Protein Purification and cDNA Cloning of suGF1, a Sea Urchin Nuclear DNA-Binding Factor

Sonja Daniela Scherer

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Biochemistry

University of Cape Town

October 1997

CONTENTS

ABSTRACT

Protein Purification and cDNA Cloning of suGF1, a Sea Urchin Nuclear DNA-Binding Factor

The upstream regulatory regions of numerous genes contain contiguous deoxyguanosine residues (G·C-rich sequences) which have been implicated in the regulation of gene expression, since they may involve alterations in their DNA structure, the binding of G-string factors and in some cases even the displacement of a nucleosome positioned over this area. A poly(dG).poly(dC)-binding protein (suGF1) has previously been identified and purified on a small scale from embryonic nuclear extracts of the sea urchin *Parechinus angulosus* (1, 2). suGF1 binds *in vitro* to the H1-H4 intergenic region of the early histone gene battery, and the recognition site contains 11 contiguous Gs which are incorporated into a positioned nucleosome core *in vitro*. suGF1 may be a member of a family of G-string factors which could be involved in the developmental regulation of unrelated genes in various organisms.

Prior to the commencement of this project no protein or DNA sequence information was available on the protein. The main objective of this thesis was to obtain the cDNA and the primary amino acid sequence for suGF1. Using this information, additional aims were to determine the developmental distribution of the protein and obtain insight into the molecular basis of the regulatory function of suGF1 by analysis of the primary protein structure and expression of the cDNA.

A large scale purification system was established to purify suGF1, essentially using affinity purification. The purification procedure involved isolating nuclei from 14 hour sea urchin embryos and extracting their nuclear proteins. These were fractionated by a combination of a phosphocellulose ion exchange step and poly(dG).poly(dC) affinity chromatography, and finally SDS-PAGE. In total, the purification yielded about 600 ng of suGF1 protein. The 57 kDa protein (about 300 ng) was excised from a Coomassie stained gel and treated by enzymatic cleavage to generate peptides, three of which were identified by mass spectral analysis.

The protein isolation and cDNA cloning strategies were developed concurrently. Since the protein sequence was only available towards the end of the project, extensive work was performed to obtain the cDNA prior to the identification of the protein sequence.

The suitability of suGF1 as a candidate for the DNA-ligand screening approach was established by characterising its DNA-binding specificity and affinity. A sea urchin recombinant DNA library constructed in lambda ZAP was expressed in bacteria and protein replica filters were screened with radiolabelled recognition site DNA containing multiple copies of the binding site. Several putative cDNA clones which recognise the DNA-binding site sequence-specifically were detected after two rounds of screening. These clones were further characterised by DNA sequencing and database comparisons. Some of the clones have reading frames which do not correspond to any previously identified proteins in the database. However, despite extensive attempts to express these putative positive clones in various expression systems, including bacterial systems, eukaryotic COS cells and *in vitro* expression with rabbit reticulocyte lysate, no protein expression was obtained to enable further characterisation of the clones, which were subsequently shown to be false positives.

During the course of the project, and prior to obtaining the amino acid sequence for suGF1, another G·C-binding factor, SpGCF1, was isolated from the sea urchin *Strongylocentrotus purpuratus* (3). Since the DNA-binding characteristics of SpGCF1 were very similar to suGF1, we predicted that they may be homologous proteins. Degenerate PCR primers were designed to the cDNA sequence of SpGCF1 to obtain a partial DNA sequence of a homologous protein in *P.angulosus* using genomic DNA as template. The 2.1 kb full length *P.angulosus* cDNA homologue to SpGCF1 was obtained by 5' and 3' RACE performed on cDNA prepared from total RNA isolated from *P.angulosus* embryos. Finally, the mass spectral analysis of native suGF1 confirmed that the isolated cDNA codes for suGF1 protein isolated from *P.angulosus* embryos, and that suGF1 and SpGCF1 are indeed orthologues.

Analysis of the primary structure of suGF1 shows that the full length protein has a precise molecular weight of 57.2 kDa, whereas SpGCF1 has a predicted molecular weight of 54.6 kDa. The difference in molecular weights of the two proteins is mostly accounted for by the 27 amino acids difference in their primary structure. Overall the proteins are 94 % similar on the amino acid level, and the highest degree of conservation is retained within the respective putative DNA-binding domains. There is no overall homology to any other proteins reported in the databases. The homology on the cDNA level is 84 % similar over a region of 1989 nucleotides, which includes the entire coding region of 513 amino acids as well as some 5' and 3' untranslated sequence. The sequences of the two cDNAs are most

diverse in the untranslated regions. The 5' untranslated sequence of suGF1 does not contain any characteristic consensus initiation signals, however at least two in frame stop codons precede the first methionine in the predicted open reading frame. The 3' untranslated region is very short and has no distinguishing features.

suGF1 consists of several characteristic domains. The putative DNA-binding domain is situated centrally in the protein and is strongly basic. It is closely associated with potential heptad repeats of hydrophobic residues, which are also found in other regions of the protein. These repeats are commonly associated with the leucine zipper class of DNA-binding proteins, as well as proteins which adopt "coiled-coil" structures. The N-terminus of the protein is characterised by nine tandem copies of a pentapeptide repeat (N/SVSMP), which is unique to suGF1 and SpGCF1 and to which no function can be assigned yet. Another characteristic of the N-terminus is its high proline content, which is associated with the activation domains of various other transcription factors. A third prominent feature of the N-terminus is the presence of multiple methionine residues, which may act as alternative initiation codons during translation. This is a common feature of genes which have TATA-less promoters, and therefore lack strong initiation consensus sequences.

The expression of suGF1 from the isolated cDNA using rabbit reticulocyte lysate was shown by SDS-PAGE to contain several translated products of approximately 57 kDa, 53 kDa, 49 kDa, 42 kDa and 39 kDa. The multiple protein products are present in a ratio of about 4 : 3 : 1 : 2 : 2, and are consistent with multiple translation start sites within the cDNA. Gel shift retardation assays of the *in vitro* translated full length cDNA showed the same characteristic protein doublet obtained with native purified protein and nuclear extracts, and was also consistent with multiple protein products being formed from several start sites within the cDNA. Gel shifts of the expressed suGF1 proteins from truncated cDNAs showed that the DNA binding domain is contained in the region of nt 760 - 1662, and that the ability of suGF1 to recognise its cognate DNA-binding site does not require homodimerisation. RT-PCR showed that the mRNA transcript for suGF1 is present in eggs, 4 to 45 hour embryos, as well as adult testes and muscle tissue, however it is absent in ovaries.

ACKNOWLEDGEMENTS

I wish to thank

My supervisors, Professor Janet Hapgood and Professor Wolf Brandt, for their guidance, support and encouragement during the course of this project.

My parents for their continual enthusiasm, support and interest in my work.

The Foundation for Research Development and the University Research Committee for financial assistance.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

aa	amino acid(s)
AP1	Adaptor primer 1 (ClonTech)
AMPS	Ammonium peroxodisulphite
β-gal	β-galactosidase protein
bp	base pair(s)
BGP1	beta globin protein 1
BSA	Bovine Serum Albumin
BSAP	B-cell specific activator protein
BTEB	BTE-binding protein
Buffer C	Column buffer (0.X buffer C denotes buffer C containing 0.X M KCl)
CAT	Chloramphenicol actyl transferase
cDNA	Complementary DNA
C/EBP	CAAT / enhancer binding protein
CID	Collision induced dissociation
COS-1 cells	African green monkey kidney cells
CyIIIa	Cytoskeletal III actin gene
ddNTP	Dideoxyribonucleotide triphosphate
DEPC	Diethylpyrocarbonate
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethylsulfoxide
ddNTP	Dideoxyribonucleotide triphosphate
dNTP	Deoxyribonucleotide triphosphate
dpm	Disintegrations per minute
ds	Double stranded
DTT	Dithiothreitol
E.coli	*Escherichia coli*
EDTA	Ethylenediaminetetra-acetic acid
EGFR	Epidermal growth factor receptor
EGTA	Ethyleneglycol-bis-(2-amino-ethyl eyther)N,N'-tetra-acetic acid
E/H fragment	Double-stranded DNA restriction fragment obtained from *EcoR1 / HindIII* digestion of plasmid pHP2
EKLF	erythroid Krüppel-like zinc finger protein
EMSA	Electrophoretic mobility gel shift
EtBr	Ethidium Bromide
FCS	Fetal Calf Serum
fig	Figure
G	Guanine residue of double stranded DNA
GATA-1	erythroid restricted zinc finger protein
G·C	Guanine·Cytosine- rich (predominantly guanine on one strand)
GCF	GC-binding factor
GCG	Genetics Computer Group
GST	Glutathione S-transferase
G-string	Oligo(dG).oligo(dC) region
Hepes	4-(2-Hydroxyethyl)-1-piperazineethane sulphonic acid
H4TF-1	Histone 4 Transcription factor -1
HS	hypersensitive site(s)
IF-1(2)	Inhibitory factor 1 (2)
IPTG	Isopropyl-β-D-thiogalactopyranoside
kb	kilo base
K_d	Dissociation constant
kDa	kilo Dalton
LB	Luria-Bertani medium
LC	Liquid chromatography
LCR	locus control region
LIP / LAP	Liver-enriched transcriptional inhibitor and transcriptional activator protein

MDR	human multidrug resistance
MMLV	Monkey murine leukemia virus
MOPS	Morpholinopropanesulfonic acid
mRNA	messenger RNA
MS	Mass spectrometry
MS/MS	tandem mass spectrometry
MW	Molecular weight
NE	Nuclear extract
NF-E2	Nuclear Factor-Erythroid 2
NF-κB	Nuclear Factor kappa B
NGFI-A(C)	Nerve growth factor induced protein A(C)
NP-40	Nonidet-P40
NS-oligo	Nonspecific oligodeoxyribonucleotide
nt	Nucleotide(s)
oct	Octamer binding protein
oligo	Oligodeoxyribonucleotide
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
P. angulosus	*Parechinus angulosus*
PCR	Polymerase chain reaction
p[d(I-C)]	Polydeoxyinosinic-deoxycytidylic acid
PEG	Polyethylene glycol
pfu	plaque forming units
pit-1	Pituitary specific protein 1
PMSF	Phenylmethylsulfonyl fluoride
poly(dG).poly(dC)	Polydeoxyguanidylic-polydeoxycytidilic acid
pur.pyr	purine.pyrimidine
RACE	Rapid Amplification of cDNA Ends
S. purpuratus	*Strongylocentrotus purpuratus*
SDS	Sodium dodecyl sulfate
Sp1	Stimulatory protein 1
Sp-oligo	Specific oligodeoxyribonucleotide
SP1/SP2	gene specific primers 1 and 2, respectively
SpGCF1	*Strongylocentrus purpuratus* G·C binding factor 1
SSAP	Stage specific activator protein
SSC	Saline sodium citrate
suGF1	Sea urchin G-binding Factor 1
TAE	Tris-Acetate-EDTA
TBE	Tris-Borate-EDTA
TCA	Trichloroacetic acid
TE	Tris-EDTA
TEMED	N,N,N',N'-Tetramethylethylene-diamine
TGE	Tris-Glycine-EDTA
Tris	2-Amino-2-(hydroxylmethyl)-1,3-propanediol
TSAP	Tissue specific activator protein
TSB	Transformation and Storage Buffer
X-gal	5-Bromo-4-chloro-3-indolyl-β-D-galactoside
YY-1	Yin-Yang-1

CHAPTER 1

Introduction

Investigations regarding the fundamental processes of development have been approached by the analysis of gene control mechanisms, showing that the differential control of gene transcription depends on a complex array of interactions within the *cis*-regulatory regions of the genes. These elements are recognised and bound by sequence-specific DNA-binding proteins, which are of central importance to transcriptional regulation and the deciphering of structural and regulatory information encoded in genomes (4, 5, 6). Transcription factors initiate and control procedures such as transcription, replication and site-specific recombination of DNA. The control functions are executed by the occupancy of target sites, the interaction of transcription factors with one another and the basal transcription apparatus (7). They determine the activity and specificity of enzymatic assemblies on DNA, by either activating or, in some cases, repressing the initiation of the general transcription apparatus (8). These DNA-binding proteins differ in their DNA-sequence specificities and in the way they themselves are regulated (9). For instance, some transcription factors are activated in response to physiological stimulation (eg hormones), whilst others are restricted to particular cell types or are expressed at certain stages of development only (10). Different combinations of these sequence-specific factors at specific target sites function to achieve highly complex and unique patterns of gene expression.

1.1 Gene Regulation Via Transcription Factors Binding to G·C-Rich DNA

DNA sequences containing characteristic homo(pur).(pyr) stretches, for instance G·C-rich regions found upstream of several unrelated eukaryotic genes, have been implicated in gene regulation (11, 12). Several investigations have reported that G-binding factors (found in a variety of tissues and organisms) are able to associate with these sequences (1, 3, 12, 13, 14, 15, 16). The relation amongst G-binding proteins, their biological significance with respect to gene regulation and the similarities amongst their DNA recognition sites need to be established via structural and functional investigations. It has been proposed

that G·C-rich DNA sequences, and homo(pur).(pyr) stretches in general, may either function to stabilise or hinder factor binding, and they could therefore act as conformational switches which are modulated by DNA-binding factors (11, 12, 17, 18). G·C-rich sequences are able to form unusual DNA structures (such as triple helices) *in vitro*. In general homo(pur).(pyr) regions are associated with the formation of unusual DNA structures *in vitro*, depending on the length of the homopurine stretch, the chemical environment and the degree of superhelical stress (19, 20). Another characteristic of these sequences is that they are often nuclease sensitive *in vivo*, a feature which is closely associated with promoter and enhancer functions, as a result of altered chromatin structure in the vicinity of actively transcribed genes (21). These hypersensitive domains are thought to result from the disruption or displacement of nucleosomes by transcription factors which bind sequence-specifically (22).

The molecular mechanisms whereby the transcriptional machinery can gain access to DNA is central to gene regulation. The DNA of genes is wrapped around histone proteins to form nucleosomes and the chromatin fibre. Chromatin structure is dynamic and can be altered during regulatory events (23), as elucidated by the inhibitory effects of higher order chromatin structures on transcription, whilst nuclease hypersensitive promoter regions (depleted of nucleosomes) are associated with transcriptional activity (24). This is exemplified by numerous systems, such as the yeast PHO5 gene promoter (25) and the β-globin genes (26). Recruitment of the transcriptional machinery may directly result in nucleosome remodelling (8). However there is also evidence that numerous genes code for proteins which are required for regulation and normal transcription by functioning as activators or repressors. Some of these factors can interact directly with nucleosomes to stabilise or destabilise them, thereby effecting their function of repression or activation (27). Examples of candidate complexes which appear to mediate factor loading on the DNA template include the SWI/SNF genes from *Saccharomyces cerevisiae* (28, 29) and *Drosophila melanogaster* nucleosome remodelling factor (NURF).

Multiple transcription factors bring regulatory information to the gene and execute their respective biochemical control functions by binding to DNA target sites, ancillary proteins and to the basal transcription apparatus (30). The regulation of gene expression is executed via distinct *cis*-regulatory regions (containing several transcription factor target binding sites) which have a modular organisation (31). Individual modules (each with a specific and individual function) interact with each other, as well as the basal promoter, in a variety of combinations to determine diverse spatial, temporal and quantitative patterns of gene expression (5). Adaptor proteins or direct interactions of transcription factors may

mediate the intercommunication of distant regulatory elements (sometimes separated by several hundred base pairs) by causing the intervening DNA to form loops. Direct evidence for DNA looping was first obtained by electron microscopy in a bacterial system (32), and since then several observations have confirmed these findings. DNA bending is another method which facilitates the communication between distant regulatory elements situated at distant sites which can be correlated with stimulation of promoter activity (33) and transcriptional initiation (34).

1.2 Examples of G·C-Rich Promoter Sequences

G·C-rich sequences have been identified in the regulatory domains of many different genes, in particular several housekeeping genes appear to have homo(pur).(pyr) stretches which are S1 nuclease sensitive and which are able to bind factors. Commonly, genes which possess multiple G·C-boxes in their 5' flanking regions may have multiple transcription initiation sites and lack a TATA-box (35).

The c-*myc* promoter, whose transcriptional activity correlates with a change in chromatin conformation (36) and is DNase I hypersensitive during transcription (37), has G·C-rich elements associated with promoter activity. These elements are bound by several zinc finger proteins (such as Sp1 and Zif87) and, together with their cognate proteins, function to regulate c-*myc* expression (38, 39). The type I collagen genes are coordinately regulated, and they, too, have been associated with a change in chromatin structure during transcriptional activity (40). The promoters of these genes are highly conserved amongst mammals (41), and it has been reported that G·C-rich elements may play an important role in enhancing the activities of the mouse, rat and human collagen gene proximal promoters (42, 43, 44). These binding sites are targeted by different classes of DNA-binding proteins, some of which are likely to be ubiquitous or cell specific transcription factors. Two G-binding factors (IF-1 (see section 1.3) and C-Krox (45)) have been identified as binding to the collagen promoter. The ornithine decarboxylase gene (whose activity regulates polyamine biosynthesis) is also highly conserved amongst human, rat and mouse (46). It contains G·C-rich sequences in the region -345 to -93, which is critical to the gene's expression. This region contains five Sp1 sites, and binds multiple nuclear proteins whose cooperative interactions may regulate expression. For instance Sp3 is able to inhibit Sp1-mediated *trans*-activation of the ornithine decarboxylase gene (47), and it is thought that the ratio of the two proteins in the cell is critical in the regulation of expression. The gastrin EGF response element (gERE) is a G·C-rich element with the target

site 5'-GGGGCGGGGTGGGGGG-3'. Sp1 is one of several factors which binds to this region, and it is thought to function in the developing and neoplastic stomach (48). Further, the human multidrug resistance (MDR1) promoter has two G·C-rich boxes in the 5'flanking region (-110 to -103 and -61 to -43), which both modulate promoter activity but have functionally distinct roles. The -110 G·C-box functions as a transcriptional "repressor" binding site, whereas the -50 G·C-box activates the basal promoter activity via the binding of Sp1 (49). Other examples of genes which have G·C-rich elements in their promoters include the luteinising hormone receptor (50), and the collagen II gene (51). This promoter interacts with various nuclear proteins and contains regulatory elements which include two G·C-boxes required for enhancer mediated transcription. A protein-mediated loop structure between the promoter and enhancer is implicated in the regulation of transcription of the gene.

The globin genes (especially the β-globin genes), have served as one of many model systems used to study developmental gene regulation (26). The mammalian and chick globin gene families form clusters of α– and β-like genes on two distinct chromosomal loci. Their sequential expression at distinct stages of development is exclusive to erythroid cells. The strict expression pattern is controlled mainly at the transcriptional level (26), although there is a complex regulatory interplay between far upstream regions, promoters, 3' flanking regions, as well as regulatory regions within the genes themselves, which may be mediated by transcription factors (52).

High level expression is probably controlled by regions far upstream (53), whereas tissue and stage specificity is conferred by proximal gene control regions (54). The locus control region (LCR) is a dominant control region located far upstream of the globin structural genes (~ 6 - 18 kb) (55) characterised by four DNase I hypersensitive sites (HS 1, 2, 3 and 4) (53). It is thought that the hypersensitive regions of the LCR interact sequentially with the individual globin genes as a unit or 'holocomplex' within the locus (56). Other HS activities include strong enhancer properties and a chromatin opening function (53). The LCR represents domains of altered chromatin structure, possibly due to the disruption or displacement of one or more nucleosomes by transcription factors (22, 57, 58), which is reminiscent of actively transcribed genes. The precise arrangement and spacing of the binding motifs present in the LCR (53, 59) affect DNA looping, which is mediated by transcription factors binding to both DNA elements within each HS site and other proteins at more proximal elements (53). DNA looping and protein-induced bending, and therefore proteins which function in the transcriptional regulation of globin gene expression during erythroid differentiation, may provide a mechanism whereby

multiple *cis*-elements (which are physically well separated along a stretch of DNA) can cooperate in the regulation of a downstream gene (60), possibly by direct interactions with the transcriptional machinery (61). Similar to other genes, the globin gene promoters have numerous sequence motifs recognised by various transcription factors, examples include the TATA motif, CCAAT, CACCC and GATA-binding sites (62). The speculation that these proteins may interact to form higher order protein-DNA structures is supported not only by the multiple occurrence of some regulatory sites in close association with each other (eg CACCC and GATA-1 sites, or GATA-1 and YY-1 elements (63)) but also by the finding that several protein-protein interactions occur amongst the associated transcription factors. For example GATA-1 (a zinc finger protein erythroid restricted protein (64, 65)) can self-associate (63, 66) and interact synergistically with other transcriptional activators (viz the Krüppel-like zinc finger protein EKLF and Sp1) to activate transcription (62). Other interactions include TAL1 / SCL and RBTN2 (67). The process of gene switching may rely on the stable interaction between the promoters of individual globin genes and specific subdomains of the LCR. These interactions are probably facilitated by protein-protein associations amongst lineage specific, ubiquitous and / or stage specific factors, possibly resulting in and stabilising DNA looping (62).

Numerous other factors whose precise role in erythroid differentiation is unknown have been identified. Examples include Nuclear Factor-Erythroid 2 (NF-E2), a basic-leucine zipper which binds AP-1 like sites (52, 68), Nuclear Factor-Erythroid 4 (NF-E4) which functions by binding downstream of the β-globin genes (69), and TAL1/SCL and RBTN2, which are required for normal erythroid development (70, 71, 72). More widely expressed factors which contribute to erythroid differentiation and development include GATA-2 (73) and Yin-Yang-1 (YY-1) (74). Erythroid cells also have several members of Krüppel-like factors, eg Sp1, YY-1, EKLF and BKLF / TEF-2 (26). The Erythroid Krüppel-like factor (EKLF) is one of several proteins which binds the CACCC motif (in general 5'-CCNCNCCCN-3'), and is thought to be involved in gene switching (75).

The chick adult β-globin gene promoter has been analysed extensively in an attempt to understand how chromatin structure relates to gene transcription. The proximal promoter is characterised by a nuclease hypersensitive domain (nt -260 to -60) (76) containing many small DNA elements which bind proteins. In particular this domain contains a G·C-rich region of 16 - 18 consecutive G residues (77), which, in supercoiled plasmids, is able to form unusual DNA structures (78). No functional role has been assigned to this G-string yet, however, it is postulated to have a role in gene regulation and it could function *in*

vivo as a conformational switch to aid β-globin gene expression, since it is able to bind factors (eg BGP1, see section 1.3). The G·C-rich region can undergo DNA conformational changes, and the displacement / exclusion of a nucleosome is associated with this region, all of which are features associated with actively transcribed genes (11).

1.3 Examples of G·C-Binding Proteins

A variety of transcription factors have been reported which bind to G·C-boxes, these include both positive regulators and negative *trans*-acting factors. It is likely that various factors which have the same or similar sequence specificity can interact with G·C-rich target sites to elicit different functions, allowing flexibility in the regulation of transcription. Some examples of G·C-rich binding factors are listed below.

The beta globin protein 1 (BGP1) is a well studied example of a G-binding factor which is thought to be involved in globin gene regulation (see section 1.3 and (11)). It is a protein with a characteristic molecular weight of 66 kDa, isolated from chicken erythrocytes. The minimum recognition sequence for BGP1 is a G_7-string (11), and its DNA-binding ability has an absolute requirement for zinc. Specifically, the BGP1 binding site lies within the borders of the positioned nucleosome. The tissue-specific expression of the chick adult β-globin gene correlates with the alteration in chromatin structure of the G-string. While the gene is silent, a nucleosome is positioned over the G-string, however in the actively expressed gene the nucleosome is absent and BGP1 binds to the G-string instead (11). More recent evidence shows that BGP1 does not function to create a nucleosome-free promoter (69). BGP1 is a tissue and developmental stage specific factor which has been implicated in gene switching, however there is no evidence for its biological role.

Analysis of the various *cis*-acting elements in the mouse α1(I) collagen gene promoter indicates that several factors may be involved in the coordinate regulation of these genes (79). In particular, it was shown that two transcriptional repressors, inhibitory factors 1 and 2 (IF-1 and IF-2), specifically bind to segments which have strong promoter activity. The factor IF-1 binds to a G_7-string (similar to an Sp1 target site) in the α1(I) and α2(I) collagen gene promoters, and is implied in mediating the developmental regulation of the respective genes (14).

The human late histone H4 gene-specific transcription factor, H4TF-1, is a DNA-binding protein thought to be involved in the promoter regulation of the histone gene H4 (80). Two H4 specific transcription factors, viz H4TF-1 and H4TF-2, were co-purified using ion exchange and affinity chromatography. H4TF-1 potentiates the expression of the target gene by interacting with G·C-rich sequences which are required for the maximal expression of the H4 gene (80).

The BTE-binding proteins (BTEB and BTEB-2) have three repeated zinc finger motifs which are similar to Sp1 (81). BTEB is a ubiquitous protein, whereas BTEB-2 is found predominantly in the placenta and testis. (The latter was cloned from human placenta using the BTEB cDNA as a probe.) BTEB stimulates promoters with repeated G·C-box sequences, and BTEB-2, too, is a G·C-rich DNA-binding protein (81). The binding specificity of both proteins is identical to the Sp1 protein (see section 1.3.1). Indeed, the three amino acids within the classical zinc finger structure considered important for DNA sequence recognition are invariant in BTEB-2, BTEB and Sp1. However, immunological supershift experiments can distinguish between the binding of the BTE-binding proteins and Sp1, since (apart from their DNA-binding zinc finger domains, which exhibit a 72 % between Sp1 and BTEB and a 59 % similarity between Sp1 and BTEB-2) these proteins have little or no similarity.

BTEB-2 has been relatively well characterised. It is a 219 amino acid protein, containing three domains specific to transcriptional regulators. The DNA binding domain consists of three zinc finger domains at the protein's C-terminus (these are 59 % similar to the Sp1 zinc finger domain (81)). BTEB-2 also has a basic region which partially identifies with the basic domain of proteins characterised by the helix-loop-helix and leucine zipper motifs. This protein has a proline-rich region (16 out of 67 residues) between amino acids 44 - 110, which is very likely to constitute a transcriptional activation domain, as in other proline-rich DNA-binding proteins (82). Transient expression studies indicated that BTEB-2 is able to activate the expression of CAT (chloramphenicol acetyl-transferase) on G·C-box containing reporter plasmids, ie BTEB-2 is implied as a transcriptional activator (81).

G·C-binding Factor (GCF) is a 91 kDa protein isolated from A431 cells (83). It has a characteristic basic region at its N-terminus which functions as the DNA-binding domain. The protein recognises G·C-rich sequences. (The binding site is 5'-GCGGGGC-3', which can also be recognised by Sp1 and ETF.) Cotransfection experiments imply that GCF acts as a sequence-specific repressor. It may function either by competing with various activators for DNA-binding sites, or alternatively, it may interact with other

proteins to achieve repression. This protein has an acidic C-terminus which is required for its full activity (83). It is also characterised by two leucine zipper motifs which are proposed to facilitate dimerisation and are common to DNA-binding proteins such as C/EBP, Myc, Fos, GCN4 and Jun (84).

The *trans*-acting epidermal transcription factor (ETF) of 120 kDa binds to G·C-rich regions and has been found to specifically stimulate the transcription of the epidermal growth factor receptor (EGFR) gene (35). The promoter for this gene characteristically lacks a TATA-box and CCAAT box, yet it is highly G·C-rich. ETF (together with Sp1) stimulates its transcription. The ETF protein is implied as a specific transcription factor for several promoters which do not contain TATA elements, since it appears that the presence of a TATA element interferes with the functions of ETF (35). It recognises various G·C-rich sequences (including stretches of poly(dG).poly(dC)) with similar affinities. The core binding sequence of 5'-CCCC-3' was deduced from various binding sites. Otherwise it has a very loose sequence requirement, ie no rigorous consensus sequence has been derived for its binding.

1.3.1 A Subfamily of Zinc Finger Proteins with G·C-Rich Binding Sites

The Cys_2/His_2 class of zinc finger DNA-binding proteins is characterised by the particular sequence of amino acids which forms structural domain(s) that bind a zinc (II) ion. This domain has a β–hairpin followed by an helix (85) and constitutes an important eukaryotic DNA-binding domain, usually characterised by tandem repeats of a conserved motif of 20 - 30 amino acids. The variation of the amino acid sequence within each zinc finger domain contributes individually to the DNA-binding affinity and specificity of the protein as a whole (ie the fingers appear to be modular in nature), which is one of the factors determining DNA-binding preferences and degeneracy of binding within target sites (86). The best understood members of this class of proteins belong to a subset of DNA-binding proteins which bind to relatively G·C-rich target DNA sequences, examples include Sp1 (87), members of the Wilm tumour family (eg Zif268 or NGFI-A, Krox-20,24 and Efr21,2) and yeast ADR1 (88). ADR1 and Zif268 are two of the best studied members of this protein subfamily, and they have revealed much that is known about the zinc finger-DNA interaction. The ADR1 protein has a DNA-binding domain which is characterised by two zinc fingers, and generally the protein binds DNA target sites as a dimer (89). The target sites have approximate dyad symmetry and the consensus binding site appears to be 5'-TTGGAG-3'. The three zinc finger domains of Zif268 (also called NGFI-A or *egr1*) preferably bind to a consensus

sequence 5'-GCGGGGGCG-3'. The Zif268 protein was bound to cognate oligonucleotides and, using a combination of altered amino acid sequences, as well as variations within the DNA target sequences, has provided much structural information regarding zinc finger protein-DNA interactions (90). For instance, it was determined that three particular amino acid positions within each zinc finger are the main contributors in determining the DNA site preference of the protein. This accounts for the DNA-binding site degeneracy of a single protein and the many aspects of variations in the known binding sites of the members of this class of DNA-binding proteins in general.

The zinc finger DNA-binding protein Sp1 (stimulatory protein 1) is one of the first mammalian transcription factors that was characterised (87). It is a ubiquitous protein of about 100 kDa, which is abundant in most cell lines, but it varies substantially in different tissues during development (91). The protein binds G·C-rich target sites and related motifs (92) present in many cellular and viral promoters. The DNA-binding domain of this protein is characterised by three zinc fingers (87), and it binds in the major groove of the DNA target site (93). The G·C-rich target sites for Sp1 are variations of the sequences 5'-GGGCGG-3' and 5'-GGGGCGGG-3'. Sp1 binding sites often appear in promoter regions in clusters, allowing the Sp1 protein to act synergistically through adjacent binding sites (94), ie enhanced promoter *trans*-activation results from the mediation of at least two binding sites (95). Sp1 not only interacts with itself to form multimeric complexes (95), it has also been implicated in the interaction with other transcription factors, such as NF-κB (96), GATA-1 (the major erythroid transcription factor (62)), as well as select components of the basal transcription apparatus (98). It is therefore able to interact and cooperate with other proteins to direct expression. The promoters and enhancers of numerous genes contain multiple binding sites for the same or for several different sequence-specific proteins (located within several hundred base pairs of each other), allowing additive and even synergistic activation of transcription, as exhibited by distal and proximal Sp1 sites (99). The ability of Sp1 to self-associate (100) suggests a mechanism whereby distant DNA segments are joined by looping of the intervening DNA as a result of stabilising protein-protein interactions. These interactions amongst proteins from multiple binding sites confer transcriptional responsiveness.

The Sp1 protein is involved in both constitutive transcription and the regulation of several inducible genes (49, 101). It is an activator protein, whose transactivation function lies in the glutamine-rich N-terminus, containing the A and B domains (102, 103). The two other domains (C, which is weakly basic, and D) are adjacent to the zinc finger region. They also influence the protein's transcriptional activation

function (103). In particular, it has been found that domain D plays a key role in the synergistic action of Sp1 mediation (95).

As indicated above, many promoters contain elements which are capable of binding Sp1 in addition to other transcription factors. Sp1 levels are abundant in the cell, and the protein coordinately regulates many different genes, therefore the access of Sp1 to the promoter must be regulated in order to ensure some specificity of control. Several mechanisms may be involved in these regulatory mechanisms, including phosphorylation or glycosylation (104, 105), regulation of the affinity of Sp1 for DNA (106), alteration of its *trans*-activation potential (107), regulation of its nuclear levels (108) and regulation of its concentration relative to other transcription factors (93).

More recently it was discovered that the DNA-target sites bound by Sp1 are bound with equal affinity by two human proteins, Sp3 and Sp4 (109). Sp4 is expressed in certain cells within the brain, and may have a role in the expression of certain genes within these cells, however the natural target genes of Sp4 *in vivo* have not been identified yet. Together, Sp1, Sp3 and Sp4 represent a family of G·C-box binding proteins which have very similar structural features (110). They have very conserved zinc finger DNA-binding domains, which are close to the C-terminus. Their N-termini are characterised by stretches of glutamine- and serine- / threonine-rich amino acids. The striking structural homology between Sp1, Sp3 and Sp4 does not reflect functional equivalence, though (110). Sp3 is a transcriptional inhibitor (110, 111) or an activator (112, 113), therefore the role of SP3 may be context- or cell-dependent. Studies on the transcriptional properties of Sp4 have revealed that it is an activator protein like Sp1, but it does not act synergystically through adjacent binding sites (108) since it lacks the functionally active domain for this property. Sp4 exhibits various other unique properties with respect to Sp1, for instance it may be a target for Sp1 activation (110).

The nerve growth factor-induced early response gene encodes a Cys_2/His_2 zinc finger protein of 50 kDa called NGFI-C (114), whose induction is very rapid but transient, as shown by induction of PC12 cells using NGF (Nerve Growth Factor). NGFI-C is a G-binding protein containing three zinc fingers, and characteristically binds to the site 5'-GCGGGGGCG-3' from which it is able to strongly activate transcription as shown by CAT reporter constructs (114). Three other proteins bind to the same target site (viz NGFI-A, Krox-20 and the Wilm's tumour gene product), and have considerable homology to NGFI-C in the DNA-binding domain only. Similarly Sp1 also has a high homology over this region. Another

characteristic feature of NGFI-C is its unusually high proline composition (25 %) over 77 amino acids, which is reminiscent of a transcriptional activation domain as found in several other transcription factors, such as AP-2, Oct-2 and c-Jun (114), and hence this region may have an analogous function.

1.4 Sea Urchins: Model Systems Used to Study Developmental Regulation

1.4.1 Regulatory Genes and *Trans*-Regulatory Factors in Sea Urchin Embryogenesis

Sea urchin embryos share some basic characteristics with a wide variety of invertebrates (115, 116). They belong to the echinoderms, which, of all the invertebrates, are considered to be most closely related to chordates and hence to vertebrates (117). The ready availability, structural simplicity and rapid development of sea urchin embryos have made them a model system for investigations regarding the molecular mechanisms underlying differentiation, region specification, transcriptional regulation and morphogenesis by focusing on transcriptional regulators (118). Several sea urchin embryo *trans*-regulatory factors (which may have a broader role within developmental control) have been identified as a result of characterising spatially and temporally controlled structural genes involved in development.

1.4.2 Spatial Regulatory Systems in Sea Urchin Embryos

Examples of putative regulatory elements controlling the timing and localisation of gene expression in sea urchins (reviewed by Maxson and Tan (1994) (118)) may restrict the expression of certain genes to the aboral ectoderm, the skeletogenic mesenchyme, as well as the vegetal plate (and its derivatives). Various regulatory factors which bind these elements have been identified. Some have been isolated and their cDNAs obtained.

The CyIIIa actin gene is an example of an aboral ectoderm specific gene whose expression is spatially, temporally and quantitatively controlled in the embryo (119). It is activated during late cleavage stage and serves as a marker for aboral ectoderm specification. The *cis*-regulatory domain (divided into the proximal, middle and distal modules, which function separately but are quantitatively interdependent on each other) spans more than 2.3 kb and binds a minimum of nine transcription factors at more than

twenty randomly distributed sites (31). The proximal module activates early gene expression, while the middle module controls late spatial expression. The distal cluster contains a high occurrence of SpGCF1 sites (these have G·C-rich sequences), which are thought to mediate a positive function (3). The CyIIIa target sites have been used to isolate several nuclear DNA-binding proteins, and the functional significance of most of the interactions (except that of the SpGCF1 protein, see section 1.4.4.2) are known. Examples of proteins which interact in this region include SpGCF1 (see section 1.4.4.2), SpP3A2 (which has a basic helix-loop-helix DNA-binding domain (120)), and SpP3A1 (120), which is a zinc finger protein that binds to specific P3A sites. These sites probably have a negative spatial regulatory function throughout the regulatory domain (119), confining expression of CyIIIa to the aboral ectoderm (121). They are also contained in the promoters of other genes which have different expression patterns to CyIIIa actin, eg skeletogenic SM50 and L1 H2B histone genes (122). The P1 protein which binds to the CyIIIa actin gene promoter is uncharacterised as yet, however the P1 target site is required for normal functioning. SpRunt1 is another factor which provides a major positive input, whereas negatively acting factors SpP7II (which binds to the negative spatial regulator P7II (123)), and SpZ12-1 confine expression to the aboral ectoderm from gastrulation onwards (31).

Another example of a gene whose expression is limited to the aboral ectoderm includes Spec2a (124). The promoter elements which specify the temporal and spatial regulation of this gene include an upstream region and three orthodenticle (Otx) sites (the consensus binding site is TAATCC) which bind a group of homeoproteins (SpOtx) from sea urchin and which have counterparts in Drosophila and mouse. SpOtx may function in the activation of the Spec2a gene in the aboral ectoderm (124).

The Endo 16 gene codes for a cell surface glycoprotein whose function is unknown, and its expression is restricted to the vegetal plate of the blastula stage embryo, continues throughout the archenteron (to which the vegetal plate gives rise) in gastrulation (125), and transcription is eventually shut down in all other regions except the midgut where it is increased. A 2.2 kb fragment upstream of the start site determined correct reporter gene expression (125) and the specific sites of protein-DNA interaction within the *cis*-regulatory domain have been mapped in order to elucidate which ones are necessary or sufficient for normal expression. The output of each module is not well defined yet, but it appears that the positive regulatory functions of the proximal and distal modules are curbed by negative interactions preventing incorrect expression in the adjacent skeletogenic and ectodermal territories (126). Studies show that a minimum of 13 factors (whose identities and functions have mostly not been determined yet)

bind to 30 discontinuously distributed target sites, including many SpGCF1 sites (31). Most of the modules also include SpGCF1 sites which could function in intermodule communication (see section 1.4.4.2).

The muscle cell lineage in sea urchin development may be regulated by SUM-1, a member of the Myo-D family of basic helix-loop-helix transcription factors (127). The temporal expression pattern of this factor coincides with the ability of a subset of secondary mesenchyme cells of the vegetal plate to differentiate into muscle cells during gastrulation (128). It is considered as one of several influences which lead to the commitment of secondary mesenchyme cells into muscle cells, since it is able to bind and *trans*-activate muscle specific enhancer elements in sea urchin embryos, as shown by microinjection experiments.

1.4.3 Temporally Regulated Sea Urchin Histone Genes

The early sea urchin histone genes are tandemly repeated in the genome, ie there are multiple copies of a gene battery, each of which codes for one H1, H2A, H2B, H3 and H4 gene. During early embryogenesis the early histone genes are coordinately expressed in a distinct temporal pattern. However, after blastulation, sea urchins express the late histones and the early genes are switched off, never to be expressed again. The expressed genes have nuclease hypersensitive intergenic spacers, whereas the shutdown of the genes correlates with the presence of a well defined nucleosome pattern. The control mechanisms governing this gene switch have not been elucidated yet. Several *cis*-elements have been defined for both the early (α) and late (β) histone genes (129). The early histone genes are expressed at very high levels from oogenesis through to blastula stage embryos, partially as a result of positive regulatory elements, many of which have been identified in the promoters of (α) H1, H2A, H2B, H3 and H4 genes (130, 131, 132, 133, 134). Negative elements in the promoters of the H3 and H4 genes cause inactivation of expression of the early genes in late-stage embryos (135, 136) after the initiation of late histone gene expression which is characteristic of blastula to adult stages.

Nuclear proteins which bind to these DNA regulatory sequences have been implicated in the developmental regulation of the sea urchin histone genes. For instance, SpOct (a POU II class gene (137)) is the major octamer binding protein in early sea urchin embryos, and is able to bind the octamer site of the αH2B gene which is essential for the correct timing of expression. Additional evidence implying that

this transcription factor may be a key regulator in the temporal control is suggested by the correlating temporal expression patterns for both SpOct and the αH2B gene. Several elements responsible for activating gene expression and candidate regulatory proteins have also been identified for the sea urchin late (β) histone genes (138, 139, 140, 141, 142). SSAP (stage specific activator protein) belongs to the class of RNA-binding proteins (118) and is a very strong candidate for the specific stimulation of transcription of the βH1 gene in blastula stage embryos, since it is able to bind the βH1 temporal enhancer, which is a stage specific enhancer (139). The L1 H2B gene is *trans*-activated by a homeodomain protein of the Abdominal B class which binds a 3' enhancer element (142), and may be encoded by a cDNA (*Hbox4*) which is similar to *Hox-C9* (143). It may therefore also function in pattern formation, as do other homeobox gene products. The developmental profile for the *Hbox4* mRNA closely resembles that of L1 H2B, reinforcing the association between them. The H2A-2 and H2B-2 histone genes have degenerate sets of elements in their promoters to which a putative temporal regulatory protein TSAP (tissue specific activator protein) binds (141). This protein has not been characterised further, however another protein, designated BSAP (B-cell specific activator protein) with identical DNA-binding specificity has been identified, which may function in B-cell specific gene expression and B-cell differentiation (144). The protein suGF1 (see section 1.4.4.1) binds a G·C-rich region in the H1-H4 spacer *in vitro*, which can also be occupied by a nucleosome (145) and therefore may play a role in modulating the stability of the nucleosome, as well as in the regulation of the early histone genes.

1.4.4 Examples of G-Binding Proteins in Sea Urchins

1.4.4.1 suGF1: a Nuclear Sea Urchin DNA-Binding Factor

Sea urchin G-binding factor 1 (suGF1) is a 59.5 kDa sea urchin DNA-binding protein which has been identified in 14 hour sea urchin (*Parechinus angulosus*) embryos (145). This factor is present in the early embryonic developmental stages. It has high binding affinity and specificity for G·C-rich DNA sequences and it binds to both poly(dG).poly(dC) and oligo(dG).oligo(dC) containing sequences by reproducibly forming two characteristic bands in electrophoretic mobility gel shift assays (1). The consensus recognition sequence was established as 5'-GGGNGGG-3' or 5'-GGGGGGC-3' (1), however it is possible that there is further degeneracy within the binding site. Both bands formed by electrophoretic mobility gel shift assays contain the suGF1 protein, however it has not been established whether the two bands

arise as a result of posttranscriptional modification or truncation of the protein. The formation of these bands is strongly dependent on ionic strength (optimal DNA-binding occurs at 175 mM KCl) and the binding is also dependent on divalent cations, since dialysis of suGF1 with EDTA partially abolishes DNA-binding (1). This originally suggested that suGF1 may belong to the zinc finger class of DNA-binding proteins. The divalent cations appear to be very tightly complexed with the factor, as the purification buffers do not need to be supplemented with divalent ions such as Zn^{2+}. suGF1 is able to form sequence-specific multimers via suGF1-suGF1 interactions (145), and it is possible that the protein could be involved in looping DNA, thereby bringing distant regulatory regions into close proximity.

1.4.4.1.1 suGF1 Interacts with the H1-H4 Intergenic Region

The H1-H4 intergenic region of the early histone gene battery contains a run of 11 Gs, and provides a binding site for suGF1 *in vitro*. The interaction of suGF1 with this region has been extensively studied and characterised (1, 2, 145). The sea urchin early histone gene battery is developmentally regulated and the genes are coordinately expressed in a distinct temporal pattern. Expression of the histone genes is associated with intergenic regions which are nuclease hypersensitive, whereas the shutdown of gene expression correlates with well-defined spaced nucleosomes (146). Thus alterations in chromatin structure of the early histone gene battery correlate with the temporal expression pattern of the early histone genes. Specifically, it has been established that there is a nucleosome positioned over the H1-H4 fragment *in vitro*, and the G_{11}-string lies close to the dyad of the positioned nucleosome core. The nucleosome positioning signal lies over the sequence $(GA)_{16}(G)_{11}$ which, *in vitro*, has the ability to form an unusual DNA structure (triple helix), under conditions of negative superhelical stress and low pH (20, 147).

Using a combination of footprinting and methylation interference studies a model for the binding interaction of suGF1 with the G_{11}-string in the H1-H4 intergenic fragment has been developed (1). It was shown that suGF1 approaches the double helix mainly from one side, and is closely associated with the DNA over about 1.5 helical turns. A bulky structure of the protein protrudes into the major groove and contacts the central Gs located in this region, as well as the phosphate backbone bordering on the run of Gs (1). Both the free and the bound DNA are curved, which can be deduced from the periodic narrowing of the minor groove on one side of the DNA helix, and the periodic widening of the major groove on the

opposite side, as elucidated by hydroxyradical footprinting. The binding of suGF1 to the DNA is consistent with the direction of curvature of the free DNA when wrapped around a nucleosome core, and suGF1 binding stabilises the curvature of the DNA fragment, implying that the protein could play a role in the alteration of chromatin structure of the H1-H4 intergenic fragment (145). It appears that suGF1 has a single preferential frame of binding to the intergenic fragment (2). The latter is able to form several distinct nucleosome core species which have different translational settings but very similar rotational placing, in that the octamer surface is positioned away from the suGF1 binding site. Despite the fact that the rotational setting of the nucleosome core maximally exposes the suGF1 binding site, the binding of suGF1 and a nucleosome core are mutually exclusive in most nucleosome species, due to steric hindrances (2). The more internal the recognition site is within the core DNA, the more the histone proteins of the positioned nucleosome sterically clash with the binding of suGF1 to its recognition site. It is therefore possible that the DNA binding site has to be exposed to the suGF1 protein in a regulated fashion, unless the transcription factor gains access to the DNA shortly after replication (2).

1.4.4.1.2 suGF1 Is Implicated In Gene Regulation

There has been no direct evidence for a biological role for suGF1 yet. However there is strong indirect evidence for its role as a transcriptional regulatory protein. It is possible that the protein may be involved in gene regulation of the sea urchin histone gene battery via alterations in chromatin structure within the pur.pyr region (145). suGF1 binds sequence-specifically and with high affinity to oligo(dG).oligo(dC) regions. The G_6-string in the LpS1β gene promoter also provides a recognition site for suGF1. This G·C-rich sequence is known to be a *cis* positive regulatory element, mutations within the G-string abolish transcriptional activity in functional assays (16). suGF1 interacts with the G_6-string by forming the same characteristic bands as it does with the G_{11} sequence in the H1-H4 intergenic fragment. These complexes are very similar to the ones formed by the ectoderm G-string factor (16) when interacting with the same probe. Not only do these proteins have indistinguishable target site specificities, they also bind the probe under the same optimum ionic strength. The similarities observed between these proteins strongly suggest a structural or functional relationship between them.

The chicken β globin gene promoter and the H1-H4 intergenic fragment also have several properties in common. For instance, the globin gene promoter contains a G-string (16 - 18 Gs) which lies on the border

of a positioned nucleosome and the factor BGP1 (see section 1.3) binds in this region. Even though it is possible that there are several frames in which BGP1 binds to its recognition site in the β-globin gene promoter, methylation interference studies established that the main interaction occurs with the central 7 Gs. The binding of factors to the G-string appears to correlate with a conformational change in the DNA and the disappearance of the nucleosome (11), all of which are closely associated with the expression of the β-globin genes. Similarly, the H1-H4 fragment contains a G_{11}-string, which lies within the positioned nucleosome and close to the dyad of symmetry, and the main interaction of suGF1 is with the centrally located Gs within the recognition site. The state of expression of the early histone genes is also correlated with a change in chromatin structure. suGF1 is able to bind the G_{11}-string in the H1-H4 intergenic region and the β-globin G-string with equal affinity. Indeed, suGF1and BGP1 produce identical footprints on the β-globin gene promoter. Hence it appears that the biochemical and DNA binding properties of the two factors are very similar, and it is possible that suGF1 and BGP1 are structurally or functionally related, despite the fact that there are several differences between them (such as distribution, size and requirement of zinc for DNA-binding).

The implication of suGF1 as a transcription factor raises several issues. For instance it could be proposed that suGF1 may be a member of a family of G-string factors involved in developmental regulation of several unrelated genes in various organisms, possibly by functioning in the alteration of chromatin structure. It is possible that suGF1 may have a similar role in gene regulation as BGP1 (11), or it could be the *Parechinus angulosus* homologue of one of several other sea urchin G-binding proteins, such as the ectoderm specific G-string factor (16), or SpGCF1 (3), since there is evidence that several unrelated sea urchin genes share upstream G·C-rich regions which may be involved in their regulation (16).

1.4.4.2 Other G-Binding Proteins in Sea Urchin Embryos

The Lps1β gene is a sea urchin cell lineage specific gene (16). The gene promoter contains a G_6-string, which is a positive *cis*-regulatory element. Functional assays have shown that mutations within the G·C-rich sequence abolish *cis* activity (16) and therefore this region is postulated to be important in sea urchin development. The Lps1β *cis*-regulatory domain is bound by different nuclear proteins, one of which is a G-string binding factor. This G-binding protein is an ectoderm specific factor present in the nuclear extracts of gastrula and blastula stage embryos (16), ie it is differentially localised in the embryo. EMSAs

performed with this protein and the G·C-rich promoter exhibit a characteristic slow mobility doublet. It has been suggested that the ectoderm specific factor may be similar to the mammalian IF-1 (see section 1.3) factor which binds the α1 and α2 collagen gene promoters (14), or it could be a homologue of suGF1 (see section 1.4.4.1).

Another example of G·C-rich target sites in sea urchin genes which bind nuclear proteins is a C_4 core element (referred to as either a "P8", "P2" or "SpGCF1" site) which has been identified to occur severalfold throughout the entire gene regulatory regions of both the Endo16 gene and the CyIIIa cytoskeletal actin gene. These target sites occur on genes which are expressed in different embryonic territories (3). Most of what is known about the SpGCF1 protein derives from the CyIIIa actin gene, which contains clusters of SpGCF1 sites in the distal regulatory domain in a variety of patterns, and often in close proximity to other DNA-binding sites (3). The biological function of the SpGCF1 sites has not been determined yet, however experiments using transgene chloramphenicol acetyl transferase (CAT) constructs and *in vivo* competition with excess target sites indicate that SpGCF1 sites have a positive transcriptional activating function (119, 148). The SpGCF1 sites are bound specifically by five proteins present in blastula-stage nuclear extracts isolated from *Strongylocentrotus purpuratus* as shown by EMSAs (3, 121). The protein which gives rise to the slowest migrating complex in the assay has been identified as a 55 kDa factor (SpGCF1), and from quantitative simulation studies it is predicted that the protein binds DNA as a dimer (149).

Distinct transcription factors bind to several regulatory sites in both the CyIIIa and Endo16 gene promoters, and some are able to bind at multiple sites, for instance SpGCF1, whose binding sites occur in clusters and often lie close to binding sites for other transcription factors (3). The regulatory domain of the CyIIIa gene is divided into three distinct regions with separate functions, all of which are required for normal embryonic expression of the CyIIIa genes (123, 119). The proximal region regulates the spatial expression of the gene, whereas the distal region controls the level of expression, and the middle region controls spatial and temporal expression, but only in conjunction with portions of the proximal module (31). This implies that gene expression may be influenced by intercommunication between the regulatory sites, for instance via the interaction of positive and negative regulatory transcription factors which bind to target sites contained on the proximal and middle modules (31). EMSAs and electron microscopy indicate that SpGCF1 molecules (which have target binding sites throughout the regulatory domain) may associate with each other to form multimers (149). If this reflects the *in vivo* situation, the DNA-bound

protein molecules could associate with each other, promoting the formation of loops in the intervening DNA (3). Thus SpGCF1 could function in bringing distant regulatory domains into close proximity, allowing stable functional complexes constituted of distal transcription factors and regulatory elements to form and intercommunicate, thereby increasing productive interactions and stimulating gene expression (31). SpGCF1 may be involved in developmental regulation of several unrelated genes, which could account for its prevalence in sea urchin embryos and for the positive function of the SpGCF1 target sites.

1.5 Purification and Identification of Transcription Factors

Characterisation and identification of a newly isolated protein is often reinforced by analysis of its primary structure, which aids in the cloning of the corresponding gene or cDNA (see section 1.6). It is estimated that DNA-binding proteins and transcription factors constitute about 0.001 % of total cellular protein (92). Calzone et al (1988) (150) estimate that the minimum prevalence of factors binding to the 5' CyIIIA cytoskeletal gene in the late cleavage-stage sea urchin embryo are as few as several hundred to thousand molecules per nucleus. Therefore the rarity of these transcription factors makes their purification to homogeneity a time-consuming and relatively difficult task, requiring enormous amounts of starting material and highly selective purification strategies of high yield. Several sensitive techniques have been developed whereby sequence-specific DNA-binding proteins can be detected, characterised or monitored during their biochemical fractionation (eg nitrocellulose filter binding, DNase 1 footprinting, and EMSAs (151)). The mobility gel shift assay is the most common technique used to detect the interaction of sequence-specific proteins with small amounts of a particular DNA fragment in solution (152). Tightly bound protein-nucleic acid complexes are stable during electrophoresis under non-denaturing conditions, and even weak complexes can often be resolved because they are stabilised by the "cage effect" created by the gel matrix, ie the diffusion of the protein away from the DNA is hindered by the matrix, which in turn favours reassociation (153). The binding takes place in solution in the presence of nonspecific competitor DNA, and once subjected to electrophoresis through polyacrylamide, the mobility of the protein-DNA complex is retarded with respect to the unbound DNA. As the two populations of DNA (free and protein-bound) enter the gel matrix, they are separated physically, and the bound protein can no longer affect the mobility of the free DNA (154). This results in the formation of discrete bands by individual protein-DNA complexes which have lower mobility than the free DNA

fragment. The difference in mobility of the discrete populations of bands is usually analysed by autoradiography of radiolabelled DNA (151).

The development of sequence-specific DNA-affinity chromatography (92) has greatly facilitated the purification of low abundance transcription factors in general. This method was originally developed using nonspecific DNA attached to cellulose or agarose supports (155). However several variations (different combinations of specific or nonspecific DNA sequences linked to several types of resins) have been described and used successfully to isolate a variety of transcription factors, including Sp1, AP-1, AP-2, NF-κB, Pit-1 and CBP-1 (155). Several types of solid support resins can be used, eg cellulose (156) or agarose (92), and DNA can be linked to them using a variety of interactions, such as CNBr, biotin-avidin and streptavidin (155, 157). Commonly DNA affinity chromatography involves preparation of an agarose resin activated by cyanogen bromide, and subsequently the DNA (containing the relevant protein binding site) is covalently linked to the solid support (155). The linked DNA is generally one of two types. Either it is a plasmid containing multiple protein binding sites (156) or it constitutes a catenated, synthetic, double stranded oligonucleotide representing the high affinity protein binding site. Oligonucleotides are generally prepared for coupling to the resin by annealing complementary strands, 5' phosphorylation, and ligation of the oligonucleotides to form oligomers (158, 155, 159).

A crude protein sample which is to be purified using DNA affinity chromatography needs to be nuclease-free, therefore the purification of transcription factors is generally approached by partial purification of a crude protein extract using conventional chromatography. This enables the removal of nucleases and other contaminants in the sample (92). Usually the protein sample is pretreated with a nonspecific competitor DNA to which the protein of interest has very low affinity, and finally the mixture is applied to the affinity column at low ionic strength. The retained proteins (which are specifically bound) are eluted by increasing salt concentration (155), whereas proteins having little or no affinity for the DNA-resin flow through the column. The expected enrichment for an affinity purified factor lies between 500- to 1000- fold (155). Subsequent separation of the sample by SDS-PAGE will usually suffice to purify a protein for most applications, including protein sequencing. Several DNA-binding proteins have been isolated to homogeneity using above procedure, examples include SpP3A2 (120), SpOct (137) and SpGCF1 (3).

Cloning a gene or cDNA coding for a transcription factor is usually achieved by producing sufficient protein sequence information to synthesise oligonucleotide probes in order to screen genomic or cDNA libraries (see section 1.6.1.1). Classical sequencing techniques (ie Edman degradation) involve the removal of one amino acid at a time from the N-terminus of the intact protein (or from internal peptides generated by proteolytic cleavage) by means of chemical reagents, with subsequent analysis of the released amino acid derivative (160). The traditional microsequencing technique sets the lower limit in sensitivity. Most protein purification strategies involve a combination of various fractionation techniques (chromatography, selective precipitation, dialysis, etc) yielding a highly purified protein sample which is often only available in minute and limited quantities (eg nanogram levels) and not always in a form that can be directly submitted for sequence analysis. Adsorptive losses make it difficult to handle low quantities of protein, which emphasises the need for highly sensitive analytical techniques in order to identify protein sequences (161). Mass spectrometry provides an alternative strategy to classical sequencing (162). In addition to detailed primary sequence information, the mass spectral results can yield purity evaluation, as well as molecular mass determinations of proteins (161). Two of the main advantages posed by mass spectrometry over classical sequencing are (i) mass spectrometry operates on picomole to femtomole sensitivities for both molecular mass determinations and partial to complete amino acid sequence information (163), and (ii) a mixture of peptides can be introduced into the mass spectrometer (164). Site specific proteolysis of the protein of interest generates small quantities of peptides which can easily be handled by combining microcolumn liquid chromatography (LC) with tandem mass spectrometry (MS/MS) (165). Electrospray ionization is used to introduce the peptide ions to the mass spectrometer, and ions of a single mass to charge ratio are targeted selectively for collision induced dissociation (CID) using an inert gas, which leads to the fragmentation of the peptides, generating a reproducible and generally predictable signal pattern characteristic of the specific amino acid sequence (161). The mass spectrometer approach has been successfully applied to both proteins and peptides (166) and the spectrum which is generated is used to search a protein database in order to identify the protein of interest, which involves correlating the predicted fragment ions from a database with the experimental data (166). This approach has only been possible by the development of computer programs which aid the interpretation of tandem mass spectra (166). The outcome of the search is a ranked list of the highest scoring amino acid identities from the database (161). The specificity of the database search is improved by taking several features of the protein into account, which, amongst others, includes the species of origin, mass of protein, inspection of non-assigned peaks in the peptide map, etc (166).

1.6 Strategies for Obtaining the cDNA of Transcription Factors

1.6.1 cDNA Cloning by Screening Recombinant DNA Libraries

DNA replication and the transcriptional regulation of gene expression is usually mediated by the interaction of DNA regulatory regions and DNA-binding proteins (ie by a combination of *cis*- and *trans*-acting elements). The majority of *cis*-acting elements are localised to the 5' flanking and intronic regions of genes and they often constitute highly conserved sequence motifs which may be common to many genes (some examples include the Z-box, the X-box, the Y-box, the octamer and the "TATA"box, which are all highly conserved sequence motifs among class II genes (167)). Investigating the functional role of the interaction of *trans*-acting factors with their cognate *cis*-elements relies largely on the ability of *trans*-acting elements within crude nuclear extracts to bind DNA regulatory sequences. Although these proteins can be retrieved in a pure form (see section 1.5) it is often difficult and tedious to isolate them in sufficient quantities which allow further biochemical studies. Therefore isolation of the gene or cDNA encoding the DNA-binding activity can facilitate the analysis of the structural and functional relationships between *cis*- and *trans*-acting elements, merely by overproducing regulatory proteins using recombinant DNA clones of their encoding DNA (168). The molecular aspects of gene regulation and functional significance of these motifs can be explored further using transfection and transgenic studies. However these investigations rely on the use of recombinant DNA clones encoding the regulatory proteins. Thus in order to understand the molecular mechanisms underlying transcriptional regulation, it is important to isolate and analyse cDNA clones encoding DNA-binding proteins.

The general approach for obtaining the sequence of a gene or mRNA is by screening a DNA library. The latter should be a faithful representation of the complexity, size and sequence of the genomic DNA or mRNA population it was derived from. This can however be a problem, especially if the DNA of interest represents a small fraction of the total target DNA, for instance when the gene is present as a single copy in the entire genome, or a rare cDNA needs to be isolated from a complex mRNA population (169). Since there is no technology available for the manipulation and propagation of RNA sequences the best alternative is to synthesise cDNA and prepare recombinant libraries. A variety of factors must be given careful consideration when screening a library. Firstly, the strategy for constructing a cDNA library is important. Factors which affect the quality, size and type of library include the integrity of the mRNA (it

must be undegraded and DNA-free), the abundance of the clone of interest (high cloning efficiency is required if the sequence of interest represents a small fraction of the total target DNA), and the screening method (this determines which type of vector the library is cloned into). Optimally a cDNA library is made from a tissue or set of cells which express the highest mRNA levels to ensure that there is at least one cDNA clone for each mRNA present in the cell. Usually one aims to achieve a cDNA library which contains at least five times more recombinants than the lowest abundance of each mRNA. This ensures that even the rarest mRNA is represented, and it takes into account the screening efficiency (eg expression screening reduces the number of identifiable clones by 1/6th) (151).

Screening methods vary from simply sequencing several individual clones until the clone of interest is identified, through ordinary hybridisation methods, to more complex methods of expression screening. The method devised to isolate a clone of interest depends on each individual case, and is determined, amongst other factors, by the type of recombinant library and the frequency of the desired clone within that library (151). The usefulness of a recombinant library depends on the ability to screen a large number of clones at one time and to identify the correct one within the population. This can be achieved by spreading a library (or large numbers of colonies) on agar plates and preparing replica copies of the plates by transferring them onto filters. The latter can then be screened by a variety of methods, such as *in situ* hybridisation screening or by binding to antibodies, as well as using DNA ligands.

1.6.1.2 Hybridisation Screening and Immunoscreening

A pure protein (see section 1.5) can be used to generate a specific antiserum, or it can be analysed for its partial amino acid sequence, from which oligonucleotide primers can be designed. Both the antiserum and the primers in turn can be used to screen a cDNA library by expression or hybridisation respectively. This should generate the desired recombinant clone, either because it expresses a segment of protein which can be recognised by an antibody, or because it hybridises to a specific nucleic acid probe. Another way of identifying cDNAs coding for sequence-specific DNA-binding proteins is to use the filter binding protocol (see section 1.6.1.2) developed by Singh et al (1988) (170).

Current protein microsequencing technology has made it convenient to isolate cDNA clones using degenerate oligonucleotide probes without raising antibodies against the purified protein. Taking codon

utilisation frequencies into consideration, the protein sequence is reverse-translated and a best-match DNA sequence is generated. An oligonucleotide is synthesised from the above information, which can be used to screen a cDNA library. The screening is normally performed on bacterial colonies containing plasmids or cosmids (151), or on bacteriophage plaques, using the in situ plaque hybridisation technique described by Benton and Davis (1977) (171). The phage are multiplied in host bacteria at high density within a thin layer of agarose which is spread onto agar plates. The phage particles and unpackaged DNA adsorb onto the filter which is applied to the agarose. Duplicate filter membranes are generated this way, they are treated with sodium hydroxide to destroy the phage particles and the denatured DNA remains fixed to the filter. The filter is prehybridised in a solution which saturates the filter's nonspecific DNA binding capacity, after which the desired clones can be detected by their ability to hybridise to a labelled DNA probe (the denatured DNAs reanneal with complementary strands approximately 25°C below the melting temperature Tm). It is critical that the DNA probe is unique (ie it may not contain any reiterated or vector sequences) to ensure that it correctly matches the sequence of interest. The nucleic acid probe may take several forms, eg a plasmid containing a previously cloned fragment, synthetic oligodeoxynucleotides, RNA or single stranded cDNA made from reverse transcribing mRNA (172). Excess probe, incorrectly matched sequences and non-sequence specific interactions are washed off the filter to reduce the signal to noise ratio in the ensuing autoradiography, which should identify authentic signals on duplicate filters (151). Autoradiography of the replica filters eliminates most artifacts, and often allows the identification of a single clone within a population of millions of other clones. The positive signals are matched with the corresponding regions on the agar plates and a plug of agar is picked from the correct area of the plate. The isolated positives are usually rescreened at a lower density, as the confluence of the agar surface in the first screening round does not allow the identification and isolation of an individual plaques or colonies (173).

Using antibodies in the screening procedure is another method whereby clones encoding specific proteins can be identified in a recombinant library. This is analogous to hybridisation screening with radioactive DNA probes, except that the plaques are screened with antibodies specific to the desired proteins. The cDNA library is cloned in an expression vector (eg λGT11 or λZAP), and the insert DNA is detected indirectly by the protein, which is generally induced to express from the cloned segment as part of a fusion protein. Expression of the particular recombinant is only initiated after the host's growth is firmly established, such that toxic proteins do not affect the growth (172). The immunological screening of fusion proteins produced either by plasmids or phage involves firstly the synthesis and immobilisation of

the antigenic material to the filter, and secondly the detection procedure using antibodies. Antibodies can either be visualised by labelling with ^{125}I directly, or by applying secondary antibody conjugates (172). The quality of the antibody probe is important, and high titre antibodies produce better signals than low titre antibodies. Both monovalent and polyvalent antibody populations have been used to successfully isolate clones of interest (172). Polyvalent antibodies usually recognise more than a single epitope, which is advantageous. However, they also contain components which bind to antigens normally produced by *E. coli*. These components can be removed from the antibody preparation by incubation with bound bacterial lysate and subsequent elution of the antibody. It is important to note that demonstrating that a cDNA encodes an antigenic determinant does not prove that it codes for the protein of interest, since the cDNA may code for a sequence which is related only at the protein product structure. Therefore the desired sequence should be tested with at least one additional distinguishing property in order to identify the clone unequivocally (151).

1.6.1.2 DNA Ligand Screening

Both screening strategies described above (see section 1.6.1.1) depend on the availability of substantial amounts of starting material for purifying the protein (see section 1.5). A novel but similar screening strategy was developed by Singh et al (1988) (170) which is analogous to immunological screening, however it obviates purification of a sequence specific protein for the purpose of isolating its cDNA or gene since it relies on a selectable function in order to identify the desired clones. The DNA ligand screening technique is designed to directly detect clones which encode sequence specific DNA-binding proteins, by simply using a cDNA expression library and a labelled DNA recognition site probe. The feasibility of the screening strategy was tested using a model system where a λGT11 phage recombinant (λEB), encoding a fusion protein of the DNA-binding domain of the Epstein-Barr virus nuclear antigen (EBNA-1), was detected with a DNA recognition site probe (170). The ease of isolation of this clone implied that this strategy could facilitate the isolation of genes for transcription factors and would be generally useful in the cloning and analysis of sequence-specific DNA-binding proteins (174). Since the initial application of the DNA ligand screening technique many mammalian cDNA clones encoding a variety of sequence-specific DNA binding proteins (with varying DNA-binding motifs) have been isolated; eg H2TF1/NF-κB, Oct-2, E12, Xbp, IRF-1, MLTF, and CREB, confirming that this method is indeed useful and not restricted to a particular subclass of DNA-binding domains. (Singh et al (1989)

(174) list many of the clones encoding transcriptional regulatory proteins which have been isolated from λGT11 cDNA libraries using recognition site DNA probes.) Other examples include the cDNAs for YB3 protein from Xenopus (175) and SpP3A1 from *Strongylocentrotus purpuratus* (122), and the cDNAs for YB-1 and dbpA were isolated independently by several groups (176, 177). It has been reported that even clones encoding proteins which bind DNA as homodimers can be detected via DNA ligand expression screening (174). Expression vectors such as λGT11 or λZAP allow insertion of foreign DNA into the β–gal structural gene, *lacZ*, and promote the expression of hybrid proteins. The hybrid proteins are then blotted onto duplicate nitrocellulose filters, and the screening procedure essentially involves the direct probing of the replica filters from the cDNA expression library with radiolabelled, sequence-specific binding site DNA. Several variables influence the successful identification of bacterially expressed DNA-binding clones using radioactive DNA as ligand. The requirements or conditions that need to be fulfilled include the functional expression of the DNA-binding domain in *E.coli* as a β-galactosidase fusion protein, a strong interaction between the DNA-binding domain and recognition site DNA (170), and the protein of interest must be able to bind the DNA independent of associations with other proteins or complexes. Several procedures greatly enhance the detection of cDNAs which encode sequence-specific DNA-binding proteins. For instance, it was demonstrated that the sensitivity of detection was greatly improved by the use of a multisite DNA probe, as well as the use of nonspecific competitor DNA such as sonicated, denatured calf thymus DNA (178, 179). The inclusion of excess nonspecific competitor DNA reduces the background, as well as the detection of nonspecific DNA-binding proteins. Using a test system where the DNA-binding domain of an enhancer binding protein (C/EBP) was fused in frame to the β-galactosidase gene of bacteriophage λGT11, Vinson et al (1988) (168) determined that processing the nitrocellulose filters through a denaturation / renaturation regimen using 6 M guanidine hydrochloride increased the level of binding between the ligand and the recombinant fusion protein and enhanced the selectivity of the interaction. Bacterial expression of recombinant proteins often leads to deposition of insoluble protein precipitates which need to be exposed to chaotropic agents to be solubilised (168). Guanidine hydrochloride dissociates these protein aggregates, and it may also overcome the problem of incorrect protein folding in *E.coli*. Therefore the denaturation / renaturation protocol leads to improved detection signals by allowing more of the correctly folded protein molecules to access the DNA ligand. In this report the authors also tested the importance of using a catenated DNA probe as ligand, showing that the catenated probe (consisting of a higher density of protein binding sites) yielded an appreciably enhanced detection signal over the monomeric ligand. It is likely that a catenated probe is able to tether several immobilised, bacterially expressed DNA-binding proteins on a single DNA molecule, which may

alleviate rapid probe dissociation and increase the stability of the protein-DNA complexes (174). A comparison of the signals obtained when using a ^{32}P-labelled DNA-binding site probe and a secondary antibody conjugate containing horseradish peroxidase shows that the DNA binding site probe has higher sensitivity, but the signal is probably comparative to that obtained when using an ^{125}I-labelled antibody (174).

1.6.2 cDNA Cloning by PCR

The polymerase chain reaction (PCR) has promoted the development of a variety of analytical procedures which can be used for detection, measurement and characterisation as a result of its power to amplify extremely small amounts of DNA. Its versatility ranges from basic PCR amplification reactions, to mutagenesis and screening of libraries (180). Prior to the development of PCR library screening techniques mainly involved plating lambda based libraries on a lawn of *E.coli*, with subsequent transfer of the phage particles or their expressed recombinant proteins to nitrocellulose filters, which were then screened with DNA hybridisation probes, antibodies or DNA ligands (see section 1.6.1). This technique is relatively time-consuming, it often results in numerous false positives and sometimes it may not be sensitive enough to detect low abundance mRNAs. PCR, in combination with degenerate or specific primers can be used to isolate families of related clones from recombinant DNA libraries, within a short time and without the use of radioactive labels. Generally a cDNA or genomic library can be plated (at relatively low density) and the phage from each plate represent the starting material for PCR screening. Positive aliquots are identified and are replated at lower and lower densities with subsequent PCR screening until a single phage plaque can be identified as a positive. PCR-based cDNA screening techniques are applicable to both bacterial and phage libraries (180).

cDNA amplification using 5' and 3' RACE (Rapid Amplification of cDNA Ends) is a technique which allows cloning of full-length cDNAs without the need to construct or screen a cDNA library. It relies on the simultaneous PCR amplification of both 5' and 3' cDNA ends performed on the same template (181), and is capable of amplifying large templates which have high fidelity with respect to the original RNA. The only requirement for successful application of 5' / 3' RACE is that a single short DNA sequence (at least 21 - 24 nt) must be known for the cDNA of interest, such that gene specific primers for both the 5' and 3' RACE reactions can be synthesised. In addition, the amplifications require a source of RNA in

which the transcript of interest is definitely present. Briefly, the technique involves synthesis of double - stranded cDNA from either total or poly-A^+ RNA using a combination of reverse transcriptase, *E.coli* polymerase I and RNase H. T4 DNA polymerase is used to create blunt-ended ds cDNA, which is ligated to adaptors with specific adaptor primer (AP) sites. The uncloned library of adaptor ligated ds cDNA can be used in combination with primers specific to the adaptor sequence and gene specific primers (a pair of antisense and sense primers) to amplify the respective 5' and 3' RACE fragments from the cDNA template of interest. Exponential amplification is achieved by using a DNA polymerase which is suitable for long distance PCR, and the two separate fragments generated in this way can be combined to form a full length cDNA. Full length cDNAs can be achieved using various techniques, such as simple ligation and cloning of the two products, PCR amplification of the full length product or by a fusion reaction (in this case the 5' and 3' fragments are annealed and simultaneously act as template and primers in a thermal cycling reaction). The full length product can then be cloned and analysed to verify its nucleotide sequence.

1.7 Aim of This Investigation

In order to elucidate the biochemical mechanisms which govern gene transcription, factors regulating gene promoters must be purified and characterised, since they are central to the understanding of regulation and control of gene expression, replication and recombination. Many examples of G·C-rich promoter elements have been cited. These elements, together with the factors that bind them, are implicated in gene regulation. A detailed understanding of the structural and functional relationships amongst factors which bind G·C-rich sequences may elucidate analogies in their biological roles and DNA recognition functions.

suGF1 is implicated as a member of a family of G-string factors with related functions. The main objective of this investigation is to purify and clone suGF1. This could possibly elucidate a functional significance for suGF1 by identifying the factor via its primary protein structure and cDNA sequence. Analysis of the molecular structure of the protein may lead to the identification of characteristic motifs similar to those of previously identified transcription factors. This may enable the classification of suGF1 as a possible member of a family of proteins which are related by their primary structure and DNA-binding specificity. The strategies for purification and cloning of suGF1 are developed concurrently in

this project, using the DNA recognition site as affinity probe in both approaches. Detection of the cDNA for suGF1 is attempted by the direct cloning method developed specifically for DNA-binding proteins by Singh et al (1988) (170), with the aim to eliminate the need for isolating large amounts of native protein. A PCR strategy is employed as second cloning approach, in order to isolate a *P.angulosus* clone based on a *S.purpuratus* homologue, which potentially represents a suGF1 clone. Identification of positive clones can be achieved by analysis of the DNA-binding specificities and affinities by expression of the putative positive clones. No information is available on the requirements of post-translational modifications or heterodimer formations regarding the suGF1-DNA interaction, therefore, despite the apparent ease of the direct cloning method by the DNA-ligand screening approach, it is not guaranteed to be successful with respect to every transcription factor. Therefore the cloning strategy and the protein purification of suGF1 are addressed concurrently. The latter involves a combination of ion exchange and affinity chromatography, and SDS-PAGE as a final purification step before primary structure determination of the protein. Using both these approaches it should be possible to identify the protein sequence and cDNA sequence for suGF1.

By analysing the primary structure of suGF1 several characteristic domains related to the functional properties of the factor may be elucidated. For instance the putative DNA-binding domain (eg zinc finger or homeodomain, amongst others), a possible dimerisation domain (which is often associated with leucine zipper proteins and helix-loop-helix functional motifs), and perhaps a transcriptional activation domain (this is usually characterised by regions which are serine / threonine-, glutamine / threonine- or proline-rich) are speculated to be present in the factor. The identification and isolation of the cDNA coding for suGF1 should lead to recognition of domains within the protein via similarities in their primary structures, allowing comparisons to be drawn with other G-binding factors. In addition, a knowledge of the cDNA coding for the suGF1 protein should allow verification of the developmental distribution of the factor, and expression of the cDNA using truncated DNA templates may enable further identification of the position of certain functional domains (eg the DNA-binding domain and a dimerisation domain) with respect to the primary protein structure.

CHAPTER 2

Materials and Methods

2.1 Materials

All chemical reagents and solvents used were analytical grade, unless otherwise stated. The source of the materials was not important unless specified. All solutions, glassware and plastics were sterilised by autoclaving or sterile filtering. All water was double distilled.

2.2 Plasmid Propagation and Isolation

2.2.1 Competent Cells

Competent bacterial cells were either purchased from Pharmacia or they were prepared using a method described by Chung et al (1989) (182). The cells were grown to early log phase (OD_{600} = 0.3 - 0.6) in Luria Bertani (LB) broth. They were pelleted by centrifugation (1000 X g for 10 minutes at 4°C). The cell pellet was resuspended in 1/10 volume of transformation and storage buffer (TSB) (LB broth (pH 6.1) containing 10 % (w/v) PEG (MW = 4 000), 5 % (v/v) DMSO and 20 mM Mg^{2+} (10 mM $MgCl_2$ and 10 mM $MgSO_2$)) at 4°C. They were then incubated on ice for approximately 10 minutes, and either stored at -70°C in aliquots, or used immediately for the transformation procedure.

2.2.2 Transformation of Competent Cells

Competent cells (purchased commercially) were transformed according to the supplier's recommendations. JM109 High Efficiency Cells (Promega) were used to transform ligation reactions. Competent cells were mixed evenly and a 50 µl aliquot of cells was combined with 2 µl of ligation reaction and incubated on ice for 20 minutes. The cells were heat shocked for 45 - 50 seconds in a waterbath at 42°C, and placed on ice for 2 minutes. The transformed cells were supplemented with 950 ml SOC medium (room temperature), and incubated for 1.5 hours at 37°C with shaking. An

aliquot of each transformation culture (100 µl) was plated onto LB plates containing ampicillin (50 µg/µl), IPTG (0.5 mM) and X-Gal (80 µg/ml). The plates were incubated at 37°C overnight.

Competent cells (100 µl) prepared by the method of Chung et al (1989) (182) (see section 2.2.1) were mixed with plasmid DNA and incubated on ice for 60 minutes. The cells were supplemented with 900 µl TSB containing 20 mM glucose, and incubated at 37°C for 60 minutes to express the antibiotic resistance gene. Transformants were selected by plating the cells on LB plates containing the relevant antibiotic.

2.2.3 Plasmid DNA Mini-Preparation by Boiling Method

Single bacterial colonies containing plasmid DNA were picked from fresh agar plates and inoculated into 10 ml LB containing the appropriate antibiotic. Cells were grown overnight at 37°C, and 1.5 ml of the culture was pelleted (12 000 rpm, 1 minute). The cell pellet was resuspended in 290 µl STET (8 % (w/v) sucrose. 0.5 % Triton X-100, 50 mM EDTA (pH 8), 10 mM Tris.HCl (pH 8)) and 10 µl lysozyme (20 mg/ml) was added to the cell suspension, which was incubated on ice for 5 minutes. The cells were lysed by boiling for 1 minute, and the cell debris was precipitated by centrifugation for 20 minutes at room temperature (12 000 rpm). The sticky pellet was removed with a toothpick and the DNA was precipitated by addition of an equal volume of isopropanol. The DNA was recovered by centrifugation (12 000 rpm, 20 minutes, 4°C), and the pellet was washed with 70 % ethanol. The dry DNA pellet was resuspended in 20 µl TE and stored at -20°C.

2.2.4 Large Scale Plasmid Isolation

Plasmids were propagated in the appropriate *E.coli* strains which were grown in Luria-Bertani growth medium (LB) containing the relevant antibiotic. Plasmids were either isolated by the triton lysis method (151), or using Wizard Midipreps DNA Purification System (Promega) according to the supplier's recommendations.

2.2.4.1 Triton Lysis Method

Briefly, 2 litres of bacterial culture were centrifuged for 30 minutes at 5 000 rpm (JA 14 rotor, Beckman). The pellet was resuspended in 15 ml 50 mM Tris.HCl (pH 7.5), and the suspension was

incubated for 30 minutes with 1 ml lysozyme (10 mg/ml), 10 minutes with 1 ml EDTA (0.5 M), 20 minutes with 100 µl RNase A (20 mg/ml), 200 µl Triton (10 % (v/v)) and centrifuged for 45 minutes at 20 000 rpm (JA20 rotor). The supernatant was extracted three times with an equal volume of neutralised phenol, and twice with an equal volume of chloroform. The sample was adjusted to 300 mM sodium acetate and the DNA was precipitated with 2.5 volumes absolute ethanol for 30 minutes at -70°C. DNA was recovered by centrifugation at 20 000 rpm (JA20 rotor) for 20 minutes.

The pellet was washed with 70 % (v/v) ethanol, dried and the plasmid was banded in a caesium chloride / ethidium bromide gradient (VTi 65 rotor (Beckman), 55 000 rpm, 16 hours, caesium chloride from Sigma). This step was performed twice if supercoiled plasmid only was required. Ethidium bromide was removed from the recovered plasmid by repeated extractions with isoamyl alcohol (151). The supercoiled plasmid was dialysed against TE (pH 7.5) and precipitated as above. Plasmid was stored in aliquots in TE (pH 7.5) at -20°C.

2.2.4.2 DNA Isolation Using Wizard Midipreps Columns (Promega)

Bacterial cultures (100 ml) were pelleted for 15 minutes at 4°C (14 000 rpm, JA14 rotor (Beckman)). The cells were resuspended in 3 ml cell resuspension solution (10 mM Tris.HCl (pH 8), 10 mM EDTA (pH 8) and 100 µg/ml RNase A), and lysis was achieved by addition of 3 ml of cell lysis solution (0.2 N NaOH, 1 % (w/v) SDS). Plasmid was released from the lysed cells by gentle swirling, and the mixture was supplemented with neutralisation solution (1.32 M KOAc (pH 4.8)). Cell debris and chromosomal material were precipitated for 15 minutes at 4°C (15 000 rpm, JA20.1 rotor (Beckman)). The supernatant was retained and supplemented with 10 ml Wizard Midipreps DNA Purification Resin (7 M Guanidine HCl), and passed directly over a Wizard Midipreps column. Elution of solvents was achieved by application of a vacuum to the column, which was subsequently washed twice with 15 ml column wash solution (8.3 mM Tris.HCl (pH 7.5), 83 mM NaCl, 2 mM EDTA and 58 % (v/v) ethanol) and dried before the DNA was eluted with 300 µl TE (pH 8.0) at 65°C by spinning the column for 2 minutes.

2.2.5 Recovery of Single Stranded DNA from pBluescript

Single stranded DNA was isolated according to the Stratagene instruction manual for pBluescript Exo / Mung DNA Sequencing system. pBluescript vector containing the DNA fragment of interest was transformed into XL1-*Blue* cells, which were grown on LB / ampicillin / tetracycline plates. Single

colonies were picked and 10 ml starter cultures were grown overnight at 37°C in LB / ampicillin / tetracycline. LB (2 ml) containing 2 µl purified M13 phage was inoculated with 20 µl of the starter culture, and the mixture was incubated at 37°C for one hour. The culture was supplemented with kanamycin (230 µg/ml final concentration) and the culture was grown at 37°C overnight. The cells were pelleted at 12 000 rpm for 5 minutes (4°C), and the phage-containing supernatant was mixed with 0.2 volume 15 % (w/v) PEG (8000) / 2.5 M NaCl. The mixture was vortexed gently and incubated at room temperature for 15 minutes. The phage were pelleted at 12 000 rpm for 5 minutes (4°C) and the supernatant was removed carefully. The pellet was resuspended in 100 µl TE (pH 8) by vortexing vigorously. The phage were extracted once with 50 µl phenol, followed by a chloroform extraction (50 µl). The DNA was precipitated with 0.1 volume 3 M sodium acetate (pH 5.2) and 2.5 volumes absolute ethanol. The mixture was incubated at room temperature for 15 minutes and the DNA was pelleted at 12 000 rpm for 10 minutes (4°C). The DNA was washed with 70 % ethanol, evaporated to dryness and resuspended in 20 µl TE (pH 8).

2.3 Sanger Di-Deoxy DNA Sequencing

Enzymatic sequence analysis was performed using the TaqTrac sequencing system (Promega), alternatively the DNA samples were sequenced using an Automated DNA Sequencer.

2.3.1 Denaturation of Double Stranded DNA

DNA (approximately 5 µg) was pipetted into a microfuge tube, the volume was adjusted to 18 µl with sterile water and 2 µl of a 2 M NaOH solution was added. The mixture was incubated at 37°C for 6 minutes. The reaction was neutralised with 4 µl of 3 M sodium acetate (pH 5.2), and 150 µl of absolute ethanol was added. The DNA was pelleted by centrifugation at 12 000 rpm for 30 minutes (4°C), and the pellet was washed with 70 % ethanol. The DNA was dried by evaporation and resuspended in 17 µl water.

2.3.2 Sequencing Using the Two Step Extension / Labelling Procedure

Primer (2 pmol) and single stranded or denatured doublestranded DNA (5 µg) were combined in 1 X Taq DNA polymerase buffer (50 mM Tris.HCl (pH 9) and 10 mM $MgCl_2$) containing 2 µl extension / label mix (7.5 µM of each dGTP, dTTP and dCTP), and annealed by incubation at 37°C for 11

minutes. The extension / labelling reaction was carried out at 37°C for 11 minutes by adding 1 µl [α-^{35}S]dATP (approximately 10 µCi/µl, Amersham) and 1 µl Taq DNA polymerase (2.5 U/µl) to the DNA / primer mixture. For each set of sequencing reactions 1 µl of each ddNTP was aliquoted into a microcentrifuge tube and 6 µl of the extension / label reaction was added. The mixture was incubated at 70°C for 16 minutes and 4 µl of stop solution (10 mM NaOH, 95 % (v/v) formamide, 0.05 % (w/v) bromophenol blue and 0.05 % (w/v) xylene cyanol) was added to each tube. The reactions were heated to 90°C for 5 minutes before resolving them on a 6 % sequencing gel (see section 2.3.3).

2.3.3 Sequencing gels

6 % sequencing gels (5.7 g acrylamide (Merck), 0.3 g bisacrylamide (BDH Biorad), 48 g urea (Merck), 10 ml of 10 X TBE, 40 ml water, 45 µl of 50 % AMPS and 45 µl TEMED) were pre-electrophoresed for 30 minutes at 90 W. Samples were electrophoresed for 1 - 6 hours at 90 W depending on where the DNA sequence of interest was situated on the template DNA.

2.4 Synthesis and Annealing of Oligodeoxyribonucleotides

Oligodeoxyribonucleotides (oligonucleotides) were synthesised on a Beckman Systems 1+ DNA Synthesizer and purified by established procedures (151). Concentrations were determined spectrophotometrically. The molar extinction coefficient for each oligonucleotide was estimated from the extinction coefficients of the individual bases (151). Complementary strands of the specific and nonspecific oligonucleotides (see fig 2.1 for their respective nucleotide sequences) were annealed at a molar ratio of 1:1, by incubating at 88°C for 2 min, 65°C for 10 min, 37°C for 10 min, 25°C for 5 min, and finally placing the sample on ice.

2.5 Enzymatic Manipulations and Radioactive Labelling of DNA

2.5.1 Restriction Enzyme Digests

Typically restriction enzyme digests were performed with plasmid DNA containing 1 X reaction buffer, 1 - 2 units / µg of DNA of each restriction enzyme, and the final volume was adjusted with water. The mixture was incubated at 37°C for one hour or longer, followed by addition of 6 X loading buffer (0.25 % bromophenol blue, 0.25 % xylene cyanol, 30 % glycerol in water). The digests were

SPECIFIC OLIGO (Sp)

```
5' gatcAGAGAGGGGGGGGGGAGGGAGAATT     3'
3'     TCTCTCCCCCCCCCCCTCCCTCTTAActag 5'
```

NONSPECIFIC OLIGO (NS)

```
5'     TCAGGTCATGGCCACTGTGACGTCTTctag 3'
3' gatcAGTCCAGTACCGGTGAGAGTGCAGAA     5'
```

Fig 2.1 Sequences of the Synthetic Oligodeoxyribonucleotides

The sequences of the specific (Sp) and nonspecific (NS) synthetic double stranded 30 bp oligonucleotides are shown in capital letters. The oligos have 4 base single-stranded overhangs (small letters) to enable multimerisation. The specific oligo contains 26 bp of the H1-H4 intergenic region of the early histone gene battery of *P.miliaris* (h22) (see fig 2.2), whereas the nonspecific oligo contains a random sequence.

```
  gaattctc atgtttgaca gcttatcatc gccctgactg agtcgagccc
  cttaagag tacaaactgt cgaatagtag cgggactgac tcagctcggg
   Eco RI
                       -440       -430       -420
aattcgagct cggtacccCA CGTAGAGGAA AAGAGAGTTA TACCACTCCT
ttaagctcga cggatgggGT GCATCTCCTT TTCTCTCAAT ATGGTGAGGA

-410       -400       -390       -380       -370
GACATGAAAC ACACTCAATT CAACATATTT AGAGGAAGGG AGAGAGAGAG
CTGTACTTTG TGTGAGTTAA GTTGTATAAA TCTCCTTCCC TCTCTCTCTC

-360       -350       -340       -330       -320
AGAGAGAGAG AGAGAGAGAG AGGGGGGGGG GGAGGGAGAA TTGCCCAAAA
TCTCTCTCTC TCTCTCTCTC TCCCCCCCCC CCTCCCTCTT AACGGGTTTT

-310       -300       -290       -280       -270
CACTGTAAAT GTAGCGTTAA TGAACTTTTC ATCTCATCGA CTGCGCGTGT
GTGACATTTA CATCGCAATT ACTTGAAAAG TAGAGTAGCT GACGCGCACA

-260       -250
ATAAGGATGA TTATAAGCTg gggatcctgt agagtcgacc tgcaggcatg
TATTCCTACT AATATTCGAc ccctaggaga tctcagctgg acgtccgtac

caagctgggc tcgacttagt cagggtcacc gataagctt   Watson
gttcgacccg agctgaatca gtcccagtgg ctattcgaa   Crick
                                  Hind III
```

Fig 2.2 DNA Sequence of the E/H Fragment

Part of the sequence of plasmid pHP2 (shown in small letters) contains a 201 bp insert (capital letters) from the H1-H4 intergenic region of the *P.miliaris* early histone gene battery (h22) (20). Numbering is with respect to the major cap site of the mRNA of H4 denoted +1 (183). A 335 bp *EcoRI / HindIII* fragment (E/H fragment) was prepared from pHP2 and radiolabelled on one strand as described in the text.

analysed on agarose gels of appropriate concentration. When plasmids from minipreps were digested, 1 µl RNase A (20 mg/ml) was also added to the restriction digest reaction.

2.5.2 Isolation and Radioactive Labelling of DNA Fragments

A 335 bp *Eco R1 / Hind III* (E/H) fragment containing the binding site of suGF1 was prepared from pHP2 (20). (This fragment includes 11 contiguous G residues on one strand, see fig 2.2.) Fragments obtained by restriction enzyme digestion were resolved on 1% agarose gels in TAE (0.04 M Tris-acetate, 0.002 M EDTA) containing ethidium bromide. The relevant bands were visualised with a handheld UV lamp (λ = 315 nm) and excised from the gel. The DNA was purified from the agarose using the Wizard PCR Preps DNA Purification System (Promega). Agarose slices were transferred to microcentrifuge tubes and 1 ml Wizard PCR Preps DNA Purification Resin was added. The agarose was melted at 65°C and applied to a Wizard Minicolumn, which was washed with 2 ml of 80 % (v/v) isopropanol. The DNA was eluted from the dry column by applying 50 µl TE (pH 8) and centrifugation at 12 000 rpm for 20 seconds. The sample was adjusted to 0.3 M sodium acetate and the DNA was precipitated by adding 2.5 volumes absolute ethanol for 30 minutes at -70°C. The DNA was recovered by centrifugation at 15 000 rpm (Beckman JA20.1) for 20 minutes, washed with 70 % (v/v) ethanol and stored in TE (pH 8) at -20°C. The DNA concentration was determined spectrophotometrically or by ethidium bromide spotting.

Restriction fragments were 3' end-labelled by a Klenow fill-in reaction. The Watson strand of the E/H fragment could be labelled selectively by filling in the *HindIII* site using [α-^{32}P]dCTP (10 µCi/µl, Amersham) as radioactive nucleotide. Specific activity of fragments was typically 12 000 to 50 000 dpm/ng.

2.5.3 Nick Translation of cDNA

cDNA was labelled by nick translation with [α-^{32}P]dCTP according to the protocol described by Maniatis et al (1982) (169). Typically an incubation was performed in 1 X Nick Translation Buffer (50 mM Tris.HCl (pH 7.5), 10 mM $MgSO_4$, 1 mM DTT, 0.05 mg/ml BSA) containing 1 µl cDNA (0.5 mg/ml), 3 µl nucleotide mix (2 mM each dATP, dGTP, dTTP), 3 µl [α-^{32}P]dCTP (10 µCi/µl), 1 µl DNaseI (0.4 µU/ml) and 1 µl DNA polymerase I (500 U/ml). The reaction was allowed to proceed

at 16°C for one hour, it was terminated by the addition of 2 µl of 0.5 M EDTA (pH 8). The specific activity of the labelled probes was 0.1 - 1.9 x 10^8 dpm/µg DNA.

2.5.4 Labelling of DNA Using the Amersham MegaPrime Kit

DNA fragments were isolated from low melting point agarose gels as described above (see section 2.5.2). The DNA template (25 ng) was combined with 5 µl random primers (Amersham) and the mixture was denatured at 95°C - 100°C for 5 minutes. The reaction cocktail, including nucleotides (4 µl of each dGTP, dTTP and dATP (10 mM)), 10 X reaction buffer (5µl), [α-^{32}P]dCTP (5 µl) and 2µl Klenow enzyme (2 U/µl) was added to the denatured DNA at room temperature. The volume was adjusted to 50 µl with water. The reaction was incubated at 37°C for 30 minutes and was terminated by addition of 0.5 M EDTA (2 µl). The volume was adjusted to 100 µl with TE and the labelled DNA was separated from the unincorporated nucleotides by chromatography on a Sephadex G-50 spin column. Specific activity of the labelled DNA was typically ~ 1 X 10^9 dpm/µg.

2.5.5 Sephadex G-50 Chromatography

The labelled DNA was passed over a 1 ml spin column containing Sephadex G-50 in TE solution (pH 8) in order to remove the unincorporated nucleotides. The spin column was prepared by plugging a disposable 1 ml syringe with sterile glass wool and filling it with Sephadex G-50. The column was washed with three column volumes of TE (pH 8) and centrifuged on a benchtop centrifuge for 4 minutes after each wash. The DNA sample was applied to the column in a volume of 100 µl and eluted by centrifugation for 4 minutes. The labelled DNA was present in the void volume and the amount of radioactivity was measured by Cerenkov counting.

2.6 RNA Isolation and Manipulations

2.6.1 RNase-free Plasticware, Glassware and Solutions

All glassware was baked at 260°C for four hours. Plasticware was soaked in a solution of 3 % (v/v) H_2O_2 for ten minutes and washed thoroughly with RNase-free water, unless sterile disposable plasticware was available. RNase-free water was double distilled Milli-Q water (passed through carbon, ion exchange and organic scavenger cartridges), filtered directly into a baked glass bottle and

autoclaved for 30 minutes. All RNase-free solutions were prepared using baked glassware, chemicals set aside for RNA work only and RNase-free water. Gloves were worn at all times when working with RNA and the accompanying solutions.

2.6.2 RNA Isolation Procedure

Total RNA was isolated using the Guanidinium thiocyanate / CsCl gradient technique proposed by Chirgwin et al (1979) (184). Guanidinium thiocyanate stock solution (4 M guanidinium thiocyanate, 0.5 % (w/v) sodium lauroylsarcosine, 25 mM sodium citrate (pH 7), 100 mM ß-Mercaptoethanol) was filtered through a 0.22 µm filter. Sea urchin eggs or embryos (grown for 4 hours, 9 hours, 14 hours, 21 hours, 30 hours or 45 hours) were allowed to settle at 4°C. Adult tissue (testes, ovaries or muscle) was frozen in liquid nitrogen and ground to a powder using a mortar and pestle. Guanidinium thiocyanate stock solution (3.3 ml) was added to 1 ml of settled eggs, embryos or adult tissue. The cells were resuspended in the guanidinium thiocyanate solution using a Pasteur pipette. The lysed cell mix was transferred to a 5 ml handheld Dounce homogenizer and the suspension was homogenized for 25 - 30 strokes. Each homogenate was added to 1.32 g CsCl and shaken to dissolve the CsCl. The mixture was layered over a 1.2 ml cushion of 5.7 M CsCl solution (5.7 M CsCl, 0.1 M EDTA (pH 7.5)) in a polyallomer tube. The RNA was pelleted by centrifugation at 36 000 rpm (105 000g) in a Beckman SW65Ti rotor for 16 hours at 20°C in a Beckman L-65 Ultracentrifuge. The supernatant was aspirated until approximately 500 µl remained. The centrifuge tube was cut above the level of the remaining solution and the latter was removed by inverting the centrifuge tube, so as not to disturb the RNA pellet. This method ensures the complete separation of RNA (found in the pellet) from contaminating DNA and protein. The pellet was resuspended in 300 µl RNase-free water, transferred to an Eppendorf vial and vortexed to mix. The solution was incubated at 65°C briefly and spun down in a microfuge. The RNA was subsequently precipitated by the addition of 0.1 volume 3 M sodium acetate (pH 5.2) and 2.5 volumes absolute ethanol and stored at -20°C. The RNA was recovered from the ethanol suspension by centrifugation for 20 minutes at 15 000 rpm and 4°C. The pellets were washed with 75 % ethanol, air dried and dissolved in 100 µl RNase-free water. A_{260} in water was used to determine the RNA concentration, and a ratio of $A_{260}/A_{280} = 2$ was ensured. The RNA was stored as an aqueous solution at -20°C.

2.6.3 Selection of Poly-A^+ RNA

Oligo(dT)-cellulose (0.5 g) was swollen in 0.1 M NaOH. A 1 ml column was poured in a silanized pasteur pipette. The column was washed with 10 ml RNase-free water and 1 X column loading buffer (20 mM Tris.HCl (pH 7.6), 0.5 M NaCl, 1mM EDTA, 0.1 % (w/v) SDS), until the pH of the column effluent was less than 8. Sea urchin embryo (14 hour) total RNA (2 mg) was resuspended in RNase-free water and heated to 65°C for 5 minutes. An equal volume of 2 X loading buffer was added, the sample was cooled to room temperature and applied to the column. The flow-through was collected, heated again to 65°C, cooled and reapplied to the column. The eluate was reapplied twice in the same fashion. The column was washed with 10 column volumes of loading buffer. The poly-A^+ RNA was eluted with 4 column volumes of elution buffer (10 mM Tris.HCl (pH 7.5), 1 mM EDTA, 0.05 % (w/v) SDS). Fractions (1 ml) were collected and the A_{260} of each fraction was determined. The poly-A^+ RNA was selected again by oligo(dT)-cellulose chromatography, by adjusting the NaCl concentration of the eluted mRNA to 0.5 M and repeating the chromatography procedure. The poly-A^+ RNA was precipitated at -20°C by adding 0.1 volume 3 M Na-acetate (pH 5.2) and 2.5 volumes absolute ethanol. The pellet was rinsed in 70 % ethanol, and resuspended in RNase-free water.

2.6.4 RNA Gel Electrophoresis

For RNA gels the electrophoresis apparatus was soaked in 3 % (v/v) H_2O_2 for 1 hour, rinsed with methanol and washed with RNase-free water.

2.6.4.1 Denaturation of RNA by Glyoxal Method

RNA was fractionated under denaturing conditions according to standard procedures (169). RNA samples were incubated in 1 M glyoxal, 50 % (v/v) DMSO, 10 mM NaH_2PO_4 (pH 7) for 60 minutes at 50°C. The RNA samples were cooled to 4°C and 0.2 volumes sample application buffer (50 % (v/v) glycerol, 10 mM NaH_2PO_4 (pH 7), 0.25 % (w/v) bromophenol blue) was added to them. The samples were fractionated through a vertical 1 % agarose gel in 10 mM NaH_2PO_4 (pH 7), at 3 - 4 V/cm with buffer recirculation. The gel was stained for 10 minutes with 33 µg/ml acridine orange in 10 mM NaH_2PO_4 (pH 7), and destained overnight in buffer only.

2.6.4.2 Denaturation of RNA by Formamide Method

The integrity of RNA samples was checked on 1 % agarose / formaldehyde gels based on a modification of the method described in Ausubel et al (1987) (151). Gels were prepared as follows: 0.5 g agarose, 5 ml 10 X MOPS (0.4 M morpholinopropanolsulfonic acid (pH7), 100 mM sodium acetate, 10 mM EDTA) and 36 ml DEPC water. The agarose was heated, cooled to 55°C and 8.4 ml 37 % (v/v) formaldehyde was added. The gel was poured in the fumehood and covered with 1 X MOPS buffer once it had set. The RNA samples were precipitated by ethanol and resuspended in 25 µl loading buffer containing 72 µl formamide, 16 µl 10 X MOPS, 26 µl 37 % (v/v) formaldehyde, 18 µl water, 10 µl 80 % (v/v) glycerol and 8 µl saturated Bromophenol Blue. The RNA samples were electrophoresed at 3.5 V/cm for 3 hours in 1 X MOPS running buffer.

2.6.5 Northern Analysis

2.6.5.1 Northern Transfer Procedure

A piece of Nylon membrane (Hybond N) was cut to fit the dimensions of the gel (see section 2.6.4) and floated on deionised water, it was then submerged and wet thoroughly. The membrane was soaked in transfer buffer (10 x SSC) until used. The gel was kept in low ionic strength buffer prior to the transfer. The RNA was transferred from the gel to the membrane by capillary action. A piece of Whatman 3MM paper was cut 10 - 20 cm longer than the gel, saturated with transfer buffer and placed on a glass plate. The ends of the paper wick were draped into a buffer reservoir. The gel was laid onto the wick of chromatography paper, and the prewet nylon membrane was placed on it, followed by 2 - 3 pieces of gel blot paper cut to fit the gel. Any air bubbles trapped between the layers were removed by rolling a pipette back and forth over every layer. The sides of the gel were surrounded by clingwrap film to prevent the paper on top of the gel from coming into contact with the lower layer of gel blot paper. A stack of paper towels was placed on the gel blot paper, and the blot was secured with a light weight. The transfer was carried out overnight, and the transfer efficiency was checked by staining the gel with acridine orange (33 µg/ml). The membrane was washed with 5 x SSC for 5 minutes after the transfer was complete, and the RNA was crosslinked to the membrane under UV light (λ = 254 nm).

2.6.5.2 Northern Hybridisation Procedure

The membrane was placed in a heat sealable bag with 0.25 ml/cm^2 prehybridisation buffer (50 % (v/v) formamide, 5 x Denhardt's reagent, 10 mM NaH_2PO_4 (pH 7), 5 x SSC, 0.1% SDS, 5 mg/ml denatured low molecular weight DNA) for 16 hours at 42°C with gentle agitation. The prehybridisation solution was removed from the blot and replaced with 0.1 ml/cm^2 hybridisation solution (50 % (v/v) formamide, 5 x Denhardt's reagent, 10 mM NaH_2PO_4 (pH 7.0), 5 x SSC, 0.1 % (w/v) SDS, 5 mg/ml denatured low molecular weight *E.coli* DNA) together with the relevant heat denatured, nick translated or MegaPrime labelled probe (5 - 20 ng/ml). Hybridisation was carried out at 42°C for 12 - 24 hours with gentle agitation.

2.6.5.3 Washes and Autoradiography

The blot was removed from the plastic bag and washed as follows: twice in 2 x SSC, 0.1 % SDS for 5 minutes at room temperature, twice in 2 x SSC, 0.1 % SDS for one hour at 65°C, and the final washes were in 2 x SSC for 2 minutes at room temperature. The membrane was blotted slightly, sealed in a plastic bag and autoradiographed.

2.7 Synthesis of cDNA by Reverse Transcription of RNA

RNA-dependent DNA polymerase (reverse transcriptase) was used to transcribe mRNA into cDNA. Either total RNA or poly-A^+ RNA (both 5 µg) was treated with 5 U DNase 1 (10 U/µl Boehringer Mannheim) in 1 X MMLV Reverse Transcriptase Buffer (15 mM $MgCl_2$, 375 mM KCl, 250 mM Tris.HCl (pH 8) (Promega)), containing 0.5 µl RNasin (40 U/µl, Boehringer Mannheim) for 30 minutes at 37°C. The RNA was extracted once with phenol (pH 4) and once with chloroform. It was adjusted to 0.3 M Na-acetate and precipitated with 2.5 volumes absolute ethanol. The RNA was pelleted for 20 minutes at 14 000 rpm, washed with 70 % ethanol and air dried before resuspending it in 15 µl RNase-free water. RNA (1 µg) was denatured at 70°C for 10 minutes, and placed on ice. Reverse transcription mix (1 µl dNTPs (10 mM), 4 µl 5 X reaction buffer (15 mM $MgCl_2$, 375 mM KCl, 250 mM Tris.HCl (pH 8)), 0.5 µl RNase inhibitor (40 U/µl), 1 µl random primer (200 µg/ml), 1 µl MMLV Reverse Transcriptase (200 U/µl) was added to each RNA sample in a final volume of 20 µl. The samples were stirred gently, incubated at room temperature for 10 minutes and placed at 37°C for 50 minutes. Finally they were heated to 70°C for 15 min. The synthesised cDNA was kept on ice until it was amplified by PCR or stored at -20°C.

2.8 cDNA Library Expression Screening Using a DNA Ligand

2.8.1 Catenated DNA Probes

Double stranded oligonucleotide probes used to expression screen the cDNA library were generated by annealing complementary synthetic DNA strands (see section 2.4). Each ds oligonucleotide was phosphorylated according to standard reaction conditions specified by the supplier. Typically a phosphorylation reaction contained 120 µg oligonucleotides, 1 X T4 polynucleotide kinase buffer (50 mM Tris.HCl (pH 8.2), 10 mM $MgCl_2$, 0.1 mM EDTA, 5 mM DTT), 100 mM ATP, 50 µCi [γ-^{32}P]ATP, and 30 µl T_4 polynucleotide kinase (1 U/µl, Boehringer Mannheim). The reaction was allowed to proceed at 37°C for 3 hours. The enzyme was inactivated by heating it to 70°C for 10 minutes, and the DNA was purified by organic extraction. The DNA was precipitated by the addition of 0.1 volume 3 M sodium acetate and 2.5 volumes absolute ethanol. The DNA was pelleted (12 000 rpm, 20 minutes, 4°C), washed with 70 % ethanol and resuspended in a final volume of 100 µl, containing 66 mM Tris.HCl (pH 7.5), 5 mM $MgCl_2$, 1 mM DTT, 1 mM ATP and 50 U T_4 DNA ligase (5 U/µl). The ligation reaction was incubated for 16 hours at 15°C, and the catenated DNA attained a mean length of 180 bp.

Radiolabelled probe was prepared using the Prime-a-Gene Labelling System (Promega). The DNA template (25 ng) was denatured at 95°C - 100°C for 2 minutes and combined with 1 X labelling buffer (50 mM Tris.HCl (pH 8), 5 mM $MgCl_2$, 2 mM DTT, 200 mM Hepes (pH 6.6) and 5 A_{260} units/ml random hexadeoxyribonucleotides), 20 µM of each nonlabelled dNTP, 400 µg/ml nuclease-free BSA, 5 µl [α-^{32}P]dCTP (50 µCi, 3000 Ci/mmol) and 1 µl Klenow enzyme (5 U/µl). The volume was adjusted to 50 µl with sterile water, and the reaction was incubated at room temperature for 60 minutes. The reaction was stopped by heating it to 95°C - 100°C for 2 minutes and addition of 2 µl 0.5 M EDTA. The labelled DNA was separated from the unincorporated nucleotides by size exclusion on a Sephadex G-50 spin column (see section 2.5.5). Specific activity of the DNA template was typically 0.4 - 1 X 10^8 cpm/µg.

2.8.2 Preparation of Nitrocellulose Filter Replicas

The DNA ligand screening method was carried out according to a protocol described by Singh et al (1989) (174). The *E.coli* host strain BB4 (Stratagene) was grown to saturation in LB medium containing 0.2 % (w/v) maltose and 10 mM $MgSO_4.7H_2O$ at 37°C. Aliquots of the culture (500 µl)

were infected with 3 - 5 X 10^4 pfu of a λZAP cDNA expression library (a gift from Prof. E. Davidson (Caltech)), which was derived from 24 hour *Strongylocentrotus purpuratus* embryos. The bacterial cells were incubated for 15 minutes at 37°C to allow phage adsorption. Top agarose (9 ml), equilibrated to 47°C, was added to each aliquot of infected cells. The mixture was inverted twice and spread on prewarmed LB / tetracycline plates (150 mm). The LB plates were incubated at 42°C for about 3 hours until tiny plaques were visible. Each LB plate was overlayed with a nitrocellulose filter (Amersham) which had been soaked in 10 mM IPTG for 30 minutes and air dried. The LB plates were incubated at 37°C for 6 hours. The position of the filter was marked on each plate. The filters were lifted off the plates, air dried for 15 minutes at room temperature and immersed in binding buffer (25 mM Tris.HCl (pH 7.9), 25 mM NaCl, 5 mM $MgCl_2$, 0.5 mM DTT) supplemented with 6 M guanidine hydrochloride (GuHCl). The filters were incubated with gentle shaking at 4°C for 10 minutes. All filters were processed in the same petri dish. This step was repeated with fresh binding buffer containing 6 M GuHCl. The second wash was supplemented with an equal volume of binding buffer without GuHCl, and the filters were incubated for 5 minutes. The 100 % dilution step was repeated four times, and the final step was followed by 2 washes with unsupplemented binding buffer for 5 minutes at 4°C. The filters were blocked for 30 minutes at 4°C with a solution containing 3 % (w/v) BSA, 50 mM Tris.HCl (pH 7.5), 50 mM NaCl, 1 mM EDTA, 1 mM DTT and 0.05 % (v/v) Tween-20. This solution was replaced with binding buffer containing 0.15 % (w/v) BSA and 0.05 % (v/v) Tween-20, the filters were immersed for 1 minute at 4°C and then screened with radiolabelled probe.

2.8.3 Screening of Nitrocellulose Filter Replicas

Filters were screened in batches by incubating them in 25 ml binding buffer containing 10^7 cpm of radiolabelled specific DNA probe, and 150 μg poly[d(I-C)] (Boehringer Mannheim). The binding reaction was incubated for 1 hour at 4°C with gentle shaking. The filters were washed with 4 changes of binding buffer (7.5 minutes each wash, 4°C). The filters were air dried on blotting paper and autoradiography was performed overnight (-70°C) with an intensifying screen.

2.8.4 Identification and Purification of Sequence Specific Clones

The presumptive positive plaques were identified by aligning autoradiographs of duplicate filters and identifying overlapping signals. Agarose plugs corresponding to positive signals were stabbed out of the plates with a pasteur pipette and secondary phage stocks were prepared according to Maniatis et al

(1982) (169). The agarose cores containing the positive plaques were placed in 1 ml suspension medium (100 mM NaCl, 0.2 % $MgSO_4.7H_2O$, 50 mM Tris.HCl (pH 7.5) and 2 % (w/v) gelatin) containing one drop of chloroform, and incubated at room temperature for 2 hours. The secondary phage stock were stored at 4°C indefinitely. The secondary phage stock (ca. 5 X 10^3 pfu) were mixed with an aliquot (200 µl) of *E.coli* BB4 cells (overnight culture grown in LB, supplemented with 10 mM $MgSO_4.7H_2O$ and 0.2 % (w/v) maltose) and incubated for 15 minutes at 37°C. Top agarose (3 ml) equilibrated to 47°C was added to the mixture and inverted twice before spreading onto prewarmed LB / tetracycline plates (100 mm). The nitrocellulose filter replicates were prepared as above (see section 2.7.2). The secondary filters were screened with the wildtype recognition site DNA probe (concatenated specific oligo), as well as a control DNA probe (nonspecific oligonucleotide) which lacks the DNA-binding site (see fig 2.1 for sequences). Phage which were detected specifically with the wild-type recognition probe but not the control DNA were plaque purified.

2.9 Lambda Zap Automatic Excision Process

The pBluescript vector (containing the cDNA inserts of interest) was excised from the secondary phage stock according to the Stratagene ExAssist / SOLR System manual. The two *E.coli* host strains, SOLR™ and XL1-*Blue*, were revived by streaking them onto LB plates containing kanamycin (20 µg/ml) and tetracycline (12.5 µg/ml) respectively, and incubating them overnight at 37°C. Single colonies were inoculated into 5 ml LB supplemented with 0.2 % (w/v) maltose and 10 mM $MgSO_4$, and grown overnight at 30°C to prevent overgrowing. The cells were pelleted at 1000 - 2000 X g for 10 minutes, and gently resuspended in 0.5 volumes 10 mM $MgSO_4$. The following components were combined in a test tube: 200 µl XL1-*Blue* cells (OD_{600} = 1), 100 µl isolated phage stock, and 1 µl ExAssist (Stratagene) helper phage (> 1 X 10^6 pfu/ml). The mixture was incubated at 37°C for 15 minutes, 3 ml of 2 X YT medium (10 g NaCl, 10 g yeast extract, 16 g bactotryptone per liter) was added and the mixture was incubated at 37°C for 3 - 5 hours without shaking. The tubes were then heated at 70°C for 20 minutes and spun for 15 minutes at 4000 X g (Beckman JA20.1 rotor). The supernatant, containing the plasmid packaged as a filamentous phage particle, was decanted into a sterile tube and stored at 4°C for up to 2 months. The rescued phagemid was plated by aliquotting 200 µl of SOLR™ (Stratagene) cells (OD_{600} = 1) into separate tubes and adding either 1 µl or 50 µl of phage stock to each tube. The tubes were incubated at 37°C for 15 minutes, and the cells (100 µl) were plated on LB / ampicillin plates and incubated at 37°C overnight. To maintain the pBluescript plasmid, colonies were restreaked onto new LB / ampicillin plates. Glycerol stocks for longterm storage were kept at -70°C.

2.10 Preparation of Genomic DNA from Sea Urchin Sperm

Genomic DNA was isolated according to the method described by Ausubel et al (1987) (151). Adult sea urchins (*P.angulosus*) were induced to spawn by injecting them with 0.5 M KCl (5 ml). Sperm was collected (2 ml) and washed twice with PBS. The sperm was pelleted by centrifugation at 500 X g, and the supernatant was discarded. The sperm was resuspended in at least 2 volumes digestion buffer (100 mM NaCl, 10 mM Tris.HCl (pH 8), 25 mM EDTA (pH 8), 0.5 % (w/v) SDS, and 0.1 mg/ml proteinase K). Cell lysis was achieved by shaking the solution at 50°C for 12 - 18 hours. Genomic DNA was extracted by mixing the lysed sample with an equal volume of PCI (phenol:chloroform:isoamyl alcohol = 25:24:1, pH 8). The phases were separated by centrifugation at 12 000 rpm for 2 minutes in a microfuge. The organic extraction step was repeated four times. The DNA was precipitated by addition of 0.5 volume of 7.5 M ammonium acetate and 2 volumes 100 % ethanol (room temperature). DNA was recovered by centrifugation at 1 700 X g for 2 minutes, and washed with 70 % ethanol. The pellet was air dried and resuspended in TE. The DNA was stored at -20°C.

2.11 The Polymerase Chain Reaction (PCR)

The polymerase chain reaction was used to amplify specific segments of DNA *in vitro*. Either genomic DNA, cDNA or plasmid DNA was used as a template for the amplification. Generally, a basic protocol (as described by Ausubel et al (1987) (151)) was applied. First the samples were denatured at 94°C, then the primers were annealed to the single-stranded DNA (annealing temperatures depend on primer sequence and length) and finally the extension reactions using Taq DNA polymerase were performed at 72°C or 68°C. The 3 steps were repeated 25 to 40 times in a Stratagene Robocycler Gradient 40.

2.11.1 Amplification of Specific DNA-Fragments

The polymerase chain reaction was carried out according to Ausubel et al (1987) (151) and was modified depending on the template DNA. PCR using genomic DNA or cDNA was performed in a final reaction volume of 50 µl in 0.5 ml PCR tubes. The reaction mix was prepared by combining template DNA (100 - 1000 ng), 5 µl 10 X PCR reaction buffer (200 mM Tris.HCl (pH 8.4), 500 mM KCl, GibcoBRL), 1 µl dNTPs (10 mM), 0.4 µl Taq (5 U/µl Taq DNA polymerase from *Thermus aquaticus* YT1, GibcoBRL), 1 µl of each primer (20 pmol/µl) and adjusting the final $MgCl_2$

concentration to 0.5 - 4 mM. Sometimes DMSO, formamide and / or glycerol were included in the reactions. The samples were adjusted to 50 µl with nuclease-free water and covered with 50 µl mineral oil. Thermal cycling took place in a Stratagene Robocycler Gradient 40. Generally the programme was set for 35 cycles; each cycle constituted an incubation at 94°C for 1 min, followed by a 55°C incubation for 1 min, and the final step was a 1 min incubation at 72°C or 68°C. The final step included a 10 min incubation at 72°C or 68°C.

2.11.2 Colony Screening for Recombinant Plasmid by PCR

For screening of bacterial colonies by PCR a single colony of a freshly plated transformed bacterial culture was placed into 20 µl of sterile water. Bacterial mix (5 µl) was transferred into a 0.5 ml PCR tube and each screening reaction was prepared by adding 0.5 µl 10 mM dNTP, 2 µl 10 X PCR buffer (200 mM Tris.HCl (pH 8.4), 500 mM KCl), 0.6 µl 50 mM $MgCl_2$, 0.4 µl Taq polymerase (5 U/µl Taq DNA polymerase, GibcoBRL) and 1 µl of each primer (20 pmol). The final reaction volume was adjusted to 20 µl with water and samples were covered with 50 µl of mineral oil. Thermal cycling was performed as described above (the annealing temperature was 50°C instead of 55°C) in a Stratagene Robocycler Gradient 40. Glycerol stocks of each colony were made by adding of 1 ml LB containing 50 mg/ml ampicillin to the remaining 15 µl of water (starter solution). The cell suspension was incubated overnight at 37°C with shaking and glycerol was added to a final concentration of 20 % (v/v). The bacterial glycerol stocks were stored at -20°C or at -70°C.

2.11.3 Rapid Amplification of cDNA Ends (RACE)

The cDNA amplification protocol outlined in the ClonTech Marathon RACE manual was followed for all reactions. Reactions were performed on ice unless otherwise indicated.

2.11.3.1 First Strand cDNA Synthesis

Total RNA (1 µg) isolated from 14 hour sea urchin embryos (*P.angulosus*) was combined with 1 µl cDNA synthesis primer (10 µM). The volume was adjusted to 5 µl with nuclease-free water, and the contents of the tube was mixed and spun briefly. The mixture was incubated at 70°C for 2 minutes, and cooled on ice. First strand synthesis was performed by combining the RNA / primer mix with first strand buffer (1 X), 1 mM dNTP mix, and 10 U MMLV reverse transcriptase (all ClonTech) in a total

volume of 10 µl. The contents of the tube was mixed briefly and collected at the bottom of the tube. The reaction was incubated at 42°C for 1 hour. First strand synthesis was terminated by placing the tube on ice.

2.11.3.2 Second Strand Synthesis

First strand cDNA (10 µl) was combined with 48.4 µl sterile water, 1 X second-strand buffer (ClonTech), 1.6 µl dNTP mix (10 mM) and 5 X second-strand enzyme cocktail containing RNase H (0.25 U/µl), *E.coli* polymerase I (6 U/µl) and *E.coli* DNA ligase (1.2 U/µl). The mixture was incubated at 16°C for 90 minutes. T4 DNA polymerase (10 units) was added to the reaction and the incubation was continued at 16°C for 45 minutes. Second strand synthesis was terminated by addition of 4 µl EDTA / glycogen mix (ClonTech) to the reaction. The DNA was extracted with 100 µl of phenol:chloroform:isoamyl alcohol (25:24:1), followed by a chloroform:isoamyl alcohol (24:1) extraction. The aqueous layer was removed and supplemented with 0.5 volume 4 M ammonium acetate and 2.5 volumes 96 % (v/v) ethanol. The DNA was precipitated in a microfuge at 14 000 rpm (room temperature) for 20 minutes. The DNA pellet was washed with 300 µl 80 % (v/v) ethanol, and the supernatant was removed. The pellet was air dried to evaporate the residual ethanol. The precipitate was dissolved in 10 µl water and the cDNA (2 µl) was analysed on an 1.2 % agarose / EtBr gel with suitable DNA size markers.

2.11.3.3 Adaptor Ligation

The double stranded cDNA (5 µl) was combined with 2 µl Marathon cDNA Adaptor (10 µM, ClonTech), 2 µl 5 X DNA ligation buffer (25 mM Tris.HCl (pH 7.8), 5 mM $MgCl_2$, 0.5 mM DTT, 0.5 mM ATP and 5 % (w/v) PEG (MW 8000)), and 1 µl T4 DNA ligase (1 U/µl). The reaction was incubated at 16°C overnight, and the DNA ligase was inactivated by heating to 70°C for 5 minutes. The adaptor ligated cDNA was diluted 1:10 with Tricine / EDTA buffer, and heated to 94°C for 2 minutes to denature the double stranded cDNA. The tube was cooled on ice and stored at -20°C.

2.11.3.4 PCR Amplification of 5' and 3' cDNA Ends

Each PCR reaction was performed in a final volume of 50 µl, and contained 5 µl diluted (1:100) adaptor ligated ds cDNA, 1 X KlenTaq PCR buffer (40 mM Tricine-KOH (pH 9.2), 15 mM KOAc,

3.5 mM $Mg(OAc)_2$ and 75 µg/ml BSA, ClonTech), 1 µl dNTP mix (10 mM), 1 X Advantage KlenTaq Polymerase Mix (1 % glycerol, 0.8 mM Tris.HCl (pH 7.5), 1 mM KCl, 0.5 mM $(NH_4)_2SO_4$, 2 µM EDTA, 0.1 mM β-mercaptoethanol, 0.005 % Thesit, ClonTech) and 1 µl of the appropriate primers (10 µM). The 5' and 3' RACE reactions were performed using a combination of adaptor primer 1 (AP1, Clontech) and individual gene specific primers (SP2 and SP1 respectively, see Appendix VII). Three negative control PCR reactions were performed using 5 µl diluted adaptor ligated ds cDNA and one of the API, SP1 and SP2 primers individually. The PCR reaction mixtures were overlayed with two drops of mineral oil and the thermal cycling was performed in an automatic thermocycler (Stratagene Robocycler Gradient 40). The programme was set for 30 cycles, each cycle involved a 30 second denaturation at 94°C, a 60°C annealing step for 30 seconds, and the extension was a 4 minute incubation at 68°C. Finally the sample was incubated at 68°C for 10 minutes. Aliquots (5 µl) of each sample were analysed on an 1.2 % agarose / EtBr gel. RACE products were subsequently characterised by Southern blot analysis (see section 2.12) and automated DNA sequencing.

2.11.3.5 Fusion of 5' and 3' RACE Products to Form Full Length cDNA

Full length cDNA was generated by fusion of the 5' and 3' RACE PCR fragments (see section 2.11.3.4), as described in the ClonTech protocol. The 5' and 3' RACE PCR products were purified on an 1 % (w/v) low melting point agarose gel (Seaplaque) in TAE buffer with ethidium bromide (0.3 µg/ml). The DNA was separated from the agarose using a Wizard PCR Preps Column (Promega) and resuspended in a final volume of 50 µl. The 5' / 3' fusion reaction was performed in 0.5 ml PCR tubes, containing 50 ng of both 5' and 3' PCR products (see section 2.11.3.4), 1 X KlenTaq PCR buffer (40 mM Tricine-KOH (pH 9.2), 15 mM KOAc, 3.5 mM $Mg(OAc)_2$ and 75 µg/ml BSA, 0.5 µl dNTP mix (10 mM) and 1 X Advantage KlenTaq Polymerase Mix (1 % (v/v) glycerol, 0.8 mM Tris.HCl (pH 7.5), 1 mM KCl, 0.5 mM $(NH_4)_2SO_4$, 2 µM EDTA, 0.1 mM β-mercaptoethanol, 0.005 % Thesit). All reagents were from ClonTech. The volume was adjusted to 20 µl with water. The contents of the tube was overlaid with 2 drops of mineral oil and thermal cycling was performed in a Robocycler Gradient 40 (Stratagene) at 94°C for 30 seconds and 68°C for 30 minutes. The program was set for 10 cycles. The fusion product was analysed on an 1 % (w/v) agarose gel in TBE. The fused full-length cDNA was diluted (1:100) with Tricine / EDTA (10 mM Tricine-KOH (pH 8.5), 0.1 mM EDTA) and 5 µl of the dilution was used in a PCR amplification, containing 1 X KlenTaq PCR buffer, 1 µl dNTP (10 mM), 1 µl AP1 (10 µM), 1 µl cDNA synthesis primer (10 µM) and 1 X KlenTaq Polymerase Mix (all reagents from ClonTech). The final volume of the reaction was adjusted to 50 µl with water. The mixture was overlaid with 2 drops of mineral oil, incubated at 94°C for 1 minute, and thermal cycling

was performed for 15 cycles. Each cycle was 30 seconds at 94°C, 30 seconds at 55°C followed by 68°C for 5 minutes. The PCR amplification reaction was analysed on an 1 % (w/v) agarose gel in TBE. The full length product was purified on a preparative low melting point agarose gel (Seaplaque) in TAE, and the full length cDNA sample was recovered using a Wizard PCR Prep column (Promega).

2.12 Southern Blot Analysis

The Southern blotting procedure was performed as described in the Amersham Hybond booklet. DNA samples were electrophoresed in an 1 % (w/v) agarose / TBE gel. After electrophoresis the agarose gel was placed in denaturing solution (1.5 M NaCl, 0.5 M NaOH) for 30 minutes at room temperature with gentle agitation. The gel was rinsed with water and placed in neutralisation buffer (1.5 M NaCl, 1 mM EDTA, 0.5 M Tris.HCl (pH 7.2) for 15 minutes at room temperature with gentle agitation. The water wash and the neutralisation step were repeated once. For the capillary blot a glass dish was filled with blotting buffer (20 X SSC). The platform was covered with a wick made from three sheets of Whatman 3MM filter paper saturated with 5 X SSC. The gel was placed on the wick and surrounded with cling wrap to prevent the blot from drying out. A sheet of Hybond N+ membrane (Amersham) was cut to the exact size of the gel and placed on the gel after wetting it in 5 X SSC. The membrane was covered with three sheets of Whatman 3MM paper cut to size and wetted with blotting buffer. A stack of absorbent paper towels was placed on top of the 3MM paper (approximately 5 cm high). A glass plate with a 1 kg weight was placed on top of the paper towels and the transfer was allowed to proceed for 16 hours. The blotting apparatus was dismantled and the membrane was marked with a pencil to allow later identification of the tracks. The membrane was washed briefly in 2 X SSC to remove any adhering agarose. It was then placed (with the DNA side up) on three layers of Whatman 3MM paper soaked in 0.5 M NaOH for 20 minutes in order to fix the DNA. The membrane was immersed in 5 X SSC for less than a minute. The membrane was placed in a hybridisation box with 80 ml of prehybridisation solution (6 X SSC, 0.4 % (w/v) SDS, 5 X Denhardt's, 20 mM NaH_2PO_4 (pH 7.5), 0.5 mg/ml denatured herring sperm DNA) at 60°C and was gently agitated for one hour. The prehybridisation solution was replaced with hybridisation buffer (6 X SSC, 20 mM NaH_2PO_4 (pH 7.5), 0.5 mg/ml denatured herring sperm DNA) containing the heat denatured labelled probe (50 ng at 0.5 X 10^6 dpm/ng). The membrane was incubated in hybrisation buffer for 3 hours at 60°C. It was washed three times in 6 X SSC / 0.1 % SDS for 10 minutes at 42°C, three times in 1 X SSC / 0.1 % SDS for 10 minutes at 60°C, and three times in 0.1 X SSC / 0.1 %

SDS for 10 minutes at 65°C. The membrane was wrapped in plastic and autoradiographed overnight at -70°C.

2.13 Cloning of cDNA into Plasmid Vectors

2.13.1 Ligation of PCR Products into T-Vectors

PCR products were ligated into the pMOS Blue T-vector (Amersham) or pGEM-T vector (Promega) according to the supplier's recommendations. The pMOS Blue T-vector ligation reaction was prepared by combining 1 µl ligase buffer (10 X), 0.5 µl DTT (100 mM), 0.5 µl ATP (10 mM), 1.0 µl vector (50 ng/µl) and 0.5 µl T4 DNA ligase (2 - 3 Weiss units) with the PCR products in a final volume of 10 µl. All reagents were available in the Amersham pMOS Blue T-vector kit. Insert DNA was either purified by Wizard™ PCR preps (see section 2.5.2) or used directly in low melting agarose slices. The reactions were stirred gently with pipette tips and incubated at 16°C overnight. The ligation reactions were stored at 4°C until they were transformed into bacteria (see section 2.2.2)

The pGEM-T vector system was used as outlined in the pGEM-T Vector Systems manual (Promega). PCR products were gel purified (see section 2.5.2) prior to the ligation reaction. Each ligation reaction contained 1 X T4 DNA Ligase Buffer (30 mM Tris.HCl (pH 7.8), 10 mM $MgCl_2$, 10 mM DTT and 1 mM ATP), 50 ng pGEM-T Vector, 3 Weiss units T4 DNA ligase, and 150 ng PCR product. The final volume was adjusted to 10 µl using nuclease-free water. The reactions were mixed by pipetting, and they were incubated at 16°C for 2 - 18 hours, 2 µl of the ligation mixture was transformed into bacteria (see section 2.2.2).

2.13.2 Subcloning cDNA Inserts

cDNA fragments obtained by DNA expression screening were subcloned into the prokaryotic expression vectors pET-29b(+) (Novagen), pGEX-3X (Pharmacia) and the eukaryotic expression vector pCIS (185). The cDNA fragments were released from pBluescript using restriction enzymes *EcoR1*, *BamH1* / *Sal1* or *Sac1* / *Xho1*, which did not cut the inserts. The prokaryotic expression vectors were linearised with the same combinations of enzymes. The cDNA insert cloned into the pCIS vector was released from the pET-29b(+) vector using the *Xba1* / *Not1* sites. cDNA inserts were separated from the original vector DNA by electrophoresis on 1 % (w/v) low melting point agarose /

TAE gels. The DNA fragments were excised from the gel with a sterile blade after visualisation of the bands using UV light (λ = 342 nm). DNA was purified from the agarose using the Gene-Clean protocol (USB). Essentially the DNA / agarose was dissolved in 3 volumes 6 M NaI at 55°C for 5 minutes, and 5 µl glass powder suspension was added. The DNA was allowed to adhere to the glass powder by incubating the mixture on ice for 5 minutes. The suspension was centrifuged for 5 - 10 seconds (12 000 rpm) and the glass powder pellet was rinsed with 50 % ethanol rinse buffer. (A 50-fold excess of buffer to glass powder was used.) The suspension was centrifuged for 5 - 10 seconds, and the wash step was repeated twice. After the final centrifugation, the glass powder was resuspended in 1 - 2 volumes TE and the DNA was eluted by incubation at 55°C for 2 - 5 minutes. The DNA concentrations were estimated by comparison to DNA standards dotted onto agarose / ethidium bromide plates. Ligation reactions were performed according to the pMOS*Blue* T-vector kit (Amersham, and see section 2.13.1), ensuring a vector : insert ratio of 1 : 10. Vector (pET-29b(+), pGEX-3X or pCIS (see Appendix VIII)) was combined with the appropriate insert in a final ligation reaction volume of 10 µl and incubated at 16°C for 16 hours. The ligation mixture (5 µl) was used to transform competent JM109 or DH5α cells (see section 2.2.2), which were then plated on LB plates containing the appropriate antibiotic.

2.14 Bacterial Target Gene Expression

2.14.1 Recombinant Protein Expression from pBluescript

Pilot experiments were performed as described in the Stratagene manual in order to optimise expression of the β-gal fusion protein from pBluescript. Several *E.coli* strains (eg XL1-*Blue*, JM109, DH5α and SOLR^{TM} cells) containing the pBluescript plasmid with the insert of interest were grown to mid-log phase (OD_{600} = 0.2) at 37°C. IPTG (100 mM) was added to a final concentration of 1 mM. The cells were grown to stationary phase (OD_{600} = 1) and pelleted at 1600 X g (Beckman, JA 20.1 rotor) for 15 minutes. The cell pellet was resuspended 1 : 4 (w : v) in lysis buffer (50 mM Tris.HCl (pH 8), 1 mM EDTA, 1 mM PMSF, 10 % (w/v) sucrose). Lysozyme was added to a final concentration of 1 mg/ml and the cells were incubated on ice for 10 minutes. The cells were supplemented with Triton X-100 (0.1 % final concentration) and the lysed cells were pelleted at 20 000 rpm (Beckman, JA 20.1 rotor) for one hour. The supernatant contained the soluble protein and was stored at -70°C. Both the supernatant and the cell pellets were analysed on SDS gels, which were either Coomassie stained or silver stained (see section 2.24).

2.14.2 Recombinant Protein Expression from pET-29b(+)

The recombinant pET-29b(+) plasmids were established in a variety of *E.coli* strains (viz HB101, JM109 and DH5α) by transforming the respective cells and allowing them to grow on LB / kanamycin plates at 37°C. The plasmids were isolated from these hosts and transformed into two different expression hosts, viz BL21DE3 or BL21DE3(pLysS) (both Novagen).

Pilot experiments establishing the optimum conditions of recombinant protein expression were performed according to the pET System Manual (Novagen). Expression conditions were optimized by varying several conditions, eg *E.coli* hosts in which the plasmid was established, expression host, culture volume, length of induction and temperature of induction (16°C - 37°C). Single bacterial colonies were picked from fresh LB / kanamycin plates, and grown to saturation in 10 ml LB / kanamycin overnight. Fresh LB / kanamycin (1.8 ml) was then inoculated with 200 µl of starter culture and grown for 2 hours until OD_{600} = 0.4 - 0.6. The cells were induced to express recombinant protein with a final concentration of 1 mM IPTG. The induction was allowed to proceed between 1 and 6 hours. Protein induction was assayed by removing 200 µl of cell culture and electrophoresing the cell pellet by SDS-PAGE.

2.14.3 Purification of Recombinant Proteins Expressed from pET-29b(+)

2.14.3.1 Isolation of Soluble Protein Fraction

Isolation of the soluble protein fraction from *E.coli* cells was performed according to a method for recombinant protein isolation from bacteria (Stratagene manual). A 10 ml culture of BL21DE3 cells (OD_{600} = 0.6) containing the recombinant pET-29b(+) vector was induced to express recombinant protein (see section 2.14.2). Cells were pelleted (6 000 X g, 15 minutes, JA20.1 rotor (Beckman)), and the supernatant was removed. All subsequent steps were performed on ice. The cell pellet was resuspended in 2 ml lysis buffer (50 mM Tris.HCl (pH 8), 1 mM EDTA, 1 mM PMSF and 10 % (w/v) sucrose). The cells were lysed by the addition of lysozyme (1 mg/ml final concentration). The mixture was incubated on ice for 10 minutes, and supplemented with Triton X-100 (final concentration 0.1 % (v/v)). The mixture was incubated on ice for 10 minutes and the cell debris was pelleted by centrifugation (18 000 rpm, 50 minutes, JA20.1 rotor (Beckman)). Part of the supernatant was retained and stored in aliquots at -70°C, the remaining supernatant was dialysed against dialysis

buffer (20 mM Tris.HCl (pH 8), 100 mM KCl, 0.2 mM EDTA, 20 % (v/v) glycerol, 4 mM $MgCl_2$ and 2 mM $ZnCl_2$), divided into aliquots and stored at -70°C.

2.14.3.2 Affinity Purification of Soluble Recombinant Protein Using a Ni^{2+} Column

Recombinant protein was induced to express as described above (see section 2.14.2), and proteins were extracted in binding buffer (40 mM imidazole, 4 M NaCl, 160 mM Tris.HCl (pH 7.9)) supplemented with 4 mM PMSF and 20 µg/ml leupeptin. The cells were placed in an icebath and sonicated in small bursts until the suspension was no longer viscous. The preparation was centrifuged for 20 minutes (39 000 X g) to remove the cell debris. The protein extraction procedure was sometimes varied by the addition of lysozyme (1 mg/ml final concentration) and Triton X-100 (0.1 % final concentration), and the length of sonication was also varied. All steps were performed at 4°C. The soluble protein extract was then subjected to affinity chromatography over a 2.5 ml Ni^{2+}NTA agarose column which had been washed with 3 column volumes of sterile water, 5 column volumes of 1 X charge buffer (50 mM $NiSO_4$), and 3 column volumes of 1 X binding buffer (5 mM imidazole, 0.5 M NaCl, 20 mM Tris.HCl (pH 7.9)). The bacterial cell extract was loaded onto the column and the flow rate was adjusted to 420 µl/min. The column was washed with 25 ml of 1 X binding buffer, 15 ml of 1 X wash buffer and the protein was eluted with 15 ml of 1 X elution buffer (1 M imidazole, 0.5 M NaCl, 20 mM Tris.HCl (pH 7.9)). The eluate was collected in 1 ml fractions, which were stored at -70°C. Aliquots of each fraction were TCA precipitated and analysed by 12 % SDS-PAGE (see section 2.24).

2.14.3.3 Inclusion Body Isolations

Several methods were applied in order to isolate expressed recombinant protein from inclusion bodies.

Cells (100 ml of culture) were induced to express recombinant protein as outlined in the pET Manual. The cells were pelleted (5000 X g, 5 minutes) and the supernatant was drained completely. The cells were resuspended in 4 ml 1 X binding buffer (5 mM imidazole, 0.5 M NaCl, 20 mM Tris.HCl (pH 7.9)). The resuspended cell pellet was sonicated in brief bursts to shear the DNA. The nonviscous lysate was centrifuged at 20 000 X g for 15 minutes to remove the debris. The supernatant was removed and the pellet was resuspended in 20 ml of 1 X binding buffer (sonication was necessary to resuspend the pellet). The inclusion bodies were pelleted as above and the centrifugation /

resuspension / sonication step was repeated several times in order to release more trapped protein. The supernatant from the final centrifugation was removed and the pellet was resuspended in 5 ml 1 X binding buffer containing 6 M guanidine HCl. The mixture was incubated on ice for one hour to completely dissolve the protein. The remaining insoluble material was removed by centrifugation at 39 000 X g for 20 minutes. The supernatant was either stored at -70°C for further analysis and column purification, or it was dialysed into dialysis buffer and then stored at -70°C. A similar isolation procedure was followed where the guanidine hydrochloride in the binding buffer was replaced with 6 M urea.

Alternatively cells were induced to express recombinant protein as above, and the cell fractions were analysed according to Höög et al (1991) (122). The cells were pelleted at 5000 X g for 5 minutes. The supernatant was discarded and the pellet was resuspended in 1/50 volume of Buffer Z (20 mM Hepes (pH 7.4), 40 mM KCl, 0.1 mM EDTA, 1 mM DTT, 20 % (v/v) glycerol). Lysozyme was added (final concentration 2 mg/ml). The cells were incubated for 2 hours on ice and 1 mM PMSF was added to the mixture. The cell lysis was completed by 4 - 5 bursts of sonication. The suspension was supplemented with 1 mM DTT, 0.5 % (v/v) NP-40 and 5 % (w/v) sucrose. The insoluble material was removed by centrifugation at 10 000 X g for 15 minutes and the soluble supernatant was supplemented with 0.1 volume 4 M ammonium sulphate and glycerol (20 % (v/v) final concentration). The mixture was centrifuged at 100 000 X g to remove ribosomes and other particles, and the soluble supernatant was either stored at -70°C or dialysed into dialysis buffer (20 mM Tris.HCl (pH 8), 100 mM KCl, 0.2 mM EDTA, 20% glycerol (v/v), 4 mM $MgCl_2$ and 2 mM $ZnCl_2$) and then stored at -70°C.

A third method of inclusion body isolation was performed according the method outlined by Lin and Cheng (1991) (186). A 20 ml cell culture (OD_{600} = 0.4 - 0.6) was induced to express recombinant protein by supplementing it with 1 mM IPTG. The induction was allowed to proceed for 3 hours at 37°C. The cells were harvested by centrifugation at 5 000 rpm for 15 minutes. The cell pellet was frozen at -70°C and subsequently thawed on ice, it was then resuspended in 1 ml Buffer A (20 mM Tris.HCl (pH 7.5), 20 % (w/v) sucrose and 1 mM EDTA). The suspension was incubated on ice for 10 minutes. The cells were pelleted at 4 000 X g for 20 minutes and resuspended in ice cold water to release spheroplasts, which were then pelleted at 8 000 X g and resuspended in 250 µl Buffer P (1 X PBS, 5 mM EDTA, 1 µg/ml leupeptin, 20 µg/ml aprotinin and 0.5 mM PMSF). The cell membranes were lysed by sonication (3 high intensity bursts for 10 seconds) and RNase A (100 µg/ml) and DNase 1 (400 µg/ 10 ml) were added to the lysate. The mixture was incubated at room temperature

for 10 minutes and 800 µl of Buffer P was added. The inclusion bodies were pelleted at 13 000 X g for 30 minutes (4°C) and resuspended in 800 µl Buffer W (1 X PBS, 25 % (w/v) sucrose, 5 mM EDTA and 1 % (v/v) Triton X-100). The suspension was incubated on ice for 10 minutes and the inclusion bodies were pelleted for 10 minutes at 25 000 X g. The inclusion body wash procedure was repeated twice and finally the inclusion bodies were resuspended in 250 µl Buffer D (50 mM Tris.HCl (pH 8), 5 M guanidine HCl, and 5 mM EDTA). The protein aggregates were sonicated for 5 second pulses in order to solubilise them. The proteins were incubated on ice for one hour and centrifuged for 30 minutes at 12 000 X g. The supernatant was added to 2.5 ml Buffer R (50 mM Tris.HCl (pH 8.0), 1 mM DTT, 20 % (v/v) glycerol, 1 µg/ml leupeptin, 20 µg/ml aprotinin and 0.5 mM PMSF). The mixture was gently stirred overnight to renature the recombinant proteins and the supernatant was clarified by centrifugation at 13 500 X g for 30 minutes. The supernatant (containing soluble protein) was analysed by SDS-PAGE and aliquoted to store at -70°C.

2.14.4 Recombinant Protein Expression from pGEX-3X

The procedure outlined in the GST Gene Fusion System Manual (Pharmacia) was used in order to screen recombinants for the expression of fusion proteins. pGEX-3X recombinant plasmids were transformed into different *E.coli* hosts (MC1061 and DH5α), and several colonies from each host were picked into separate tubes containing 10 ml LB / ampicillin. Liquid cultures were grown at 37°C overnight. The starter culture (200 µl) was used to inoculate 1.8 ml LB / ampicillin. Cells were grown at 37°C for 1 hour until OD_{600} = 0.4 - 0.6. Recombinant protein expression was induced by addition of 20 µl IPTG (100 mM). Cells were incubated for 3 hours at 37°C, they were then pelleted for 1 minute in a microfuge and the pellet was resuspended in 200 µl SDS sample application buffer (1 X). The samples (20 µl) were analysed by 12 % SDS-PAGE.

2.15 Eukaryotic Recombinant Gene Expression in COS-1 Cells

(All reagents were at 37°C before use.) HBS buffer (pH 7.1) was 137 mM NaCl, 5 mM KCl, 0.7 mM NaH_2PO_4, 21 mM Hepes. Hepes-DMEM was 10 mM Hepes in DMEM (Dulbecco's Modified Eagle's Medium). The DEAE Dextran stock solution was 3 mg/ml in HBS buffer, filter sterilised and stored at 4°C. The chloroquine stock solution (10 mM) was made freshly and filter sterilised. The chloroquine DMEM solution was 200 µM chloroquine and 2 % fetal calf serum in DMEM.

COS-1 cells were plated at a cell density of 7 X 10^6 cells per flask. They were supplemented with 25 ml DMEM and 10 % fetal calf serum (FCS), and allowed to grow at 37°C for 24 hours. The HBS / DEAE dextran (3 mg/ml) was diluted 10-fold in DMEM / P-5 and 80 µg DNA sample (pCIS vector only or vector containing insert) was added to 10 ml HBS / DEAE dextran / DMEM / P-5. The medium was aspirated from the cells and 15 ml Hepes / DMEM / P-5 was added to each flask, which was also aspirated from the cells. The DNA solution (as prepared above) was added to the cells and they were incubated at 37°C for 4 hours in a 10 % CO_2 incubator. The medium was aspirated from the cells and 20 ml of chloroquine / DMEM solution was added to each flask. The cells were incubated at 37°C for 50 minutes. The medium was aspirated from the cells, and the cells were washed gently using 15 ml serum-free DMEM / P-5. The medium was aspirated from the cells and replaced with 20 ml DMEM / 10 % FCS / P-5. The flasks were incubated at 37°C for 18 - 24 hours and processed into cell extracts and nuclear extracts using a method described by Jiang et al (1995) (187). Cells were placed in Buffer H (140 mM NaCl, 4 mM KCl, 20 mM Hepes, 1 mg/ml BSA, 8.3 mM glucose, 1 mM DTT and 0.5 mM PMSF) containing 1 mM EDTA for 10 minutes at 37°C. The cells were scraped from the dishes using a rubber policeman and pelleted. The cell pellet was washed in 10 ml PBS, and resuspended in 2 X Buffer H containing 20 % (v/v) glycerol. The cells were exposed to 4 cycles of rapid freezing (dry ice) and thawing. The cellular debris and nuclei were separated from the cellular extract by centrifugation at 12 000 rpm for 15 minutes (4°C). The whole cell extract was aliquoted and stored at -70°C. Nuclear pellet was resuspended in 50 µl of 2 X Buffer H / 40 % (v/v) glycerol, and an equal volume 2 M KCl was added to the suspension dropwise, whilst mixing. The mixture was rolled at 4°C for 60 minutes and the extracts were centrifuged for 20 minutes at 12 000 rpm. The supernatant was removed and stored in aliquots at -70°C.

2.16 *In Vitro* Coupled Transcription / Translation

Eukaryotic *in vitro* translations were performed using either the TNT T3 Coupled Reticulocyte Lysate System or the TNT T7 Quick Coupled Reticulocyte Lysate System (both Promega). The kits were used according to the supplier's recommendations. All components were stored at -70°C. The lysate was stored in aliquots and never frozen and thawed more than twice. All reactions were performed using RNase-free glassware, plasticware and chemicals. The DNA templates were pBluescript or pGEM-T vectors containing the cDNA inserts of interest. "Translational grade" ^{35}S-methionine (15 mCi/ml, Amersham) or Trans ^{35}S-label^{TM} (~ 10 mCi/ml, ICN) was used when radiolabelling the protein products. Sometimes an amino acid mixture containing no radiolabel was used. The reaction components (25 µl TNT Rabbit Reticulocyte Lysate, 2 µl TNT Reaction Buffer, 1 µl TNT RNA

Polymerase, 1 µl amino acid mixture Minus Methionine (1 mM), 4 µl ^{35}S-methionine (15 mCi/ml), 1 µl RNasin Ribonuclease Inhibitor (40 U/µl, Promega), and 1 µg of DNA template) were assembled in a 1.5 ml microcentrifuge tube. The volume was adjusted to 50 µl with RNase-free water. The reaction was incubated for 2 hours at 37°C and the samples were stored at -20°C. Samples were analysed by SDS-PAGE (see section 2.24) and EMSA (see section 2.21).

2.17 Growth of Sea Urchin Embryos

Sea urchins (*P. angulosus*) were collected in rock pools on the West Coast of the Cape Peninsula, at Melkbos Beach. The sea urchins were induced to spawn by injecting them with 5 ml 0.5 M KCl. The eggs were collected and filtered through two layers of cheesecloth, after which they were washed three times with filtered sea water. (Sea water was filtered through Whatman 3MM paper.) 50 ml sperm (1:500 dilution) was added per litre of 4 % (v/v) egg suspension which contained 100 mg/l penicillin and 50 mg/l streptomycin in sea water. The cultures were shaken at 180 rpm (21°C) for 14 hours. All subsequent steps were performed at 4°C. The cultures were allowed to settle, centrifuged (4 000 rpm, 1 sec, JA14 rotor) and washed three times with 0.5 M KCl. The embryos were either frozen at -70°C in 2 to 3 volumes nuclear storage buffer (Buffer A containing 25 % (v/v) glycerol) or processed immediately after washing once with 2 - 3 volumes 0.25 M sucrose, 10 mM Tris.HCl (pH 8), 0.1 mM EDTA. (Buffer A is 15 mM Tris.HCl (pH 8), 65 mM KCl, 15 mM NaCl, 0.15 mM spermine, 0.5 mM spermidine, 0.2 mM EDTA, 0.2 mM EGTA, 10 mM ß-mercaptoethanol and 0.1 mM PMSF.)

2.18 Preparation of Nuclei

Nuclei were prepared using several different methods.

2.18.1 Method by Morris and Marzluff (1983) (188)

All steps were performed at 4°C. The embryos were washed three times with 4 volumes 0.5 M KCl and once with 0.25 M sucrose in 10 mM Tris.HCl (pH 8), 0.1 mM EDTA, after which they were resuspended in Buffer A containing 0.32 M sucrose. The suspension was homogenised with a tight dounce for twenty strokes and cell breakage was monitored by light microscopy. The homogenate was adjusted to 1.8 M sucrose by adding the required volume of 2.3 M sucrose in Buffer A. The

homogenous suspension was centrifuged at 90 000 g (Beckman ultracentrifuge) for 50 minutes. The nuclei were either processed directly or frozen in nuclei storage buffer at -70°C.

2.18.2 Hexylene Glycol Method

A method based on several procedures was followed (1). All steps were carried out at 4°C. Embryos were washed once with Hexylene glycol in Buffer A (HexA), resuspended in the same buffer and rolled for two hours. The suspension was homogenised for twenty strokes with a tight dounce and the cell breakage was monitored by microscopy. The intact nuclei were pelleted at 5 000 rpm for one minute, after which they were washed with HexA buffer and resuspended in the minimum volume of the same buffer. The suspension was adjusted to 1.8 M sucrose by adding the required volume of 2.3 M sucrose in Buffer A. The nuclei were pelleted by centrifugation at 90 000 g (Beckman Ultracentrifuge) for 50 minutes. The nuclei were processed directly.

2.18.3 Method by Calzone et al (1991)

Preparation of nuclei was carried out essentially as described (120). The fresh embryos were washed with 1 M glucose, after which they were resuspended in buffer D / 0.36 M sucrose. Buffer D was 10 mM Tris.HCl (pH 8), 1 mM EDTA, 1 mM EGTA, 1 mM spermidine, 1 mM PMSF. At this stage the embryos were frozen at -70°C. The frozen embryos were crushed and uniformly thawed for further processing. The nuclei were collected by centrifugation at 2 500 g for 40 minutes and washed in Buffer D three times. The nuclei were subsequently washed three times in Buffer D containing 0.1 % (v/v) NP-40. The nuclei were resuspended in 3 - 3.7 pellet volumes lysis buffer (10 mM HEPES (pH 7.9), 1 mM EDTA, 1 mM EGTA, 1 mM spermidine-Tris-HCl, 1 mM DTT and 10 % (v/v) glycerol). The nuclei were processed directly.

2.19 Preparation of Nuclear Extracts

All steps were performed at 4°C. Nuclei from 1 litre of fourteen-hour culture were resuspended in 32 ml lysis buffer (15 mM Tris.HCl (pH 8), 100 mM KCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 1 mM DTT, and 0.1 mM PMSF). Ammonium sulfate (4 M) was added dropwise and with immediate mixing to a concentration of 0.4 M over a period of 10 minutes. The solution was rolled for 30 minutes at 4°C and centrifuged at 90 000 g for 45 minutes. The pellet was discarded and 0.25 g/ml solid $(NH_4)_2SO_4$ was

added to the supernatant. The suspension was rolled for 45 minutes and centrifuged at 90 000 g for 15 minutes. The pellet was resuspended in 2 ml 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) for every litre of original 4 % (v/v) culture. The extract was dialysed for 5 hours against 200 volumes of the same buffer, centrifuged for 20 minutes at 15 000 rpm and the supernatant ("nuclear extract") stored in aliquots at -70°C. The protein concentration typically ranged between 5 and 15 mg/ml.

2.20 Protein Determination with the Folin Ciocalteu Reagent

Working standards were prepared from bovine serum albumin (Boehringer) by making a series of solutions containing 0 - 10 µg of BSA. Assays were performed in triplicate for the standard protein solutions and in duplicate for the unknown protein.

The protein solutions were adjusted to 1 ml with water and 10 µl sodium deoxycholate (1.76 % (w/v)) was added to each solution. Dilutions were mixed well and incubated for 15 minutes at room temperature. Proteins were precipitated by the addition of 333 µl TCA (24 % (w/v)) and centrifugation at 18 000 rpm (Beckman, JA 20.1 rotor) for 50 minutes at 4°C. The supernatant was removed carefully and each protein pellet was resuspended in 1 ml Lowry reagent C (100 volumes Na_2CO_3 (2 % (w/v)) in 0.1 N NaOH, 1 volume Cu_2SO_4 (1 % (w/v)), and 1 volume disodium tartrate (2 % (w/v))). This was followed by the addition of 100 µl Folin-Ciocalteu phenol reagent (Merck) to each solution with rapid mixing. The reactions were allowed to proceed for 75 minutes in the dark, after which the optical density readings were determined at $\lambda = 660$ nm.

2.21 Electrophoretic Mobility Gel Shift Assays

Electrophoretic mobility shift assays (EMSAs) were carried out essentially as described by Fried and Crothers (1984) (153) and Garner et al (1981) (189)). In the standard EMSA, 1 ng of end-labelled DNA restriction fragment was incubated with variable amounts of protein for 30 minutes at 4°C in EMSA incubation buffer (16 mM Tris.HCl (pH 8), 175 mM KCl, 1.6 mM $MgCl_2$, 1 mM EDTA, 16 % (v/v) glycerol, 0.8 mM DTT, 0.4 mM PMSF, 0.5 µg p[d(I-C)] (Boehringer) and 1 µg BSA (Molecular Biology Grade, Boehringer) in a total volume of 25 µl.

Nondenaturing 4 % polyacrylamide gels (acrylamide Merck, bisacrylamide Biorad) (22 cm X 18.5 cm X 0.15 cm) were pre-electrophoresed at 30 mA for 2 hours. The electrophoresis buffer was

changed and the EMSA incubation mixtures were loaded directly onto the gels. Electrophoresis was overnight at 30 mA per gel. A buffer system consisting of TGE (50 mM Tris.HCl (pH 8), 380 mM glycine (Merck), 2 mM EDTA) was employed. Gels were dried and exposed to a preflashed X-ray film with an intensifying screen at -70°C.

The binding specificity of protein for the E/H fragment was determined by using double stranded DNA deoxyoligonucleotides (oligos) containing a G·C-rich region and mutations thereof as unlabelled competitors in the mobility gel shift assay. The E/H fragment was used as radiolabelled probe. The mobility gel shift incubations were carried out as above, however the DNA competitors (present in various ratios to unlabelled DNA) were included in the reaction cocktail (see individual experiments, Chapter 3). Gels were dried and analysed by autoradiography or by Instant Imager 2024. The data provided a measure of the relative affinity and specificity of DNA-protein-binding which could be evaluated by Scatchard analysis.

2.22 Synthesis of Poly(dG).Poly(dC)-Affinity Matrix

A trace of poly(dG).poly(dC) (Boehringer) was labelled with [γ-^{32}P]dATP and T_4 Polynucleotide Kinase (Boehringer) after removal of 5' phosphates with Calf Intestinal Phosphatase (151), and mixed with the unlabeled poly(dG).poly(dC). Approximately 1.3 mg of this homopolymer preparation was coupled to approximately 20 ml Sepharose CL-4B (Pharmacia) by the cyanogen bromide method, essentially as described by Kadonaga (1990) (158) and Kadonaga and Tjian (1986) (92).

Sepharose CL-4B (15 ml settled bed volume) was washed with 900 ml H_2O in a 60 ml scintered glass funnel, transferred to a 25 ml cylinder and adjusted to 20 ml with H_2O. The slurry was transferred to a 150 ml beaker in a 15°C waterbath over a magnetic stirrer in a fume cupboard. CNBr (1.1 g, Riedel-de Haen) dissolved in 2 ml N,N-dimethylformamide was added dropwise over 1 minute to the stirring slurry. Sodium hydroxide (30 µl, 5 M) was immediately added, followed by another addition of 30 µl every 10 seconds for 10 minutes, to a final volume of 1.8 ml. Ice-cold H_2O (100 ml) was immediately added and the mixture was poured into a 60 ml scintered glass funnel under suction. Care was taken not to suck the resin to a dry cake. The resin was washed three times with 100 ml ice-cold H_2O and once with 100 ml ice-cold 10 mM potassium phosphate (pH 8), and the thick slurry was immediately transferred to a silanized SS34 tube (Sorvall). The DNA (1.5 ml) was added immediately, and the slurry was rolled at room temperature for 16 hours.

The resin was transferred to a scintered glass funnel and washed with 2 x 100 ml H_2O, 1 X 100 ml 1 M ethanolamine (pH 8), and rolled in 4 ml ethanolamine solution for 6 hours at room temperature to inactivate the unreacted CNBr-activated Sepharose. The resin was finally washed with 100 ml each of 10 mM potassium phosphate (pH 8), 1 M potassium phosphate (pH 8), 1 M KCl, H_2O and column storage buffer (10 mM Tris.HCl (pH 7.5), 1 mM EDTA, 0.3 M NaCl, 0.04 % (w/v) sodium azide. The resin was stored at 4°C in column storage buffer.

The coupling efficiency was estimated to be approximately 90 % by comparing the level of radioactivity in the first few millilitres of the wash (after the overnight coupling step) with that of the washed resin.

2.23 Purification of Native suGF1

Column buffers containing different concentrations of potassium chloride are referred to as "0.X buffer C", where 0.X buffer C is 20 mM Tris.HCl (pH 8.0), 2 mM $MgCl_2$, 0.2 mM EDTA, 20% (v/v) glycerol, 0.5 mM PMSF, 1 mM DTT and containing 0.X M KCl. All chromatographic steps were performed at 4°C and all fractions were stored at -70°C between manipulations.

2.23.1 P11 Phosphocellulose Chromatography

P11 phosphocellulose chromatography was based on a method described by Dailey et al (1988) (190).

P11 phosphocellulose (Whatman) was swollen in a large volume of distilled water. The resin was stirred in 5 volumes 0.5 M NaOH for 30 minutes and rinsed with water until the eluate reached pH 8. The resin was stirred in 5 volumes 0.5 M HCl for 30 minutes, followed by a wash with water until the eluate reached pH 4. The resin was resuspended in 2 volumes 0.05 M Tris.HCl (pH 7.9) and stirred for 15 minutes. The suspension was adjusted to pH 7.9 with 6 M KOH. The resin was packed in a column of radius 2.2 cm and a bed volume of approximately 180 ml was ensured (1 column volume). The column was equilibrated overnight with 0.1 Buffer C containing no $MgCl_2$. The flow rate was always 60 ml/hour. Nuclear extract (protein concentration between 5 and 15 mg/ml) was loaded onto the column and washed with 2.4 column volumes of 0.1 Buffer C (containing no $MgCl_2$). The bound protein was eluted stepwise with 3 column volumes 0.3 Buffer C lacking $MgCl_2$, 3 column volumes 0.5 Buffer C and 1 column volume 0.8 Buffer C. The eluate was collected in 15 ml fractions and the

elution was monitored spectrophotometrically at λ = 280 nm. The fractions (1 µl aliquots) were monitored for suGF1 activity by EMSA in 250 mM KCl. The column was regenerated by washing it with 10 column volumes of 2.5 M KCl until the resin was white. When not in use, the column was stored in 50 mM Tris.HCl (pH 8), 100 mM KCl and 0.04 % (w/v) sodium azide.

2.23.2 Poly(dG)·Poly(dC) Affinity Chromatography

The poly(dG)·poly(dC) affinity matrix was packed in a column of radius 4.75 mm and bed volume approximately 9 ml. The flow rate was 0.5 ml/min and 9 ml fractions were collected at 4°C. Column buffers containing different concentrations of potassium chloride are referred to as "0.X buffer C", where 0.X buffer C is 20 mM Tris.HCl (pH 8), 2 mM $MgCl_2$, 0.2 mM EDTA, 20 % (v/v) glycerol, 0.5 mM PMSF, 1 mM DTT and containing 0.X M KCl. All buffers were supplemented with 0.01 % (v/v) NP-40. The column was equilibrated in 0.35 buffer C. P11 column fractions exhibiting suGF1 activity in EMSAs were pooled and adjusted to 0.35 buffer C by addition of 0.0 buffer C (and 0.01 % NP-40), incubated with 400 µl p[d(I-C)] (1 mg/ml) for 10 minutes and loaded onto the column. The flow-through was collected in a single fraction, and the column was washed with 5 column volumes of 0.35 buffer C. Bound proteins were eluted in a stepwise fashion with 8 column volumes of 0.55 buffer C, 5 column volumes of 0.7 buffer C, and 3 column volumes of 1.0 buffer C. Aliquots (5 µl) of each fraction were monitored for suGF1 activity in EMSAs. The fractions were stored at -70°C between manipulations. The column was regenerated at room temperature by washing with 300 ml column regeneration buffer (CRB) (10 mM Tris.HCl (pH 8), 1 mM EDTA, 2.5 M NaCl, 1 % (v/v) NP-40) followed by 300 ml column storage buffer (CSB) (10 mM Tris.HCl (pH 8), 1 mM EDTA, 0.3 M NaCl, 0.04 % (w/v) sodium azide). The matrix was stored at 4°C.

2.23.3 TCA Precipitation of Proteins

Protein solutions were adjusted to a volume of 500 µl with water, and TCA (final concentration 20 % (v/v)) was added. The samples were placed on ice for 60 min and centrifuged for 30 min at 15 000 rpm, 4°C. The supernatant was discarded and the pellet was resuspended in cold 1 ml acidified acetone (0.05 % (v/v) HCl). The samples were centrifuged as above. The protein pellet was washed with 1 ml acetone by spinning at 15 000 rpm and 4°C for 20 min, resuspended in 1 X SDS sample application buffer (see section 2.24) and neutralised with NaOH, if necessary. The samples were boiled in the presence of a reducing agent and analysed by SDS-PAGE.

2.24 SDS Polyacrylamide Gel Electrophoresis and Silver Staining

Samples were boiled in SDS sample application buffer (0.0625 M Tris.HCl (pH 6.8), 2 % (w/v) SDS (Sigma), 10 % (v/v) glycerol, 5 % (v/v) ß-mercaptoethanol, 0.001 % (w/v) bromophenol blue) and loaded onto a 10 % or 12 % SDS gels (acrylamide (Merck) : bisacrylamide (Sigma) = 30 : 0.8). The gels were electrophoresed at constant voltage (180 V) until the ion front reached the end of the gel. The gels were fixed in 50 % (v/v) methanol containing 10 % (v/v) acetic acid for a minimum of 45 minutes. They were either stained in Coomassie Brilliant Blue or silverstained. Silver staining was performed in clean glass dishes, and all solutions were made up freshly. The gels were washed with deionized water and soaked in 100 ml DTT (5 µg/ml) for 30 minutes followed by a further 30 minutes in 0.1 % (w/v) silver nitrate. The gel was briefly rinsed in a small amount of deionized water before washing it twice with carbonate developing solution (3 % (w/v) Na_2CO_3 containing 0.5 ml 37 % (v/v) formaldehyde per litre of developing solution). The gel was covered with 100 ml developing solution and agitated slowly until it reached the desired level of staining. The reaction was stopped by adding 5 ml 2.3 M citric acid to the gel in developer solution.

2.25 Mass Spectral Protein Sequencing

2.25.1 In Gel Digestion

The Coomassie-stained protein band representing suGF1 was precisely excised from the acrylamide gel (see 2.24), cut into small cubes and rinsed twice with 100 µl water for 5 - 10 minutes to remove the SDS and acid. The pH was ensured to be 6 - 7. The same volume of acetonitrile / water (1:1) was added to remove the Coomassie dye and incubated for 10 - 20 minutes. This step was repeated three times. The residual water was extracted from the gel with pure acetonitrile by incubating for 10 minutes. The acetonitrile was removed and replaced with 30 - 50 µl digestion buffer (NH_4CO_3, NMM) containing 0.5 µg trypsin. The digestion was performed at 37°C for 6 - 12 hours. The supernatant was recovered and the gel pieces were extracted twice with 0.1 % TFA (20 - 30 minutes). The volume of the combined extracts was reduced to 5 µl in a speed-vac.

2.25.2 LC-MS Analysis

LC-MS (microcolumn liquid chromatography) and MS/MS (tandem mass spectrometry) spectra were recorded on a Finnigan LCQ ion trap mass spectrometer equipped with an electrospray (ESI) ion source. LC separation of the peptides was performed on a 300 μm X 12 cm C18 column at a flow rate of 4 μl/min. The peptides were eluted by a gradient of acetonitrile (10 - 40 %) in 0.05 % TFA (40 min) and on line introduced into the ESI source. Full scan MS spectra (m/z 350 to m/z 1700) were collected continuously and production of MS/MS spectra (collision induced dissociation (CID) of individual peptides) was triggered by a peptide signal intensity above a preset threshold (3.0E4 ions). The signal switched the instrument to isolate the parent ion, to perform CID fragmentation and to scan the product ions. This sequence of scan events (full scan - MS/MS scan) was repeated every 4 seconds.

2.25.3 Computer Analysis of MS/MS Spectra

Cross correlation analysis of the data was performed by using the "Sequest" program (166) package and the OWL database (147 000 entries).

2.26 Autoradiography

Sequencing gels, SDS gels and nondenaturing PAGE gels (gel mobility shift assays) were vacuum dried onto Whatman 3MM blotting paper before they were exposed to X-ray films. Either Cronex 4 X-ray film or MP™-Film were used in X-ray cassettes. For gels containing the isotope ^{32}P, the films were preflashed twice on either side and exposed to the gel in the presence of an intensifying screen at -70°C. The intensifying screens were not necessary if ^{35}S was used and exposure was performed at room temperature. Exposure times ranged from overnight to one week.

Gels were either analysed by autoradiography or Instant Imager 2024. The Instant Imager 2024 is an electronic alternative to autoradiography on film or phosphor screens. It is a fully automated system which quantifies radioactivity distributed on flat samples using the microchannel array detector.

CHAPTER 3

Cloning the cDNA for suGF1

3.1 Introduction

Since at the start of the project there was no sequence information available on either the suGF1 protein or the cDNA coding for it, the DNA ligand screening procedure was an obvious choice in the cloning strategy of this protein. This choice was supported by the fact that suGF1 not only binds to the G·C-rich H1-H4 intergenic spacer with high affinity and specificity, but also by the ease with which native suGF1 can be denatured and subsequently renatured to regain its specific DNA-binding ability when subjected to Southwestern analysis (1, 191), indicating that suGF1 may be an ideal candidate for cloning by the DNA ligand screening approach. suGF1 is proposed to be one of several sea urchin G-binding proteins (such as SpGCF1 or the ectoderm specific factor) involved in the regulation of unrelated sea urchin genes, implying that it could be a member of a family of G·C-rich DNA-binding proteins (see section 1.4.4). Given the similarities between suGF1 and SpGCF1 with respect to their size and DNA-binding specificity (see sections 1.4.4.1 and 1.4.4.2, respectively), it was of interest to investigate whether these two proteins are possibly homologous. Therefore a second cloning approach, viz a PCR screening strategy (see section 1.6.2), was used at a later stage in order to clone a cDNA representing the *P.angulosus* homologue of SpGCF1 (see section 1.4.4.2).

3.2 DNA Binding Properties of suGF1

In order to establish the conditions for isolating the cDNA of suGF1 using the DNA ligand screening approach, the affinity and specificity of the interaction between suGF1 and the G·C-rich E/H fragment (see fig 2.2) derived from the H1-H4 intergenic region was investigated using mobility gel shift assays, which result in the formation of two characteristic suGF1-DNA complexes, B1 and B2 (fig 3.1). EMSAs were performed using suGF1 in crude nuclear extract (isolated from 14 hour *P.angulosus* embryos) and the labelled E/H fragment prepared from the histone gene battery (20) in the presence of various amounts of unlabelled oligonucleotide and DNA competitors (fig 3.1). The oligonucleotides contained either the wild-type histone H1-H4 intergenic G_{11}-string (specific oligo)

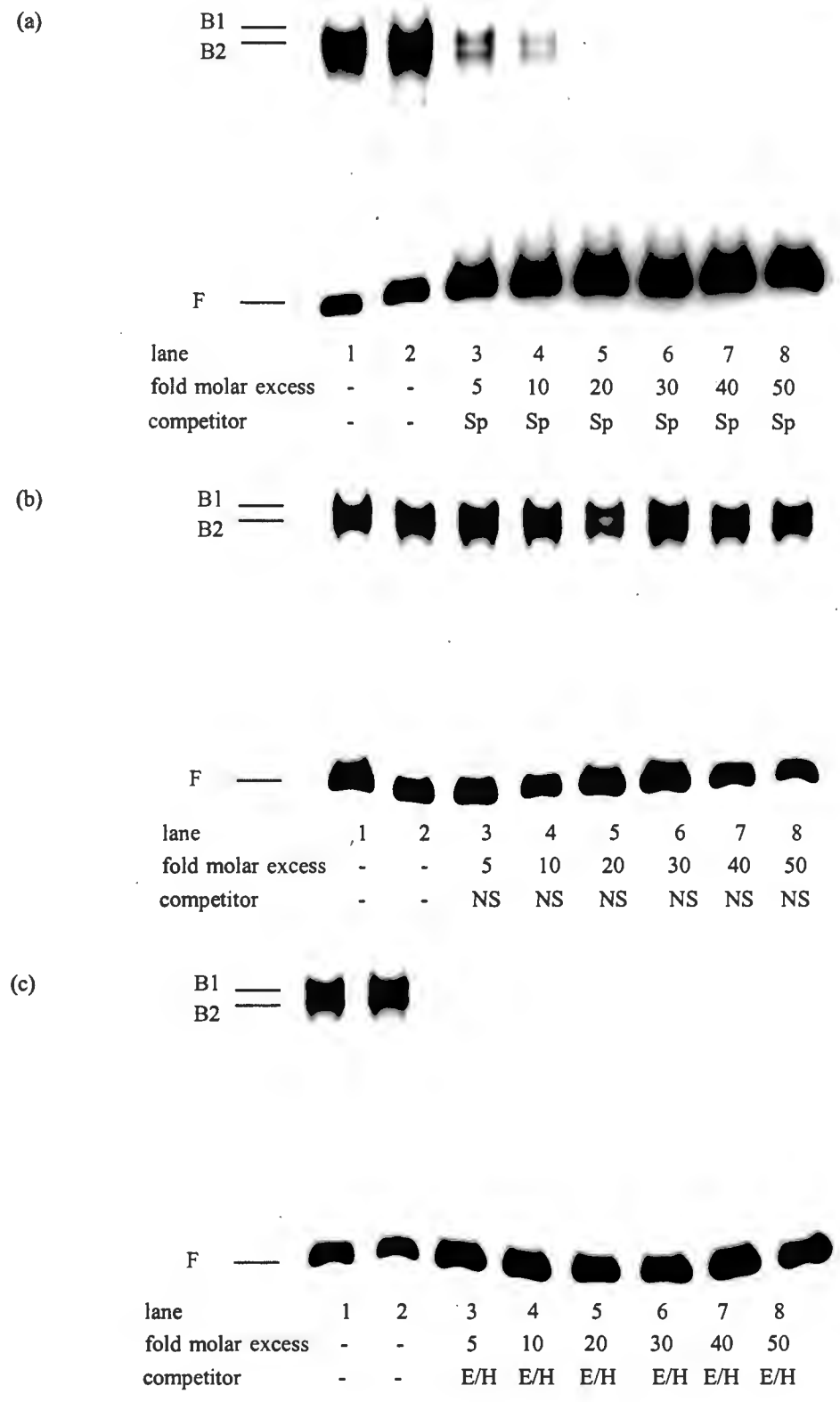


Fig 3.1 suGF1 Interacts Sequence-Specifically with G,C-Rich DNA

Electrophoretic mobility shift assays were performed with suGF1 (2 μg nuclear extract) in the presence of oligonucleotide or DNA competitors. The formation of the suGF1 complexes (B1 and B2) is competed away by the specific (Sp) oligonucleotide (a) and the E/H fragment (c), but not by the nonspecific (NS) oligonucleotide (b). The competition assays (see (a), (b) and (c)) were performed both in the absence of competitors (lanes 1 and 2) and using increasing amounts (5 - 10 fold molar excess) of each competitor (lanes 3 - 8). F is free labelled DNA probe, B1 and B2 are suGF1- DNA complexes.

(fig 3.1 (a)), or a random mutation of the binding site (nonspecific oligo) (fig 3.1 (b)). The nucleotide sequences of the oligonucleotides and the E/H fragment are shown in fig 2.1 and fig 2.2 respectively. The oligos were used in a 5- to 50- fold molar excess in the competition gel shift assays. EMSAs performed with unlabelled specific oligonucleotide competitor (fig 3.1 (a), lanes 2 - 8) show that formation of complexes B1 and B2 is competed away efficiently, even at low molar ratios of competitor with respect to labelled probe (5 - 10 M excess, fig 3,1 (a), lanes 3 and 4). The extent of the competition compares favourably with a similar assay using unlabelled E/H fragment as DNA competitor (fig 3.1 (c), lanes 2 - 8). Binding is competed for at very low molar excess of unlabelled to labelled E/H fragment (5 M excess, fig 3.1 (c), lane 3). In contrast, the same amounts of unlabelled nonspecific oligo do not compete for complex formation at all (fig 3.1 (b), lanes 2 - 8), showing that suGF1 interacts in a highly specific manner with the G·C-rich region of the specific oligo.

Above findings were reinforced by a quantitative competition gel shift assay performed with the specific and nonspecific competitors in a 2 - 10 fold molar excess with respect to the labelled E/H probe (fig 3.2). A constant amount of suGF1 (2 μg nuclear extract) was incubated with 1 ng E/H fragment, in the presence of increasing amounts of competitor oligonucleotides. The amount of free DNA and the protein-DNA complexes formed at each different concentration of competitor were quantified by Instant Imager 2024 (table 3.1). The specific oligonucleotide exhibits more than 50 % competition at a 5 - 10 fold molar excess (fig 3.2, lanes 5 - 7 and table 3.1), whereas the nonspecific competitor (fig 3.2, lanes 8 - 13 and table 3.1) does not compete for the formation of complexes B1 and B2 at all. These results show that suGF1 binds with very high specificity to G·C-rich DNA, since binding can be competed away at very low molar excess of G·C-rich competitor, whereas suGF1 has no affinity for random DNA sequences.

The kinetic and equilibrium constants for protein-DNA interactions in solution imply that only proteins with relatively high binding constants will complex the DNA long enough to withstand the wash protocols. Therefore the kinetics of the suGF1-DNA interaction was investigated using Scatchard analysis. The end-labelled E/H probe (1 ng) was incubated with a constant amount of 14 hour nuclear extracts (2 μg) in the presence of increasing amounts of unlabelled E/H fragment as competitor (fig 3.3). The competitor DNA was present in a 1- to 35- fold molar excess relative to the radiolabelled E/H fragment. The amount of radioactive DNA present in complexes B1 and B2 in each lane was quantitated using an Instant Imager 2024 (table 3.2). A Scatchard analysis of this data is shown in Appendix I. The dissociation constant (K_d) was determined to be 3.6×10^{-10}M, confirming that suGF1 (in crude nuclear extracts) binds with very high affinity to the G_{11}-string present in the binding-site.

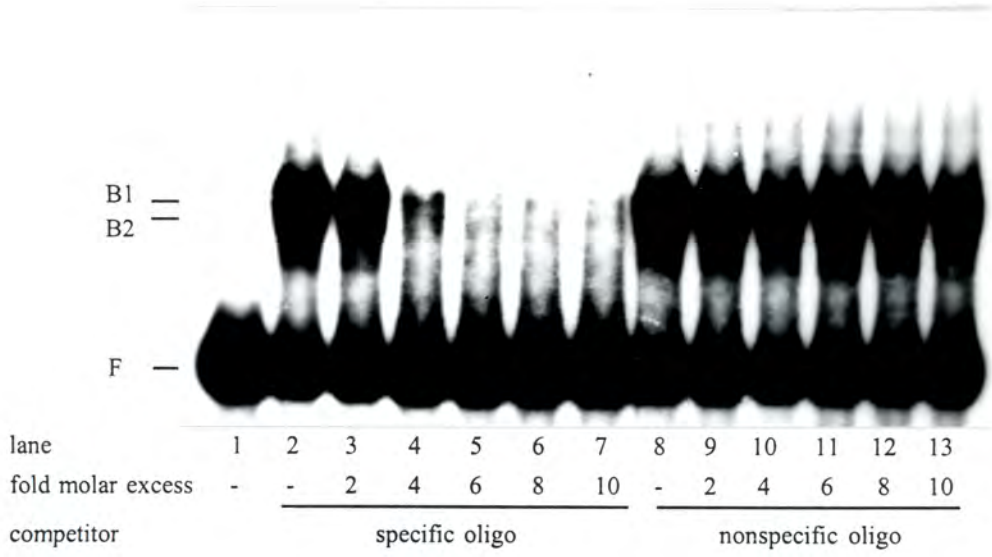


Fig 3.2 Quantitative Competition EMSA Using Specific and Nonspecific Oligonucleotide Competitors

The amount of suGF1 binding in each lane in the presence of increasing amounts (2 - 10 fold molar excess) of specific and nonspecific oligonucleotide competitors (lanes 3 - 7 and 8 - 13 respectively), was quantified by Instant Imager 2024 (see table 3.1). Both the specific and nonspecific competitors were unlabelled. F is free labelled probe, B1 and B2 are suGF1-DNA complexes.

free labelled E/H fragment (cpm)	protein-DNA complexes (cpm)	unlabelled DNA competitor (fold Molar excess)
		(Specific oligo competitor)
69.761	1.268	0
33.179	24.134	0
39.733	16.721	2
58.712	5.996	4
62.112	4.282	6
63.437	3.727	8
58.734	3.735	10
		(Nonspecific oligo competitor)
42.403	23.446	0
37.001	24.569	2
27.443	17.674	4
37.784	24.341	6
40.582	22.017	8
35.899	14.943	10

Table 3.1 Quantitation of the Amount of Labelled DNA in the Unbound and Protein-Bound Fractions in the Presence of Different Concentrations of Unlabelled Specific and Nonspecific Oligonucleotide Competitors

Gel shift reactions were performed in the presence of increasing amounts of both specific and nonspecific competitor oligonucleotides (2 - 10 fold molar excess) and the amount of labelled probe in the unbound and protein-bound fractions (see fig 3.2) was quantified by Instant Imager 2024.

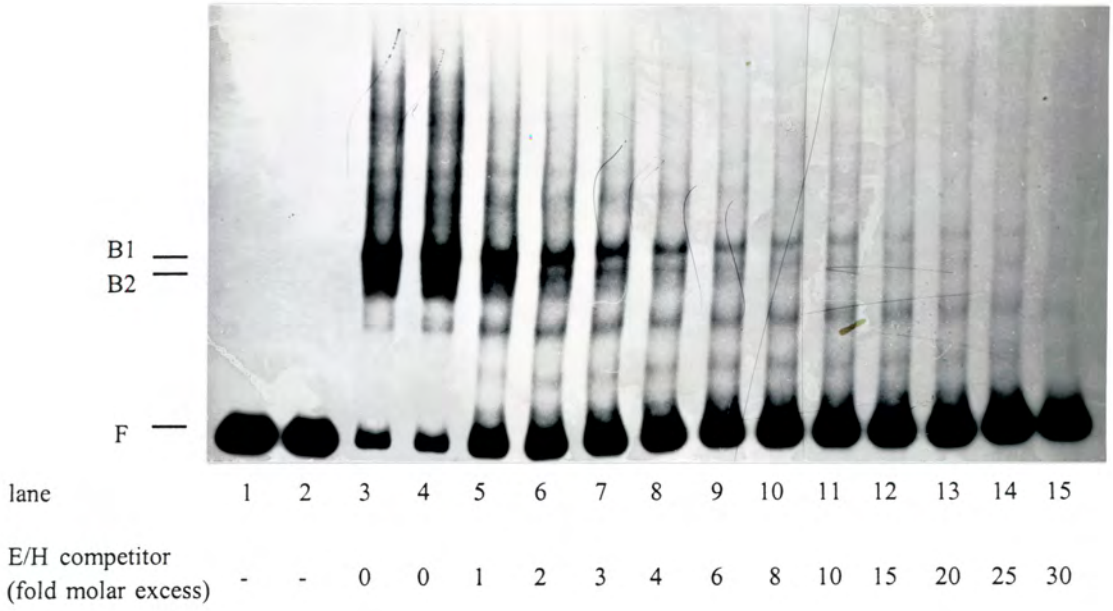


Fig 3.3 <u>The Dissociation Constant of suGF1 with Respect to the G-String in the H1-H4 Intergenic Region was Determined by Quantitative Competition EMSA</u>

Labelled E/H fragment (1 ng) was incubated in the absence of protein (lanes 1 and 2) and with a constant amount of suGF1 (2 μg nuclear extract) in the presence of increasing amounts (1 - 35 fold molar excess) of unlabelled G-string competitor (lanes 3 - 15). The suGF1-bound DNA was separated from the free DNA by electrophoresis, and both were quantified by Instant Imager 2024, as indicated in table 3.2. B1 and B2 are the suGF1-DNA complexes, F is free labelled DNA.

free labelled E/H fragment)	protein-DNA complexes (B1 and B2)	total counts of labelled E/H fragment	unlabelled E/H fragment		
(cpm	(cpm)	(cpm)	fold Molar excess	[M]	error [M]
360.419	5.420	366	0	0.00E+00	0.00E+00
372.706	5.995	379	0	0.00E+00	0.00E+00
44.536	117.211	162	0	0.00E+00	0.00E+00
40.932	97.977	139	0	0.00E+00	0.00E+00
116.268	71.791	188	1	1.7E-10	4.86E-11
142.372	48.478	191	2	3.4E-10	9.72E-11
168.918	24.226	193	3	5.12E-10	1.46E-10
184.950	28.449	213	4	6.8E-10	1.94E-10
194.173	17.298	211	6	1.02E-9	2.92E-10
216.307	10.869	227	8	1.36E-9	3.89E-10
230.372	12.091	242	10	1.71E-9	4.86E-10
243.155	6.194	249	15	2.56E-9	7.29E-10
260.504	5.625	266	20	3.41E-9	9.72E-10
259.504	5.579	265	25	4.26E-9	1.22E-9
201.143	9.464	211	35	5.97E-9	1.70E-9
		234 1.7E-10[a]			
		67 0.5E-10[b]			

a) In a 25 µl reaction volume, 1 ng of the 335 bp E/H fragment represents 234 cpm and the reaction is 1.7 X 10^{-10} M in the labelled DNA fragment.
b) The error for the total counts is 67 cpm, and therefore the error for the molarity of the E/H fragment in the reaction is 0.5 X 10^{-10} M.

Table 3.2 Quantitation of the Amount of Labelled DNA in the Unbound and Protein-Bound Fractions in the Presence of Increasing Amounts of Unlabelled E/H Competitor DNA

The amount of labelled DNA present in the unbound and the suGF1-complexed fractions (in the presence of increasing amounts of unlabelled E/H fragment as shown in fig 3.3) was quantified by Instant Imager 2024, in order to determine the dissociation constant for suGF1 via Scatchard analysis (see Appendix I).

Formation of the suGF1-DNA complex was investigated in the presence of various amounts and types of nonspecific competitor DNA, viz calf thymus DNA, *E.coli* DNA and poly[d(I-C)] (fig 3.4). Even at low concentrations (0.5 - 2 μg) both *E.coli* DNA (lanes 1 - 3) and calf thymus DNA (lanes 4 - 6) compete substantially for the formation of suGF1-DNA complexes B1 and B2. In contrast, higher amounts of poly[d(I-C)] can be included in the incubation without competing for the specific DNA complexes (fig 3.4, lanes 7 and 9). Generally, calf thymus DNA has been used to isolate several clones encoding DNA-binding proteins successfully, and it is the preferred nonspecific competitor DNA used in the DNA ligand screening procedure (174). However, given the fact that it competes so easily for the formation of complexes B1 and B2 in the suGF1-DNA interaction, the preferred nonspecific competitor DNA used in combination with the suGF1 protein is poly[d(I-C)], since it has a low capacity for competing for the formation of specific complexes B1 and B2, and it prevents the formation of nonspecific protein-DNA complexes (see fig 3.4, lanes 7 - 9).

3.3 DNA Ligand Screening a Sea Urchin Embryonic cDNA Library

A cDNA expression library is a crucial starting material for the DNA ligand screening procedure. Attempts to have a *Parechinus angulosus* cDNA expression library produced commercially from 14 hour poly-A^+ or total RNA (see section 3.5) were unsuccessful (Clontech), since the recombinant inserts repeatedly contained small inserts (< 2 kb) and only about 300 000 independent clones, which is a low yield and does not comprise a representative library. Therefore an alternative source (viz a cDNA library derived from a different species of sea urchin) was used. Professor E. Davidson (Caltech) kindly provided a custom synthesised λZAP cDNA library from 24 hour (late blastula / early gastrula) *Strongylocentrotus purpuratus* embryos for which 1.3 X 10^6 independent clones were obtained (see Appendix II). This library was constructed using both oligo(dT)- and random-primed synthesis. The inserts were cloned into the *EcoR1* site of the λZAPII vector and their sizes ranged from 1.0 to 4 - 5 kb.

An outline of the steps involved to identify suGF1 cDNA clones from the recombinant expression library using the DNA ligand screening strategy is depicted in fig 3.5. The first stage involved the identification of recombinant clones which detect the binding site DNA probe. These initial positive clones were subjected to a second round of screening in which the hybrid proteins were blotted onto duplicate filters and individual filters were probed with either the specific binding-site DNA probe, or with a DNA probe that lacks the binding site. The second screening procedure selectively identified clones which specifically bind to the given recognition site.

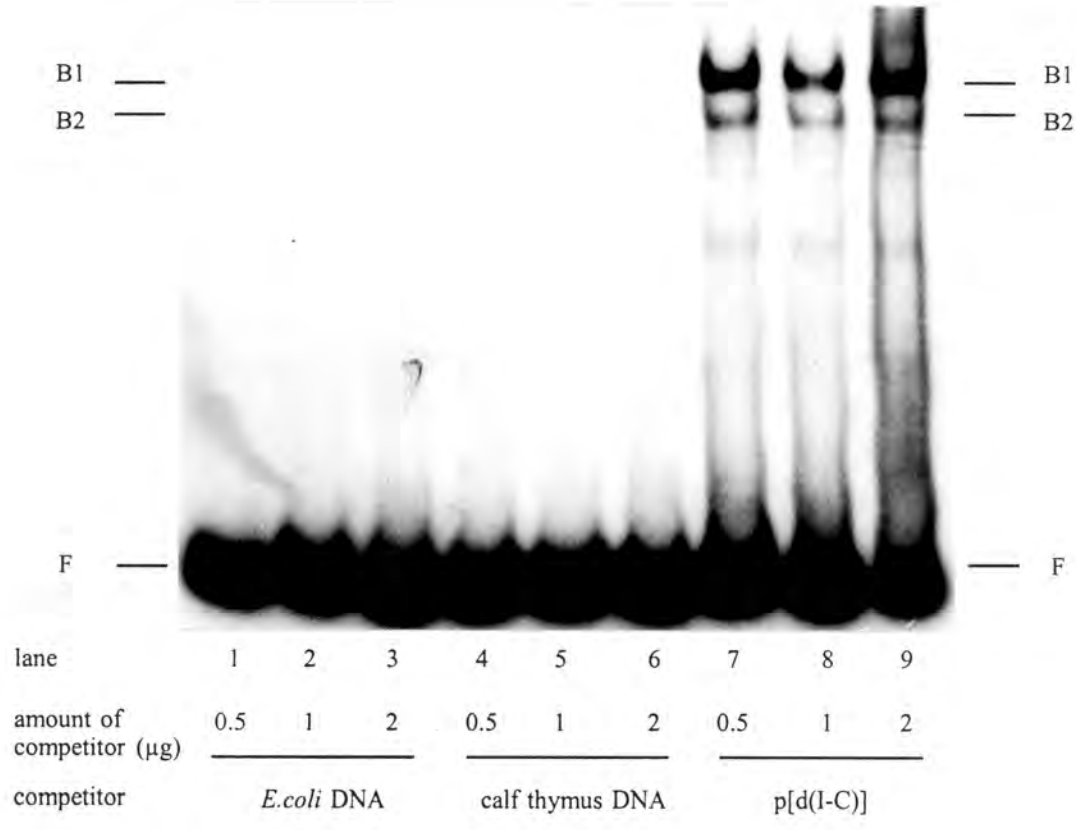


Fig 3.4 suGF1 Has Very Low Affinity for p[d(I-C)]

The suitability of several nonspecific DNA competitors was investigated by incubating suGF1 in nuclear extract with labelled E/H fragment in the presence of varying amounts of *E.coli* DNA (lanes 1 - 3), sonicated calf thymus DNA (lanes 4 - 6) and poly[d(I-C)] (lanes 7 - 9). The amounts of the competitors are indicated in the individual lanes. F is free labelled DNA, B1 and B2 are suGF1-DNA complexes.

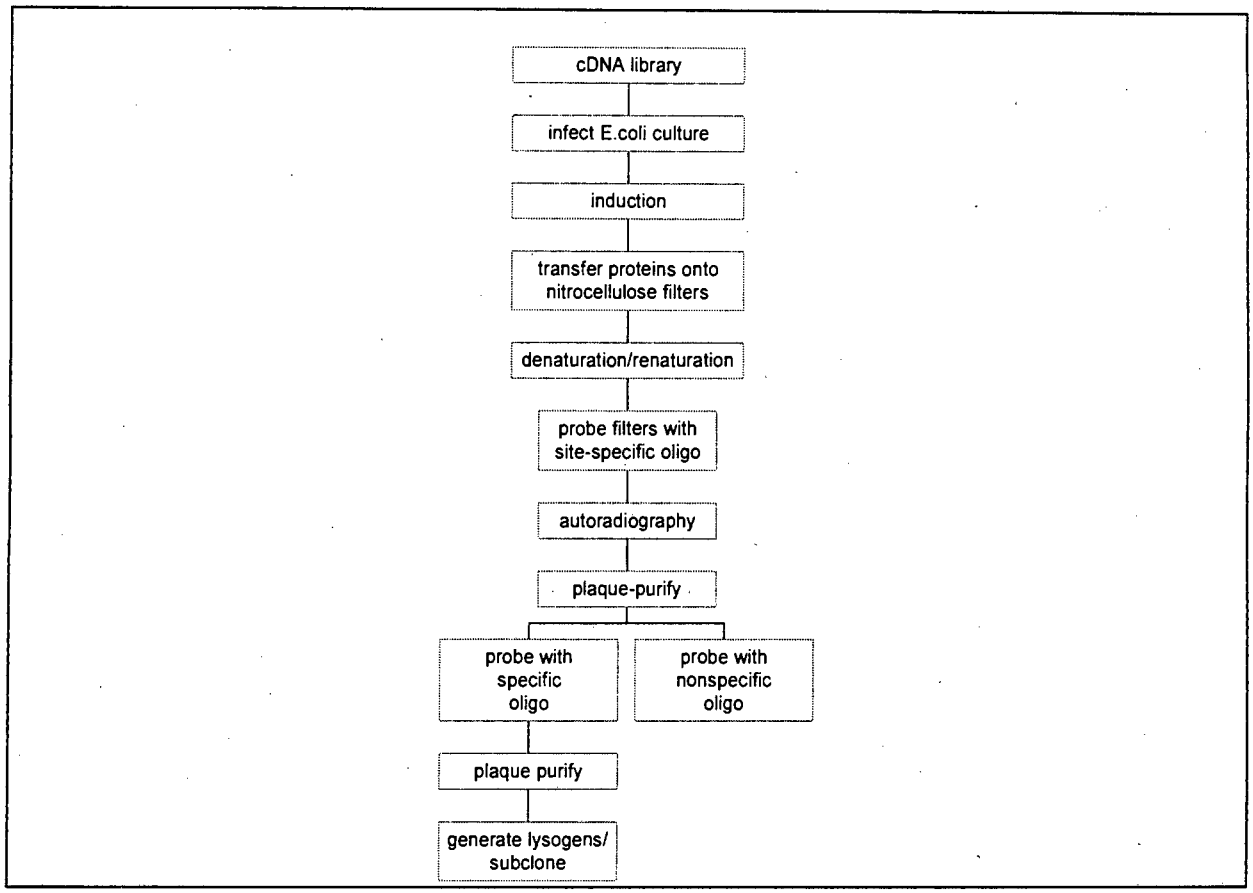


Fig 3.5 An Outline of the Steps Involved in the DNA Ligand Screening Strategy

The DNA ligand screening strategy allows cloning of sequence-specific DNA-binding proteins from a bacteriophage cDNA expression library. The specific oligo represents the recognition site probe, whereas the non-specific oligo is a control probe which lacks the recognition site, or contains a mutant version thereof.

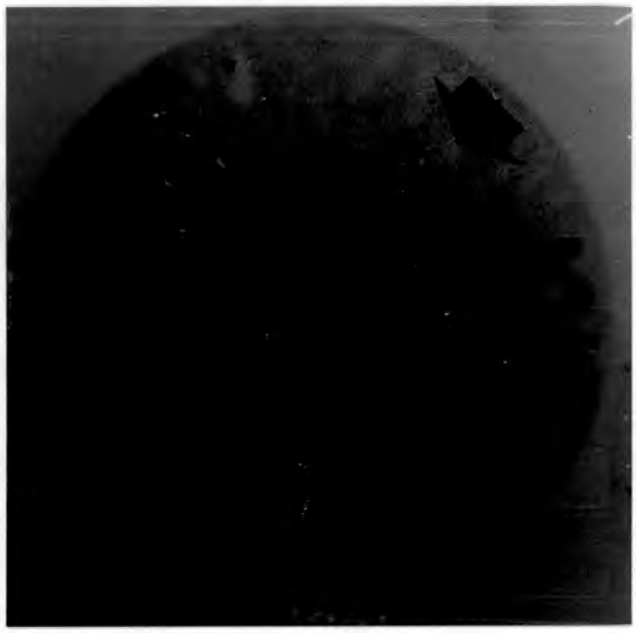
In order to improve the efficiency of detection of prospective cDNA clones, both the specific and nonspecific oligonucleotide probes (see fig 2.1) were catenated extensively using DNA ligase (151). The mean length of the catenated oligonucleotides was ~ 200 bp. Both the catenated DNA probes were labelled to a specific activity of ~ 10^8 dpm/μg with [α–^{32}P]dCTP. Using a DNA probe with above specific activity, and assuming a 1:1 stoichiometry for the protein-DNA complex, it should be possible to detect 10^{-2} fmol of protein in a plaque, considering that the level of expression of the *lacZ* fusion gene in a single phage plaque exceeds these amounts (174). A standard *E.coli* host strain (BB4) was used throughout the screening procedure to produce a fusion protein between the products of the β–galactosidase gene and the cloned cDNAs. The amplified library was plated at a density of 1 X 10^3 pfu/plate and the plates were overlaid with nitrocellulose filters saturated with IPTG to induce the expression of the fusion proteins. A total of 6 X 10^5 plaques was screened. The filters were subjected to a denaturation / renaturation protocol using 6 M guanidine hydrochloride, which facilitates the

correct folding of a larger fraction of the bacterially expressed fusion proteins (168, 174). The filters were subsequently incubated with radiolabelled recognition site probe in the presence of poly[d(I-C]) to prevent nonspecific binding.

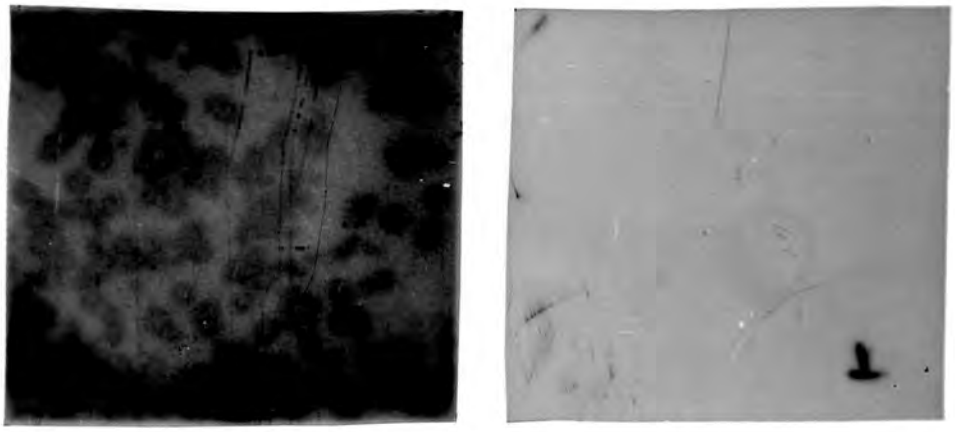
The screening procedure was performed as outlined by Singh et al (1989) (174) (see section 2.8 and fig 3.5). Typical signals generated by the DNA-binding site probe in the first round of screening are illustrated in fig 3.6 (a). The recombinant hybrid proteins were blotted onto duplicate nitrocellulose filters, and the resulting autoradiographs were superimposed in order to identify presumptive positive DNA-binding signals. The first round of screening resulted in the identification of 19 putative positive plaques, from screening a total of 6 X 10^5 plaques. The autoradiographs were aligned with the original LB plates to identify the presumptive positive plaques, which were then plaque purified according to standard procedures (169). The secondary phage stocks generated in this way were used to identify sequence specific clones by blotting the expressed fusion proteins onto replica filters. One filter was screened with the recognition site probe (specific oligo) whereas the other filter was screened with the DNA probe lacking the binding site (nonspecific oligo) (see fig 3.5). A typical result of a sequence specific clone identified by the second round of screening is shown in fig 3.6 (b). Binding is specific with respect to the recognition site probe, whereas there is no interaction with the DNA probe lacking the G·C-rich binding site, thereby satisfying the criteria depicted in fig 3.5. In this manner the number of putative positive clones was finally reduced to four (referred to as Clones 2, 6, 11 and 16).

λZAP recombinants, unlike those derived from λGT11, do not easily form lysogens, therefore the cloned proteins are not usually isolated in the form of extracts derived from induced lysogens of *E.coli* cells. Instead it is recommended that the pBluescript phagemids (containing the cDNAs of interest) are excised from λZAP and that the cDNA inserts are cloned into an expression vector for further analysis of the protein products. The four positive plaques generated by the DNA ligand screening procedure were plaque purified, the Bluescript plasmid (containing the cDNA insert) was excised from λZAP and analysed by restriction enzyme digestion with *EcoR1*. The individual DNA fragments were separated by electrophoresis on an agarose gel in order to determine the size of each insert by comparing it to a lambda *EcoR1 / HindIII* digest (fig 3.7). The insert sizes are estimated to be 2.4 kb (Clone 2, lane 5), 0.85 kb (Clone 6, lane 7), 0.9 kb (Clone 11, lane 9) and 2.2 kb (Clone 16, lane 11).

(a)



(b)



SPECIFIC PROBE NONSPECIFIC PROBE

Fig 3.6 DNA Ligand Screening the cDNA Library Yielded Positive Signals from Several Clones

(a) The 24 hour *S.purpuratus* cDNA library was plated at a density of 1 x 10^3 pfu / plate in the first round of the DNA ligand screening process. The recombinant hybrid proteins were blotted onto duplicate nitrocellulose filters and screened with radiolabelled specific oligonucleotide. The duplicate blots were exposed to autoradiography and presumptive positive plaques were identified by superimposing duplicate autoradiographs. An example of a putative positive plaques is marked by the arrow.

(b) Phage stocks generated from the putative positive plaques identified in the first round of screening (a) were screened for sequence specific clones by blotting the fusion proteins onto replica filters and screening one filter with the recognition site probe (specific oligo) and the other with the DNA probe lacking the binding site (nonspecific oligo). Several clones (four) revealed sequence specific binding to the recognition site probe, whereas they did not interact with the DNA probe lacking the G.C-rich binding site thereby satisfying the criteria outlined in fig 3.5.

3.4 DNA Sequence Analysis of the Putative Positive Clones Generated by the DNA Ligand Screening Procedure

The 5' and 3' termini of each clone (see section 3.3) were subjected to several rounds (between 4 and 7) of manual sequencing (occasionally automated DNA sequencing was used), in order to obtain DNA sequence information (see Appendices III - VI). Analyses of the partial DNA sequences of the four clones and subsequent homology comparisons using algorithms such as Bestfit (from the GCG program) revealed that each clone represents an independent isolate of a different cDNA. Additional homology searches were performed with programs such as FastA (GCG) (this compares the query sequence to nucleotide sequences in the GenEMBL database), BLASTN and BLASTP (192), as well as BLASTX (193) which perform comparisons on the public sequence databases. The BLASTN program is optimised to find nearly identical nucleotide sequences, whereas BLASTX is used for database similarity searches of protein coding regions. The query sequence for each search was filtered. This process eliminates low complexity regions and thereby avoids nonspecific pairwise alignment. A summary of the analysis of each of the clones is presented in table 3.3.

cDNA	**Clone 2**	**Clone 6**	**Clone 11**	**Clone 16**
length (bp)	2 200	850	900	2 400
5' nt sequence	236 nt	655 nt	519 nt	622 nt
3' nt sequence	223 nt	159 nt	301 nt	156 nt
total length sequenced	459 nt	764 nt	820 nt	778 nt
5' ORF	yes	3 partial ORFs	yes	3 partial ORFs
3' ORF	partial ORF	yes	yes	yes
database search result	3' untranslated region of *S.purpuratus* CyIIb actin gene	Krüppel-like Zinc-finger protein in *Caenorhabditis elegans*	mRNA of a G-box binding factor (GBF) in *Dictyostelium discoideum*	basic-helix-loop-helix-leucine zipper transcription factor in *Caenorhabditis elegans*
% homology	96 %	63 %	56 %	64 %
length of overlap	205 nt	96 nt	166 nt	14 amino acids
reference	Lee et al (1984) (194)	Smith et al (1994) (195)	Schnitzler et al (1994) (196)	Wilson et al (1994) (197)

Table 3.3 Summary of the Analysis of Clones 2, 6, 11 and 16 Isolated by the DNA Ligand Screening Method

The table lists the insert size of each clone, the length of its known DNA sequence, an indication whether either cDNA end has an open reading frame and the most relevant database homology score for each clone.

Clone 2 has a 2.2 kb cDNA insert (fig 3.7, lane 5) of which 459 nt were sequenced. Partial DNA sequences of the 5' and 3' termini (236 and 223 nt respectively) are shown in Appendix III (a) and (b) respectively. The partial DNA sequences were analysed for putative open reading frames (ORFs)

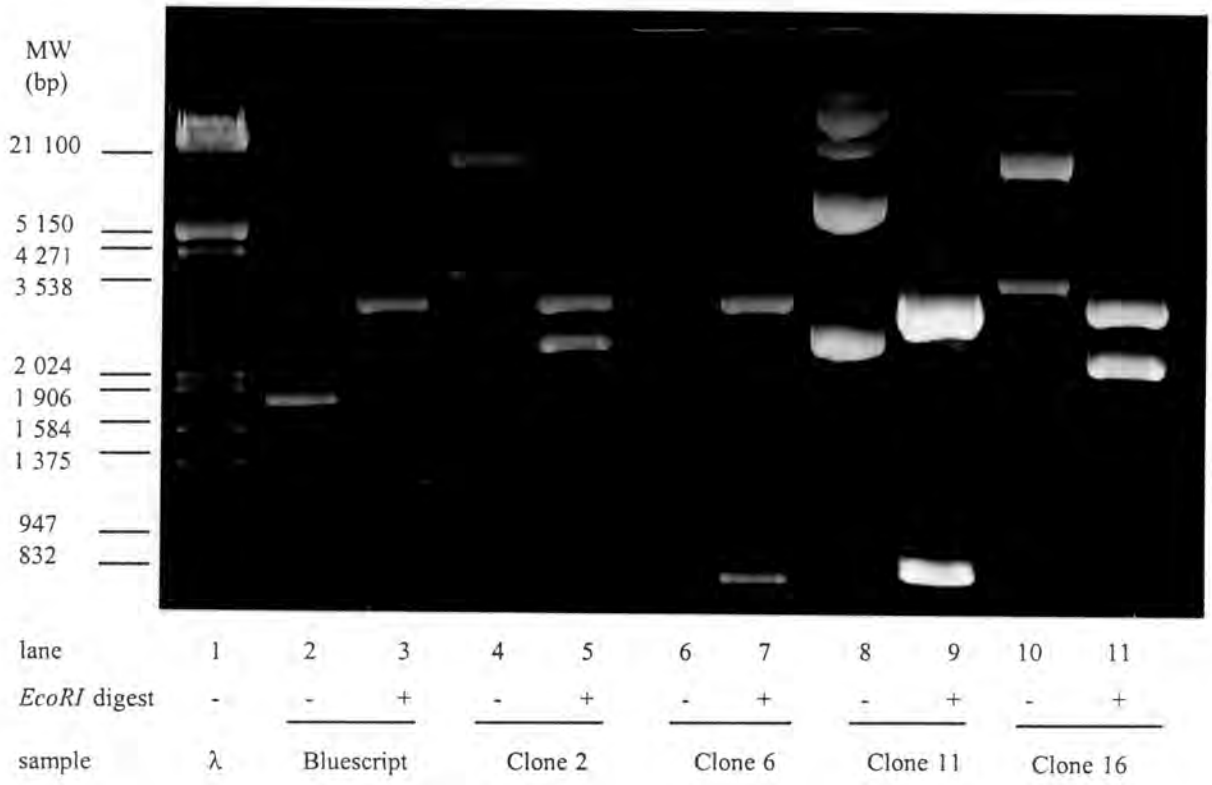


Fig 3.7 Restriction Analysis of the Clones Isolated by the DNA Ligand Screening Method

cDNA inserts isolated by the DNA ligand screening method (section 2.8) were released from the Bluescript plasmid using *EcoR1*. Supercoiled and *EcoR1* digested Bluescript (lanes 2 and 3 respectively), and supercoiled Clones 2, 6, 11 and 16 (lanes 4, 6, 8 and 10) and their respective *EcoRI* restriction digests (lanes 5, 7, 9 and 11) were separated by electrophoresis on an 1 % agarose gel. The size of each insert was estimated by comparison to a lambda *EcoR1 / HindIII* digest (lane 1). The insert sizes are estimated to be 2.4 kb (Clone 2, lane 5), 0.85 kb (Clone 6, lane 7), 0.9 kb (Clone 11, lane 9) and 2.2 kb (Clone 16, lane 11).

using the 'MAP' algorithm in the GCG program, which revealed three putative ORFs for the 5' terminal sequence (see Appendix III (c)), one of these (reading frame "b") has a putative start codon at nt 59, the other two reading frames (beginning at nt 85 and nt 146) do not start with methione residues. The ORFs code for 59, 50 and 27 amino acids respectively, and they continue to the end of the known sequence for the 5' terminus, however their full lengths were not established since the cDNA was only sequenced partially. The 3' terminus of Clone 2 (see Appendix III (d)) has a single ORF of significant length (nt 4 - 174), however it is interrupted (this could be due to erroneous sequencing results, or it could indicate the end of the ORF for this clone). The 5' and 3' DNA sequences were both subjected to FastA searches in the GCG program, as well as BLASTN searches. The 236 nt sequence obtained for the 5' terminus shows a striking and significant homology to the 3' untranslated region of the *Strongylocentrotus purpuratus* CyIIb actin gene (194). The homology extends over 205 nt and has 96 % identity. Several other scores are listed in both the nucleotide and protein homology searches. For instance, the former shows a 73 % homology between the 5' terminus of Clone 2 and a zinc finger protein (CEZF) in *C.elegans* (198). The homology ranges over 49 nt (in the translated region of Clone 2) of which 36 nt are identical. The homology searches performed with the 223 nt 3' terminal sequence of Clone 2 revealed no significant homologies using either nucleotide or protein database searches. One interesting score was a 56 bp overlap of 64 % identity with a posttranslationally regulated *Drosophila* chorion transcription factor, CF2 (199). The homology exists between the 3' untranslated region of the transcription factor and the 3' terminus of Clone 2, however the sequences lie in opposite orientation with respect to the each other.

The insert size of Clone 6 is 850 bp (see fig 3.7, lane 7). Almost the entire clone was sequenced (764 bases in total, see Appendix IV (a) and (b)). The MAP algorithm (GCG) was used to analyse the clone for ORFs, as shown in Appendix IV (c) and (d). The 5' terminus (655 nt) of Clone 6 does not appear to have a single continuous ORF, however three individual ORFs overlap partially. One of these (nt 54 - 237) codes for 61 amino acids, a second one (nt 286 - 478) codes for 64 amino acids, and a third ORF codes for 42 amino acids (nt 392 - 518) which partially overlaps with the second ORF (see Appendix IV (c)). None of the ORFs begin with putative ATG start codons, nor do any of them continue to the end of the known DNA sequence for the 5' terminus. Protein database homology searches revealed no substantial scores for the amino acid sequences corresponding to the 5' region of this clone, whereas the most interesting score for the nucleotide searches showed a 96 nt overlap (within the first ORF described above) of 63 % identity with a Krüppel-like Zinc finger protein in *C.elegans* (195) using FASTA (GCG). The 3' terminal DNA sequence of Clone 6 shows an ORF of 49 amino acids spanning nt 34 - 181, and potentially represents the end of the coding sequence for

this clone. Neither the protein nor nucleotide database homology searches showed scores which were either significant, or relevant with respect to DNA-binding proteins.

Clone 11 contains an insert of 900 bp (see fig 3.7, lane 9). About 820 nt of the entire clone were sequenced (see Appendix V (a) and (b)). The 5' terminal DNA sequence (519 nt) has a putative ATG start codon positioned at nt 72, which continues into a single long ORF to the end of the known DNA sequence of the 5' terminus (Appendix V (c)). The 3' terminal nucleotide sequence (301 nt) was also analysed for ORFs, of which three were identified. These code for 51, 47 and 21 amino acids, they all overlap with each other and continue to the end of the known DNA sequence for the 3' terminus (see Appendix V (d)). It appears that Clone 11 represents a full length cDNA clone, which has a 5' untranslated region (72 nt), a coding region of about 680 nt and a 3' untranslated region (~ 150 nt). Nucleotide and protein database homology searches revealed that the 5' terminal nucleotide sequence scored low homologies to several DNA-binding proteins, for instance 36 amino acids had a 44 % identity with heat shock protein 70 from *Chlamydia trachomatis* (200), whereas 166 nt within the 5' ORF of this clone had a 56 % identity with the mRNA of a G·C-box binding factor (GBF) in *Dictyostelium discoideum* (196). Another score includes the engrailed-like homeodomain protein (smox-2) mRNA from *Schistosoma mansoni* (201), which has an identity of 75 % with 53 nt of the 5' ORF of Clone 11.

A total of 778 nt were sequenced for Clone 16, which has a 2.4 kb insert (fig 3.7, lane 11). The 5' terminal (622 nt) and 3' terminal (156 nt) sequences for this clone are shown in Appendix VI (a) and (b). The 'MAP' algorithm (GCG) was used to establish the ORFs for both the 5' and 3' termini. There is a long 5' untranslated region followed by an ATG codon at nt 285 (see Appendix VI (c)), which may be indicative of a putative start codon. The nucleotide sequence codes for three separate discontinuous ORFs which are interleading. This could imply that there may be one continuous ORF, taking into account that DNA sequencing is subject to several artifacts. Alternatively it is possible that protein translation only starts at nt 501, which represents the beginning of the last ORF, however this does not continue to the end of the known DNA sequence of the 5' terminus. The 3' terminus also has an ORF, which lies between nt 13 - 156 (see Appendix VI (d)), and may therefore be indicative of the end of the coding region for this clone. The nucleotide and protein database homology searches revealed no relevant scores for the 5' terminal sequence, however the 3' terminus has an 18 amino acid overlap of 44 % identity with the human NF-κB TF subunit (202). In addition there is a 64 % identity (over 14 amino acids) with the cDNA of a basic-helix-loop-helix leucine zipper transcription factor in *C.elegans* (197).

The DNA ligand screening technique yielded four independent isolates of different cDNAs based on the ability of their recombinant proteins to preferentially interact with the specific recognition site probe instead of the mutant nonspecific probe (see section 3.3). Partial DNA sequence analysis confirmed that at least three of the clones (2, 11 and 16) potentially encoded recombinant proteins of substantial length, as these clones had ORFs at their 5' ends. Database homology studies indicated that three of the clones (6, 11 and 16) were unique cDNAs which scored homologies to other DNA-binding proteins. The DNA-binding specificities of the recombinant proteins encoded by these clones required further investigation to verify whether any of them correlated with suGF1 (see Chapter 4).

3.5 A PCR Cloning Strategy Was Used to Amplify a cDNA Sequence Potentially Encoding suGF1

A PCR strategy was used to amplify the cDNA sequence encoding the *P.angulosus* homologue for SpGCF1 (a transcription factor present in *S.purpuratus* embryos, see section 1.4.4.2), which represents a candidate homologous protein to suGF1. Genomic DNA prepared from sea urchin sperm (section 2.10), as well as cDNA generated from 14-hour sea urchin embryo RNA (section 2.6.1), were used as templates for the PCR amplifications. The integrity of the RNA was verified on an agarose / formaldehyde gel (section 2.6.4.2) as shown in fig 3.8 (a). Both the 14-hour poly-A^+ RNA (lane 1) and the 14-hour total RNA (lane 2) are undegraded. The poly-A^+ RNA, as expected, has a slight smear associated with it, and the 28S and 18S ribosomal bands in the sample are faint. There is no smear apparent for the total RNA (fig 3.8 (a), lane 2) and the intensities of the 28S and 18S ribosomal RNA bands are present in a ratio of about 1.5 : 1. cDNA was produced by using 2 μg of either poly-A^+ RNA or total RNA. The integrity of the genomic DNA isolated from *P.angulosus* sea urchin sperm (section 2.10) was verified on an agarose gel (fig 3.8 (b), lane 2) and was judged to be intact since it forms a sharp band in the region of 21 kb.

PCR amplification (see section 2.11.1) was performed on both genomic DNA and cDNA from 14 hour *P.angulosus* sea urchin embryos using three different combinations of two degenerate primer-pairs. The design of degenerate PCR primers (see Appendix VII) was based on the SpGCF1 cDNA sequence (3). Degenerate primers 1S and 1A were designed by Zeller et al (1995) (3), whereas degenerate primers 2S and 2A were designed by J. Hapgood (personal communication). Specific primers SP1, SP2, SP3 and SP4 were based on the DNA sequences of PCR products amplified from *P.angulosus* genomic DNA and cDNA. Amplification with primer-pair 1S/1A is predicted to result in a fragment of 106 bp, a combination of the primers 2S/2A should give a 372 bp fragment, and the size of the product generated by the primers 1S/2A is predicted to be 642 bp (see Appendix VII). PCR

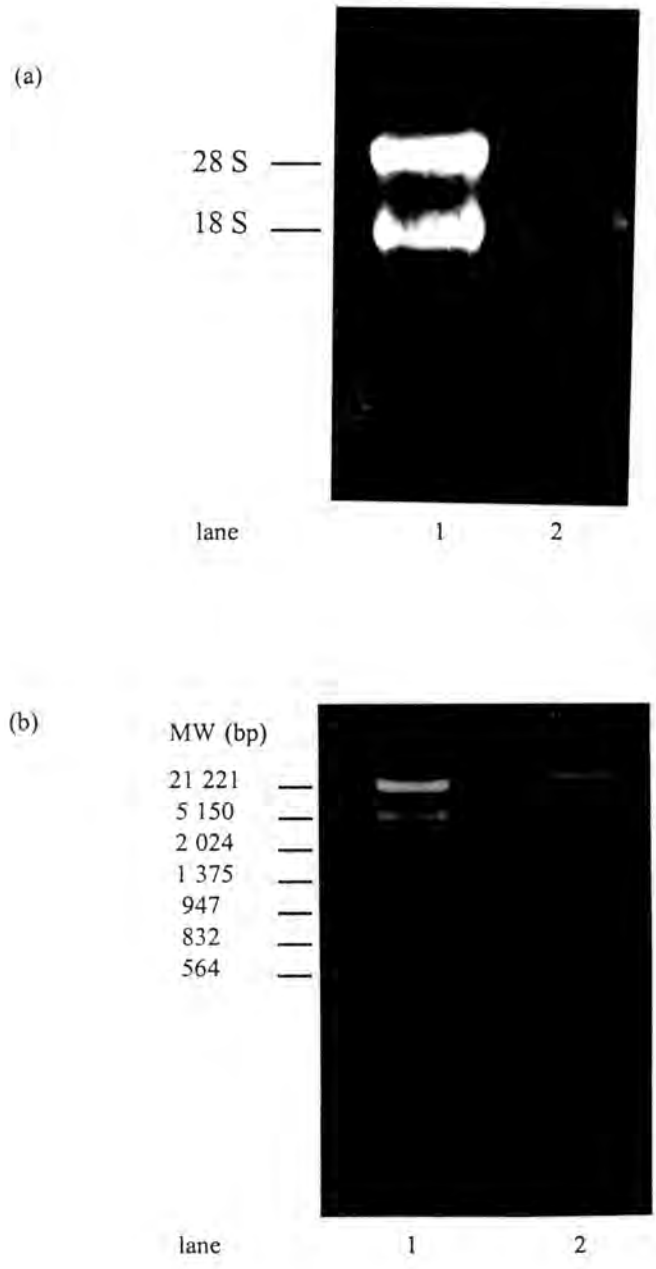


Fig 3.8 Analysis of the Integrity of RNA and Genomic DNA Isolated from Sea Urchin (*P.angulosus*)

RNA isolated from 14 hour sea urchin embryos (*P.angulosus*) (a) and genomic DNA (b) isolated from sea urchin sperm were used as starting material in the PCR reactions. (a) Total RNA (lane 1) and poly-A^+ RNA (lane 2) from 14 hour embryos were analysed (1 µg each) on an 1 % agarose / formaldehyde gel. The 28S and 18S ribosomal RNA bands are marked. (b) Genomic DNA (1 µg) (lane 2) was intact as indicated by its high molecular weight as compared to an *EcoRI / HindIII* digest of lambda (lane 1).

amplifications using genomic sea urchin DNA as template and primer-pair 1S/1A resulted in the successful amplification of a 106 bp DNA fragment (see fig 3.9, lanes 8 and 9), showing firstly that the *P.angulosus* genome contains a gene homologue to SpGCF1, and secondly that there is no intron present between primers 1S and 1A in this gene. The 372 bp fragment using primer pairs 2S/2A could not be amplified from genomic DNA using similar conditions (fig 3.9, lanes 10 - 13), implying that a large intron may be present between primers 2S and 2A. Attempts to optimise PCR reactions using genomic DNA as template and primer pairs 1S/2A and 2S/2A respectively, by varying the concentrations of $MgCl_2$ and DMSO, did not result in products of the predicted size. Amplification of cDNA, generated from 14 hour sea urchin embryo RNA, using the primer pair 1S/2A was unsuccessful, however the primer-pairs 1S/1A and 2S/2A both generated PCR products of the correct size (viz 106 bp and 372 bp), implying that the sea urchin *P.angulosus* expresses a protein homologue to SpGCF1. The 106 bp and 372 bp DNA fragments generated by PCR were gel purified, cloned into the pMOS Blue T-vector (see Appendix VIII (b)), and transformed into host bacteria. Several colonies generated by the ligation / transformation procedure were checked for positive insertion by direct PCR screening of the bacterial cells (section 2.11.2). The cloned PCR products were replicated in bacterial cultures, followed by plasmid isolation (section 2.2.4.2) using Wizard Midipreps. Plasmid inserts were analysed by DNA sequencing (section 2.3) using the T7 and U19 primers from the vector. Several clones from each PCR product were sequenced automatically, all of them revealed the same respective sequences for PCR products generated from the primer combinations 1S/1A and 2S/2A. The DNA sequences of the two individually amplified fragments from *P.angulosus* were compared to the DNA sequence coding for SpGCF1 as published by Zeller et al (1995) (3) using a computer programme (GCG). The comparison of the sequences shows a very high homology (~ 94 % and 92 % respectively) between the two sea urchin species (see Appendix X (a) and (b)).

Gene specific primers SP1 and SP2 (see Appendix VII) were designed from the *P.angulosus* DNA sequences obtained for the 106 bp and 372 bp PCR products respectively (see Appendix X) in order to amplify the full length cDNA by application of 5' and 3' RACE (Rapid Amplification of cDNA Ends, see section 1.6.2). The integrity of the specific primers was verified by PCR amplification of the overlap region between SP1 and SP2 (see Appendix X), which generated a DNA fragment with a predicted size of 421 bp from *P.angulosus* cDNA and genomic DNA (see fig 3.10 (a), lanes 4 -7 and (b), lanes 3 - 7 respectively). Both these PCR products were isolated from a preparative low melting point agarose gel and cloned into the pMOS-T vector (see Appendix VIII). Plasmid DNA was isolated and the 421 bp insert was sequenced from several colonies. All were found to contain the same sequence, which was used in a homology comparison with respect to the SpGCF1 sequence (see Appendix X (c)), showing that the 421 bp fragment has a very high sequence homology (~ 92 %) to

lane	1	2	3	4	5	6	7	8	9	10	11	12	13
DNA (µg)	-	-	-	0.1	0.5	1	1	1	1	0.5	1	1	1
$MgCl_2$ (mM)	-	2	2	1	1	1.5	2	2	2.5	1.5	1.5	1.5	2
DMSO (%)	-	-	-	-	-	-	-	2.5	5	-	-	-	-
primer pair				1S / 1A						2S / 2A			

Fig 3.9 PCR Amplifications from Sea Urchin Genomic DNA Using Degenerate Primers

PCR amplification conditions using genomic DNA (isolated from *P.angulosus* sea urchin sperm) were optimised by varying the $MgCl_2$ and DMSO concentrations (lanes 4 to 13). The 106 bp fragment (see lanes 8 and 9, marked by an arrow) was amplified with primer pair 1S/1A (lanes 4 - 9), whereas the expected 372 bp fragment could not be amplified using the primer pair 2S/2A (lanes 10 - 13). Negative PCR controls without DNA (lanes 2 and 3) were performed with primer pairs 1S/1A and 2S/2A, respectively. A 123 bp ladder (lane 1) was used as molecular weight marker. The different conditions under which the PCR amplifications were performed are listed (see individual lanes), viz amount of DNA (µg), % (v/v) DMSO, and $MgCl_2$ concentration (mM).

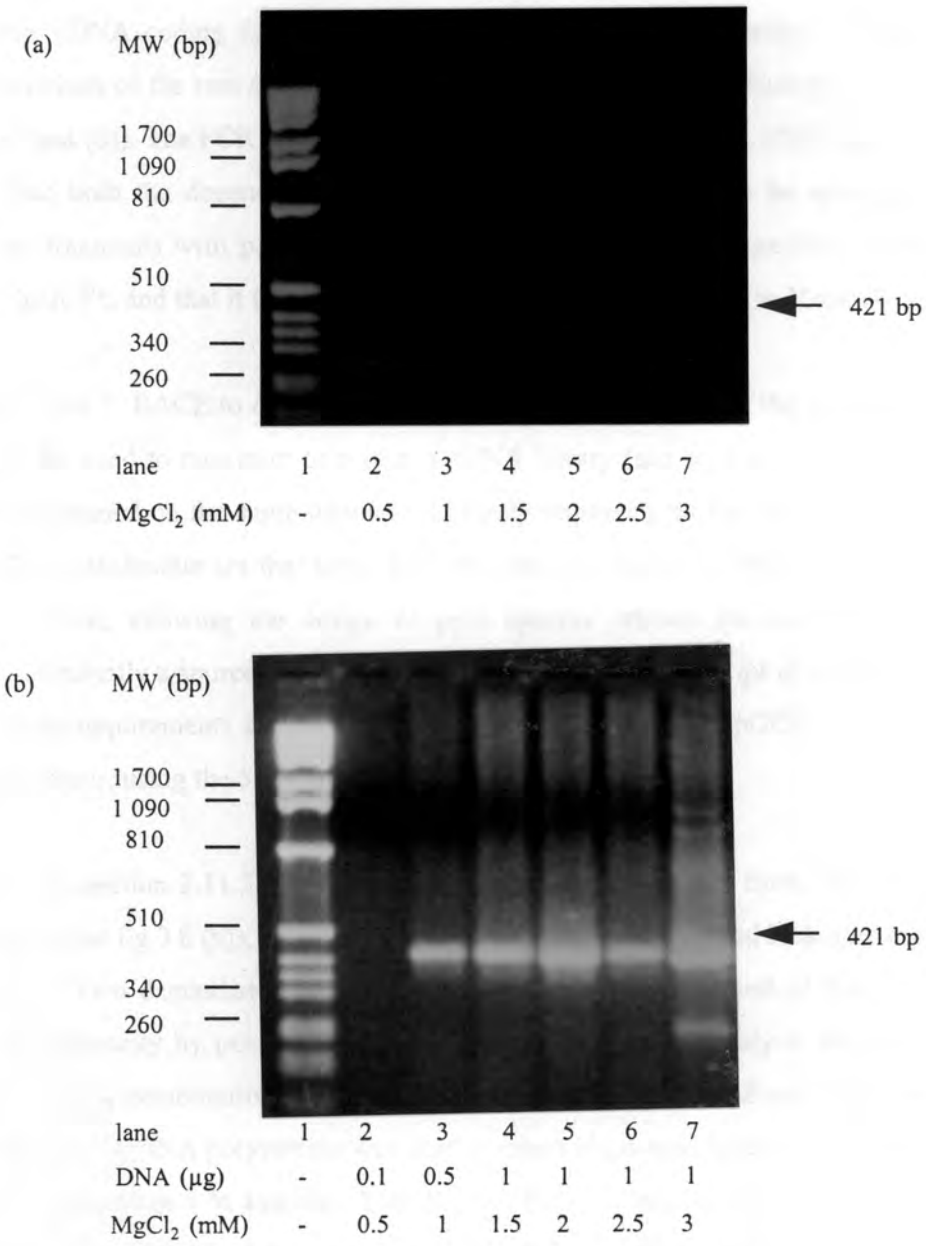


Fig 3.10 Amplification of a 421 bp fragment from *P.angulosus* cDNA and Genomic DNA Using Specific Primers SP1 / SP2

(a) The primer pair SP1/SP2 was used to PCR amplify a 421 bp fragment (lanes 2 - 7) from the cDNA of 14 hour *P.angulosus* embryos. The different $MgCl_2$ concentrations used in each reaction are indicated (see individual lanes). Negative and positive controls (without cDNA (lane 2) and with cDNA (lane 3) respectively) were performed with degenerate primers 1S/1A. PCR products (421 bp) are indicated with an arrow (lanes 4, 5 and 6). A *PstI* digest of lambda was used as molecular weight marker (lane 1).

(b) PCR amplification of the 421 bp fragment from genomic DNA (isolated from sea urchin sperm) using primer pair SP1/SP2 (lanes 3 - 7) was optimised by using different concentrations of genomic DNA, $MgCl_2$ and DMSO, as indicated in each lane. A negative control (without genomic DNA) is shown in lane 2. A *PstI* digest of lambda was used as molecular weight marker (lane 1).

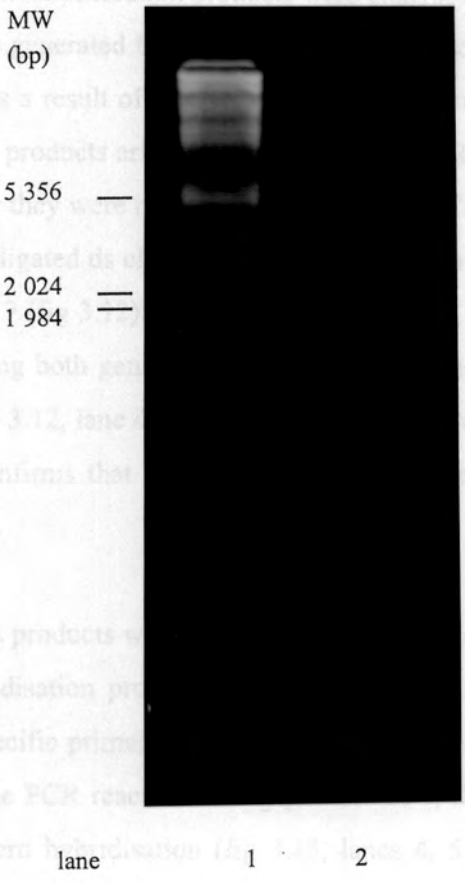


Fig 3.11 Analysis of cDNA Generated from Sea Urchin Embryo Total RNA

cDNA obtained by reverse transcription of total RNA (from 14 hour sea urchin embryos) using MMLV reverse transcriptase was analysed on an 1 % agarose gel (lane 2). The molecular weight marker (lane 1) is an *EcoRI / HindIII* digest of phage lambda.

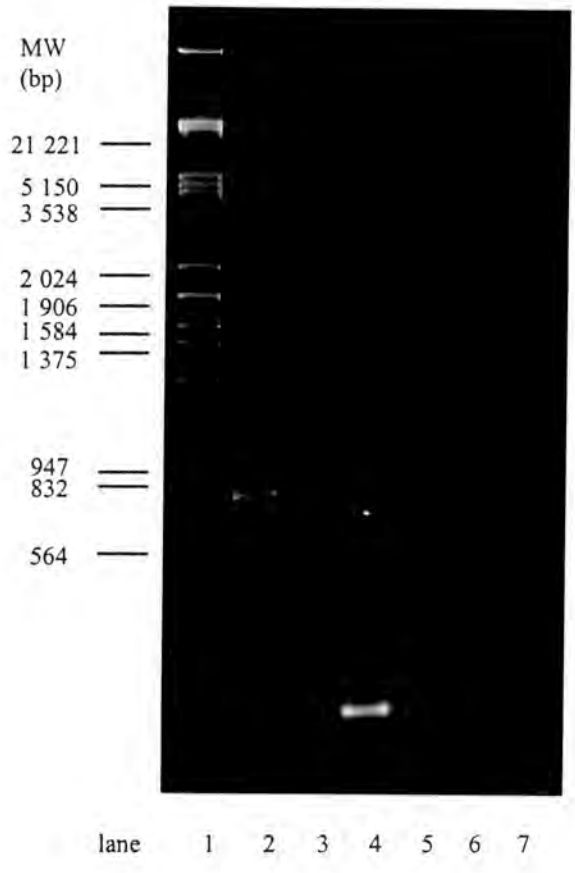


Fig 3.12 RACE Generates Multiple 5' and 3' DNA Fragments

The DNA fragments generated by 5' and 3' RACE reactions (see section 2.11.3) using sea urchin cDNA (fig 3.11) were analysed on an 1 % agarose gel (lanes 2 and 3 respectively). The 421 bp product (lane 4) generated by amplification of cDNA with primer pair SP1/SP2 (see fig 3.10) served as a positive control, whereas negative controls (lanes 5, 6 and 7) represent PCR reactions using the three respective primers SP1, SP2 and AP1 (Clontech) individually, under the same amplification conditions. The molecular weight marker (lane 1) is an *EcoRI / HindIII* digest of phage lambda.

has a major band in the expected region (viz ~ 1.3 kb), and there is an additional strong signal in the region of 3 kb. The two bands probably arose as a result of different polyadenylation sites (see section 6.1.3).

The two major fragments resulting from the 5' and 3' RACE reactions (marked by the arrows in fig 3.13) were recovered from a low melting point agarose gel, and the fragments were purified using Wizard PCR Preps columns (section 2.5.2). The fragments were identified further by sequencing them automatically. The isolated 5' RACE product was sequenced with the T7 primer (an annealing site is situated on the adaptor ligated to the cDNA, see Appendix IX) and the 3' RACE fragment was sequenced with a gene specific primer (SP1). Homology comparisons established that the two fragments resulting from 5' and 3' RACE have a very high identity with the SpGCF1 cDNA sequence, viz 74 % for the 5' RACE product and 92 % for the 3' RACE product (see Appendix XI). This indicates that together these fragments represent the *P.angulosus* homologue of SpGCF1 from *S.purpuratus*. Comparing the combined length (~ 2.1 kb) of the 5' and 3' RACE fragments to the SpGCF1 DNA sequence reveals that it is likely that the full length *P.angulosus* clone has been amplified, since the 5' RACE product extends at least 50 bp further into the 5' untranslated region than the cDNA for SpGCF1 (see Appendix XI), and judging from its complete length (~ 1.3 kb) the 3' RACE product extends at least 300 bp beyond the 3' end of the open reading frame of SpGCF1. Thus the 5' and 3' RACE products were characterised by Southern blot analysis and their correct identity was confirmed by sequencing the fragments.

The generation of a ~ 2.1 kb full length clone potentially encoding suGF1 involved fusing the 5' and 3' RACE fragments (see section 2.11.3.5) which was facilitated by the region of overlap between the fragments (the sense and the antisense primers are separated by 421 bp (see Appendix VII)). The final product of the fusion / amplification reaction was analysed by agarose gel electrophoresis (fig 3.14, lane 2). The full length product was isolated from a low melting point agarose gel, purified using a Wizard PCR preps column and the fragment was cloned into pGEM-T (a T/A type PCR cloning vector, see Appendix VIII (c)) using the A overhang incorporated by Taq polymerase on the PCR products. DNA from several colonies was isolated to confirm the correct size of the insert and subsequently the plasmids containing inserts were sequenced (see section 3.6).

3.6 DNA Sequence Analysis of the PCR-Generated cDNA Clone

A consensus DNA sequence was derived for the full length *P.angulosus* clone by automatic sequencing of several independent colonies (see fig 3.15). A minimum of three colonies was

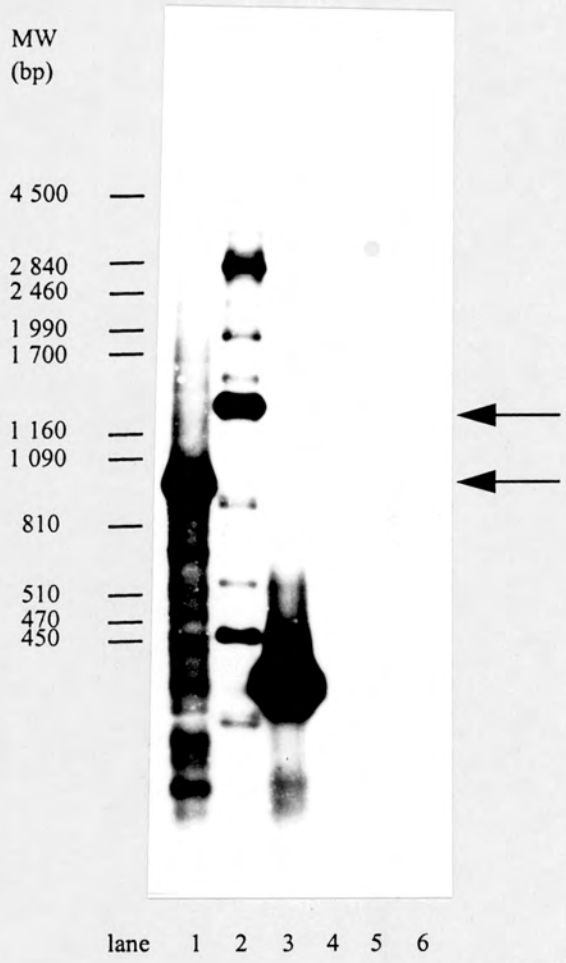


Fig 3.13 Southern Analysis of the 5' and 3' RACE Products

The 5' and 3' RACE products as well as the control reactions were separated on an 1 % agarose gel and blotted onto a nitrocellulose filter (see section 2.12). Southern analysis was performed using the radioactively labelled 421 bp fragment representing the region between the gene specific primers (SP1 and SP2), which were used in the amplification reactions. Multiple RACE products (lane 1 (5' RACE) and lane 2 (3' RACE)), as well as the 421 bp fragment (lane 3) hybridised to the probe. The main RACE products (see lanes 1 and 2) with correctly predicted sizes are marked by arrows. The negative control reactions performed with primers SP1, SP2 and AP1 individually (lanes 4, 5 and 6 respectively) did not hybridise to the probe. The molecular weight marker (a *Pst I* digest of lambda) is marked in the margin.

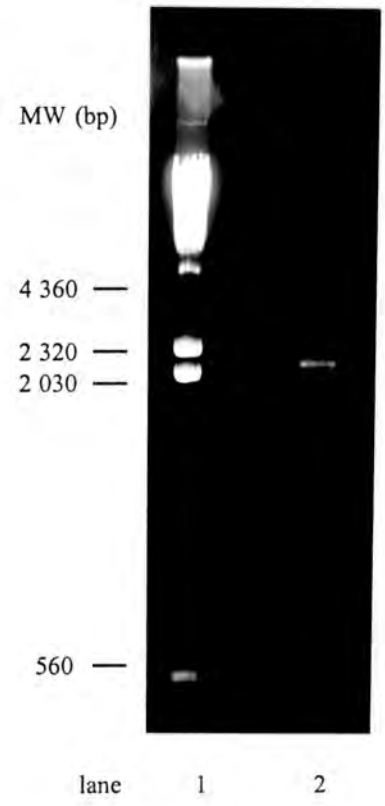


Fig 3.14 Fusion of the 5' and 3' RACE Fragments Yielded a Full Length Clone

The 5' and 3' RACE products were fused in their region of overlap and extended by a thermal cycling reaction in which both products simultaneously acted as primer and template. The resultant full length product of ~ 2.1 kb (lane 2) was PCR amplified, gel purified and cloned into the pGEM-T vector. The molecular weight marker (lane 1) is a *HindIII* digest of phage lambda.

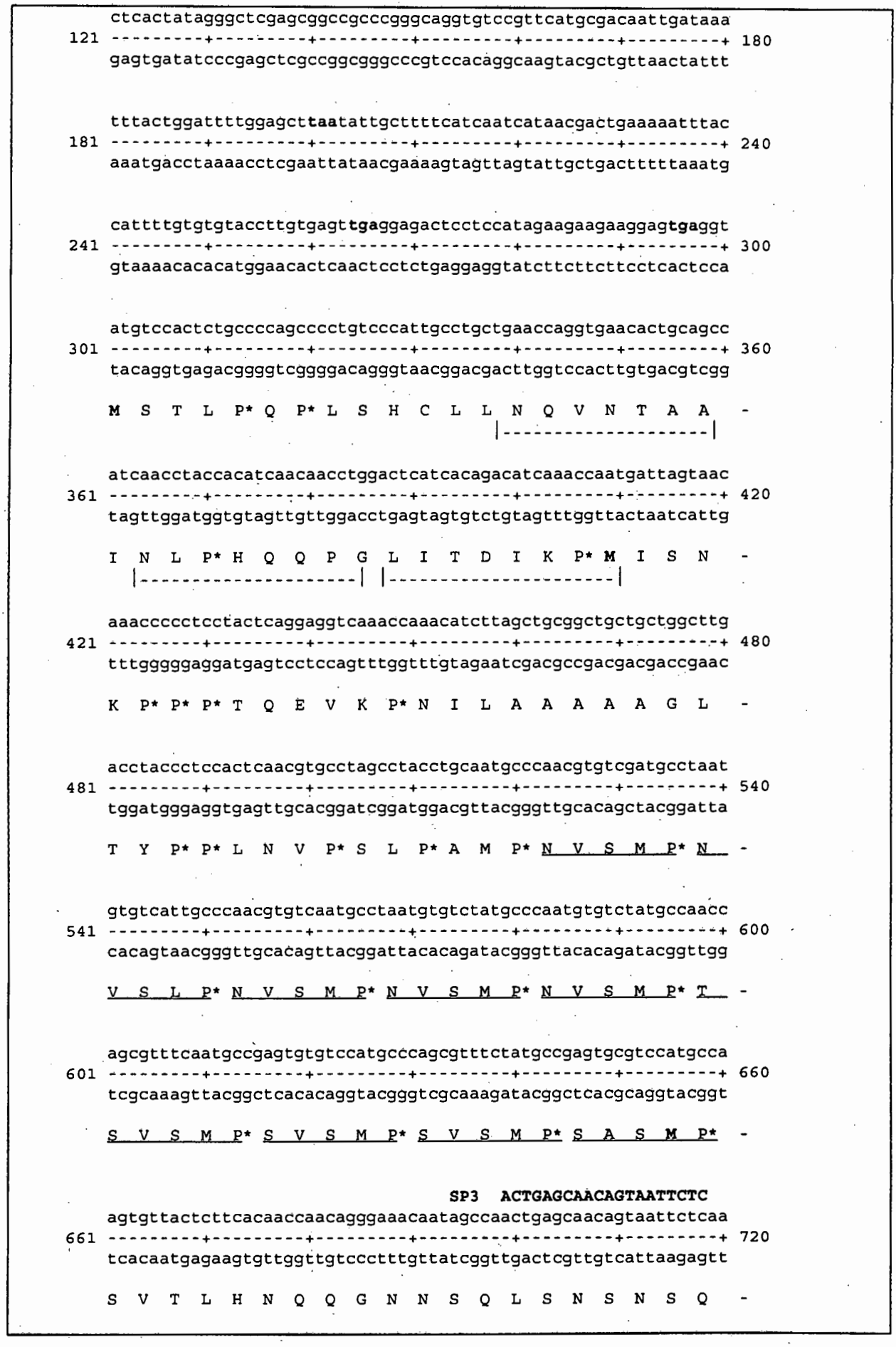
sequenced in both directions, using the T7 and Sp6 primers of the pGEM-T vector (see Appendix X). In addition, internal sequencing was achieved by primer walking, ie internal gene specific primers were designed (primers SP3 and SP4, see Appendix IX and fig 3.15) in order to obtain continuous sequence information. Several nucleotide positions in the sequences showed degeneracy within the three clones. These regions were resequenced on two additional independent clones in order to eliminate any errors in the consensus sequence for the entire full length clone (fig 3.15). A homology comparison between the *P.angulosus* clone identified by the PCR strategy and the cDNA sequence coding for SpGCF1 (3) using the GCG computer programme, revealed that the full length *P.angulosus* clone includes the entire coding region, as well as some 5' (~ 300 nt) and 3' (~ 500 nt) untranslated sequence information. An alignment of the two sequences (see Appendix XIV) shows an 84 % identity over a region of 1989 nt. The *P.angulosus* cDNA has a single open reading frame of 1542 nt, which can be translated into a protein of 514 amino acids (see fig 3.15), with a molecular weight of 57 kDa.

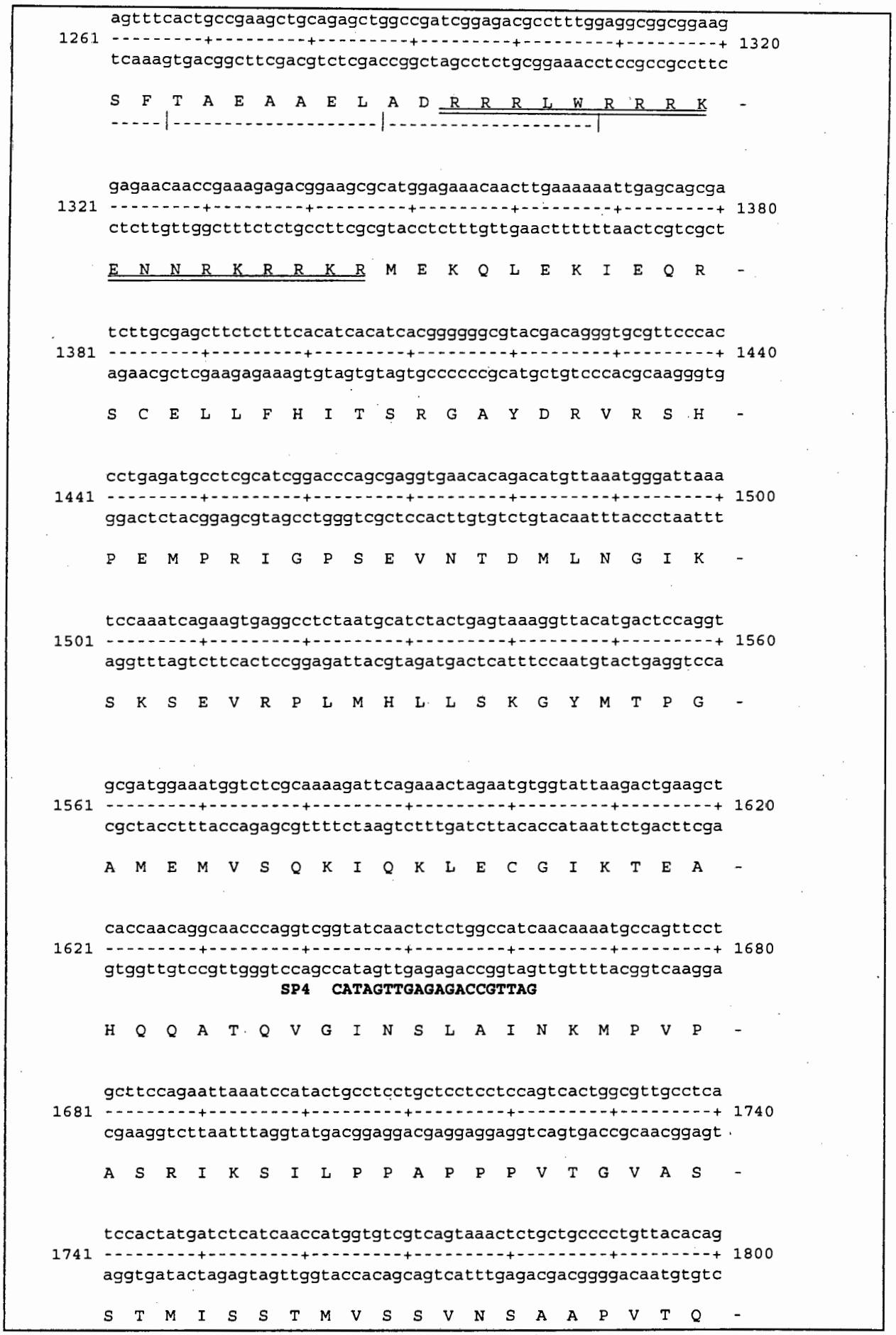
In conclusion, several clones potentially encoding suGF1 were generated using a combination of a PCR strategy (see section 3.5) and DNA ligand screening a cDNA expression library (see section 3.4). Further determination of the DNA-binding specificity of all the recombinant proteins encoded by the putative positive clones was necessary in order to establish whether any of the cDNA clones isolated by both these methods correctly represent the cDNA coding for suGF1 (see Chapter 4).

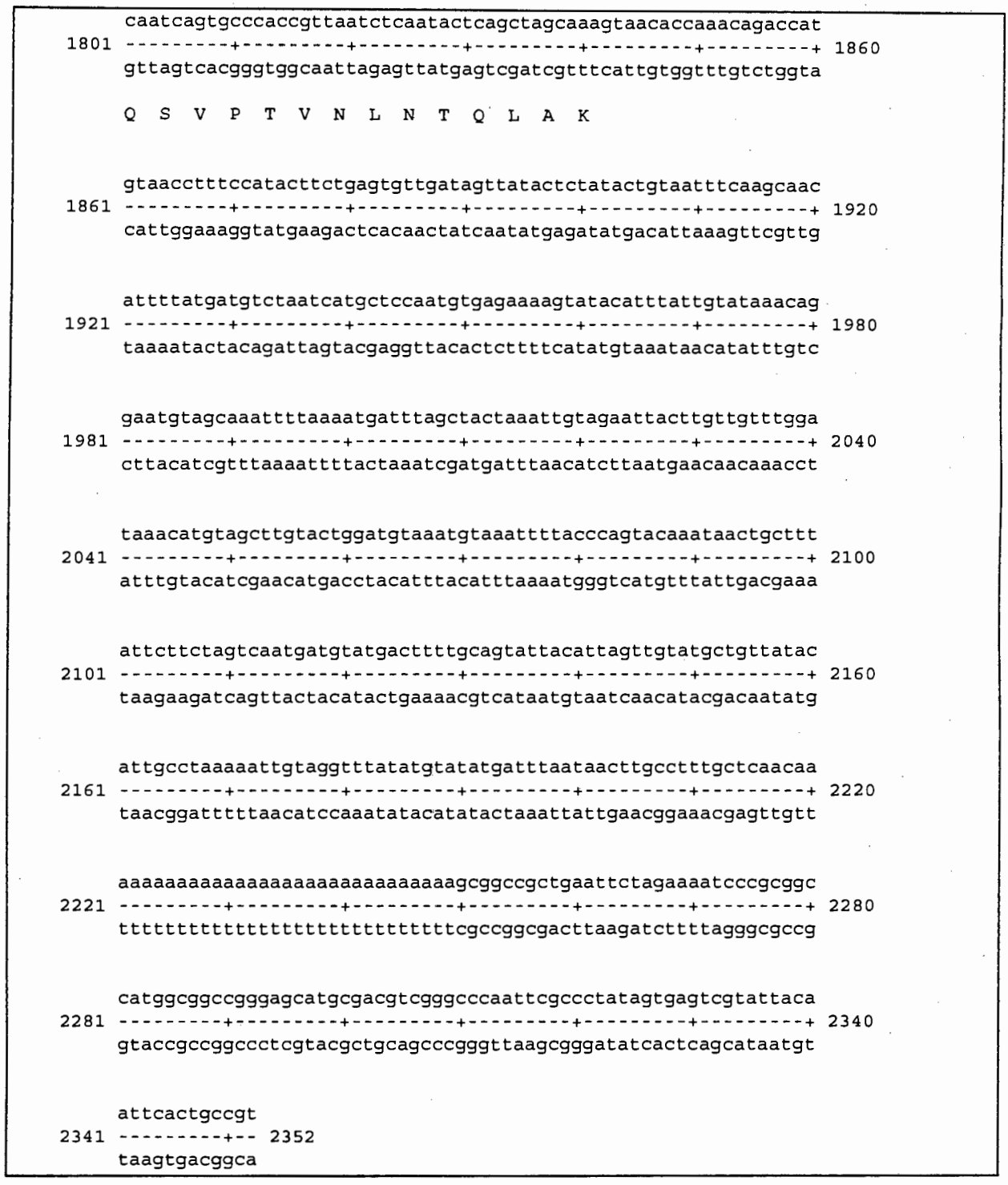
Fig 3.15 Sequence Analysis of the Full Length *P.angulosus* cDNA Clone Isolated by the PCR Strategy

A full length clone potentially encoding suGF1 was isolated from *P.angulosus* using a PCR strategy (see section 3.5). The consensus sequence for the clone was derived by sequencing a minimum of three independent colonies. The specific primer pair (SP1 / SP2) was used to PCR amplify the 3' and 5' ends of the cDNA respectively, whereas primers SP3 and SP4 were used to obtain internal DNA sequence information. (All primers are shown in bold print.) The 5' untranslated region has three stop codons (bold print) in frame with a single open reading frame, which codes for a 514 amino acid protein. The N-terminus is rich in proline residues (marked by asterisks), and contains several methionine residues (bold print) which may present alternate translation start sites. The protein is characterised by nine pentapeptide repeats (N/SVSMP), which are underlined. It also has a basic domain representing a putative DNA-binding domain, which is double underlined and it has heptad repeats which are marked by the broken lines.

```
      ttacgccaagctatttaggtgacactatagaatactcaagctatgcatccaacgcgttgg
    1 ---------+---------+---------+---------+---------+---------+ 60
      aatgcggttcgataaatccactgtgatatcttatgagttcgatacgtaggttgcgcaacc


      gagctctcccatatgggtcgacctgcaggcggccgcactagtgattccatcctaatacga
   61 ---------+---------+---------+---------+---------+---------+ 120
      ctcgagagggtatacccagctggacgtccgccggcgtgatcactaaggtaggattatgct
```


ctcactatagggctcgagcggccgcccgggcaggtgtccgttcatgcgacaattgataaa
121 ---------+---------+---------+---------+---------+---------+ 180
gagtgatatcccgagctcgccggcgggcccgtccacaggcaagtacgctgttaactattt
tttactggattttggagcttaatattgcttttcatcaatcataacgactgaaaaatttac
181 ---------+---------+---------+---------+---------+---------+ 240
aaatgacctaaaacctcgaattataacgaaaagtagttagtattgctgacttttaaatg
cattttgtgtgtaccttgtgagttgaggagactcctccatagaagaagaaggagtgaggt
241 ---------+---------+---------+---------+---------+---------+ 300
gtaaaacacacatggaacactcaactcctctgaggaggtatcttcttcttcctcactcca
atgtccactctgccccagcccctgtcccattgcctgctgaaccaggtgaacactgcagcc
301 ---------+---------+---------+---------+---------+---------+ 360
tacaggtgagacggggtcggggacagggtaacggacgacttggtccacttgtgacgtcgg
M S T L P* Q P* L S H C L L N Q V N T A A -
atcaacctaccacatcaacaacctggactcatcacagacatcaaaccaatgattagtaac
361 ---------+---------+---------+---------+---------+---------+ 420
tagttggatggtgtagttgttggacctgagtagtgtctgtagtttggttactaatcattg
I N L P* H Q Q P G L I T D I K P* M I S N -
aaacccccctcctactcaggaggtcaaaccaaacatcttagctgcggctgctgctggcttg
421 ---------+---------+---------+---------+---------+---------+ 480
tttgggggaggatgagtcctccagtttggtttgtagaatcgacgccgacgacgaccgaac
K P* P* P* T Q E V K P* N I L A A A A A G L -
acctaccctccactcaacgtgcctagcctacctgcaatgcccaacgtgtcgatgcctaat
481 ---------+---------+---------+---------+---------+---------+ 540
tggatgggaggtgagttgcacggatcggatggacgttacgggttgcacagctacggatta
T Y P* P* L N V P* S L P* A M P* N V S M P* N -
gtgtcattgcccaacgtgtcaatgcctaatgtgtctatgcccaatgtgtctatgccaacc
541 ---------+---------+---------+---------+---------+---------+ 600
cacagtaacgggttgcacagttacggattacacagatacgggttacacagatacggttgg
V S L P* N V S M P* N V S M P* N V S M P* T -
agcgtttcaatgccgagtgtgtccatgcccagcgtttctatgccgagtgcgtccatgcca
601 ---------+---------+---------+---------+---------+---------+ 660
tcgcaaagttacggctcacacaggtacgggtcgcaaagatacggctcacgcaggtacggt
S V S M P* S V S M P* S V S M P* S A S M P* -
SP3 ACTGAGCAACAGTAATTCTC
agtgttactcttcacaaccaacagggaaacaatagccaactgagcaacagtaattctcaa
661 ---------+---------+---------+---------+---------+---------+ 720
tcacaatgagaagtgttggttgtccctttgttatcggttgactcgttgtcattaagagtt
S V T L H N Q Q G N N S Q L S N S N S Q -


agtttcactgccgaagctgcagagctggccgatcggagacgcctttggaggcggcggaag
1261
1320
tcaaagtgacggcttcgacgtctcgaccggctagcctctgcggaaacctccgccgccttc
S F T A E A A E L A D R R R L W R R R K -
gagaacaaccgaaagagacggaagcgcatggagaaacaacttgaaaaaattgagcagcga
1321
1380
ctcttgttggctttctctgccttcgcgtacctctttgttgaacttttttaactcgtcgct
E N N R K R R K R M E K Q L E K I E Q R -
tcttgcgagcttctctttcacatcacatcacgggggcgtacgacagggtgcgttcccac
1381
1440
agaacgctcgaagagaaagtgtagtgtagtgccccccgcatgctgtcccacgcaagggtg
S C E L L F H I T S R G A Y D R V R S H -
cctgagatgcctcgcatcggacccagcgaggtgaacacagacatgttaaatgggattaaa
1441
1500
ggactctacggagcgtagcctgggtcgctccacttgtgtctgtacaatttaccctaattt
P E M P R I G P S E V N T D M L N G I K -
tccaaatcagaagtgaggcctctaatgcatctactgagtaaaggttacatgactccaggt
1501
1560
aggtttagtcttcactccggagattacgtagatgactcatttccaatgtactgaggtcca
S K S E V R P L M H L L S K G Y M T P G -
gcgatggaaatggtctcgcaaaagattcagaaactagaatgtggtattaagactgaagct
1561
1620
cgctacctttaccagagcgttttctaagtctttgatcttacaccataattctgacttcga
A M E M V S Q K I Q K L E C G I K T E A -
caccaacaggcaacccaggtcggtatcaactctctggccatcaacaaaatgccagttcct
1621
1680
gtggttgtccgttgggtccagccatagttgagagaccggtagttgttttacggtcaagga
SP4 CATAGTTGAGAGACCGTTAG
H Q Q A T Q V G I N S L A I N K M P V P -
gcttccagaattaaatccatactgcctcctgctcctcctccagtcactggcgttgcctca
1681
1740
cgaaggtcttaatttaggtatgacggaggacgaggaggaggtcagtgaccgcaacggagt
A S R I K S I L P P A P P P V T G V A S -
tccactatgatctcatcaaccatggtgtcgtcagtaaactctgctgcccctgttacacag
1741
1800
aggtgatactagagtagttggtaccacagcagtcatttgagacgacggggacaatgtgtc
S T M I S S T M V S S V N S A A P V T Q -


caatcagtgcccaccgttaatctcaatactcagctagcaaagtaacaccaaacagaccat
1801 ---------+---------+---------+---------+---------+---------+ 1860
gttagtcacgggtggcaattagagttatgagtcgatcgtttcattgtggtttgtctggta
Q S V P T V N L N T Q L A K
gtaacctttccatacttctgagtgttgatagttatactctatactgtaatttcaagcaac
1861 ---------+---------+---------+---------+---------+---------+ 1920
cattggaaaggtatgaagactcacaactatcaatatgagatatgacattaaagttcgttg
attttatgatgtctaatcatgctccaatgtgagaaaagtatacatttattgtataaacag
1921 ---------+---------+---------+---------+---------+---------+ 1980
taaaatactacagattagtacgaggttacactcttttcatatgtaaataacatatttgtc
gaatgtagcaaattttaaaatgatttagctactaaattgtagaattacttgttgtttgga
1981 ---------+---------+---------+---------+---------+---------+ 2040
cttacatcgtttaaaattttactaaatcgatgatttaacatcttaatgaacaacaaacct
taaacatgtagcttgtactggatgtaaatgtaaattttacccagtacaaataactgcttt
2041 ---------+---------+---------+---------+---------+---------+ 2100
atttgtacatcgaacatgacctacatttacatttaaaatgggtcatgtttattgacgaaa
attcttctagtcaatgatgtatgacttttgcagtattacattagttgtatgctgttatac
2101 ---------+---------+---------+---------+---------+---------+ 2160
taagaagatcagttactacatactgaaaacgtcataatgtaatcaacatacgacaatatg
attgcctaaaaattgtaggtttatatgtatatgatttaataacttgcctttgctcaacaa
2161 ---------+---------+---------+---------+---------+---------+ 2220
taacggatttttaacatccaaatatacatatactaaattattgaacggaaacgagttgtt
aaaaaaaaaaaaaaaaaaaaaaaaaaaagcggccgctgaattctagaaaatcccgcggc
2221 ---------+---------+---------+---------+---------+---------+ 2280
ttttttttttttttttttttttttttttcgccggcgacttaagatcttttagggcgccg
catggcggccgggagcatgcgacgtcgggcccaattcgccctatagtgagtcgtattaca
2281 ---------+---------+---------+---------+---------+---------+ 2340
gtaccgccggccctcgtacgctgcagcccgggttaagcgggatatcactcagcataatgt
attcactgccgt
2341 ---------+-- 2352
taagtgacggca

3.7 Developmental Distribution of the mRNA Transcript Corresponding to the PCR-Generated cDNA in *P.angulosus*

To determine the developmental distribution of the RNA transcript corresponding to the PCR-generated clone (see section 3.5), total RNA was isolated from *P.angulosus* eggs, several developmental embryonic stages (4, 9, 14, 21, 30 and 45 hour embryos), as well as adult tissue (muscle, ovary and testis). The RNA was analysed by Northern blotting and reverse-transcriptase (RT) PCR. Fractionation of RNA under denaturing conditions (see section 2.6.4.1) followed by acridine orange staining showed that the RNA was intact, as determined by the presence of distinct undegraded 28S and 18S ribosomal bands. Northern blotting of the fractionated RNA was performed with a radiolabelled probe containing part of the putative suGF1 cDNA (viz the overlap region between specific primers SP1 and SP2 (see Appendix IX, and fig 3.15)). Hybridisation signals resulting from Northern blots were uninterpretable as the probe appeared to bind over a wide range of molecular weights for each RNA sample (data not shown). An alternative approach using RT-PCR was followed. Briefly, this technique involved treating the RNA samples with DNase 1 and subsequently with MMLV reverse transcriptase (see section 2.7). cDNA generated from each RNA sample was amplified in a PCR reaction using specific primers SP1 and SP1, generating the characteristic 421 bp fragment described above. PCR amplifications were analysed on an agarose gel, revealing that the putative suGF1 RNA is present in *P.angulosus* eggs, embryonic stages (4 - 45 hour embryos), as well as adult muscle and testis tissue (see individual lanes in fig 3.16). In contrast, ovary tissue appears to be deficient in the RNA transcript of interest (see fig 3.16, lane 11).

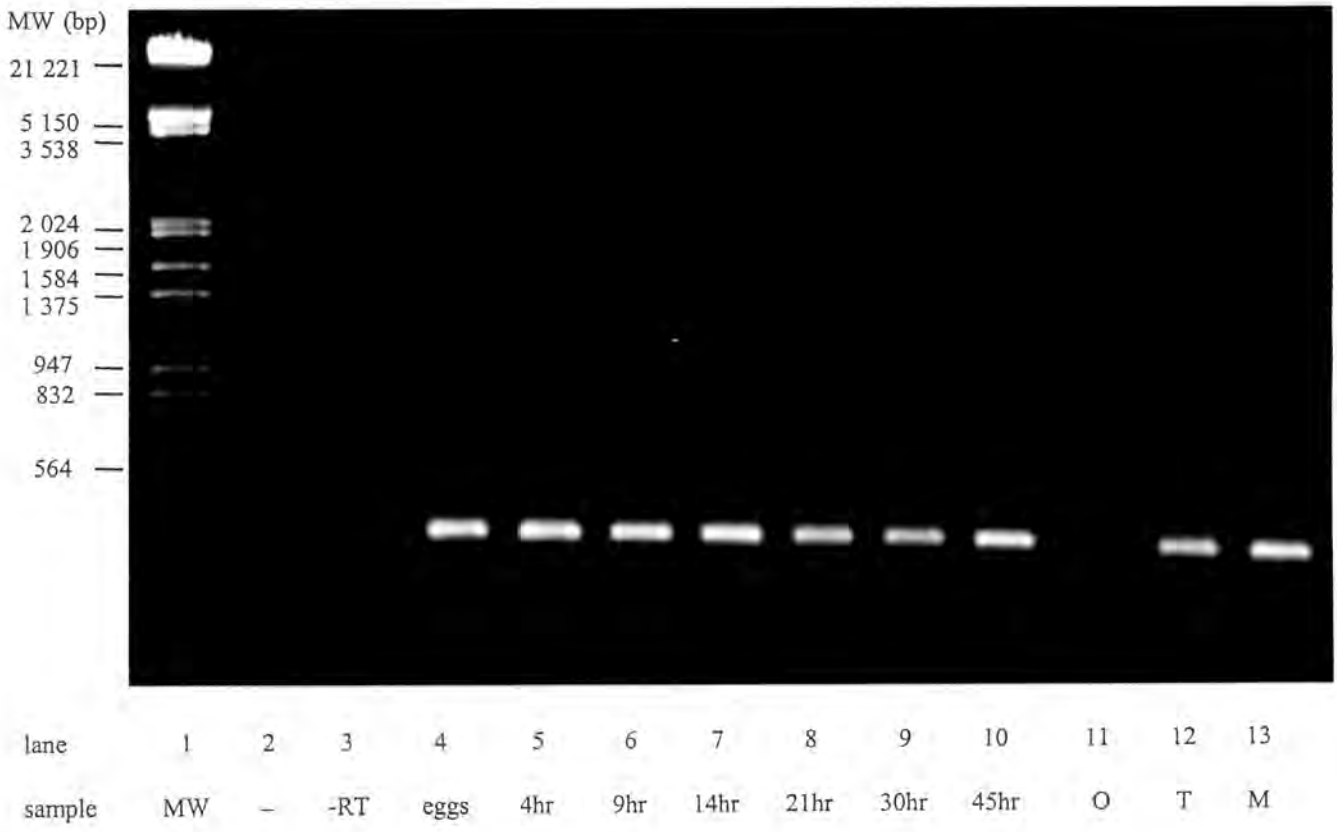


Fig 3.16 Analysis of the Distribution of the mRNA Transcript Corresponding to the PCR-Generated Clone

mRNA isolated from sea urchin eggs, embryos (4 - 45 hours) and adult tissue (muscle (M), ovaries (O) and testes (T)) was reverse transcribed using MMLV reverse transcriptase (see section 2.7). The resulting cDNA was used in PCR amplifications in combination with specific primers SP1 and SP2 to amplify the characteristic 421 bp fragment. The mRNA transcript of interest is present in eggs (lane 4), 4 - 45 hour embryos (lanes 5 - 10), testes (lane 12) and muscle tissue (lane 13). The transcript is absent in ovaries (lane 11). Control PCR amplifications performed in the absence of cDNA (lane 2), and with 14 hour mRNA not treated with reverse transcriptase (lane 3) did not yield DNA products. A molecular weight marker (*EcoRI / HindIII* digest of lambda) is shown in lane 1.

CHAPTER 4

RECOMBINANT PROTEIN EXPRESSION

4.1 Introduction

Biologically important eukaryotic proteins are often cloned and overexpressed to produce large amounts of material. Commonly, the expression of large quantities of proteins encoded by cloned genes can be achieved in *E.coli*, since this system is easy to manipulate, inexpensive to maintain, grows quickly and represents one of the best understood organisms in nature (151, 204). Several other systems such as yeast, mammalian cells, baculovirus, plants, transgenic animals and eukaryotic *in vitro* transcription / translation systems are also available for recombinant protein expression (205). These systems generally rely on a similar basic approach. A foreign gene is inserted into an expression vector (eg bacteriophage, plasmid or virus), which has several characteristic features including (i) a selectable marker to actively maintain the foreign gene, (ii) a transcription promoter (eg *lac*, *trp*, *tac*, T7, etc) whose induction is tightly controlled and maximises the efficiency of transcription, (iii) translational control signals, eg a ribosome binding site and signals for the initiation of translation, and (iv) a polylinker which simplifies the insertion of a foreign gene into the vector (151). The choice of the expression vector is often determined by its promoter (promoters may be controlled by various mechanisms, eg temperature shift, chemical induction or metabolic response, etc) (151). Expression constructs are transformed into appropriate expression systems, which, ideally, should facilitate both the expression and purification of foreign target peptides. However, recombinant protein expression is often associated with problems such as instability, insolubility and inactivity of the target protein. These problems may be overcome by conditions determined for each target protein individually, such as the synthesis of hybrid proteins, the use of host strains with decreased proteolytic capacity, or exploitation of the formation of inclusion bodies, which may sometimes be solubilised and renatured to gain functional activity by use of denaturing agents.

Unfortunately there are no set methods or rules which guarantee the successful expression of cloned proteins in a useful form, and each new gene presents its own unique expression problems. This means that the expression of each gene has to be tested and optimised individually, possibly using several host expression systems and varying the expression conditions for each. Therefore, since the

expression of recombinant proteins is an inexact science, a trial and error approach was used to investigate the expression of the isolated cDNA clones (see Chapter 3) potentially encoding suGF1.

4.2 Expression of Clones Obtained by the DNA Ligand Screening Approach

4.2.1 Recombinant Protein Expression Using the Prokaryotic Expression Systems pBluescript, pET and pGEX

Recombinant protein expression of Clones 2, 6, 11 and 16 in Bluescript (see Appendix VIII (a)) was induced by addition of IPTG to mid log phase cells, which were grown to stationary phase and subsequently assayed for recombinant protein expression by comparison to uninduced cells using SDS-PAGE (see section 2.24). Various techniques (see section 2.14.1) were used in attempts to detect recombinant protein expression from the bacterial cells. The most direct assay involved lysis of the induced stationary phase cells (which were pelleted by centrifugation) in SDS loading dye with subsequent analysis by SDS-PAGE. A second method involved lysis of induced stationary phase cells using a combination of lysozyme and Triton X-100, and the bacterial lysate (separated from the cell debris by centrifugation) was analysed by SDS-PAGE for the presence of recombinant proteins. A third method promoted the release of proteins from the periplasmic space by high speed centrifugation of induced stationary phase cells. Again the supernatant was analysed by SDS-PAGE after TCA precipitation of the proteins. Each method involved protein induction using several different *E.coli* host strains, eg SOL^R, JM109, XL-1 *Blue* and DH5α, and analysis was performed on multiple colonies derived from each host strain. Various other parameters were optimised, such as varying the temperature and length of induction, and using different volumes of culture, ranging from 2 ml to 50 ml. In addition, protein extracts were separated by various percentages SDS-PAGE (the gels were stained using both Coomassie Brilliant Blue and silver). Analysis of protein extracts was also attempted by EMSA and DNase 1 footprinting, which are highly sensitive techniques used to detect DNA-binding proteins. However, none of the strategies outlined above resulted in the detection of induced recombinant proteins (data not shown), despite the fact that target proteins from Bluescript inserts should be reasonably stable, since they are expressed as fusion proteins to the amino terminus (~ 20 - 50 amino acids) of the α-complementing portion of the β-galactosidase gene (206).

The four clones (2, 6, 11 and 16) were subsequently subcloned from pBluescript SK into pET-29b(+) (see Appendix XIV (a) for details on the pET system and Appendix XIV (b) for details on the

subcloning strategy). Several parameters were varied in order to obtain and optimise expression of these constructs. For instance, the inserts were subcloned using several unique restriction sites, the resulting pET constructs were established in various different *E.coli* host strains (JM109, HB101 and DH5α) and subsequently induced to express in two different host strains BL21DE3 and pLysS (see Appendix XIV for details). Multiple colonies from each expression strain were analysed for expression, the conditions were also altered with respect to temperature and length of induction, as well as volume of culture. A plasmid referred to as the "induction control", which codes for the β-galactosidase protein and has matching elements (ie promoter, selective marker, etc), was provided with the pET-29b(+) vector (see Appendix VIII). This plasmid allowed convenient testing of induction. Optimisation of expression (see Appendix XIV (b) for experimental details) resulted in the expression of a single clone only. Clone 11 was expressed successfully as a ~ 25 kDa protein in the BL21DE3 expression host, as depicted by the arrow in fig 4.1. Expression was independent of both the restriction site used in the subcloning procedure (data not shown) and the host strain the construct was originally established in (fig 4.1, lanes 4 - 6 (DH5α), lanes 7 - 9 (HB101) and lanes 10 - 12 (JM109)). Both uninduced (lanes 4, 7 and 10) and induced cells (lanes 5, 6, 8, 9, 11 and 12) were harvested after 3 hours, resuspended in SDS sample application buffer and analysed by SDS-PAGE (silver stained). Protein molecular weight standards are indicated (lane 1), and uninduced and induced total cell protein from the induction control plasmid were analysed, too (lanes 2 and 3, respectively). The asterisk indicates the position of the induced β-gal protein, which, although expressed very strongly, was not always easily distinguishable from the background, as it often discoloured into a light yellow band upon silver staining. Expression of Clone 11 was analysed by a time course of induction at various temperatures (16°C, 20°C, 30°C and 37°C, see Appendix XVI (b) for details), which established that the recombinant protein is maximally induced at 37°C after 3 - 4 hours of induction and at 30°C after 5 - 6 hours of induction. Recombinant protein expression could not be detected for any of the other clones (2, 6 and 16), irrespective of the conditions used to optimise induction (eg changes in temperature, time, volume of culture), the cloning site used and initial or expression host. However expression of the β-gal protein from the control plasmid was always detected successfully in these assays (data not shown).

In order to confirm the DNA-binding ability of Clone 11 several strategies were attempted to generate recombinant protein in a soluble form (see section 2.14.3 and Appendix XIV (c)). For the first method, induced cells were lysed using a combination of lysozyme and Triton X-100, the cellular debris was separated from the soluble proteins in the supernatant and analysed by SDS-PAGE (see Appendix XIV (c) for details), showing that Clone 11 can be induced as a soluble protein, however this result was not reproducible. Extracts containing soluble recombinant protein were titrated in

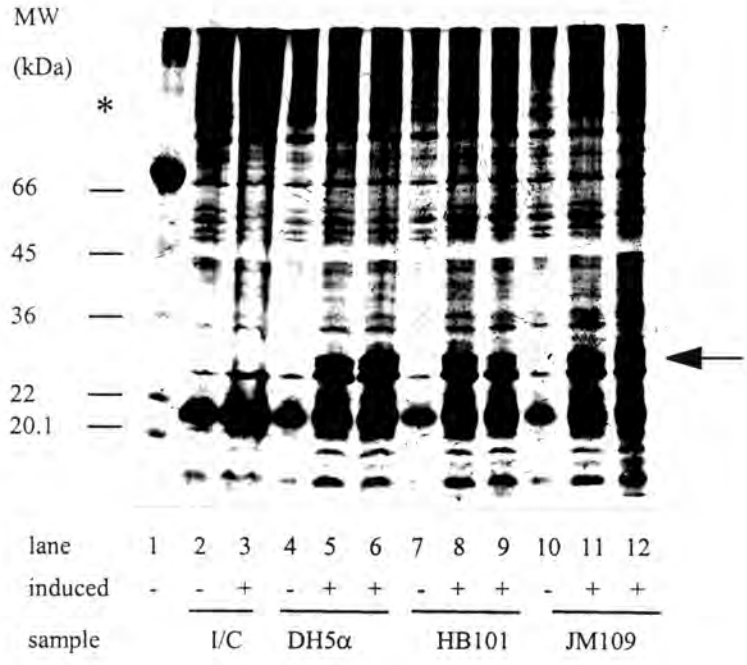


Fig 4.1 Clone 11 Expresses a 25 kDa Recombinant Protein

Bacterial cell extracts containing recombinant protein product generated from the pET-29b(+) / Clone 11 expression construct were analysed by 10 % SDS-PAGE (silver stained). The BL21DE3 expression host was induced to express the recombinant protein, after establishing the expression construct in various *E.coli* host strains, viz DH5α (lanes 4 - 6), HB101 (lanes 7 - 9) and JM109 (lanes 10 - 12). Uninduced cell extracts are shown in lanes 4, 7 and 10, whereas the induced extracts are in lanes 5 and 6 (DH5α), 8 and 9 (HB101) and 11 and 12 (JM109). The arrow shows the position of the induced recombinant protein expressed by Clone 11. Extracts from cells containing the induction control plasmid (I/C) are shown in lanes 2 (uninduced cells) and 3 (induced cells), the position of the induced β-gal protein is marked by the asterisk. The protein molecular weight standards are indicated (lane 1).

EMSAs using the E/H fragment as labelled DNA probe (see section 2.21), indicating that the induced recombinant protein was inactive with respect to DNA-binding activity (data not shown), despite attempts to improve recombinant protein folding by dialysis of the bacterial extracts into dialysis buffer (see section 2.19) containing 4 mM $MgCl_2$ and 2 mM $ZnCl_2$. (Prokaryotic gene expression may lead to incorrect folding of eukaryotic proteins or loss of their biological activity (151).) Further attempts to improve the method of soluble recombinant protein isolation were unsuccessful (see Appendix XIV (c) for details).

Purification of soluble recombinant protein was approached using a second method, which exploits the fusion of cloned inserts to the His-Tag on the pET plasmid (see Appendix VIII) by subsequent Nickel column chromatography of protein extracts. Cells were induced, pelleted and resuspended in column binding buffer (see section 2.14.3.2) containing protease inhibitors. The protein sample was processed by a combination of sonication, washing the pellet and differential centrifugation (for experimental details see Appendix XIV (c)). Protein extracts containing recombinant protein from Clone 11 were subjected to Nickel chromatography, however the recombinant protein could not be discerned in the elution profile resulting from the column, implying that insufficient protein may have been released into the supernatant for subsequent detection. Similarly when above procedure was repeated with induced β-gal protein from the induction control plasmid, the β-gal protein could also not be detected in the elution profile (see Appendix XIV (c) for details). These results imply that either insufficient protein was released by the sonication steps, and therefore was not detectable in the elution profile, or that the Nickel column was prone to ion leaching during the wash steps, leading to the loss of the protein-metal complexes. Alternatively these complexes may have leached out gradually over several of the eluting fractions. Thus Nickel chromatography did not appear to enrich the recombinant proteins, and therefore it was decided to approach the protein purification by first denaturing the insoluble recombinant protein using denaturing agents, in order to release greater amounts of protein.

It appeared that insoluble inclusion bodies (ie dense aggregates of highly concentrated expressed target protein) contained the major fraction of the expressed protein. These were used as a source for further purification of the recombinant protein. Several methods involving inclusion body isolations and renaturation of recombinant protein were examined (see Appendix XIV (d) for experimental details). Briefly, one method (pET System Manual and section 2.14.3.3) involved resuspension of the induced cells in binding buffer, followed by brief bursts of sonication, several washes and subsequent resuspension in binding buffer (see section 2.14.3.2) supplemented with 6 M guanidine HCl or 6 M urea. Resuspension of protein was aided by repeated sonication and subsequently it was dialysed into

dialysis buffer (see section 2.19) in order to remove the guanidine HCl. A second inclusion body isolation method (described by Calzone et al (1991) (120), and see section 2.14.3.3), involved lysis of induced cells using lysozyme (aided by sonication), and the suspension was supplemented with NP-40 and sucrose. Soluble proteins were selectively precipitated and the resuspended proteins were dialysed into dialysis buffer (see section 2.19). Inclusion body isolation using a third method (outlined by Lin and Cheng (1991) (186) and see section 2.14.3.3) required formation of spheroplasts from induced cells. Lysis was achieved by sonication, and nucleic acids were removed enzymatically. Crude inclusion bodies were pelleted and washed several times, resuspended by sonication and denatured in 5 M guanidine HCl. Proteins were renatured overnight by dilution of the denaturant, and subsequently dialysed into the buffer of choice (dialysis buffer). Despite several attempts to solubilise expressed recombinant protein by means of extensive sonication combined with denaturation agents such as urea and guanidine HCl, none of the methods outlined above (see Appendix XIV (d) for experimental details) successfully dissociated the inclusion bodies containing the recombinant protein expressed from Clone 11. The recombinant protein was continuously present in the pelleted fractions rather than the solubilised fractions (data not shown). Therefore, it appears that inclusion body isolations in combination with denaturation / renaturation techniques are not compatible with the isolation of the recombinant protein expressed by Clone 11. Overall, the above experiments indicate that the expression of the isolated sea urchin clones (see section 3.4) is not successful in combination with the pET-29 system, since firstly only one clone was successfully expressed (at very low levels) and secondly the respective recombinant protein could not be solubilised or renatured.

The expression of the four clones (2, 6, 11 and 16) was therefore attempted in the pGEX expression system. The clones were subcloned from pBluescript into the pGEX-3X vector (see Appendix VIII) using the *EcoR1* site in order to retain the same original reading frame. The correct orientation of the inserts was confirmed by restriction enzyme analysis. Recombinant protein expression was induced chemically by means of IPTG and induction was performed in several hosts, eg JM109, DH5α and MC1061. Several colonies of each clone (originating from different expression hosts) were analysed for recombinant protein production by SDS-PAGE, however only the characteristic 26 kDa glutathione S-transferase (GST) protein was expressed and no fusion proteins were generated, implying that none of the constructs were expressed successfully in the pGEX-3X vector (data not shown). Despite its eukaryotic origin GST is generally expressed at very high levels in *E.coli* as a soluble homodimeric protein with a MW of 26 kDa (207). Expression is induced chemically and is controlled by the *tac* promoter.

4.2.2 Eukaryotic Protein Expression in Mammalian Cells

Transient transfection studies were performed with Clone 11 subcloned into the pCIS plasmid (see Appendices XV). The expression construct was transfected into COS-1 cells (derived from African green monkey kidney cells) using the DEAE-dextran method (see section 2.15). Cells were allowed to express for 24 - 48 hours after which they were harvested and processed into whole cell extracts and nuclear extracts (see section 2.15), which were analysed by SDS-PAGE (see Appendix XV for details). No difference was observed between untransfected and transfected cells for either the nuclear or whole cell extracts, implying that recombinant protein expression may have been unsuccessful. Analysis of the extracts via EMSA using the labelled E/H probe revealed no DNA-binding activity in any of the extracts either (data not shown). The results imply that either the mammalian expression system is inefficient in this case (possibly due to low transfection success), such that the expression of the recombinant protein could not be ascertained or, alternatively, the protein could not be expressed in this host environment. Generally eukaryotic genes or cDNAs should express efficiently in eukaryotic systems, resulting in the proper translation and processing of the expressed product (208). Instead of performing a detailed trouble shooting analysis (see Appendix XV for a discussion of the possible underlying problems associated with the system), the problem of eukaryotic expression was readdressed using a different eukaryotic system, viz *in vitro* coupled transcription / translation (see section 2.16 and 4.2.3).

4.2.3 *In Vitro* Eukaryotic Transcription / Translation

Recombinant proteins from Clones 2, 6. 11 and 16 in the Bluescript plasmid (see Appendix VIII) were expressed in conjunction with a rabbit reticulocyte lysate transcription / translation system (see section 2.16). Transcription was effected from the T3 promoter in all the clones, including the luciferase control plasmid which was provided with the system. Eukaryotic translation of the proteins was generally performed in the presence of translational grade ^{35}S-methionine. The resultant protein products were analysed by SDS-PAGE with subsequent autoradiography (fig 4.2). Protein molecular weight markers are indicated in the margin, a control reaction was performed in the absence of a plasmid (lane 1), and the control plasmid resulted in translation of the 61 kDa luciferase protein (lane 2). Protein products from Clones 2, 6, 11 and 16 are shown in lanes 3, 4, 5 and 6 respectively. Clone 2 (2.2 kb) resulted in a protein of molecular weight about 48 kDa (lane 3), the protein arising from Clone 11 (0.9 kb) is about 25 kDa (lane 5), and Clone 16 (2.4 kb) has a protein of about 45 kDa (lane 6). The protein sizes are in approximate agreement with the sizes of the cDNA fragments encoding

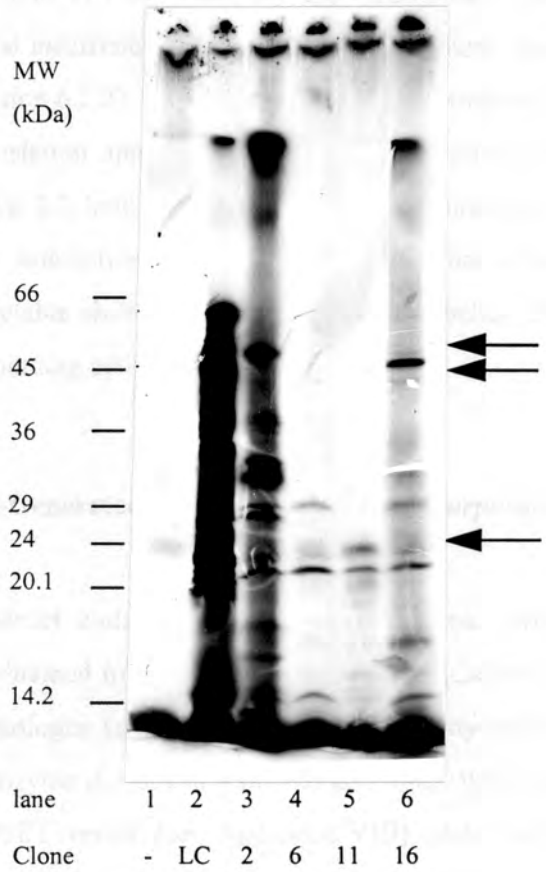


Fig 4.2 *In Vitro* Transcription / Translation Yielded Recombinant Proteins from the Clones Isolated by the DNA Ligand Screening Procedure

A rabbit reticulocyte lysate system (see section 2.16) was used to express recombinant proteins from Clones 2, 6, 11 and 16 in the presence of translational grade ^{35}S-methionine. Transcription was effected from the T3 promoter on the Bluescript plasmid, as well as the luciferase control plasmid which was provided with the system (Promega). The resultant protein products were analysed by 10 % SDS-PAGE with subsequent autoradiography. A negative control was performed without plasmid DNA (lane 1), and the luciferase control protein (LC) forms a clear product at about 61 kDa (lane 2). Protein products from Clones 2, 6, 11 and 16 are shown in lanes 3, 4, 5 and 6 respectively. Clone 2 (2.2 kb) encodes a protein of molecular weight ~ 48 kDa (lane 3), the protein arising from Clone 11 (0.9 kb) is ~ 25 kDa (lane 5), and Clone 16 (2.4 kb) has a protein of ~ 45 kDa (lane 6) as shown by the arrows. No unique protein bands corresponding to a protein encoded by Clone 6 (0.85 kb) could be identified (lane 4). Standard protein molecular weight markers are indicated in the margin.

them. No unique band corresponding to a protein encoded by Clone 6 (0.85 kb) could be identified (lane 4). This correlates with previous indications that this clone may not have an open reading frame (see section 3.4 and Appendix IV). Alternatively the recombinant protein may have a very low methionine content, such that insufficient ^{35}S label was incorporated, thereby preventing visualisation of the product (see also section 6.2.2). The amounts of protein produced from the four clones by the *in vitro* transcription / translation application appear to be relatively low when compared to the luciferase control protein (fig 4.2, lane 2). Expressed target proteins were titrated in EMSAs in order to assay for protein-DNA interactions, however it appears that either insufficient protein was expressed to discern a detectable shift in the mobility of the labelled DNA, or the expressed target proteins do not have DNA-binding activity (data not shown).

4.3 Expression of the PCR-Generated cDNA Clone and its *S.purpuratus* Homologue

A pRSET expression construct coding for the SpGCF1 protein (which represents a candidate homologue to suGF1) was obtained from Professor E. Davidson (Caltech) prior to the isolation of the full length *P.angulosus* homologue (see section 3.5). The integrity of this expression construct was confirmed by a restriction enzyme digest using *BamHI* and *BglII*, which resulted in the release of a ~ 1.7 kb insert from the pRSET vector (see Appendix VIII) (data not shown). Expression of the SpGCF1 construct was attempted in bacterial expression hosts BL21DE3 and pLysS, in order to correlate the identity of native suGF1 and recombinant SpGCF1. Induction of recombinant SpGCF1 expression was performed according to conditions outlined by Zeller et al (1995) (3). Expression of the recombinant protein could not be observed by comparing total protein from uninduced and induced cells as analysed by SDS-PAGE (fig 4.3, compare lanes 6 with lanes 7 and 8). However, when recombinant proteins were isolated from inclusion bodies (see section 2.14.3.3), induced cells appeared to have a stronger expression pattern in the region of 42 kDa marked by the arrow (fig 4.3, compare lanes 2 with lanes 3 and 4), which corresponds to one of the expected sizes of the SpGCF1 protein (3). This suggested that expression of SpGCF1 may have been successful. The DNA-binding activity of induced proteins was analysed by EMSA using the E/H fragment, however the pattern of protein-DNA complexes was not unique for the induced extract with respect to the uninduced protein fraction (data not shown), implying that either recombinant SpGCF1 was not being expressed or it was expressed in both uninduced and induced cells as a result of a leaky promoter. Therefore the specificity of the protein-DNA complexes was investigated using EMSAs performed with specific and nonspecific ds oligonucleotides and E/H fragment (see fig 2.1 and fig 2.2, respectively) as cold DNA competitors. Competition EMSAs revealed that neither the induced nor the uninduced protein

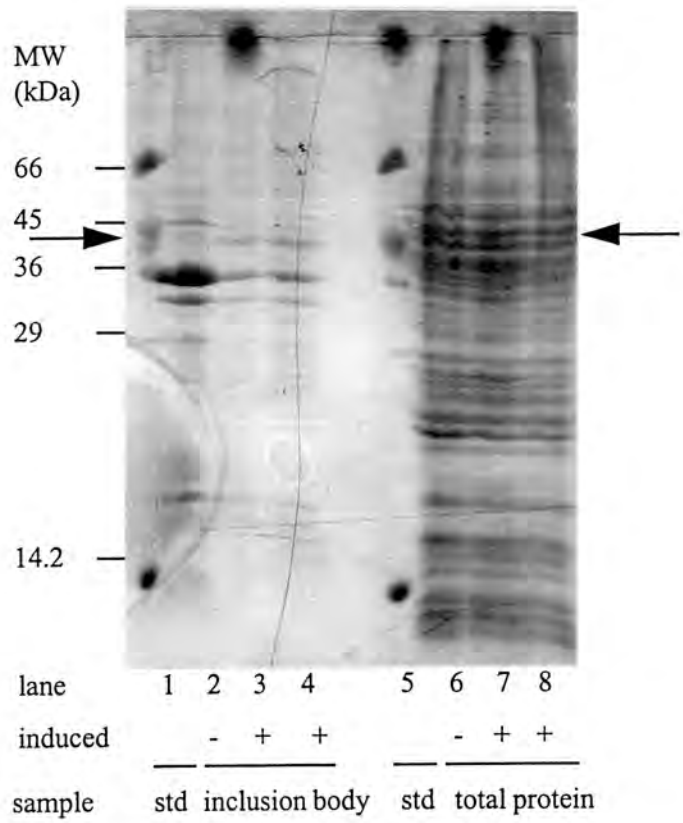


Fig 4.3 Inclusion Bodies Were Isolated from Bacterial Cells Induced to Express Recombinant SpGCF1

The induced protein pattern obtained from two different expression hosts containing the SpGCF1 expression construct was analysed by 10 % SDS-PAGE. The protein pattern of uninduced BL21DE3 pLysS cells before the purification of inclusion bodies is shown in lane 6, whereas lanes 7 (BL21DE3) and 8 (BL21DE3 pLysS) show a lysate of induced total bacterial proteins. Inclusion bodies were isolated from uninduced (lane 2 (BL21DE3 pLysS)) and induced bacterial hosts (lane 3 (BL21DE3 pLysS) and lane 4 (BL21DE3)). A stronger expression pattern was observed in the region of 42 kDa for induced inclusion body fractions (lanes 3 and 4), as indicated by the arrows. Standard protein molecular weight markers are marked (lanes 1 and 5).

extracts contained protein-DNA complexes which were sequence-specifically competed for by the G·C-rich competitors. This, together with the fact that none of the protein-DNA complexes were unique to the induced fraction, implies that recombinant expression of SpGCF1 was unsuccessful, and therefore comparisons between recombinant SpGCF1 and native suGF1 were impeded. Expression of the PCR-generated *P.angulosus* clone was not attempted in a bacterial system, since neither the SpGCF1 construct, nor Clones 2, 6, 11 and 16 had expressed successfully in bacteria.

Instead, both the PCR-generated *P.angulosus* clone, and its SpGCF1 homologue were successfully expressed using *in vitro* eukaryotic coupled transcription / translation (see section 2.16). Transcription of both the *P.angulosus* and *S.purpuratus* constructs (cloned into pGEM-T and pRSET respectively) was directed by the T7 bacteriophage promoter present in the respective plasmids (see Appendix VIII), whereas translation was achieved with eukaryotic signals in the rabbit reticulocyte lysate, which was supplemented with ^{35}S-methionine. Both constructs were used in supercoiled and linear conformations. The pGEM-T vector was linearised using the restriction enzyme *SacI*, whereas the pRSET vector was linearised with *BglII*. The protein products were analysed by SDS-PAGE (see fig 4.4), and visualised by subsequent autoradiography of the dried gel. The protein molecular weight markers are indicated in the margin, a transcription / translation reaction was performed in the absence of DNA (lane 1), and the luciferase protein was translated from a positive control plasmid with a T7 promoter (lane 2). The 61 kDa luciferase protein represents the clear band of highest molecular weight (lane 2), and several lower molecular weight bands (with lower intensities) can be observed in the same lane. Proteins expressed from both the supercoiled and linearised *P.angulosus* clone (see section 3.5) show the same pattern of protein bands (fig 4.4, compare lanes 3 and 4, respectively), however the intensities of the bands differ, indicating that transcription / translation is more efficient from a supercoiled plasmid (lane 3). The protein band of highest molecular weight for the *P.angulosus* clone (lanes 3 and 4) represents a ~ 55 - 57 kDa protein. This correlates both with the predicted size (57 kDa) of the translated clone (see section 3.5), and previous estimates of 59.5 kDa for native suGF1 as analysed by SDS-PAGE (191). Several other unique protein bands are present in the region of 36 - 66 kDa (fig 4.4, compare lanes 3 and 4 with lane 2), implying that this clone may code for several protein isoforms which could arise from internal translation start sites present in the cDNA (see section 3.5 and fig 3.15). Using the Peptidesort program in GCG, analysis of the multiple putative start sites coded for by the cDNA reveals that the first ATG start codon correlates with a full length protein of 57 kDa, whereas N-terminally truncated protein isoforms of molecular weights 53 kDa, 45 kDa, 42 kDa and 39 kDa may potentially be formed from alternative ATG start codons further downstream in the cDNA (see fig 3.15). These protein sizes correlate with several bands observed in lane 3 (marked by dots). The marked protein bands occur in a ratio of about 4 : 3 : 1 : 2 :

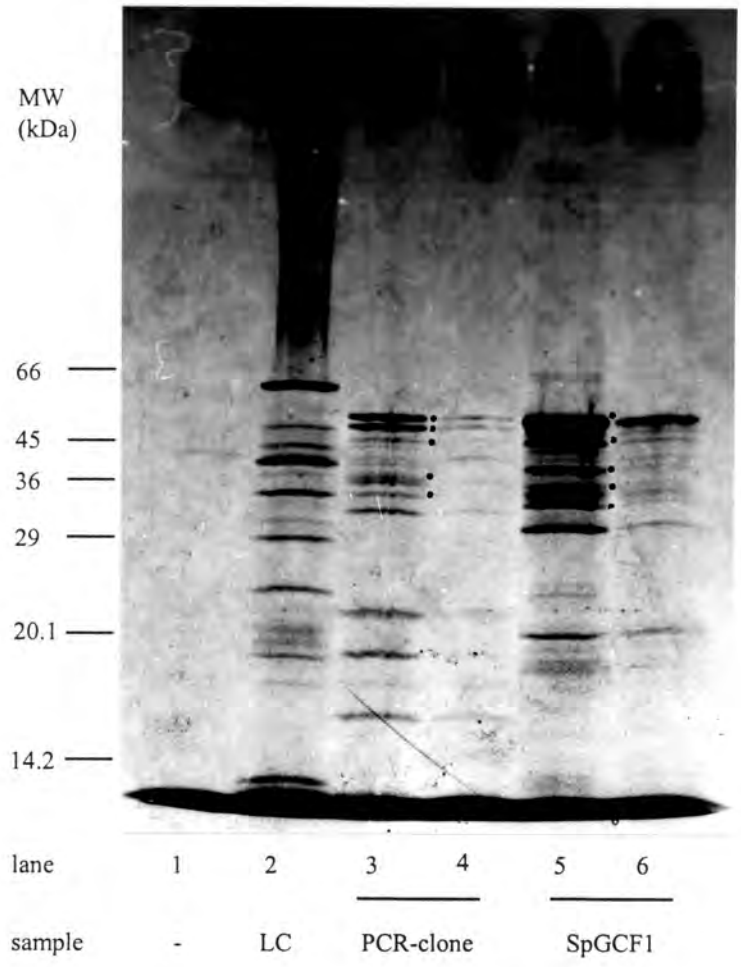


Fig 4.4 Recombinant Proteins from the PCR-Generated Clone and Its *S.purpuratus* homologue were Expressed by *In Vitro* Transcription / Translation

The expressed protein products resulting from *in vitro* transcription / translation of the PCR-generated clone (lanes 3, 4) and its *S.purpuratus homologue* (lanes 5, 6) were analysed by 10 % SDS-PAGE. Expression from supercoiled (lane 3) and linearised (lane 4) plasmid containing the PCR-generated clone, as well as supercoiled (lane 5) and linearised (lane 6) SpGCF1 constructs, resulted in several unique protein products, as marked by the dots. Protein molecular weight standards are indicated in the margin, an incubation without plasmid DNA (lane 1) and the luciferase positive control plasmid (LC) (lane 2) were analysed, too.

2 in order of decreasing molecular weight. Expression of the supercoiled and linearised SpGCF1 constructs was analysed in a similar manner (fig 4.4, lanes 5 and 6 respectively). The highest molecular weight protein obtained for SpGCF1 was also in the vicinity of ~ 55 - 57 kDa, which corresponds to the predicted molecular weight of 55 kDa for this factor (3). Similarly to the *P.angulosus* homologue, several unique protein bands of lower molecular weight are observed for recombinant SpGCF1 (marked by dots in lane 5, fig 4.4). These correspond in molecular weight (viz 55 kDa, 50 kDa, 43 kDa, 40 kDa and 37 kDa) to the multiple protein products observed for SpGCF1 as a result of alternate internal translational start codons present in the cDNA (3).

The *in vitro* generated expression products described above (derived from homologous sea urchin clones) were analysed by standard EMSA conditions (see section 2.21) using the E/H fragment as labelled DNA probe (fig 4.5). Native suGF1 (nuclear extract) exhibited the characteristic doublet (B1 and B2) when bound to the E/H fragment (fig 4.5, lanes 2 and 3). Free DNA (F) is shown in lane 1. Different amounts (1.5 µl and 3.5 µl) of the *in vitro* translated protein from the *P.angulosus* clone were used in EMSAs (fig 4.5, see lanes 4 and 5 respectively) forming several protein-DNA complexes (marked by asterisks). These complexes appear in an approximate ratio of 4 : 3 : 1 : 2 : 2 (in order of increasing mobility) which is reminiscent of the pattern of the protein isoforms discussed above (see fig 4.4, lane 3). Therefore, it appears that the *in vitro* translation system is able to produce several protein isoforms from the *P.angulosus* cDNA. The two protein-DNA complexes of lowest mobility correspond to complexes B1 and B2 formed by native suGF1 protein (fig 4.5, compare lanes 4, 5 with lanes 2, 3), showing that the PCR-generated *P.angulosus* cDNA clone is a likely candidate coding for the suGF1 protein. Complexes C2 and C3 present in nuclear extract (fig 4.5, lanes 2 and 3) are two of three complexes which have been observed in nuclear extract preparations on numerous occasions, apart from the specific complexes B1 and B2 (see section 5.2.3.2). Three higher mobility complexes are also observed in the *in vitro* translated protein sample (fig 4.5, lanes 4 and 5), and two of these may correspond to complexes C2 and C3 in nuclear extract, as they have very similar mobilities. This implies that *in vivo*, suGF1 may be expressed as a transcription factor with several isoforms which differ in their N-termini. The *in vitro* generated protein forms virtually identical complexes to native suGF1, indicating that the native protein probably does not undergo post-translational modifications. The SpGCF1 protein was also titrated in EMSAs using the E/H fragment as labelled DNA probe (fig 4.5, lanes 6 and 7). Several prominent protein-DNA complexes are formed. They are present in a ratio of 5 : 2 : 1 : 2 in order of increasing mobility and appear to correspond to the protein bands observed by SDS-PAGE analysis (see discussion above and fig 4.4, lanes 5 and 6). The protein-DNA complex formed by the full length SpGCF1 protein (marked by an arrow in fig 4.5, lane 7) corresponds in mobility to the doublet (B1 and B2) formed by native suGF1,

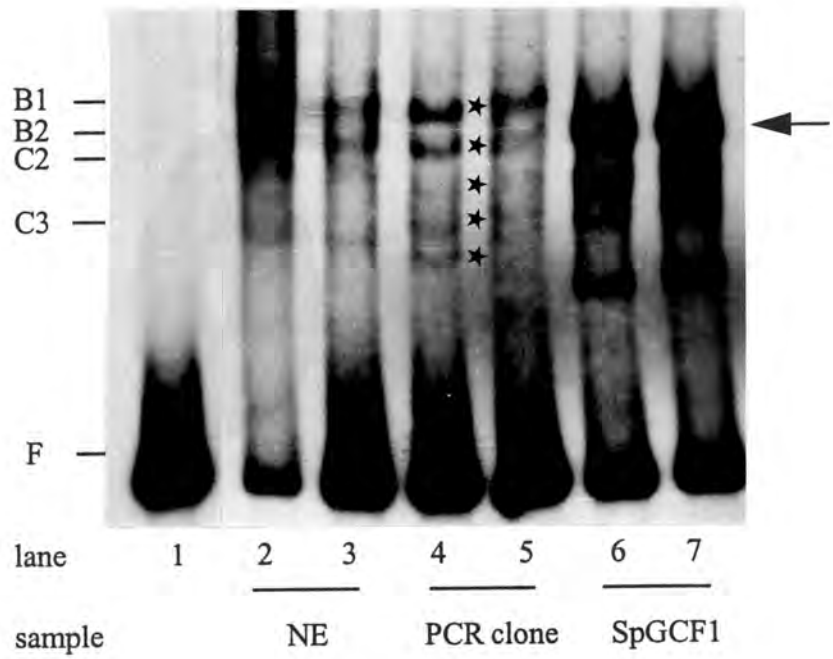


Fig 4.5 Expression Products from the PCR-Generated cDNA Clone and Its *S.purpuratus* Homologue Form Multiple Protein-DNA Complexes

In vitro translated expression products from the PCR-generated clone (lanes 4, 5) form several distinct protein-DNA complexes when analysed by EMSA (marked by the asterisks). Similarly, the expression products from the *S.purpuratus* homologue (lanes 6, 7) also generate multiple protein-DNA complexes (the lowest mobility complex is marked by an arrow). Native suGF1 (lanes 2, 3) from nuclear extract (NE) forms a characteristic doublet (B1 and B2), and other complexes, C2 and C3, are commonly observed in nuclear extract, too. F is free labelled DNA (lane 1).

however it does not form a doublet. Three higher mobility protein-DNA complexes are also observed for the SpGCF1 translated protein, which probably correspond to truncated isoforms of the SpGCF1 protein as a result of internal translation start sites in the cDNA (3). Two of the higher mobility protein-DNA complexes of SpGCF1 (fig 4.5, lanes 6 and 7) correspond in mobility to some of the protein-DNA complexes formed by the homologous *P.angulosus* recombinant protein, however the highest mobility protein-DNA complexes of the homologous clones differ substantially in mobility (fig 4.5, compare lanes 4, 5 with lanes 6, 7). Another obvious difference between the two homologous recombinant proteins is that their isoforms are not expressed in the same ratios, which is reflected by both SDS-PAGE analysis (fig 4.4, compare lanes 3 and 5), as well as EMSA (fig 4.5, compare lanes 4 and 6).

Clearly, the pattern of protein-DNA complexes formed by the *P.angulosus* recombinant protein is virtually identical to the protein-DNA complexes (B1 and B2) formed by native suGF1 (fig 4.5, compare lanes 2, 3 with lanes 4, 5). Therefore, it is appears that the PCR-based cloning strategy has resulted in the successful isolation of a *P.angulosus* clone encoding the suGF1 protein, which is likely to be an orthologue of SpGCF1 from *S.purpuratus*.

4.4 DNA-Binding Analysis of *In Vitro* Expressed Truncated Proteins from the PCR-Generated Clone

DNA fragments, coding for different sized truncated protein products, were amplified from the full length PCR-generated clone and cloned into the pGEM-T vector (see Appendix VIII). The recombinant plasmids were used to express the truncated proteins by *in vitro* transcription / translation in the presence of ^{35}S-methionine (see section 2.16), and the protein products were analysed by SDS-PAGE and subsequent autoradiography (see fig 4.6). Primers SP3/SP4 (see fig 3.15) amplified a 963 bp DNA product, encoding a protein of ~31 kDa (fig 4.6, lane 3), whereas primer SP1 in conjunction with the pGEM-T vector primer SP6 (see Appendix VIII) amplified a DNA fragment of 1.2 kb encoding a ~ 41 kDa protein (fig 4.6, lane 4). The DNA fragments code for single unique protein products, as shown by the arrows (fig 4.6). These protein products differ mainly in the length of their C-terminus (see fig 3.15). *In vitro* transcription / translation control reactions were performed in the absence of plasmid DNA (fig 4.6, lane 1), and using the luciferase control plasmid provided with the system, which codes for the 61 kDa luciferase protein (lane 2). Standard protein molecular weight markers are indicated in the margin.

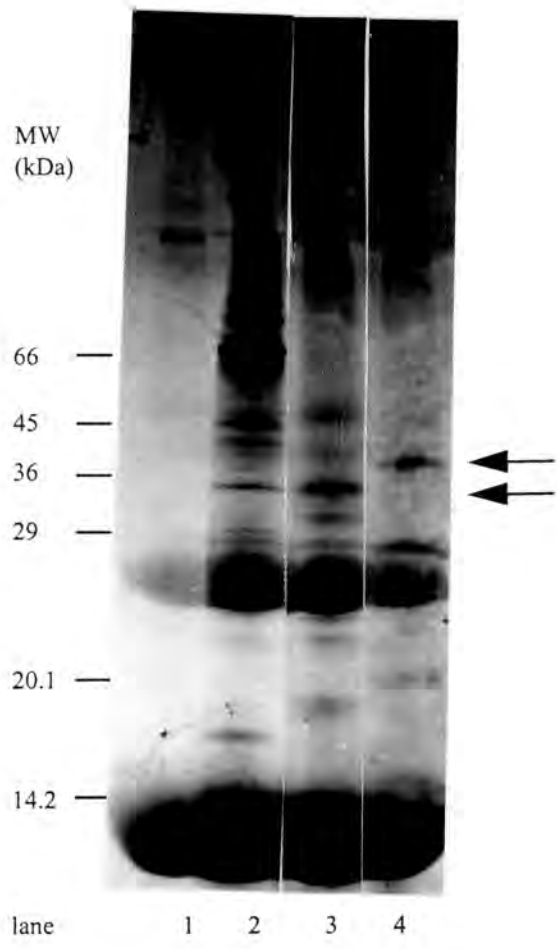
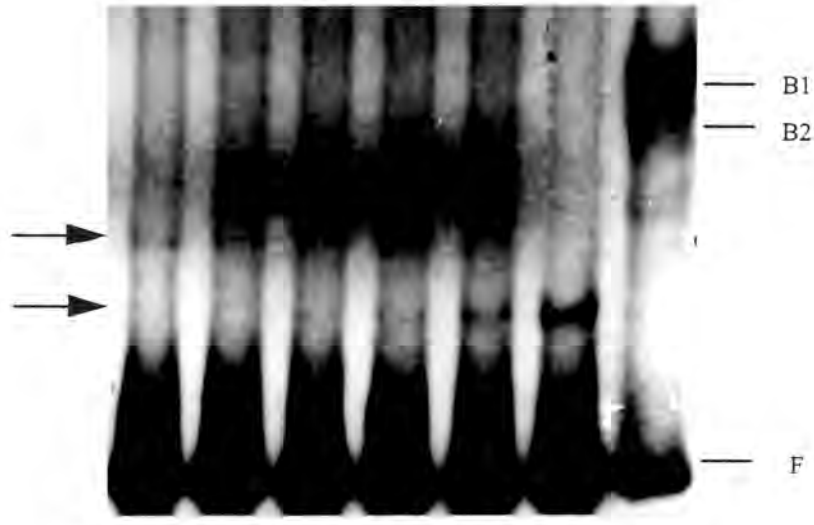


Fig 4.6 Individual DNA Fragments Amplified from the PCR-Generated Clone Encode Two Different Size Truncated Proteins

The PCR-generated clone was used as a template in separate PCR reactions to generate a 963 bp fragment using gene specific primers SP3 and SP4, and a 1.2 kb fragment using gene specific primer SP1 and the pGEM-T plasmid primer Sp6. The DNA fragments were cloned into the pGEM-T plasmid (see Appendix VIII) and proteins were expressed from them by *in vitro* transcription / translation. The *in vitro* expression products (5 µl) were analysed by 12 % SDS-PAGE and subsequent autoradiography. The 963 bp fragment yielded a protein of ~ 31 kDa (lane 3), whereas the 1.2 kb resulted in a protein of ~ 41 kDa (lane 4) as indicated by the arrows. Control reactions included an incubation in the absence of plasmid DNA (lane 1), and translation of the 61 kDa luciferase protein (lane 2) from the control plasmid provided in the system.

EMSAs of the individual truncated proteins show that both the 41 kDa and the 31 kDa products are able to form distinct single lower mobility protein-DNA complexes with the E/H fragment (fig 4.7, lanes 1, 2 and lane 6 respectively) as indicated by the arrows. This suggests that the DNA-binding domain of the PCR-generated clone is located centrally in the protein. DNA-binding analysis of a combination of both truncated protein products (fig 4.7, lanes 3 - 5) results in the formation of two distinct protein-DNA complexes corresponding in mobility to the protein-DNA complexes formed by the individual protein products (compare lanes 3 - 5 with lanes 1 and 6). The absence of an intermediate protein-DNA complex suggests that the two different sized proteins do not associate with each other in this assay, and are able to bind DNA as monomers, ie this protein does not homodimerise in order to recognise the DNA-binding site.



lane	1	2	3	4	5	6	7
41 kDa (µl)	5	4	4.5	4	3	-	NE
31 kDa (µl)	-	-	0.5	1	2	1	NE

Fig 4.7 <u>The Protein Encoded by the PCR-Generated Clone Has a Centrally Located DNA-Binding Domain and Recognises DNA as a Monomer</u>

In vitro expressed truncated protein products (see fig 4.6) were analysed by EMSA with 1 ng of E/H fragment. Reactions performed with either protein individually show that both the 41 kDa protein (lanes 1 and 2, 4 µl and 5µl of protein respectively) and the 31 kDa protein (lane 5, 1 µl protein) form single protein-DNA complexes with distinct mobilities, as indicated by the arrows. When analysed together (lanes 2 - 4, see individual lanes for amounts of respective proteins used), the truncated proteins retained their individual protein-DNA complexes without forming a protein-DNA complex of intermediate mobility. Nuclear extract (lane 7, NE) forms the characteristic suGF1-DNA doublet (B1 and B2). F is free labelled DNA.

CHAPTER 5

PURIFICATION AND SEQUENCING OF suGF1

5.1 Introduction

The characterisation of sequence specific DNA-binding proteins and ultimately the identification of genes which code for them, is hampered by the rarity of these factors in biological samples, which bears several implications on the purification strategy, since enormous amounts of starting material are required if the pure protein is used for techniques such as sequencing. Therefore the protein purification strategy needs to be highly selective, and should have as high a yield as possible. The purification of suGF1 followed a general approach for transcription factors, whereby the crude nuclear protein extract was partially purified using conventional chromatography, enabling the removal of nucleases and other contaminants in the sample. This procedure was combined with a high enrichment affinity chromatography step, which required pretreatment of the protein sample with a nonspecific competitor DNA to which suGF1 has very low affinity. Specifically bound suGF1 was eluted by increased salt concentration, followed by separation of the protein using SDS-PAGE. This sufficed to purify suGF1 for protein sequencing. Several other DNA-binding proteins have been isolated to homogeneity using similar procedures, examples include SpP3A2 (120), SpOct (137) and SpGCF1 (3).

Identification of suGF1 was achieved by analysis of its primary structure using the mass spectrometer approach, which provides detailed primary sequence information and identified the protein of interest by correlating the experimentally generated spectrum with predicted fragment ions from a protein database search (166), thereby generating a search outcome in the form of a ranked list of the highest scoring amino acid identities from the database.

5.2 Isolation of Native suGF1

5.2.1 Electrophoretic Mobility Gel Shift Assays

suGF1 binds sequence-specifically to oligo(dG).oligo(dC) sequences as analysed by EMSA (see section 3.2). This assay was used after each step of the purification procedure to monitor the presence of suGF1. The ease with which the assay can be performed allows many samples to be analysed simultaneously in a rapid and highly sensitive fashion (femtomole quantities of DNA-binding proteins can be detected routinely (151)), which makes the technique particularly suitable for monitoring the purification of this DNA-binding protein. Assaying suGF1 activity by gel shift analysis during the purification was not quantitative, it merely served as a qualitative confirmation that the protein was present. Generally, accurate quantitation of the relative activity of a DNA-binding protein present at each purification step is complicated by several factors. For instance it is known that the optimal binding conditions of a protein can vary considerably during the purification procedure (159), crude and pure protein preparations often differ with respect to their increased sensitivity to oxidative or chemical damage (209) and the pure protein has increased tendencies to adhere to surfaces (159) even in the presence of carrier proteins or detergents.

5.2.2 Lowry Protein Concentration Measurements

Protein concentrations of nuclear extracts were measured by the micro modified Lowry method (210). Using this method, the protein concentration is proportional to the absorbance at λ = 660 nm in the range of 1 - 50 μg, even in the presence of interfering chemicals such as sucrose, Tris, Tricine, glycerol and EDTA, since the method utilises a quantitative precipitation step by combining Na-deoxycholate with TCA and subsequent quantitation by the standard Lowry procedure. The micro modified Lowry method is highly sensitive, but nonspecific (210). Generally, the nuclear extracts measured in this way had a protein concentration ranging from 5 - 15 mg/ml, which agrees well with previous observations for the purification of suGF1 (1) and other DNA binding proteins, such as SpP3A2 (120).

5.2.3 Fractionation of suGF1

Although the suGF1-DNA interaction has been characterised in detail (145, 2), the structure of the protein has remained unknown. Isolating the pure protein is the most direct approach to the

biochemical characterisation of the factor. The development of DNA affinity chromatography makes the purification of transcription factors relatively straightforward, despite the low prevalence of these factors in the cell (92, 156). suGF1 had previously been isolated on a small scale (191), however several changes were introduced to this purification method, including changes to the method of nuclei isolation, reduction in the number of affinity chromatography passes, and isolation of the protein from an SDS gel. An application like protein sequencing sets several constraints (such as volume and buffer composition) on the protein sample (211), and these were taken into account during the purification procedure.

5.2.3.1 Optimisation of Nuclear Protein Extraction

Sea urchin embryo cultures present an ideal system in which large amounts of material can be grown easily in a synchronised fashion. The resultant embryos can (for most purposes) be regarded as cells in suspension. For the purposes of isolating suGF1, sea urchin embryos (*P.angulosus*) were grown for 14 hours, the embryos were harvested and processed for nuclei, from which nuclear proteins were extracted. This served as a starting material in the purification of suGF1.

The isolation of nuclei is a critical step in the preparation of nuclear extracts which contain active DNA-binding proteins. Nuclei should be intact and relatively clean. Several methods have been described whereby nuclei are obtained from intact cells (212, 188, 141). Most commonly these methods involve the isolation of nuclei using either a nonionic detergent, or resuspension of cells in a hypotonic medium combined with homogenisation. Several methods described for the isolation of nuclei from sea urchin embryos involve a variation of the latter method (1, 120, 188). Resuspension of the cells / embryos in a hypotonic medium serves to swell the cells and increases their fragility, which facilitates cell breakage by homogenisation (151). Homogenisation is the most critical step in the preparation of nuclei: too extensive homogenisation will result in the premature rupturing of the nuclei, which may result in leaching of nuclear proteins, whereas insufficient homogenisation may lead to very dirty preparations which have cytoplasmic or other contamination. Hence the conditions of nuclei / nuclear extract preparation were determined for optimum ease of handling large amounts of embryos without compromising the quality of the extracts.

Several methods for isolating nuclei were compared (see section 2.18 for experimental details). Generally, the buffer conditions for all three methods were very similar, and contained a combination of multivalent cations such as Mg^{2+}, Ca^{2+}, spermine and / or spermidine. All buffers contained

chelating agents (either EDTA or EGTA), as well as one or several protease inhibitors. Briefly, the first method (188) involved washing the embryos with 0.5 M KCl, followed by a wash with a hypotonic sucrose solution. The embryos were lysed by homogenisation in an isotonic sucrose solution containing multivalent cations which prevent the swelling of chromatin and hence the premature rupturing of the nuclei. The purified nuclei were collected by centrifugation over a sucrose bed, and resuspended in a low ionic strength buffer. The second method (1) involved a similar procedure, however the embryos were washed with a buffer containing hexylene glycol, the cells were swollen for an extended time in the same buffer, followed by disruption of the cells by homogenisation. The nuclei were purified by centrifugation through a high density sucrose solution. The third method (120) involved washing the embryos with a cold glucose solution, a subsequent wash with an isotonic sucrose solution and then the embryos were frozen. The frozen embryos were crushed before thawing, and the nuclei were collected by centrifugation, followed by 2 - 3 washes with a buffer containing NP-40.

Nuclear extracts were prepared from nuclei by two ammonium sulphate fractionation steps, which firstly involved the preparation of an ammonium sulphate fraction from lysed nuclei and extraction of soluble protein from chromatin by centrifugation. Secondly, the soluble proteins were differentially precipitated by increased salt concentration. Resuspension of the protein pellet was followed by dialysis to remove excess salt and the remaining insoluble proteins were precipitated by centrifugation. The soluble fraction (nuclear extract) contained suGF1. The protein concentration of each nuclear extract preparation was determined by the Lowry protein determination method (see sections 2.20 and 5.2.2), and was within a range of 3 - 15 mg/ml for all three methods used to generate nuclei, indicating that each method yielded comparable total amounts of protein. The quality of the nuclear extracts was assessed by titrating the protein concentration in EMSAs (fig 5.1). Although the latter assessment was not intended as a quantitative measure of the DNA-binding activity present (it seems that other contaminating proteins in the extracts relative to suGF1 affect the amount of sequence specific binding (191)), nevertheless these assays provided an indication of which extraction method yielded the clearest results with respect to suGF1 activity. Nuclei prepared by high speed centrifugation of embryo homogenates through concentrated sucrose solutions generally have higher purity nuclear extracts with distinctive suGF1 binding activity (fig 5.1, lanes 1 - 8) compared to nuclear extracts generated from nuclei released by crudely crushing the embryos (lanes 9 - 12). At higher protein concentrations (4 - 10 μg total protein), suGF1 binding activity is masked by a high molecular weight smear (lanes 1, 2, 5, 6, 9 and 10), whereas at lower concentrations (0.5 - 1.5 μg of total protein) the binding activity shows two characteristic protein-DNA complexes, B1 and B2 (lanes 3, 4, 7 and 8). The nuclear extract generated from the crude nuclei preparation

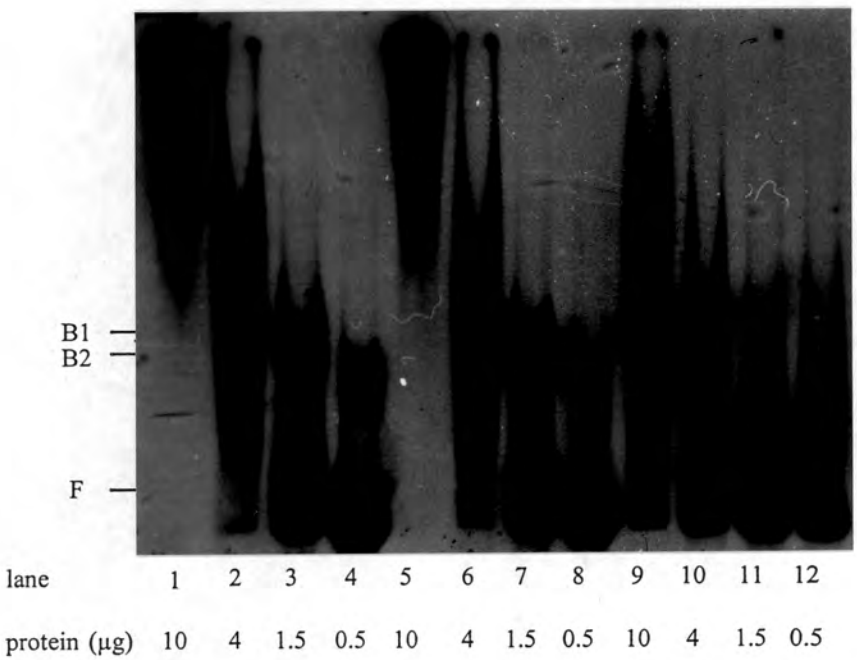


Fig 5.1 Nuclear Proteins Isolated from Three Different Nuclei Preparations were Compared by EMSA

Nuclear extracts from 14 hour *P.angulosus* embryos were generated from nuclei isolated by three different methods (see section 2.18). The protein concentrations in the nuclear extracts (lanes 1 - 4, method by Morris and Marzluff (1983) (188), lanes 5 - 8, modified hexylene glycol method (1), lanes 9 - 10, method by Calzone et al (1991) (120)) were titrated in order to assess the distinct protein-DNA complexes (see individual lanes for final amounts of protein used in each incubation).

appears to have much less suGF1 activity compared to the other preparations (compare lanes 3, 4, 8 and 9 to lanes 11 and 12), which implies that the suGF1 protein was inefficiently extracted from the nuclei, or that the nuclear extract contains an excessive amount of nonspecific inhibitory proteins which were extracted alongside the nuclear proteins.

In 1.3 - 1.8 M sucrose nuclei are sufficiently dense and large enough to sediment under the centrifugal forces which are applied, whereas unbroken cells and subcellular organelles float to the top of the centrifuge tube and form a pellicle - most of the cytosolic material can be removed from the nuclei in this way, leaving very pure nuclei from which a high quality nuclear extract can be derived (see fig 5.1, lanes 1 - 8). It was therefore decided that the extraction of nuclear proteins was best achieved using homogenised nuclei isolated by high speed centrifugation, and as there was no qualitative difference between nuclei isolated in buffers containing sucrose (188) and those containing hexylene glycol (fig 5.1 compare lanes 1, 2, 3 and 4 with lanes 5, 6, 7 and 8), it was decided to use the sucrose isolation procedure, as this was more cost-effective.

5.2.3.2 P11 Phosphocellulose Chromatography

The P11 cation exchange column was chosen as a first step in the chromatography of suGF1 because of its superior yield and enrichment compared to other columns, and because it is able to separate suGF1 from several contaminating activities which bind to the E/H fragment (191).

Sea urchin embryo nuclear extracts from a total of about 120 liters of 14 hour cultures were prepared using the method by Morris and Marzluff (1983) (188) (see section 5.2.3.1 and section 2.18.1). The nuclear extracts were prepared in batches, and each batch was applied individually to a 180 ml P11 phosphocellulose column (see section 2.19). Briefly, the column was washed with 450 ml of 0.1 buffer C, and the proteins were eluted by increasing the ionic strength in a stepwise fashion to 0.3, 0.5 and 0.8 M KCl respectively. Fractions of 15 ml were collected. Protein elution was monitored by absorbance readings at 280 nm against a reagent blank (151) in order to obtain an elution profile (fig 5.2). The profile shows that a large proportion of the protein was not retained by the column at 100 mM KCl, and with each stepwise increase in the KCl concentration more protein was eluted, usually within about 12 fractions (just over one column volume), implying that fractionation took place efficiently. The profile of suGF1 elution (as monitored by EMSA) is shown in fig 5.3. EMSA incubations with 5 µl aliquots of the P11 column fractions were performed at 250 mM KCl in order to decrease the competition from nonspecific / low specific DNA-binding proteins. Generally every fifth

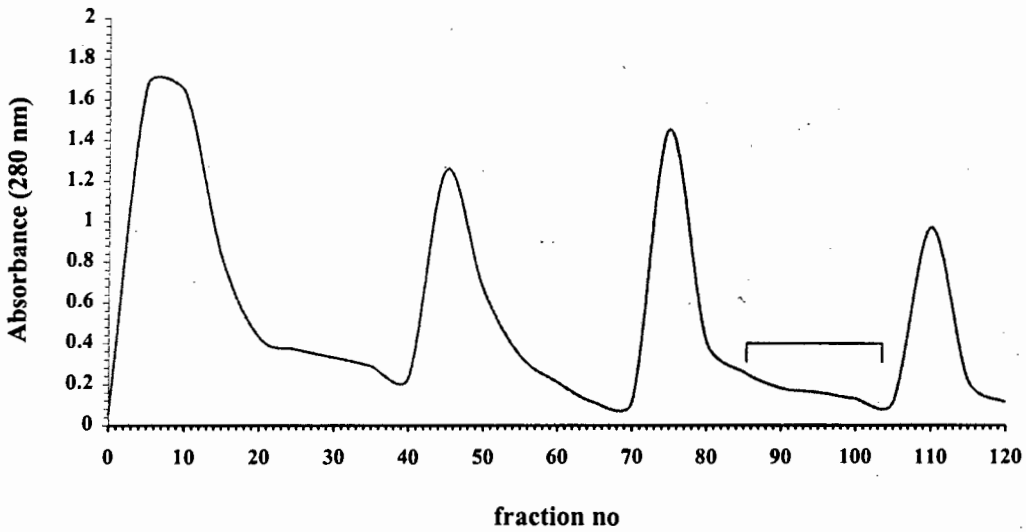


Fig 5.2 P11 Phosphocellulose Protein Elution Profile

Nuclear extracts from 14 hour *P.angulosus* embryos were loaded onto a P11 phosphocellulose ion exchange column (bed volume 180 ml). The column was washed with 2.5 column volumes 0.1 buffer C, and bound proteins were eluted stepwise with increasing KCl concentration, viz 3 column volumes 0.3 buffer C (starting after fraction number 40), 3 column volumes 0.5 buffer C (starting after fraction number 70) and 1 column volume 0.8 buffer C (starting after fraction number 107). The proteins were collected in 15 ml fractions, and the absorbance at 280 nm of each fraction was measured against a reagent blank. An elution profile of absorbance reading at 280 nm (ordinate) vs fraction number (abscissa) was plotted. The fractions containing suGF1 binding activity (fractions 85 - 107) are labelled by the bracket. These fractions were pooled for further purification.

fraction eluted from the P11 column was monitored for suGF1 activity by EMSA. Within the 0.5 buffer C elution range (fractions 72 - 108), where suGF1 was expected to elute (191) every second fraction was tested by EMSA (fig 5.3, lanes 17, 20 - 33). This established which fractions could be pooled for further purification.

A standard nuclear extract incubation (fig 5.3, lane 1) was always analysed by EMSA in parallel to the column fractions in order to gage the specific suGF1-DNA complexes, B1 and B2, in the elution profile. In addition to these specific complexes, generally three other factor-DNA complexes could be distinguished (they are referred to as complexes C1, C2 and C3 (fig 5.3, lane 1)). These complexes have been observed in numerous nuclear extract preparations (J. Hapgood, D. Patterton - private communication, and (191)). It appears that these complexes are fractionated out by P11 phosphocellulose chromatography. Initially it was speculated that these factors either bind the E/H fragment nonspecifically or bind to other sequences in the E/H fragment. Later evidence suggested that some of the higher mobility complexes could represent N-terminally truncated isoforms of suGF1 (see section 6.2.3). Further analysis of the nature of these complexes was, however, not within the scope of this project.

Several features were observed in the EMSA elution profile of the P11 column (fig 5.3). For instance, no suGF1 activity could be detected in the flow-through. A protein exhibiting some DNA-binding activity elutes during the application of 0.3 buffer C to the column (fig 5.3, lane 12). However this protein does not correspond to suGF1, rather it appears to correlate with complex C3. suGF1 elutes in the 0.5 buffer C range, which corresponds to fractions (~72 - 108), and in this region every second fraction was monitored by EMSA. suGF1 activity can first be detected in fraction 85, it increases from fraction 89 onwards, and it continues through to fraction 107 (fig 5.3, lanes 22 - 33), but is no longer present during the elution of 0.8 buffer C (fig 5.3, lanes 34 - 36). Other protein-DNA complexes are formed in the 0.5 buffer C elution range, too. For instance complex C1 can be observed in fractions 91 - 99 (fig 5.3, lanes 25 - 29), whereas complex C2 elutes in fractions 85 - 89 (fig 5.3, lanes 22 - 24). There also appears to be a high molecular weight smear of shifted probe in fractions 83 and 85 (fig 5.3, lanes 21 - 22), as well as an additional complex C4 in fraction 83 (fig 5.3, lane 21). It was decided to pool fractions 89 - 107 (which contain the bulk suGF1 activity), in order to eliminate most of the contaminating factors which can bind the E/H fragment, and yet maximise the amount of suGF1 retained.

Fig 5.2 and fig 5.3 are representative elution profiles obtained for P11 chromatography of suGF1. Fractions from other P11 chromatographic runs were analysed and pooled in a similar fashion, with

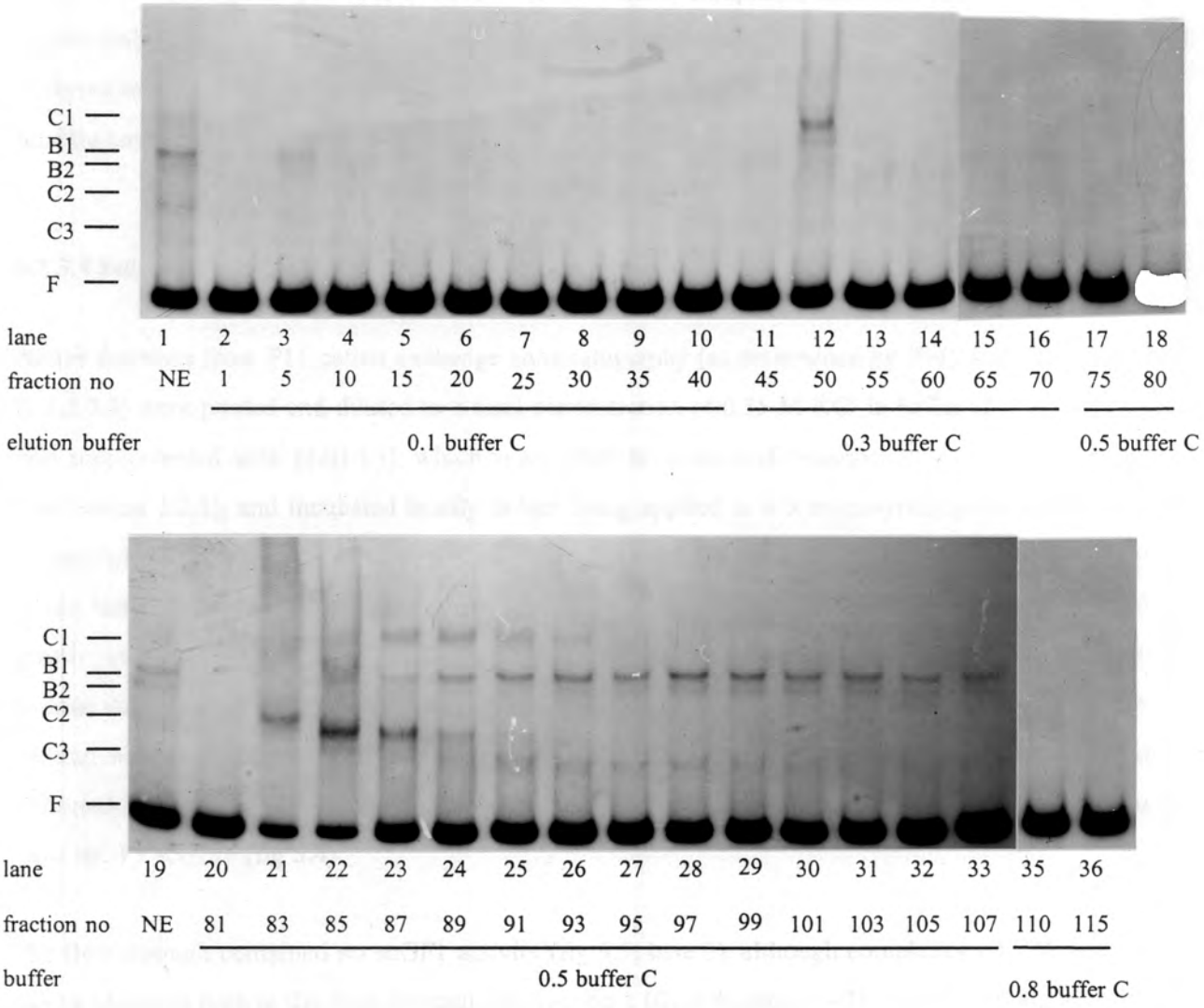


Fig 5.3 Elution Profile of the DNA-Binding Activities from the P11 Phosphocellulose Column

An elution profile of the DNA-binding activities eluted from the P11 cation exchange column was generated by incubating aliquots (5 µl) of every fifth fraction eluted with 0.1, 0.3 and 0.8 buffer C (lanes 2 - 16 and lanes 34, 35) and every second fraction eluted with 0.5 buffer C (lanes 17, 18 and 20 - 33) with 1 ng of labelled E/H fragment in EMSA incubation buffer containing 250 mM KCl. The fraction numbers refer to the same fractions as in fig 5.2. Nuclear extract (NE, 2 µg) is present in lanes 1 and 19. F is free labelled DNA probe, B1 and B2 are suGF1-DNA complexes. C1, C2 and C3 are factor-DNA complexes distinct from B1 and B2.

minor differences in the positions of elution of the various complexes described above (see fig 5.3). Occasionally some fractions would exhibit smearing when analysed by EMSA, these fractions were analysed using a lower concentration of the eluted protein, in order to resolve and identify the DNA-binding complexes.

5.2.3.3 Poly(dG).Poly(dC)-Affinity Chromatography

Active fractions from P11 cation exchange chromatography (as determined by EMSA and described in 5.2.3.2) were pooled and diluted to a final concentration of 0.35 M KCl in buffer C. The solution was supplemented with p[d(I-C)], which is an excellent nonspecific competitor for suGF1-binding (see section 3.2.1), and incubated briefly before being applied to a 9 ml poly(dG).poly(dC)-affinity column which was pre-equilibrated in 0.35 buffer C. Two factors, viz the use of p[d(I-C)] and loading of the sample at high ionic strength, dramatically increase the capacity of the affinity column, and greatly impede nonspecific binding (155). The flow-through (FT) was collected separately, the column was washed with 5 column volumes of 0.35 buffer C and bound proteins were eluted from the column with sequential washes of 0.55, 0.7 and 1.0 buffer C (see section 2.23.2). The eluate was collected in 9 ml fractions and 1 µl aliquots were analysed by mobility gel shift assays in order to trace suGF1 activity (fig 5.4).

The flow-through contained no suGF1 activity (fig 5.5, lane 2), although complexes C1, C2 and C3 can be observed both in the flow-through and fraction 1 (fig 5.4, lanes 2 - 3). These contaminants did not bind to the affinity column. suGF1 was eluted in 0.7 buffer C as shown in fractions 14, 15 and 16 (fig 5.4, lanes 16 - 18). The fractions containing suGF1 activity were pooled for further analysis.

5.2.3.4 TCA Precipitation and SDS Gel Electrophoresis

Active fractions eluted from each single affinity chromatography run were pooled, small aliquots of the pooled fractions were TCA precipitated (see section 2.23.3) and analysed by SDS-PAGE with subsequent silver staining in order to estimate both the amount of protein eluted from each column run and the level of contamination of each pooled sample. A typical example of an aliquot of suGF1 precipitated from a single affinity column is shown in fig 5.5 (lane 5). The amount of suGF1 (~ 50 ng) was compared to several BSA standards (50 - 200 ng) run concurrently on the gel (fig 5.5 (a), lanes 2 - 4). Active fractions eluted from the poly(dG).poly(dC) affinity column still contain minor traces of contaminating proteins which can be illustrated by SDS electrophoresis and subsequent

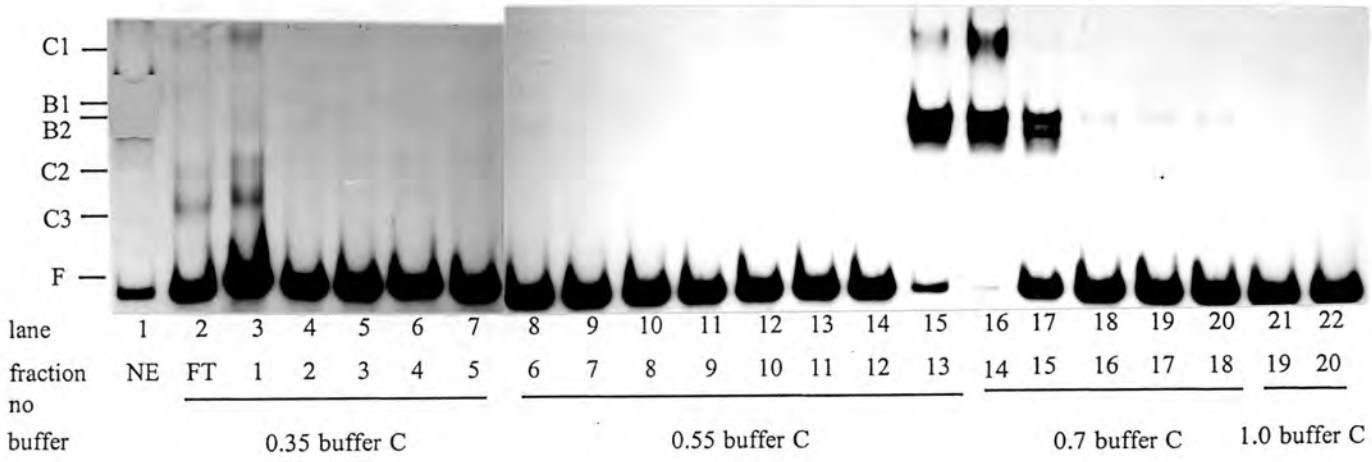


Fig 5.4 Elution Profile of the DNA-Binding Activities from the Poly(dG).Poly(dC)-Affinity Column

An elution profile of the DNA-binding activities eluted from the poly(dG).poly(dC)-affinity column was generated by incubating aliquots (1 μl) of the flow-through (FT) and each fraction collected with 1 ng of labelled E/H fragment in EMSA incubation buffer containing 175 mM KCl. Fraction numbers and elution buffers are indicated. Nuclear extract (NE, 2 μg) is present in lane 1. F is free labelled DNA, B1 and B2 are suGF1-DNA complexes. C1, C2 and C3 refer to the same complexes as in fig 5.3, and are distinct from B1 and B2.

silver staining of larger amounts of the pooled active fractions (data not shown). Proteins contaminating the suGF1 preparation had to be eliminated in order to identify the (partial) amino acid sequence of suGF1. Therefore the final step in the purification procedure involved electrophoretic separation of the pooled, precipitated affinity column fractions containing suGF1 on a denaturing SDS gel. SDS gel electrophoresis is the final chromatographic step in the purification of several DNA-binding proteins, since this provides the protein in a form pure enough for direct use in applications such as sequencing (3, 137, 120).

The active fractions eluted from several individual affinity column runs were pooled and TCA precipitated by incubating the proteins in a final concentration of 20 % (w/v) TCA for 60 minutes on ice. The precipitated proteins were washed first with acidified acetone (0.02 % (v/v) HCl), followed by a wash with pure acetone. The protein pellet was dried and then dissolved in 30 µl SDS sample application buffer containing β-mercaptoethanol. After heat treating the protein sample for 5 minutes at 100°C it was loaded onto a 10 % SDS gel and separated by electrophoresis for 2.5 hours at 180 V. The SDS gel was Coomassie stained in order to visualise the separated proteins. The final purified suGF1 protein product is shown in fig 5.6 (lane 8), and appears to be a single pure band when stained by Coomassie. The amount of suGF1 loaded on the gel was estimated to be 300 ng compared to a series of BSA standards (20 - 500 ng) (fig 5.6, lanes 2 - 7). The Coomassie-stained suGF1 protein band was excised precisely from the acrylamide gel and rinsed twice in water, before further treatment for mass spectrometric analysis (see section 5.3).

5.3 Mass Spectrometric Protein Sequencing

5.3.1 Sample Preparation for Mass Spectral Analysis

suGF1 was excised from a Coomassie-stained SDS gel, and its primary structure was analysed by Dr R.W. Frank (Zentrum für Molekulare Biologie, Universität Heidelberg) using mass spectrometry (see section 5.3.2). The protein sample was prepared for mass spectroscopic analysis as described in section 2.25.1. Briefly, the acrylamide slice containing the excised protein band was cut into small cubes and rinsed with water in order to remove the SDS and to ensure a pH between 6 and 7. The Coomassie dye was removed by repeated addition of 100 µl acetonitrile / water (1 : 1) and incubation for 20 minutes. Excess water was extracted from the gel by addition of pure acetonitrile, which was subsequently replaced by digestion buffer containing 0.5 µg trypsin. Trypsin is a protease which cleaves specifically at the carboxy-terminal side of lysine and arginine, excluding lysine-proline and

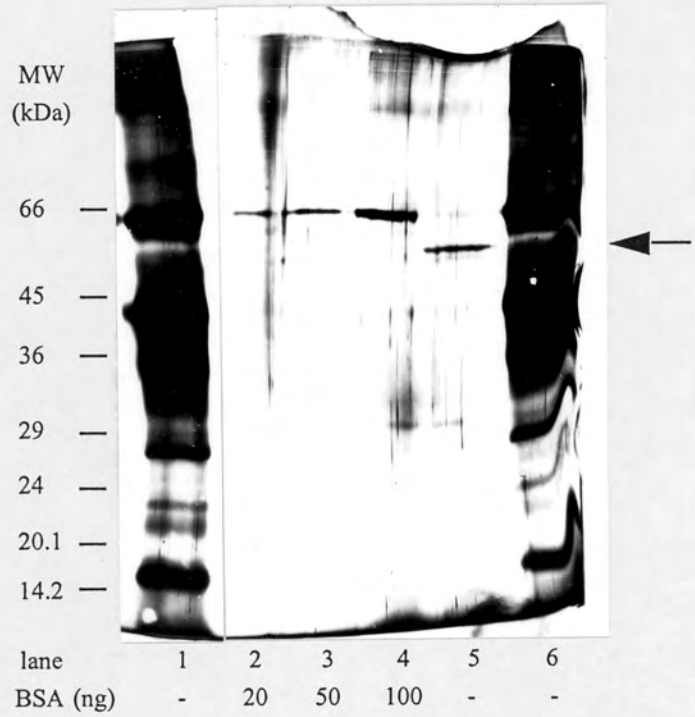


Fig 5.5 Chromatographic Fractionation Results in an Enriched Preparation of suGF1

The suGF1-containing fractions eluted from each poly(dG).poly(dC) column were pooled, precipitated with TCA (20 % final concentration) and resuspended in a small amount of water (10 - 20 µl), an aliquot of which was separated by SDS-PAGE (lane 5) and compared to BSA standards (20 - 100 ng, lanes 2 - 4) in order to estimate the amount of suGF1 eluted from each fractionation. The protein standards are indicated (lanes 1 and 6). The gel was silver stained (see section 2.24).

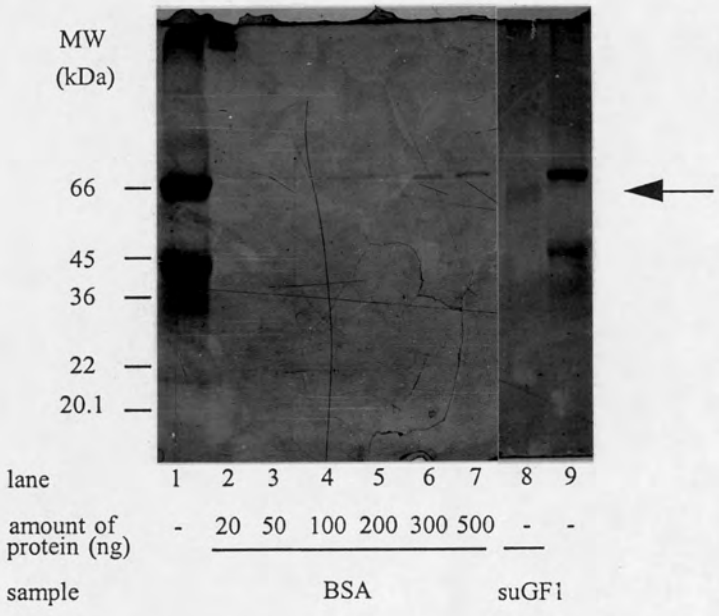


Fig 5.6 Preparative SDS-PAGE of suGF1 Prior to Mass Spectral Sequence Analysis

The pooled, TCA precipitated fractions containing suGF1 (lane 8) were subjected to a final chromatographic step using SDS-PAGE. The gel was Coomassie stained. The amount of suGF1 (~ 300 ng) was estimated by comparison to BSA standards (20 - 500 ng) as indicated in the individual lanes (lanes 2 - 7). The protein was excised from the gel precisely and processed further for mass spectral sequence analysis (see section 2.25). The protein molecular weight standards are indicated in lanes 1 and 9.

arginine-proline linkages. Trypsin is often the preferred protease since it results in complete enzymatic digestion of the protein (due to its high specificity) and it is able to digest insoluble as well as adsorbed proteins (213). The trypsin digestion was allowed to proceed for 6 - 12 hours at 37°C, and the supernatant containing the resultant peptides was recovered. The volume was reduced to 5 µl in a speed-vac.

5.3.2 Mass Spectrometry and Sequence Assignment

The trypsin digested protein sample was analysed by LC-MS and MS/MS (microcolumn liquid chromatography mass spectrometry and tandem mass spectrometry, respectively). The spectra were recorded on a Finnigan LCQ ion trap mass spectrometer equipped with an electrospray (ESI) ion source (214). LC separation of the peptides was performed on a C18 column, and the peptides were eluted by a gradient of acetonitrile (10 - 40 %) in 0.05 % TFA. The separated peptides were on line introduced into the ESI source. Full scan MS spectra were collected continuously, and the production of MS/MS spectra was triggered by peptide signal intensity above a preset threshold of 3.0E4 ions. The signal induced the instrument to isolate the parent ion, which was subsequently fragmented by collision induced dissociation (CID) and the product ions were scanned.

A mass analyser scan of the HPLC gradient identified several of the trypsin generated peptides from the gel slice. The mass spectrometric peptide map is shown in fig 5.7. In total 10 peptides were analysed by MS/MS, three peptides could not be identified, whereas four of the peptides were identified as trypsin (these are marked "TRY" on the mass spectrum, fig 5.7). The other three, marked by an asterisk in fig 5.7, were identified as peptide derivatives of suGF1. These peptide fragments were detected as sequence specific ions in the MS/MS analysis, as shown by the mass spectra of figures 5.8, 5.9 and 5.10 (panel (a)). The spectra were generated by collision induced dissociation of the parent ions to produce daughter ions, and the tandem mass signals observed were generated by recording the mass to charge ratio of the fragment ions (166). The low mass ions are characteristic of the amino acids present in the peptides and the mass difference between the signal of each sequence ion is proportional to the mass of the following amino acid residue in the sequence. The sequence can therefore be interpreted relatively easily by identifying the N-terminal amino acid fragment and counting the mass difference between the subsequent sequence ions. This is illustrated in figures 5.8, 5.9 and 5.10 (panel (b)) where the primary structure of each individual peptide is listed, as interpreted from the mass spectra in panel (a) of the same figures. The three peptides each constitute 18, 18 and 15 amino acid residues respectively, and the MS/MS spectrum for each of these

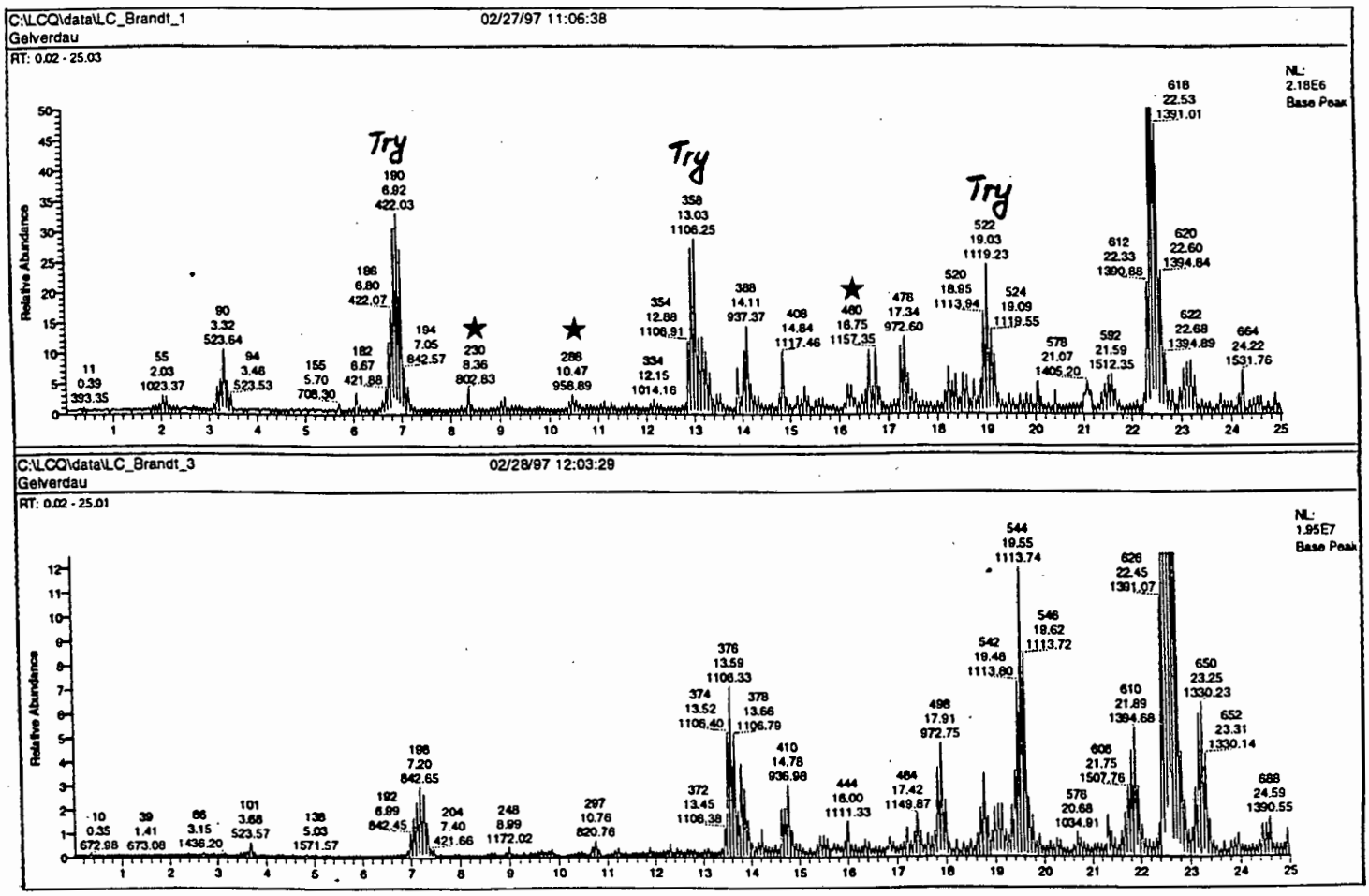
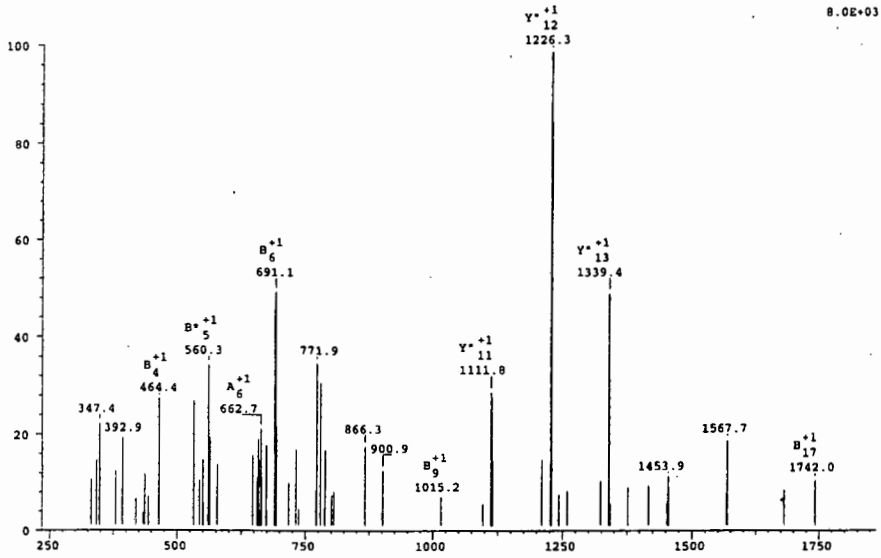


Fig 5.7 Mass Spectrometric Peptide Map Generated by a Tryptic Digest of suGF1

suGF1 was excised precisely from a SDS-gel, the gel slice containing the protein was washed and treated with trypsin to generate peptides. These were isolated by reverse HPLC and a mass analyzer scan of the HPLC gradient revealed four peaks which were subsequently assigned to trypsin peptides (marked "Try" in the scan), as well as three peptide peaks which resulted from trypsin digestion of suGF1 (marked with asterisks).

a



b

Seq Name : None
Composition: C82 H138 N28 O25
Exact Mass : 1915.03906
Bull&Breese: 1190 Charge: 1
Spec1 Name : brandt
Mass Modify: 0.0
N-terminal : Free Amino

Seq Length : 18(0)
Avg Mass : 1916.17273
HPLC Index : 27.7
Spec2 Name : myco
Derivatize : None
C-terminal : Free Acid

No.	Seq	A	B	B*	Bo	Y*	Y*	Yo	No.
1	Val	72.1	100.1	83.0	82.1	1916.0	1899.0	1898.0	18
2	His	209.1	237.1	220.1	219.1	1817.0	1800.0	1799.0	17
3	Asn	323.2	351.2	334.2	333.2	1679.9	1662.9	1661.9	16
4	Leu	436.3*	464.3*	447.2	446.3	1565.9	1548.9	1547.9	15
5	Leu	549.4*	577.3*	560.3*	559.3	1452.8	1435.8	1434.8	14
6	Ile	662.4*	690.4*	673.4*	672.4	1339.7*	1322.7	1321.7	13
7	Asn	776.5	804.5*	787.4*	786.5	1226.6*	1209.6*	1208.6	12
8	Pro	873.5	901.5	884.5	883.5	1112.6*	1095.6	1094.6	11
9	Asn	987.6	1015.6*	998.5	997.6	1015.5*	998.5	997.5	10
10	Ser	1074.6	1102.6	1085.6	1084.6	901.5	884.5	883.5	9
11	Gly	1131.6	1159.6	1142.6	1141.6	814.5	797.4	796.4	8
12	Val	1230.7	1258.7*	1241.7*	1240.7	757.4	740.4	739.4	7
13	Ala	1301.7	1329.7	1312.7	1311.7	658.4*	641.3	640.4	6
14	Gln	1429.8	1457.8	1440.8	1439.8	587.3	570.3	569.3	5
15	Asn	1543.8	1571.8	1554.8	1553.8	459.3	442.2	441.3	4
16	Val	1642.9	1670.9	1653.9	1652.9	345.2	328.2	327.2	3
17	Ala	1713.9	1741.9*	1724.9	1723.9	246.2	229.1	228.1	2
18	Arg	1870.0	1898.0	1881.0	1880.0	175.1	158.1	157.1	1

c

Feb 28 1996 16:37 lc_brandt_1.0449.0451.2.out

lc_brandt_1.0449.0451.2.out 145 - 162
SEQUEST v.822, Copyright 1993-96
Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates
Licensed to Finnigan MAT
02/28/96, 04:29 PM, 8 min. 4 sec. on icis.zmbh.uni-heidelberg.de
mass=2226.4(+2), fragment tol.=0.00, mass tol.=3.00, AVG
# amino acids = 46028980, # proteins = 147165, # matched peptides = 41400
immonium (HFYWH) = (00000), total inten. = 4676.6, lowest Sp = 25.9
ion series nA nB nY ABCDVWXYZ: 0 1 1 0.5 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
rho=0.200, beta=0.075, top 10, /usr/users/finnigan/database/owl.seq
(M# +16.0) Enzyme:Trypsin

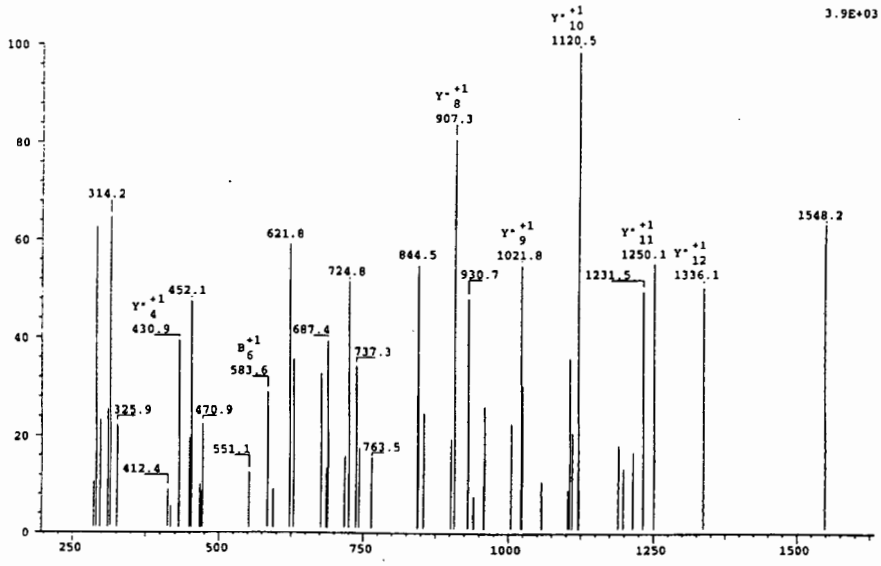
#	Rank/Sp	(M+H)+	Cn	deltCn	C*10^4	Sp	Ions	Reference	Peptide
1.	1 / 1	2227.5	1.0000	0.0000	1.0688	102.3	9/51	SPU18784	(R)QLFYNDVAM#QLYNSDFNK
2.	2 /128	2223.4	0.8554	0.1446	0.9143	36.4	6/60	CEF45H72	(R)EKDVDGDGAVFGVGLDVIFNR
3.	3 / 92	2226.4	0.8444	0.1556	0.9026	38.7	6/60	YNK1_CAEE	(R)SSIPGQTAHATGETDTVRQLR
4.	4 / 43	2226.5	0.7773	0.2227	0.8308	46.7 *	7/54	A55455	(K)TIINQVLSQTQIEFANHLR
5.	5 / 3	2228.5	0.7766	0.2234	0.8301	73.1	8/51	AAR2_YEAS	(K)N##EEKDGAKFENIVHNFK
6.	6 /373	2226.6	0.7569	0.2431	0.8090	27.9	6/63	UL62_HCMV	(R)TPPGKGGKRVEPTGTCVRAGEK
7.	7 /204	2225.5	0.7448	0.2552	0.7960	32.7	6/57	ITAL_HUMA	(K)NCGEDKKCEANLRVSFSPAR
8.	8 / 34	2223.6	0.7405	0.2595	0.7915	48.8	7/54	HH61HMEAR+1	(K)KCSMGKQNLESNRFCQGLR
9.	9 /170	2228.5	0.7193	0.2807	0.7689	33.9	7/54	YLP3_CAEE	(K)TEEEFKAALKEKSPSLYER
10.	10 /155	2226.5	0.6955	0.3045	0.7434	34.9	6/54	HELI_HSVE	(R)LFESTASRGEFENIRALER

1. SPU18784 SPU18784 NID: g755247 - purple urchin.
2. CEF45H72 CEF45H7 NID: g508088 - Caenorhabditis elegans.
3. YNK1_CAEEL HYPOTHETICAL 98.3 KD PROTEIN R10E12.1 IN CHROMOSOME III. - CAENORHABDITIS ELEGANS.
4. A55455 steroidogenic acute regulatory protein precursor, mitochondrial - mouse
5. AAR2_YEAST A1 CISTRON SPLICING FACTOR AAR2. - SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
6. UL62_HCMVA HYPOTHETICAL PROTEIN UL62. - HUMAN CYTOMEGALOVIRUS (STRAIN AD169).
7. ITAL_HUMAN LEUKOCYTE ADHESION GLYCOPROTEIN LFA-1 ALPHA CHAIN PRECURSOR (LEUKOCYTE FUNCTION ASSOCIATED MOLECULE 1, ALPHA CHAIN) (CD11A) (INTEGRIN ALPHA- L). - HOMO SAPIENS (HUMAN).
8. HH61HMEARH26 HH6IMMEARH NID: g451932 - Human herpesvirus 6 (strain U1102) DNA.
9. YLP3_CAEEL HYPOTHETICAL 20.0 KD PROTEIN F02A9.3 IN CHROMOSOME III. - CAENORHABDITIS ELEGANS.
10. HELI_HSVEB PROBABLE HELICASE. - EQUINE HERPESVIRUS TYPE 1 (STRAIN AB4P). (EHV-1).

Fig 5.8 Mass Spectral Analysis of Peptide 1

(a) The tandem mass spectrum for the peptide was generated by recording the mass to charge ratio of the fragment ions. The peaks at low masses are characteristic of the amino acids present in the peptide, whereas the high group masses result from R-group loss from the intact protein. The peaks between the low and the high mass ions result from sequence specific fragmentation of the peptide. The mass difference between the signal of each low mass sequence ion is proportional to the mass of the following amino acid residue in the sequence. (b) The primary structure of the peptide is interpreted and compiled from the fragment ions of the mass spectrum in (a). (c) The outcome of the comparison between the tandem mass spectrum generated by the peptide of interest and the OWL protein database is presented as a ranked list of candidates and the highest priority score is the closest identity to the unknown protein. As can be seen from the highest ranking score, the original linear sequence of the peptide identifies with the SpGCF1 protein from the sea urchin *S. purpuratus* (3).

a



b

Seq Name : None
Composition: C67 H114 N18 O25 S1
Exact Mass : 1602.79224
Bull&Breese: 220 Charge: 1
Spec1 Name : brandt
Mass Modify: 0.0
N-terminal : Free Amino

Seq Length : 15(0)
Avg Mass : 1603.81201
HPLC Index : 6.0
Spec2 Name : myco
Derivatize : None
C-terminal : Free Acid

No.	Seq	A	B	B*	Bo	Y*	Y*	Yo	No.
1	Ile	86.1	114.1	97.1	96.1	1603.8	1586.8	1585.8	15
2	Gly	143.1	171.1	154.1	153.1	1490.7	1473.7	1472.7	14
3	Pro	240.2	268.2*	251.1	250.2	1433.7	1416.7	1415.7	13
4	Ser	327.2	355.2	338.2	337.2	1336.6*	1319.6	1318.6	12
5	Glu	456.2	484.2	467.2	466.2*	1249.6*	1232.6	1231.6	11
6	Val	555.3	583.3*	566.3	565.3	1120.6*	1103.5	1102.6	10
7	Asn	669.4	697.4*	680.3	679.3	1021.5*	1004.5	1003.5*	9
8	Thr	770.4	798.4	781.4	780.4	907.5*	890.4*	889.4	8
9	Asp	885.4	913.4*	896.4	895.4*	806.4	789.4	788.4	7
10	Mso	1032.5	1060.5	1043.4	1042.5	691.4	674.4	673.4	6
11	Leu	1145.6	1173.5	1156.5	1155.5	544.3	527.3	526.3	5
12	Asn	1259.6	1287.6	1270.6	1269.6	431.3*	414.2	413.3	4
13	Gly	1316.6	1344.6	1327.6	1326.6	317.2	300.2	299.2	3
14	Ile	1429.7	1457.7	1440.7	1439.7	260.2	243.2	242.2	2
15	Lys	1557.8	1585.8	1568.8	1567.8	147.1	130.1	129.1	1

c

Feb 28 1996 14:49 lc_brandt_1.0289.0293.2.out

lc_brandt_1.0289.0293.2.out 194 - 211
SEQUEST v.B22, Copyright 1993-96
Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates
Licensed to Finnigan MAT
02/28/96, 02:42 PM, 7 min. 19 sec. on icis.zmbh.uni-heidelberg.de
mass=1916.8(+2), fragment tol.=0.00, mass tol.=3.00, AVG
# amino acids = 46028980, # proteins = 147165, # matched peptides = 41633
immonium (HFYWM) = (00000), total inten. = 6518.8, lowest Sp = 122.2
ion series nA nB nY ABCDVWXYZ: 0 1 1 0.5 1.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
rho=0.200, beta=0.075, top 10, /usr/users/finnigan/database/owl.seq
(M# +16.0) Enzyme:Trypsin

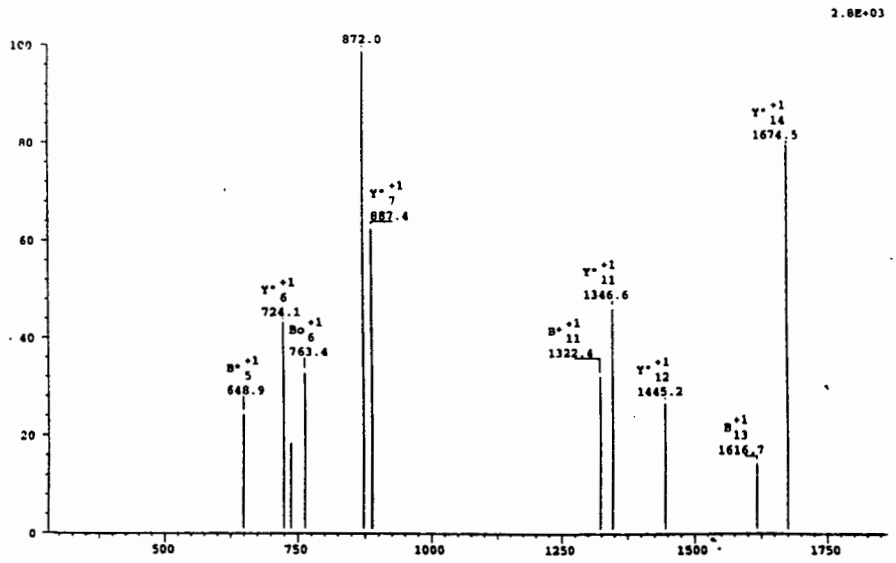
#	Rank/Sp	(M+H)+	Cn	deltCn	C*10^4	Sp	Ions	Reference	Peptide
1.	1 / 2	1917.2	1.0000	0.0000	3.7380	682.3	22/51	SPU18784	(R)VHNLLINPNSGVAQNVAR
2.	2 / 9	1917.3	0.7029	0.2971	2.6274	356.0	16/48	DMU26445	(K)TPKHLLLIDPEGVVRTK
3.	3 / 1	1917.2	0.6933	0.3067	2.5914	704.1	20/48	U006911	(K)FDNSIIDSSM#VNFALVK
4.	4 / 3	1916.5	0.6373	0.3627	2.3822	645.4	22/48	HUMFNRAS	(R)RPPLLPLLLLLLPPPPR
5.	5 / 7	1918.3	0.6058	0.3942	2.2646	393.3	17/51	S59742	(R)M#VNLSLGGDPNLILASNR
6.	6 / 8	1916.2	0.5960	0.4040	2.2280	357.9	16/42	CNA2_YEAS	(K)FLSDNVKINKEYWMK
7.	7 / 5	1917.2	0.5716	0.4284	2.1368	426.8	18/48	BTU16749	(R)RVAHNNLM#AM#AKDIGDK
8.	8 / 11	1916.2	0.5626	0.4374	2.1028	332.5	16/48	THTR_ECOL+1	(R)RGHIPGALNVPWTELVR
9.	9 / 51	1919.3	0.5212	0.4788	1.9481	226.3	13/48	HRDB_STRC+3	(R)GM#LFLDLIQEGNLGLIR
10.	10 / 56	1917.3	0.5163	0.4837	1.9299	214.7	15/48	ALBU_CHIC	(R)MMSNLCSQQDVFSGKIK

1. SPU18784 SPU18784 NID: g755247 - purple urchin.
2. DMU26445 DMU26445 NID: g1079555 - fruit fly.
3. U006911 U00691 NID: g392795 - Dictyostelium discoideum .
4. HUMFNRAS HUMFNRAS NID: g182709 - Human placenta. cDNA to mRNA, clone lambda-P7 [1], and umbilical
5. S59742 S1f1 protein - yeast (Saccharomyces cerevisiae)
6. CNA2_YEAST 1',5'-CYCLIC-NUCLEOTIDE PHOSPHODIESTERASE 2 (EC 3.1.4.17) (PDEASE 2) (HIG H-AFFINITY CAMP PHOSPHODIESTERASE). - SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
7. BTU16749 BTU16749 NID: g567063 - cow.
8. THTR_ECOLI PUTATIVE THIOSULFATE SULFURTRANSFERASE (EC 2.8.1.1) (RHODANESE-LIKE PROTE IN). - ESCHERICHIA COLI.
9. HRDB_STRCO RNA POLYMERASE PRINCIPAL SIGMA FACTOR HRDB. - STREPTOMYCES COELICOLOR.
10. ALBU_CHICK SERUM ALBUMIN PRECURSOR. - GALLUS GALLUS (CHICKEN).

Fig 5.9 Mass Spectral Analysis of Peptide 2

(a) The tandem mass spectrum for the peptide was generated by recording the mass to charge ratio of the fragment ions. The peaks at low masses are characteristic of the amino acids present in the peptide, whereas the high group masses result from R-group loss from the intact protein. The peaks between the low and the high mass ions result from sequence specific fragmentation of the peptide. The mass difference between the signal of each low mass sequence ion is proportional to the mass of the following amino acid residue in the sequence. (b) The primary structure of the peptide is interpreted and compiled from the fragment ions of the mass spectrum in (a). (c) The outcome of the comparison between the tandem mass spectrum generated by the peptide of interest and the OWL protein database is presented as a ranked list of candidates and the highest priority score is the closest identity to the unknown protein. As can be seen from the highest ranking score, the original linear sequence of the peptide identifies with the SpGCF1 protein from the sea urchin *S. purpuratus* (3).

a



b

Seq Name : None
Composition: C100 H144 N24 O32 S1
Exact Mass : 2225.00977
Bull&Breese: -2630 Charge: 1
Spec1 Name : DERIVED SPECTRUM
Mass Modify: 0.0
N-terminal : Free Amino

Seq Length : 18(0)
Avg Mass : 2226.45190
HPLC Index : 64.7
Spec2 Name : tat5223_7
Derivatize : None
C-terminal : Free Acid

No.	Seq	A	B	B*	Bo	Y*	Y°	Yo	No.
1	Gln	101.1	129.1	112.0	111.1	2226.0	2209.0	2208.0	18
2	Leu	214.2	242.2	225.1	224.1	2098.0	2080.9	2079.9	17
3	Phe	361.2	389.2	372.2	371.2	1984.9	1967.8	1966.9	16
4	Tyr	524.3	552.3	535.3	534.3	1837.8	1820.8	1819.8	15
5	Asn	638.3	666.3	649.3*	648.3*	1674.7*	1657.7	1656.7	14
6	Asp	753.4	781.4	764.3*	763.3*	1560.7	1543.7	1542.7	13
7	Val	852.4	880.4	863.4	862.4	1445.7*	1428.6	1427.7	12
8	Ala	923.5	951.5	934.4	933.4	1346.6*	1329.6	1328.6	11
9	Hso	1070.5	1098.5	1081.5	1080.5	1275.6	1258.5	1257.6	10
10	Gln	1198.6	1226.6	1209.5	1208.5	1128.5	1111.5	1110.5	9
11	Leu	1311.6	1339.6	1322.6*	1321.6*	1000.5	983.4	982.5	8
12	Tyr	1474.7	1502.7	1485.7	1484.7	887.4*	870.4*	869.4	7
13	Asn	1588.7	1616.7*	1599.7	1598.7	724.3*	707.3	706.3	6
14	Ser	1675.8*	1703.8	1686.7	1685.8	610.3	593.3	592.3	5
15	Asp	1790.8	1818.8	1801.8	1800.8	523.3	506.2	505.2	4
16	Phe	1937.9	1965.9	1948.8	1947.9	408.2	391.2	390.2	3
17	Asn	2051.9	2079.9	2062.9	2061.9	261.2	244.1	243.1	2
18	Lys	2180.0	2208.0	2191.0	2190.0	147.1	130.1	129.1	1

c

Feb 28 1996 14:08 lc_brandt_1.0231.0235.2.out

369 – 383

lc_brandt_1.0231.0235.2.out
SEQUEST v.B22, Copyright 1993-96
Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates
Licensed to Finnigan MAT
02/28/96, 01:59 PM, 8 min. 34 sec. on icis.zmbh.uni-heidelberg.de
mass=1604.7(+21), fragment tol.=0.00, mass tol.=1.00, AVG
# amino acids = 46028990, # proteins = 147165, # matched peptides = 19140
immonium (HFYWH) = (00000), total inten. = 9599.7, lowest Sp = 108.9
ion series nA nB nY ABCDVWXYZ: 0 1 1 0.5 1.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
rho=0.200, beta=0.075, top 10, /usr/users/finnigan/database/owl.seq
(M8 +16.0) Enzyme:Trypsin

#	Rank/Sp	(M+H)+	Cn	deltCn	C*10^4	Sp	Ions	Reference		Peptide
1.	1 / 9	1604.8	1.0000	0.0000	2.0791	247.8	11/42	SPU18784		(R)IGPSEVNTGM*SLNGIK
2.	2 /180	1603.7	0.7854	0.2146	1.6330	138.1	10/39	MZELIPASE		(K)IPGYMKSSGEEDRR
3.	3 / 29	1603.9	0.7678	0.2322	1.5963	204.1	9/36	JC4093		(R)LQNKVNLM*KELEK
4.	4 /147	1603.9	0.7509	0.2491	1.5613	144.5	9/39	CPNANH4		(R)LGGKIIFLDVKDER
5.	5 / 76	1603.7	0.6924	0.3076	1.4396	169.8	9/33	CEZC5043		(R)YRKVDSRDNRDR
6.	6 / 80	1606.9	0.6856	0.3144	1.4255	169.2	11/42	PUR7_CAMA		(R)ISAYDVIMANGIPMK
7.	7 / 52	1604.9	0.6631	0.3369	1.3786	182.7	10/42	CELT25861		(K)LAELGVFILFDGPGR
8.	8 /110	1603.8	0.6492	0.3508	1.3497	155.0	10/39	PCCA_RAT		(R)QEDIPSCMAVECR
9.	9 /361	1603.8	0.6322	0.3678	1.3145	120.0	9/39	DROSIST		(R)QLLERGYPAATLDR
10.	10 /246	1604.8	0.6237	0.3763	1.2967	131.1	8/39	S21533	+1	(R)DLKPENVFISKSAR
16.	16 / 1	1605.0	0.5965	0.4035	1.2402	469.7	14/39	MOBS_THIF		(R)KILVGALVLDNN#ER

1. SPU18784 SPU18784 NID: g755247 - purple urchin.
2. MZELIPASE MZELIPASE NID: g532622 - Zea mays (strain TX5855) 3-day-old germinating se ed scutellum cDNA
3. JC4093 signal recognition particle receptor alpha chain homolog - Bacillus subtilis
4. CPNANH4 CPNANH NID: g853805 - Clostridium perfringens.
5. CEZC5043 CEZC504 NID: g897712 - Caenorhabditis elegans.
6. PUR7_CANAL PHOSPHORIBOSYLAMIDOIMIDAZOLE-SUCCINOCARBOXAMIDE SYNTHASE (EC 6.3.2.6) (SA ICAR SYNTHETASE). - CANDIDA ALBICANS (YEAST).
7. CELT25861 CELT2586 NID: g1109898 - Caenorhabditis elegans strain=Bristol N2.
8. PCCA_RAT PROPIONYL-COA CARBOXYLASE ALPHA CHAIN PRECURSOR (EC 6.4.1.3) (PCCASE) (PROP ANOYL-COA:CARBON DIOXIDE LIGASE) (FRAGMENT). - RATTUS NORVEGICUS (RAT).
9. DROSIST DROSIST NID: g401734 - Drosophila melanogaster (strain Oregon R) DNA.
10. S21533 protein kinase PKN2 (EC 2.7.1.-) - Myxococcus xanthus

Fig 5.10 Mass Spectral Analysis of Peptide 3

(a) The tandem mass spectrum for the peptide was generated by recording the mass to charge ratio of the fragment ions. The peaks at low masses are characteristic of the amino acids present in the peptide, whereas the high group masses result from R-group loss from the intact protein. The peaks between the low and the high mass ions result from sequence specific fragmentation of the peptide. The mass difference between the signal of each low mass sequence ion is proportional to the mass of the following amino acid residue in the sequence. (b) The primary structure of the peptide is interpreted and compiled from the fragment ions of the mass spectrum in (a). (c) The outcome of the comparison between the tandem mass spectrum generated by the peptide of interest and the OWL protein database is presented as a ranked list of candidates and the highest priority score is the closest identity to the unknown protein. As can be seen from the highest ranking score, the original linear sequence of the peptide identifies with the SpGCF1 protein from the sea urchin *S. purpuratus* (3).

peptides was subjected to a database comparison by computer analysis. The data were cross-correlated using the "Sequest" program package (166) in conjunction with the OWL database (147 000 entries). This is a development whereby the uninterpreted tandem mass spectra of peptides can be correlated with amino acid sequences in a database. The protein database is searched for linear amino acid sequences whose predicted mass-to-charge ratios correlate within a certain mass tolerance to the fragment ions observed experimental data of the tandem mass spectrum (166). This approach conveniently interprets the tandem mass spectrum with respect to known sequences in a protein database. The result of the comparison between the tandem mass spectra (generated by the protein peptides of interest) and the OWL protein database is shown in figures 5.8, 5.9 and 5.10 (panel (c)). The outcome of the database comparison is presented as a ranked list of candidates, and the highest priority score is the closest identity to the unknown protein. The database comparison confirms that the three peptides which originate from suGF1 unambiguously identify with the single protein SpGCF1, which is a transcription factor isolated from the sea urchin *Strongylocentrotus purpuratus* (3 and see section 1.4.4.2)). The suGF1 peptides, whose identities were established by mass spectrometry, were aligned with the protein sequence of SpGCF1, as illustrated in fig 5.11, showing that the peptides from suGF1 correlate with the amino acid sequence of SpGCF1 (3). In addition, a comparison of the amino acid sequences of these peptides shows full agreement with the amino acid sequence deduced for the cDNA clones isolated by means of PCR-RACE from *P.angulosus* mRNA (see section 3.5). These results not only confirm that the protein suGF1, isolated from *Parechinus angulosus*, is an orthologue of SpGCF1 (as speculated in Chapter 3), it also confirms that the cDNA generated by the PCR strategy (see section 3.5) codes for the transcription factor suGF1.

Fig 5.11 Alignment of the Three suGF1 Peptides Identified by Mass Spectral Analysis with SpGCF1

The amino acid sequences of three peptides originating from suGF1 were identified by MS/MS spectral analysis (see figures 5.8, 5.9 and 5.10). Alignment of the peptide primary structures with the corresponding amino acid sequence of SpGCF1 from *Strongylocentrotus purpuratus* (3), using the programme "Bestfit" (GCG), shows that the two proteins have identical sequences in these regions.

CHAPTER 6

DISCUSSION

6.1 Cloning the cDNA for suGF1

6.1.1 DNA Binding Properties of Native suGF1

The ability of a transcription factor to distinguish between its cognate DNA-binding site and an unrelated sequence is an important requirement for the identification of clones using the DNA ligand screening approach (see fig 3.5, and (170)). suGF1 discriminates specifically between its cognate G·C-rich DNA binding site and random mutations thereof. The results of the Scatchard analysis (see section 3.2) show that suGF1 in nuclear extract exhibits a very low dissociation constant ($K_d \sim 3.6 \times 10^{-10}$ M) with respect to the G_{11}-string in the H1-H4 intergenic fragment. The DNA-binding properties of suGF1 (as established by analysis with EMSA) satisfy the criteria for this transcription factor to be identified using the DNA ligand screening approach, since the expressed protein binds the recognition site with high specificity and should be able to withstand the wash protocol without a loss in signal as a result of its low dissociation constant. In addition, the identification of sequence-specific suGF1-DNA complexes is selectively enhanced by using the nonspecific DNA competitor poly[d(I-C)], the use of which in the probe solution should reduce both nonspecific protein-DNA interactions and the overall background.

Mobility gel shift assays performed with several competitor DNAs show that suGF1 has similar affinity for the native DNA binding site present in the E/H fragment and the specific oligo containing the G_{11}-string (see section 3.2), since both these DNA probes are able to compete for the formation of the suGF1-DNA complex when present in very low concentrations. In contrast, suGF1 has no affinity for random DNA sequences, as can be seen by the competition EMSA performed with the nonspecific oligo. The high affinity site oligonucleotide (specific oligo) is therefore an appropriate recognition site probe for the DNA ligand screening method.

It is unclear whether the equilibrium and kinetic constants of a protein-DNA interaction are the same in solution as they are for the binding of a DNA probe to a matrix of protein immobilised on a filter, however it may be possible to isolate recombinants encoding proteins with binding constants of 10^{-9} M or lower (174). Even though the functional screening of expression libraries using a DNA ligand is not restricted to a particular subclass of DNA-binding domains, the successful screening of recombinants may be restricted to proteins with relatively strong binding constants, since only these have the ability to form complexes with half-lives long enough to withstand the 30 minute wash protocol (174).

Poly[d(I-C)] is an excellent nonspecific competitor which can be used in combination with suGF1-DNA interactions to reduce the formation of nonspecific protein-DNA complexes, since even relatively high amounts of this competitor do not compete for the formation of the specific suGF1-DNA complexes (see fig 3.4). This is in contrast to other nonspecific competitors, such as sonicated calf thymus DNA and *E.coli* DNA, which compete with high affinity for the formation of suGF1-DNA complexes at substantially lower concentrations. Therefore it was advisable to use poly[d(I-C)] as nonspecific competitor in the DNA ligand screening procedure.

6.1.2 Analysis of the cDNA Clones Generated by DNA Ligand Screening

The DNA ligand screening technique yielded four unique putative positive bacteriophage plaques, whose inserts were excised in the pBluescript phagemid. This facilitated subcloning, restriction analysis and DNA sequencing of the four clones. The inserts were also analysed for their ability to express recombinant protein products (see section 6.2). A total of 6 X 10^5 phage in the cDNA library were screened, which yielded 19 positive phage in the first round of screening, and 4 remained positive after the second round of screening for specific DNA-binding signals. The DNA sequences and their respective translated amino acid sequences were subjected to database homology searches using either FASTA (GCG) or the BLAST search engines. The characteristic features of the clones and the most interesting homology scores are listed in table 3.3. The homology scores implied that most of the clones (viz Clones 6, 11 and 16) are novel sequences which may potentially encode DNA-binding proteins. Some of the clones had partial open reading frames. Sequencing errors may have masked the presence of continuous open reading frames, as these clones were mainly sequenced manually. Since the identification of the clones relied on the sole requirement that the expressed proteins identify the target DNA-binding site with high specificity, it could be postulated that the isolated clones share a certain structural / functional motif with each other and possibly with suGF1.

It was therefore pertinent to gain further insight into these clones by analysing their expressed protein products (see Chapter 4 and section 6.2).

The cDNAs isolated by the DNA ligand screening procedure were not suGF1 clones. The reasons for the lack of success with the screening method are discussed below. A transcription factor like suGF1 represents a very small fraction of the total target RNA, therefore it is fundamentally important for the recombinant library to completely represent the original mRNA population in terms of sequence, size and complexity (ie it must contain at least one cDNA clone representing each mRNA in the cell). The abundance of the mRNA of interest, and the chosen screening method determine the size of the library, which is larger than the number of clones that will be screened (the number of identifiable clones is reduced by at least $1/6^{th}$ in the case of suGF1 since reading frames must be considered in the DNA ligand screening procedure). A representative mammalian cDNA library would contain about 10^6 - 10^7 independent recombinants from 100 ng of cDNA (151). Sea urchins are lower eukaryotic organisms with a less complex mRNA population, therefore the number of independent clones may be less. Thus, a cloning efficiency of 1.3 X 10^6 for the 24 hour *S.purpuratus* cDNA library is indicative of a representative library, and the number of putative positive clones (four) obtained from this library compares favourably with several results outlined in the literature. For instance, Singh et al (1988) (170) screened 2.5 X 10^5 plaques in total, of which two were identified as true positives, Hasegawa et al (1991) (177) screened a total of 6 X 10^5 clones, five of these proved to be independent isolates of identical clones. Didier et al (1988) (167) screened a total of 1 X 10^7 plaques, where the initial screening yielded 34 positive plaques and three remained positive after two rounds of screening. Singh et al (1988) (170) estimate that the frequency of a positive phage clone is about 1 : 10^5, which correlates well with the results obtained in this investigation. The failure to pick up the clone of interest by the ligand screening method may have been due to the absence of the relevant transcript in the cDNA library. However from the size of the library and the considerations above this is unlikely, unless the RNA transcript of interest was lost during amplification of the 24 hour *S.purpuratus* cDNA library. Representation of the SpGCF1 transcript in this library was not certain, as the cDNA encoding SpGCF1 was originally obtained by hybridisation screening a different *S.purpuratus* library derived from 4 hour embryos (3). However the presence or absence of the SpGCF1 cDNA could have been verified by PCR amplification using the degenerate primers (see Appendix VII and section 3.5).

The DNA ligand screening method provides a powerful means for isolating cDNA clones (174). However it has several limitations associated with it. For instance, this technique relies on the expression of functional fusion proteins in bacteria, which implies that the recombinant protein may

not rely on any form of post-translational modification or association with other distinct subunits (the problem of heterodimer formation can only be overcome if the subunit of interest binds the DNA probe with detectable affinity by itself (174)). Furthermore, the protein must be folded in the correct configuration when expressed in the host. The *lac*Z promoter in the λZAP vector drives the expression of inserts as fusion proteins to the amino terminal part of the β–galactosidase gene. Although the carrier protein in this case is usually only 20 - 50 amino acids, it is possible that it could inhibit the correct folding of the protein encoded by the insert, thereby preventing access of the DNA-binding domain to the cognate DNA-binding site. Hence, incorrect folding of recombinant suGF1 or conformational constraints with respect to its DNA-binding domain as a result of fusion to the β-gal protein could potentially explain the unsuccessful cloning of the suGF1 cDNA. It is well known that some clones cannot be expressed in an active configuration in *E.coli* (174), as they may preferentially form insoluble protein aggregates preventing access of the DNA probe. Exposing expressed recombinant proteins to a guanidinium hydrochloride denaturation / renaturation protocol may renature the binding specificity of a fraction of the expressed proteins (168), provided the proteins are amenable to renaturation. It is possible that expressed proteins (such as C/EBP (168)), which were used to develop the conditions for the DNA ligand screening procedure, are unusually amenable to renaturation. Previously, Southwestern analysis of suGF1 indicated that the DNA-binding ability of this protein is easily renatured from the unfolded protein (1), suggesting that suGF1 would be a suitable candidate for the DNA ligand expression screening technique. However, it is possible that recombinant suGF1 in the form of a fusion protein may not be as amenable to renaturation as the native protein. For instance, two recombinant sea urchin proteins (SpGCF1 and SpP3A2), which are known to be expressed in the form of inclusion bodies, could only be restored to about 0.1 % of their DNA-binding activity (3, 120), showing that renaturation of the recombinant protein may be a limiting factor in the detection of its respective cDNA. Our own unsuccessful attempts to express recombinant SpGCF1 in bacterial hosts (see section 4.3) indicate that the clone of interest may not have been expressed in the cDNA library at all.

The bacteriophage expression vector λZAP has a high cloning efficiency, with a cloning capacity of inserts up to 10 kb in length and it directs very high levels of fusion protein expression in recombinant phage (about 1 % of total protein mass) (174). Together with a high specific activity ^{32}P-labelled recognition site probe this implies a high detection limit, which should result in the positive identification of clones encoding DNA-binding proteins by application of the DNA ligand screening procedure. However, another limitation associated with the DNA-ligand screening method pertains to the kinetics of interaction between a DNA ligand and a protein immobilised on a filter, which have not been defined rigorously. Therefore one cannot be sure that all expressed DNA-binding proteins

will be able to withstand the wash protocol in the screening method (174). One would assume that proteins which exhibit strong binding may be suited to this cloning procedure. suGF1 is able to recognise its DNA binding site sequence-specifically and with high affinity, which allows the factor firstly to distinguish between the binding site and a random mutation, and secondly to form a stable interaction with the binding site. Therefore suGF1 should have been suitable for cloning via the DNA ligand screening procedure. The stability of the protein-DNA interaction and the sensitivity of the screening methodology should have been enhanced further by the use of multisite, catenated recognition DNA site probes which can simultaneously interact with several immobilised proteins.

A possible reason for the detection of false positives obtained for suGF1 screening was the presence of p[d(I-C)]. Nonspecific competitor DNA was included in the probe screening solution in order to reduce the background signal from the filters and to block the interaction (and hence minimise the detection) of nonspecific protein with the labelled DNA probe. Binding to either double or single stranded DNA is an undesirable feature of the DNA ligand screening procedure, which may be enhanced by the presence of p[d(I-C)] (174). However it is also possible that the probe itself can undergo structural alterations (eg formation of ss DNA) during processing of the filters (174) which may also result in the detection of false positives.

6.1.3 Analysis of the PCR-Generated Clone

Successful amplification of DNA fragments (106 bp and 421 bp) from *Parechinus angulosus* genomic DNA showed conclusively that a gene homologue for SpGCF1 is present in the *P.angulosus* genome. In addition, PCR amplification of DNA fragments (106 bp, 327 bp and 421 bp) from cDNA showed that RNA for the SpGCF1 homologue is transcribed in 14 hour *P.angulosus* embryos, implying that the protein homologue is expressed in early embryonic stages. A PCR strategy, using a combination of 5' and 3' RACE, was used to amplify two individual fragments representing the full length cDNA coding for the homologue. The correct identity of these fragments was confirmed by Southern analysis and by the high DNA sequence homology to SpGCF1 cDNA (73 % and 92 % for the 5' and 3' RACE fragments respectively). The full length clone (~ 2.1 kb), generated by fusion of the two individual fragments was sequenced automatically, allowing a consensus cDNA sequence to be derived for the clone. The full length homologous cDNAs from *P.angulosus* and *S.purpuratus* showed 84 % homology. The high conservation of this factor amongst the two sea urchin species, together with the fact that the respective proteins are expressed in the developing sea urchin embryos, implies that this transcription factor could have an important function in the regulation of genes during the

development of sea urchin embryos. Further, the primary structure of the *P.angulosus* clone was derived from the single open reading frame (1542 nt) of the cDNA, showing that it codes for a 514 amino acid protein, of molecular weight 57 kDa. This protein is unique, apart from its 94 % homology to SpGCF1.

The successful application of the RACE technique relies on very limited DNA sequence information with respect to the template of interest (such that gene specific primers for both the 5' and 3' RACE reactions can be synthesised) and a source of RNA in which the transcript of interest is definitely present. These requirements were all met by the *P.angulosus* homologue of SpGCF1 as confirmed by the PCR strategy described above. Application of RACE resulted in the formation of multiple products for both reactions, which is not unusual. Possible reasons for multiple products generated by 5' RACE include the presence of alternative transcription start sites, whereas some of the multiple products generated by 3' RACE may stem from different polyadenylation sites. Both the 5' and 3' RACE reactions may generate multiple products arising from alternative splice sites, or the amplification of related genes, which would give rise to several homologous cDNAs. The latter is not a very likely situation for the *P.angulosus* PCR-generated clone, as it was previously found that SpGCF1 is a single copy gene (3). The products generated by the RACE reactions were characterised further in order to distinguish whether they were real or artifactual results. The latter are classified as either incomplete or nonspecific, which can be ascribed to a variety of factors. Artifactual results may arise as a result of premature termination of first strand synthesis, ie pausing of the reverse transcriptase, which causes multiple 5' RACE fragments, or because degraded RNA is used as a starting material (usually this results in multiple 5' products). Other reasons include nonspecific priming during RACE-PCR, as well as high G·C content of the template. The amplification products of interest were identified by Southern hybridisation which revealed that both the 5' and 3' RACE products contained bands of the expected sizes. A larger 3' RACE product (~ 3 kb) was also observed and can probably be attributed to a product with a longer poly-A^+ tail, ie the oligo(dT) priming may have taken place at different mRNA sites during the synthesis of the cDNA. The isolated 5' and 3' RACE products were further characterised by sequence analysis, which proved that the fragments were highly homologous to the SpGCF1 cDNA (~ 73 % for the 5' RACE product and ~ 92 % for the 3' RACE product). This confirmed that the fragments of interest with the correct sizes had been identified. The two separate fragments were then combined to form the full length ~ 2.1 kb cDNA representing the *P.angulosus* homologue of SpGCF1.

The full length, double stranded products were finally amplified and cloned, and three independent clones were sequenced completely (using automated techniques), enabling a consensus sequence to

be derived for the PCR amplified clone. Nucleotide positions which showed sequence degeneracies within the *P.angulosus* cDNA were resequenced from several independent clones to confirm the consensus sequence and eliminate any sequencing errors. The degeneracies may have arisen due to inaccurate nucleotide incorporations as a result of the PCR amplification reactions, since it is well known that Taq polymerase is prone to misincorporations. Alternatively, the differences in the clones could result from differences in the original mRNA templates, which may have arisen as a result of transcription products from different alleles. It appears that SpGCF1 is coded for by a single gene (3), therefore the sequence degeneracy is unlikely due to transcription products resulting from different gene members.

6.2 Recombinant Protein Expression

6.2.1 *In Vivo* Recombinant Protein Expression of cDNA Clones Isolated by the DNA Ligand Screening Technique

The four clones which were isolated by the DNA ligand screening technique (see section 3.3) were not compatible with most of the expression systems employed, as judged by the lack of detectable recombinant protein products in the bacterial pBluescript and pGEX systems, as well as the eukaryotic COS cell expression system. This may be attributed to a variety of reasons. Since none of the recombinant protein products were detectable in either the pBluescript or the pGEX systems, nor the pET system's pLysS strain (which is not protease deficient) it is possible that the target proteins are very protease-sensitive and their stability is easily compromised in a foreign host environment. Both the pBluescript system and the pGEX system should offer enhanced stability to recombinant proteins, since they express target proteins as fusion products of the truncated β-galactosidase protein and the glutathione S-transferase protein respectively. Conditions such as altering the expression host in order to reduce host proteolytic activity did not improve the lack of detection of the recombinant proteins in these systems, implying that other problems could also have affected the expression. For instance, it is possible that the codon usage in these clones is not ideally suited to bacterial expression, or the translational signals were not optimal, or possibly the mRNA was very unstable. Alternatively, mRNA secondary structure formation could have blocked transcription, or detection may have been impeded by very low expression levels or instability of the target protein. Most of these factors can be addressed by the choice of the expression system, as well as the choice of the expression vector. Another impeding factor in the analysis of the four clones was that the cDNA sequence encoding suGF1 (which was ultimately what I was looking for) was unknown, therefore the

identification and characterisation of the clones relied solely on the basis of the DNA-binding activity they displayed. Hence, from the tentative analysis of the DNA sequences of the four clones, the cDNAs could not be characterised fully and it was not obvious whether the cDNAs included long regions of 5' and 3' untranslated sequences, which could pose potential problems with the expression of fusion proteins.

Despite trying to optimise expression of the four clones by investigating them in a variety of expression environments, only a single clone (Clone 11) responded positively to the expression conditions. Clone 11 was expressed as a 25 kDa protein from the pET-29b(+) expression construct in BL21DE3 cells. Optimal expression conditions ranged from 3 - 6 hours at 30°C - 37°C. Isolation of the recombinant protein product was attempted using several procedures (eg soluble protein extraction, Nickel column chromatography and inclusion body isolations), however the protein was generally expressed as an insoluble protein precipitate in the form of inclusion bodies, whose solubilisation proved unsuccessful despite a variety of denaturation conditions using potent denaturing agents.

A variety of systems are available for recombinant protein expression, such as bacteria, yeast, mammalian cells, baculovirus, plants, transgenic animals and eukaryotic *in vitro* transcription / translation systems (151, 205). However the cloning of eukaryotic proteins does not always result in high level expression, and some systems may not result in any expression of desired clones at all (151), therefore one often needs to experiment with recombinant protein expression using a trial and error approach by individually testing a variety of expression systems and optimising their conditions. Several bacterial expression systems differing in their promoter systems were employed to investigate expression of the four clones isolated by the DNA ligand screening approach. These included the pBluescript *lacZ* promoter, the *tac* promoter in the pGEX system, and the T7*lac* promoter system present in pET. The different expression plasmid systems all potentially provided enhanced stability to the target protein, as expression is in the form of fusions to proteins such as the *E.coli* β-galactosidase protein (pBluescript), glutathione S-transferase (pGEX) or a histidine tag (pET-29). Over and above providing enhanced stability of expressed proteins, these fusions should improve the solubility, extraction and purification of expressed target proteins. For optimisation of expression of the four clones (2, 6, 11 and 16) attention was also given to other factors, such as the choice of expression hosts (stringency of host requirement is usually determined by the type of expression vector employed), and expression was investigated under different conditions of temperature, induction time and volume of culture. Further, an eukaryotic expression system (a pCIS expression

construct in COS cells) was also used to analyse recombinant protein expression, however it, too, was unsuccessful.

6.2.2 *In Vitro* Recombinant Expression of cDNA Clones Generated by the DNA Ligand Screening Technique

A rabbit reticulocyte lysate system was used to successfully express ^{35}S-labelled recombinant proteins from Clones 2, 11 and 16. Analysis of the protein products by SDS-PAGE and autoradiography showed that Clone 2 (2.2 kb) codes for a protein of molecular weight about 48 kDa, the protein arising from Clone 11 (0.9 kb) is 25 kDa (which was verified by expression in the pET system) and Clone 16 (2.4 kb) encodes a protein of about 45 kDa. The protein sizes are in approximate agreement with the sizes of the cDNA fragments encoding them. These results strongly suggest that the cDNA clones did contain open reading frames. However they were masked by only partial sequence analysis of the clones, and also by erroneous sequencing (mainly for Clone 16), possibly due to manual sequencing techniques. No unique protein band corresponding to a protein encoded by Clone 6 (0.85 kb) could be identified, implying that either this clone does not have an open reading frame coding for a protein (as implied by the analysis of the cDNA sequence), or it has a very low methionine content, such that insufficient ^{35}S label was incorporated into the protein product, thereby preventing its detection. Analysis of the *in vitro* generated recombinant proteins by EMSA did not exhibit any gel shift activity, suggesting that the proteins were either inactive in this respect, or that the protein translation efficiency was so low that the DNA-binding activity could not be visualised. Low translation efficiency may arise due to low quality of mRNA, RNase contamination, the presence of inhibitors in the reaction, or suboptimal Mg^{2+} concentrations.

6.2.3 *In Vitro* Expression of the PCR-Generated cDNA Clone

The PCR-generated clone was expressed using an eukaryotic *in vitro* transcription / translation system, generating a full length protein of 57 kDa, which correlates in size with the protein predicted to arise from the single long open reading frame of the cDNA, and also with the size of 59.5 kDa for native purified suGF1 previously estimated by SDS-PAGE (1). In addition, analysis of the expressed recombinant protein by EMSA proved to have the same characteristic protein-DNA doublet (complexes B1 and B2) as observed for suGF1, indicating that the PCR generated clone encodes the suGF1 protein. SDS-PAGE analysis of the *in vitro* expressed protein mixture revealed several other unique protein bands of approximate molecular weights 53 kDa, 45 kDa, 42 kDa and 39 kDa, which

may represent several N-terminally truncated protein isoforms arising from alternative internal translation start sites present in the cDNA (see section 3.5). Analysis of the *in vitro* expressed recombinant proteins by EMSA revealed five protein-DNA complexes (present in a similar ratio to the multiple protein products revealed by SDS-PAGE analysis), confirming the presence of several protein isoforms encoded by the cDNA. It is likely that these isoforms result from the degradation of RNA or protein in the transcription / translation reaction, or they could arise as a result of premature termination. However, it is also possible that the *in vitro* production of multiple protein isoforms reflects the *in vivo* situation. Evidence for this is gleaned from the pattern of protein-DNA complexes formed by sea urchin nuclear extracts, which reveal at least two protein DNA-complexes of higher mobility than the suGF1 complexes, B1 and B2. The higher mobility complexes formed by native proteins compare favourably in both mobility and intensity with the higher mobility complexes formed by the *in vitro* translated protein products. *In vitro* translation systems do not provide any form of post-translational modifications, which suggest that native suGF1 is unlikely to be subject to post-translational modifications either, since the *in vitro* expressed suGF1 protein forms virtually identical complexes to native suGF1, as exhibited by formation of the characteristic protein-DNA doublet, and the similarity in both the mobilities and intensities of the respective complexes.

Expression of the homologous SpGCF1 construct was analysed in a similar manner. A full length protein of ~ 55 kDa was generated, which corresponds to the predicted size for this factor (3). Several unique protein bands of lower molecular weight (viz 50 kDa, 43 kDa, 40 kDa and 37 kDa) can be observed for SpGCF1, representing truncated protein products resulting from alternative internal translational start codons present in the cDNA (3). Analysis by EMSA resulted in the formation of several prominent protein-DNA complexes, also present in a similar ratio to the multiple protein bands observed by SDS-PAGE. The single protein-DNA complex formed by the full length SpGCF1 protein corresponds in mobility to the doublet formed by native suGF1. Three higher mobility protein-DNA complexes were observed for the SpGCF1 translated protein, corresponding to truncated isoforms of the SpGCF1 protein. Two of these complexes correspond in mobility to protein-DNA complexes formed by the homologous *P.angulosus* recombinant protein, showing that the homologous proteins are translated in a similar fashion, although analysis by SDS-PAGE and EMSA reflects that the isoforms of the respective homologues are not expressed in the same ratios. In addition, the highest mobility protein-DNA complexes of the homologous clones differ substantially in mobility.

Expression of recombinant suGF1 was achieved by application of eukaryotic *in vitro* coupled transcription / translation in preference to bacterial expression, as the latter was unsuccessful with

both recombinant SpGCF1 and the clones generated by the DNA-ligand expression screening method. This suggested that bacterial expression may not be suited to clones encoding G·C-rich binding factors. In comparison, the *in vitro* translation system resulted in successful expression of most of the clones isolated by the DNA-ligand screening technique, as well as recombinant SpGCF1. Therefore, *in vitro* translation using a cell free protein synthesising system posed an attractive alternative to *in vivo* recombinant expression for the suGF1 clone. Not only did it avoid complicating artifacts related to prokaryotic expression of eukaryotic genes, it also produced sufficient recombinant suGF1 for verification of the clone's open reading frame, and successful identification of recombinant suGF1-DNA interaction as analysed by EMSA.

6.3 Protein Purification of Native suGF1

6.3.1 Purification Strategy

The suGF1 protein was purified to homogeneity on a large scale from *P.angulosus* embryos. The generation of high quality nuclear extracts was the first step in the purification procedure, and it was found that nuclear extracts obtained from nuclei which were isolated by centrifugation through dense sucrose gradients consistently showed least contaminating proteins relative to suGF1, as judged by EMSA. Therefore this was the method of choice when generating the starting material for subsequent fractionation of suGF1. Previous observations showed that contaminating proteins may affect the amount of sequence-specific binding which can be observed in nuclear extracts (191). The isolation procedure for suGF1 followed a general approach outlined for transcription factor purification (96). Nuclear extracts were fractionated by P11 chromatography (cation exchange), which promoted the removal of nucleases and other contaminants in the sample. Conventional ion exchange chromatography was combined with DNA affinity chromatography, which exhibits a very high enrichment for suGF1 (191) and generally for many other sequence-specific DNA-binding proteins (3, 120, 96, 155). The cation exchange purification step exhibited only a 10-fold enrichment (191), however it represented a crucial step in the fractionation of suGF1. Nuclear extracts which were only fractionated by affinity chromatography and SDS-PAGE reflected numerous protein contaminants with respect to suGF1 (data not shown), making it impossible to identify the protein band of interest. Thus, despite proteins only requiring low purity in order to establish their sequence from SDS gels (120), in practice extensive enrichment may be needed, depending on the sample of interest. SDS gel electrophoresis was the final chromatographic step in the purification of suGF1, since it has successfully provided several other DNA-binding proteins in a pure enough form for applications such as sequencing. Examples of DNA-binding proteins which have been isolated by similar

procedures include SpP3A2 (120), SpOct (137) and SpGCF1 (3). In total, about 600 ng of suGF1 was purified by the method outlined above, of which about 300 ng was separated by SDS-PAGE and subjected to further analysis by mass spectrometry.

6.3.2 Identification of suGF1 by Mass Spectral Analysis

Purified suGF1 (~ 300 ng) was excised from a Coomassie-stained SDS gel and subjected to trypsin digestion, which generated peptides by the cleavage of suGF1. The resulting peptides were subjected to tandem mass spectrometry. Three of the ten peptides on the mass spectral peptide map (see fig 5.7) were identified unambiguously as derivatives of the SpGCF1 protein, as can be seen from the ranked list of database homologies in figures 5.8, 5.9 and 5.10. (Four of the ten peptides were identified as derivatives of trypsin, whereas the other three could not be identified.) The three suGF1 peptides had lengths of 18, 18 and 15 amino acids respectively and their primary sequences identified 100 % with the sequence of SpGCF1 over the corresponding primary structure. This confirms that the isolated protein suGF1 is the *P.angulosus* homologue of SpGCF1 from *S.purpuratus*, and that the cDNA sequence of the clone generated by application of RACE (see section 3.5) encodes suGF1.

The development of mass spectrometry with regard to the sequence analysis of proteins and peptides poses several advantages over classical Edman degradation, since it is more sensitive and it enables the sequencing of peptides in mixtures, which reduces the number of manipulations required. This is particularly advantageous for the identification of a transcription factor such as suGF1, as this protein is present in minute quantities *in vivo* and requires vast amounts of biological starting material for purification to homogeneity. Mass spectral analysis substantially reduced the amount of material required for protein sequencing of suGF1, as it only requires picomole to femtomole quantities. The purification strategy for suGF1 involved a combination of selective precipitation, dialysis and several chromatographic manipulations, yielding a purified protein sample available in very low quantities. Adsorptive losses make it difficult to handle low quantities of proteins, therefore it was advantageous using a highly sensitive technique in order to identify the suGF1 protein sequence. Another advantage of mass spectral analysis (in conjunction with computer manipulation) is that the signal patterns resulting from the peptides of interest are correlated directly with the predicted fragment ions from a database. Thus the sequencing data is interpreted and analysed in context with other known protein sequences in a specific manner which may be more precise than conventional database comparisons.

6.4 Analysis of suGF1 Primary Structure

6.4.1 Characteristic Features of the suGF1 cDNA and Protein Sequence

The length of the cDNA coding for suGF1 (as isolated by 5' and 3' RACE) is about 2.1 kb, of which 300 bp represent 5' untranslated region, the coding region has 1542 bp and the 3' untranslated region has about 250 bp. The short 3' untranslated region presumably originates from a cDNA with a longer poly-A^+ tail (as judged from the SpGCF1 homologue), and probably arises due to alternate priming of the mRNA by the oligo(dT) primer during synthesis of the cDNA. Thus, no distinguishing features can be attributed to the 3' untranslated sequence of the suGF1 cDNA. The 5' upstream region contains three stop codons which are in frame with the predicted open reading frame of the suGF1 cDNA. The lack of strong initiation signals is often a feature associated with cDNA sequences containing several alternative start sites within the protein coding region (42). Indeed, the cDNA for suGF1 exhibits several internal ATG codons positioned at nt 409, 655, 733 and 808 (see fig 3.15). It has been proposed that ribosomes which fail to initiate at the first ATG start codon can begin translation downstream at alternate start sites (this is referred to as the "ribosome scanning mechanism" as described by Descombes and Schibler (1991) (215)), and may result in the formation of truncated protein products arising from a single mRNA species. As predicted from the cDNA sequence, the N-terminal suGF1 protein sequence is characterised by the presence of multiple methionine residues. *In vitro* translation of suGF1 indicates that some of these methionine residues are associated with the formation of several translation products which retain their DNA-binding ability (see section 6.3). This implies that *in vivo*, suGF1 may be translationally regulated at the initiation level, more specifically, it is possible that the mRNA may be scanned by the ribosomal complex to select one of several translation initiation codons (9). It is possible that the characteristic doublet formed by native suGF1 when analysed by EMSA results from N-terminally truncated protein isoforms, which have a very similar molecular weight and therefore cannot be differentiated by SDS-PAGE under the conditions described in this project. Several other examples of proteins which differ in their N-terminal lengths and are transcribed from a single mRNA include C/EBP (216) and antagonists LIP and LAP (215). Similarly, it has been suggested that several truncated SpGCF1 proteins appear to be formed in this way (3).

suGF1 has an open reading frame of 1542 bp (nt 301 - 1843, see fig 3.15) which codes for 514 amino acids, comprising a full length protein of molecular weight 57.2 kDa. The predicted molecular weight

of suGF1 correlates well with the previously predicted molecular weight of 59.5 kDa (1). suGF1 is characterised by a very high proline content (41 proline residues in total, which constitute 8 % of the overall amino acid composition). More than half of the proline residues (25) occur in the N-terminal region. Proline-rich domains are reminiscent of transcription factor activation domains, and are postulated to have an important function in the activation potential of DNA-binding proteins. There are several examples of proline-rich domains which are implicated in certain classes of transcriptional activation domains, such as NGFI-C (114), CTF/NF-I and BTEB (82), and NF-E2 (217). Other examples include c-Jun (114) and C/EBP (218) and the mammalian transcription factors AP-2, OCT-2 and the serum response factor. Therefore, by analogy, this implies that part of the proline-rich N-terminus of suGF1 could represent an activation domain.

The N-terminus of suGF1 is further characterised by a tandem repeat of nine pentapeptides (see fig 3.16), which contain many of the methionine and proline residues as discussed above. Database searches show that these repeats (N/SVSMP) are unique to the suGF1 and SpGCF1 protein (3), and no function can be ascribed to them at this point.

6.4.2 DNA-Binding of *In Vitro* Translated suGF1

The structural features important for the DNA recognition of suGF1 are contained in the central region of the protein, as shown by expression of a truncated suGF1 polypeptide of 31 kDa coded for by nt 699 - 1662 (see fig 3.15), which retained its DNA-binding activity when analysed by EMSA. Previously it was speculated that suGF1 is a zinc finger protein, as evidence from EMSAs indicates that this protein has a requirement for divalent cations (191). However, the primary sequence of suGF1 shows no evidence for a zinc finger structure in the DNA-binding domain. Instead, the region containing the DNA-binding domain comprises a region with 17 amino acids which are highly basic residues (underlined in fig 3.15). Basic amino acids are often associated with DNA-binding domains (219, 220), and probably form part of the DNA-binding domain for suGF1. Several other G·C-binding proteins are associated with basic DNA-binding domains. Examples include, amongst others GCF (83), CTF/NF-1 (82) and NSEP-1 (221). suGF1 is also reminiscent of the transcription factor BTEB-2 whose basic region partially identifies with the basic domain of proteins characterised by the helix-loop-helix and leucine zipper motifs (82).

The N-terminal region of the suGF1 basic domain is a combination of hydrophobic residues, which may constitute two potential heptad repeats of 22 and 63 amino acids each (aa 251 - 273 and 275 -

328). suGF1 also has a third potential heptad of repeats located in the N-terminus (aa 13 - 37) (see fig 3.15). Although no function has been assigned to these repeats either, it appears that they are commonly found in proteins which can adopt the coiled-coil structure, such as myosin (222) and the leucine zipper class of DNA-binding proteins (84). The latter is characterised by both a very basic DNA-binding domain and a strict heptad of leucine residues, which constitute the leucine zipper and mediate dimerisation, a feature which is central to DNA-binding within this family of transcriptional regulators. This leads to the speculation that the potential heptad repeats in the suGF1 protein may form a dimerisation domain which participates either in the formation of homodimers of suGF1 molecules, or may assist in the interaction of suGF1 with accessory proteins. Previous results showed that suGF1 forms discrete multimers in EMSAs in the presence of excess protein (145) and several other transcription factors are also known to form multimeric complexes (eg Sp1). Indeed, Zeller et al (1995b) (149) suggest that SpGCF1 is a transcription factor which binds DNA as a homodimer, and using electron microscopy studies this group predicts that SpGCF1 molecules associate with each other. Therefore direct and implicative evidence suggests that multimerisation may be inherent to the function of suGF1. This was investigated by EMSAs using a combination of two different sized truncated suGF1 proteins. The results indicate that suGF1 does not bind DNA as a homodimer. The two polypeptides of MW 31 kDa (described above) and 41 kDa (nt 762 - 1842) differ in the length of their C-termini, and both contain the putative dimerisation domain, as well as an active DNA-binding domain. Individually these proteins exhibit distinct electrophoretic mobilities when bound to DNA, however when they are co-analysed in EMSAs, no intermediary electrophoretic mobility can be observed. This shows that the differently sized suGF1 polypeptides do not associate with each other in the assay, indicating that the region comprising aa 153 - 455 is devoid of a dimerisation domain, and that DNA-sequence recognition functions independently of the putative dimerisation domain. These results suggest that native suGF1 protein binds DNA as a monomer, and that DNA-sequence recognition of suGF1 is not determined by homodimerisation. However, this does not exclude the possibility that suGF1 may either associate with itself or other accessory proteins. Previous results (145) show that purified suGF1 does form multiple protein-DNA complexes, most likely via protein-protein interactions. It is possible that part of the function of suGF1 occurs via its interaction with accessory proteins (and therefore the formation of heterodimers). However, if so, these accessory proteins are not required for DNA-binding, since the same gel-shift pattern is observed for the native suGF1 in crude nuclear extract and expressed full length recombinant suGF1. The formation of heterodimers between the *in vitro* generated recombinant suGF1 and sea urchin accessory proteins would not take place in the rabbit reticulocyte lysate *in vitro* cell free extract. In addition, native suGF1 purified by DNA-affinity chromatography shows only one protein band of 59.5 kDa when analysed by SDS-PAGE (1). However it is possible that heterodimer formation which is not necessary

for DNA-binding, but necessary for some other function, does occur *in vivo* but the *in vitro* DNA-binding conditions are not suitable for their detection. Although further experiments are required to show whether suGF1 does indeed dimerise, our results indicate that suGF1 recognises DNA as a monomer. Our results do not support the dimer model developed by Zeller et al (1995b) (149), whereby quantitative simulation of gel shift results with recombinant protein were used to predict that SpGCF1 binds DNA in the form of homodimers only. This model excludes the formation of monomeric SpGCF1 complexes. However these authors did not show conclusive experimental evidence for heterodimer formation.

6.4.3 Comparison Between suGF1 and Other Sea Urchin Transcription Factors

suGF1 does not show any significant homology of its DNA or protein sequence to other sea urchin transcription factors, apart from SpGCF1 (3). A comparison between the full length cDNAs coding for suGF1 and SpGCF1 (protein homologues from sea urchin species *P.angulosus* and *S.purpuratus*) shows that they exhibit 84 % identity over a region of 1989 nucleotides, which partially includes the 5' and 3' untranslated regions, as well as the entire coding regions. suGF1 exhibits a longer 5' untranslated region (~ 100 bp), however its poly-A^+ tail is truncated with respect to SpGCF1. The greatest homology (about 89 %) is present in the coding regions of the homologues, both the 5' and 3' untranslated regions exhibit less conservation of nucleotide sequence than the coding regions.

A comparison between the derived amino acid sequences for suGF1 and SpGCF1 (3) reveals an extremely conserved primary structure of 94 % identity between the proteins. All the essential features of suGF1 (see section 6.4.1) are reflected in the SpGCF1 homologue (3). This implies that the homologous proteins are functionally probably identical. The 94 % identity between the two proteins shows that the DNA homology of 89 % between the analogous cDNAs can mainly be attributed to nucleotide base substitutions in degenerate positions. suGF1 codes for a protein of molecular weight 57 kDa and contains 514 amino acids, whereas SpGCF1 comprises 486 amino acids with a molecular weight of 55 kDa. The difference in the molecular weights between the homologues can mainly be attributed to the 28 amino acid difference in their primary structure. The two proteins exhibit highest identity in the central region, which contains the DNA-binding domain. In this region there is a single amino acid residue difference between the proteins. The remaining C-terminal and N-terminal domains are less conserved, for instance the C-terminus exhibits differences between 9 of the amino acid residues, and the suGF1 protein contains 11 additional amino acids in this region. The N-termini show even more variation amongst the homologues, for instance, suGF1 contains two

additional pentapeptide repeats with respect to SpGCF1, and overall the N-termini have 25 non-identical amino acids, of which 1 substitution has preserved its charge, and 9 others have retained their neutrality (see Appendix XIII). Interestingly, three of the amino acids which have undergone changes are prolines within the putative activation domain of SpGCF1. In the suGF1 homologue these positions are represented by two alanine residues and a tyrosine residue respectively. However overall, suGF1 is characterised by five more proline residues than SpGCF1. There is evidence from deletion experiments performed with C/EBP (223) that activation resulting from prolines is not necessarily restricted to a specific domain or local region, in which case the activation potential of some proteins may rather depend on overall proline content. In total, suGF1 contains five additional proline residues (two in the N-terminus and three in the C-terminus) with respect to its SpGCF1 homologue (3), implying that suGF1 may have a higher activation potential than SpGCF1.

6.4.4 suGF1 and Other G·C-Binding Factors

suGF1 may exist in several N-terminally truncated protein isoforms, possibly resulting from translational control at the initiation level. All the isoforms are able to bind G·C-rich DNA (see section 6.2.3). If native suGF1 protein translation is initiated from the internal ATG start codons situated within the cDNA, it may result in the formation of several suGF1 polypeptides which differ in the length of their activation domains and therefore in their ability to activate. It could be proposed that the shorter proteins may compete for binding with the longer proteins, and therefore effectively act as competitive repressors, as has been shown for C/EBP and c-Myc. For instance, all species of c-Myc examined exist in two protein isoforms (224) which differentially regulate transcription (225). The difference in their *trans*-activation abilities is ascribed to the differing amino termini. Only the longer protein contains the *trans*-activation domain which is proposed to undergo a conformational change, possibly allowing differential interaction with other proteins (such as the transcriptional machinery), which could explain the functional difference between the two protein isoforms. Similarly, the C/EBP family consists of several transcription factors important in the regulation of growth and differentiation of a number of cell types (226). A single mRNA species serves as template for the translation of two protein isoforms of C/EBP, viz a full-length and N-terminally truncated isoform (216, 215) which differentially modulate transcriptional activity. The full length protein is a potent activator, whereas the truncated isoform is a repressor or an activator with low activity, depending on the promoter context. The ratios of the two isoforms may be regulated by the activity of the translation initiation factor eIF-2 (9). SpGCF1 also appears to give rise to several truncated proteins present in different ratios, most of which are able to bind DNA (3).

suGF1 has a characteristically high proline content. More than half of the proline residues (25) occur in the N-terminal region. Proline-rich domains are a characteristic feature of certain classes of transcriptional activation domains and are implicated in the functional activation potential of several DNA-binding proteins. Examples of transcription factors which have proline-rich activation domains include NGFI-C, which has an unusually high proline composition of 25 % over 77 amino acids (114), CTF/NF-I and BTEB-2 both have proline-rich regions (the latter has 16 prolines over a region of 67 residues) (82), and NF-E2 has a N-terminal proline-rich transactivation domain which is required to activate globin gene expression (217). Further examples include c-Jun (114), the mammalian transcription factors AP-2, OCT-2 and the serum response factor, as well as C/EBP (218). Therefore, by analogy, suGF1 is implied to have an activation domain at its N-terminus which is proline-rich, although the overall proline content of suGF1 may also act as an important contributor to functional activation potential, since evidence from deletion experiments performed with C/EBP (223) implies that activation resulting from prolines is not necessarily restricted to a local region, but may depend on overall proline content of the transcription factor.

Structural features important for the sequence-specific binding of suGF1 are contained in the central region of the protein which contains a characteristic domain of highly basic amino acid residues (aa 332 - 349, underlined in fig 3.15). Basic amino acids are often associated with DNA-binding domains (219, 220). Several G·C-binding proteins contain basic DNA-binding domains, too. For example BTEB-2 has a basic region which partially identifies with the basic domains of proteins characterised by the helix-loop-helix and leucine zipper motifs (82), and GCF is characterised by the basic region at its N-terminus which functions as the DNA-binding domain (83). This protein also has two leucine zipper motifs, which are proposed to facilitate dimerisation and are common to DNA-binding proteins such as the *fos/jun* system (227), C/EBP (228), CREB (229), Myc and GCN4 (84). Other examples of proteins which have basic DNA-binding domains (but are not associated with the leucine zipper motifs) include CTF/NF-1 (82) and NSEP-1 (221).

6.5 Developmental Distribution of the suGF1 mRNA Transcript

The finding that suGF1 binds *in vitro* to the G·C-rich region in the spacer between H1 and H4 histone genes of the early histone gene battery, which is developmentally regulated in sea urchins, raises the question whether suGF1 functions as part of a set of mechanisms ensuring tight regulation of these genes. Binding of suGF1 to this element *in vitro* does not necessarily correlate with the protein(s) that

functionally interact with the element *in vivo*. The early histone gene battery is coordinately expressed in a distinct temporal pattern during early embryogenesis. It is expressed up to late blastula stage embryos. In *P.angulosus* embryos the switch from the early to the late set of histones occurs between 9 and 12 hours after fertilisation (230). Co-expression of suGF1 with transcription of these genes would support an important regulatory role for suGF1 in the controlled expression of the histone genes. Therefore the similarity between the temporal pattern of suGF1 mRNA and the expression pattern of the histone genes was investigated by Northern analysis and RT-PCR.

The RT-PCR experiment showed that mRNA transcripts for suGF1 are present in *P.angulosus* eggs, 4 - 45 hour embryos, muscle and testis tissue. However they are absent in ovaries. RT-PCR gives qualitative insight about the location of the RNA transcripts of interest, however this experiment cannot give a quantitative indication about the relative differences in the amounts of transcripts amongst the different tissues / embryonic stages. The experiment was not performed in a semiquantitative manner either, since no suitable internal standard could be found. Candidate genes which generally serve as internal controls, eg actin genes, histone genes etc, are developmentally regulated in the sea urchin embryo. Further, the RT-PCR experiment cannot be used to extrapolate in which tissue / embryonic stages the functional protein (translated from the RNA transcript of interest) is expressed. It simply serves to indicate the presence of the RNA transcripts.

Analysis of *P.angulosus* sea urchin RNA by Northern blotting resulted in uninterpretable signals, as the radioactive probe appeared to hybridise over a wide range of molecular weights for each RNA sample. This implies that the probe may have had homology with some other abundant transcripts, which may have masked authentic signals from the suGF1 transcripts. Zeller et al (1995) (3) used quantitative RNase protection assays to measure the transcript prevalence of SpGCF1. Absence of Northern analysis for *S.purpuratus* RNA may indicate that this group encountered similar problems with Northern blotting of SpGCF1 transcripts. (suGF1 transcripts were not analysed by RNase protection assays due to consistently inadequate incorporation of the radiolabel into the antisense probe.) RNase protection assays detected the presence of the SpGCF1 transcript in *S.purpuratus* ovaries and about 5000 copies of the transcript in unfertilised eggs. Transcript prevalence reached maximum levels in 9 hr embryos and steadily decreased to about 1500 copies per embryo at 72 hours (3). Adult *S.purpuratus* tissue (other than ovaries) was not analysed for the presence of SpGCF1 transcripts. Analysis of the presence of suGF1 transcripts in *P.angulosus* embryos by RT-PCR correlates with the distribution of SpGCF1 transcripts in *S.purpuratus* as analysed by RNase protection assays. (Note that the *P.angulosus* and *S.purpuratus* embryos are grown under different conditions and their rates of development differ, viz *P.angulosus* embryos develop at almost twice the

rate with respect to *S.purpuratus* embryos.) Unfortunately analysis of suGF1 transcripts was not quantitative, therefore no comparisons can be drawn about the relative abundance of suGF1 and SpGCF1 transcripts. Distribution of suGF1 and SpGCF1 transcripts does, however, not correlate with respect to ovaries (suGF1 transcripts are absent in this tissue). It is possible that SpGCF1 transcripts were located in the ovaries of *S.purpuratus* embryos as a result of incomplete depletion of eggs from this tissue.

The presence of suGF1 transcripts in both early and late embryonic stages, as well as adult tissue suggests that the functional suGF1 protein is unlikely to have a role in the temporal regulation of the early histone gene battery as was speculated previously (see section 1.4.3). The early histone genes are switched off in blastula stage embryos (9 hours after fertilisation, 230). However, the wide distribution of the suGF1 mRNA transcript throughout several embryonic stages and adult tissues implies that it could indeed play a role in transcriptional regulation of several unrelated genes with G·C-rich promoters. Thus suGF1 may have a role as a general transcriptional regulatory protein.

6.6 Conclusions and Perspectives

Although there is no direct evidence for the biological role of suGF1 it appears that it may function as a general transcriptional regulator of several unrelated genes in sea urchin development. The presence of suGF1 mRNA transcripts in the early sea urchin embryonic stages, as well as adult muscle and testis tissue cannot be taken as direct evidence that the functional protein is expressed in these stages / tissues. However, previous evidence shows that suGF1 is a prevalent nuclear transcription factor in embryonic tissue (1). It has been implicated as a transcription factor with respect to its functional abilities (it binds sequence specifically to G·C-rich DNA sequences which have been shown by others to function as *cis*-regulatory elements) and as a result of its similar DNA-binding specificity to other G·C-rich DNA-binding factors, such as BGP1 (11), the ectoderm specific factor (16) and its *S.purpuratus* homologue SpGCF1 (3), which have all been implicated as transcriptional regulators. Several genes in different organisms have upstream G·C-rich regions, hence suGF1 may be a member of a family of G-binding factors which is involved in the regulation of unrelated genes.

The cDNA encoding suGF1 was successfully cloned using a PCR-strategy, based on the homology of suGF1 to a G·C-binding protein present in *S.purpuratus* embryos. The integrity of the clone was confirmed by eukaryotic *in vitro* expression of the recombinant protein, which showed identical DNA-binding properties to native suGF1 in crude nuclear extracts suggesting that native suGF1 does not undergo post-translational modifications. In addition, the true identity of the cDNA sequence was

confirmed by partial identification of the primary structure of native suGF1. Functional cDNA cloning of suGF1 using the DNA-ligand screening technique proved unsuccessful, either as a result of enhanced detection of false positives by use of p[d(I-C)] in the protocol, or more likely because recombinant suGF1 protein is not compatible with expression in a bacterial system.

Analysis of the primary structure of the suGF1 protein has given further support towards the proposed biological role for suGF1 as a regulator in gene expression. The region of sequence-specific DNA-binding has been identified in the suGF1 primary structure (nt 762 - 1662). A truncated suGF1 protein corresponding to this DNA region retains its DNA-binding ability. Many G·C-rich binding proteins are classified as zinc-finger proteins, however suGF1 does not have a Zinc finger DNA-binding domain, despite its requirement for divalent cations in order to bind DNA (191). The suGF1 DNA-binding domain contains highly basic amino acid residues which is a feature common to several transcription factors. A putative proline-rich transcriptional activation domain has been identified in the N-terminus of suGF1, implying a functional activity for this protein over and above its DNA-interaction. Activation domains may function in contacting other proteins, which could present another feature in the function of suGF1, since this protein has been shown to form discrete suGF1-suGF1 multimers (1), similar to transcription factors like Sp1. The results show that recognition of the cognate DNA-binding site of this protein does not rely on obligate homodimerisation or heterodimerisation. However the suggestion that suGF1 does potentially form associations with accessory factors can be gleaned from its primary structure, which comprises several potential heptad repeats reminiscent of a class of proteins which form "coiled-coil" conformations, and also leucine zipper proteins. The heptad repeat motif is closely associated with protein-protein interactions (ie homodimer or heterodimer formations), and thus it ties in with a putative dimerisation domain for suGF1, potentially allowing suGF1 to interact with different general transcription factors or with subunits of the transcriptional machinery. The inability to detect such putative interactions may have been due to the *in vitro* assay conditions.

Analysis of the primary structure of suGF1 implies that different isoforms of this protein may exist, ie the suGF1 protein may be translationally regulated by the ribosome scanning mechanism (215). This form of transcription factor regulation (ie translational control) and how it regulates the activity of transcription factors is not well understood at this stage, and still requires future attention. However, it could be proposed that suGF1 may be a promoter selective regulator which modulates the activity of promoters via interactions of the different protein isoforms, which could either act as activators with differing activational strength, or the truncated forms could act as competitive inhibitors.

Identification of suGF1 and determination of its primary structure has provided several insights about its mechanism of action. However, several questions remain to be answered. For instance, does suGF1 function by forming contacts between distant DNA elements, and does multimerisation with itself or other factors result in looping of DNA *in vivo*? Does the formation of stable complexes increase productive interactions which may stimulate expression? Can suGF1 interact with nucleosomes or with the basal transcription apparatus to effect transcriptional regulation? By making use of the cloned cDNA of suGF1 and application of *in vitro* functional assays, many of the above suggestions and questions could be explored and addressed further, to elucidate the biological role associated with suGF1, and possibly other G-binding factors.

CHAPTER 7

REFERENCES

1) Hapgood, J. and D. Patterton (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.

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

3) Zeller, R.W., J.A. Coffman, M.G. Harrington, R.J. Britten, and E.H. Davidson (1995). SpGCF1, a sea urchin embryo DNA-binding protein, exists as five nested variants encoded by a single mRNA. *Dev. Biol.* **169:**713-727.

4) Mitchell, P.J. and R. Tjian (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA-binding proteins. *Science* **245:**371-378.

5) Johnson, P.F., and S.L. McKnight (1989). Eukaryotic transcriptional regulatory proteins. *Biochemistry* **58:**799-839.

6) Saltzman, A.G. and R. Weinmann (1989). Promoter specificity and modulation of RNA polymerase II transcription. *FASEB J.* **3:**1723-1733.

7) Fassler, J.S. and G.N. Gussin (1996). Promoter and Basal Transcription Machinery in Eubacteria and Eukaryotes: Concepts, Definitions and Analogies. *Methods in Enzymology* **273:**3-29.

8) Ptashne, M. and A. Gann (1997). Transcriptional activation by recruitment. *Nature* **386:**569-577

9) Calkhoven, C.F. and G. Ab (1996). Multiple steps in the regulation of transcription-factor level and activity. *Biochem. J.* **317:**329-342

10) Jackson, S.P. (1992). Identification and characterization of eukaryotic transcription factors. In *Gene Transcription A Practical Approach.* Harnes, B.D. and S.J. Higgins, eds (IRL Press).

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

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

13) Emerson, B.M., J.M. Nickol and T.C Fong (1989). Erythroid-specific activation and derepression of the chick beta-globin promoter *in vitro*. *Cell* **57:**1189-1200.

14) Karsenty, G. and B. deCrombrugghe (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.

15) Davis, T.L., A.B. Firulli and A.J. Kinniburgh (1989). Ribonucleoprotein and protein factors bind to an H-DNA forming c-*myc* DNA element. Possible regulators of the c-*myc* gene. *Proc. Natl. Acad. Sci.* **86:**9682-9686.

16) Xiang, M., S.-Y. Lu, M. Musso, G. Karsenty and W.H. Klein (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.

17) Kohwi, Y., S.R. Malkhosyan and T. Kohwi-Shigematsu (1992). Intramolecular dG.dG.dC triplex detection in *Escherichia coli* cells. *J. Biol. Chem.* **223**:817-822.

18) McKeon, C., A. Schmidt, and B. deCrombrugghe (1984). A sequence conserved in both the chicken and mouse α2(I) collagen promoter contains sites sensitive to S1 nuclease. *Biol. Chem.* **259**:6636-6640.

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

20) Patterton, H.-G. and C. von Holt (1993). Negative supercoiling and nucleosome cores. II. The effect of negative supercoiling on the positioning of nucleosome cores *in vitro*. *J. Mol. Biol.* **229**:367-655.

21) Stadler, J., A. Larsen, J.D. Engel, M. Dolan, M. Groudine and H. Weintraub (1980). Tissue specific DNA cleavages in the globin chromatin domain introduced by DNase1. *Cell* **20**:451-460.

22) Elgin, S.R.C. (1995). The formation and function of DNAse I hypersensitive sites in the process of gene activation. *J. Biol. Chem.* **263**:19259-19262.

23) Schmid, A., K.-D. Fascher and W. Horz (1992). Nucleosome disruption at the yeast PHO5 promoter upon PHO5 induction occurs in the absence of DNA replication. *Cell* **71**:853-864.

24) Gross, D.S. and W.T. Garrard (1988). Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**:159-197.

25) Leuther, K.K., J.M. Salmeron and S.A. Johnson (1993). Genetic evidence that an activation domain of GAL4 does not require acidity and may form a β-sheet. *Cell* **72**:575-585.

26) Baron (1997). Transcriptional control of globin gene switching during vertebrate development. *Biochimica et Biophysica Acta* **1351**:51-72.

27) Kingston, R.E., C.A. Bunker and A.N. Imbalzano (1996). Repression and activation by multiprotein complexes that alter chromatin structure. *Genes Dev.* **10**:905-920.

28) Wison, C.J., D.M. Chao, A.N. Imbalzano, R.E. Kingston and R.A. Young (1996). RNA polymerase II holoenzyme contains SWI/SNF regulators involved in chromatin remodelling. *Cell* **84**:235-244.

29) Winston F. and M. Carlson (1992). Yeast SNF/SWI transcriptional activators and the SPT/SIN chromatin connection. *Trends in Genet.* **8**:387-391.

30) Tjian, R. and T. Maniatis (1994). Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**:5-8.

31) Kirchhamer, C.V., C.-H. Yuh and E.H. Davidson (1996). Modular *cis*-regulatory organisation of developmentally expressed genes: Two genes transcribed territorially in the sea urchin embryo and other examples. *Proc. Natl. Acad. Sci. (USA)* **93**:9322-9328.

32) Busby, S. and R. Ebright (1994). Promoter structure, promoter recognition and transcription activation in prokaryotes. *Cell* **79**:743-746.

33) Perez-Martin J. and M. Espinosa (1994). Correlation between DNA bending and transcriptional activation at a plasmid promoter. *J. Mol. Biol.* **241**:7-17.

34) Bracco, L., D. Kotlarz, A. Kolb, S. Diekman and H. Buc (1989). Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli*. *EMBO J.* **8**:4289-4296.

35) Kageyama, R. and I. Pastan (1989). Nuclear Factor ETF Specifically Stimulates Transcription from Promoters without a TATA Box. *J. Biol. Chem.* **264**:15508-15514.

36) Siebenlist, U., P. Bressler and K. Kelly (1988). Two distinct mechanisms of transcriptional control operate on the c-myc differentiaion in HL60 cells. *Mol. Cell. Biol.* **8**:867-874.

37) Kinniburgh, A.J., Firulli, A.B. and R. Kolluri (1994). DNA triplexes and regulation of the c-*myc* gene. *Gene* **149**:93-100.

38) Pyrc, J.J., K.H. Moberg and D.J. Hall (1992). Isolation of a novel cDNA encoding a zinc finger protein that binds to two sites within the c-*myc* promoter. *Biochemistry* **31**:4102-4110.

39) Bossone, S.A., C. Asselin, A.M. Patel and K.B. Marcu (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.

40) Beck, K.M., A.H. Seekamp, G.R. Askew, Z. Mei, C.M. Farrell, S. Wank and L.N. Lukens (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. *Nucl. Acids. Res.* **19**:4975-4982.

41) Liau, G., D. Szapary, C. Setoyama and B. deCrombrugghe (1986). Restriction enzyme digestions identify discrete domains in the chromatin around the promoter of the mouse alpha 2(I) collagen gene. *J. Biol. Chem.* **261**:11362-11368

42) Hasegawa, T., X. Zhou, L.A. Garrett, E.C. Ruteshauser, S.N. Maity and B. deCrombrugghe (1996). Evidence for three major transcription activation elements in the proximal mouse proα2(1) collagen promoter. *Nucl. Acids. Res.* **24**:3253-3260.

43) Kovacs, A., J.C. Kandala, K.T. Weber and R. Guntaka (1996). Triple Helix-forming Oligonucleotide Corresponding to the Polypyrimidine Sequence in the Ratα1(I) Collagen Promoter Specifically Inhibits Factor Binding and Transcription. *J. Biol. Chem.* **271**:1805-1812.

44) Tamaki, T., K. Ohnishi, L.C. Hart and E.C. LeRoy (1995). Characterization of a G.C-rich containing Sp1 binding site(s) as a constitutive responsive element of the alpha 2(I) collagen gene in human fibroblasts. *J. Biol. Chem.* **270**:4299-4304.

45) Karsenty, G. and R.W. Park (1995). Regulation of type I collagen gene expression. *Int. Rev. Immunol.* **12**:177-185.

46) Kumar, A.P., P.K. Mar, B. Zhao, R.L. Montgomery, D.-C. Kang and A.P. Bulter (1995). Regulation of Rat Ornithine Decarboxylase Promoter Activity by Binding of Transcription Factor Sp1. *J. Biol. Chem.* **270**:4341-4348.

47) Kumar, A.P. and A.P. Bulter (1997). Transcription factor Sp3 antagonises activation of the ornithine decarboxylase promoter by Sp1. *Nucl. Acids. Res.* **25**:2012-2019.

48) Merchant, J.L., A. Shiotani, E.R. Mortensen, D.K. Shumaker and D.R. Abraczinskas (1995). Epidermal Growth Factor Stimulation of the Human Gastrin Promoter Requires Sp1. *J. Biol. Chem.* **270**:6314-6319.

49) Cornwell, M.M. and D.E. Smith (1993). Sp1 activates the MDR1 promoter through one of two distinct G-rich regions that modulate promoter activity. *J. Biol. Chem.* **268**:19505-19511.

50) Tsai-Morris, C.-H., Y. Geng, E. Buczko and M.L. Dufau (1995). Characterization of Diverse Functional Elements in the Upstream Sp1 Domain of the Rat Luteinizing Hormone Receptor Gene Promoter. *J. Biol. Chem.* **270**:7487-7494.

51) Savagner, P., P.H. Krebsbach, O. Hatano, T. Miyashita, J. Liebman and Y. Yamada (1995). Collagen II promoter and enhancer interact synergistically through Sp1 and distinct nuclear factors. *DNA Cell Biol. (United States)*. **14**:501-510.

52) Stamatoyannopoulos, G. and A. Nienhuis (1994). Hemoglobin Switching. In *The Molecular Basis of Blood Diseases.* (G. Stamatoyannopoulos, A. W. Nienhuis, P. Leder and P.W. Majerus, eds), W.B. Sounders, Philadelphia. pp107-155.

53) Grosveld, F., M. Antonion, M. Berry, E. De Boer, N. Dillon, J. Ellis, P. Fraser, J. Hurst, A. Imam, D. Meijer, S. Philipsen, S. Pruzin, J. Strouboulis and D. Whyatt (1993). Regulation of human globin gene switching. *Cold Spring Harbor Symp. Quant. Biol.* **58**:7-13.

54) Trepicchio, W.L., M.A. Dyer and M.H. Baron (1993). Developmental regulation of the human embryonic beta-like globin gene is mediated by synergistic interactions among multiple tissue- and stage-specific elements. *Mol. Cell. Biol.* **13**:7457-7468.

55) Grosveld, F., G.B. Van Assendelft, D. Greavs and G. Kollias (1987). Position-independent, high level expression of the human beta-globin gene in transgenic mice. *Cell* **51**:975-985.

56) Wijgerde, M., F. Grosveld and P. Fraser (1995). Transcription complex stability and chromatin dynamics *in vivo*. *Nature* **377**:209-213.

57) Emerson, B.M. and G. Felsenfeld (1984). Specific factor conferring nuclease hypersensitivity at the 5' end of the chicken adult beta globin gene. *Proc. Natl. Acad. Sci (USA)* **81**:95-99.

58) Strauss, E.C. and S.H. Orkin (1992). *In vivo* protein-DNA interactions at hypersensitive site 3 of the human beta globin locus control region. *Proc. Natl. Acad. Sci. (USA)* **89**:5809-5813.

59) Orkin, S.H. (1995). Hematopoiesis: how does it happen? *Curr. Opin. Cell Biol.* **7**:870-877.

60) Dyer, M.A., R. Naidoo, P. Hayes, C.J. Larson, G.L. Verdine and M.H Baron (1996). A DNA-binding protein interacts with an essential upstream regulatory element in the human embryonic beta-like globin gene. *Mol. Cell. Biol.* **16**:829-838.

61) Amrolia, P.J. (1993). Identification of two novel regulatory elements within the 5'-untranslated region of the human A gamma-globin gene. *J. Biol. Chem.* **270**:12892-12898.

62) Merika, M. and S.H. Orkin (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.

63) Yant, S.R., W. Zhu, D. Millinoff, J.L. Slighton, M. Goodman and D.L. Gumucio (1995). High affinity YY1 binding motifs: identification of two core types (ACAT and CCAT) and distribution of potential binding sites within the human beta globin cluster. *Nucl. Acids Res.* **23**:4353-4362.

64) Evans, T. and G. Felsenfeld (1989). The erythroid-specific transcription factor Eryf1: a new finger protein. *Cell* **58**:877-885.

65) Tsai, F.-Y., G. Keller, F.C. Kuo, M. Weiss, J. Chen, M. Rosenblatt, F.W. Alt and S.H. Orkin (1989). An early hematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**:221-226.

66) Crossley, M., M. Merika and S.H. Orkin (1995) Self-association of the erythroid transcription factor GATA-1 mediated by its zinc finger domain. *Mol. Cell. Biol.* **15**:2448-2456.

67) Valge-Archer, V.E., H. Osada, A.J. Warren, A. Forster, J. Li, R. Baer and T.H. Rabbitts (1994). The LIM protein RBTN2 and the basic helix-loop-helix protein TAL-1 are present in a complex in erythroid cells. *Proc. Natl. Acad. Sci (USA)* **91**:8617-8621.

68) Andrews, N.C., H. Erdjument-Bromage, M.B. Davidson, P. Tempst and S.H. Orkin (1993). Erythroid transcription factor NF-E2 is a hemopoietic-specific base-leucine zipper protein. *Nature* **362**:722-728.

69) Barton, M.C., N. Madani and B.M. Emerson (1993). The erythroid protein cGATA-1 functions with a stage-specific factor to activate transcription of chromatin-assembled beta-globin genes. *Genes &Dev.* **7**:1796-1809.

70) Robb, L., C.C. Drinkwater, D. Metcalf, R. Li, F. Kontgen, N.A. Nicola and C.G. Begley (1995). Hematopoietic and lung abnormalities in mice with a null mutation of the common beta subunit of the receptors for granulocyte-macrophage colony-stimulating factor and interleukins 3 and 5. *Proc Natl Acad Sci U S A* **92**:9565-9569.

71) Shivdasani, R.A., E.L. Mayer and S.H. Orkin (1995). Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* 1995 **373**:432-434.

72) Warren, A.J., W.H. Colledge, M.B. Carlton, M.J. Evans, A.J. Smith and T.H. Rabbitts (1994). The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. *Cell* **78**:45-57.

73) Yamamoto, M., L.J. Ko, M.W. Leonard, H. Beug, S.H. Orkin and J.D. Engel (1990). Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. *Genes. Dev.* **10**:1650-1662.

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

75) Miller, I.J. and J.J. Bieker (1993). A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. *Mol. Cell. Biol.* **5**:2776-2786.

76) Lewis, C.D., S.P. Clark, G. Felsenfeld and H. Gould (1988). An erythroid-specific protein that binds to the poly(dG) region of the chicken β-globin gene promoter. *Genes Dev.* **2**:863-873.

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

78) Kohwi, Y. (1989). Cationic metal-specific structures adopted by the poly(dG) region and the direct repeats in the chicken adult beta A globin gene promoter. *Nucleic Acids Res.* **17**:4493-4502.

79) Karsenty, G., P. Golumbek and B. de Crombrugghe (1988). Point mutations and small substitution mutations in three different upstream elements inhibit the activity of the mouse alpha 2(I) collagen promoter. *Biol Chem* **263**:13909-13915.

80) Dailey, L., S.M. Hanly, R.G. Roeder and N. Heintz (1986). Distinct transcription factors bind specifically to two regions of the human histone H4 promoter. *Proc Natl Acad Sci (USA)* **83**:7241-7245.

81) Sogawa, K., H. Imataka, Y. Yamasaki, H. Kusume, H. Abe and Y. Fuji-Kuriyama (1993). cDNA cloning and transcriptional properties of a novel GC box-binding protein BTEB-2. *Nucl. Acids. Res.* **21**:1527-1532.

82) Mermod, N., E.A. O'Neill, T.J. Kelly and R. Tjian (1989). The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell* **58**:741-753.

83) Kageyama, R. and I. Pastan (1989). Molecular Cloning and Characterization of a Human DNA Binding Factor That Represses Transcription. *Cell* **59**:815-825.

84) Landschulz, W.H., P.F. Johnson and S.L. McKnight (1988). The leucine zipper: a hypothetical structure common to a new class of DNA-binding proteins. *Science* **240**:1759-1764.

85) Berg, J.M. (1992). Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. *Proc. Natl. Acad. Sci. (USA)* **89**:11109-11110.

86) Kriwacki, R.W., S.C. Schultz, T.A. Steitz and J.P. Caradonna (1992). Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1. *Proc. Natl. Acad. Sci. (USA)* **89**:9759-9763.

87) Kadonaga, J.T., C. Carner, F.R. Masiarz and R. Tjian (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA-binding domain. *Cell* **51**:1079-1090.

88) Hartshorne, T.A., H. Blumberg and E.T. Young (1986). Sequence homology of the yeast regulatory protein ADR1 with *Xenopus* transcription factor TFIIIA. *Nature (London)* **320**:281-287.

89) Eisen, A., W.E. Taylor, H. Blumberg and E.T. Young (1988). The yeast regulatory protein ADR1 binds in a zinc-dependent manner to the upstream activating sequence of ADH2. *Mol. Cell. Biol.* **8**:4552-4556.

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

91) Saffer, J.D., S.P. Jackson and S.J. Thurston (1990). SV40 stimulates expression of the transacting factor Sp1 at the mRNA level. *Genes Dev.* **4**:659-666.

92) Kadonaga, J.T., and R. Tjian (1986). Affinity purification of sequence specific DNA-binding proteins. *Proc. Natl. Acad. Sci. (USA)* **11**:20-23.

93) Nehls, M.C, M.L. Grapilon and D.A. Brenner (1992). NF-I / SP1 switch elements regulate collagen alpha 1(I) gene expression. *DNA Cell. Biol.* **11**:443-452.

94) Al-Asadi, R., E.C. Yi and J.L. Merchant (1995). Sp1 affinity for GC-rich elements correlates with ornithine decarboxlyase activity. *Biochem. Biophys. Res. Commun.* **214**(2):324-330.

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

96) Mastrelangelo, I.A., A.J. Courey, J.S. Wall, S.P. Jackson and P.V.C. Hough (1991). DNA looping and Sp1 multimer links: A mechanism for transcriptional synergism and enhancement. *Proc. Natl. Acad. Sci. (USA)* **88**:5670-5674.

97) Perkins, N.D., N.L. Edwards, C.S. Duckett, A.B. Agranoff, R.M. Smid and G.J. Nabel (1993). A cooperative interaction between NF-kappa B and Sp1 for HIV-1 enhancer activation. *EMBO J.* **12**:3551-3558.

98) Gunther, M., T. Frebourg, M. Laithier, N. Fossar, M. Bouziane-Ouartini, C. Lavialle and O. Brison (1995). An Sp1 binding site and the minimal promoter contribute to the overexpression of the cytokeratin 18 gene in the tumorigenic clones relative to that in the nontumorigenic clones of a human carcinoma cell line. *Mol. Cell. Biol.* **15**:2490-2499.

99) Anderson, G.M. and S.O. Freytag (1991). Synergistic Activation of a Human Promoter *in vivo* by Transcription Factor Sp1. *Mol. Cell. Biol.* **11**:1935-1943.

100) Su, W., S. Jackson and R. Tjian (1991). DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes Dev.* **5**:820-826.

101) Geiser, A.G., K.J. Busam, S.J. Kim, R. Layatis, M.A. O'Reilly, R. Webbink, A.B. Roberts and M.B. Sporn (1993). Regulation of the transforming growth factor -beta1 and -beta3 promoters by transcription factor Sp1. *Gene (Amst)* **129**:223-228.

102) Kadonaga, J.T., A.J. Courey, J. Ludika and R. Tjian (1988). Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science* **242**:1566-1570.

103) Courey, A.J. and R. Tjian (1988). Analysis of Sp1 *in vivo* reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* **55**:887-898.

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

105) Jackson, S.P, J.J. MacDonald, S. Lees-Miller and R. Tjian (1990). GC-box binding induces phosphorylation of Sp1 by a DNA-binding protein kinase. *Cell* **63**:155-165.

106) Thiesen, H.J. and C. Bach (1991). Transition metals modulate DNA-protein interactions of Sp1 zinc finger domain with its cognate target site. *Biochem. Biophys. Res. Commun.* **176:**551-557.

107) Kim, S.J., U.S. Onwuta, Y.I. Lee, R. Li, M.R. Botchan and P.D. Robbins (1992). The retinoblastoma gene product regulates Sp1-mediated transcription. *Mol. Cell. Biol.* **12:**2455-2463.

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

109) Hagen, G., S. Müller, G. Beato and G. Suske (1992). Cloning by recognition site screening of two novel GT-box binding proteins: a family of Sp1 related genes. *Nucl. Acids. Res.* **20:**5519-5525.

110) Hagen, G., J. Dennig, A. Preiss, M. Beato and G. Suske (1995). Functional Analyses of the Transcription Factor Sp4 Reveal Properties Distinct from Sp1 and Sp3. *J. Biol. Chem.* **270:**24989-24994.

111) Majello, B., P. DeLuca, G. Suske and L. Lania (1995). Differential transcriptional regulation of c-myc promoter through the same DNA-binding sites targeted by Sp1-like proteins. *Oncogene* **10:**1841-1848.

112) Udvadia, A., D.J. Templeton and J.M. Horowitz (1995). Functional interactions between the retinoblastoma (Rb) protein and Sp-family members: superactivation by Rb requires amino acids necessary for growth repression. *Proc. Natl. Acad. Sci. (USA)* **92:**3953-3957.

113) Liang, Y.X., D.F. Robinson, J. Dennig, G. Suske and W.E. Fahl (1996). Characterization of mutations in the beta subunit of the mitochondrial F1-ATPase that produce defects in enzyme catalyst assembly. *J. Biol. Chem.* **271:**11792-11797.

114) Crosby, S.D., J.J. Puetz, K.S. Simburger, T.J. Fahrner and J. Milbrandt (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.

115) Davidson, E.H. (1990). How embryos work: a comparative view of diverse modes of cell fate specification. *Development* **108:**365-389.

116) Davidson, E.H. (1991). Spatial mechanisms of gene regulation in metazoan embryos. *Development* **113:**1-26.

117) Bullough, W. S. (1958). Practical Invertebrate Anatomy. The Macmillan Press LTD, London. p 412.

118) Maxson, M. and H. Tan (1994). Promoter analysis meets pattern formation: transcriptional regulatory genes in sea urchin embryogenesis. *Current Opinion in Genetics and Development* **4:**678-684.

119) Franks, R.R., R. Anderson, J.G. Moore, B.R. Hough-Evans, R.J. Britten and E.H. Davidson (1990). Competitive titration in living sea urchin embryos of regulatory factors required for expression of the *CyIIIa* actine gene. *Development* **110:**31-40.

120) Calzone, F.J., N. Thézé, P. Thiebaud, R.L. Hill, R.J. Britten and E.H. Davidson (1988). Developmental appearance of factors that bind specifically to *cis*-regulatory sequences of a gene expressed in the sea urchin embryo. *Genes Dev.* **2:**1074-1088.

121) Coffman, J.A. and E.H. Davidson (1992). Expression of Spatially Regulated Genes in the Sea Urchin Embryo. *Curr. Opin. Genet. Dev.* **1:**136-146.

122) Höög, C., F.J. Calzone, A.E. Cutting, R.J. Britten and E.H. Davidson (1991). Regulatory Factors of the Sea Urchin Embryo.II. Two Dissimilar Factors, P3A1 and P3A2, Bind to the Same Target Sites that are Required for Early Territorial Gene Expression. *Development* **112:**351-364.

123) Hough-Evans, B.R., R.R Franks, R.W. Zeller, R.J. Britten and E.H. Davidson (1990). Negative spatial regulation of lineage specific *CYIIIa* actin gene in the sea urchin embryo. *Development* **110:**41-50.

124) Mao, C.A., L. Gan and W.H. Klein (1994). Multiple Otx binding sites required for expression of the *Strongylocentrotus purpuratus* Spec2a gene. *Dev Biol* **165**:229-242.

125) Yuh, C.H., A. Ransick, P. Martinez, R.J. Britten and E.H. Davidson (1994). Complexity and organization 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.

126) Yuh, C.H. and E.H. Davidson (1996). Modular *cis*-regulatory organization of Endo16, a gut-specific gene of the sea urchin embryo. *Development* **122**:1069-1082.

127) Venuti, J.M., L. Goldberg. T. Chakraborty. E.N. Olson and W.H. Klein (1991). A Myogenic Factor from Sea Urchin Embryos Capable of Programming Muscle Differentiation in Mammalian Cells. *Proc. Natl. Acad. Sci. (USA)* **88**:6219-6223.

128) Venuti, J.M., L. Gan, M.T. Kozlowski and W.H. Klein (1993). Developmental Potential of Muscle Cell Progenitors and the Myogenic Factor SUM-1 in the Sea Urchin Embryo. *Mech. Dev.* **41**:3-14.

129) Maxson, R., R. Cohen, L. Kedes and T. Mohun (1983). Expression and Organisation of Histone Genes. *Annu. Rev. Genet.* **17**:239-277.

130) Fei, H. and G. Childs (1993). Temporal embryonic expression of the sea urchin early H1 gene is controlled by sequences immediately upstream and downstream of the TATA element. *Dev. Biol.* **155**:383-395.

131) Palla, F., C. Casano, L. Albanese, L. Anello, F. Gianguzza, M.G. DiBernardo, C. Bonura and G. Spinelli (1989). *Cis*-acting elements of the sea urchin histone H2A modular binding transcription factors. *Proc. Natl. Acad. Sci. (USA)* **86**:6033-6037.

132) Palla, C. Bonura, L. Anello, C. Casano, M. Ciaccio and G. Spinelli (1993). Sea urchin early histone H2A modular binding factor 1 is a positive transcription factor also for the early histone H3 gene. *Proc. Natl. Acad. Sci (USA).* **90**:6854-6858.

133) Bell, J., B.R. Char and R. Maxson (1992). An octamer element is required for the expression of the alpha H2B histone gene during the early development. *Dev.Biol.* **150**:363-371.

134) Lee, I.J., L. Tung, D.A. Bumcrot and E.S. Weinberg (1991). UHF-1, a factor required for maximal transcription of early and late sea urchin histone H4 genes: analysis of promoter binding sites. *Mol. Cell. Biol.* **11**:1048-1061.

135) DiLiberto, M., Z.C. Lai, H. Fei and G. Childs (1989). Developmental control of promoter-specific factors responsible for the embryonic activation and inactivation of the sea urchin early histone H3 gene. *Genes & Dev.* **3**:973-985.

136) Tung, L., I.J. Lee, H.L. Rich and E.S. Weinberg (1990). Positive and negative transcriptional regulatory elements in the early H4 histone gene of the sea urchin. *Nucleic Acids Res.* **18**:7339-7348.

137) Char, B.R., J.R. Bell, J.D. Dovala, J.A. Coffman, M.G. Harrington, J.C. Becerra, E.H. Davidson, F.J. Calzone and R. Maxson (1993). SpOct, a Gene Encoding the Major Octamer-Binding Protein in Sea Urchin Embryos: Expression Profile, Evolutionary Relationships, and DNA-binding of Expressed Protein. *Dev. Biol.* **158**:350-363.

138) Lai, Z.C., D.J. DeAngelo, M. DiLiberto and G. Childs (1989). An embryonic enhancer determines the temporal activation of a sea urchin late H1 gene. *Mol. Cell. Biol.* **9**:2315-2321.

139) DeAngelo, D.J., J. DeFalco and G. Childs (1993). Purification and Characterisation of the Stage-Specific Embryonic Enhancer-Binding Protein SSAP-1. *Mol. Cell. Biol.* **13**:1746-1758.

140) Zhao, Z., A.M. Colin, J.B. Bell, M.B. Baker, B.R. Char, and R.E. Maxzon (1990). Ontogenic activation of a sea urchin late H2B histone gene by an enhancer-like element located 3' of the gene. *Mol. Cell. Biol.* **10**:6730-6741.

141) Barberis, A., G. Superti-Furga, L. Vitelli, I. Kemler and M. Busslinger (1989). Developmental Tissue Specific Regulation of a Novel Transcription Factor of the Sea Urchin. *Genes Dev.* **3**:663-675.

142) Zhao, Z., G. Vasant, J. Bell, T. Humphreys and R. Maxson (1991). Activation of the L1 Late H2B Histone Gene in Blastula-Stage Sea Urchin Embryos by Antennapedia-Class Homoeprotein. *Mech. Dev.* **34**:21-28.

143) Erselius, J.R., M.D. Goulding and P. Gruss (1990). Structure and expression pattern of the murine *hox*-3.2 gene. *Development* **110**:629-642.

144) Adams, B., P. Dorfler, A. Aguzzi, Z. Kozmik, P. Urbanek, I. Maurer-Fogy and M. Busslinger (1992). *Pax*-5 Encodes Transcription Factor BSAP and is Expressed in B-Lymphocyte, the Developing CNS and Adult Testis. *Genes Dev.* **6**:1589-1607.

145) Patterton, D. and J.P. Hapgood (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.

146) Wu, T.-C. and R.T. Simpson (1985). Transient alteration of the chromatin structure of the sea urchin early histone genes during embryogenesis. *Nucl. Acids. Res.* **13**:6185-6203.

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

148) Flytzanis, C.N., R.J. Britten and E.H. Davidson (1987). Ontogenic activation of a fusion gene introduced into sea urchin eggs. *Proc Natl Acad Sci (USA)* **84**:151-155.

149) Zeller, R.W., J.D. Griffith, J.G. Moore, C.V. Kirchhamer, R.J. Britten and E.H. Davidson (1995b). A multimerising transcription factor of sea urchin embryos capable of looping DNA. *Proc. Natl. Acad. Sci. (USA)* **92**:2989-2993.

150) Calzone, F.J., N. Theze, P. Thiebaud, R.L. Hill, R.J. Britten and E.H. Davidson (1988). Developmental appearance of factors that bind specifically to *cis*-regulatory sequences of a gene expressed in the sea urchin embryo. *Genes Dev.* **2**:1074-1088.

151) Ausubel, F.M., R. Brent, R.E. Kingston, D.D. Moore, J.G. Seideman, J.A. Smith and K. Struhl (1987). *Current Protocols in Molecular Biology*. John Wiley and Sons, New York.

152) Fried, M. (1989). Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis* **10**:366-376.

153) Fried, M. and D.M. Crothers (1984). Kinetics and mechanism in the reaction of gene regulatory proteins with DNA. *J. Mol. Biol.* **172**:263-282.

154) Taylor, J.D., A.J. Ackroyd and S.E. Halford (1994). The Gel Shift Assay for the Analysis of Protein-DNA Interactions. In *DNA-Protein Interactions, Principles and Protocols (Methods in Molecular Biology, Volume 30)*. Kneale, G.G. (ed.)

155) Kadonaga, J.T. (1991). Purification of Sequence-Specific Binding Proteins by DNA-Affinity Chromatography. *Methods in Enzymology* **208**:10-23.

156) Rosenfeld P.J. and T.J. Kelly (1986). Purification of Nuclear Factor 1 by DNA recognition site affinity chromatography. *J. Biol. Chem.* **261**:1398-1408.

157) Gabrielsen, O.D., G.Hornes, L. Korsnes, A. Rvet and T.B. Øyen (1989). Magnetic DNA affinity purification of yeast transcription factor J - a new purification principle for the ultrarapid isolation of near homogeneous factor. *Nucl. Acids. Res.* **17**:6253-6267.

158) Kadonaga, J.T. (1990) Sequence-specific DNA chromatography. *DNA and Protein Engineering Techniques* **2**:82-87.

159) Sorger, P.K., G. Ammerer and D. Shore (1989). Identification and purification of sequence-specific DNA-binding proteins. In *Protein Function: A Practical approach.* Creighton, T.E. ed. (IRL Press at Oxford University) pp 199-223.

160) Edman, P. and G. Begg (1967). A protein sequenator. *Eur. J. Biochem.* **1**:80

161) Anderson, J.S, B. Svensson and P. Roepstorff (1996). Electrospray ionization and matrix assisted laser desorption / ionization mass spectrometry: powerful analytical tools in recombinant protein chemistry. *Nature Biotechnology* **14**:449-454.

162) Hunt, D.F., J.R. Yates III, J. Shabinowitz, S. Winston and C.R. Hauer (1986). Protein Sequencing by Tandem Mass Spectrometry. *Proc. Natl. Acad. Sci. (USA)* **83**:6233-6237.

163) Mann, M. (1990). Electrospray: Its Potential and Limitations as an Ionization Method for Biomolecules. *Organic Mass Spectrometry* **25**:575.

164) Bieman, K. (1992). Mass spectrometry of peptides and proteins. *Ann.Rev. Biochem.* **61**:977-1010.

165) Arnott, D., J. Shabanowitz and D.F. Hunt (1993). Mass spectrometry of protein and peptides: sensitive and accurate mass measurement and sequence analysis. *Clin. Chem.* **39**:2005-2010.

166) Eng, J.K., A.L. McKormack and J.R. Yates III (1994). An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *American Society for Mass Spectrometry* **5**:976-989.

167) Didier, D.K., J. Schiffenbauer, S.L. Woulfe, M. Zacheis, B.D. Schwartz (1988). Characterisation of the cDNA encoding a protein binding to the major histocompatibility complex class II Y box. *Proc. Natl. Acad. Sci. (USA)* **85**:7322-7326.

168) Vinson, C.R., K.L. LaMarco, P.F. Johnson, W.H. Landschulz and S.L. McKnight (1988). *In situ* detection of sequence specific DNA-binding activity specified by a recombinant bacteriophage. *Genes Dev.* **2**:801-806.

169) Maniatis, T., E.F. Fritsch and J. Sambrook (1982). Molecular Cloning: A Laboratory Manual. Cold Spring Harbour, New York.

170) Singh, H., J.H. LeBowitz, A.S. Baldwin Jr and P.A. Sharp (1988). Molecular Cloning of an Enhancer Binding Protein: Isolation by Screening of an Expression Library with a Recognition Site DNA. *Cell* **52**:415-423.

171) Benton, W.D. and R.W. Davis (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. *Science* **196**:180-182.

172) Huynh, T.V., R.A. Young and R.W. Davies (1985). Constructing and screening cDNA libraries in λGT10 and λGT11. In *DNA Cloning: A Practical Approach (Volume 1)*. Glover, D.M., ed (IRL Press).

173) Kaiser, K. and N.E. Murray (1985). The Use of Phage Lambda Expression Vectors in the Construction of Representative Genomic DNA Libraries. In *DNA Cloning: A Practical Approach (Volume 1)*. Glover, D.M., ed (IRL Press).

174) Singh, H., R.G. Clerc and J.H. LeBowitz (1989). Molecular Cloning of Sequence-Specific DNA Binding Proteins Using Recognition Site Probes. *BioTechniques* **7**:252-261.

175) Cohen, I. and W.F. Reynolds (1991). The Xenopus YB3 protein binds the B box element of the class III promoter. *Nucl. Acids. Res.* **19**:4753-4759

176) Sakura, H., T. Moukawa, F. Imamoto, K. Yasuda and S. Ishii (1988). Two human genes isolated by a novel method encode DNA-binding proteins containing a common region of homology. *Gene* **73**:499-507.

177) Hasegawa, T., P.W. Doetsch, K.K. Hamilton, A.M. Martin, S.A. Okenquist, J. Lenz and J.M Boss (1991). DNA-binding properties of YB-1 and dbpA: binding to double stranded, single stranded and abasic site containing DNAs. *Nucl. Acids. Res.* **19**:4915-4920.

178) Clerk, R.G., L.M. Corcoran, J.H. LeBowitz, D. Baltimore and P.A. Sharpe (1988). The B-cell specific Oct2-protein contains POU and homeobox-type domains. *Genes Dev.* **2**:1570-1581.

179) Staudt, L.M., R.G. Clerc, H. Singh, J.H. LeBowitz, P.A. Sharp and D. Baltimore (1988). Cloning of a lymphoid specific cDNA encoding protein binding the regulatory octamer DNA motif. *Science* **241**:577-580.

180) King, M.W. (1997). Rapid and Nonradioactive Screening of Recombinant Libraries by PCR. In *PCR Cloning Protocols, From Molecular Cloning to Genetic Engineering (Methods in Molecular Biology, Vol 67).* White, B.A. (ed.) pp 331-338.

181) Chenchik, A., F. Moqadam and P. Siebert (1995). Marathon cDNA amplification: A new method for cloning full-length cDNAs. *ClonTechniques* **1**:5-8.

182) Chung, C.T., S.L. Niemela and R.H. Miller (1989). One-step preparation of competent *Escherichia coli*: Transformation and storage of bacterial cells in the same solution. *Proc. Natl. Acad. Sci. (USA)* **86**:2172-2175.

183) Hentschel, C.C. and M.L. Birnstiel (1981). The organization and expression of histone gene families. *Cell* **25**:301-313.

184) Chirgwin, J.M., A.E. Przybyla, R.J. MacDonald, and W.J. Rutter (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* **18**:5294-5299.

185) Gorman, C.M., D.R. Gies and G. McCray (1990). Transient production of proteins using an adenovirus transformed cell line. *DNA and Protein Engineering Techniques.* **2**: 3-10.

186) Lin, K-H. and S.Y. Cheng (1991). An efficient method to purify active eukaryotic proteins from the inclusion bodies in *Escherichia coli. BioTechniques* **11**:748-753.

187) Jiang, G., L. Nepomuceno, K. Hopkins and F.M. Sladek (1995). Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol. Cell. Biol.* **15**:5131-5143.

188) Morris, G.F. and W.F. Marzluff (1983). A factor in sea urchin eggs inhibits transcription in isolated nuclei by sea urchin RNA polymerase III. *Biochemistry* **22**:645-653.

189) Garner, M.M. and A. Revzin (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* **9**:3047-3060.

190) Dailey, L., S.M. Hanly, R.G. Roeder and N. Heintz (1988). Distinct transcription factors bind specifically to two regions of the human histone H4 promoter. *Proc. Natl. Acad. Sci. (USA)* **83**:7241-7245.

191) Patterton, D. (1992) Masters thesis. University of Cape Town.

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

193) Gish, W. and D.J. States (1993). Identification of protein coding regions by database similarity searches. *Nat. Genet.* **3**:266-272.

194) Lee, J.J., R.J. Shott, S.J. Rose, T.L. Thomas, R.J. Britten and E.H. Davidson (1984). Sea Urchin Actin Gene Subtypes: Gene Number, Linkage and Evolution. *J. Mol. Biol.* **172**:149-176.

195) Smith et al (1994), unpublished. Database accession number Z30662.

196) Schnitzler, G.R., W.H. Fischer and R.A. Firtel (1994). Cloning and characterization of the G-box binding factor, an essential component of the developmental switch between early and late development in *Dictyostelium*. *Genes Dev.* **8**:502-514.

197) Wilson et al (1994). Database accession number U41020. *Nature* **368**:32-38.

198) Saha, V., T. Chaplin, A. Gregorini. P. Ayton and B.D. Young (1995). The leukemia-associated-protein (LAP) domain, a cysteine-rich motif, is present in a wide range of proteins, including MLL, AF10 and MLLT6 proteins. *Proc. Natl. Acad. Sci. (USA)* **92**:9737-9741.

199) Hsu, T., J.A. Gogos, S.A. Kirsh and F.C. Kafatos (1992). Multiple Zinc Finger Forms Resulting from Developmentally Alternative Splicing of a Transcription Factor Gene. *Science* **257**:1946-1950.

200) Birkelund, S., A.G. Lundemose and G. Christiansen (1990). The 75-kilodalton cytoplasmic *Chlamydia trachomatis* L2 polypeptideis a DnaK-like protein. *Immun* **58**:2098-2104.

201) Webster, P.J. and T.E. Mansour (1992). Conserved classes of homeodomains in *Schistosoma mansoni*, an early bilateral metozoan. *Mech. Dev.* **38**:25-32.

202) Hillier et al (unpublished). Database accession number P25980.

203) Borson, N.D., W.D. Sato and L.R. Drewes (1992). A lock-docking oligo(dT) primer for 5' and 3' Race PCR. *PCR Methods and Applications* **2**:144-148.

204) Balbás, P. (1997). Designing Expression Plasmid Vectors in *E.coli*. In *Recombinant Gene Expression Protocols* (Methods in Molecular Biology, Volume 62).

205) Tuan, R.S. (1997). Overview of Experimental Strategies for the Expression of Recombinant Proteins. In *Recombinant Gene Expression Protocols. (Methods in Molecular Biology, Volume 62).*

206) Short, J.M. (1988). Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucl. Acids. Res.* **16**:7583-7600.

207) Ford, K.G., A.J. Whitmarsh and D.P. Hornby (1994). Overexpression and Purification of Eukaryotic Transcription Factors as Glutathione-S-Transferase Fusions in *E.coli*. In *DNA-Protein Interactions: Principles and Protocols (Methods in Molecular Biology, Volume 30)*, Kneale, G.G., ed.

208) Kaufman, R.J. (1997). Overview of Vector Design for Mammalian Gene Expression. In *Recombinant Gene Expression Protocols (Methods in Molecular Biology, Volume 62).*

209) Rhodes, D. (1989). Analysis of sequence specific DNA-binding proteins. In *Protein Function: A Practical Approach.* Creighton, T.E. ed. (IRL Press at Oxford Univesity) pp177-198.

210) Bensadouan, A. and D. Weinstein (1976). Assay of proteins in the presence of interfering materials. *Anal. Biochem.* **70**:241-250.

211) Matsudaira, P. (1993). In *A Practical Guide to Protein and Peptide Purification for Microsequencing.* Matsudaira, P. (ed).

212) Hewish, D.R. and L.A. Burgoyne (1973). Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem. Biophys. Res. Commun.* **52**:504-510.

213) Stone, K.L. and K.R. Williams (1993). Enzymatic Digestion of Proteins. In *A Practical Guide to Protein and Peptide Purification for Microsequencing.* Matsudaira, P. (ed.)

214) Griffin, P.R., J.A. Coffman, L.E. Hood and J.R. Yates III (1991). Structural Analysis of Proteins by Capillary HPLC Electrospray Tandem Mass Spectrometry. *Intl. J. Mass Spectrom. Ion Proc.* **111**:131.

215) Descombes, P. and U. Schibler (1991). A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. *Cell* **67**:569-579.

216) Ossipow, V., P. Descombes and U. Schibler (1993). CCAAT/enhancer-binding protein mRNA is translated into multiple proteins with different transcription activation potentials. *Proc. Natl. Acad. Sci. (USA)* **90**:8219-8223.

217) Bean, T.L. and P.A. Ney (1997). Multiple regions of p45 NF-E2 are required for beta-globin gene expression in erythroid cells. *Nucleic Acids Res* **25**:2509-2515.

218) Friedman, A.D. and S.L. McKnight (1990). Identification of two polypeptide segments of CCAAT/enhancer-binding protein required for transcriptional activation of the serum albumin gene. *Genes Dev.* **4**:1416-1426.

219) Davis, R.L., P. Cheng, A.B. Lassar and H. Weintraub (1990). The MyoD DNA binding domain contains a recognition code for muscle specific gene activation. *Cell* **60**:733-764.

220) Voronova, A. and D. Baltimore (1990). Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. *Proc. Natl. Acad. Sci. (USA)* **87**:4722-4726.

221) Kolluri, R. and A.J. Kinniburgh (1991). Full length cDNA sequence encoding a nuclease-sensitive element DNA binding protein. *Nucleic Acids Res.* 19: 4771.

222) LeBlanc, J.M. and L.A. Leinwand (1991). The diversity of myosin-based contractile systems in eukaryotic cells. *Am. Zool.* **31**:514-521.

223) Pei, D.Q. and C.H. Shih (1991). An "attenuator domain" is sandwiched by two distinct transactivation domains in the transcription factor C/EBP. *Mol. Cell. Biol.* **11**:1480-1487.

224) Hann, S.R., M.W. King, D.L. Bentley, C.W. Anderson and R.N. Eisenman (1988). A non-AUG translational initiation in c-myc exon 1 generates a N-terminally distinct protein while synthesis is disrupted in Burkit's lymphomas. *Cell* **34**:185-195.

225) Hann, S.R., M. Dixit, R.C. Sears and L. Sealy (1994). The alternatively initiated c-Myc proteins differentially regulate transcription through a non-canonical DNA-binding site. *Genes Dev.* **8**:2441-2452.

226) Cao, Z., R. Umek and S. McKnight (1991). Regulated expression of three C/EBP isoforms during adipose conversion of 3T3-L1 cells. *Genes Dev.* **5**:1538-1552.

227) Abate, C., D. Lak, R. Gentz, F.J. Rauscher III and T. Curran (1990). Expression and purification of the leucine zipper and DNA-binding domains of Fos and Jun: both Fos and Jun contact DNA directly. *Proc. Natl. Acad. Sci. (USA)* **87**:1032-1036.

228) Johnson, R.S., S.A. Martin, K. Bieman, J.T. Stults and J.T. Watson (1987). Novel Fragmentation Process by Collision-induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine. *Anal. Chem* **59**:2621-2625.

229) Dwarki, V.J., M. Montimy and I.M. Verma (1990). Both the basic region and the "leucine zipper" domain of the cyclic AMP response element binding (CREB) protein are essential for transcriptional activation. *EMBO J.* **9**:225-232.

230) de Groot, P. (1982). Masters thesis. University of Cape Town.

231) Voet, D. and J.G. Voet (1992). Biochemie. VCH Verlagsgesellschaft mbH, Weinheim. pp 856-876. *Proc. Natl. Acad. Sci. (USA)* **88:**6219-6223.

232) Studier, F.W. and B.A. Moffatt (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189:**113-130.

233) Studier, F.W., A.H. Rosenberg, J.J. Dunn and J.W. Dubendorff (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods in Enzymology* **185:**60-89.

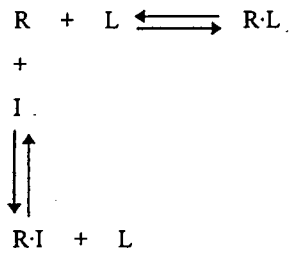
234) Grodberg, J. and J.J. Dunn (1988). ompT encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. *J. Bacteriology* **170:**1245-1253.

235) Moffatt, B.A. and F.W. Studier (1987). T7 lysozyme inhibits transcription by T7 RNA polymerase. *Cell* 49:221-227.

Appendix I

Scatchard Analysis

The model used to investigate the receptor-ligand (or protein-DNA) competition studies is analogous to the Michaelis-Menten Competitive Inhibition Rate Law which describes enzyme inhibition (231).



where I is the competing ligand whose dissociation constant with the receptor is expressed as:

$$K_I = \frac{[R][I]}{[R \cdot I]}$$

by derivation:

$$[R \cdot L] = \frac{[R]_T[L]}{K_L\left(1 + \frac{[I]}{K_I} + [L]\right)} \qquad (1)$$

The relative affinities of a ligand and an inhibitor may therefore be determined by dividing equation (1) in the presence of inhibitor with that in the absence of inhibitor:

$$\frac{[R \cdot L]_I}{[R \cdot L]_0} = \frac{KL = [L]}{K_L\left(1 + \frac{[I]}{K_I} + [L]\right)} \qquad (2)$$

When this ratio is 0.5 (50 %) inhibition, the competitor concentration is referred to as $[I_{50}]$.
Solving equation (2) for K_I at 50 % inhibition:

$$K_I = \frac{[I_{50}]}{1 + \frac{[L]}{K_L}}$$

where I = L, and I = unlabelled E/H fragment
L = free labelled E/H fragment

solving equation (2) for K_I results in

$$\Rightarrow \quad K_I = K_D \qquad (3)$$

(3) in (2):

$$K_D = [I_{50}] - [L]$$

where $[I_{50}]$ = concentration of unlabelled E/H fragment at 50 % competition
$[I_{50}] = (4 \pm 2) \times 10^{-10}$ M (see graph)
[L] = the corresponding concentration of free labelled E/H fragment

$\Rightarrow$ at 50 % competition $[L] = (4 \pm 2) \times 10^{-11}$ M (53 cpm)

since (234 ± 67) cpm corresponds to $(1.75 \pm 0.5) \times 10^{-10}$ M (see Table 3.2)

$$\Rightarrow \quad K_D = (4 \pm 2) \times 10^{-10}\ M - (0.4 \pm 0.2) \times 10^{-10}\ M$$

$$K_D = (3.6 \pm 2) \times 10^{-10}\ M$$

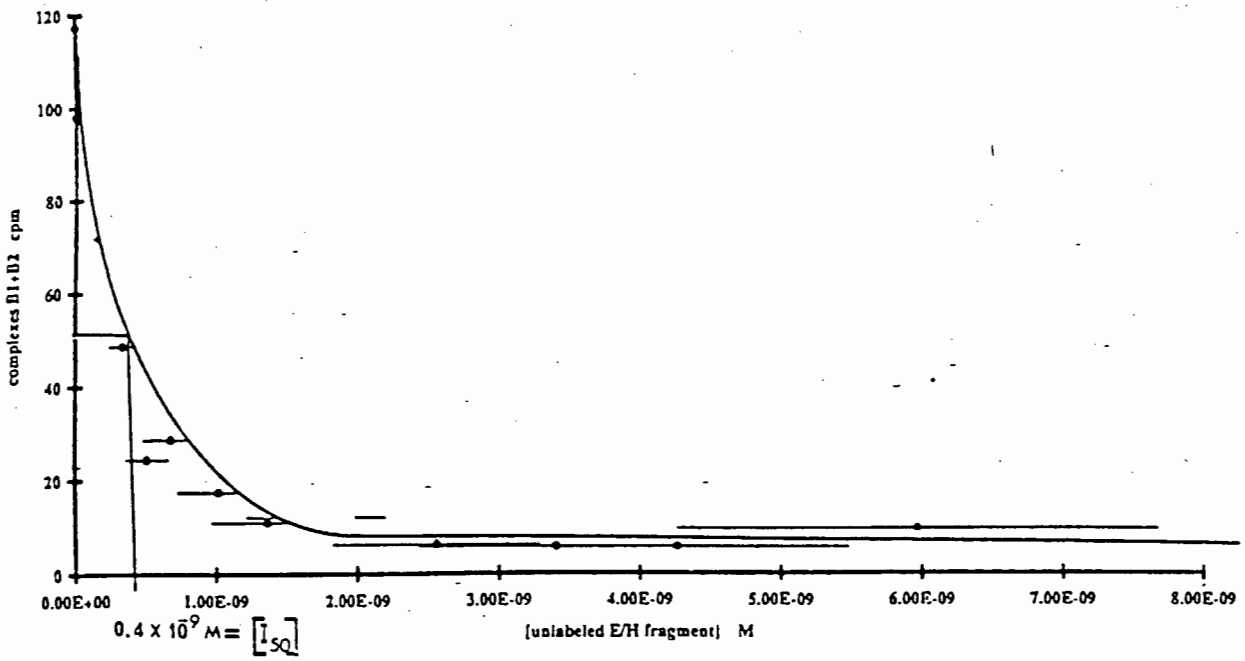


Fig (i) Binding of suGF1 to the E/H Fragment

The amount of radioactivity in the unbound DNA and suGF1-DNA complexes (see fig 3.3) was determined by Instant Imager 2024 (see table 3.2). The amount of radioactivity in the suGF-DNA-complexes was plotted against the amount of radioactivity in the unbound fraction in order to determine the dissociation constant of suGF1 (in nuclear extract) with respect to the G-string in the E/H fragment, using Scatchard analysis.

Strongylocentrotus purpuratus cDNA library

DATE: April 7, 1993

TO: DR. ROBERT ZELLER, Caltech

FROM: Dr. Kenneth Fong, Director of Custom Library Synthesis Dept.
Performed by: Cynthia Chang Ph.D., Research Scientist III

SUBJECT: Custom synthesis of cDNA library in Lambda Zap II oligo dT-primed (x), random-primed (x).

TITER & VOLUME: ca 10^{10}/ml (4 xl ml)

STORAGE: 4°C or -70°C when aliquoted for long-term storage after adding 7% DMSO or 50% sterile glycerol.

SOURCE OF mRNA: 24H

CLONING VECTOR: Lambda Zap II (See map on reverse side)

CLONING SITE: Eco Rl

SELECTION CRITERIA: Clear from blue (parental) plaques*

% OF CLEAR PLAQUES: 85%

NUMBER OF INDEPENDENT CLEAR PLAQUES (Clones): 1.3 x 10^6

INSERT SIZE RANGE: 1.0 to 4-5 Kb (as revealed on an autoradiogram before cDNA was cloned into Zap II arms)
Average: 1.9 Kb
(see PCR data on insert sizes)

USAGE: For immunoscreening or probe screening, plate an appropriate dilution of the library on *E. coli* strain BB4 as described in theLibrary Protocol Handbook enclosed.

* Clear plaques are recombinant phage clones and blue plaques are parental lambda Zap II phage. Occasionally, particular recombinant phage plaques produce a small but detectable blue color. The detection of blue plaques can be achieved by adding 35 µl of 40 mg/ml X-Gal (Cat. # 8060) to 2.5 ml of LB soft agar before plating.

(X-Gal is NOT soluble in water. It can be dissolved in Dimethylformamide (DMF). DMF dissolves plastic surfaces. X-Gal dissolved in DMF should be added directly into soft agar).

Appendix III

DNA Sequence Analysis of Clone 2

Clone 2 (a 2.2 kb insert) was sequenced with primers T3 and T7 from the Bluescript plasmid giving partial sequence information for (a) the plus strand and (b) the minus strand, respectively. Analysis of putative open reading frames for the plus strand (c) and the minus strand (d) was performed using the 'MAP' algorithm in the GCG program.

```
(a)     plus strand (primer T3)
        length sequenced: 236 nt

        1  ttatatttgt ttgtatatga aatatataca ataccttgtg agagcttcaa
       51  acatataaat gttaaggatt cgttcaagtt atgaagaaaa aataatcttt
      101  caaggttttt tgaaggattt cctagcattc aggttaacaa ttttgaaaaa
      151  tatgatttgt ggcaattttg gcatccaatt tgtttacgaa acgtgttcag
      201  aaacaatatg ccaggcacat ttatcacact ctacta

(b)     minus strand (primer T7)
        length sequenced: 223 nt

        1  tcgaattccg gaccattttt tttctatcgc tgcctgatta gctgtaggtt
       51  taatgatagt tttgatgagt agtttagttc tcgtatcgac caacaaatgt
      101  cctaaaaaca gtggtctttt ttctgaaaaa tcgcctggtg ttgccgctaa
      201  agcgtttgaa attaaatcgg cca

(c) analysis of translational reading frame for the plus strand

         ttatatttgtttgtatatgaaatatatacaataccttgtgagagcttcaaacatataaat
        1 ---------+---------+---------+---------+---------+---------+
          aatataaacaaacatatactttatatatgttatggaacactctcgaagtttgtatattta

b                                                                    M

          gttaaggattcgttcaagttatgaagaaaaaataatctttcaaggttttttgaaggattt
       61 ---------+---------+---------+---------+---------+---------+
          caattcctaagcaagttcaatacttctttttttattagaaagttccaaaaaacttcctaaa
a                               R  K  N  N  L  S  R  F  F  E  G  F
b          L  R  I  R  S  S  Y  E  E  K  I  I  F  Q  G  F  L  K  D  F

          cctagcattcaggttaacaattttgaaaaatatgatttgtggcaattttggcatccaatt
      121 ---------+---------+---------+---------+---------+---------+
          ggatcgtaagtccaattgttaaaacttttatactaaacaccgttaaaaccgtaggttaa
a         P  S  I  Q  V  N  N  F  E  K  Y  D  L  W  Q  F  W  H  P  I
b          L  A  F  R  L  T  I  L  K  N  M  I  C  G  N  F  G  I  Q  F
c                                        F  V  A  I  L  A  S  N  L

          tgtttacgaaacgtgttcagaaacaatatgccaggcacatttatcacactctacta
      181 ---------+---------+---------+---------+---------+------ 236
          acaaatgctttgcacaagtctttgttatacggtccgtgtaaatagtgtgagatgat
a         C  L  R  N  V  F  R  N  N  M  P  G  T  F  I  T  L  Y
b          V  Y  E  T  C  S  E  T  I  C  Q  A  H  L  S  H  S  T
c           F  T  K  R  V  Q  K  Q  Y  A  R  H  I  Y  H  T  L  L

(d) analysis of translational reading frame for the minus strand

          tcgaattccggaccatttttttttctatcgctgcctgattagctgtaggtttaatgatagt
        1 ---------+---------+---------+---------+---------+---------+
          agcttaaggcctggtaaaaaaaagatagcgacggactaatcgacatccaaattactatca
            R  I  G  S  W  K  K  E  I  A  A  Q  N  A  T  P  K  I  I  T

          tttgatgagtagtttagttctcgtatcgaccaacaaatgtcctaaaaacagtggtctttt
       61 ---------+---------+---------+---------+---------+---------+
          aaactactcatcaaatcaagagcatagctggttgtttacaggatttttgtcaccagaaaa
            K  I  L  L  K  T  R  T  D  V  L  L  H  G  L  F  L  P  R  K

          ttctgaaaaatcgcctggtgttgccgctaatgccgaaaaaaaattacttttctgtcatgt
      121 ---------+---------+---------+---------+---------+---------+
          aagacttttagcggaccacaacggcgattacggctttttttaatgaaaagacagtaca
            E  S  F  D  G  P  T  A  A  L  A  S  F  F  N  S  K  Q  *  T

          caactgtttcgtttaccacaagcgtttgaaattaaatcggcca
      181 ---------+---------+---------+---------+--- 223
          gttgacaaagcaaatggtgttcgcaaactttaatttagccggt
            L  Q  K  T  *  W  L  R  K  F  N  F  R  G
```

Appendix IV

DNA Sequence Analysis of Clone 6

Clone 6 (an 850 bp insert) was sequenced with primers T3 and T7 from the Bluescript plasmid giving partial sequence information for (a) the plus strand and (b) the minus strand, respectively. Analysis of putative open reading frames for the plus strand (c) and the minus strand (d) was performed using the 'MAP' algorithm in the GCG program.

```
(a)     plus strand (primer T3)
        length sequenced: 655 nt


       1  cattttcctc tctaaagaaa attgctcata aggacagctt catacacact
      51  taagttttta tttggaatat tacaatcttt aaaaatatca aaatagggct
     101  actcaatcca ttgtattcaa agaaattggc atcctttcga tctaggagaa
     151  agggatccgc tcagttttta gaagtcataa aatttgaatt tttcctaaac
     201  tttaagaaac aacccatgaa tattcatatc aataggtaaa atttggtata
     251  atcgcaagta aactacatcc agattgcaat gaatgagtga tcacagtttt
     301  ctggccatat gttgagatac acctcgaaac caattttaag tcggtacatt
     351  gcatagatgg aatcgacata acacatcaac atgctagtgt gaggttctct
     401  cttgagaaaa agaacagtac ttgtatcgtg tgtgggggat ttggtctgtt
     451  acctcaaagt ctggcggcat tcattggtta agttttcact ctcccacata
     501  cttcaaaggt ttctgatata atgtcaaggg aggcaaagag ataatcgttt
     551  gcatttattg gttaaaagat ggtttcataa ctctatgacc tttgaccttt
     601  gaggtgaaag gttaaacgca ctcaactatg ctgttctatg ccacacctaa
     651  gaccc



(b)     minus strand (primer T7)
        length sequenced: 181 nt


       1  aaaaaaaaaa ttctcattcc aatgttctct taaaagagac cactgatatt
      51  agatatgctt ctttcgactc ttgtcgacct ttgacctagc tgccatattt
     101  tgaaagaaat tctatgtgtg ataaaagcag tttgttggat agtgtttatt
     151  tcaatcgttt tataccgagc gggtcttagg t



(c) analysis of translational reading frame for the plus strand

          cattttcctctctaaagaaaattgctcataaggacagcttcatacacacttaagttttta
       1  ---------+---------+---------+---------+---------+---------+
          gtaaaaggagagatttcttttaacgagtattcctgtcgaagtatgtgtgaattcaaaaat
c                                                           V  F  I

          tttggaatattacaatctttaaaaatatcaaaatagggctactcaatccattgtattcaa
      61  ---------+---------+---------+---------+---------+---------+
          aaaccttataatgttagaaatttttatagttttatcccgatgagttaggtaacataagtt
c           W  N  I  T  I  F  K  N  I  K  I  G  L  L  N  P  L  Y  S  K

          agaaattggcatcctttcgatctaggagaaagggatccgctcagtttttagaagtcataa
     121  ---------+---------+---------+---------+---------+---------+
          tctttaaccgtaggaaagctagatcctctttccctaggcgagtcaaaaatcttcagtatt
c           K  L  A  S  F  R  S  R  R  K  G  S  A  Q  F  L  E  V  I  K

          aatttgaatttttcctaaactttaagaaacaacccatgaatattcatatcaataggtaaa
     181  ---------+---------+---------+---------+---------+---------+
          ttaaacttaaaaaggatttgaaattctttgttgggtacttataagtatagttatccattt
c           F  E  F  F  L  N  F  K  K  Q  P  M  N  I  H  I  N  R

          atttggtataatcgcaagtaaactacatccagattgcaatgaatgagtgatcacagtttt
     241  ---------+---------+---------+---------+---------+---------+
          taaaccatattagcgttcatttgatgtaggtctaacgttacttactcactagtgtcaaaa
b                                                       V  I  T  V  F

          ctggccatatgttgagatacacctcgaaaccaattttaagtcggtacattgcatagatgg
     301  ---------+---------+---------+---------+---------+---------+
          gaccggtatacaactctatgtggagctttggttaaaattcagccatgtaacgtatctacc
b          W  P  Y  V  E  I  H  L  E  T  N  F  K  S  V  H  C  I  D  G
```

```
        aatcgacataacacatcaacatgctagtgtgaggttctctcttgagaaaaagaacagtac
    361 ---------+---------+---------+---------+---------+---------+
        ttagctgtattgtgtagttgtacgatcacactccaagagagaactctttttcttgtcatg
b        I  D  I  T  H  Q  H  A  S  V  R  F  S  L  E  K  K  N  S  T
c                                    G  S  L  L  R  K  R  T  V  L

        ttgtatcgtgtgtgggggatttggtctgttacctcaaagtctggcggcattcattggtta
    421 ---------+---------+---------+---------+---------+---------+
        aacatagcacacacccccctaaaccagacaatggagtttcagaccgccgtaagtaaccaat
b        C  I  V  C  G  G  F  G  L  L  P  Q  S  L  A  A  F  I  G
c         V  S  C  V  G  D  L  V  C  Y  L  K  V  W  R  H  S  L  V  K

        agttttcactctcccacatacttcaaaggtttctgatataatgtcaagggaggcaaagag
    481 ---------+---------+---------+---------+---------+---------+
        tcaaaagtgagagggtgtatgaagtttccaaagactatattacagttccctccgtttctc
c         F  S  L  S  H  I  L  Q  R  F  L  I

        ataatcgtttgcatttattggttaaaagatggtttcataactctatgacctttgaccttt
    541 ---------+---------+---------+---------+---------+---------+
        tattagcaaacgtaaataaccaatttttctaccaaagtattgagatactggaaactggaaa

        gaggtgaaaggttaaacgcactcaactatgctgttctatgccacacctaagaccc
    601 ---------+---------+---------+---------+---------+----- 655
        ctccactttccaatttgcgtgagttgatacgacaagatacggtgtggattctggg
```

(d) analysis of translational reading frame for the minus strand

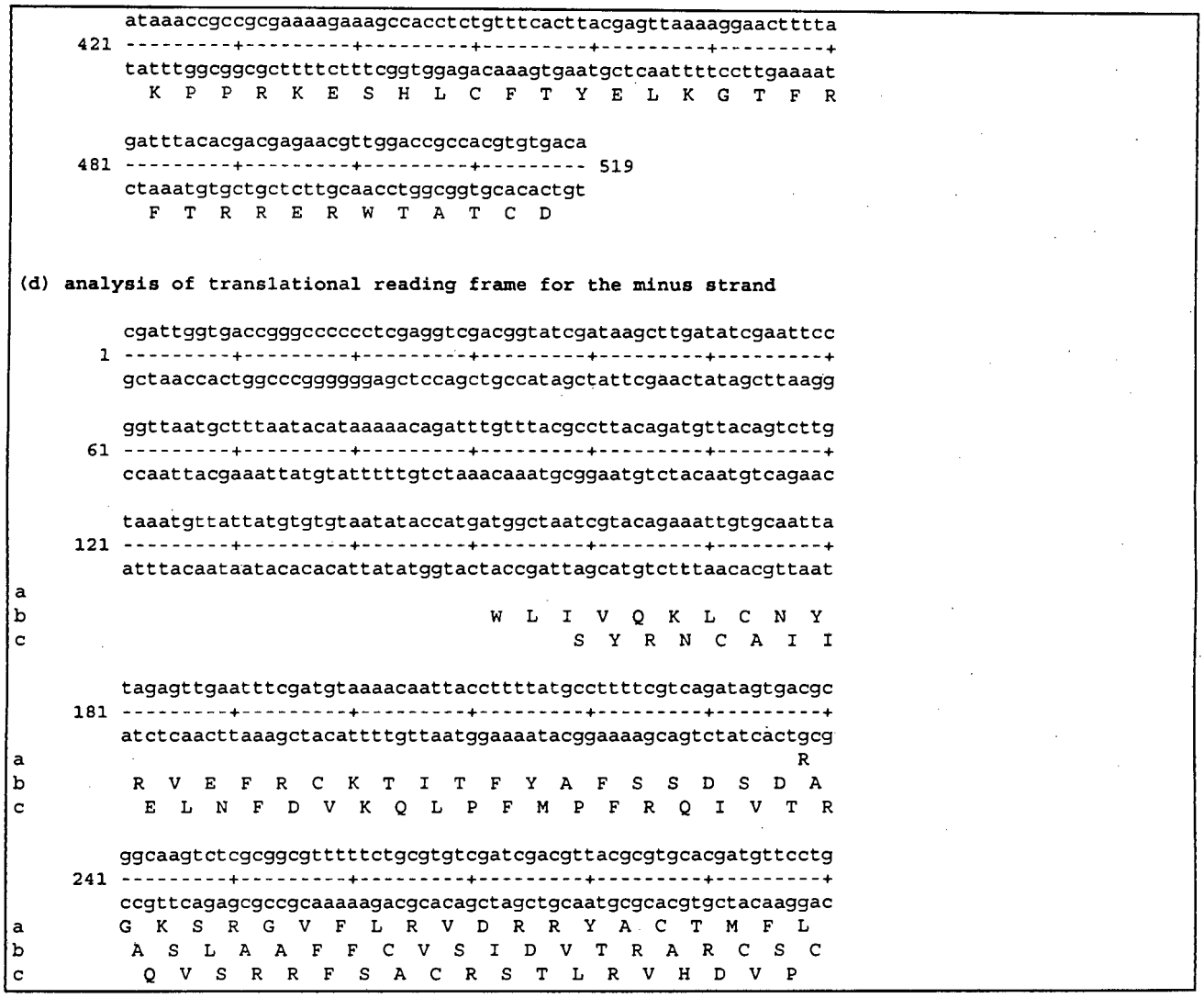
```
        aaaaaaaaaattctcattccaatgttctcttaaaagagaccactgatattagatatgctt
      1 ---------+---------+---------+---------+---------+---------+
        ttttttttttaagagtaaggttacaagagaattttctctggtgactataatctatacgaa
                                          F  L  G  S  I  N  S  I  S

        ctttcgactcttgtcgacctttgacctagctgccatattttgaaagaaattctatgtgtg
     61 ---------+---------+---------+---------+---------+---------+
        gaaagctgagaacagctggaaactggatcgacggtataaaacttttctttaagatacacac
         R  E  V  R  T  S  R  Q  G  L  Q  W  I  K  F  S  I  R  H  T

        ataaaagcagtttgttggatagtgtttatttcaatcgttttataccgagcgggtcttagg
    121 ---------+---------+---------+---------+---------+---------+
        tattttcgtcaaacaacctatcacaaataaagttagcaaaatatggctcgcccagaatcc
         I  F  A  T  Q  Q  I  T  N  I  E  I  T  K  Y  R  A  P  R  L
```

Appendix V

Sequence Analysis of Clone 11

Clone 11 (a 900 bp insert) was sequenced with primers T3 and T7 from the Bluescript plasmid giving partial sequence information for (a) the plus strand and (b) the minus strand, respectively. Analysis of putative open reading frames for the plus strand (c) and the minus strand (d) was performed using the 'MAP' algorithm in the GCG program.

```
(a)     plus strand (primer T3)
        length sequenced: 514 nt

       1  cccggctgat atcaagaata gaaacaaagt ttaacaaaga atgtgcgaac
      51  tgtgttctta tcttgaatgt tcatgacgac gacgacgagg acgaaaatga
     101  tgatgatgtt aatgatagtg ggtggtggtg gtgattgtga caatgttgag
     151  gatgttgatg gtgatgatat tgttgatggt gcattcatta tgactgtaat
     201  gacggcaacg gtcaatgatg atgttggtga taatgacgat ggtgataatg
     251  atgatcattg cggcgacaat tatattggtg atgaggaaaa ggagactaat
     301  gacgggtccc atccacaact tgtacaaaca attggcctcg tcgaactagg
     351  tggtatactc gggaaatcat caaaagatct aactgtttta aaggttcgtt
     401  cagatatgac actgtataaa ccgccgcgaa aagaaagcca cctctgtttc
     451  acttacgagt taaaaggaac ttttagattt acacgacgag aacgttggac
     501  cgccacgtgt gaca


(b)     minus strand (primer T7)
        length sequenced: 183 nt

       1  tttaatacat aaaaacagat ttgtttacgc cttacagatg ttacagtctt
      51  gtaaatgatt attatgtgtg taatatacca tgatggctaa tcgtacagaa
     101  attgtgcaat tatagagttg aatttcgatg taaaacaatt acctttttatg
     151  ccttttcgtc agatagtgac gcggcaagtc tcg


(c) analysis of translational reading frame for the plus strand

         ggaattccggctgatatcaagaatagaaacaaagtttaacaaagaatgtgcgaactgtgt
       1 ---------+---------+---------+---------+---------+---------+
         ccttaaggccgactatagttcttatctttgtttcaaattgtttcttacacgcttgacaca

         tcttatcttgaatgttcatgacgacgacgacgaggacgaaaatgatgatgatgttaatga
      61 ---------+---------+---------+---------+---------+---------+
         agaatagaacttacaagtactgctgctgctgctcctgcttttactactactacaattact
                    M  F  M  T  T  T  T  R  T  K  M  M  M  M  L  M  I

         tagtgggtggtggtggtgattgtgacaatgttgaggatgttgatggtgatgatattgttg
     121 ---------+---------+---------+---------+---------+---------+
         atcacccaccaccaccactaacactgttacaactcctacaactaccactactataacaac
           V  G  G  G  G  D  C  D  N  V  E  D  V  D  G  D  D  I  V  D

         atggtgcattcattatgactgtaatgacggcaacggtcaatgatgatgttggtgataatg
     181 ---------+---------+---------+---------+---------+---------+
         taccacgtaagtaatactgacattactgccgttgccagttactactacaaccactattac
           G  A  F  I  M  T  V  M  T  A  T  V  N  D  D  V  G  D  N  D

         acgatggtgataatgatgatcattgcggcgacaattatattggtgatgaggaaaaggaga
     241 ---------+---------+---------+---------+---------+---------+
         tgctaccactattactactagtaacgccgctgttaatataaccactactccttttcctct
           D  G  D  N  D  D  H  C  G  D  N  Y  I  G  D  E  E  K  E  T

         ctaatgacgggtcccatccacaacttgtacaaacaattggcctcgtcgaactaggtggta
     301 ---------+---------+---------+---------+---------+---------+
         gattactgcccagggtaggtgttgaacatgtttgttaaccggagcagcttgatccaccat
           N  D  G  S  H  P  Q  L  V  Q  T  I  G  L  V  E  L  G  G  I

         tactcgggaaatcatcaaaagatctaactgttttaaaggttcgttcagatatgacactgt
     361 ---------+---------+---------+---------+---------+---------+
         atgagccctttagtagttttctagattgacaaaatttccaagcaagtctatactgtgaca
           L  G  K  S  S  K  D  L  T  V  L  K  V  R  S  D  M  T  L  Y
```


ataaaccgccgcgaaaagaaagccacctctgtttcacttacgagttaaaaggaactttta
421 ---------+---------+---------+---------+---------+---------+
tatttggcggcgcttttctttcggtggagacaaagtgaatgctcaattttccttgaaaat
K P P R K E S H L C F T Y E L K G T F R
gatttacacgacgagaacgttggaccgccacgtgtgaca
481 ---------+---------+---------+--------- 519
ctaaatgtgctgctcttgcaacctggcggtgcacactgt
F T R R E R W T A T C D
(d) analysis of translational reading frame for the minus strand
cgattggtgaccgggcccccctcgaggtcgacggtatcgataagcttgatatcgaattcc
1 ---------+---------+---------+---------+---------+---------+
gctaaccactggcccgggggagctccagctgccatagctattcgaactatagcttaagg
ggttaatgctttaatacataaaaacagatttgtttacgccttacagatgttacagtcttg
61 ---------+---------+---------+---------+---------+---------+
ccaattacgaaattatgtatttttgtctaaacaaatgcggaatgtctacaatgtcagaac
taaatgttattatgtgtgtaatataccatgatggctaatcgtacagaaattgtgcaatta
121 ---------+---------+---------+---------+---------+---------+
atttacaataatacacacattatatggtactaccgattagcatgtctttaacacgttaat
a
b W L I V Q K L C N Y
c S Y R N C A I I
tagagttgaatttcgatgtaaaacaattacctttttatgccttttcgtcagatagtgacgc
181 ---------+---------+---------+---------+---------+---------+
atctcaacttaaagctacattttgttaatggaaaatacggaaaagcagtctatcactgcg
a R
b R V E F R C K T I T F Y A F S S D S D A
c E L N F D V K Q L P F M P F R Q I V T R
ggcaagtctcgcggcgtttttctgcgtgtcgatcgacgttacgcgtgcacgatgttcctg
241 ---------+---------+---------+---------+---------+---------+
ccgttcagagcgccgcaaaaagacgcacagctagctgcaatgcgcacgtgctacaaggac
a G K S R G V F L R V D R R Y A C T M F L
b A S L A A F F C V S I D V T R A R C S C
c Q V S R R F S A C R S T L R V H D V P

Appendix VI

DNA Sequence Analysis of Clone 16

Clone 16 (a 2.4 kb insert) was sequenced with primers T3 and T7 from the Bluescript plasmid giving partial sequence information for (a) the plus strand and (b) the minus strand, respectively. Analysis of putative open reading frames for the plus strand (c) and the minus strand (d) was performed using the 'MAP' algorithm in the GCG program.

```
(a)     plus strand (primer T3)
        length sequenced: 622 nt

       1  gggatctcta ggagaaatga aagatggtta agcaagcttt aaacttatca
      51  gggattttgc attctcacat ggtacaaaat tgtgtattgc tttctctgtt
     101  gctgctaaag gagcgataca ttgtacctat atttcactcc tatattcatc
     151  acggcaaatt tatactccaa acttgagaca cgatttcgta gccgtagacc
     201  gcagccaaaa ttttgtctat tattgatcac ttgtgttcat cttgcacttc
     251  tgctgttttg taattcttgt gctattaatt tcaaatgtgt tttctgtttg
     301  tggtgctgct caaaattgta atcgatctac atattcttaa agtgaaggtt
     351  acgatttta taatctatct gtacaaacaa aatgaagggt gaaagtatct
     401  tgttgtatag gcatattatg tatgatcgta taaatgtaaa aaagcacaca
     451  cagaaaaaaa atctcaaaat gttatttatt tgtgtgtatg gccatctaag
     501  atgtactatg tacatgtttt ctgtaccaat ctggaaagga ccatggggaa
     551  agaattacta cttcccaccc tcccaccaag tttaacggcc caacataata
     601  tttgaaaaaa aaactattaa cc

(b)     minus strand (primer T7)
        length sequenced: 156 nt

       1  ttttccttta caacgcacag gatatttcac ttcccgtatc ttcatcccca
      51  tccaaggatc ctgttatatg aaatggatga cttgggtgtg gtacatgtac
     101  gtcatttttt aaacagaaat caacagatac agtacctcca cagaatgtaa
     151  aacatg


(c) analysis of translational reading frame for the plus strand

          gggatctctaggagaaatgaaagatggttaagcaagctttaaacttatcagggattttgc
        1 ---------+---------+---------+---------+---------+---------+
          ccctagagatcctctttactttctaccaattcgttcgaaatttgaatagtccctaaaacg

          attctcacatggtacaaaattgtgtattgctttctctgttgctgctaaaggagcgataca
       61 ---------+---------+---------+---------+---------+---------+
          taagagtgtaccatgttttaacacataacgaaagagacaacgacgatttcctcgctatgt

          ttgtacctatatttcactcctatattcatcacggcaaatttatactccaaacttgagaca
      121 ---------+---------+---------+---------+---------+---------+
          aacatggatataaagtgaggatataagtagtgccgtttaaatatgaggtttgaactctgt

          cgatttcgtagccgtagaccgcagccaaaatttttgtctattattgatcacttgtgttcat
      181 ---------+---------+---------+---------+---------+---------+
          gctaaagcatcggcatctggcgtcggttttaaaacagataataactagtgaacacaagta

          cttgcacttctgctgttttgtaattcttgtgctattaatttcaaatgtgttttctgtttg
      241 ---------+---------+---------+---------+---------+---------+
          gaacgtgaagacgacaaaacattaagaacacgataattaaagtttacacaaaagacaaac
a
b
c                                                   M  C  F  L  F  V

          tggtgctgctcaaaattgtaatcgatctacatattcttaaagtgaaggttacgattttta
      301 ---------+---------+---------+---------+---------+---------+
          accacgacgagttttaacattagctagatgtataagaatttcacttccaatgctaaaaat
a
b
c           V  L  L  K  I  V  I  D  L  H  I  L  K  V  K  V  T  I  F  I

          taatctatctgtacaaacaaaatgaagggtgaaagtatcttgttgtataggcatattatg
      361 ---------+---------+---------+---------+---------+---------+
          attagatagacatgtttgttttacttcccactttcatagaacaacatatccgtataatac
a                           M  K  G  E  S  I  L  L  Y  R  H  I  M
b
c           I  Y  L  Y  K  Q  N  E  G
```

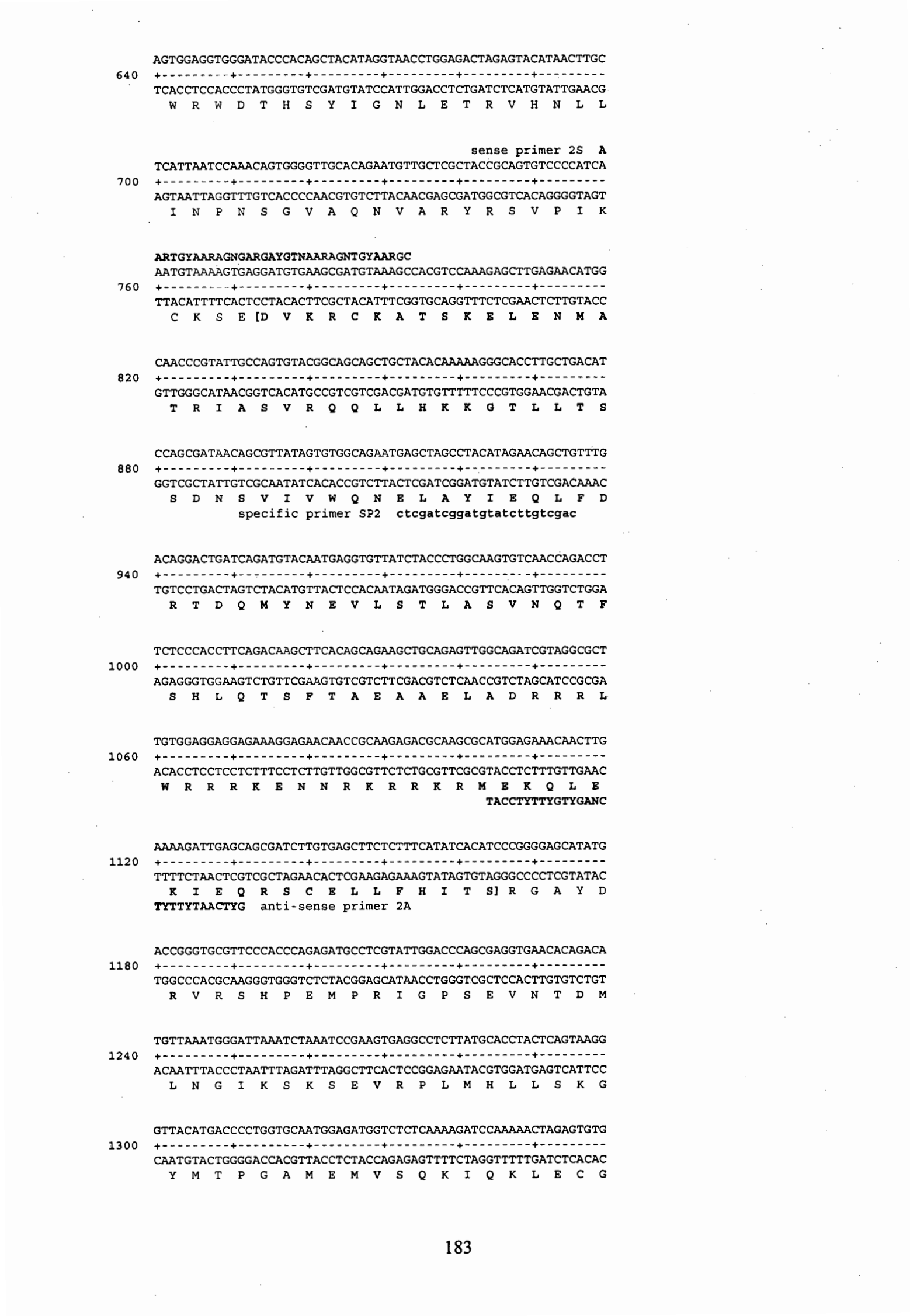
```
        tatgatcgtataaatgtaaaaaagcacacacagaaaaaaaatctcaaaatgttatttatt
    421 ---------+---------+---------+---------+---------+---------+
        atactagcatatttacatttttttcgtgtgtgtctttttttttagagttttacaataaataa
a       Y  D  R  I  N  V  K  K  H  T  Q  K  K  N  L  K  M  L  F  I
b
c


        tgtgtgtatggccatctaagatgtactatgtacatgttttctgtaccaatctggaaaggga
    481 ---------+---------+---------+---------+---------+---------+
        acacacataccggtagattctacatgatacatgtacaaaagacatggttagacctttcct
a       C  V  Y  G  H  L  R  C  T  M  Y  M  F  S  V  P  I  W  K  G
b
c                        M  Y  Y  V  H  V  F  C  T  N  L  E  R  T

        ccatggggaaagaattactacttcccaccctcccaccaagtttaacggcccaacataata
    541 ---------+---------+---------+---------+---------+---------+
        ggtacccctttcttaatgatgaagggtgggagggtggttcaaattgccgggttgtattat
a       P  W  G  K  N  Y  Y  F  P  P  S  H  Q  V
b
c          M  G  K  E  L  L  L  P  T  L  P  P  S  L  T  A  Q  H  N  I

        tttgaaaaaaaaactattaacc
    601 ---------+---------+-- 622
        aaacttttttttgataattgg


(d) analysis of translational reading frame for the minus strand

        ttttcctttacaacgcacaggatatttcacttcccgtatcttcatccccatccaaggatc
      1 ---------+---------+---------+---------+---------+---------+
        aaaaggaaatgttgcgtgtcctataaagtgaagggcatagaagtaggggtaggttcctag
                 *  L  A  C  S  I  E  S  G  T  D  E  D  G  D  L  S

        ctgttatatgaaatggatgacttgggtgtggtacatgtacgtcatttttaaacagaaat
     61 ---------+---------+---------+---------+---------+---------+
        gacaatatactttacctactgaacccacaccatgtacatgcagtaaaaaatttgtcttta
        G  T  I  H  F  P  H  S  P  H  P  V  H  V  D  N  K  L  C  F

        caacagatacagtacctccacagaatgtaaaacatg
    121 ---------+---------+---------+------ 156
        gttgtctatgtcatggaggtgtcttacattttgtac
        D  V  S  V  T  G  G  C  F  T  F  C
```

Appendix VII

Primers Designed for PCR Amplification of the SpGCF1 Homologue in *P.angulosus*

The double stranded cDNA sequence coding for recombinant SpGCF1 protein in *S.purpuratus* embryos (3) is aligned with the predicted amino acid sequence. The N-terminal region is a proline rich domain and the prolines are marked with asterisks (*). The sequence corresponding to the minimum DNA binding domain (3) spans from amino acids 223 - 353 (printed bold) and is contained in square brackets. Several degenerate primer pairs, viz 1S/1A (3) and 2S/2A (J. Hapgood, personal communication) have been designed corresponding to this cDNA (shown in bold above the sequence). The specific primer pair SP1/SP2 and primers SP3 and SP4, designed according to the DNA sequence for the *P.angulosus* homologue, are also aligned to the SpGCF1 cDNA.

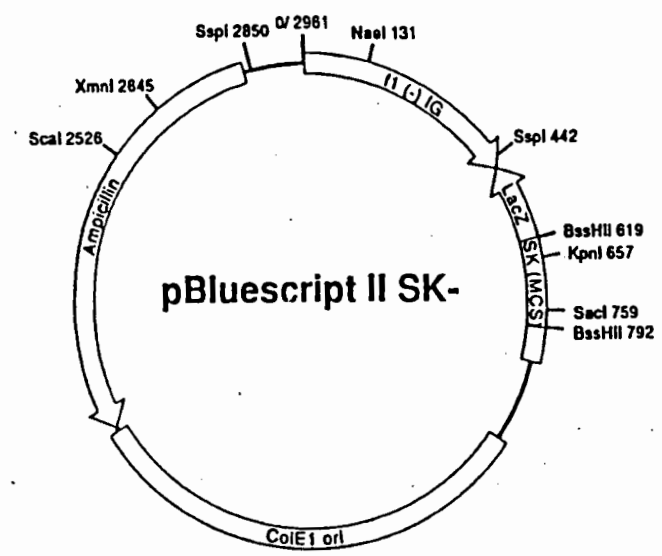
```
      AGTGAGGTATGTCCACTCTGCCCCAGCCCCTTTCCCACTGCCTGCTGAACCAGGTACACC
100   +---------+---------+---------+---------+---------+---------
      TCACTCCATACAGGTGAGACGGGGTCGGGGAAAGGGTGACGGACGACTTGGTCCATGTGG
            M  S  T  L  P* Q  P* L  S  H  C  L  L  N  Q  V  H  P*

      CCGCTCTCAACCTGCCCCAGACAGGGGTCATCACAGACATCAAGCCCATGATCAGTAATA
  160 +---------+---------+---------+---------+---------+---------
      GGCGAGAGTTGGACGGGGTCTGTCCCCAGTAGTGTCTGTAGTTCGGGTACTAGTCATTAT
       A  L  N  L  P* Q  T  G  V  I  T  D  I  K  P* M  I  S  N  K

      AACCTCCTACACAGGAGGTCAAACCAAACATCCTAGCAACTGGCTTGCCCTATCCTCCAC
  220 +---------+---------+---------+---------+---------+---------
      TTGGAGGATGTGTCCTCCAGTTTGGTTTGTAGGATCGTTGACCGAACGGGATAGGAGGTG
        P* P* T  Q  E  V  K  P* N  I  L  A  T  G  L  P* Y  P* P* L

      TCAACGTGCCTAGGCTACCCGTCATGCCCAATGTGTCTCTGCCTAGTGTCTCTATGCCGA
  280 +---------+---------+---------+---------+---------+---------
      AGTTGCACGGATCCGATGGGCAGTACGGGTTACACAGAGACGGATCACAGAGATACGGCT
        N  V  P* R  L  P* V  M  P* N  V  S  L  P* S  V  S  M  P* S

      GTGTGTCTATGCCCAATGTCTCCATGCCCAACGCATCCATGCCCAGTGTTTCGATGCCCA
  340 +---------+---------+---------+---------+---------+---------
      CACACAGATACGGGTTACAGAGGTACGGGTTGCGTAGGTACGGGTCACAAAGCTACGGGT
        V  S  M  P* N  V  S  M  P* N  A  S  M  P* S  V  S  M  P* N

                                              specific primer SP3  actga
      ATGTGTCCATGCCAAGTATTCCTCATCACAACTTACAGGGTAACTTAGGCCAATTACTCA
  400 +---------+---------+---------+---------+---------+---------
      TACACAGGTACGGTTCATAAGGAGTAGTGTTGAATGTCCCATTGAATCCGGTTAATGAGT
        V  S  M  P* S  I  P* H  H  N  L  Q  G  N  L  G  Q  L  L  N

      gagcaacagtaattctc                     specific primer SP1  catc
                  sense oligo 1S  CGGGATCCCATGCCNAAYGARTTYCTNCAYC
      ACAACAGTAATTCCCAAAAAATGTCCCAAATGAAAAAGTGCCCCAACGAGTTTTTACATC
  460 +---------+---------+---------+---------+---------+---------
      TGTTGTCATTAAGGGTTTTTTACAGGGTTTACTTTTTCACGGGGTTGCTCAAAAATGTAG
        N  S  N  S  Q  K  M  S  Q  M  K  K  C  P  N  E  F  L  H  Q

      agaatccacaaagtgagcg
      A
      AGAATCCACAAAGTGAGCGACAGCTTTTCTACAACGACGTAGCCATGCAACTGTATAACA
  520 +---------+---------+---------+---------+---------+---------
      TCTTAGGTGTTTCACTCGCTGTCGAAAAGATGTTGCTGCATCGGTACGTTGACATATTGT
        N  P  Q  S  E  R  Q  L  F  Y  N  D  V  A  M  Q  L  Y  N  S
                                                 TACGTYGANATRTTRT

      GTGACTTCAACAAGTTTGCTTCCAAGAAGGGATTTCATGGCTACCTGTTAGAGCAACAGA
  580 +---------+---------+---------+---------+---------+---------
      CACTGAAGTTGTTCAAACGAAGGTTCTTCCCTAAAGTACCGATGGACAATCTCGTTGTCT
        D  F  N  K  F  A  S  K  K  G  F  H  G  Y  L  L  E  Q  Q  K
      TRCTRAACCCATGGGG anti-sense primer 1A
```

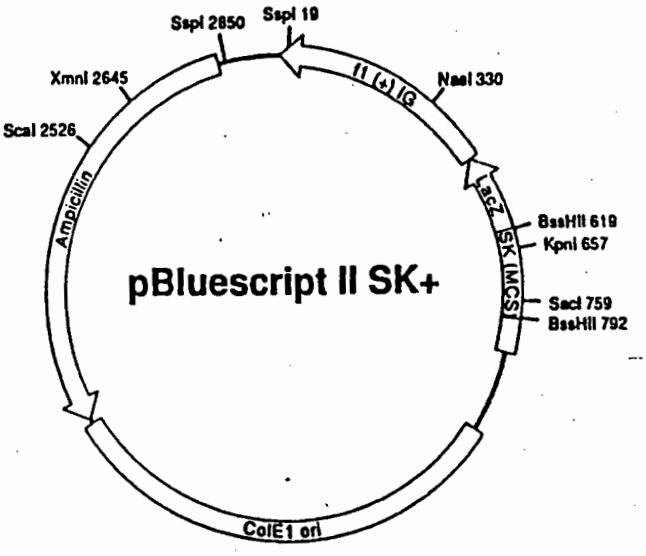

AGTGGAGGTGGGATACCCACAGCTACATAGGTAACCTGGAGACTAGAGTACATAACTTGC
640
TCACCTCCACCCTATGGGTGTCGATGTATCCATTGGACCTCTGATCTCATGTATTGAACG
W R W D T H S Y I G N L E T R V H N L L
sense primer 2S A
TCATTAATCCAAACAGTGGGGTTGCACAGAATGTTGCTCGCTACCGCAGTGTCCCCATCA
700
AGTAATTAGGTTTGTCACCCCAACGTGTCTTACAACGAGCGATGGCGTCACAGGGGTAGT
I N P N S G V A Q N V A R Y R S V P I K
ARTGYAARAGNGARGAYGTNAARAGNTGYAARGC
AATGTAAAAGTGAGGATGTGAAGCGATGTAAAGCCACGTCCAAAGAGCTTGAGAACATGG
760
TTACATTTTCACTCCTACACTTCGCTACATTTCGGTGCAGGTTTCTCGAACTCTTGTACC
C K S E [D V K R C K A T S K E L E N M A
CAACCCGTATTGCCAGTGTACGGCAGCAGCTGCTACACAAAAAGGGCACCTTGCTGACAT
820
GTTGGGCATAACGGTCACATGCCGTCGTCGACGATGTGTTTTTCCCGTGGAACGACTGTA
T R I A S V R Q Q L L H K K G T L L T S
CCAGCGATAACAGCGTTATAGTGTGGCAGAATGAGCTAGCCTACATAGAACAGCTGTTTG
880
GGTCGCTATTGTCGCAATATCACACCGTCTTACTCGATCGGATGTATCTTGTCGACAAAC
S D N S V I V W Q N E L A Y I E Q L F D
specific primer SP2 ctcgatcggatgtatcttgtcgac
ACAGGACTGATCAGATGTACAATGAGGTGTTATCTACCCTGGCAAGTGTCAACCAGACCT
940
TGTCCTGACTAGTCTACATGTTACTCCACAATAGATGGGACCGTTCACAGTTGGTCTGGA
R T D Q M Y N E V L S T L A S V N Q T F
TCTCCCACCTTCAGACAAGCTTCACAGCAGAAGCTGCAGAGTTGGCAGATCGTAGGCGCT
1000
AGAGGGTGGAAGTCTGTTCGAAGTGTCGTCTTCGACGTCTCAACCGTCTAGCATCCGCGA
S H L Q T S F T A E A A E L A D R R R L
TGTGGAGGAGGAGAAAGGAGAACAACCGCAAGAGACGCAAGCGCATGGAGAAACAACTTG
1060
ACACCTCCTCCTCTTTCCTCTTGTTGGCGTTCTCTGCGTTCGCGTACCTCTTTGTTGAAC
W R R R K E N N R K R R K R M E K Q L E
TACCTYTTYGTYGANC
AAAAGATTGAGCAGCGATCTTGTGAGCTTCTCTTTCATATCACATCCCGGGGAGCATATG
1120
TTTTCTAACTCGTCGCTAGAACACTCGAAGAGAAAGTATAGTGTAGGGCCCCTCGTATAC
K I E Q R S C E L L F H I T S] R G A Y D
TYTTYTAACTYG anti-sense primer 2A
ACCGGGTGCGTTCCCACCCAGAGATGCCTCGTATTGGACCCAGCGAGGTGAACACAGACA
1180
TGGCCCACGCAAGGGTGGGTCTCTACGGAGCATAACCTGGGTCGCTCCACTTGTGTCTGT
R V R S H P E M P R I G P S E V N T D M
TGTTAAATGGGATTAAATCTAAATCCGAAGTGAGGCCTCTTATGCACCTACTCAGTAAGG
1240
ACAATTTACCCTAATTTAGATTTAGGCTTCACTCCGGAGAATACGTGGATGAGTCATTCC
L N G I K S K S E V R P L M H L L S K G
GTTACATGACCCCTGGTGCAATGGAGATGGTCTCTCAAAAGATCCAAAAACTAGAGTGTG
1300
CAATGTACTGGGGACCACGTTACCTCTACCAGAGAGTTTTCTAGGTTTTTGATCTCACAC
Y M T P G A M E M V S Q K I Q K L E C G

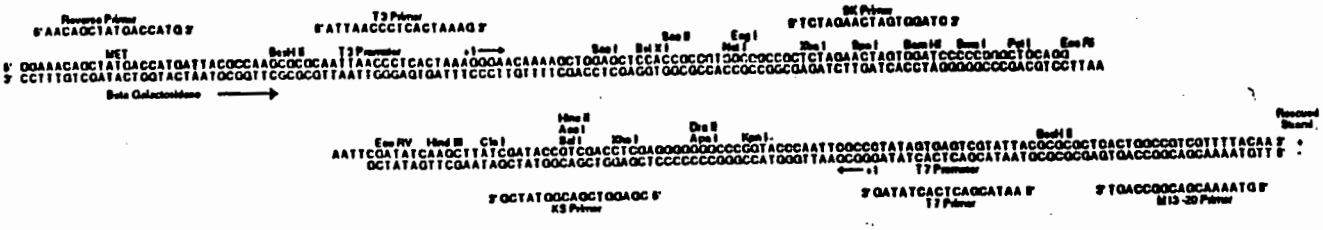
```
      GTATTAAGACTGAAGCGCACCAACAGGCAACCCAGGTTGGTATCAACTCCCTGTCGATCA
1360  +---------+---------+---------+---------+---------+---------
      CATAATTCTGACTTCGCGTGGTTGTCCGTTGGGTCCAACCATAGTTGAGGGACAGCTAGT
        I  K  T  E  A  H  Q  Q  A  T  Q  V  G  I  N  S  L  S  I  N
                         specific primer SP4  catagttgagagaccgttag

      ACAAAATTACAGCACCTGCTTCAGAGCTAAACTCCATACTGCCTCCTGTCACTGGAATTG
1420  +---------+---------+---------+---------+---------+---------
      TGTTTTAATGTCGTGGACGAAGTCTCGATTTGAGGTATGACGGAGGACAGTGACCTTAAC
        K  I  T  A  P  A  S  E  L  N  S  I  L  P  P  V  T  G  I  A

      CCTCATCAAATATGGTGTCATCTGTAAACTCAGCTGTGACACAACAATCAGTGCCCACAG
1480  +---------+---------+---------+---------+---------+---------
      GGAGTAGTTTATACCACAGTAGACATTTGAGTCGACACTGTGTTGTTAGTCACGGGTGTC
        S  S  N  M  V  S  S  V  N  S  A  V  T  Q  Q  S  V  P  T  V

      TAAATCTTAACACTCAATTAGCGAAGTAAAGACATTTTAACCAAGTCACAGCGACTTTGC
1540  +---------+---------+---------+---------+---------+---------
      ATTTAGAATTGTGAGTTAATCGCTTCATTTCTGTAAAATTGGTTCAGTGTCGCTGAAACG
        N  L  N  T  Q  L  A  K
```

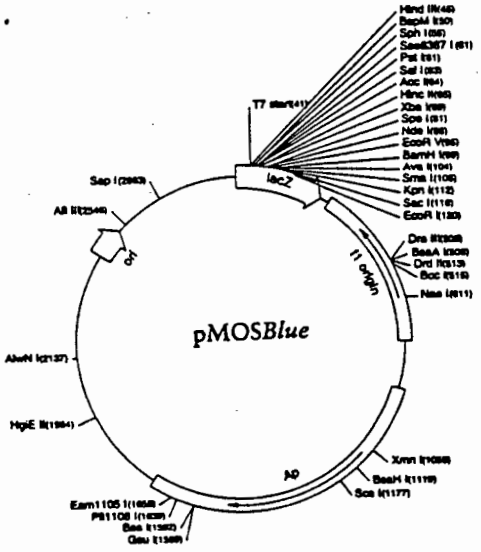
Appendix VIII

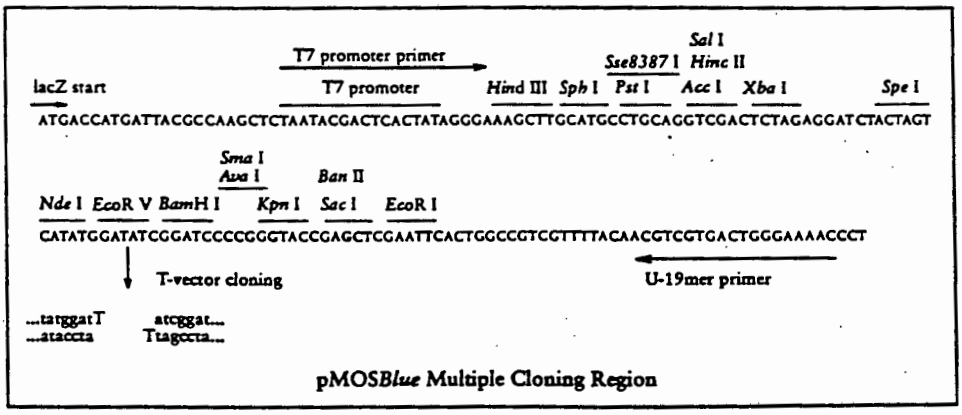
Plasmid Maps

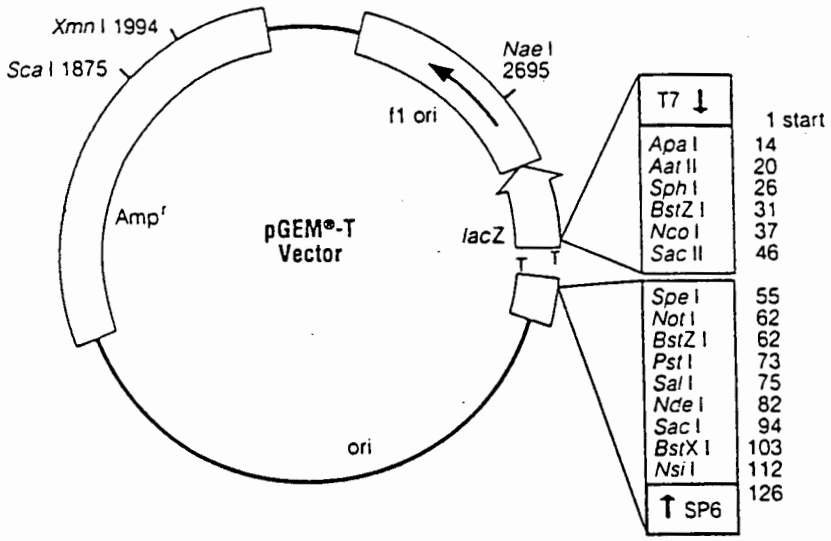


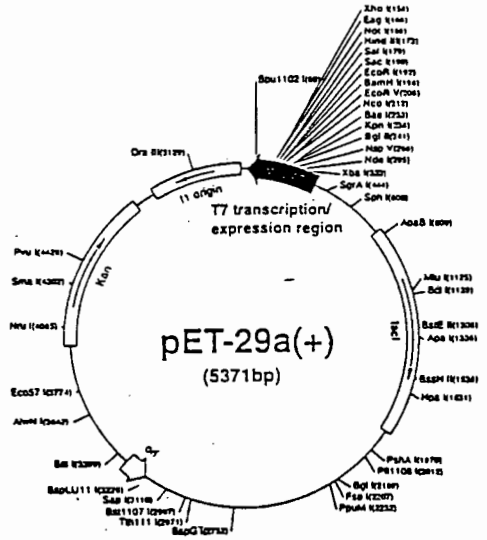


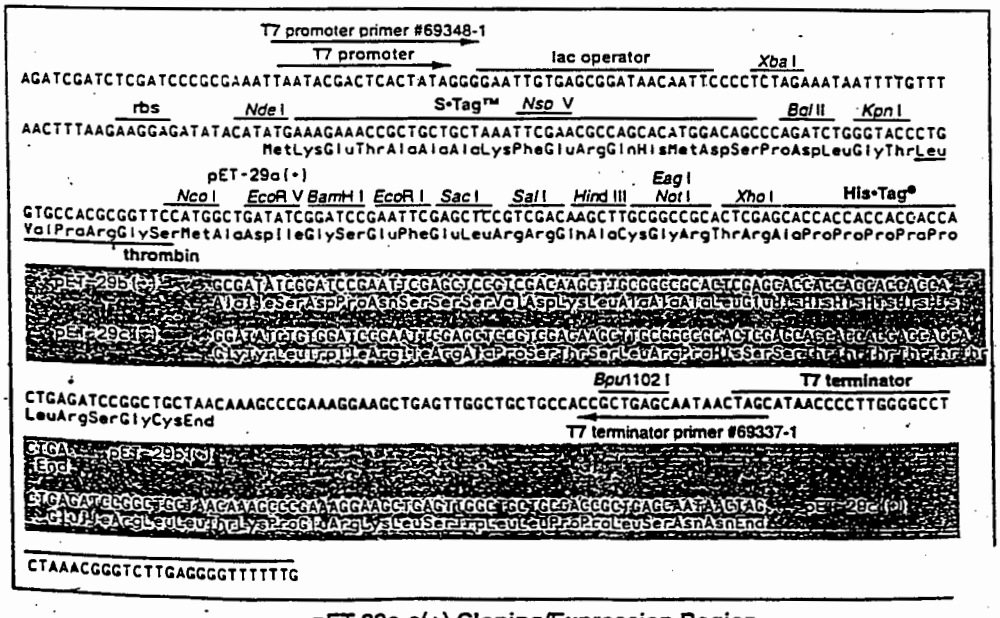




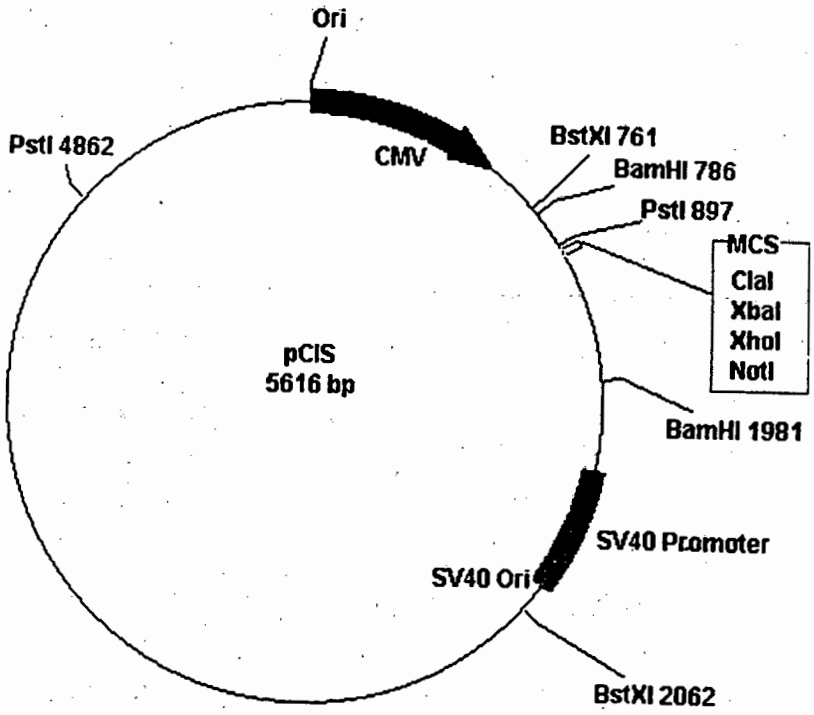


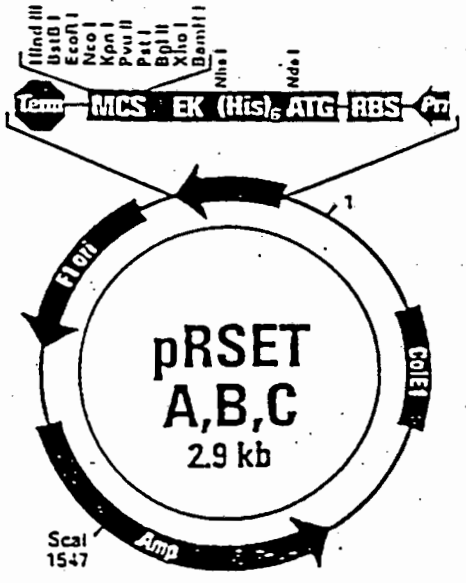






pET-29a-c(+) Cloning/Expression Region

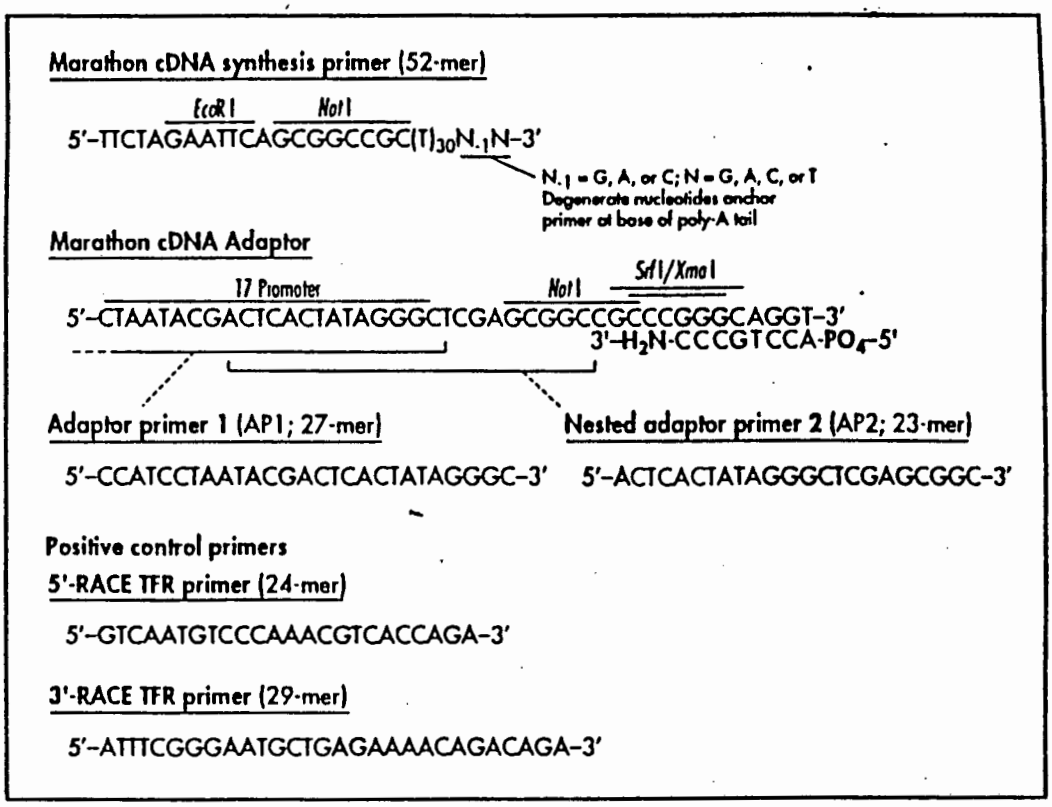




- Multiple cloning sites in three different reading frames for insertion of the gene of interest in frame with the Xpress™ N-terminal peptide.
- Simplified analysis with a monoclonal antibody specific for the Xpress™ peptide (page 62).
- F1 origin for rescue of ssDNA which binds the T7 primer (sense strand). For convenient generation of RNA transcripts, see the InvitroScript™ kit on page 9.

Appendix IX

Adaptor Primer (AP1)



Sequences of the Marathon cDNA Synthesis primer, the Marathon cDNA Adaptor, and the AP1 and AP2 primers, and the positive control TFR primers (all ClonTech).

Appendix X

Sequence Analysis of PCR-generated DNA Fragments

DNA sequences from *P.angulosus (P.ang)* corresponding to (a) the 106 bp fragment amplified using the degenerate primer pair 1S/1A, (b) the 372 bp fragment amplified using the degenerate primer pair 2S/2A and (c) the 421 bp fragment amplified using the specific primer pair SP1/SP2, were compared to the DNA sequence for SpGCF1 from *S.purpuratus* (*S.purp*) using a computer program (GCG). The numbers for the nucleotide sequence of the *P.angulosus* clone were allocated with respect to the numbering of the full length clone (see fig 3.15), whereas the numbers for the nucleotide sequence of the *S.purpuratus* clone correspond to the numbers allocated to the full length cDNA for SpGCF1 shown in Appendix VII and Zeller et al (1995) (3). The degenerate primers 1S/1A and 2S/2A used to amplify the respective fragments in (a) and (b) were not included in the sequence comparison. The *P.angulosus* sequences are printed in small letters, whereas the *S.purpuratus* sequences are printed in capital letters. The percentage similarity between the individual PCR fragments and the *S.purpuratus* clone are shown above each alignment respectively.

```
(a) Percentage Similarity: 93.877

S.purpuratus   520  AGAATCCACAAAGTGAGCGACAGCTTTTCTACAACGACGTAGCCATGCAA  569
                    |||||||||||||||||||| ||||| |||||||||||| ||||||||||||
P.angulosus    763  agaatccacaaagtgagcgtcagctattctacaacgatgtagccatgcaa  812


(b) Percentage Similarity: 91.912

S.purpuratus   796  CGTCCAAAGAGCTTGAGAACATGGCAACCCGTATTGCCAGTGTACGGCAG  845
                    |||| || ||||| ||||| ||||||||| ||||||||||||||| |||
P.angulosus    1040 cgtcaaaggagctggagaatatggcaaccggtattgccagtgtacgacag  1089

               846  CAGCTGCTACACAAAAAGGGCACCTTGCTGACATCCAGCGATAACAGCGT  895
                    |||||||| |||||||||||||||||||| |||| ||||||||| || |
               1090 cagctgctgcacaaaaagggcaccttgctaacat.cagcgataatagtg.  1137

               896  TATAGTGTGGCAGAATGAGCTAGCCTACATAGAACAGCTGTT  937
                    ||||||   |||||||||||||| ||||||||||||||| ||
               1138 tatagt...gcagaatgagctag.ctacatagaacagctatt  1175


(c) Percentage similarity: 91.640

S.purp  546  TTCTACAACGACGTAGCCATGCAACTGTATAACAGTGACTTCAACAAGTT  595
             |||||||| || |||||||||| |||||||||||||||||||||||||
P.ang   790  ttctacaatgatgtagccatgcagctgtataacagtgacttcaacaagtt  839

        596  TGCTTCCAAG.AAGGGATTTCATGGCTACCTGTTAGAGCAACAGAAGTGG  644
             |||||||||| |||| |||||||||||||||||||| |||||||||
        840  tgcttccaagaaaggaatttcatggctacctgttagagcagcagaagtgg  889

        645  AGGTGGGATACCCACAGCTACATAGGTAACCTGGAGACTAGAGTACATAA  694
             || |||||||||||||||||||||||||||||||||| |:||| |||||
        890  agatgggatacccacagctacataggtaacctggagaccanagtccataa  939

        695  CTTGCTCATTAATCCAAACAGTGGGGTTGCACAGAATGTTGCTCGCTACC  744
             |||||||| |||||||||||||||||| || || |||||||| || |
        940  cttgctcatcaatccaaacagtggggttgcccaaaacgttgctcgatatc  989

        745  GCAGTGTCCCCATCAAATGTAAAAGTGAGGATGTGAAGCGATGTAAAGCC  794
             |||| ||||| |||||||||||||| || ::|||||||||||||| |||||
        990  gcagcgtcccaatcaaatgtaaaagcgaanntgtgaagcgatgtgaagcc  1039

        795  ACGTCCAAAGAGCTTGAGAACATGGCAACCCGTATTGCCAGTGTACGGCA  844
             ||||| || ||||| ||:|| |||||||| ||||||||||||||||: ||
        1040 acgtcaaaggagctgganaatatggcaacgcgtattgccagtgtacnaca  1089

        845  GCAGCTGCTACA  856
             |||||||| ||
        1090 gcagctgctgca  1101
```

Appendix XI

Partial Sequence Analysis of the 5' and 3' RACE Products

The DNA sequences obtained for the ~ 900 bp 5' RACE product (a) and the ~ 1.5 kb 3' RACE product (b) amplified from *P.angulosus* (*P.ang*) cDNA have 74 % and 92 % identity to the SpGCF1 (3) sequence from *S.purpuratus* (*S.purp*) respectively. The homology comparisons were performed using the computer program GCG. The numbering of the nucleotide sequences for both SpGCF1 and the *P.angulosus* sequences correspond to the numbering allocated to the respective full length clones (see Appendix VII and see fig 3.15).

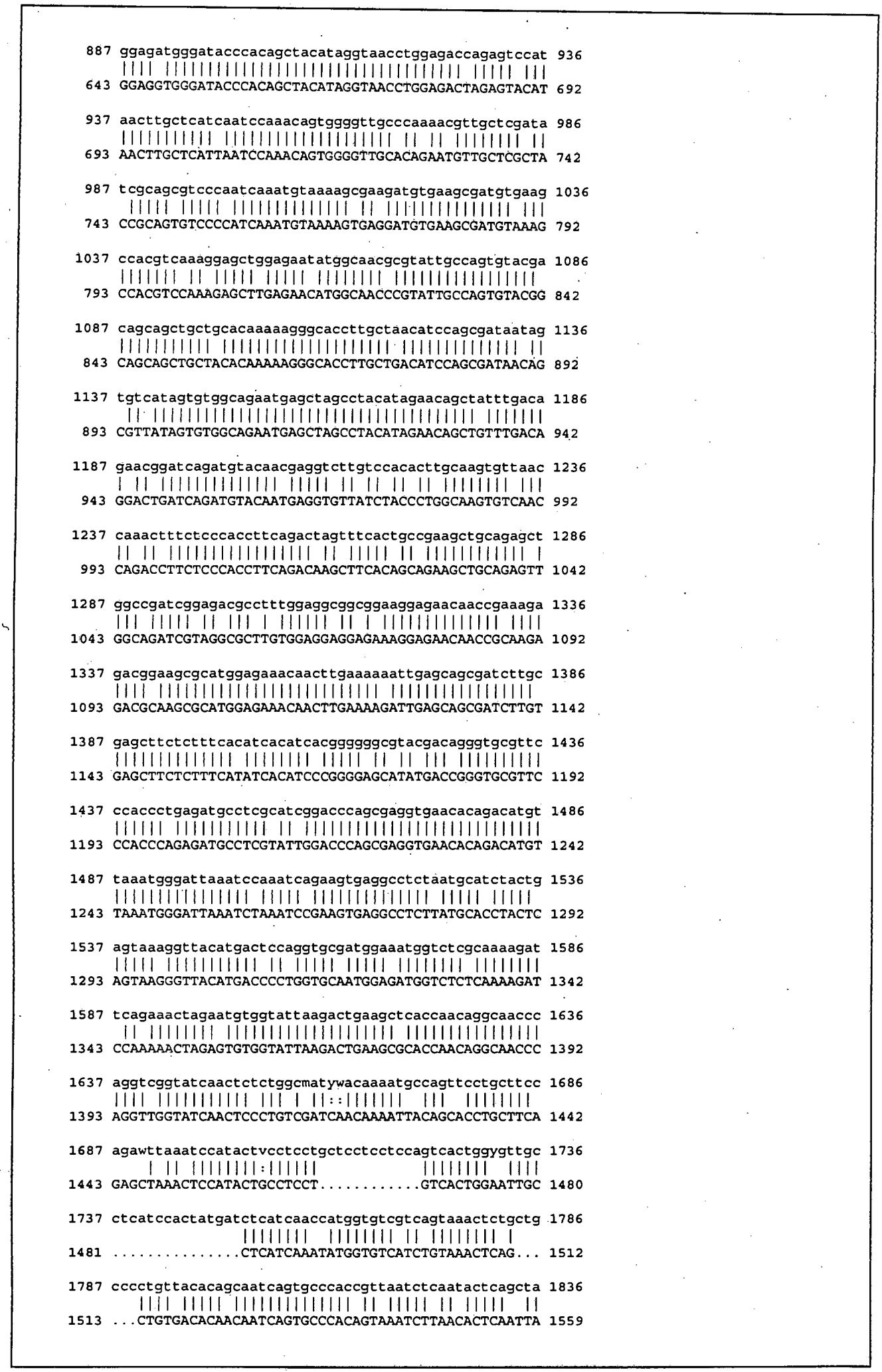
```
(a)     5' RACE fragment: Percentage Identity 73.626

S.purp     5  TTTGGGGGCATAATTTTGCTATTGATCAAGGATAGCGGGCCCGAATTTAC  54
              ||| || || |||| ||||| || |||||  ||| |       || ||
P.ang    190  ttttggagcttaatattgcttttcatcaatcataactactgaaaaattta  239

          55  TCATTTT........TTAGTGACTTGACGAGGATCCAACA.GAGGTGAGT  95
               ||||||         | || | | || |||  |||  || ||    ||
         240  ccattttgtgtgtcccttgttagtggaggagactcctccatgaaagaagg  289

          96  GAGGAGTGAGGTATGTCCACTCTGCCCCAGCCCCTTTCCCACTGCCTGCT  145
               |||||||||||| ||||| ||||||||||||||| ||||| ||||||||
         290  aaggagtgaggtttgtccgctctgccccagcccctgtcccattgcctgct  339

         146  GAACCAGGTACAC...CCCGCTCTCAACCTGCCC......CAGACAGGGG  186
               ||||  |||  ||    | ||  ||||||| |||      ||  ||||
         340  gaacgtggtgaacactgcagccatcaacctaccccatcaacaaccaggac  389

         187  TCATCACAGACATCAAGCCCATGATCAGTAATAAACCTCCTACACAGGAG  236
               |||||||||||||||| || ||||| ||||| ||||| ||| |   |||
         390  tcatcacagacatcaaaccaatgattagtaacaaacccccctcctactgag  439

         237  G.....TCAAACCAAACATCCTAGC...........AACTGGCTTGCCCT  270
               |     || |||| ||| || ||||           |||||||| ||
         440  ggaggttccaacccaacttcttagcctgcggctgcttgctggcttgaccc  489

         271  ATCCTCC 277
               | |||||
         490  accctcc 496

(b)     3' RACE fragment: Percentage Identity: 92.063

S.purp   561  GCCATGCAACTGTATAACAGTGACTTCAACAAGTTTGCTTCCAAGAAGGG  610
              |||||||| || ||||||||||||||||||||||||||||||||||||
P.ang    805  gccatgcagctctataacagtgacttcaacaagtttgcttccaagaagga  854

         611  ATTTCATGGCTACCTGTTAGAGCAACAGAAGTGGAGGTGGGATACCCACA  660
              |||||||||||||||||||||||| |||||||||| ||||||||||||
         855  atttcatggctacctgttagagcagcagaagtggagatgggatacccaca  904

          661  GCTACATAGGTAACCTGGAGACTAGAGTACATAACTTGCTCATTAATCCA  710
               |||||||||||||||||||| ||||| |||||||||||||| ||||||
          905  gctacataggtaacctggagaccagagtccataacttgctcatcaatcca  954

          711  AACAGTGGGGTTGCACAGAATGTTGCTCGCTACCGCAGTGTCCCCATCAA  760
               |||||||||||||| || || |||||||| || ||||| ||||| |||||
          955  aacagtggggttgcccaaaacgttgctcgatatcgcagcgtcccaatcaa 1004

          761  ATGTAAAAGTGAGGATGTGAAGCGATGTAAAGCCACGTCCAAAGAGCTTG  810
               ||||||||| || ||||||||||||||| |||||||||| || ||||| |
         1005  atgtaaaagcgaagatgtgaagcgatgtgaagccacgtcaaaggagctgg 1054

          811  AGAACATGGCAACCCGTATTGCCAGTGTACGGCAGCAGCTGCTACACAAA  860
               |||| |||||||| |||||||||||||||| |||||||||| ||||||
         1055  agaatatggcaacgcgtattgccagtgtacgacagcagctgctgcacaaa 1104

          861  AAGGGCACCTTGCTGACATCCAGCGATAACAGCGTTATAGTGTGGCAGAA  910
               |||||||||||||| |||||||||||||| || || ||||||||||||||
         1105  aagggcaccttgctaacatccagcgataatagtgtcatagtgtggcagaa 1154

          911  TGAGCTAGCCTACATAGAACAGCTGTTT  938
               ||||||||||||||||||||||||| |||
         1155  tgagctagcctacatagaacagctattt 1182
```

Appendix XII

DNA Homology Comparison Between the PCR-generated *P.angulosus* Clone and the cDNA Encoding SpGCF1

The consensus DNA sequence representing the full length SpGCF1 homologue present in *P.angulosus* was aligned and compared to the SpGCF1 DNA sequence (3) using the computer programme GCG. The *P.angulosus* sequence (*P.ang*) is printed in small letters and the *S.purpuratus* (*S.purp*) sequence is shown in capital letters. The numbers for the nucleotide sequence of the *S.purpuratus* clone correspond to the numbers allocated to the full length cDNA as shown in Appendix IX and Zeller et al (1995) (3), and the numbers allocated to the consensus DNA sequence for the *P.angulosus* homologue correspond to the nucleotide position as determined by sequence analysis.

```
Percentage Similarity: 83.763

P.ang  190 atttggagcttaatattgcttttcatcaatcataacgactgaaaaattta 239
S.purp   5 TTTGGGGGCATAATTTTGCTATTGATCAAGGATAGCGGGCCCGAATTTAC  54

       240 ccattttgtgtgtaccttgtgagttgaggagactcctccatagaagaaga 289
        55 TCATTTT........TTAGTGACTTGACGAGGATCCAACAGAGGTGAGTG  96

       290 aggagtgaggtatgtccactctgccccagcccctgtcccattgcctgctg 339
        97 AGGAGTGAGGTATGTCCACTCTGCCCCAGCCCCTTTCCCACTGCCTGCTG 146

       340 aaccaggtgaacactgcagccatcaacctaccacatcaacaacctggact 389
       147 AACCAGGTACAC...CCCGCTCTCAACCTGCC......CCAGACAGGGGT 187

       390 catcacagacatcaaaccaatgattagtaacaaacccctcctactcagg 439
       188 CATCACAGACATCAAGCCCATGATCAGTAATAAA...CCTCCTACACAGG 234

       440 aggtcaaaccaaacatcttagctgcggctgctgctggcttgacctaccct 489
       235 AGGTCAAACCAAACATCCTA.........GCAACTGGCTTGCCCTATCCT 275

       490 ccactcaacgtgcctagcctacctgcaatgcccaacgtgtcgatgcctaa 539
       276 CCACTCAACGTGCCTAGGCTACCCGTCATGCCCAATGTGTCTCTGCCTAG 325

       540 tgtgtcattgcccaacgtgtcaatgcctaatgtgtctatgcccaatgtgt 589
       326 T..............GTCTCTATGCCGAGTGTGTCTATGCCCAATGTCT 360

       590 ctatgccaaccagcgtttcaatgccgagtgtgtccatgcccagcgtttct 639
       361 CCATGCCCAAC..................GCATCCATGCCCAGTGTTTCG 392

       640 atgccgagtgcgtccatgccaagtgttactcttcacaaccaacagggaaa 689
       393 ATGCCCAATGTGTCCATGCCAAGTATTCCTCATCACAACTTACAGGGTAA 442

       690 caatagcca...actgagcaacagtaattctcaacggctgtcccaaatga 736
       443 CTTAGGCCAATTACTCAACAACAGTAATTCCCAAAAAATGTCCCAAATGA 492

       737 agaaatgccccaatgagtttctccatcagaatccccaaagtgagcgtcag 786
       493 AAAAGTGCCCCAACGAGTTTTTACATCAGAATCCACAAAGTGAGCGACAG 542

       787 ctattctacaatgatgtagccatgcagctgtataacagtgacttcaacaa 836
       543 CTTTTCTACAACGACGTAGCCATGCAACTGTATAACAGTGACTTCAACAA 592

       837 gtttgcttccaagaaggaatttcatggctatctgttagagcagcagaagt 886
       593 GTTTGCTTCCAAGAAGGGATTTCATGGCTACCTGTTAGAGCAACAGAAGT 642
```


887 ggagatgggatacccacagctacataggtaacctggagaccagagtccat 936
643 GGAGGTGGGATACCCACAGCTACATAGGTAACCTGGAGACTAGAGTACAT 692
937 aacttgctcatcaatccaaacagtggggttgcccaaaacgttgctcgata 986
693 AACTTGCTCATTAATCCAAACAGTGGGGTTGCACAGAATGTTGCTCGCTA 742
987 tcgcagcgtcccaatcaaatgtaaaagcgaagatgtgaagcgatgtgaag 1036
743 CCGCAGTGTCCCCATCAAATGTAAAAGTGAGGATGTGAAGCGATGTAAAG 792
1037 ccacgtcaaaggagctggagaatatggcaacgcgtattgccagtgtacga 1086
793 CCACGTCCAAAGAGCTTGAGAACATGGCAACCCGTATTGCCAGTGTACGG 842
1087 cagcagctgctgcacaaaaagggcaccttgctaacatccagcgataatag 1136
843 CAGCAGCTGCTACACAAAAAGGGCACCTTGCTGACATCCAGCGATAACAG 892
1137 tgtcatagtgtggcagaatgagctagcctacatagaacagctatttgaca 1186
893 CGTTATAGTGTGGCAGAATGAGCTAGCCTACATAGAACAGCTGTTTGACA 942
1187 gaacggatcagatgtacaacgaggtcttgtccacacttgcaagtgttaac 1236
943 GGACTGATCAGATGTACAATGAGGTGTTATCTACCCTGGCAAGTGTCAAC 992
1237 caaactttctcccaccttcagactagtttcactgccgaagctgcagagct 1286
993 CAGACCTTCTCCCACCTTCAGACAAGCTTCACAGCAGAAGCTGCAGAGTT 1042
1287 ggccgatcggagacgcctttggaggcggcggaaggagaacaaccgaaaga 1336
1043 GGCAGATCGTAGGCGCTTGTGGAGGAGGAGAAAGGAGAACAACCGCAAGA 1092
1337 gacggaagcgcatggagaaacaacttgaaaaaattgagcagcgatcttgc 1386
1093 GACGCAAGCGCATGGAGAAACAACTTGAAAAGATTGAGCAGCGATCTTGT 1142
1387 gagcttctctttcacatcacatcacgggggcgtacgacagggtgcgttc 1436
1143 GAGCTTCTCTTTCATATCACATCCCGGGGAGCATATGACCGGGTGCGTTC 1192
1437 ccaccctgagatgcctcgcatcggacccagcgaggtgaacacagacatgt 1486
1193 CCACCCAGAGATGCCTCGTATTGGACCCAGCGAGGTGAACACAGACATGT 1242
1487 taaatgggattaaatccaaatcagaagtgaggcctctaatgcatctactg 1536
1243 TAAATGGGATTAAATCTAAATCCGAAGTGAGGCCTCTTATGCACCTACTC 1292
1537 agtaaaggttacatgactccaggtgcgatggaaatggtctcgcaaaagat 1586
1293 AGTAAGGGTTACATGACCCCTGGTGCAATGGAGATGGTCTCTCAAAAGAT 1342
1587 tcagaaactagaatgtggtattaagactgaagctcaccaacaggcaaccc 1636
1343 CCAAAAACTAGAGTGTGGTATTAAGACTGAAGCGCACCAACAGGCAACCC 1392
1637 aggtcggtatcaactctctggcmatywacaaaatgccagttcctgcttcc 1686
1393 AGGTTGGTATCAACTCCCTGTCGATCAACAAAATTACAGCACCTGCTTCA 1442
1687 agawttaaatccatactvcctcctgctcctcctccagtcactggygttgc 1736
1443 GAGCTAAACTCCATACTGCCTCCT.....GTCACTGGAATTGC 1480
1737 ctcatccactatgatctcatcaaccatggtgtcgtcagtaaactctgctg 1786
1481 .....CTCATCAAATATGGTGTCATCTGTAAACTCAG... 1512
1787 cccctgttacacagcaatcagtgcccaccgttaatctcaatactcagcta 1836
1513 ...CTGTGACACAACAATCAGTGCCCACAGTAAATCTTAACACTCAATTA 1559

```
1837 gcaaagtaacaccaaacagaccatgt......aacctttccatacttctg 1880
     || ||||||   ||     |||| ||       || || ||| | | | |
1560 GCGAAGTAAAGACATTTTAACCAAGTCACAGCGACTTTGCCACATTGCCG 1609

1881 agtg.ttgatagt............tatactctatactgtaatttcaag 1916
     |||| |||| | |             | ||||| | ||||   ||| ||
1610 AGTGTTTGACATTGAGTAGGCTGTACTCTACTCCACACTG...TTTTAAC 1656

1917 caacattttatgatgtctaatcatgctccaatgtgagaaaagtatacatt 1966
     ||||||| ||| |||| | | ||| |||  |  || | ||| | ||||||
1657 CAACATTGTATTATGTATGAGCATACTCTTACATG.GCAAAATGTACATT 1705

1967 tattgta.taaacaggaatgtagcaaattttaaaatgatttagctactaa 2015
     |||| || ||  ||||||||| |||    || |   ||| || | |  |
1706 TATTATATTATGCAGGAATGTTGCATCAGTTTA...GATCTAACCAAAAC 1752

2016 attgtagaattacttgttgtttggataaacatgtagcttgtactggatgt 2065
     | | |   |||||||||| |||              || |||  | |||
1753 AGTTTTAGATTACTTGTTCTTT.............TTTTTACCAGGTGT 1788

2066 aaatgtaaattttacccagtacaaat 2091
     | ||||||||||| |  ||  || |
1789 ACATGTAAATTTTAACTGGTGAAATT 1814
```

Appendix XIII

Comparison Between suGF1 and SpGCF1 Primary Sequence

The predicted amino acid sequences corresponding to the full length suGF1 clone from *P.angulosus* (*P.ang*) and the SpGCF1 clone from *S.purpuratus* (*S.purp*) (3) were aligned using GCG. They show a homology of 94 %. Identical amino acid residues in both sequences are connected with a solid line, those which differ between the two proteins are marked in bold type, and are connected by a colon (conservation of charge), or by a single dot (charge is not retained). Nonidentical amino acids are shown in bold type. The numbering of the SpGCF1 amino acids corresponds to the numbering allocated in Appendix VII and Zeller et al (1995) (3).

```
Percentage Similarity: 94.227

P.ang   1 MSTLPQPLSHCLLNQVNTAAINLPHQQPGLITDIKPMISNKPPPTQEVKP 50
          |||||||||||||||||: :|:||| |.|:|||||||||||| ||||||||
S.purp  1 MSTLPQPLSHCLLNQVH.PALNLP..QTGVITDIKPMISNK.PPTQEVKP 46

         51 NILAAAAAGLTYPPLNVPSLPAMPNVSMPNVSLPNVSMPNVSMPNVSMPT 100
            |||   |.||.|||||||.||.|||||:|.||:|.|||||||||||
         47 NIL...ATGLPYPPLNVPRLPVMPNVSLPSVSMPSVSMPNVSMPN..... 88

        101 SVSMPSVSMPSVSMPSASMPSVTLHNQQGNNSQ.LSNSNSQRLSQMKKCP 149
                  .||||||||..||||:. || ||| :| |.|||||::|||||||
         89 ......ASMPSVSMPNVSMPSIPHHNLQGNLGQLLNNSNSQKMSQMKKCP 132

        150 NEFLHQNPQSERQLFYNDVAMQLYNSDFNKFASKKEFHGYLLEQQKWRWD 199
            |||||||||||||||||||||||||||||||||:|||||||||||||||
        133 NEFLHQNPQSERQLFYNDVAMQLYNSDFNKFASKKGFHGYLLEQQKWRWD 182

        200 THSYIGNLETRVHNLLINPNSGVAQNVARYRSVPIKCKSEDVKRCEATSK 249
            ||||||||||||||||||||||||||||||||||||||||||||.||||
        183 THSYIGNLETRVHNLLINPNSGVAQNVARYRSVPIKCKSEDVKRCKATSK 232

        250 ELENMATRIASVRQQLLHKKGTLLTSSDNSVIVWQNELAYIEQLFDRTDQ 299
            |||||||||||||||||||||||||||||||||||||||||||||||||
        233 ELENMATRIASVRQQLLHKKGTLLTSSDNSVIVWQNELAYIEQLFDRTDQ 282

        300 MYNEVLSTLASVNQTFSHLQTSFTAEAAELADRRRLWRRRKENNRKRRKR 349
            ||||||||||||||||||||||||||||||||||||||||||||||||||
        283 MYNEVLSTLASVNQTFSHLQTSFTAEAAELADRRRLWRRRKENNRKRRKR 332

        350 MEKQLEKIEQRSCELLFHITSRGAYDRVRSHPEMPRIGPSEVNTDMLNGI 399
            ||||||||||||||||||||||||||||||||||||||||||||||||||
        333 MEKQLEKIEQRSCELLFHITSRGAYDRVRSHPEMPRIGPSEVNTDMLNGI 38

        400 KSKSEVRPLMHLLSKGYMTPGAMEMVSQKIQKLECGIKTEAHQQATQVGI 449
            ||||||||||||||||||||||||||||||||||||||||||||||||||
        383 KSKSEVRPLMHLLSKGYMTPGAMEMVSQKIQKLECGIKTEAHQQATQVGI 432

        450 NSLAINKMPVPASRNKSILPPAPPPVTGVASSTMISSTMVSSVNSAAPVT 499
            |||.|||:..|||  .|||    |||||:|     ||.||||||||  ||
        433 NSLSINKITAPASELNSIL....PPVTGIA.....SSNMVSSVNSA..VT 471

        500 QQSVPTVNLNTQLAK 514
            ||||||||||||||
        472 QQSVPTVNLNTQLAK 486
```

Appendix XIV

Recombinant Protein Expression in pET

(a) Description of the pET Expression System

The pET system provides translation vectors which can generally be used for expression of eukaryotic target genes (232). These vectors are available in three different reading frames with respect to the *BamH1* site in the polylinker. In addition to a versatile multiple cloning site, pET vectors have several other features such as peptide "tags" to which the target protein can be fused. These are useful for detection and purification. The pET-29b(+) plasmid contains sequences coding for the His-Tag (an oligohistidine domain) which, when fused to the C-terminus of the target protein, can be used for convenient purification of the recombinant protein. The His-Tag can be released from the target protein by cleavage at the thrombin site. Another "tag" available on the pET-29b(+) vector is the S-Tag peptide (this forms an N-terminal fusion with the target protein) and an antibody generated against this region is available, allowing quantitation, detection and affinity purification of the expressed protein. Stop signals, as well as a downstream T7 transcription terminator, are available in all three reading frames on the plasmid. The plasmid codes for the kanamycin resistance gene, which is read in the opposite direction to the target gene, preventing read-through transcription from the T7 promoter. This turn prevents the accumulation of the kanamycin gene product (pET System Manual). Finally the presence of an f1 origin of replication on the plasmid allows for the production of ss plasmid DNA.

Expression constructs are first established in a host which does not contain T7 RNA polymerase. This ensures plasmid stability by preventing potentially toxic genes from being expressed. Several cloning hosts are suitable, three of which (viz HB101, JM109 and DH5α) were used to establish the expression constructs for Clones 2, 6, 11 and 16 (see section 4.2.1). The absence of a T7 RNA polymerase source in the cloning host reduces the background target protein synthesis because the host enzymes do not recognise (or initiate from) the T7 promoter (232). Target genes remain transcriptionally inactive in the uninduced state until they are transferred to an expression host containing T7 RNA polymerase. Generally these are lysogens of the DE3 bacteriophage, which is a λ derivative, and as a result these host cells have a DNA fragment containing the *lac*I gene, the *lac*UV5 promoter and the gene for T7 RNA polymerase (232). The T7 RNA polymerase gene is under control of the inducible *lac*UV5 promoter which can be induced by addition of IPTG to the growth medium. Induction results in the expression of T7 RNA polymerase, which in turn transcribes the target DNA in the plasmid. However, even in the absence of IPTG the *lac*UV5 promoter may allow some expression of T7 RNA polymerase, and, when dealing with toxic genes, this may be sufficient to prevent the establishment of plasmids in the expression host. Therefore some pET plasmids (eg the pET-29 series) have been developed with the T7*lac* promoter (233), in which case the plasmid contains a *lac* operator sequence downstream of the T7 promoter. The natural promoter and coding sequence for the *lac* repressor (*lac*I) is oriented such that the T7*lac* promoter and the *lac*I promoters diverge, hence the *lac* repressor acts both at the *lac*UV5 promoter in the host chromosome to repress transcription of the T7 RNA polymerase gene, and at the T7*lac* promoter in the vector to block transcription of the target gene by any T7 RNA polymerase that is made (233). Some of the most common expression hosts used to express target proteins from pET plasmids are the two bacterial strains BL21(DE3) and BL21(DE3)pLysS. Both these strains were used in the expression studies (see section 4.2.1). The BL21(DE3) strain lacks the *lon* protease and the *ompT* outer membrane protease which helps reduce the degradation of proteins during purification (234), and thereby improves the stability of proteins. The pLysS strain of BL21(DE3) contains a plasmid which produces a small amount of T7 lysozyme. This enzyme has the bifunctional properties of cutting a specific bond in the peptidoglycan layer of the *E.coli* cell wall, as well as acting as a natural inhibitor of T7 RNA polymerase, thereby providing extra stability to target genes (235). The low amount of lysozyme produced has very little effect on target gene expression once the T7 RNA polymerase is induced, since more polymerase is produced than can be inhibited by the lysozyme. Lysozyme is unable to pass through the inner cell membrane; relatively high levels of the enzyme can be tolerated by the cells and therefore rapid of lysis of cells can be induced by combining the lysozyme with other treatments that would normally not cause cell lysis, eg freeze-thaw, chloroform, and mild detergent (eg 0.1 % Triton X-100) (pET System Manual).

(b) Optimisation of Recombinant Protein Expression Using the pET-29b(+) Vector

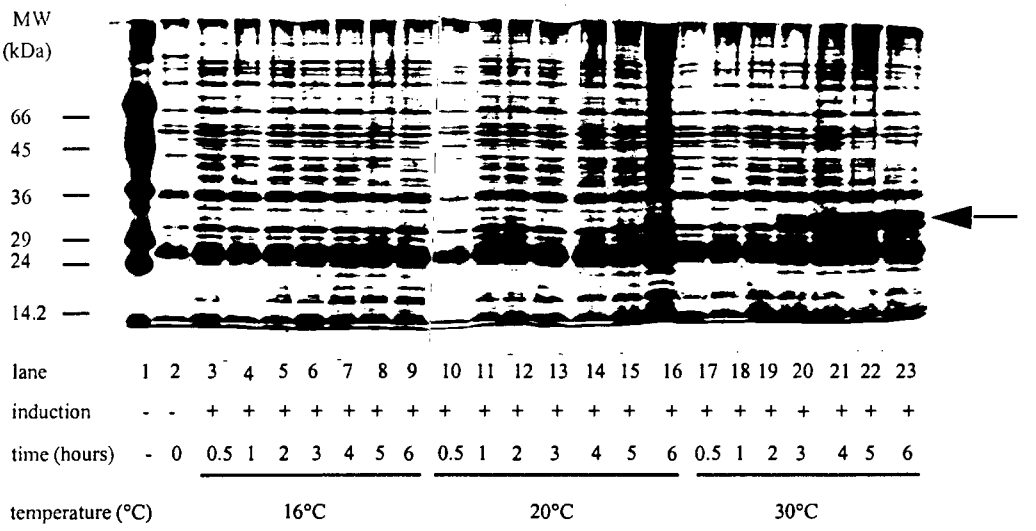
The four clones (2, 6, 11 and 16) generated by the DNA ligand screening technique were subcloned from pBluescript SK into pET-29b(+) (Novagen) using restriction site combinations *SacI / BamHI* (Clone 2) and *SalI / XbaI* (Clones 6, 11 and 16). These combinations allowed for directional cloning of the inserts. The inserts were also released from Bluescript with *EcoRI* and subcloned into pET-29b(+) using the same site. Correct orientation of the inserts was ensured by restriction mapping. The "b" reading frame was chosen in the pET-29 vector series in order to retain the same reading frame as originally provided by the Bluescript plasmid. The expression constructs were established in several different *E.coli* host strains, viz JM109, HB101 and DH5α, and expression of the constructs was induced in the host strains BL21DE3 and pLysS using several colonies originating from each original host, and subsequently placed in both expression hosts. A plasmid referred to as the "induction control" which has matching elements to the pET-29 vector, and codes for the β-galactosidase protein (pET System Manual) was provided in the DE3 lysogen host and was always used as a positive control to test induction by IPTG. Apart from optimising both the initial and expression hosts for Clones 2, 6, 11 and 16, other conditions used to optimise expression included varying the culture volumes (0.2 ml to 50 ml), the temperature of induction (16°C, 20°C, 30°C and 37°C), and length of induction (0.5 to 6 hours). The percentage SDS gel used to analyse the total cell protein was varied in order to optimise the separation of proteins and the gels were stained with both Coomassie and silver (see sections 2.26). Optimisation of expression resulted in the expression of a single clone in the pET plasmid. Clone 11 was expressed successfully as a ~ 25 kDa protein in the expression host BL21DE3 (for details see fig 4.1). Expression of Clone 11 was further optimised by performing a time course of induction at various temperatures. Log phase cells were induced at 16°C, 20°C, 30°C (fig (ii) (a)) and 37°C (fig (ii) (b)) and aliquots of induced cells were removed at hourly intervals for 6 hours. Uninduced and induced cells were analysed by SDS-PAGE. Recombinant protein expression from Clone 11 is maximally induced at 37°C for 3 - 4 hours (fig (ii) (b), lanes 7 - 8) and at 30°C for a period of 5 - 6 hours (fig (ii) (a), lanes 24 - 25). Induction at 16°C (fig (ii) (a) lanes 3 - 9) can only be detected very faintly after six hours, and at 20°C (lanes 10 - 16) it is detected at low levels from about 5 hours onwards. At both 30°C and 37°C the induction occurs strongly after three hours and continues with increasing time. Protein molecular weight standards are shown in lane 1 of fig (ii) (a) and (b), and the uninduced and induced proteins resulting from the induction control plasmid are shown in lanes 2 and 3 respectively (fig (ii) (b)).

(c) Soluble Forms of Recombinant Protein Expression

Several methods were applied in attempts to isolate recombinant protein from Clone 11 in a soluble form in order to confirm its DNA-binding ability. The first method (Stratagene Protocols) involved harvesting the cells (induction with IPTG at mid log phase, growth at 37°C for 3 - 4 hours), resuspension in lysis buffer (section 2.14.3.1) which was supplemented with lysozyme and Triton X-100 to lyse the cells, and subsequent separation of the cellular debris from the soluble proteins in the supernatant by centrifugation. Soluble protein extracts were analysed by SDS-PAGE and silver stained, as shown in fig (iii). Lane 1 represents protein molecular weight markers, lanes 2 and 3 are uninduced and induced cell extracts from the control plasmid (β-gal is shown by the asterisk), lanes 4 and 6 are uninduced extracts from Clone 11, whereas lanes 5 and 7 represent soluble protein extracts derived from induced cells of Clone 11 (recombinant protein is marked by the arrow). It appears that both the β-gal protein (lane 3) and the protein from Clone 11 (lanes 5 and 7) are at least partially soluble. These extracts were titrated in gel shift experiments with the E/H fragment (see section 2.21) however the induced protein showed no DNA-binding ability (data not shown). This could possibly be attributed to either too much detergent present in the extract, lack of Mg^{2+} and Zn^{2+} ions during isolation, or incorrect folding of the recombinant protein. These potential problems were addressed by dialysing the bacterial extracts into dialysis buffer (see section 2.19) containing 4 mM $MgCl_2$ and 2 mM $ZnCl_2$, however this did not alter the DNA binding ability of the extracts. Several attempts to improve the isolation of soluble recombinant protein (by varying the temperature, length of induction, changing the expression host) were not successful, and the induction and isolation methods of soluble protein appeared to be unreliable, since the induced proteins could not reproducibly be detected in the supernatant.

The purification of soluble recombinant protein was attempted by an alternative method, which exploits the fusion of recombinant proteins to the His-Tag on the pET plasmid, followed by Nickel column chromatography of the protein extracts (pET System

(a)



(b)

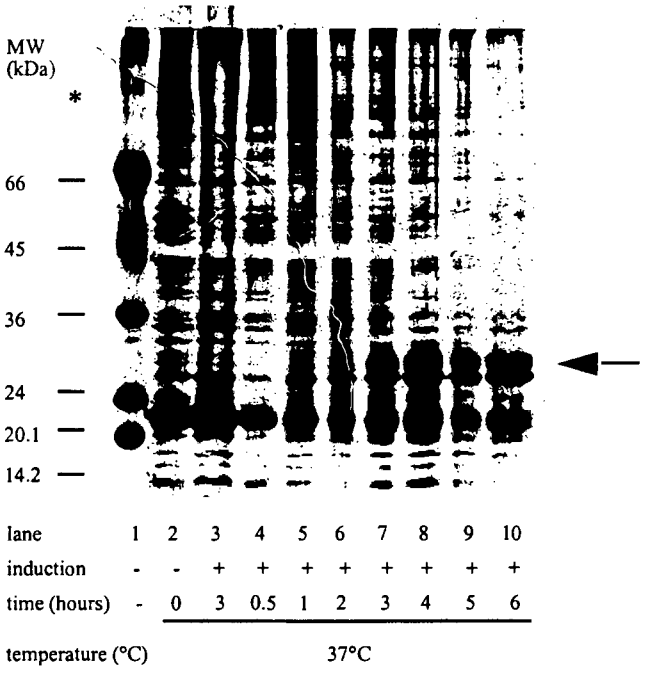


Fig (ii) Optimisation of Recombinant Protein Expression from Clone 11 in the pET-29b(+) Expression Vector

Recombinant expression from Clone 11 in pET-29b(+) in BL21DE3 cells was optimised by an induction time course over a range of temperatures (16°C, 20°C and 30°C, (a) and 37°C, (b)). Cells were grown to OD_{600} = 0.4 and induced at 0 hours (lane 2 in (a) and (b)). Aliquots of induced protein from Clone 11 for each temperature ((a) lanes 3 - 9 (16°C), lanes 10 - 16 (20°C), lanes 17 - 25 (30°) and (b) lanes 4 - 10 (37°C)) were removed at regular time intervals (0.5 - 6 hours) and analysed by 10 % SDS-PAGE with subsequent silver staining. Host cells containing the control plasmid coding for the β-gal protein (see asterisk) were induced for 3 hours at 37°C (panel (b), lane 3). Standard molecular weight markers are indicated in lane 1 in (a) and (b)).

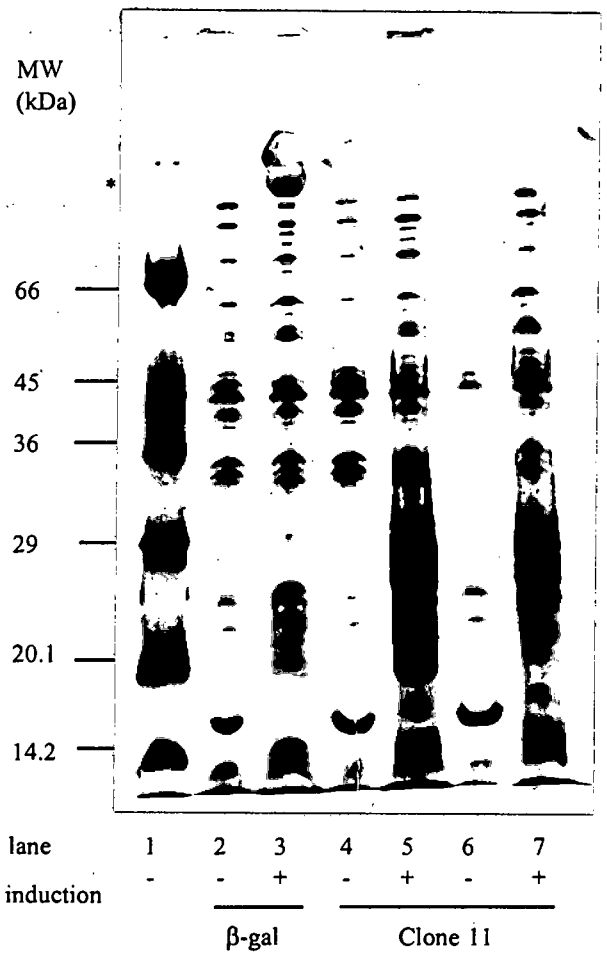


Fig (iii) <u>**Soluble Recombinant Protein Expression from Clone 11 in the pET-29b(+) Expression Vector**</u>

Soluble protein extracts from BL21DE3 cells containing the Clone 11 expression construct (lanes 4 - 7) or the induction control plasmid (lanes 2 - 3) were analysed by 10 % SDS-PAGE (silver stained). Duplicate extracts from induced cells containing Clone 11 (lanes 5 and 7) express a recombinant protein at ~ 25 kDa compared to uninduced cell extracts (lanes 4 and 6). Similarly, the β-gal protein (marked by the asterisk, see lane 3) induced from the induction control plasmid, can be distinguished above the background of uninduced cell extract (lane 2). Standard molecular weight markers are indicated (lane 1).

Manual and see section 2.14.3.2). The mid-log phase cells were subjected to standard induction conditions with 1mM IPTG at 30°C or 37°C. The cells were pelleted and resuspended in column binding buffer containing protease inhibitors (see section 2.14.3.2). The suspension was subjected to sonication in bursts (the length and number of sonication steps was varied in order to optimise the release of the recombinant proteins), and several washes of the pellet in order to firstly reduce the viscosity, and secondly to help solubilise any recombinant protein precipitates. Remaining cell debris was removed by centrifugation. Soluble protein extracts containing recombinant protein from Clone 11 were subjected to Nickel chromatography, however the amount of recombinant protein that was solubilised by sonication was probably insufficient to be visualised in the Nickel column elution profile as analysed by SDS-PAGE. These results were confirmed by repeating the same procedure with the induced β-gal protein. The successful induction of the β-gal protein in total cell extracts is shown in the Coomassie stained SDS gel (fig (iii), lane 3). Repeated sonication and wash steps released more β-gal-protein into the supernatant, these fractions were pooled and fractionated by Nickel column chromatography using gravity flow (see section 2.14.3.2) followed by SDS-PAGE analysis. The column was washed once with binding buffer and once with washing buffer. Subsequently, proteins were eluted in 15 fractions of 1 ml each using elution buffer, after which the column was stripped with a high salt solution. All the fractions collected from the column were TCA precipitated before analysis by 10 % SDS-PAGE which was silver stained as Coomassie was not sensitive enough to visualise the protein bands (data not shown). Proteins eluted from the column during both the wash steps, and the remaining proteins eluted in fractions 1 - 15, however from the profile it was not possible to gage where the β-gal protein eluted. Similar results were obtained for both the β-gal control protein as well as the recombinant protein from Clone 11, indicating that either insufficient solubilised protein was released for subsequent detection in the elution profile, or the Nickel column was prone to ion leaching. It is possible that Nickel is lost from the resin as a result of the wash steps, leading to the loss of the protein-metal complexes, or these complexes could have leached out gradually over several of the eluting fractions. These factors, together with the fact that the expressed recombinant proteins were not easily solubilised and released into the supernatant could all contribute to the fact that neither the β-galactosidase protein nor the recombinant Clone 11 protein could be discerned in the elution profile. Therefore the recombinant protein purification was approached using a different technique, viz by first denaturing the insoluble recombinant proteins using denaturing agents, in order to release greater amounts of recombinant proteins.

(d) Inclusion Body Isolation Methods

Insoluble inclusion bodies (dense aggregates of overexpressed protein which can accumulate in the cytoplasm) may be used as a source for further purification of the target protein (Lin and Cheng, 1991) (186). Their dense nature allows them to be precipitated away from other *E.coli* proteins, and sometimes, depending on the nature of the protein, they can be dissociated by strong denaturing reagents, such as urea or guanidine HCl. The solubilised protein can then be refolded by slowly removing or diluting the denaturant, and the renatured protein can be further purified.

Several methods pertaining to inclusion body isolations and renaturation of recombinant protein were applied. One method (pET System Manual, and section 2.14.3.3) involved resuspension of the induced cells in binding buffer (see section 2.14.3.2), followed by sonication of the suspension in brief bursts to aid resuspension of cells and shearing of the DNA. The resuspended cells were washed twice in binding buffer with subsequent resuspension in binding buffer supplemented with 6 M guanidine HCl or 6 M urea. Resuspension was aided by sonication, which was repeated several times in order to release more trapped proteins from the pelleted inclusion bodies. The suspension was incubated on ice to dissolve the proteins, and the supernatant was subsequently dialysed into dialysis buffer (see section 2.19) in order to remove the guanidine HCl.

A second protocol used to isolate inclusion bodies followed the method described by Calzone et al (1991) (120) (see section 2.14.3.3). This involved standard induction and expression of the recombinant protein, after which the cells were lysed by addition of lysozyme and by short bursts of sonication. The suspension was supplemented with NP-40 and sucrose, after which the insoluble material was removed by centrifugation. Soluble proteins were selectively precipitated with ammonium sulphate, and the resuspended proteins were dialysed against dialysis buffer (see section 2.19).

A third method of inclusion body isolation, developed to isolate active eukaryotic proteins, was outlined by Lin and Cheng (1991) (see section 2.14.3.3). The induced cells were pelleted and resuspended, the outer membranes were removed to form spheroplasts which were lysed by repeated sonication in short bursts. RNA and DNA were digested enzymatically. The crude inclusion bodies were pelleted and washed several times, after which they were resuspended by sonication and denatured in 5 M guanidine HCl. The suspension was supplemented with a large volume of buffer and the proteins were renatured overnight, and subsequently dialysed into the buffer of choice (dialysis buffer).

Despite several attempts to solubilise the expressed recombinant protein by means of extensive sonication, combined with denaturation agents such as urea and guanidine HCl, none of the methods outlined above successfully dissociated the inclusion bodies containing the protein expressed from Clone 11, as assessed by means of SDS-PAGE analysis. The dialysed protein extracts, obtained after extensive sonication of the inclusion bodies (and denaturation of the released proteins) showed no enrichment of the expressed recombinant protein when compared to the insoluble fraction (data not shown). The expressed recombinant protein remained in the insoluble fractions of the inclusion body, ie recombinant proteins could not be solubilised, showing that despite the denaturation / renaturation protocols these methods do not result in the successful isolation of the recombinant protein expressed by Clone 11.

Appendix XV

Recombinant Protein Expression in Eukaryotic COS Cells

Prokaryotic gene expression lacks post-translational modifications, which may lead to incorrect folding of the protein or loss of biological activity of eukaryotic recombinant proteins (151), whereas eukaryotic systems effect higher stability of the target protein because of proper folding, activation, processing and assembly (208). An eukaryotic expression plasmid requires a SV40 origin of replication for high copy number amplification in the host cell. It also needs an efficient promoter element for transcription initiation, mRNA processing signals, a polylinker for easy cloning of inserts, a selectable marker to select for stably integrated cells, and often expression vectors have plasmid backbone sequences which enable their propagation in bacterial cells (151, 208). All these requirements are met by the expression plasmid pCIS (see Appendix VIII) into which the insert from Clone 11 was cloned using the *XbaI* / *NotI* restriction sites in the multiple cloning site. Transient eukaryotic expression is limiting, though, since usually only 5 - 50 % of the cell population is able to acquire and express the DNA (151). Expression of the protein will last over a period of days to several weeks, until the DNA is lost from the population. COS cells express high levels of the SV40 large tumour (T) antigen which initiates replication of the expression plasmid from the SV40 origin (151). Transfected DNA can be amplified to exceed 100 000 copies of the plasmid per cell, which implies that very high recombinant expression levels can be achieved. Cells transfected with Clone 11 were allowed to express for 24 - 48 hours after which they were harvested and processed into whole cell extracts and nuclear extracts (see section 2.15). The extracts were analysed by SDS-PAGE, as shown in fig (iv). Whole cell extract and nuclear extract of untransfected cells are shown in lanes 2 and 3 respectively, whereas the whole cell extract and nuclear extract of transfected cells is shown in lanes 4 and 5 respectively. Several differences (marked by the arrows) are apparent between the total cell extracts and nuclear extracts for both the transfected and untransfected cell populations, however there are no distinguishable differences between untransfected and transfected cells for either type of extract generated (compare lane 2 with lane 4, and lane 3 with lane 5). This was verified by analysis of the extracts in EMSAs using the labelled E/H probe (data not shown), which revealed no DNA-binding activity in any of the extracts shown in fig (iv). These results imply that either the mammalian expression system is very inefficient, possibly due to low transfection success, such that the expression of the recombinant protein could not be ascertained or, alternatively, the protein cannot be expressed in this host environment. Since the expression of the protein was not detected in the COS cells, the vector system should have been examined in more detail to confirm the underlying problems (151). For instance, the expected structure of the expression construct should have been reconfirmed by DNA sequencing. In addition, a positive control using the same vector with another insert (which is known to express reliably in the system) would be needed to determine the transfection efficiency. The expected size of the RNA, as well as its level of expression could be analysed by Northern hybridisation. The expression could also be approached with a different vector.

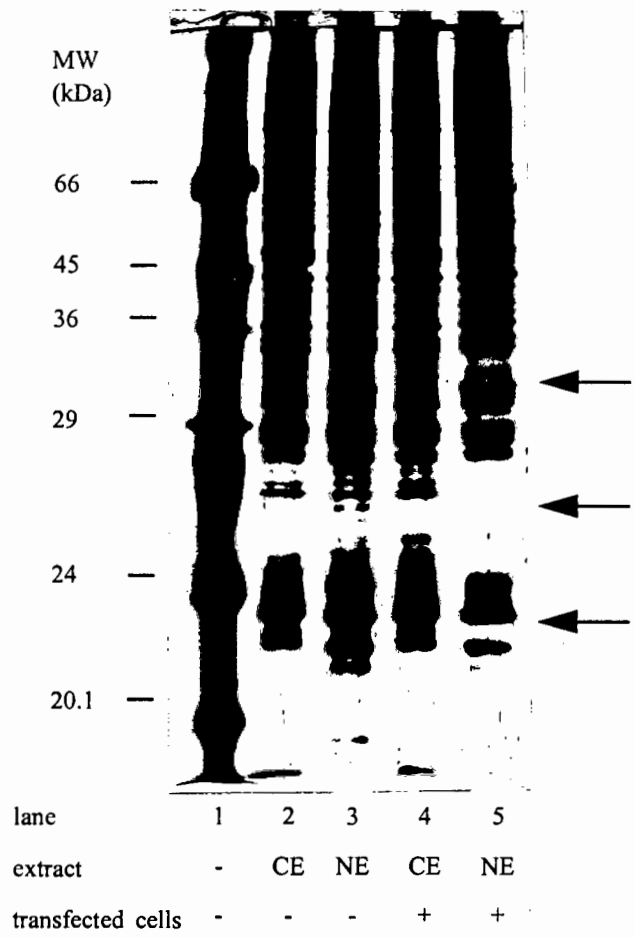


Fig (iv) **Eukaryotic Protein Expression of Clone 11 in COS Cells**

Clone 11 was subcloned from pBluescript into pCIS an the expression construct was transfected into COS cells. Untransfected (lanes 2 and 3) and untransfected cells (lanes 4 and 5) were harvested after 24 - 48 hours and processed into total cell extracts (CE, lanes 2 and 4) and nuclear extracts (NE, lanes 3 and 5). The differences observed between cell extracts and nuclear extracts are marked by arrows, however there appear to be no differences between the extracts from untransfected and transfected cell populations (compare lane 2 with 4 and lane 3 with 5). The standard molecular weight markers are indicated in lane 1.