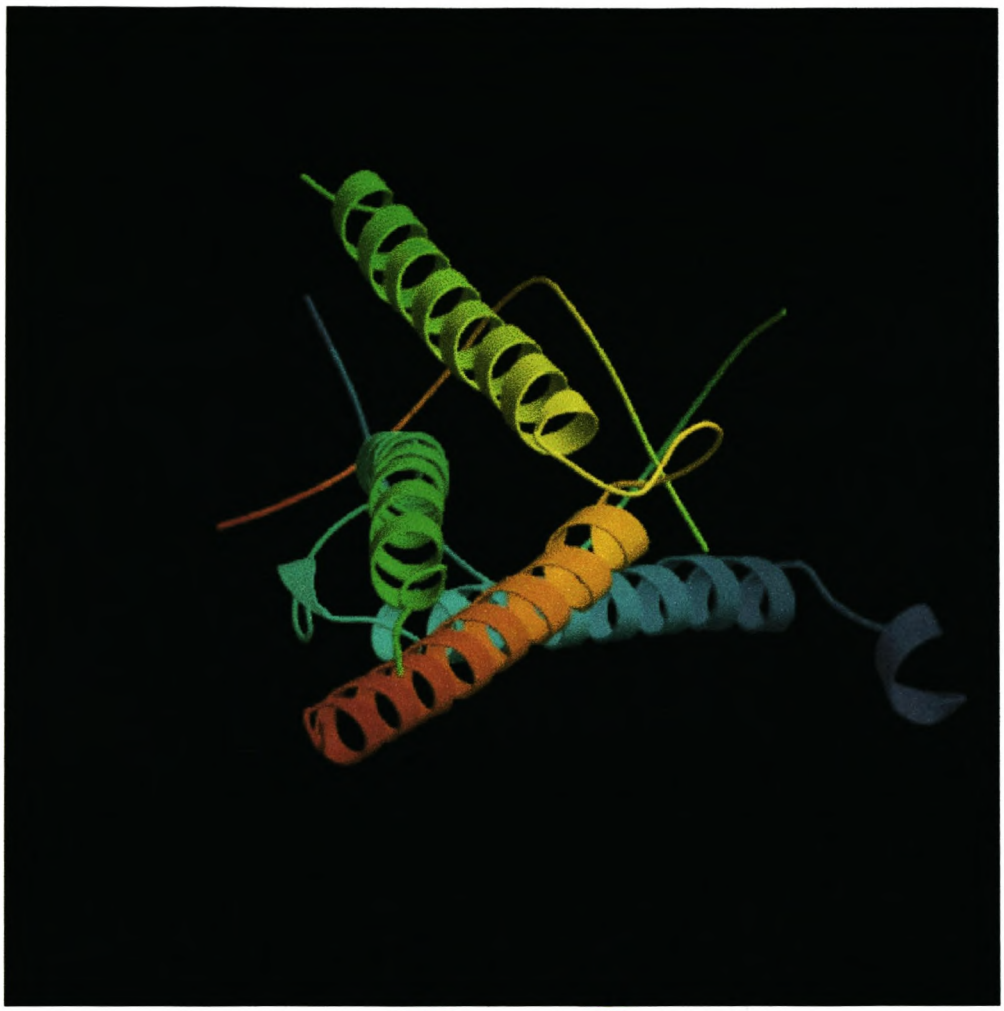


Fig. 4.14 Structure of the hORFX basic region as predicted by fold recognition modeling.

The ribbon-structure of the hORFX basic region, as predicted by fold recognition (threading), is shown. The modeling was based on the crystal-structure of human Max transcription factor that contains a helix-loop-helix DNA-binding domain, shown here in complex with its potential binding site present in a synthetic oligodeoxyribonucleotide.

Protein (PDB no.)	Structure	*In vivo* function	% Similarity to suGF1 basic region	% Identity to suGF1 basic region
1etc	Helix-Turn-Helix	Mammalian Transcription factor	61	18
2ezl	Helix-Turn-Helix	Phage Mu DNA transposition	82	28

Protein (PDB no.)	Structure	Function	% Similarity to hORFX basic region	% Identity to hORFX basic region
1hlo	Basic Helix-Loop-Helix	Mammalian Transcription factor	93	7.3

Table 4.2 A summary of the tertiary structure predictions for the suGF1 and hORFX basic regions.

The table summarises the results for threading of the suGF1 and hORFX basic regions. The general structural features and *in vivo* functions for each of the proteins are given. The mapped, local alignment scores for similarity and identity are given as a percentage value.

Chapter 5 - Results

Recombinant Protein Expression in Yeast

5.1 Introduction

The exact *in vivo* function of suGF1 is at present still an unsolved puzzle. In the light of the fact that suGF1 binds specifically to G-strings, a goal in the laboratory was to investigate a possible role for suGF1 in transcriptional regulation. Several lines of indirect evidence (see Chapter 1) support a functional role for suGF1 as a transcription factor. The classic double hybrid yeast transactivation assay (St. John, 1981) would be a good model system for studying the putative transactivation potential of suGF1. A significant advantage of this experimental setup is the fact that database searches indicated the absence of any known GC-box binding proteins in *S.cerevisiae.* Testing for suGF1 / SpGCF1 function within such a system would therefore reflect solely on the effect of the protein synthesised from the expression construct. The control samples as well as the test samples would therefore most likely exhibit relatively low levels of endogenous proteins (background), which would significantly increase the reliability of the results. As a first step in setting up the transactivation experiments in yeast, an suGF1 expression construct had to be engineered that would be expressed at high levels in a yeast cell line. A second criterion, which would be necessary to ensure the suitability of the yeast expression system, would be that the expressed suGF1 exhibit similar DNA-binding properties to the native protein.

5.2 Preparation of an suGF1 Expression Construct

The suGF1 cDNA had previously been cloned and inserted into the *XhoI–NotI* site of the pcDNAI-Amp vector (Table 2.1) by Dr S.Scherer (1997). This construct was digested in the presence of *HindIII* and *XbaI* generating the linearised pcDNAI-Amp vector and a 2.0 kb full-length suGF1 cDNA insert. The insert was gel purified and cloned into the *HindIII-XbaI* site of a 5.9 kb shuttle vector pYES2 (Invitrogen), generating the 7.9 kb pYES2-suGF1 expression construct (Table 2.1). The pYES2 vector contains the very strong, D-galactose inducible *GAL1* promoter, a T7 promoter transcription start site, an ampicillin resistance gene and a selectable marker gene, *URA3*. The integrity of the pYES2-suGF1 expression extract was analysed by restriction enzyme digestion with *HindIII* and *XbaI* (Fig. 5.1 - lane 3), which generated a 5.9 kb linearised pYES2 plasmid and the full-length suGF1 cDNA insert (2.0 kb).

5.3 Expression of Recombinant suGF1 from pYES2-suGF1

To test whether suGF1 can be recombinantly expressed in *S.cerevisiae*, the pYES2-suGF1 expression construct was transformed into the protease-deficient yeast strain Y294. The transformed cells were selectively grown on CSM dropout plates (-Ura / + Sorbitol), because the shuttle vector contains the *URA3* gene. The picked colonies were grown in the presence of D-galactose, which effects expression from the *GAL1* promoter (upstream from the inserted suGF1 cDNA) and induces expression of the cDNA from the transcription start site. Nuclear extracts were then prepared from the yeast cells. Since the *GAL1* promoter theoretically induces transcription of upstream cDNAs up to 10 000 fold (St.John, 1981), suGF1 was predicted to be present at high levels in these nuclear extracts. Nuclear extracts were analysed by SDS-PAGE to confirm the presence of

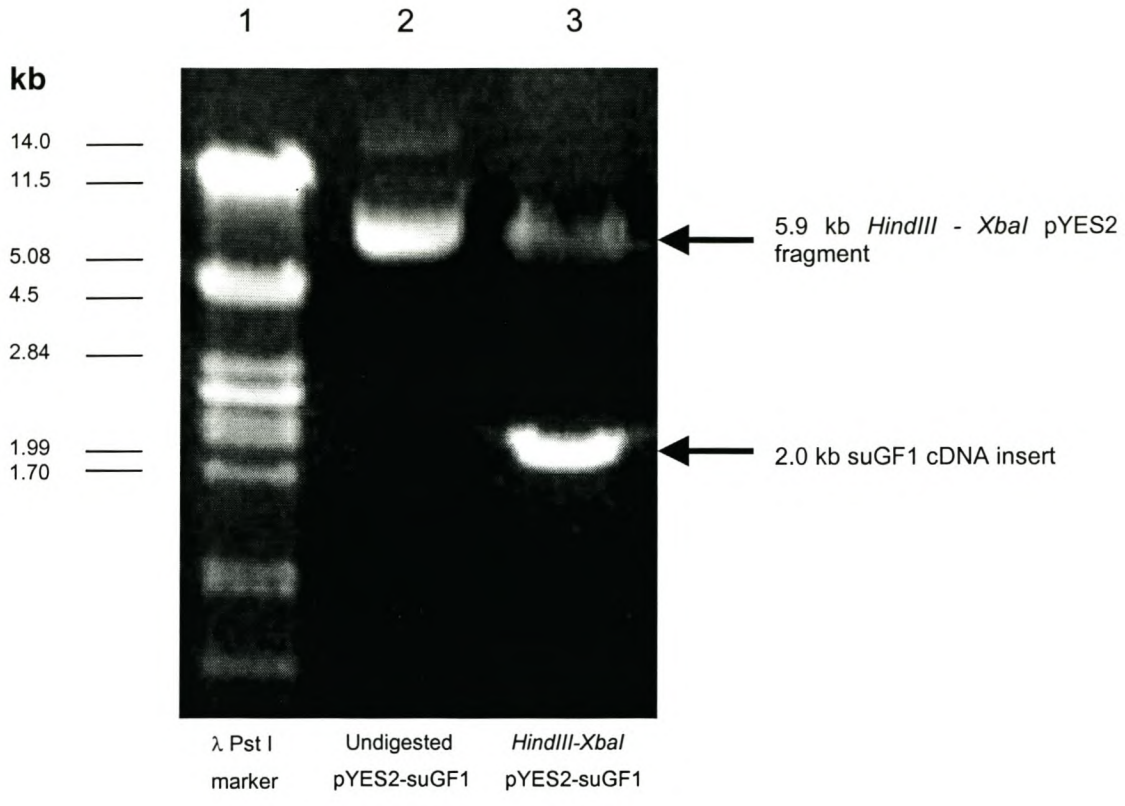


Fig 5.1 Analytical agarose gel analysis of the pYES2-suGF1 expression construct.

An ethidium bromide stained analytical agarose gel of the pYES2-suGF1 construct after digestion with *HindIII* and *XbaI* (lane 3) is shown. The 5.9 kb linearised pYES2 plasmid and the full-length suGF1 cDNA insert (2.0 kb) are shown with arrows. The undigested plasmid is shown in lane 2. A λPstI standard marker is shown in lane 1 with the respective band sizes indicated in the margin.

suGF1. Thereafter, the DNA binding properties of the recombinant suGF1 were investigated by electrophoretic mobility shift assays. As negative control the Y294 strain was transformed with the pYES2 vector containing no suGF1 cDNA, and subjected to the same experimental procedures as the extracts obtained from the cells transformed with pYES-suGF1.

The SDS-PAGE analysis (Fig. 5.2) clearly showed the presence of a series of unique bands (B1–B3) present in the lane (lane 2) containing the nuclear extracts derived from Y294 cells transformed with the pYES2-suGF1 expression construct. Relative to the marker lanes (Rainbow marker – lane 5; BSA standard – lane 6), the slowest migrating band in lane 2 has an estimated molecular weight of about 60 kDa, which is consistent with the molecular weight of suGF1 which has previously been determined to be 59.5 kDa. No bands of similar electrophoretic mobility were observed in lane 1, which contains nuclear extracts from Y294 cells transformed with pYES2 (thus no suGF1 cDNA). The approximate molecular weights of the two slower migrating bands (lane 2) correspond well with that of two of the slower migrating bands obtained during *in vitro* transcription and translation of the suGF1 protein (Fig 3.1 and Section 3.2). This suggests, as postulated before, that truncated protein products are generated due to utilisation of multiple AUG translation initiation start sites from the suGF1 mRNA transcript. The bands that are present in both lanes 1 and 2 most likely constitute endogenous, background, Y294 proteins. Lanes 3 and 4 containing whole cell extracts prepared from FY23 cells transformed with pYES2 and pYES2-suGF1 respectively, exhibit no bands unique to suGF1, although the extracts seem to be degraded and therefore somewhat smeary. The SDS-PAGE analysis of yeast nuclear extracts shows that the full-length, recombinant suGF1 protein was successfully expressed in the yeast Y294 cell line. This experiment was however only performed once and should be repeated to confirm the result.

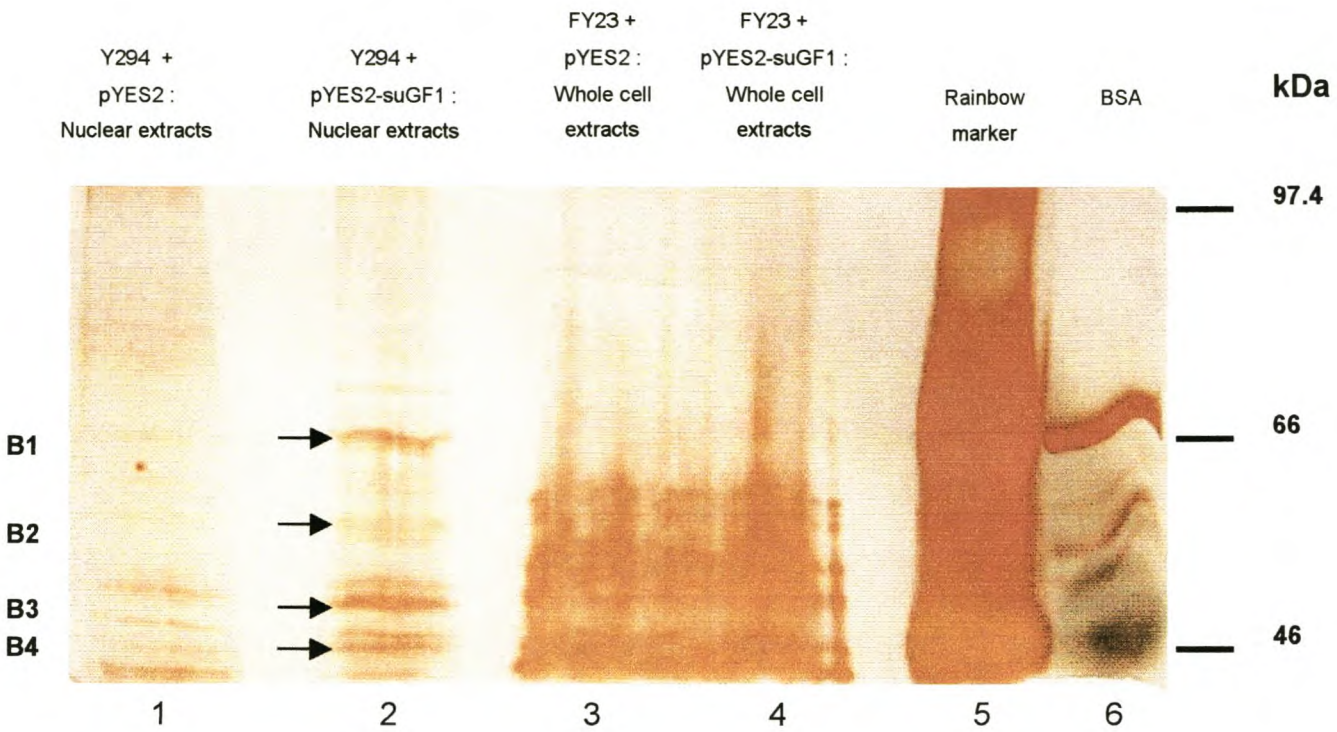


Fig 5.2 suGF1 is recombinantly expressed in Y294 yeast cells.

A silver stained SDS-PAGE gel of the yeast nuclear extracts is shown. In lane 2 that contains nuclear extracts from yeast cells transformed with the pYES2-suGF1 construct, multiple unique bands (indicated with arrows B1 - B4) representing recombinant suGF1 and truncations thereof, can be observed. The negative control reaction (lane 1) that contains nuclear extracts from Y294 cells transformed with the pYES2 vector (no suGF1 cDNA insert) showed no bands of similar electrophoretic mobility. The lanes containing whole cell extracts from FY23 cells transformed with pYES2 (lane 3) and pYES2-suGF1 (lane 4) respectively exhibit no unique bands. A rainbow marker (lane 5) and BSA standard (lane 6) are shown as standard markers. The respective sizes of the marker proteins are given in the margin. All lanes contain equal amounts of total protein i.e. 3.0 μg.

5.4 DNA-Binding Properties of the Recombinantly Expressed suGF1

The SDS-PAGE analysis of the Y294 nuclear extracts, prepared from yeast cells transformed with the pYES-suGF1 plasmid, verified the existence of at least four unique bands, which most probably represent the full-length suGF1 protein and truncations thereof. To investigate whether the recombinantly expressed suGF1 exhibited similar DNA-binding properties to native and IVT suGF1, electrophoretic mobility shift assays (EMSA) were performed. Y294 nuclear extracts and FY23 whole cell extracts were incubated with a radiolabeled, synthetic oligodeoxyribonucleotide (S-Oligo), containing a central G_{11}-string, which is present in the H1-H4 early histone gene battery of *P. miliaris*. Native and IVT suGF1 were previously shown to bind this sequence specifically (Fig. 3.5). The autoradiograph generated from the EMSA analysis of yeast whole cell and nuclear extracts, is shown in Fig. 5.3 and shows the presence of multiple protein-DNA complexes (B1–B3) in the lane containing nuclear extracts from Y294 cells transformed with the pYES2-suGF1 expression construct (lane 8). Complexes B1 and B2 are clearly competed away for by the addition of 100-fold molar excess cold S-Oligo (lane 9) as competitor DNA. This competition is however not exhibited by the addition of 100 fold molar excess of a random sequence DNA (NS-Oligo) (lane 10), showing that the binding of the recombinant protein is specific for the G-string. Lanes 5–7, containing nuclear extracts prepared from yeast Y294 cells transformed with the pYES2 vector only (no suGF1 cDNA insert), showed no protein-DNA complexes of reduced electrophoretic mobility. Furthermore, lanes 11 and 12, containing the yeast FY23 whole cell extracts incubated with the same probe, exhibited no specific bands of reduced electrophoretic mobility. The positive control reactions in lane 2 (native suGF1 present in sea urchin nuclear extracts) and lane 4 (IVT suGF1) generated characteristic protein-DNA complexes (A1-A5) of decreased electrophoretic mobility, similar to the results obtained in Fig. 3.5. The negative control reactions,

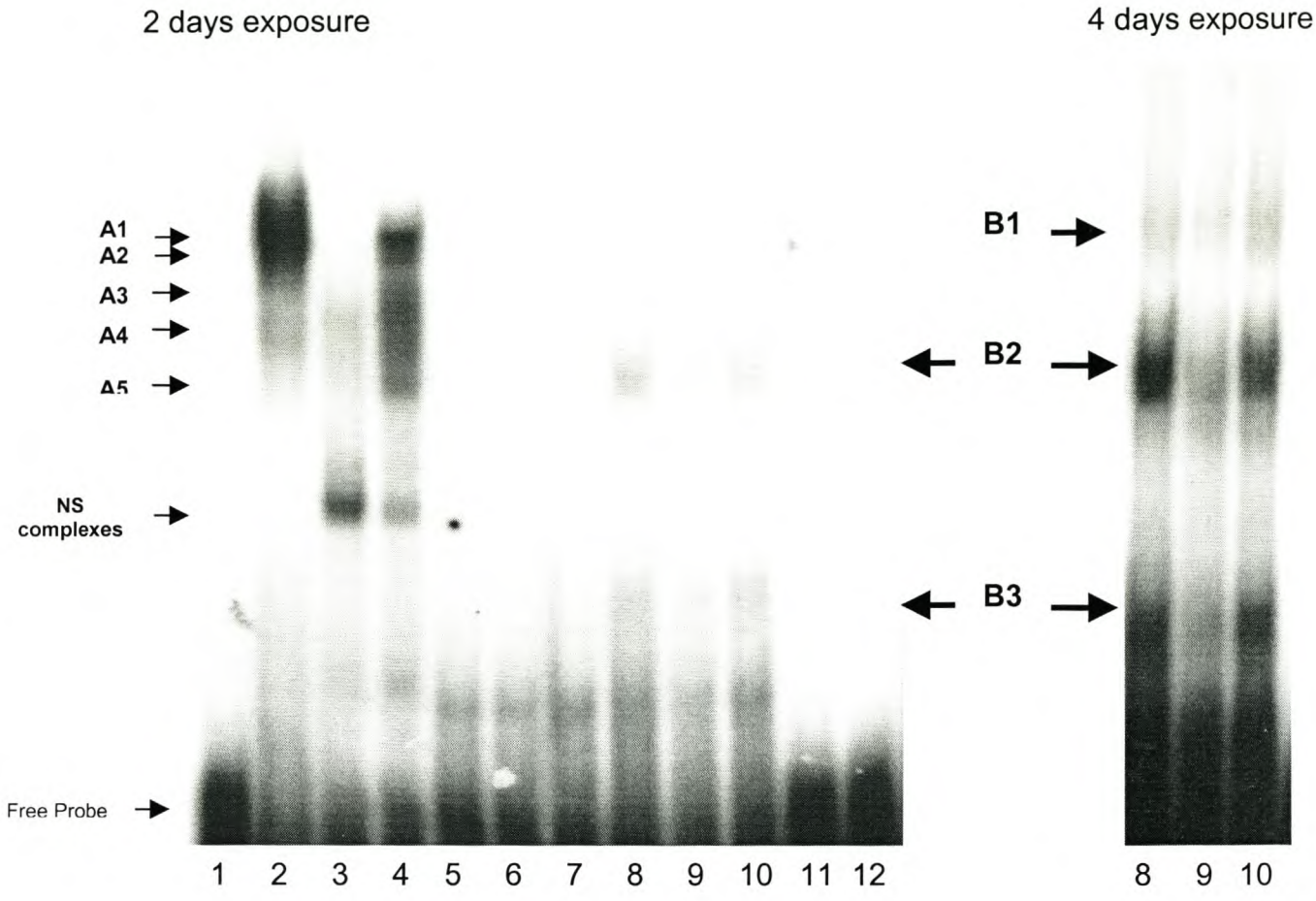


Figure 5.3 suGF1, recombinantly expressed in Y294 cells, exhibits similar DNA-binding properties to native and IVT suGF1.

An autoradiograph of the EMSA analysis of yeast nuclear extracts when incubated with a radiolabeled S-Oligo, is shown. Multiple protein-DNA complexes (B1 – B3) in lane 8, containing nuclear extracts from Y294 cells transformed with the pYES2-suGF1 expression construct, are shown. Lanes 9 and 10 contain a 100 fold molar excess of unlabeled S-Oligo and NS-Oligo respectively. Lanes 5 – 7, contain nuclear extracts prepared from yeast Y294 cells transformed with the pYES2 vector only (no suGF1 cDNA insert). Lanes 11 and 12 contain yeast FY23 whole cell extracts incubated with the same probe. The positive control reactions in lane 2 (native suGF1) and lane 4 (IVT suGF1) generated characteristic protein-DNA complexes (A1 – A5) of decreased electrophoretic mobility. Dialysis buffer and lysate, used as negative controls, are shown in lanes 1 and 3 respectively. The various protein-DNA complexes as well as the free radiolabeled probe are indicated with arrows.

containing dialysis buffer (lane 1) or rabbit reticulocyte lysate (lane 3) incubated with the same radiolabeled S-Oligo, produced no bands of real significance. The band in lane 3 (lysate) is probably a non-specific complex as this band is also observed in lane 4 (IVT suGF1 present in rabbit reticulocyte lysate). This putative protein-DNA complex might be due to the interaction between an endogenous G-string binding factor (present in the lysate) and the probe.

Chapter 6

Discussion and Conclusions

6.1 Expression and DNA-Binding Analysis of Native and IVT suGF1

The suGF1 cDNA was subjected to *in silico* expression, as well as *in vitro* transcription and translation (IVT) in a rabbit reticulocyte lysate system. The IVT suGF1 was subjected to SDS-PAGE, to analyse the protein products and compare these to those predicted by *in silico* expression. In addition, the DNA-binding properties of the IVT suGF1 were compared to those of the native suGF1 present in sea urchin nuclear extracts, by EMSAs.

Careful analysis of the suGF1 cDNA revealed the occurrence of multiple AUG translational start sites, suggesting that multiple protein products could be expressed (Fig. 3.3). The *in silico* protein products were predicted from the cDNA sequence and by comparison with the experimental results. It may well be that other ATG start codons are also utilised for initiation of translation. This is particularly valid when considering that many of the putative ATG start codons are relatively close to each other. Since determination of the molecular masses of the respective protein products from the autoradiograph of the SDS-PAGE can only be measured accurately up to two significant digits, it would be difficult to accurately assign the positions from which translation is initiated in the IVT.

SDS-PAGE analysis of the *in vitro* transcribed and translated (IVT) end products verified the presence of at least six suGF1 proteins (Fig. 3.1). The six protein products observed were calculated to have molecular masses of 58, 52, 48, 42, 35 and 32 kDa respectively. The truncations are most likely on the N-terminal side and are most probably all translation products from a single mRNA transcript due to utilisation of the multiple AUG initiation

codons. It is therefore likely that the proteins share a common hydrophobic center and basic region (putative DNA-binding domain), but differ at the N-terminus. It is possible that more, smaller truncated versions of suGF1 were also expressed. Such smaller proteins would not have been detected on the gel, as the smallest proteins would have eluted off the gel matrix during electrophoresis, due to their low molecular weight.

When considering the experimental and theoretical analysis of suGF1 cDNA expression, it is reasonable to predict that the truncated proteins would most likely represent a range of proteins that differ in their respective transactivation, multimerisation and membrane anchoring potentials (Fig. 6.1 summarises these findings). The full-length protein product, suGF1 (a), contains all the functionally significant domains. suGF1 (b), the longest of the truncated products is likely to also contain the putative multimerisation and transactivation domains, the central hydrophobic core region, as well as the established DNA-binding domain. However, it is possible that the Gram-positive membrane anchor is not present. The exact relevance of this domain (identified by sequence analysis) is still unclear. The two longest suGF1 protein products therefore, most likely, both contain all the domains essential for protein-protein interactions, transactivation and DNA-binding. The third truncation, suGF1 (c), shows a significant alteration when compared to its two longer counterparts, since the putative transactivation domain is absent. This loss might render the protein incompetent for the transactivation of its target genes. However the putative multimerisation and DNA-binding domains are still intact, suggesting an alternative role for suGF1 (c). suGF1 (d) to (f) are in essence the same, containing only the DNA-binding domain and the serine-rich C-terminus and would most likely have no role in transactivation due to the absence of the putative transactivation domain.

It is interesting to speculate on the functional significance of the truncated versions of suGF1, if they are indeed expressed *in vivo*. Indeed Zeller *et al.* (1995) documented the IVT expression of five nested variants of SpGCF1 (species homologue of suGF1) from a single mRNA molecule. These five variants had molecular masses of 55, 50, 43, 40 and 37 kDa respectively. However, the purification of suGF1 from sea urchin nuclear extracts, revealed a single band of 59 kDa upon SDS-PAGE analysis, suggesting that multiple AUG start sites are not utilised *in vivo*. EMSAs performed with purified native suGF1, however did show two bands that exhibited slight differences in their respective electrophoretic mobilities (Hapgood and Patterton, 1994). The authors proposed that a difference in post-translational modification generated two proteins with a small difference in molecular mass, which were not separated by SDS-PAGE. It is, however, possible that under specific conditions or during certain stages of development, the truncated versions of suGF1 are expressed from the same gene by utilisation of multiple AUG initiation codons. Although *in vitro* these proteins could all specifically bind G-strings (Fig. 3.4 and 3.5), they might differ in their potential to transactivate their target genes and participate in homo- and / or heterodimerisation *in vivo*. In addition, the possibility exists that these protein variants are involved in a variety of unrelated functions. This would be energetically and metabolically favorable for the sea urchin organism, since it could rely on the expression of one gene to mediate a variety of cellular functions.

Previously Hapgood and Patterton (1994) provided evidence for the high-specificity binding of suGF1, present in sea urchin embryo nuclear extracts, to contiguous deoxyguanosine residues (also called G-strings). The species homologue to suGF1, SpGCF1, was shown by Zeller *et al.* (1995) to bind a similar sequence containing a $G_4.C_4$ core element. Both native SpGCF1 and IVT SpGCF1 exhibited identical bands of reduced electrophoretic mobility during EMSA. SpGCF1 and therefore also suGF1 are believed to

be involved in the regulation of developmentally regulated genes e.g. *CyIIIa* and *Endo16*. These genes contain multiple suGF1 / SpGCF1 binding sites that constitute an essential component of their modular *cis*-regulatory regions.

Results obtained during this thesis showed that IVT suGF1 can recognise the same gene sequence as native suGF1 and produce identical bands of reduced electrophoretic mobility in EMSAs. EMSAs with a radiolabeled oligodeoxyribonucleotide probe containing the consensus suGF1 binding site (S-Oligo) (Table 2.2) clearly show the presence of multiple retarded bands that represent suGF1-DNA complexes (Fig. 3.4). Since these bands are absent in the negative control lane (lane 1), one can assume these complexes are specific for the reaction of a protein (native or IVT suGF1) with the DNA probe. Lanes 2 and 3 (containing IVT suGF1) produced identical patterns of protein-DNA complex formation to lane 4 (containing native suGF1). The formation of multiple bands (B1 to B5) of reduced electrophoretic mobility is consistent with the SDS-PAGE results of IVT suGF1 that revealed multiple protein products. The band representing the full-length IVT suGF1 protein in SDS-PAGE, is most likely the one that results in the most intense band of slowest electrophoretic mobility in EMSAs. This is based on the observation that the native, 59 kDa suGF1 protein from sea urchin nuclear extracts, produced a complex in EMSAs with the same mobility as the slowest mobility complex obtained with IVT suGF1 in EMSAs (Fig. 3.4).

When the EMSAs were repeated for native and IVT suGF1 in the presence of an *EcoRI-HindIII* (E/H) fragment obtained from the H1-H4 gene battery, again multiple bands of reduced electrophoretic mobility were observed (Fig. 3.5). Competition assays using unlabeled specific (S-Oligo) and non-specific (NS-Oligo) competitor DNA (Table 2.2) showed that suGF1 recognises the labeled DNA fragment specifically. In this experiment

two important negative control reactions were also included. The reaction in which dialysis buffer (negative control for nuclear extracts) was incubated with the radiolabeled probe, produced no bands of reduced electrophoretic mobility on the gel. The incubation of rabbit reticulocyte lysate (negative control for IVT suGF1) with the radiolabeled probe, produced a diffused band(s) exhibiting reduced electrophoretic mobility through the gel matrix. The formation of this complex or possible complexes was most probably due to the presence of an endogenous lysate protein(s). This complex(es) was however found to be non-specific as neither the addition of unlabeled specific or non-specific competitor DNA could compete away the formation of the complex. Native and IVT suGF1 produced identical gel shift patterns i.e. four specific bands that represent four specific protein-DNA complexes. It is interesting to note that native suGF1 is more readily competed away for by the addition of cold S-Oligo, compared to that of IVT suGF1. It is possible that the reticulocyte lysate contains certain endogenous proteins with a weak affinity for the probe and competitor DNA. This would result in a requirement for a higher concentration of competitor DNA to decrease the amount of IVT suGF1 bound to the probe.

Apart from the fact that the EMSA results shown in Fig. 3.4 and 3.5 both generated multiple bands (implying the formation of multiple protein-DNA-complexes), the exact pattern seems not to be a perfect match. Why would incubation of suGF1 with a synthetic consensus binding as compared to a natural DNA fragment, generate distinct patterns of protein-DNA formation? Hapgood and Patterton (1994) consistently observed this intriguing difference in electrophoretic mobility. This is most probably due to the fact that the 330 bp E/H fragment has a much higher overall negative charge, compared to the 30 bp oligodeoxyribonucleotide. The oligo-suGF1 complexes are therefore not only smaller, but also exhibit less attraction towards the positive electrode of the gel and would therefore migrate slower through the gel matrix compared to the E/H fragment-suGF1 complexes.

Due to the increased mobility of the E/H fragment-suGF1 complexes the resolving power of the gel would be greater than for the oligo-suGF1 complexes and better separation of the bands would be obtained.

The results of Fig. 3.4 and 3.5 thus clearly show that both native and IVT suGF1 can specifically bind G-strings *in vitro*. Both sources of suGF1 protein can confidently be used as positive controls, when investigating the DNA-binding properties of putative suGF1 homologues or suGF1 expressed in yeast.

6.2 The suGF1 Binding Site has the Ability to Form Unusual DNA Structures

With the advent of improved phosphotriester methodology for the synthesis of oligo-deoxyribonucleotides, combined with the improved spectrophotometric techniques available, researchers found DNA to be a highly dynamic macromolecule. Indeed evidence exists that a particular DNA molecule can comprise distinct helical forms, which presumably exist in equilibrium with each other (Palecek, 1991). DNA can adopt several different conformations depending on the relationship between the primary sequence and environmental conditions such as hydration status, chemical modification and the prevalence of counterions (Kohwi and Kohwi-Shigematsu, 1991). A good example of this is homopurine.homopyrimidine DNA sequences, such as poly(dG).(dC) tracts, which normally exhibit the classical B-conformation. When the environmental salt concentration is increased or the relative humidity decreased the polynucleotide preferentially forms a triple helical structure. The normal B-DNA (Fig. 6.2 (c and d)) undergoes a transition to the A-form (Fig. 6.2 (a and b)) in which the triplex consists of a $[poly(dG).(dC)]_2$ moiety and a single polynucleotide chain, poly(dG) (Fig. 6.3 (a and b)). The polynucleotides comprising the regular A-DNA duplex are orientated in an anti-parallel fashion, and are

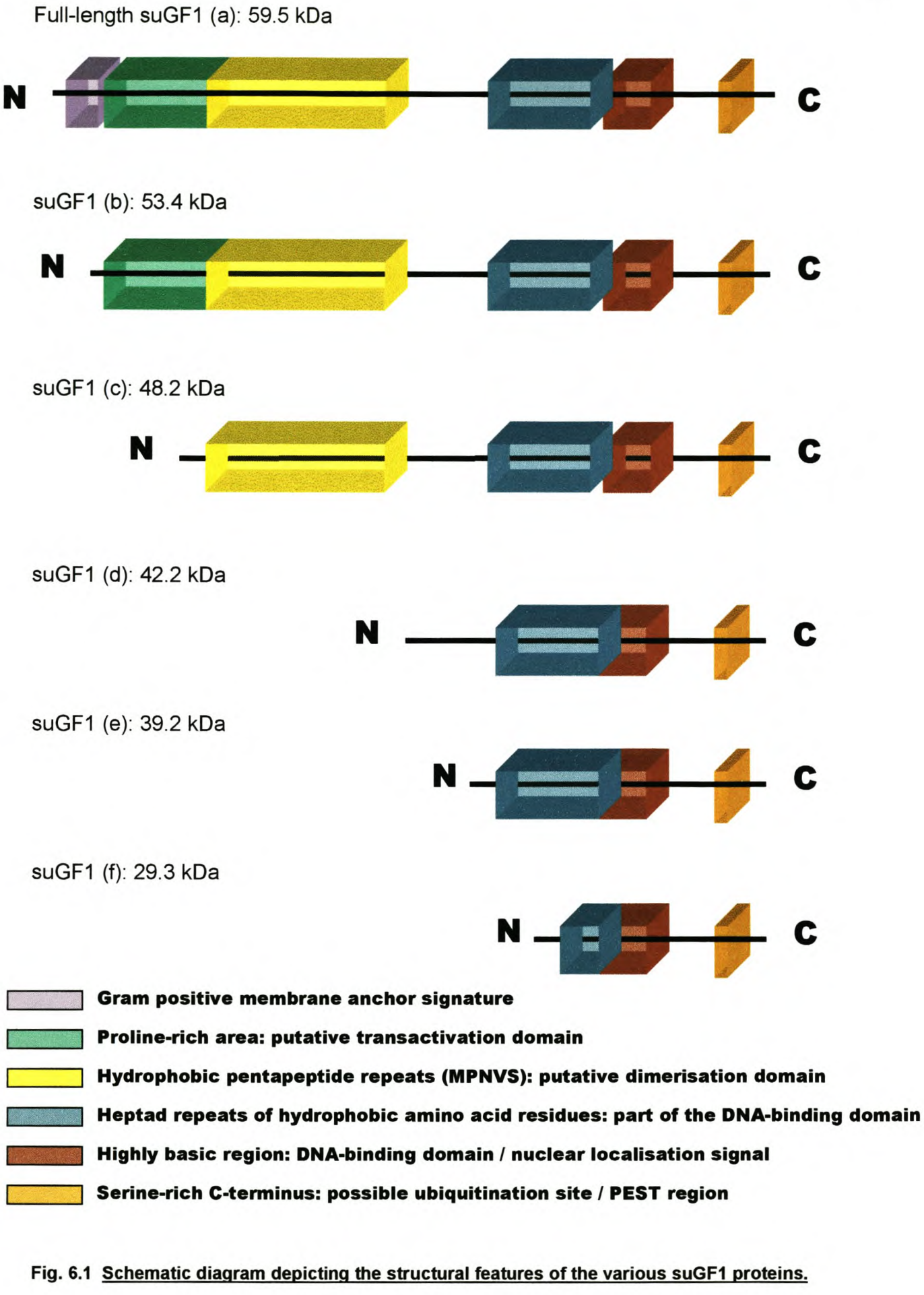


Fig. 6.1 Schematic diagram depicting the structural features of the various suGF1 proteins.

held together by classical Watson-Crick hydrogen bonds. The extra polypyrimidine strand is accommodated within the deep major groove of the A-DNA duplex and is hydrogen bonded to the polypurine (poly(dG)) strand by Hoogsteen base pairing, in which the two strands run in a parallel fashion. Recent reports suggest that triplexes do exist *in vivo* in eukaryotic cells and can influence different cellular processes such as recombination, replication and transcription (Musso *et al.*, 1998). It has been proposed that these homopurine.homopyrimidine stretches may function either to stabilise or hinder factor binding. They could therefore act as conformational switches, which are modulated by DNA-binding factors (Hobbs and Yoon, 1994; Kinniburgh *et al.*, 1994; Mayfield *et al.*, 1994; Supakar, 1997). Characteristically, these sequences are frequently nuclease sensitive *in vivo*, most likely because of disruption or displacement of nucleosomes due to binding of factors to the DNA (Kinniburgh *et al.*, 1994; Patterton and Hapgood, 1994).

Guanine quadruplexes are four stranded structures found naturally as terminating sequences at the ends of eukaryotic chromosomes (Wang and Patel, 1993). The four strands of the quadruplex associate through guanine quartets, in which each guanine uses its Watson-Crick face to hydrogen-bond to the Hoogsteen face of its neighbour (Fig. 6.3 (c and d)). Quadruplex strands may be arranged parallel or anti-parallel in several patterns, depending on the connectivity. *In vivo* the formation of these quadruplex structures is involved in replication, recombination and centromere linkage (Wang and Patel, 1993).

To investigate the potential of the suGF1 poly(dG).(dC) binding site to exist as an unusual structure under specified environmental conditions, the synthetic oligodeoxy-ribonucleotides (Table 2.2 and Appendix 2) used during EMSAs were subjected to CD analyses. Samples were dissolved in nuclease-free water (pH 7.0) before being subjected to circular polarised light in a spectropolarimeter. When considering the *in vitro* EMSA

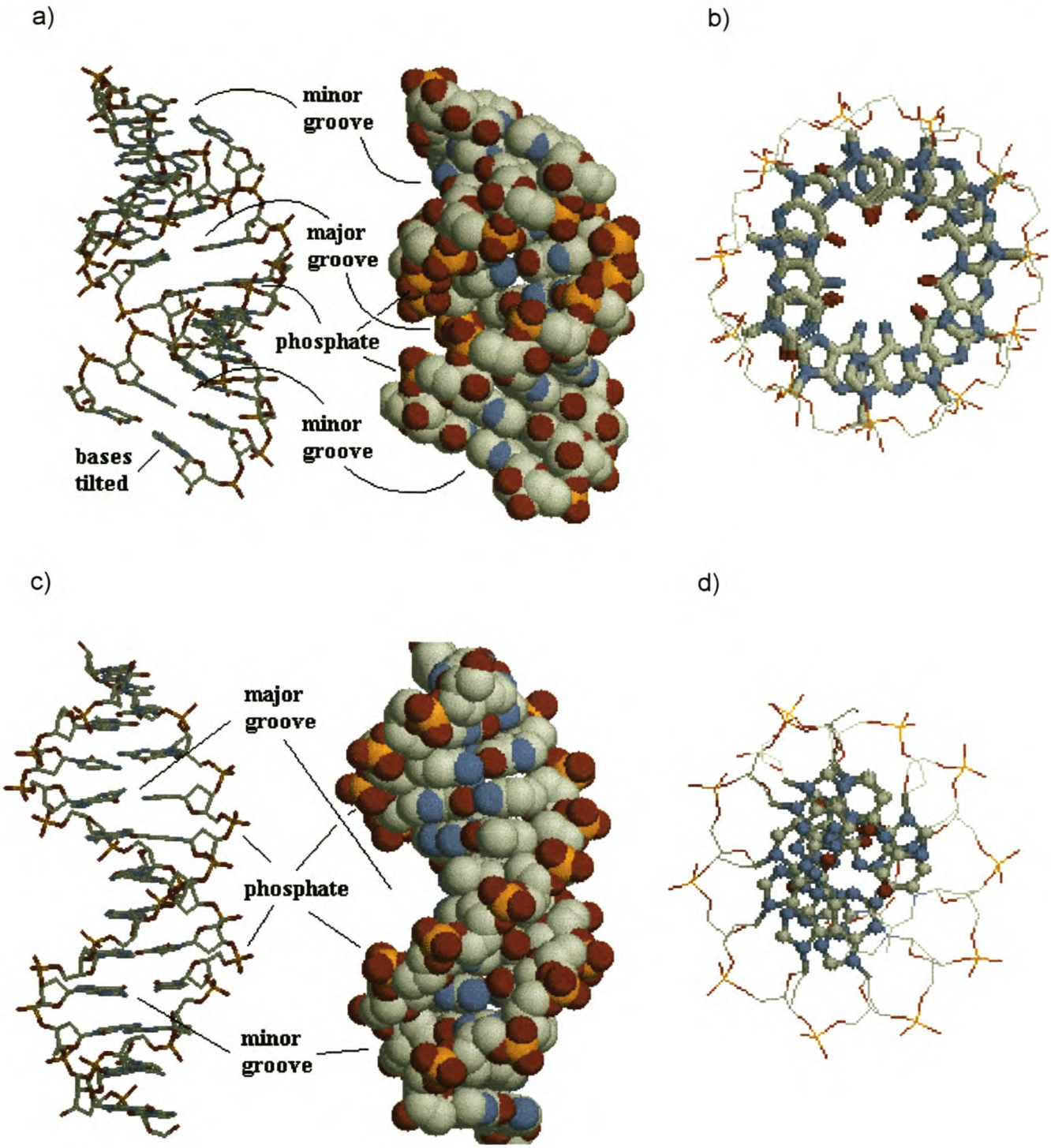


Fig. 6.2 Three-dimensional structures of classic A-DNA and B-DNA conformations.

A 3-D view of typical A-DNA (a and b) and B-DNA (c and d) as seen from the side and from the top. The major and minor grooves are indicated, as well as the relative positioning of the phosphate backbone. Taken from Nucleic Acid Architecture – http://www.chembio.uoguelph.ca/educmat/chm730.

conditions the optimal assay conditions would have been to dissolve the DNA in dialysis buffer, since this is the buffer used during EMSA (and thus the environment in which suGF1 can specifically recognise the G-string). Alternatively, a physiological buffer could have been prepared in order to imitate the *in vivo* environment of the sea urchin embryo nucleus. Unfortunately the presence of potassium, sodium and other mono- and divalent cations present in these buffer solutions, interfere with the transmission of the light wave through the sample. This would severely influence the resultant change in the elipticity of the circular polarised light, and would therefore generate inaccurate results due to high background values.

From the literature it is known that the classical CD spectrum for B-DNA obtained during circular dichroism is characterised by a positive band at higher wavelength values (maximum at 270–275 nm, zero at 247–259 nm) and a negative band at shorter wavelength values (minimum at 240–245 nm, zero at ~228 nm) (Hashizume and Imahori, 1967). Indeed, the B-DNA control included in this experiment produced a positive peak (maximum at 273 nm, zero at 251 nm) and a negative peak (minimum at 241 nm, zero at 227 nm), consistent with the data obtained from the literature. The total area of the positive peak for a B-spectrum is approximately the same as the total negative area, a feature unique to the CD spectra of B-DNA.

An A-DNA control was unfortunately not included in this experiment. However, the literature shows that this spectrum exhibits a much larger positive peak area (maximum at ~260 nm, zero at ~240 nm) and a very small negative area (minimum at 210-225 nm) (Hashizume and Imahori, 1967). Furthermore, Ikehara *et al.* (1972) showed that the CD spectra of right-handed nucleic acids typically show negative peaks in the shorter wavelength range and positive peaks in the longer wavelength range. When the reverse

spectrum is observed, i.e. negative peaks in the longer wavelength range and positive bands in the shorter wavelength range, the structure is interpreted as being left-handed (Fig. 3.7).

From the CD analysis (Fig. 3.8 and 3.9) it is clear that all the samples subjected to circular dichroism analysis exhibit positive peaks in the longer wavelength range and negative peaks in the shorter wavelength range. It can therefore be said that all these DNA samples are right-handed helices. This was expected for the control samples which have previously been shown to be right-handed helical structures. However, some GC-rich sequences belonging to the Z-form have been shown to exhibit left-handed helical orientation. It was therefore somewhat surprising to observe the right-handed helical structure obtained for the specific oligodeoxyribonucleotide (S-Oligo containing the central G_{11}-string). However, as already mentioned polypurine.polypyrimidine stretches have been shown to form triplex DNA structures with the duplex moiety of the triplex being in the A-form. The spectrum obtained for the S-Oligo exhibits a large positive peak (maximum at 273 nm, zero at 246 nm) in the higher wavelength range and a very small negative peak (minimum at 239 nm, zero at 226 nm) in the lower wavelength range. The total area of the positive peak is much larger than that of the negative peak. It therefore seems as if the S-Oligo contains a spectrum which has certain features characteristic of both the classical B-DNA (the x-axis wavelength intercepts for the positive and negative peak are almost identical to the classical B-DNA spectrum) and A-DNA (the positive peak is much larger than the negative peak). Saenger (1984) demonstrated that A-form DNA usually occurs in low ionic strength buffers and that B to A transition can occur by the addition of specific counterions. It is possible that such a transition was taking place while performing the CD analysis and that B-forms and A-forms of the S-Oligo were present at the same time, generating a CD spectrum belonging partially to the classical B-form and partially to the A-

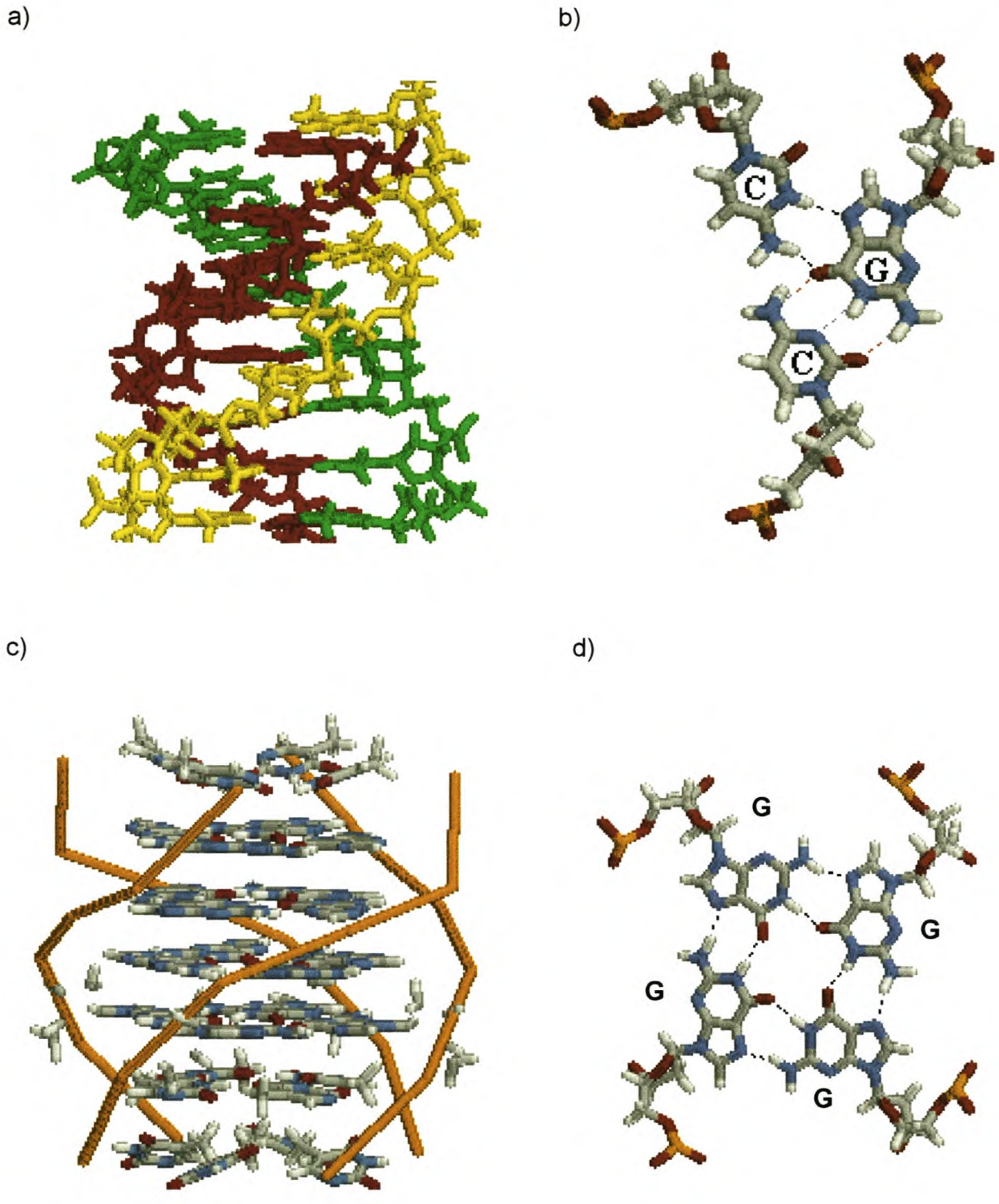


Fig. 6.3 **Three-dimensional structures of triplex and quadruplex DNA.**

A 3-D view of typical triplex (a and b) and quadruplex structured DNA molecules respectively (c and d), as seen from the side and from the top. The bases involved in the formation of these unusual DNA structures are indicated. Taken from Nucleic Acid Architecture – http://www.chembio.uoguelph.ca/educmat/chm730.

form. The CD spectrum for the control triplex DNA sample also showed similar features to that of the spectrum for the S-Oligo, suggesting that the S-Oligo might have a triple helical conformation. The duplex moiety of the triple helix would therefore most likely consist of the poly(dG).d(C) tract, while the single-stranded poly(dC) nucleotide tract fits into the major groove of the A-form duplex. The nucleotides flanking the G-string in the S-Oligo are not part of the polypurine.polypyrimidine tract and this area of the molecule might therefore exhibit a classical B-DNA form (as for the NS-Oligo), whereas the central G-string might be involved in the formation of a triple and / or quadruplicate helical structure. The change in elipticity of the circular polarised light as generated by the spectropolarimeter is a representation of the overall contents of the DNA sample, and it might be possible that different conformations of the same DNA sequence do exist in the samples analysed. This would generate CD spectra that deviate slightly from the original spectra of the individual DNA conformations.

The CD analysis for the random sequence DNA exhibited features reminiscent of a classical right-handed, B-DNA structure i.e. a positive peak at higher wavelength values and a slightly smaller (or equal) negative peak at lower wavelength values. The CD results thus indicated major structural differences between the random DNA sequence that is probably in the B-DNA conformation and the G-string oligodeoxyribonucleotide that seems to exhibit a spectrum consistent with the presence of several unusual DNA conformations in equilibrium.

It may be highly significant that the suGF1 binding site has the ability to form unusual structures such as triplexes and / or quadruplexes under certain *in vitro* conditions. An essential question that is yet to be answered, is whether these G-strings exhibit an unusual structure when suGF1 binds these sequences *in vitro* during EMSAs and *in vivo*.

The potential ability of suGF1 to specifically bind unusual DNA structures raises many interesting questions regarding eukaryotic gene regulation. An interesting implication of this would be that suGF1 may be able to discriminate between different conformations of DNA. The conformation of the DNA within a specific region may be regulated by e.g. superhelical stress, transcriptional activity in the vacinity, or the chemical environment. The ability of a specific protein to influence the expression pattern of a gene might therefore involve the protein itself as well as the conformational status of its DNA-binding site.

6.3 hORFX is not a Functional Homologue of suGF1

The apparent importance of suGF1 in developmental gene regulation steered this project into a search for a functional homologue in mammals (and more specifically humans). Exhaustive database searches showed that suGF1 / SpGCF1 exhibit no apparent sequence homology with any previously identified proteins or cDNAs from any species. This suggests that suGF1 / SpGCF1 is a novel G-string binding protein, which does not conform to any of the known families of transcription factors. This was an interesting discovery, since the suGF1 amino acid sequence exhibits features characteristic of transcription factors.

Given the relevance of mammalian models to medical science, and since no sea urchin cell lines are available yet, the identification of a mammalian functional homologue would facilitate the determination of the *in vivo* function of such a potentially important, putative, novel DNA-binding protein in mammalian cell lines. In this study sequence analysis tools were used to identify hORFX, a putative functional human homologue of suGF1. Prior to the present study, no information was available regarding the biochemical properties of

hORFX protein. Alignment of the two full-length amino acid sequences showed that they only share 15.9% global homology. However, both proteins not only contain similar domain features i.e. an N-terminal proline-rich domain (putative transactivation domain), hydrophobic amino acid repeats (putative dimerisation domain), a central region of hydrophobic heptad repeats (part of the DNA-binding domain), a highly basic region (putative DNA-binding domain) and a serine-rich C-terminus (putative PEST region), but in addition, these domains are orientated in exactly the same order within the sequence (Fig. 4.1). This prompted an investigation into the DNA-binding properties of hORFX as well as a more detailed structure prediction analysis, with a view to determining whether hORFX is a functional homologue of suGF1.

The putative DNA-binding properties of hORFX were investigated in EMSAs. Initial results indicated that hORFX does not recognise the same synthetic G-string to which suGF1 binds specifically. The autoradiograph of this experiment (Fig. 4.3) clearly shows the formation of the specific suGF1-DNA complexes, whereas in the lanes containing increasing amounts of hORFX, no specific bands of decreased electrophoretic mobility were observed. Taken together, the EMSA results obtained from incubating hORFX with a synthetic G-string oligodeoxyribonucleotide indicated that this protein can not specifically recognise the suGF1 binding site.

To investigate whether hORFX can recognise and bind specifically to a 330-bp radiolabeled E/H fragment, EMSAs were performed in the presence or absence of unlabeled competitor DNA. This fragment has been proposed to confer a natural promoter site for suGF1 and was shown in chapter 3 of this thesis to bind suGF1 specifically. The autoradiograph obtained from this experiment (Fig. 4.4) shows the formation of several indistinct protein-DNA complexes in the lanes containing hORFX. Although the full-length

hORFX has a higher molecular mass than suGF1, these complexes migrated further through the gel compared to the suGF1-DNA complexes. This change in mobility could be due to differences in overall charge, size and conformation of the complexes. If, for example, the overall charge of the protein-DNA complex is more negative, it would naturally migrate further through the gel matrix. These hORFX-DNA complexes are, however, only partially competed away for by the addition of unlabeled specific, competitor DNA. When hORFX was incubated with the radiolabeled S-Oligo (Fig.4.4) only a smear was observed in the lane containing 15 μl of IVT hORFX. It is possible that increasing the duration of electrophoresis might have separated the complex from the free probe more effectively, or that longer exposure of the film might have revealed the presence of slow-migrating bands of low intensity. The non-specific complexes are also incompletely competed away for by the addition of unlabeled non-specific competitor DNA, suggesting that hORFX binds DNA non-specifically. The affinity of hORFX for the random sequence appears to be higher than for the G-string oligodeoxyribonucleotide, as more competition is observed in the lane containing the unlabeled non-specific competitor DNA. The relative abundance of free-labeled probe in the lanes containing the hORFX protein, compared to that observed in the lane containing the positive control, native suGF1, indicates the absence of any factors that occupy the probe. Since IVT hORFX is most probably present at a much higher concentration in the lysate than suGF1 in the nuclear extracts, one would expect that a sufficient amount of hORFX is present to form a complex with the probe if hORFX was indeed a G-string binding factor. The EMSA results imply that hORFX does not bind G-strings specifically. Due to the fact that the hORFX protein is expressed in rabbit reticulocyte lysate, the protein might not undergo all the exact post-translational modifications that are necessary for G-string-binding activity. It might therefore be possible that *in vivo* or *in vitro* under different experimental conditions, hORFX has the ability to

specifically bind G-strings. However, if this protein was a functional homologue of suGF1, this would be unlikely, since IVT suGF1 does bind G-strings specifically.

Transcription factors often require the presence of divalent cations (e.g. Zinc-finger binding proteins) to bind to DNA (Bossone *et al.,* 1992). To test whether hORFX needs divalent cations to specifically bind G-strings, EMSAs were performed in the presence of increasing concentrations of $ZnCl_2$. The autoradiograph obtained from this experiment shown in Fig. 4.5, indicates the formation of at least two bands (in the lanes containing 10 μl IVT hORFX) that exhibit reduced electrophoretic mobility through the gel matrix, similar to the result shown in Fig. 4.4. These complexes have previously been shown to be non-specific. The significant aspect of this experiment was that the absence or presence of $ZnCl_2$ had no effect on the G-string binding properties of hORFX. It is possible that hORFX has an essential requirement for some other divalent cations e.g. Co^{2+} or Ca^{2+}, before attaining the ability to specifically bind G-strings. suGF1, however, binds specifically to G-strings in the absence of any divalent cations. If indeed hORFX needed a specific divalent cation to bind DNA, this would signify a major difference in the functional attributes of this protein and those of suGF1.

The interpretation of the *in vitro* EMSAs could be misleading as these experiments are performed *in vitro*. However, when considering that IVT suGF1 was shown to specifically bind a sequence containing contiguous deoxyguanosine residues and hORFX (expressed in exactly the same system) failed to generate similar complexes, it can be deduced from this investigation that hORFX is most probably not a functional homologue of suGF1.

6.4 Sequence Analysis and Structure Prediction

Two factors prompted a detailed sequence and structural analysis of suGF1 using Bioinformatics tools. Firstly, several interesting domains and apparently unique features of suGF1 spurred the analysis with a view to obtaining more information about the potential function of this protein. Secondly, a search for a mammalian homologue was required, due to a desire to obtain a protein, the function of which could be more easily investigated in mammalian cell lines in the future. In addition, cross species searches could potentially reveal important information on evolutionary conservation and function.

When incubating suGF1 present in sea urchin embryo nuclear extracts with various probes containing a sequence of contiguous deoxyguanosine residues, a classic pattern of suGF1-DNA complexes is observed. This means that under the experimental conditions for these assays, the only factor specifically recognising the G-strings is suGF1, as no other bands of decreased electrophoretic mobility are observed. It is therefore quite possible that in sea urchins only one G-string binding factor exists. Table 1.1 clearly indicates the presence of at least fourteen mammalian GC-box binding proteins, a marked difference compared to the situation in sea urchins. The apparent absence of an Sp1-like factor in sea urchins and the apparent absence of an suGF1-like factor in mammals suggests that factors binding to GC-rich DNA are species-specific i.e. that they have evolved to meet the developmental requirements particular to the developmental patterns of the organism. It is curious why so many different factors that recognise the same *cis* elements have evolved within the same species (e.g. Sp1 and the Sp1-like factors).

To investigate the possibility that a mammalian or human protein, exhibiting structural characteristics similar to suGF1 exists, the suGF1 amino acid sequence was subjected to intensive domain and structural feature analysis. Special protein analysis tools were used

to identify hORFX, a putative functional homologue to suGF1. hORFX was shown by EMSAs not to exhibit similar DNA-binding properties to suGF1. This result was quite surprising when considering the features and orientation of functional domains in the hORFX protein sequence (Fig. 4.1), the most prominent being the highly basic domains shared by both proteins. Following this finding both proteins were subjected to an in depth sequence and structure-function prediction analysis, to examine possible differences in primary, secondary and tertiary structural features.

suGF1 was found to contain a proline-rich region, pentapeptide repeats and a highly basic region. These domains suggest that suGF1 is involved in transcriptional regulation since they have been found by others to be involved in transactivation, protein-protein interaction and DNA-binding functions, respectively, in other proteins. Furthermore function-prediction programs revealed that suGF1 contains a putative nuclear localisation domain (within the basic region), as well as a Gram-positive membrane anchor signature and a C-terminal containing a PEST-region. Sequence analysis for the hORFX protein showed a potentially significant resemblance to that of suGF1. hORFX contains two proline-rich regions, a hydrophobic repeat sequence, a highly basic region and a possible C-terminal PEST region (serine-rich). Additionally hORFX also contains two bromodomains that are not present in the suGF1 protein. These bromodomains have been implicated in protein-protein interactions (Beck *et al.,* 1992). Not only do these two proteins share similar domain features, but these domains are also ordered in exactly the same pattern within the amino acid sequence (Fig. 4.1). These results, given that no other candidate proteins with an apparently higher degree of similarity could be detected in the searches, prompted further theoretical comparisons of their primary, secondary and tertiary structures, with a view to understanding the differences in their respective DNA-binding properties obtained by EMSAs.

The smoothed hydropathy patterns for the suGF1 and hORFX basic domains were generated to compare the overall hydrophobicity of these regions. Both regions containing predominantly positively charged lysine and arginine amino acid residues, exhibit a significant decline in hydrophobicity compared to the surrounding regions. The hydrophilic trough extends for the duration of the basic region and reaches similar minimum values for both proteins. When considering the overall positive charge of these regions it appeared likely that both sequences would posses the ability to bind DNA. EMSAs, however, showed that hORFX does not bind specifically to G-strings, but does bind non-specifically to DNA, probably due to the abundance of positive charges in the basic region. Specific binding to DNA is, however, usually not dependent on the presence of basic amino acid residues, but involves unique patterns of hydrogen bond donor and acceptor groups in the major groove of the DNA (Johnson and McKnight, 1989). An explanation for the difference in DNA-binding properties was sought by means of more detailed sequence and structure prediction analysis.

suGF1 has no sequence homologues in the databases available up to date, making structural analysis and prediction very difficult. hORFX, on the other hand, exhibits significant amino acid homology to the RING3 and *Drosophila fsh* proteins (none of which has been structurally defined), suggesting ORFX to be involved in developmental processes. These proteins all contain one or more bromodomains, which have been implicated in protein-protein interactions (Beck *et al.*, 1992). Structure-function predictions for hORFX suggest this protein to be a nucleoporin protein, involved in transport across the nuclear membrane. This would be consistent with the secondary-structure predictions, which show a repetition of α-helical bundles, as well as a positively charged C-terminal tail often implicated in membrane transport.

The problem therefore was that neither suGF1 nor hORFX could be subjected to classical methods of structure prediction e.g. homology modeling, due to their unique sequence features. The comparison and prediction of structure and function was therefore purely based on non-homology methods, except for the tertiary structure prediction which was based on fold recognition. The problem of predicting protein structure from sequence only, remains fundamentally unsolved despite more than three decades of intensive research efforts. However, new and promising methods in 3D, 2D, and 1D prediction have reopened the field and might shed some light on the structural features of suGF1 and hORFX. This theoretical modelling has been driven by the belief that the 3D structure of a protein is primarily determined by its amino acid sequence (Anfinsen, 1973). While it is now known that chaperones often play a role in the folding pathway, and in correcting misfolds (Corrales and Fersht, 1996, Hartl et al., 1994), it is believed that the final structure is at the free-energy minimum of the molecule. Furthermore, *in vivo,* native polypeptides undergo a series of post-translational modifications (e.g. phosphorylation and glycosylation) before actually acquiring fully functional status. Thus, in essence all the information needed to predict the native structure of a protein is contained in the amino acid sequence, but also requires knowledge of its native solution environment and possible post-translational modifications. It is however possible to gain some insight into the structural features of a specific protein when examining only the amino acid sequence, as this manifests the native model of the polypeptide. suGF1 and hORFX displayed major differences within secondary as well as tertiary structure when the predictions for their respective basic regions were compared.

Due to the physicochemical similarities between the basic regions present in the suGF1 and hORFX proteins, as well as the potential, functional importance of these domains in

interacting with DNA, these regions were subjected to secondary structure prediction. The method used for secondary structure prediction is based on the probable placement of secondary structural elements along the entire length of any given amino acid sequence, and is based purely on the chemical composition and order of the monomeric sub-units. The analysis algorithm is based on probabilistic Discrete State-space Models (DSMs), optimal filtering and smoothing algorithms as described by Stultz *et al.* (1993). The mathematical basis for the models and algorithms were determined and investigated by White *et al.* (1994). To use the PSA program, a single amino acid sequence is submitted to the server, which may be instructed to analyse the sequence in one of three ways: using Type-1, Type-2, or WD-repeat DSMs (Table 4.1). DSMs (Discrete State-space Models) define the parameters for patterns of alpha helices, strands, tight turns, and loops in specific structural classes. The basic regions present in the suGF1 and hORFX proteins were subjected to Type-1 analysis. The results for this analysis (Fig. 4.7 and 4.9) clearly illustrated major structural differences between these regions, implying different DNA-binding capacities. The relative abundance of contours preceding the suGF1 basic region (residues 320–330), in the buried and exposed helical state suggests this area to be an alpha helix. Immediately following this alpha helix, the contours reposition and are more abundant within the turn-state, suggesting this region to be a turn-like structure, after which the contours again accumulate solely in the helix state, suggesting the last section of the basic region to be in a helical conformation. Residues 332–350, constituting the suGF1 DNA-binding domain (suGF1 DBD) are therefore predicted to have a helix-(irregular turn / β-turn)$_2$-helix structure. Due to the electrostatic repulsion of positive charges (Arg and Lys residues indicated in red) the domain is most likely exposed and capable of interacting with DNA. The abundance of the positive charges might induce the formation of the irregular turn / β-turn, which protrudes from the rest of the molecule and exposes the positive charges to the exterior. This result is consistent with the fact that

suGF1 is a DNA-binding protein *in vitro* and substantiates the predicted secondary structure for the DBD as illustrated in Fig. 4.7. Interesting was the fact that the hORFX basic region seemed to exhibit a helix-loop-helix conformation, a motif that is also characteristic of certain transcription factors e.g. the basic leucine zipper DNA-binding proteins (Brownlie *et al.*, 1997).

The tertiary structure prediction (using the method of threading or fold recognition) of the suGF1 basic region was consistent with the results obtained for the secondary structure prediction. The basic region of suGF1 aligned with high percentage similarity to two entries in the databases for which a specific fold has been registered to the fold library. Both these entries were DNA-binding proteins belonging to the helix-turn-helix DNA-binding domain family. The suGF1 basic region exhibited 18% identity and 61% similarity to the basic region of the murine Ets-1 DNA-binding protein (PDB code 1etc), which belongs to the ETS family of transcription factors. This domain is characterised by a helix-turn-helix motif on a four-stranded anti-parallel β-sheet, conforming into a classical winged helix structure, which gives the protein its ability to specifically recognise the DNA-binding site (Fig. 4.12). This domain has been conserved through evolution through many species, and it might be possible that this conservation was maintained from sea urchins to humans, when considering the fold recognition study (Donaldson *et al.*, 1996). The second high confidence alignment generated from threading with the suGF1 basic region showed an even higher confidence level. The suGF1 basic region displayed 26% identity and 82% similarity to the basic region of the Mu end DNA-binding Iβ subdomain of phage Mu transposase (PDB code 2ezk). This enzyme binds to the ends of the Mu genome during assembly of higher order nucleoporin complexes. This facilitates the movement of defined segments of DNA (transposons) to distant locations within the genome. Interestingly, the Iα subdomain of this protein belongs to the winged HTH family (similar to

Ets-1), whereas the Iβ domain showed features characteristic of homeodomain HTH DNA-binding proteins. It comprises five α-helices, including the HTH motif (formed by helices 3 and 4), with the DNA recognition helix protruding from a disc-like structure (Fig. 4.14). The structural features of the basic regions of these two proteins are consistent with the secondary structure predictions for the suGF1 basic region i.e. a turn-like structure (recognition domain) that protrudes form bundles of α-helices extending out from both sides (Schumacher *et al.*, 1997). These results imply that suGF1 belongs to the HTH class of DNA-binding proteins and that this protein exhibits structural features reminiscent of previously documented DNA-binding proteins. It is possible that this region belongs to a novel subfamily of HTH motifs, as the two HTH-proteins exhibiting high confidence alignments with this region seemed to belong to distinct groups of this family. Taken together, the predictions from secondary and tertiary analysis of suGF1 are consistent with the experimental data and support the proposed role of suGF1 *in vivo*, as a DNA-binding transcription factor.

The structure prediction for the hORFX basic region (Fig. 4.9) by the PSA server clearly displayed significant structural differences in comparison to that of the suGF1 basic region. The relative abundance of contours, preceding the start to the hORFX basic region, in the buried and exposed helical states, implies that this area conforms into an alpha-helical structure. Immediately following this alpha helix, the contours reposition and are more abundant within the loop state, suggesting this region to be a loop-like structure, after which the contours again accumulate solely in the helix state, implying that the last section of the basic domain is an alpha helix. Taken together, the PSA structure prediction indicates that the basic region of the hORFX protein is a helix-loop-helix domain and is different to that of the helix-(irregular turn / β-turn)$_2$-helix predicted for the suGF1 basic region. The scattered presence of negatively charged residues in between the positive

residues might induce a closed and buried loop-like structure within the tertiary assembly of the molecule, concealing and also diminishing the net positive charge of the region, due to the electrostatic attraction between positive and negative residues. This conformation might still be able to bind DNA non-specifically via ionic interactions. However it appears to exhibit major structural differences to the suGF1 basic region that might render hORFX incapable of binding specifically to G-strings.

Again the tertiary structure prediction for the hORFX domain seemed to be consistent with the PSA secondary structure prediction results, as both methods predicted this region to be a helix-loop-helix domain. The basic region of hORFX aligned with high percentage similarity to one entry in the databases for which a specific fold has been registered to the fold library. The hORFX basic region displayed 7% identity and 93% similarity to the basic region of the human Max protein. Max belongs to the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors (PDB code 1hloA). Interestingly the hORFX contain two bromodomains which have been implicated in protein-protein interactions, similar to the Max protein that was shown to form homodimers and heterodimers via a leucine zipper motif and an unidentified dimerisation domain. In conjunction with various other regulators of gene expression e.g. Myc and Mxi1, this protein recognises the classical E-box promoter element to control different modes of transcription (Brownlie *et al.*, 1996). It might therefore be possible that hORFX indeed plays a role in gene regulation via binding to DNA. However, it may only specifically recognise a site different to that of the G-strings, whilst it may also have some non-specific DNA-binding properties. This would be consistent with the EMSA results in this thesis, showing that hORFX can non-specifically recognise the E/H fragment as well as the S-Oligo and NS-Oligo (Fig. 4.4), possibly via its predicted helix-loop-helix domain.

The sequence analysis and structure predictions for the respective basic regions of suGF1 and hORFX, as well as for the full-length amino acid sequence, indicated that there are similarities and differences between these two proteins. The different methods of theoretical prediction produced data that substantiate each other as well as the experimental data obtained within the scope of this thesis. Theoretical predictions suggest that suGF1 has a DNA-binding domain belonging to a different family to that predicted for hORFX, suggesting differences in DNA-binding specificity. These theoretical predictions support the experimental results obtained in this project, that hORFX is not a functional homologue to suGF1.

6.5 suGF1 is Expressed in Yeast and Exhibits Similar DNA-Binding Properties to Native and IVT suGF1

For the preparation of yeast nuclear extracts containing recombinantly expressed suGF1, a protease deficient strain Y294 was transformed with an suGF1 expression construct. The suGF1-expression construct (pYES2-suGF1) contained a galactose-inducible marker gene that was essential for the specific expression of suGF1 in yeast cells.

When grown in selective media containing D-galactose as main sugar source, the suGF1 cDNA was readily expressed from the Gal1 promoter in the protease-deficient strain, as shown in Fig. 5.2. The lane containing yeast cells transformed with the expression construct verified the integrity of multiple unique bands that are comparable in size to the products obtained during *in vitro* transcription and translation of the suGF1 protein (Fig. 3.1). As the full-length suGF1 migrated a little faster through the gel matrix compared to the BSA marker proteins (66 kDa) one can assume this protein product to have a molecular mass of approximately 60 kDa. This is consistent with the native, full-length

suGF1, IVT and *in silico* expressed suGF1 which have been shown to have a molecular mass of 58 to 60 kDa. The two other visible, unique bands in this lane were estimated to have molecular masses of 50 and 46 kDa respectively, compared to the marker proteins. The presence of various other background bands made it difficult to estimate whether low levels of the other suGF1 truncations are also present on the gel. The bands present in the lanes containing yeast cells transformed with only the vector (pYES2) also exhibit these background bands and represent endogenous yeast proteins that are expressed at relatively high levels under these experimental conditions. It was however significant that unique bands, representing the full-length suGF1 and truncations thereof, were observed in only the lane containing nuclear extracts from yeast cells, transformed with the suGF1 expression construct. Also intriguing is the observation that again suGF1 seemed to be expressed as multiple protein products from a single gene sequence. This is consistent with the literature that documented SpGCF1 (suGF1 species homologue) to be expressed as five nested variants from a single mRNA molecule (Zeller *et al.,* 1995a). Furthermore, experimental results obtained within the context of this research project i.e. *in vitro* transcription-translation and *in silico* expression of the suGF1 cDNA, suggested suGF1 to be expressed as multiple truncated protein products by the utilisation of multiple AUG translation start sites.

The SDS-PAGE results of the nuclear extracts from yeast transformed with the suGF1 expression construct verified the presence of a recombinantly expressed protein. Although this was an encouraging result, no evidence supporting the presence of a fully functional suGF1 protein could be gathered from this. Subsequently these extracts were subjected to EMSAs to investigate the putative G-string binding properties of the recombinantly expressed suGF1. Indeed suGF1 produced similar specific interactions with the synthetic, radiolabeled G-string probes, suggesting that this protein retained its

ability to bind to G-strings even when expressed within a heterologous environment (Fig.5.3). The lanes containing nuclear extracts from yeast cells transformed with the suGF1 expression construct produced three visible bands of decreased electrophoretic mobility, suggesting the presence of three unique suGF1-DNA complexes. These complexes were partially competed away for by the addition of unlabeled specific competitor DNA, verifying the specificity of binding. Furthermore the lanes containing nuclear extracts from yeast cells transformed with only the pYES2 construct (no suGF1 cDNA) produced no specific protein-DNA complexes, showing that the observed complexes in the lanes containing recombinant suGF1 were specific for the yeast cells transformed with the suGF1 expression construct.

Initially yeast cells that were not protease deficient were used to prepare whole cell extracts, after transformation with the same expression construct. However, the lanes containing either whole cell extracts from yeast cells transformed with the vector only or the expression construct appeared to be highly degraded on the SDS-PAGE gel, suggesting rapid proteolytic digestion of the suGF1 and endogenous yeast proteins (Fig. 5.2). The rapid proteolysis of suGF1 is consistent with the sequence prediction that the suGF1 C-terminal might constitute a putative PEST region (similar to hORFX), which would make it highly susceptible to ubiquitination and other tagging mechanisms, which would destine the factor for rapid digestion. As expected the lanes containing whole cell extracts from yeast cells transformed with either pYES2 or pYES2-suGF1 produced no protein-DNA complexes in the EMSA, supporting the idea that these extracts have undergone extensive proteolytic digestion, due to the fact that the extracts were prepared from yeast cells containing endogenous proteolytic pathways (unlike the protease-deficient strain used for nuclear extract preparation).

It can therefore be concluded that suGF1 was indeed expressed in the protease-deficient yeast cells and exhibits similar DNA-binding properties to native and IVT suGF1. An intriguing observation from the EMSAs of yeast extracts was the relative patterning of the recombinant suGF1-DNA complexes compared to the complexes containing native or IVT suGF1 respectively. Three specific complexes were observed for the recombinant suGF1, compared to the five complexes obtained for the other two protein sources. In addition the slowest recombinant suGF1-DNA complex appeared to be absent in the lanes containing native or IVT suGF1. It is therefore possible that in the yeast cell, suGF1 is predominantly expressed from three preferentially utilised AUG translational start sites, compared to the five or more sites utilised during expression of the IVT suGF1. The yeast ribosomal scanning mechanism might therefore only recognise three start sites for translation, resulting in the production of only three suGF1 protein products. This is consistent with the SDS-PAGE results for the yeast nuclear extracts that also showed three unique suGF1 bands. Notable is that the second retarded band (representing a recombinant suGF1-DNA complex), is almost as intense as the bands for the positive control lanes, and is far more prominent than the two other complexes. It is therefore likely that in yeast the second AUG translational start site is preferentially used to initiate translation, which will ultimately mean that the second suGF1 truncation is the main protein product present in yeast. Also possible, however, is the partial degradation of the nuclear extracts, resulting in the truncation of the full-length protein. If this is the case these truncations still retained the ability to specifically interact with G-strings, but might exhibit increased mobility in complex with the G-string. The post-translational modification of suGF1 in the yeast cell might also be different from the *in vivo* situation or the lysate environment. This could lead to a full-length protein of altered molecular mass, since phosphorylation, glycosylation and other post-translational modifications can drastically influence the overall mass and three-dimensional conformation of the expressed protein.

It seems nevertheless as if suGF1 was successfully expressed in the yeast system and the recombinant protein is able to bind G-strings specifically. These *in vitro* results are crucial for the ultimate outcome and correct interpretation of future transactivation experiments and lay the foundation for further investigation into the possible role of suGF1 in transcriptional regulation.

6.6 Future Perspectives

The results obtained during this research project support the hypothesis that suGF1 is a transcription factor. The recombinant expression of suGF1 in yeast laid the foundation for future transactivation assays, which might elucidate the role of suGF1 *in vivo*. The future experiments that need to performed are therefore as follows:

1) Transactivation assays to determine the transactivation potential of suGF1.
2) Further database searches for an suGF1 homologue to establish whether suGF1 is indeed a novel G-string binding protein. The identification of a mammalian functional homologue would facilitate determination of the *in vivo* function of such a potentially important protein.
3) Yeast two hybrid and pull-down assays to identify possible protein-protein interactions.
4) The determination of the 3-D structure of suGF1, using crystallographic or NMR techniques. The determination of the structure for an suGF1-G-string complex, would also yield important information, especially if suGF1 is a novel helix-turn-helix G-string binding factor, that specifically recognises unusual DNA structures.
5) Construction of shortened suGF1 cDNA expression constructs to investigate the expression and DNA-binding properties of the various truncated protein products.

This would include transactivation and protein-protein interaction assays to establish the difference in function between the truncations.

References

Akasaka K, Frudakis TN, Killian CE, George NC, Yamasu K, Khaner O and Wilt FH, 1994. Genomic organisation of a gene encoding the spicule matrix protein SM30 in the sea urchin *Strongylocentrotus purpuratus*. *J Biol Chem* **269**: 20592-20598.

Al-Asadi R, Yi EC and Merchant JL, 1995. Sp1 affinity for GC-rich elements correlates with ornithine decarboxylase activity. *Biochem Biophys Res Commun* **214**: 324-330.

Anfinsen CB, 1973. The formation and stabilisation of protein structure. *Biochem J* **128**: 737-749.

Arcot SS and Deininger PL, 1992. Protein binding sites within the human thymidine kinase promoter. *Gene* **111**: 249-254.

Asundi VK, Keister BF and Carey DJ, 1998. Organisation of the 5'-flanking sequence and promoter activity of the rat *GPC1* gene. *Gene* **206**: 255-261.

Augustin LB, Felsheim RF, Min BH, Fuchs SM, Fuchs JA and Loh HH, 1995. Genomic structure of the mouse delta opoid receptor gene. *Biochem Biophys Res Commun* **207**: 111-119.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seideman JG, Smith JA and Struhl K, 1987. Current protocols in molecular biology. *John Wiley and Sons*, New York.

Baron K, 1997. Transcriptional control of globin gene switching during vertebrate development. *Biochim Biophys Acta* **1351**: 51-72.

Barton MC, Madani N and Emerson BM, 1993. The erythroid protein cGATA-1 functions with a stage-specific factor to activate transcription of chromatin-assembled β-globin genes. *Genes Dev* **7**: 1796-1809.

Baumann M, Feederle R, Kremmer E and Hammerschmidt W, 1999. Cellular transcription factors recruit viral replication proteins to activate the Epstein-Barr virus origin of lytic DNA replication, oriLyt. *EMBO J* **18**: 6095-6105.

Beck KM, Seekamp AH, Askew GR, Mei Z, Farrell CM, Wank S and Lukens LN, 1991. Association of a change in the chromatin structure with a tissue-specific switch in transcription start sites in the *a*2(1) collagen gene. *Nucleic Acids Res* **19**: 4975-4982.

Beck S, Hanson I, Kelly A, Pappin DJ and Trowsdale J, 1992. A homologue of the Drosophila female sterile homeotic (fsh) gene in the class II region of the human MHC. *DNA Seq* **2:** 203-210.

Beverly ME, Emerson BM, Lewis CD and Felsenfeld G, 1985. Interaction of specific nuclear factors with the nuclease-hypersensitive region of the chicken adult β-globin gene: nature of the binding domain. *Cell* **41**: 21-30.

Bird AP and Wolffe AP, 1999. Methylation-induced repression-belts, braces and chromatin. *Cell* **99**: 451-454.

Birnbaum MJ, Wijnen AJ, Odgren PR, Last TJ, Suske G, Stein GS and Stein JL, 1995. Sp1 trans-activation of cell cycle regulated promoters is selectively repressed by Sp3. *Biochemistry* **34**: 16503-16508.

Birnbaum MJ, Wright KL, Wijnen AJ, Ramsey-Ewing AL, Bourke MT, Last TJ, Aziz F, Frenkel B, Rao BR and Aronin N, 1995. Functional role for Sp1 in the transcriptional amplification of a cell cycle regulated histone H4 gene. *Biochemistry* **34**: 7648-7658.

Birnboim B and Doly AC, 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res* **7:** 1513 -1523.

Blake MC, Jambou RC, Swick AG, Kahn JW and Azizkhan JC, 1990. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol Cell Biol* **10**: 6632-6641.

Bossone SA, Asselin C, Patel AM and Marcu KB, 1992. MAZ, a Zinc finger protein, binds to c-Myc and C2 gene sequences regulating transcriptional initiation and termination. *Proc Natl Acad Sci USA* **89**: 7452-7456.

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A and Cedar H, 1994. Sp1 elements protect a CpG island from *de novo* methylation. *Nature* **371**: 435-438.

Brown A, Browes C, Mitchell M and Montano X, 2000. c-abl is involved in the association of p53 and trk A. *Oncogene* **15**: 3032-3040.

Brownlie P, Ceska T, Lamers M, Romier C, Stier G, Teo H and Suck D, 1997. The crystal structure of an intact human Max-DNA complex: new insights into mechanisms of transcriptional control. *Structure* **15:** 509-520.

Chang DC, Gao PQ and Maxwell BL, 1991. High efficiency gene transfection by electroporation using a radio-frequency electric field. *Biochim Biophys Acta* **1092:** 153-160.

Chen A and Davis BH, 1999. UV irradiation activates JNK and increases αI(I) collagen gene expression in rat hepatic stellate cells. *J Biol Chem* **274**: 158-164.

Chen J, Spector MS, Kunos G and Gao B, 1997. Sp1-mediated transcriptional activation from the dominant promoter of the rat α1B adrenergic receptor gene in DDT1MF-2 cells. *J Biol Chem* **272**: 23144-23150.

Chiou-Hwa Y, Bolouri H and Davidson EH, 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**: 1896-1902.

Chu S, Blaisdell CJ, Liu MM and Zeitlin PL, 1999. Perinatal regulation of the ClC-2 chloride channel in lung is mediated by Sp1 and Sp3. *Am J Physiol* **276** (*Lung Cell Mol Physiol* 20): L614-L624.

Clare SE, Fantz DA, Kistler WS and Kistler MK, 1997. The testis-specific histone *H1t* gene is strongly repressed by a G/C-rich region just downstream of the TATA box. *J Biol Chem* **272**: 33028-33036.

Clark SP, Lewis CD and Felsenfeld G, 1990. Properties of BGP1, a poly(dG)-binding protein from chicken erythrocytes. *Nucleic Acids Res* **18**: 5119-5126.

Cook J,Gebelein B, Mesa K, Mladek A and Urrutia R, 1998. Molecular cloning and characterisation of TIEG2 reveals a new subfamily of transforming growth factor-β-inducible Sp1-like zinc finger-encoding genes involved in the regulation of cell growth. *J Biol Chem* **273**: 25929-25936.

Corrales FJ and Fersht AR, 1995. The folding of GroEL-bound barnase as a model for chaperonin-mediated protein folding. *Proc Natl Acad Sci U S A* **92:** 5326-5330.

Courey AJ and Tjian R, 1992. In *Transcriptional Regulation.* McKnight, S.L and K.R. Yamamoto, eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY): 743-749.

Crosby SD, Puetz JJ, Simburger KS, Fahrner TJ and Milbrandt J, 1991. The early response gene NGFI-C encodes a zinc-finger transcriptional activator and is a member of the GCGGGGGCG (GSG) element-binding protein family. *Mol Cell Biol* **11**: 3835-3841.

Dailey L, Hanly SM, Roeder RG and Heintz N, 1986. Distinct transcription factors bind specifically to two regions of the human histone H4 promoter. *Proc Natl Acad Sci USA* **83**: 7241-7245.

Denver RJ, Ouellet L, Furling D, Kobayashi A, Fujii-Kuriyama Y and Puymirat J, 1999. Basic transcription element-binding protein (BTEB) is a thyroid hormone-regulated gene in the developing central nervous system. *J Biol Chem* **274**: 23128-23134.

Donaldson LW, Petersen JM, Graves BJ and McIntosh LP, 1996. Solution structure of the ETS domain from murine Ets-1: a winged helix-turn-helix DNA binding motif. *EMBO J* **15:** 125-134.

Eden S, Ottosson M, Lonnroth P and Bjorntorp P, 1999. DNA methylation represses transcription *in vivo*. *Nature Genet* **22**: 203-206.

Elgin SRC, 1995. The formation and function of DNase I hypersensitive sites in the process of gene activation. *J Biol Chem* **263**: 19259-19262.

Emerson BM, Nickol JM and Fong TC, 1989. Erythroid-specific activation and derepression of the chicken beta globin promoter *in vitro*. *Cell* **57**: 1189-1200.

Felsenfeld G, Boyes J, Chung J, Clark D and Studitsky V, 1996. Chromatin structure and gene expression. *Proc Natl Acad Sci USA* **93**: 9384-9388.

Franks RR, Anderson R, Moore JG, Hough-Evans BR, Britten RJ and Davidson EH, 1990. Competitive titration in living sea urchin embryos of regulatory factors required for expression of the *CyIIIa* actin gene. *Development* **110**: 31-40.

Fuminori H, Tanaka H, Hirano Y, Hiramoto M, Handa H, Makino I and Scheidereit C, 1998. Functional interference of Sp1 and NF-κB through the same DNA binding site. *Mol Cell Biol* **18**: 1266-1274.

Ginder GD, Singal R, Little JA, Dempsey N, Ferris R and Wang SZ, 1998. Silencing and activation of embryonic globin gene expression. *Ann N Y Acad Sci* **850**: 70-79.

Haas TL, Stitelman D, Davis SJ, Apte SS and Madri JA, 1999. Egr-1 mediates extracellular matrix-driven transcription of membrane type 1 matrix metalloproteinase in endothelium. *J Biol Chem* **274**: 22679-22685.

Hagen G, Muller S, Beato M and Suske G, 1992. Cloning by recognition site screening of teo novel GT box binding proteins: a family of Sp1 related genes. *Nucleic Acids Res* **20**: 5519-5525.

Hapgood J and Patterton D, 1994. Purification of an oligo(dG)·oligo(dC)-binding sea urchin nuclear protein, suGF1: a family of G-string factors involved in gene regulation during development. *Mol Cell Biol* **14**: 1402-1409.

Hartl FU, 1994. Protein folding. Secrets of a double-doughnut. *Nature* **371:** 557-559.

Hasegawa T, Zhou X, Garrett LA, Ruteshauser EC, Maity SN and Decrombrugghe B, 1996. Evidence for three major transcription activation elements in the proximal mouse proα2(1) collagen promoter. *Nucleic Acids Res* **24**: 3253-3260.

Hashizume H and Imahori K, 1967. Circular dichroism and conformation of natural and synthetic polynucleotides. *J Biol Chem (Tokyo)* **61:** 738-749.

Hobbs CA and Yoon K, 1994. Differential regulation of gene expression *in vivo* by triple helix-forming oligonucleotides as detected by a reporter enzyme. *Antisense Res Dev* **4**: 1-8.

Hough-Evans BR, Franks RR, Zeller RW, Britten RJ and Davidson EH, 1990. Negative spatial regulation of the lineage specific *CyIIIa* actin gene in the sea urchin embryo. *Development* **110:** 41-50.

Ikehara I, Uesugi KL and Yano J, 1972. Left-handed helical polynucleotides with D-sugar phosphodiester backbones. *Nat New Biol* **240:**16-17.

Izmailova ES, Wieczorek E, Brent Perkins E and Zehner ZE, 1999. A GC-box is required for expression of the human vimentin gene. *Gene* **235**: 69-75.

Jackson SP and Tjian R, 1988. O-glycosylation of eukaryotic transcription factors: implications for mechanisms of transcriptional regulation. *Cell* **55**: 125-133.

Johnson PF and McKnight SL, 1989. Eukaryotic transcriptional regulatory proteins. *Biochemistry* **58**: 799-839.

Jones C, Wang J, Norcross M, Bohnlein E and Razzaque A, 1994. Identification and characterisation of a human herpesvirus 6 gene segment capable of transactivating the human immunodeficiency virus type 1 long terminal repeat in an Sp1 binding site-dependant manner. *J Virol* **68**: 1706-1713.

Kadonaga JT, Carner KR, Masiarz FR and Tjian R, 1987. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* **51**: 1079-1090.

Kageyama R and Pastan I, 1989a. Nuclear factor ETF specifically stimulates transcription from promoters without a TATA box. *J Biol Chem* **264**: 15508-15514.

Kageyama R and Pastan I, 1989b. Molecular cloning and characterisation of a human DNA binding factor that represses transcription. *Cell* **59**: 815-825.

Karsenty G and deCrombrugghe B, 1991. Conservation of binding sites for regulatory factors in the coordinately expressed α1(I) and α2(I) collagen promoters. *Biochem Biophys Res Commun* **177**: 538-544.

Keshet I, Lieman-Hurwitz J and Cedar H, 1986. DNA methylation affects the formation of active chromatin. *Cell* **44**: 535-543.

Killian CE and Wilt FH, 1996. Characterisation of the proteins comprising the integral matrix of *Strongylocentrotus purpuratus* embryonic spicules. *J Biol Chem* **271**: 9150-9159.

Kim SJ, Onwuta US, Lee YI, Li R, Botchan MR and Robbins PD, 1992. The retinoblastoma gene product regulates Sp1-mediated transcription. *Mol Cell Biol* **12**: 2455-2463.

Kim Y, Ratziu V, Choi S-G, Lalazar A, Theiss G, Dang Q, Kim S-J and Friedman SL, 1998. Transcriptional activation of transforming growth factor β1 and its receptors by the Krüppel-like factor Zf9/Core promoter-binding protein and Sp1. *J Biol Chem* **273**: 33750-33758.

Kinniburgh AJ, Firulli AB and Kolluri R, 1994. DNA triplexes and regulation of the c-*myc* gene. *Gene* **149**: 93-100.

Kirchhamer CV, Bogarad LD and Davidson EH, 1996a. Developmental expression of synthetic *cis* regulatory systems composed of spatial control elements from two different genes. *Proc Natl Acad Sci USA* **93**: 13849-13854.

Kirchhamer CV, Yuh C-H and Davidson EH, 1996b. Modular cis-regulatory organization of developmentally expressed genes: Two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc Natl Acad Sci USA* **93**: 9322-9328.

Kohwi Y and Kohwi-Shigematsu T, 1991. Altered gene expression correlates with DNA structure. *Genes Dev* **5**: 2547-2554.

Kohwi-Shigematsu T and Kohwi Y, 1985. Poly(dG).poly(dC) sequences, under torsional stress, induce altered DNA conformation upon neighbouring DNA sequences. *Cell* **43**: 199-206.

Koritschoner NP, Bocco JL, Panzetta-Dutari GM, Dumur CI, Flury A and Patrito LC, 1997. A novel human zinc finger protein that interacts with the core promoter element of the TATA box-less gene. *J Biol Chem* **272**: 9573-9580.

Kwon H-S, Kim M-S, Edenberg HJ and Hur M-W, 1999. Sp3 and Sp4 can repress transcription by competing with Ap1 for the core *cis*-elements on the human *ADH5/FDH* minimal promoter. *J Biol Chem* **274**: 20-28.

Lacy J, Roth G and Shieh B, 1994. Regulation of the human IgE receptor (FcεRII/CD23) by EBV. *J Immunol* **153**: 5537-5548.

Law GL, Itoh H, Law DJ, Mize GJ, Merchant JL and Morris DR, 1998. Transcription factor ZBP-89 regulates the activity of the ornithine decarboxylase promoter. *J Biol Chem* **273**: 19955-19964.

Lee CH, Murphy MR, Lee JS and Chung JH, 1999. Targeting a SWI/SNF-related chromatin remodelling complex to the beta-globin promoter in erythroid cells. *Proc Natl Acad Sci USA* **96**: 12311-12315.

Li N and Seetharam B, 1998. A 69-base pair fragment derived from human transcobalamin II promoter is sufficient for high bidirectional activity in the absence of a TATA box and an initiator element in transfected cells. *J Biol Chem* **273**: 28170-28177.

Li R, Knight JD, Jackson SP, Tjian R and Botchan MR, 1991. Direct interaction between Sp1 and the BPV enhancer E2 protein mediates synergistic activation of transcription. *Cell* **65**: 493-505.

Lisowsky T, Polosa PL, Sagliano A, Roberti M, Gadaleta MN and Cantatore P, 1999. Identification of human GC-box-binding zinc finger protein, a new Krüppel-like zinc finger protein, by the yeast one-hybrid screening with a GC-rich target sequence. *FEBS Letters* **453**: 369-374.

Livant DL, Cutting AE, Britten RJ and Davidson EH, 1988. An *in vivo* titration of regulatory factors required for expression of a fusion gene in transgenic sea urchin embryos. *Proc Natl Acad Sci USA* **85**: 7607-7611.

Macleod D, Charlton J, Mullins J and Bird AP, 1994. Sp1 sites in the mouse *aprt* gene promoter are required to prevent methylation of the CpG island. *Genes Dev* **8**: 2282-2292.

Maouche L, Lucien N, Cartron JP and Chrétien S, 1995. A CCACC motif mediates negative transcriptional regulation of the human erythropoietin receptor. *Eur J Biochem* **233**: 793-799.

Marin M, Karis A, Visser P, Grosveld F and Philipsen S, 1997. Transcription factor Sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell* **89**: 619-628.

Mastrelangelo IA, Courey AJ, Wall JS, Jackson SP and Hough PVC, 1991. DNA looping and Sp1 multimer links: A mechanism for transcriptional synergism and enhancement. *Proc Natl Acad Sci USA* **88**: 5670-5674.

Masuda ES, Yamaguchi-Iwai Y, Tsuboi A, Hung P, Arai K-I and Arai N, 1994. The transcription factor Sp1 is required for induction of the murine GM-CSF promoter in T cells. *Biochem Biophys Res Commun* **205**: 1518-1525.

Mäueler W, Kyas A, Keyl H-G and Epplen JT, 1998. A genome-derived (gaa.ttc)24 trinucleotide block binds nuclear protein(s) specifically and forms triple helices. *Gene* **215**: 389-403.

Maxam G and Gilbert M, 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**: 560-564.

Mayfield C, Ebbinghaus S, Gee J, Jones D, Rodu B, Squibb M and Miller D, 1994. Triplex formation by the human Ha-ras promoter inhibits Sp1 binding and *in vivo* transcription. *J Biol Chem* **269**: 18232-18238.

Merchant JL, Shiotani A, Mortensen ER, Shumaker DK and Abraczinskas DR, 1995. Epidermal growth factor stimulation of the human gastrin promoter requires Sp1. *J Biol Chem* **270**: 6314-6319.

Merika M and Orkin SH, 1995. Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Krüppel family proteins Sp1 and EKLF. *Mol Cell Biol* **15**: 2437-2447.

Miller IJ and Bieker JJ, 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Krüppel family of nuclear proteins. *Mol Cell Biol* **5**: 2776-2786.

Mitchel PJ and Tijian R, 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA-binding proteins. *Science* **245**: 371-378.

Mitsuhiro S, 1998. Induction of Sp1 in differentiating human embryonal carcinoma cells triggers transcription of the fibronectin gene. *Mol Cell Biol* **18**: 3010-3020.

Moore DS and Wagner TE, 1974. Double-helical DNA and RNA circular dichroism. *Biopolymers* **13:** 977-986.

Musso M, Nelson LD and Van Dyke MW, 1998. Characterization of purine-motif triplex DNA-binding proteins in Hela extracts. *Biochemistry* **37**: 3086-3095.

Nan X, Ng H-H, Johnson CA, Laherty CD, Turner BM, Eisenman RN and Bird A, 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **28**: 386-389.

Nehls MC, Grapilon ML and Brenner DA, 1992. NF-I / SP1 switch elements regulate collagen alpha 1(I) gene expression. *DNA Cell Biol* **11**: 443-452.

Ng H-H and Bird A, 1999. DNA methylation and chromatin modification. *Curr Opin Gen Dev* **9**: 153-163.

Nickol MC and Felsenfeld G, 1983. DNA conformation at the 5' end of the chicken adult beta-globin gene. *Cell* **35**: 467-477.

Nielsen SJ, Praestegaard M, Jorgensen HF and Clark FC, 1998. Different Sp1 family members differentially affect transcription from the human elongation factor 1 A-1 gene promoter. *Biochem J* **333**: 511-517.

Novak U, Ji H, Kanagasundaram V, Simpson R and Paradiso L, 1998. STAT3 forms stable homodimers in the presence of divalent cations prior to activation. *Biochem Biophys Res Commun* **247:** 558-563.

Oda E, Shirasuna K, Suzuki M, Nakano K, Nakajima T and Oda K, 1998. Cloning and characterization of a GC-box binding protein G10BP-1, responsible for repression of the rat fibronectin gene. *Mol Cell Biol* **18**: 4772-4782.

Palecek E, 1992. Local supercoil-stabilised DNA structures. *Crit Rev Biochem Mol Biol* **26:** 151-226.

Palla F, Melfi R, Gaetano L, Bonura C, Anello L, Alessandro C and Spinelli G, 1999. Regulation of the sea urchin early H2A histone gene expression depends on the modulator element and on sequences located near the 3' end. *Biol Chem* **380**: 159-165.

Pascal E and Tjian R, 1991. Different activation domains of Sp1 govern formation of multimers and mediate transcriptional synergism. *Genes Dev* **5**: 1646-1656.

Patletich NP and Pabo CO, 1991. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**: 809-816.

Patterton D and Hapgood JP, 1994. suGF1 binds in the major groove of its oligo(dG).oligo(dC) recognition sequence and is excluded by a positioned nucleosome core. *Mol Cell Biol* **14**: 1410-1418.

Patterton D, 1992. MSc thesis: Purification and characterisation of a poly(dG).poly(dC)-binding protein from *Parenchinus Angulosis*. University of Cape Town, South Africa.

Patterton H-G and Hapgood JP, 1996. The translational placement of nucleosome cores *in vitro* determines the access of the transacting factor suGF1 to DNA. *Nucleic Acids Res* **24**: 4349-4355.

Patterton H-G and von Holt C, 1993. Negative supercoiling and nucleosome cores II. The effect of negative supercoiling on the positioning of nucleosome cores *in vitro*. *J Mol Biol* **229**: 637-655.

Philipsen S and Suske G, 1999. A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. *Nucleic Acids Res* **27**: 2991-3000.

Pruzina S, Antoniou M, Hurst J, Grosveld F and Philipsen S, 1994. Transcriptional activation by hypersensitive site three of the human β-globin locus control region in murine erythroleukemia cells. *Biochim Biophys Acta* **1219**: 351-360.

Ratzlu V, Lalazar A, Wong L, Dang Q, Collins C, Shaulian E, Jensen S and Friedman SL, 1998. Zf9, a Krüppel-like transcription factor up-regulated *in vivo* during early hepatic fibrosis. *Proc Natl Acad Sci USA* **95**: 9500-9505.

Redell JB and Tempel BL, 1998. Multiple promoter elements interact to control the transcription of the potassium channel gene, KCNJ2. *J Biol Chem* **273**: 22807-22818.

Rosalia CM, Chung TE, Imataka H, Michel FJ, Badinga L and Simmen FA, 1999. Trans-activation functions of the Sp-related nuclear factor, basic transcription element-binding protein, and progesterone receptor in endometrial epithelial cells. *Endocrinology* **140**: 2517-2525.

Roush W, 1996. "Smart" genes use many cues to set cell fate. *Science* **272**: 652-653.

Saltzman AG and Weinmann R, 1989. Promoter specificity and modulation of RNA polymerase II transcription. *FASEB J* **3**: 1723-1733.

Sambrook J, Fritsch EF and Maniatis T, 1989. Molecular cloning: A laboratory manual (2nd Edition). *Cold Springs Harbor Laboratory Press*, United States of America.

Schaffner T, Zimmerman A, Keller HU, Locher GW and Cottier H, 1978. Genes and spacers of cloned sea urchin histone DNA analysed by sequencing. *Cell* **14:** 655-671.

Scharer I and Iggo YH, 1992. Mammalian p53 can function as a transcription factor in yeast. *Nucleic Acids Res* **20:** 1539-1545.

Scherer SD, 1997. PhD thesis: Protein purification and cDNA cloning of suGF1, a sea urchin nuclear DNA-binding protein. University of Cape Town, South Africa.

Schumacher S, Clubb RT, Cai M, Mizuuchi K, Clore GM and Gronenborn AM, 1997. Solution structure of the Mu end DNA-binding ibeta subdomain of phage Mu transposase: modular DNA recognition by two tethered domains. *EMBO J* **16:**7532-7541.

Seid CA, Ramachandran RK, George JM, Govindarajan V, Gonzalez-Rimbau MF, Flytzanis CN and Tomlinson CR, 1997. An extracellular matrix response element in the promoter of the LpS1 genes of the sea urchin *Lytechinus pictus. Nucleic Acids Res* **25**: 3175-3182.

Seid CA, Sater AK, Falzone RL and Tomlinson CR, 1996. A tissue-specific repressor in the sea urchin embryo of *Lytechinus pictus* binds the distal G-string element in the LpS1-beta promoter. *DNA Cell Biol* **15**: 511-517.

Shrivastava A and Calame R, 1994. An analysis of genes regulated by the multi-functional transcriptional regulator Yin-Yang-1. *Nucleic Acids Res* **22**: 5151-5155.

Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ and Cedar H, 1999. DNA methylation represses transcription in vivo. *Nature Genet* **2**: 203-206.

Sogawa. K, Kikuchi Y, Imataka H and Fujii-Kuriyama Y, 1993. Comparison of DNA-binding properties between BTEB and Sp1. *J Biochem (Tokyo)* **114**: 605-609.

St.John D and Davis TC, 1981. The organisation and transcription of the galactose gene cluster of Saccharomyces. *J Mol Biol* **152:** 285 -315.

Stokorvá J, Vojtiková K and Palecek S, 1989. Electron Microscopy of supercoiled pEJ4 DNA containing homopurine.homopyrimidine sequences. *J Biomol Struct Dyn* **6**: 893-897.

Stultz CM, White JV and Smith TF, 1993. Structural analysis based on state-space modelling. *Protein Sci* 1993 **2:** 305-314.

Supakar CS, 1997. Functional role of a conformationally flexible homopurine / homo-pyrimidine domain of the androgen receptor gene promoter interacting with Sp1 and a pyrimidine single strand DNA-binding protein. *Mol Endocrinol* **11**: 3-15.

Suzuki T, Yamamota T, Kurabayashi M, Nagai R, Yazaki Y and Horikoshi M, 1998. Isolation and initial characterisation of GBF, a novel DNA-binding zinc finger protein that binds to the GC-rich binding sites of the HIV-1 promoter. *J Biochem (Tokyo).* **124**: 389-395.

Taketani S, Mohri T, Hioki K, Tokunaga R and Kohno H, 1999. Structure and transcriptional regulation of the mouse ferrochelate gene. *Gene* **227**: 117-124.

Takimoto M, 1999. Molecular analysis of the GCF gene identifies revisions to the cDNA and amino acid sequences. *Biochim Biophys Acta* **1447**: 125-131.

Tate P, Skarnes W and Bird A, 1996. The methyl-CpG binding protein MeCP2 is essential for embryonic development in the mouse. *Nature Genet* **12**: 205-208.

Tazi J and Bird A, 1990. Alternative chromatin structure at CpG islands. *Cell* **60**: 909-920.

Viswanathan M, Yu M, Mendoza L and Yunis JJ, 1996. Cloning and transcription factor-binding sites of the human c-rel proto-oncogene promoter. *Gene* **170**: 271-276.

Wade PA, Jones PL, Veenstra GJ, Vermaak D, Kass SU, Landsberger N, Strouboulis J and Wolffe AP, 1998. Methylated DNA and MeCP2 recruit histone deacetylase to transcription. *Nature Genet* **19**: 187-191.

Wang and Patel, 1993. Solution structure of a parallel-stranded G-quadruplex DNA. *J Mol Biol* **20**: 1171-1183.

White SH, 1994. The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences. *J Mol Evol* **38:** 383-394.

Wimmer EA, Jackle H, Pfeifle C and Cohen SM, 1993. A *Drosophila* homologue of human Sp1 is a head-specific segmentation gene. *Nature* **366**: 690-694.

Winston F, Dollard C and Ricupero-Hovasse SL, 1995. Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast* **11:** 53-55.

Wishart DS, Boyko RF, Willard L, Richards FM and Sykes BD, 1994. SEQSEE: a comprehensive program suite for protein sequence analysis. *Comput Appl Biosci.* **10**:121-132.

Wray GA, 1998. Promoter logic. *Science* **279**: 1871-1872.

Xiang M, Lu S-Y, Musso M, Karsenty G and Klein WH, 1991. A G-string positive *cis*-regulatory element in the LpS1 promoter binds two distinct nuclear factors distributed non-uniformly in *Lytechinus pictus* embryos. *Development* **113**: 1345-1355.

Yamasu K and Wilt FH, 1999. Functional organisation of DNA elements regulating SM30alpha, a spicule matrix gene of sea urchin embryos. *Dev Growth Differ* **41**: 81-91.

Yenidunya A, Davey C, Clark D, Felsenfeld G and Allan J, 1994. Nucleosome positioning on chicken and human globin gene promoters *in vitro*. *J Mol Biol* **237**: 401-414.

Yuh CH and Davidson EH, 1996. Modular *cis*-regulatory organisation of *Endo16*, a gut-specific gene of the sea urchin embryo. *Development* **122**: 1069-1082.

Yuh CH, Ransick A, Martinez P, Britten RJ and Davidson EH, 1994. Complexity and organisation of DNA-protein interactions in the 5'-regulatory region of an endoderm-specific marker gene in the sea urchin embryo. *Mech Dev* **47**: 165-186.

Zacharias W, Larson JE, Klysik J, Stirdivant SM and Wells RD, 1982. Conditions which cause the right-handed to left-handed DNA conformational transitions. Evidence for several types of left-handed DNA structures in solution. *J Biol Chem* **25**: 2775-2782.

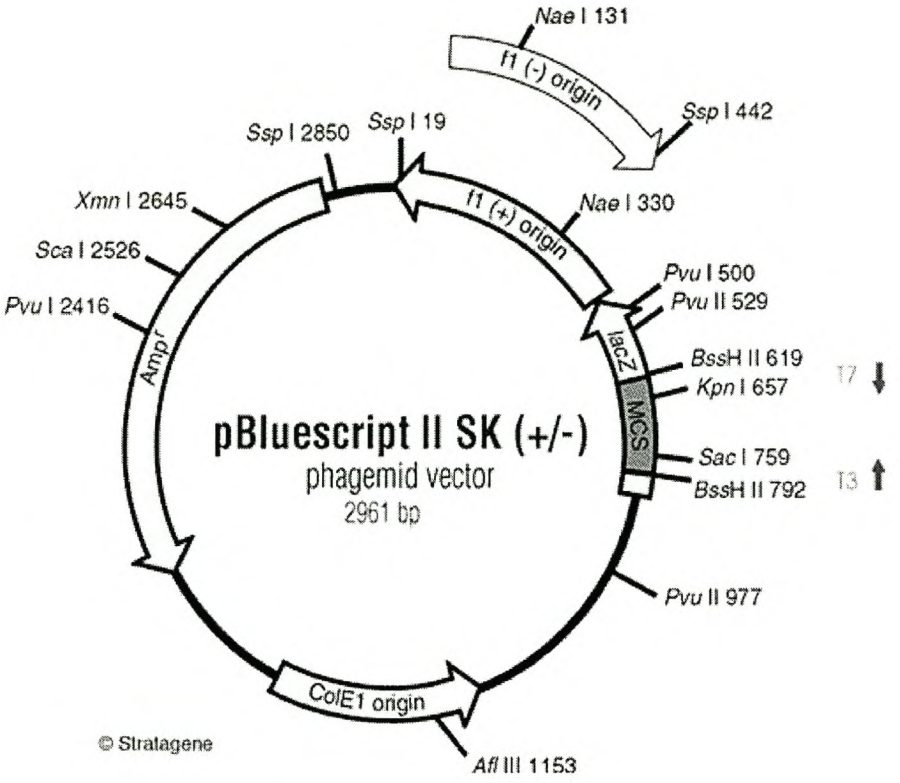
Zeller RW, Coffman JA, Harrington MG, Britten RJ and Davidson EH, 1995a. SpGCF1, a sea urchin embryo DNA-binding protein, exists as five nested variants encoded by a single mRNA. *Dev Biol* **169**: 713-727.

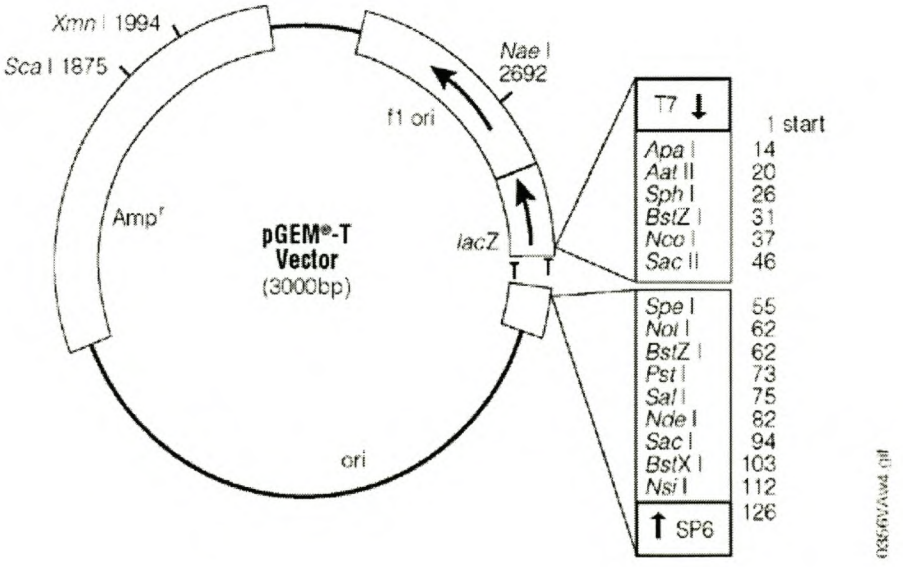
Zeller RW, Griffith JD, Moore JG, Kirchhamer CV, Britten RJ and Davidson EH, 1995b. A multimerising transcription factor of sea urchin embryos capable of looping DNA. *Proc Natl Acad Sci USA* **92**: 2989-2993.

Zor K and Selinger JH, 1996. Linearisation of the Bradford protein assay increases its sensitivity: theoretical and experimental studies. *Anal Biochem* **236:** 302-308.

Appendix I

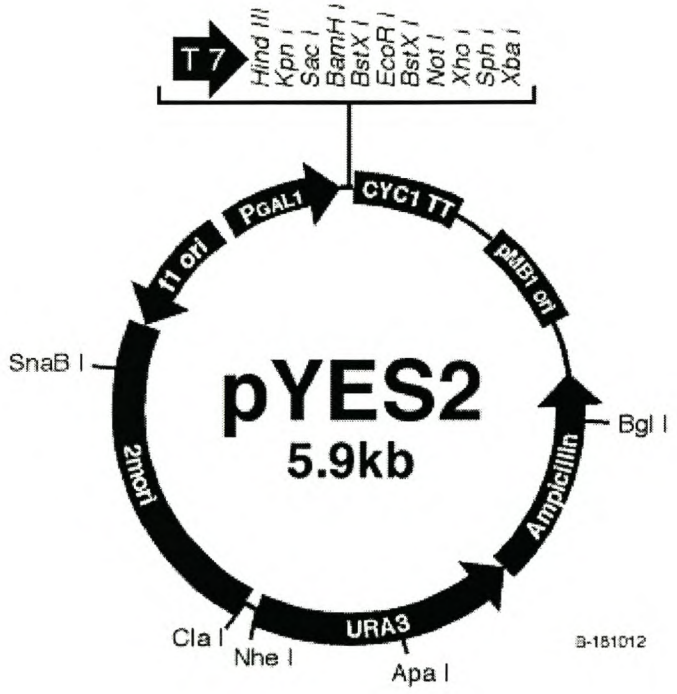
Plasmid Maps





Comments for pYES2:
5857 nucleotides

GAL1 promoter: bases 1-452
T7 promoter/priming site: bases 476-495
Multiple cloning site: bases 502-601
CYC1 transcription terminator: bases 609-857
pMB1 (pUC-derived) origin: bases 1039-1712
Ampicillin resistance gene: bases 1857-2717
URA3 gene: bases 2735-3842
2 micron origin: bases 3846-5317
f1 origin: bases 5385-5840



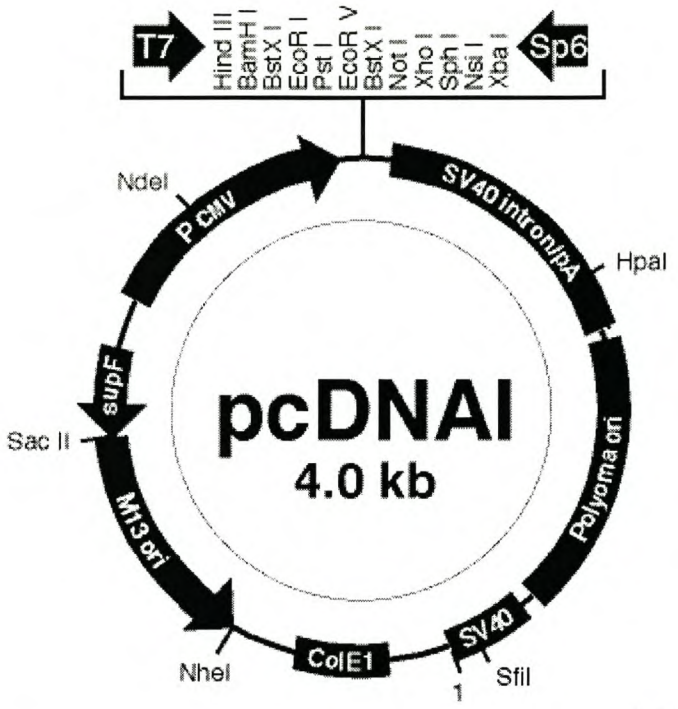
The sequence of pYES2 has been compiled from information in sequence databases, published sequences, and other sources. This vector has not been completely sequenced. If you suspect an error in the sequence, please contact Invitrogen's Technical Services Department.

U.S. Headquarters
Tel: 1-800-955-6288
Fax: 1-760-603-7201

European Headquarters
Tel: +31 (0) 50 5299 299
Fax: +31 (0) 50 5299 280

Comments for pcDNA I:
4033 base pairs

Col E1 origin: bases 1-587
M13 origin: bases 588-1182
SupF gene: bases 1183-1384
CMV promoter: bases 1517-2170
T7 promoter: bases 2171-2189
T7 primer sequence: bases 2170-2189
Polylinker: bases 2187-2306
Sp6 primer sequence: bases 2307-2325
Splice and poly A: bases 2326-3024
Polyoma origin: bases 3029-3870
SV40 origin: bases 3871-4033



vA-151002

The sequence of pcDNA I has been compiled from information in sequence databases, published sequences, and other sources. Portions of this vector have not yet been completely sequenced. If you suspect an error in the sequence, please contact Invitrogen's Technical Services Department at 800-955-6288.

U.S. Headquarters
Tel: 1-800-955-6288
Fax: 1-619-597-6201

European Headquarters
Tel: +31 (0) 5945-15175
Fax: +31 (0) 5945-15312

Appendix II

335 bp *EcoRI / HindIII* insert from the H1 - H4 intergenic region of the *P.miliaris* early histone gene battery.

gaattctc atgtttgaca gcttatcatc gccctgactg agtcgagccc
cttaagag tacaaactgt cgaatagtag cgggactgac tcagctcggg

aattcgagct cggtacccCA CGTAGAGGAA AAGAGAGTTA TACCACTCCT
ttaagctcga gccatgggGT GCATCTCCTT TTCTCTCAAT ATGGTGAGGA

GACATGAAAC ACACTCAATT CAACATATTT AGAGGAAGGG **AGAGAGAGAG**
CTGTACTTTG TGTGAGTTAA GTTGTATAAA TCTCCTTCCC TCTCTCTCTC

AGAGAGAGAG AGAGAGAGAG AGGGGGGGGG GGAGGGAGAA TTGCCCAAAA
TCTCTCTCTC TCTCTCTCTC TCCCCCCCCC CCTCCCTCTT AACGGGTTTT

CACTGTAAAT GTAGCGTTAA TGAACTTTTC ATCTCATCGA CTGCGCGTGT
GTGACATTTA CATCGCAATT ACTTGAAAAG TAGAGTAGCT GACGCGCACA

ATAAGGATGA TTATAAGCTg gggatcctct agagtcgacc tgcaggcatg
TATTCCTACT AATATTCGAc ccctaggaga tctcagctgg acgtccgtac

caagctgggc tcgacttagt cagggtcacc gataagctt
gttcgacccg agctgaatca gtcccagtgg ctattcgaa

Appendix III

Protein Sequences

1) Full-length suGF1 amino acid sequence:

MSTLPQPLSH CLLNQVNTAA INLPHQQPGL ITDIKPMISN KPPPTQEVKP
NILAAAAAGL TYPPLNVPSL PAMPNVSMPN VSLPNVSMPN VSMPNVSMPT
SVSMPSVSMP SVSMPSASMP SVTLHNQQGN NSQLSNSNSQ RLSQMKKCPN
EFLHQNPQSE RQLFYNDVAM QLYNSDFNKF ASKKEFHGYL LEQQKWRWDT
HSYIGNLETR VHNLLINPNS GVAQNVARYR SVPIKCKSED VKRCEATSKE
LENMATRIAS VRQQLLHKKG TLLTSSDNSV IVWQNELAYI EQLFDRTDQM
YNEVLSTLAS VNQTFSHLQT SFTAEAAELA DRRRLWRRRK ENNRKRRKRM
EKQLEKIEQR SCELLFHITS RGAYDRVRSH PEMPRIGPSE VNTDMLNGIK
SKSEVRPLMH LLSKGYMTPG AMEMVSQKIQ KLECGIKTEA HQQATQVGIN
SLAINKMPVP ASRIKSILPP APPPVTGVAS STMISSTMVS SVNSAAPVTQ
QSVPTVNLNT QLAK

2) Full-length spGCF1 amino acid sequence:

MSTLPQPLSHCLLNQVHPALNLPQTGVITDIKPMISNKPPTQEVKPNILATGLPYPPLNVPRLPVM
PNVSLPSVSMPSVSMPNVSMPNASMPSVSMPNVSMPSIPHHNLQGNLGQLLNNSNSQKMSQMKKCP
NEFLHQNPQSERQLFYNDVAMQLYNSDFNKFASKKGFHGYLLEQQKWRWDTHSYIGNLETRVHNLL
INPNSGVAQNVARYRSVPIKCKSEDVKRCKATSKELENMATRIASVRQQLLHKKGTLLTSSDNSVI
VWQNELAYIEQLFDRTDQMYNEVLSTLASVNQTFSHLQTSFTAEAAELADRRRLWRRRKENNRKRR
KRMEKQLEKIEQRSCELLFHITSRGAYDRVRSHPEMPRIGPSEVNTDMLNGIKSKSEVRPLMHLLS
KGYMTPGAMEMVSQKIQKLECGIKTEAHQQATQVGINSLSINKITAPASELNSILPPVTGIASSNM
VSSVNSAVTQQSVPTVNLNTQLAK

3) Full-length hORFX amino acid sequence:

MSTATTVAPAGIPATPGPVNPPPPEVSNPSKPGRKTNQLQYMQNVVVKTLWKHQFAWPFYQPVDAI
KLNLPDYHKIIKNPMDMGTIKKRLENNYYWSASECMQDFNTMFTNCYIYNKPTDDIVLMAQALEKI
FLQKVAQMPQEEVELLPPAPKGKGRKPAAGAQSAGTQQVAAVSSVSPATPFQSVPPTVSQTPVIAA
TPVPTITANVTSVPVPPAAAPPPPATPIVPVVPPTPPVVKKKGVKRKADTTTPTTSAITASRSESP
PPLSDPKQAKVVARRESGGRPIKPPKKDLEDGEVPQHAGKKGKLSEHLRYCDSILREMLSKKHAAY
AWPFYKPVDAEALELHDYHDIIKHPMDLSTVKRKMDGREYPDAQGFAADVRLMFSNCYKYNPPDHE
VVAMARKLQDVFEMRFAKMPDEPVEAPALPAPAAPMVSKGAESSRSSEESSSDSGSSDSEEERATR
LAELQEQLKAVHEQLAALSQAPVNKPKKKKEKKEKEKKKKDKEKEKEKHKVKAEEEKKAKVAPPAK
QAQQKKAPAKKANSTTTAGRQLKKGGKQASASYDSEEEEEGLPMSYDEKRQLSLDINRLPGEKLGR
VVHIIQSREPSLRDSNPDEIEIDFETLKPTTLRELERYVKSCLQKKQRKPFSASGKKQAAKSKEEL
AQEKKKELEKRLQDVSGQLSSSKKPARKEKPGSAPSGGPSRLSSSSSSESGSSSSSGSSSDSSDSE

4) Sequence of suGF1 basic region used for structure-function prediction:

SFTAEAAELA **DRRRLWRRRK ENNRKRRKRM EKQLEKIEQR** SCELLFHITS RGAYDRVRSH

5) Sequence of hORFX basic region used for structure-function prediction:

LKAVHEQLAA LSQAPVN**KPK KKKEKKEKEK KKKDKEKEKE KHKVKAEEEK KAK**VAPPAK

END