A COMPARATIVE BIOINFORMATIC ANALYSIS OF ZINC BINUCLEAR CLUSTER PROTEINS

Research Report

Submitted in fulfilment of the requirements for the Degree of

Master of Science in Bioinformatics and Computational Molecular Biology by coursework

in the

Department of Biochemistry, Microbiology and Biotechnology
Faculty of Science
Rhodes University

by

Jabulani S. Mthombeni

September 2004

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my heartfelt thanks to my supervisor Prof. R.A Dorrington for her support, encouragement and endless communication via electronic mail. Thank you for believing in me. I would like to acknowledge my co – supervisor and course co-ordinator, Prof G.L. Blatch for his constant guidance and enduring patience and advise throughout the project.

I would like to extend my sincere gratitude to, Pooja Prasdavijan for the help, support and lighter moments. Thanks to my colleagues in Labs 301 and 417 for their support, ideas and valuable time. I thank my friends in Oakdene House who were fountains of inspiration and pillars of sanity.

I would like to acknowledge the South African Institute of Bioinformatics (SANBI) for their timely and generous funding. My sincere gratitude goes to my family for always being there for me, without their support and belief in me, I would not have come this far. Thank you to Nolwazi for everything.

ABSTRACT

Members of the zinc binuclear cluster family are important fungal transcriptional regulators sharing a common DNA binding domain. Dal81p is a pleotropic zinc binuclear cluster protein involved in the induction of the *UGA* genes required for the γ-aminobutyrate nitrogen catabolic pathway in *Saccharomyces cerevisiae*. The zinc binuclear cluster domain is dispensable for function in Dal81p and little is known about other domains in this protein.

The aim of the study was to explore the zinc binuclear cluster protein family using comparative bioinformatics as a complement to biochemical and structural approaches. A database of all zinc binuclear cluster proteins was composed. A total of 118 zinc binuclear cluster proteins are reported in this work. Thirty nine previously unidentified zinc binuclear cluster proteins were found. Four homologues of Dal81p were identified by homology searching. Important sequence motifs were identified in the aligned sequences of Dal81p and its homologues. The coiled coil motif found in the Gal4p zinc binuclear cluster protein could not be identified in Dal81p and its homologues. This suggested that Dal81p did not dimerise through this structural motif as other zinc binuclear cluster proteins. Solvent accessible site that could be phosphorylated by protein kinase C or casein kinase II and the role of such sites in the possible regulation of Dal81p function were discussed.

LIST OF ABBREVIATIONS

UGA	Utilisation of GABA
UAS_{GABA}	Upstream activating Sequence (γ-aminobutyric acid)
UIS	Upstream Induction sequence
URS	Upstream Repression sequence
CGG	Cytosine - Guanine - Guanine triplet
UAS_{GATA}	Upstream activating sequence
GABA	γ-aminobutyric acid
PDB	Protein Data Bank
DNA	Deoxyribonucleic acid
PIR	Protein information Resource
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
DBJ	DNA databank of Japan

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction to Fungal Genomics

The exponential growth in protein sequence data from huge sequencing projects has driven research into the post genomic era, more precisely functional genomics. The slower nature of experimental approaches used to determine protein structure has necessitated the development of improved approaches for secondary structure determination from protein sequence data. Functional genomics seeks to develop global experimental approaches to assess gene function by making use of information in structural genomics.

The *Saccharomyces cerevisiae* genome was fully sequenced in 1996 through a worldwide collaboration involving the Sanger Institute, Stanford and St. Louis Universities (Goffeau *et al*., 1996). The *S. cerevisiae* genome was the first eukaryotic genome to be fully sequenced. To date the *S. cerevisiae* genome has an estimated 5885 protein-coding genes, approximately 43 % of these sequences have functions have been assigned but the majority still have unknown function (Goffeau *et al*., 1996; Mackiewicz *et al.*, 2002). There is an even smaller subset of proteins in the *S. cerevisiae* genome which have their structures elucidated. Todate the structural information on *S. cerevisiae* proteins available in public databases represents a fraction of the total number of protein sequences identified in *S. cerevisiae*. Detailed study of these proteins presents a challenge to understand fungal genome architecture. Bioinformatics can be used to study datasets of fungal genomic protein sequences using available structural data. Information derived from bioinformatic analysis can be used as a platform for detailed experimental study of the fungal proteins.

A comparative approach of analysing fungal protein sequences using computational tools and techniques can be used to reveal interesting novel fungal protein features.

1.2 General nitrogen metabolism in *S. cerevisiae*

Nitrogen is an essential metabolite for the synthesis of proteins and nucleic acids in the living cell. *S. cerevisiae* has evolved a mechanism of selectively discriminating between good and poor sources of nitrogen in the nutritional environment (Marzluf, 1993). In the presence of rich nitrogen sources such as ammonia or glutamate a physiological response termed catabolite repression is activated resulting in gene expression inhibition (Marzluf, 1993). The presence of poor nitrogen sources such as γ-aminobutyrate (GABA) results in induction of *UGA* genes whose products are responsible for the catabolism of the substrate (Ramos *et al.,* 1985). Nitrogen metabolism regulation has been shown to be under the influence of a cluster of regulatory gene elements of distinct nitrogen catabolic pathways involved at the transcriptional level (Marzluf, 1997).

1.2.1 GABA metabolism in *S. cerevisiae*

There are three different permeases that transport GABA into the cell while the degradation of GABA degradation requires three enzymes (Grenson *et al*., 1987). These include two non-specific permeases, proline permease encoded by the gene (*PUT4*), and a general amino acid permease encoded by the gene (*GAP4*). The other permease, Uga4p (GABA permease) encoded by *UGA4* is induced in the presence of GABA (Vissers *et al.*, 1989). The *UGA1* gene encodes GABA transaminase (γ-aminobutyrate: 2 –oxo-glutarate) involved in the initial step of GABA degradation resulting in the formation of 2 –oxo-glutarate and glutamate (Ramos *et al.,* 1985). Spontaneous conversion of 2 –oxo-glutarate yields succinate semialdehyde. Succinate semialdehyde dehydrogenase will be referred to as Uga2p for the purposes of this review (Coleman *et al.*, 2001). Conversion of the succinate semialdehyde to succinate has been shown to occur under the influence of Uga2p (succinate semialdehyde dehydrogenase) encoded by the *UGA2* gene (Coleman *et al.,* 2001). Mutant *S. cerevisiae* cells lacking either *UGA1* or *UGA2* genes grown on GABA media as the sole nitrogen source exhibited impaired growth phenotypes when compared to wild type *S. cerevisiae* cells (Coleman *et al.*, 2001). These observations showed the crucial role of these gene products in the catabolism of GABA to ensure the supply of nitrogen to the cell in repressed conditions. Expression of *UGA* genes occurs at basal

levels under non-repressive conditions coupled with the absence of metabolic precursors (Ramos *et al*., 1985). Availability of GABA in the absence of rich nitrogen sources results in induction of *UGA* gene expression at higher levels (Ramos *et al*., 1985).

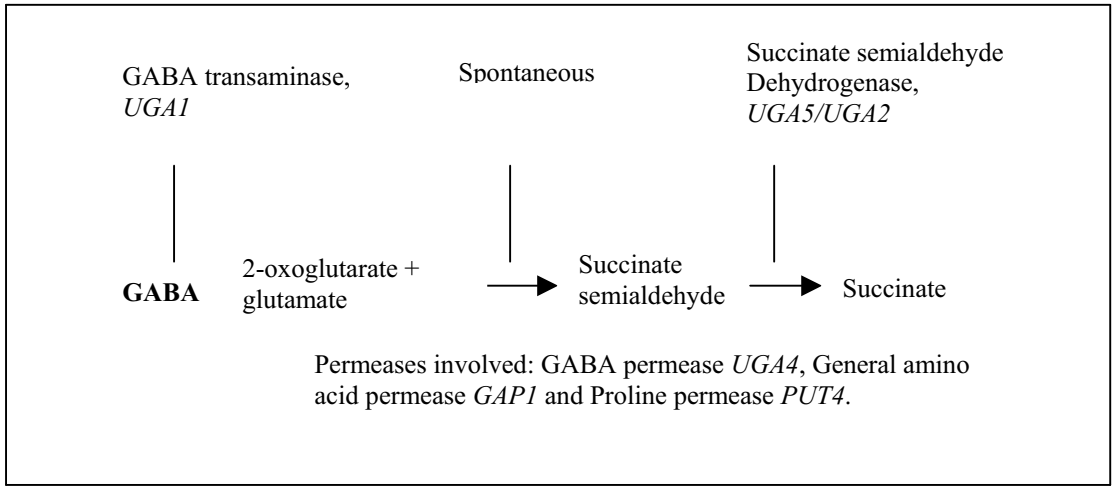

Figure 1.1 GABA metabolism pathway: Schematic representation showing metabolic enzymes encoded by the genes, permeases involved in transport of GABA into the cell and specific enzymes.

1.2.2 Transcriptional Regulation: GABA metabolism in *S. cerevisiae*

Cis- and *trans*-acting factors are known to mediate nitrogen catabolite repression (NCR) and induction in *S. cerevisiae*. *Cis*-factors are promoter elements found upstream 5' end of the *UGA g*enes to which trans-acting factors bind. Three distinct cis–acting factors have been characterised, an upstream activation sequence (*UAS*) essential for transcriptional activation, an upstream repression sequence (*URS*) maintains transcription at basal levels in the absence of an inducer and an upstream induction sequence (*UIS*) responsible for inducer specific interaction (Rai *et al.,* 1989; Yoo and Cooper, 1989). Two positive transcription factors, Gln3p and Gat1p have been characterised as trans-acting factors in NCR (Stanbrough *et al*., 1995; Coffman *et al*., 1996).

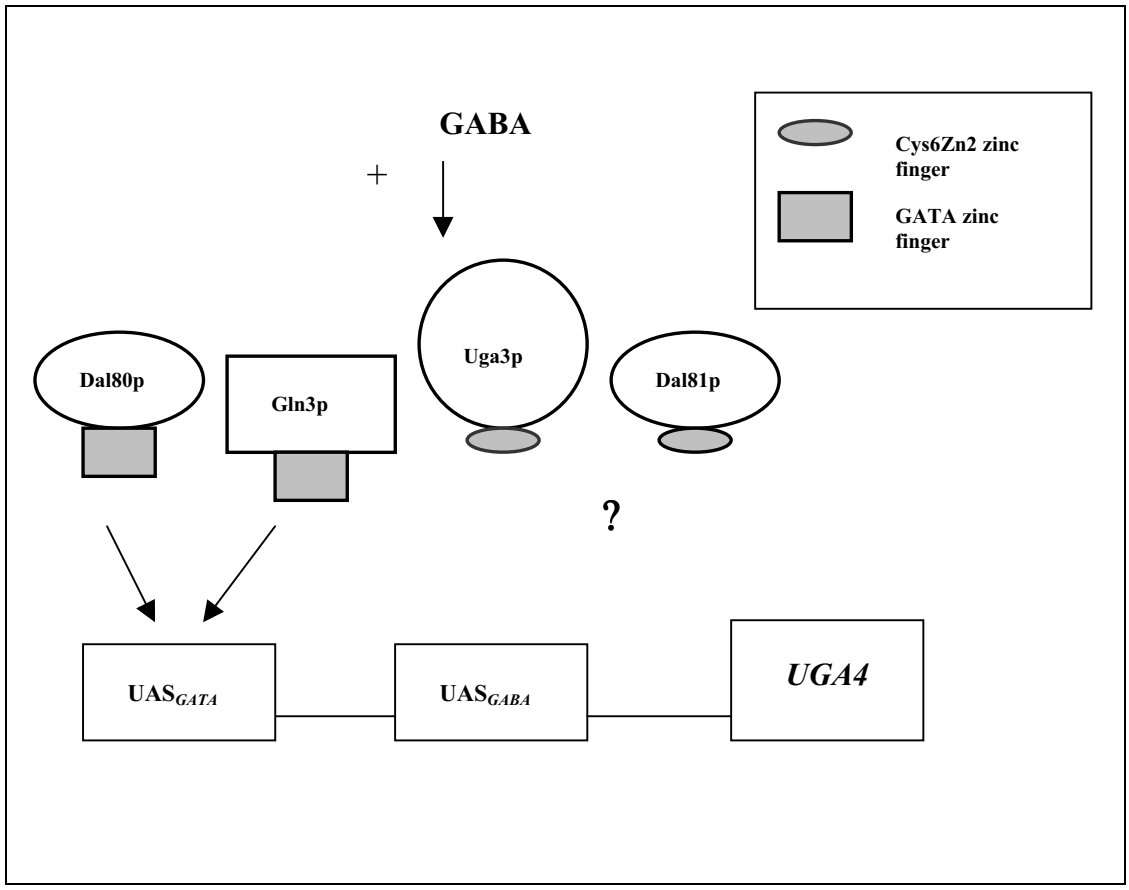

Figure 1.2: **Prototypical model illustration cascade of events of induced γ-aminobutyric acid (GABA) transcription of the *UGA4* gene (GABA permease).** The interaction of two elements (**UAS$_{GABA}$** and **UAS$_{GATA}$**) is necessary for high-level induced expression of the *UGA4* gene. Uga3p and Dal81p are believed to interact at the UAS$_{GABA}$ level, synergistically recruit Gln3p (global nitrogen regulatory factor) that competes with Dal80p (repressor) on the UAS$_{GATA}$ site in the presence of GABA (Talibi *et al.*, 1995). There is no evidence in literature that Dal81p binds directly to DNA.

A minimum of five critical transcription factors are believed to regulate GABA gene induction: Gln3p, Ure2p, Dal81p, Uga3p and Dal80p (André *et al.* 1995; Talibi *et al*., 1995). Gln3p is global nitrogen regulatory protein in *S. cerevisiae* containing a GATA DNA binding domain consisting of two homologues adjacent zinc fingers separated by a linker region (Trainor *et al*., 2000). In the presence of a rich nitrogen sources Gln3p is hyperphosphorylated and found to be bound to phosphorylated Ure2p in the cytoplasm (Cardenas *et al*., 1999). The presence of poor nitrogen sources in the nutritional environment stimulates Gln3p – Ure2p dephosphorylation cascades resulting in Gln3p cytoplasmic to nuclear migration (Cardenas *et al*., 1999). A repressor, Dal80p is believed to compete with Gln3p for GATA sequences (Rai *et*

al., 1999). Under nitrogen repressed conditions bound Dal80p ensures low level expression of nitrogen catabolism pathway genes (Rai *et al.*, 1999).

A 19 base pair GC rich sequence was identified as an UAS_{GABA} essential for Dal81p and Uga3p interaction for *UGA* induction (Figure 1.2). Uga3p has been shown to bind to the UAS_{GABA} through the everted repeat CGG – N_4 – CGG (Noel and Turcotte, 1998). Simultaneous occupation of these sites has been shown to be essential for GABA–dependent transcriptional activation *in vivo* (Idicula *et al.*, 2002). Dal81p is a pleiotrophic positive transcription factor and together with Uga3p has been shown to be essential for *UGA4* GABA - induced transcription (Andrè *et al*., 1995). Dal81p has been shown to be also required for the allophonate-triggered induction of genes in urea, allantoin and ornithine metabolism (Vissers *et al.,* 1990). Deletion of *UGA3* and *DAL81* impaired the activation of *UGA1* and *UGA4* genes (Vissers *et al.*, 1990). There has been experimental evidence showing that a combination of Dal81p and Uga3p interaction are thought regulate the induction of *UGA* genes (Vissers *et al.*, 1990).

The co-interaction between the two elements UAS_{GATA} and UAS_{GABA} is speculated to result in high level induced transcription of the *UGA4* gene (Talibi *et al.*, 1995).

1.3 Uga3p and Dal81p

Uga3p and Dal81p are members of the zinc binuclear cluster family. Uga3p is 528 amino acids long consisting of a zinc cluster domain (residues 16-46), a predicted coiled coil (residues 51-67) and a C-terminal located acidic domain (residues 504-521) (Schjerling and Holmberg, 1996). The Middle Homology Region (MHR) speculated for the role of reducing improper binding sites on DNA was undetected by computational analysis (Schjerling and Holmberg, 1996). Interestingly a C-terminal located putative WD-40 like motif was identified using computational analysis (Idicula *et al*., 2002). Repeated WD-40 motifs form a blade propeller like structure which acts as a site for protein-protein interactions in various cellular roles such as transcriptional regulation and signal transduction (Neer *et al.*, 1994). The WD-40 domain has been speculated to increase the binding specificity of Uga3p to DNA for

optimal GABA transcriptional activation (Idicula *et al*., 2002). The question of exactly how Uga3p interacts with UAS_{GABA} remains relatively unclear since there is little known structural data.

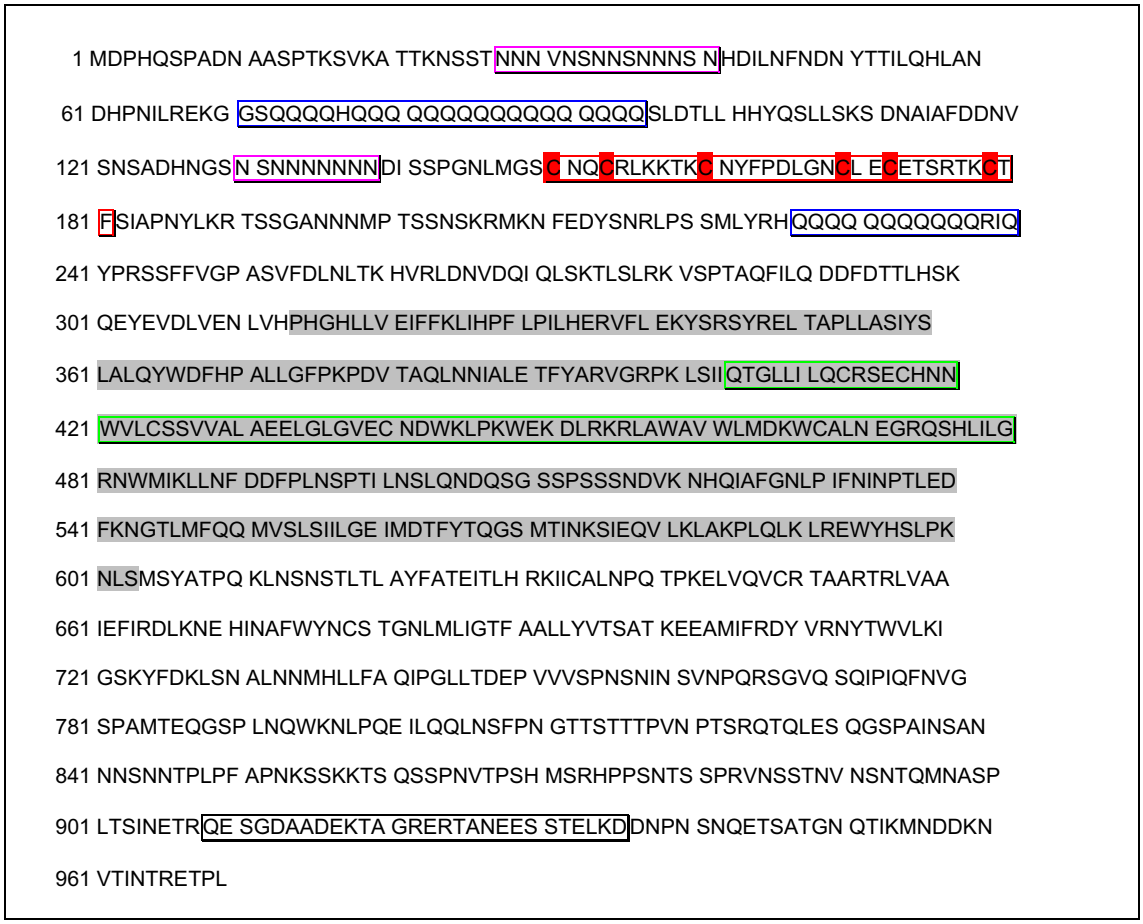

Figure 1.3: Schematic illustration of the protein sequence of Dal81p. Homopolymeric stretches shown boxed in pink, residues 28-41 and 130-138 respectively. Polyglutamine stretches shown boxed in blue, residues 73-94 and 227-237 respectively (Bricmont *et al*., 1991). Zinc cluster region shown boxed in orange (residues 149-181) while the lime boxed area shows middle homology region (405-477) (Schjerling and Holmberg, 1996).The eight motif domain (residues 314-603) is show by the grey shaded area (Poch, 1997) The acidic domain is shown by the black boxed area (Schjerling and Holmberg, 1996).

The *DAL*81 gene encodes a 970 amino acid protein in *S. cerevisiae* containing sequences homologous to the zinc cluster motif (Bricmont and Cooper, 1989). *S. cerevisiae* cells containing Dal81p mutants lacking the zinc finger domain were found to be functionally unaffected by the deletion (Bricmont *et al*., 1991). Separate deletion studies on the each of the polyglutamine stretches on Dal81p were carried out (Bricmont *et al.*, 1991). A 50 % loss of induced urea amidolyase activity was

observed in a group of Dal81p mutants (Figure 1.3) (deletion: residues 73-94) compared to wild type but there was no observed phenotype in growth when the mutants were grown in media with GABA as the sole nitrogen source. Dal81p mutants lacking a second glutamine stretch (residues 227-237) did not have a detectable phenotypic loss of function (Bricmont *et al.,* 1991). The predicted coiled coil domain was found to lie on this polyglutamine stretch (Figure 1.3, residues 227-240) (Schjerling and Holmberg, 1996).

Glutamine rich domains have been shown to be important in transcriptional regulation by interacting with other transcription factors in *S. cerevisiae* and *Homo sapiens* (Escher *et al*., 2000). The middle homology region and the acidic domain were designated to lie between residues 405-477 and residues 909-936 respectively (Schjerling and Holmberg, 1996). The eight-motif domain in Dal81p thought to be involved in the regulation of some zinc binuclear cluster proteins was found to encompass residues 314-603 (Poch, 1997).

A related fungal homologue to Dal81p, TamAp (739 amino acids long) in *Aspergillus nidulans* has been reported (Small *et al*., 2001). Deletion of the first 152 amino acids in TamAp did not have a phenotypic effect on *Aspergillus sp.* grown in media with GABA as the sole nitrogen source. Interestingly the zinc finger motif is contained in this sequence suggesting that the motif was dispensible for function (Small *et al.*, 2001). Deletion studies on most regions of TamAp resulted in loss of function; overall conformation of TamAp may be essential for function (Small *et al.*, 2001). Full-length protein sequence similarity between the Dal81p and TamAp was observed at 40 % (Small *et al.*, 2001). A higher degree of amino acid sequence similarity was observed in the zinc-binding domain (77 %) (Small *et al.*, 2001). TamAp has a characteristic serine / threonine rich region (residues 105-128), a central domain (residues 376-421) which was indispensible for TamAp function. The region exhibited a 59 % identity and 77 % similarity in amino acid residues with an equivalent region of Dal81p on a multiple sequence alignment (Small *et al.*, 2001). Another homologue Otamp encoded for by the *tamA* gene in (*Aspergillus oryzae*) was found to consist of 711 amino acids (Small *et al.*, 2001). Three potential nuclear localisation signals (NLSs) in the protein sequence of TamAp have been identified

suggesting that TamAp is located in nucleus the cell. The NLSs sequences are not found conserved in Dal81p (Small *et al.*, 2001).

1.4 Basic principles of Protein Structure

Proteins have evolved a unique architecture termed regular secondary structure formed as a result of amino acid side chain hydrogen bonding (Lesk, 2001). Globular proteins form an ordered arrangement that consists of a "hydrophobic core" and a hydrophilic surface (Branden and Tooze, 1991). Polar main-chain groups along the polypeptide interact with hydrophobic residues through hydrogen bonding. This leads to main-chain folding into the protein interior. This defines the two main structural features of a protein: α-helices and β-strands (Branden and Tooze, 1991). Regions linking chains in globular protein are termed loops or random coils. Loops lack properly defined secondary structure and have a preference for outer regions of the proteins while α-helices and β-strands have a preference for the protein core.

1.4.1 The Alpha (α –Helix)

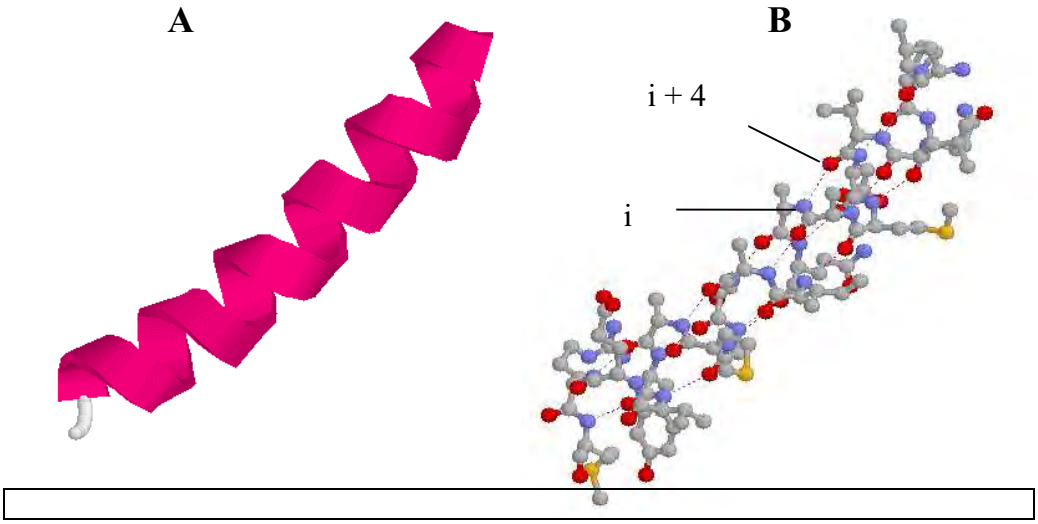

Figure 1.4: **Diagrammatic representation of a helical region** (residues 360-380) crystal structure Of Human Tyrosine-Protein Kinase C-Src (Xu *et al.*, 1997) [PDB: 1FMK]. **A**, Cartoon representation. **B**, Ball and stick representation is shown, hydrogen bonds shown as dotted lines. Residue (i) forms a hydgrogen bond with Residue (i+4). Molecular visualisation by RASMOL (Sayle and Milner-White, 1995).

A typical α-helix has 3.6 amino acid residues per turn and can either be right or left handed (Pauling *et al.,* 1951; Branden and Tooze, 1991). Hydrogen bonds are formed between the NH group of residue i with the C=O of the residue i + 4, four residues away (Figure 1.4). Steric hinderance of amino acid side chains dictates a preference for certain amino acid residues in ∝- helix formation. Glutamate is an excellent helix-forming amino acid since the side chains participate in hydrogen bonding, while proline has a sidechain that interrupts the hydrogen-bonding pattern in ∝-helices (Lesk, 2001). A consensus pattern of hydrophobic non-polar residues on one side of the helix and hydrophilic and polar residues on the other side has been observed in amphipatic α-helices (Lesk, 2001). The amphipatic α-helices present a hydrophilic segment to the aqueous environment and a hydrophobic segment away from the solvent.

1.4.2 (β) Beta Sheets

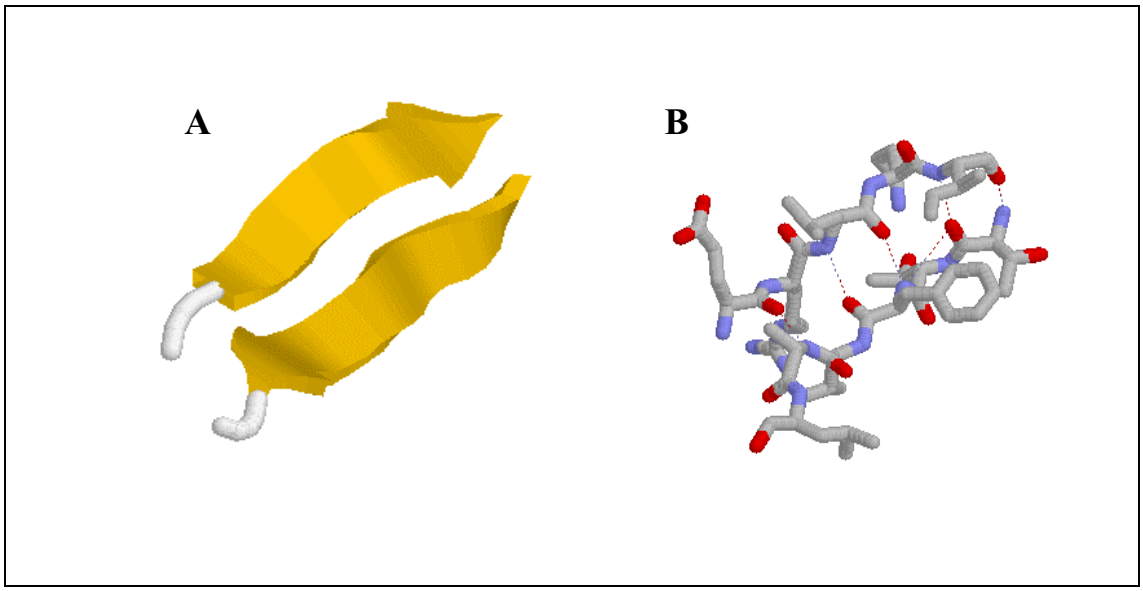

Figure 1.5: Diagrammatic representation of typical beta sheet region (residues 84-89, 106-110) Crystal Structure of Human Tyrosine-Protein Kinase C-Src (Xu *et al.*, 1997) [PDB: 1FMK]. **A**, cartoon representation shown . **B**, Stick representation shown, hydrogen bonds shown as dotted lines. Molecular visualisation by RASMOL (Sayle and Millner- White, 1995).

β-Sheets are formed through hydrogen bonding between adjacent β-strands, which are 5-10 amino acids long (Figure 1.5). Linear hydrogen bonding between amino acids on different polypeptide chains between the amino group of one β-strand with

the carboxyl group of another β-strand results in the formation of β-sheets (Pauling and Corey 1951; Branden and Tooze, 1991). Hydrogen bonding patterns in the same or opposite directions result in either parallel or antiparallel beta sheet configurations respectively. β-Strands can combine with mixed β-sheets and β-strands in globular proteins (Figure 1.6).

Loop regions are characteristically varied in length and have irregular shape (Branden and Tooze, 1991). Loops may also connect α-helices and β-sheets to form the hydrophobic core of the protein, allowing surface exposed main-chain amino and carboxyl groups to hydrogen bond with water molecules (Branden and Tooze, 1991). Analysis of homologous proteins shows that amino acid insertion and deletion events occur mostly in loop regions as compared to α-helices and β-sheets (Branden and Tooze, 1991).

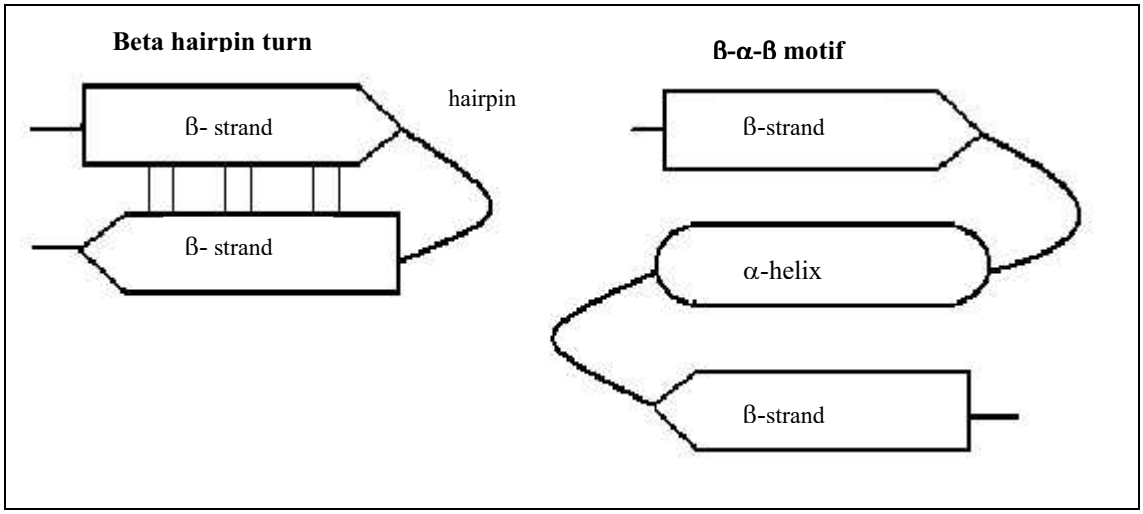

Figure 1.6: Schematic representation of supersecondary structure. Beta hairpin turn links two beta strands hydrogen bonded through adjacent mainchain. β-α-β motif consists of two parallel β-strands linked to an alpha helix oriented parallel to the β-strands. The motif forms part of the hydrophobic core (Lesk, 2001).

Beta-hairpins are short loop regions connecting adjacent strands allowing the polypeptide chain to change direction (Figure 1.6). The organisation of two or more consecutive α-helices and β-sheets give rise to protein supersecondary structures

such as the β-α-β unit which are important in forming globular structure (Lesk, 2001).

1.5 Functional protein domains

Protein domains are interactive modules involved in protein-protein / protein-nucleic acid interactions. Protein domains are important in biological processes including cell signalling, transcription, cell cycle events and immunological response (Alberts *et al.*, 2002). The growing amount of available structural data on functional protein domains in databases has lead to the prediction of potential functional sites from primary amino acid sequence. The approach of using known protein sequence profiles has been reliably used to search for potential biologically significant sites in novel proteins (Rost and Sander, 2002). The same approach can be used in zinc binuclear cluster proteins.

1.5.1 Coiled coil domains

Coiled coil motifs are the most common protein- protein interactive modules. Coiled coil domains have been implicated in biological processes such as transcription and membrane fusion (Lupas, 1996a). A typical coiled coil has two or more α-helices interacting with each other to form a super-helical twist. The core of a coiled coil is the heptad repeat, a structure formed by a linear pattern of seven amino acids (Keating *et al.*, 2001). The orientation of amino acid residues in side chains within a heptad repeat have been shown to give rise to left or right handed super-helical twists resulting in a structure with a periodicity of approximately 3.6 residues per turn.

The projection of the dimeric coiled coil (Figure 1.7) onto the helical arrangement leads a super coiled structure that consists of a matrix with hydrophobic residues in the centre of the domain with the polar /charged residues forming an interface between the centre and solvent. This characteristic pattern enables accurate prediction of this motif using computational analysis of protein sequence (Wolf *et al.*, 1997). The COILS prediction server (Lupas *et al.*, 1991) uses a technique predicting long coiled coils. Shorter coils characteristic of zinc finger proteins (Schjerling and Holmberg, 1996) were predicted less accurately when assigned with a four heptad repeat window (28 residues).

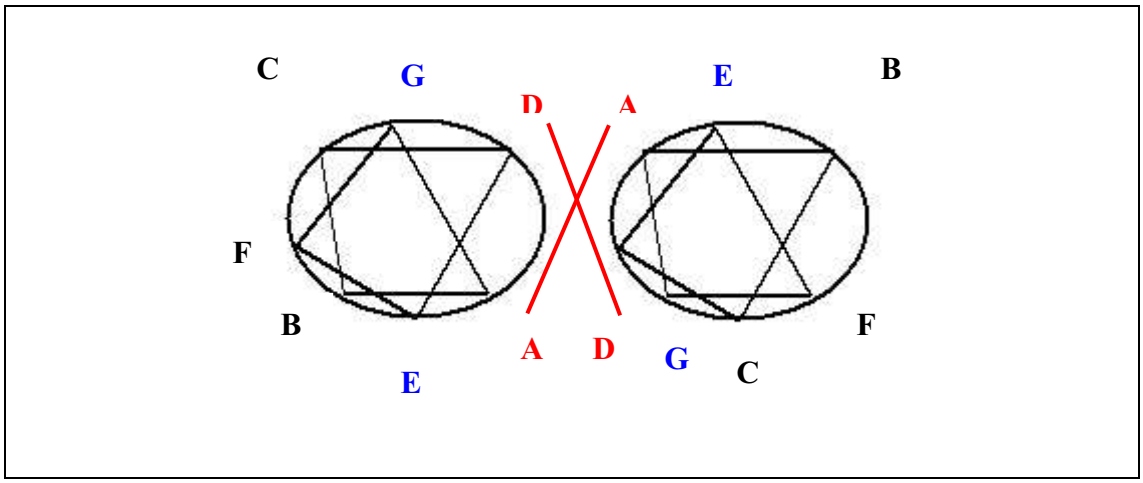

Figure 1.7: Prototypical coiled coil helical wheel representation. The heptad repeat sequence in each of the two ∝-helices is represented above $(ABCDEFG)_n$. Positions A and D (red) are preferentially occupied by hydrophobic residues. Positions E and G (blue) are preferentially occupied by charged / polar residues (Keating *et al*., 2001).

Interestingly coiled coils predicted using a heptad of two repeats (14 residues) correlated to coiled coils in Gal4p and Ppr1p as determined by X-ray crystallography (Schjerling and Holmberg, 1996).

Fewer coiled coils were predicted by another prediction method in zinc cluster protein study (Schjerling and Holmberg, 1996). The PAIRCOIL method (Berger *et al.*, 1995) was observed to predict fewer coiled coils compared to the COILS method using a window of 14 residues. The PAIRCOIL algorithm uses a fixed heptad of 28 residues (Berger *et al.*, 1995). The COILS algorithm is inherently biased towards positively charged protein sequence stretches and default settings have been shown to predict coiled coils in the absence of a heptad repeat (Lupas *et al*., 1996a). A weighting option, which assigns equal weighting to all seven residues, has been shown to have a negative effect on algorithm performance (Lupas *et al*., 1996b). The method was shown to be less accurate at detecting shorter coiled coils characteristic of zinc binuclear cluster family (Schjerling and Holmberg, 1996).

1.5.2 Potential modification sites: Phosphorylation sites

Protein phosphorylation of tyrosine, serine and threonine residues is key in cell signalling pathways and represents an important protein reversible post-translational modification (Blom *et al.*, 1999). Protein phosphorylation involves addition of a \phosphoryl group from ATP catalysed by phosphokinases. The reverse reaction resulting in phosphoryl cleavage is catalysed by phosphatases. Specificity of phosphorylation is determined by the acidic, basic or hydrophobic amino acid residues located in proximity to the phosphorylated residue (Blom *et al.*, 1999).

Transcription factors are part of a highly regulated intricate network of transcription machinery; post-translational modification is one of many mechanisms that activate a transcription factor. Phosphorylation at multiple sites has been shown to be critical for nuclear import for transcription factor Pho4 (Kaffman *et al.*, 1998; Komeili and O'Shea, 1999). Specific phosphorylation of Ser 699 residue in Gal4p is critical for optimal transcriptional activation of the *GAL* genes (Sadowski *et al.*, 1996). The identification of probable phosphorylation sites may provide an insight into the regulation of transcription factors.

1.5.3 The Nuclear Localisation Signal (NLS)

Three classes of macromolecules are primarily involved in active transport of proteins between the nucleus and cytoplasm. Some proteins bind directly to receptors while some receptors form receptor substrate complexes by binding to one or more adaptors (Mattaj and Englmeier, 1998). The formation of the receptor substrate complex facilitates for either export or import of the protein across the nuclear envelope via the nuclear pore complexes (NPC). Once the transport process has been completed, the transport complex dissociates to release the adaptors and receptor sites.

This review will focus on the nuclear import mechanism involving NLS. The NLS is typically a defined protein sequence motif recognised by a molecular mechanism of regulating protein entry into the nucleus. Two types of NLS have been reported, the monopartite and bipartite motifs (Boulikas, 1993). The monopartite motif consists of a short consensus sequence, –K (K/R)-X- (K/R)- of basic amino acid residues (Hodel

et al., 2001). Two clusters of basic residues separated by a linker of about 10-12 residues characterise the bipartite motif (Hodel *et al*., 2001).

Protein import involving the NLS occurs in two sequential stages. The first stage involves the energy independent docking of the protein at the cytoplasmic surface with importin α-protein resulting in attachment on the exterior of the nuclear pore channel (Mattaj and Englmeier, 1998). Importin α-proteins have two subunits that each have specialised functions in docking. The importin α-subunit binds to the protein by detecting the NLS sequence while the importin β-subunit interacts with the NPC (Adam *et al*., 1989). Two additional proteins, GTPase and p10/NTF2 are critical for ATP driven translocation process that facilitates import into the nucleus through the NPC (Corbett *et al*., 1995). Permeabilized cells require p10/NTF2 for effective NLS-protein nuclear import (Moore and Blobel, 1994). This protein interacts with a number of proteins during the GDP-bound state (Paschal and Gerace, 1995).

The sub-cellular location of proteins such as transcription factors has important functional implications (Nair and Rost, 2003). Fungal zinc binuclear cluster proteins exhibit diversity in nuclear import pathways, nuclear import is a process poorly understood to date (Nikolaev *et al*., 2003). The vast amount of sequence data in databases has necessitated the development of quicker automated methods to define nuclear localisation as an alternative to the slower molecular approaches (Nair and Rost, 2003). Putative nuclear localisation sites can be predicted using an algorithm that utilises nuclear localisation data from proteins with known sites in proteins (PredictNLS method: http://cubic.bioc.columbia.edu/predictNLS/) (Nair *et al.*, 2003). Computational predictive methods are relatively accurate and can handle vast mounts of data. The method has been shown to identify approximately 50% of all known nuclear localisation signals (Rost and Liu, 2003).

1.6 Zinc finger proteins – General overview of their function in fungal cells

Zinc finger domains are a common motif for specific DNA binding activities in fungi. Zinc atoms coordinate with amino acid residues from the polypeptide chain ensuring

stability of the domain. The domain is common in proteins participating in a wide range of cellular activities (Krishna *et al.*, 2003). Zinc fingers have been recently classified into eight different structural groups (Krishna *et al.*, 2003).

Table 1.1: Structural classification of zinc finger proteins

Zinc Finger fold group	Description of ligand placement
C2H2 like	Two ligands from a knuckle and two more from the C-terminus
Gag knuckle	Two ligands from a knuckle and two from a short helix or loop
Treble cleft	Two ligands from a knuckle and two more from the N-terminus
Zinc ribbon	Two ligands from two knuckles
Zn2/Cy6	Two ligands from N terminus of a helix and two more from a loop
TAZ2 domain like	Two ligands each from the termini of two helices
Zinc binding loops	Four ligands in a loop
Metallothionein	Cysteine rich metal binding loop

Description of the zinc finger proteins adapted from (Krishna *et al*., 2003)

1.6.1 Zinc binuclear cluster proteins

Bioinformatic analysis of zinc binuclear clusters has revealed five putative domains: the zinc binuclear cluster domain, the linker region, a coiled coil domain, the middle homology region and an activation domain (Schjerling and Holmberg, 1996). The zinc binuclear cluster domain (Figure 1.8) was generally N-terminally located and found to bind to DNA.

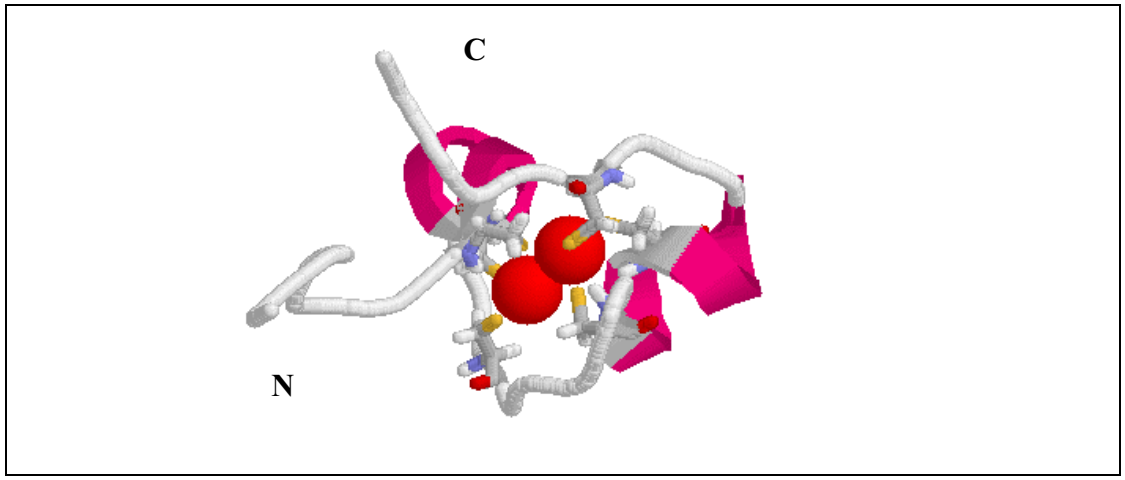

Figure 1.8: Diagrammatic representation of Gal4p DNA binding domain (Baleja *et al.*, 1997) [PDB: 1AW6]. Metal nuclei (red balls) tetrahedrally coordinated by six-cysteine residues (Yellow sticks show sulphur atoms in cysteine residues). Two short α-helices (pink) followed by an extended strand make up the domain. Molecular visualisation by RASMOL (Sayle and Milner - White, 1995).

Structural studies of the conserved DNA binding domain have revealed two zinc nuclei coordinated in a tetrahedral manner by six cysteine residues (Figure 1.8). The zinc binuclear cluster domain was found to have the general consensus sequence - $CX_2CRX_2KXKCDX_3PX_2CX_2CX_6C$-, where X denotes other amino acid residues (Schjerling and Holmberg, 1996). The folded polypeptide chain contains two α-helices separated by a short loop region from which the cysteine coordinate with the zinc nuclei. The zinc binuclear cluster binds to specific DNA triplet sequences. There are over 1200 CGG triplet sites in the yeast genome; specific binding to certain sites is critical for transcriptional activation. Many zinc cluster proteins have been shown to bind as symmetrical homodimers to a pair of CGG triplets in the promoters of regulated genes (Mamane *et al.*, 1998). Three types of DNA binding sites to which zinc binuclear cluster proteins have been characterised: the direct repeat (CGG Nx CGG), the inverted repeat (CGG Nx CCG) and the everted repeat (CCG Nx CGG), where Nx represents a variable number of nucleotides (Figure 1.9)

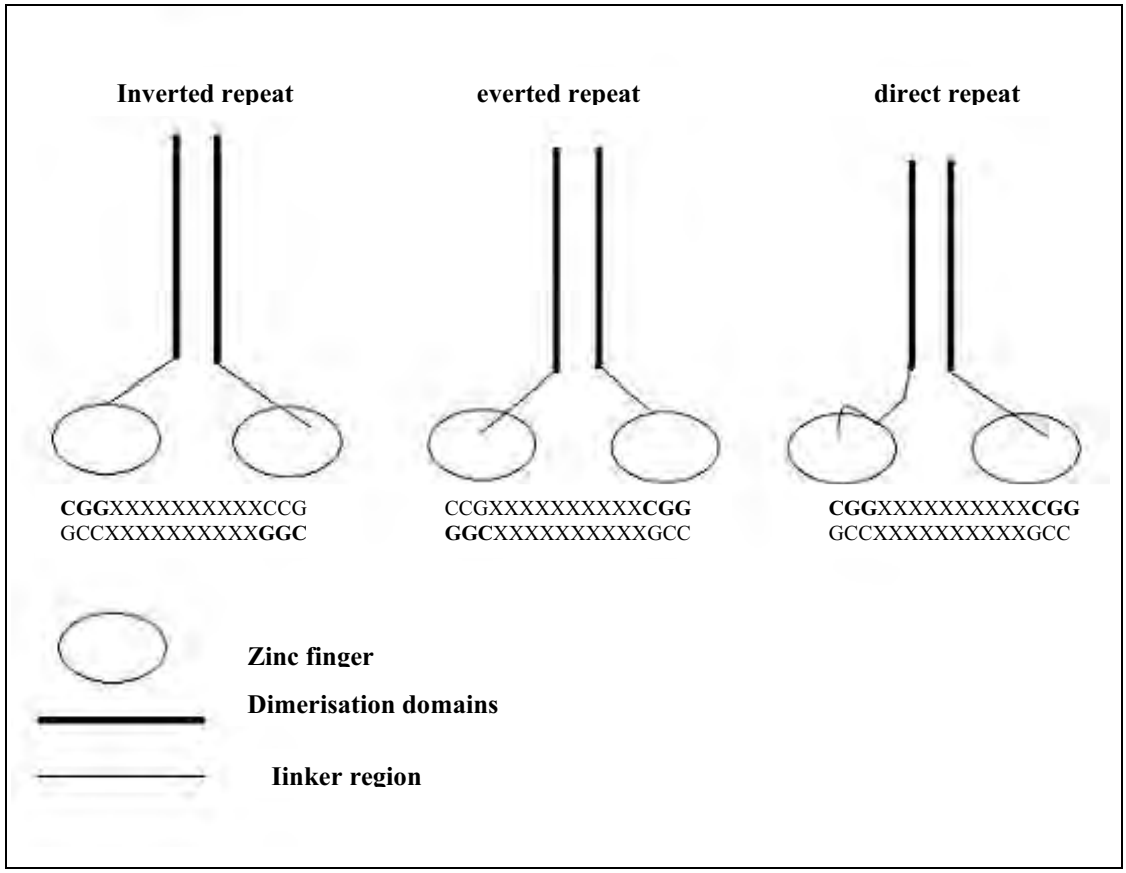

Figure 1.9: **Prototypic models of zinc binuclear cluster proteins binding to DNA sequences**. Shown are the one-dimensional modes of DNA binding as an everted, direct or inverted CGG repeat (in bold) (Mamane *et al.*, 1998).

(Hellauer *et al.,* 1996; Mamane *et al.,* 1998). Structural studies carried out on Gal4p, a well-studied zinc binuclear cluster protein revealed short coil coils interrupted with short stretches of amino acids (Marmorstein *et al.*, 1992). These coiled coil domains have been implicated in specific homodimerisation of zinc binuclear cluster proteins. A short linker region connects the coiled-coil domain with the zinc cluster. Comparison of linker regions in the zinc binuclear family did not reveal any distinct homology (Schjerling and Holmberg, 1996). The middle homology region has been thought to modulate the affinity of binding to incorrect CGG triplets by the zinc binuclear cluster domain (Schjerling and Holmberg, 1996). A C-terminal located acidic domain thought has been thought to participate in transcriptional activation in some zinc binuclear cluster proteins.

1.7 Problem Statement

Dal81p is zinc binuclear cluster protein with a pleotrophic role and is required for the induction of genes in nitrogen catabolite pathways. The zinc binuclear cluster domain has been shown to be dispensable for function in Dal81p. There is however little information known about other domains that potentially participate in regulatory cascades during inducer–dependent transcription. A comparative bioinformatic study of Dal81p and its homologues could reveal information that contributes to the understanding of the regulatory role of Dal81p in nitrogen catabolite repression.

1.8 Research Hypothesis

Dal81p and its homologues share common conserved amino acid sequence features outside the zinc binuclear cluster domain. The conserved features correlate to domains that are important for function during the inducer-dependent transcription of nitrogen catabolite genes.

1.9 Research Objectives

i) To assemble a zinc binuclear cluster protein database.

ii) To conduct a comparative study on the zinc binuclear cluster proteins in the database.

iii) To identify potential Dal81p homologues and reveal common protein sequence features amongst the homologues.

iv) To identify putative regulatory domains and predict the secondary structure of Dal81p.

CHAPTER 2

EXPERIMENTAL MATERIALS AND METHODS

2.1 General database searches

Zinc binuclear cluster proteins of fungal origin were obtained from the Entrez protein sequence cross database search engine on the 23rd November 2003 (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein). The search integrated the European Molecular Biology Laboratory (EMBL) (http://www.embl-heidelberg.de/) database, and the SWISSPROT (http://us.expasy.org/sprot/) database. The Protein Information Resource (http://pir.georgetown.edu/), a non-redundant database was searched for sequences containing the consensus zinc binuclear cluster motif.

2.2 Coiled Coil prediction

Coiled coils were predicted using the COILS algorithm (default settings) on the prediction server interface (http://www.ch.embnet.org/software/COILS_form.html) (Lupas *et al*., 1991). A fixed window of 14 amino acid residues, two repeats of a heptad was used to predict coiled coils (Schjerling and Holmberg, 1996).

2.3 Zinc binuclear cluster alignments

Multiple protein sequence alignments of the zinc finger domain were done using the GCG program (CLUSTALW 18.1), using the European Bioinformatics Institute (EBI) server interface (http://www.ebi.ac.uk/clustalw/index.html) on the 3rd December 2003 (Thomson *et al*., 1994). The Blosum62 matrix was used as the likelihood method for estimating the occurrence of pairwise subsitutions (Henikoff and Henikoff, 1992) for generating the alignments, the sequences were manually inspected and aligned further refine the alignments.

2.4 Dal81p homologue search

Dal81p homologues were obtained using the (Default settings) PSI-BLAST program (Altschul *et al.*, 1997) on the 23rd November 2003. The program used an iterative approach to build an initial profile used to find distant related proteins elusive in normal database searches.

2.5 Dal81p Homologue Multiple sequence alignments

Multiple protein sequence alignments of the five homologues were done using the GCG program (CLUSTALW 18.1), using the European Bioinformatics Institute (EBI) server interface (default settings) (http://www.ebi.ac.uk/clustalw/index.html) on the 3rd December 2003 (Thomson *et al.*, 1994). The sequences were manually inspected to further refine the alignments.

2.6 Kyte-Doolittle Hydrophobicity profiles

Hydrophobicity plots were used to predict potential protein structures using the Kyte-Doolittle method (Kyte and Dolittle, 1982). A window size of nine was used to detect similar regions from the alignments.

2.7 Protein Secondary Structure Prediction

Three neural network-based protein secondary structure predictive methods were used to analyse data obtained from the primary amino acid sequences. The approach was to use 3rd generation algorithms predictive tools that incorporate information obtained from homologues. Algorithms have an estimated prediction accuracy of over 70 % (Rost and Liu, 2003). The algorithms used included Profile fed network systems from Heidelberg (PHDsec) (Rost, 1996), PROFsec (Rost & C Sander, 1993) and Predictor of Non-Regular secondary structure (NORSp) (Liu and Rost, 2003). The prediction was done on the PredictProtein server (Rost and Liu, 2003): (http://www.embl-heidelberg.de/predictprotein/submit_def.html).

2.8 Identification of Functional Motifs

Searches for known possible functional motifs was done using PROSITE motif searches (Hofmann *et al*., 1999) employing the PredictProtein server interface (Rost and Liu, 2003) (http://www.emblheidelberg.de/predictprotein/submit_def.html). The three phosphorylation sites namely: Cyclic AMP phosphorylation sites, Protein Kinase (PKC) phosphorylation sites, Casein Kinase II (CKII) phosphorylation sites were searched for, tentative phosphorylation site where correlated to data from a neural networks based phosphorylation site predictive algorithm. The NetPhos 2.0 prediction server (Blom *et al*., 1999) http://www.cbs.dtu.dk/services/NetPhos/) assigned an estimate of the probable phosphorylation of the predicted motif. A phosphorylation score greater than 0.9 was taken to be a putative phosphorylation site. Nuclear localisation sites were searched for using PredictNLS (Nair *et al*., 2003).

2.9 Solvent accessibility Prediction

Solvent accessibility of residues was done through the PredictProtein server (http://cubic.bioc.columbia.edu/predictNLS/) (Rost and Liu, 2003) (http://www.embl-eidelberg.de/predictprotein/submit_def.html). Two distinct algorithms were used, Profile fed network systems from Heidelberg (PHDacc) (Rost and Sander, 1994) and the more recent Predictor of Non-Regular secondary structure (NORSp-solvent accessibilty) (Liu and Rost, 2003).

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Assembly of a zinc binuclear cluster protein database

A homology search (23rd November 2003) using the highly conserved zinc binuclear cluster motif sequence (-$CX_2CRX_2KXKCDX_3PX_2CX_2CX_6C$-) (Schjerling and Holmberg, 1996) yielded 118 zinc binuclear cluster proteins from sixteen different fungal genomes (Table 3.1). Most of the 39 new annotations reported in this work were from recently sequenced genomes. There were no new annotations from the *S. cerevisiae* genome (zinc binuclear cluster data-disc). The zinc nuclear data disc attached to the research report as an excel spreadsheet shows all the information used to construct the database.

The zinc binuclear cluster proteins were variable in protein sequence length, ranging from 321 amino acid residues (Cmr1p) up to 1862 residues (Gsabp). A total of 56 proteins were found in the *S. cerevisiae* genome as previously reported (Schjerling and Holmberg, 1996; Todd and Andrianopoulos, 1997).

Table 3.1: Summary of novel zinc binuclear cluster proteins

Fungal species	Estimated genes (Genome sequence published?)	Total number of zinc binuclear proteins	New annotations (Novel)	Protein Names (Novel)
Aspergillus flavus	-	1	-	-
Aspergillus fumigatus	-	1	1	Ca4p
Aspergillus nidulans	-	9	-	-
Aspergillus niger	-	2	1	Xlnrp

Table 3.1 -Continued

Fungal species	**Estimated genes (Genome sequence published?)**	**Total number of zinc binuclear proteins**	**New annotations (Novel)**	**Protein Names (Novel)**
Aspergillus oryzae	-	3	2	Amyr1p OTamp
Candida albicans	-	2	-	-
Fusarium solani	-	2	2	Ct1ap Ct1bp
Hansenula polymorpha	-	1	1	Mut3p
Kluveromyces lactis	-	2	1	Sef1p_K
Lentinula endodes		1	-	-
Margnaporthe grisea	-	2	2	Afa1p Ea2p
Neurospora crassa	10000 (Galagan *et al*., 2003)	8	5	Ac15p Cmr1p Ea1p Flufp Pro1p_N
Pichia pastoris	-	1	1	Gsabp
Saccharomyces cerevisiae	5885 (Goffeau *et al*., 1996)	56	-	-
Schizosaccharomyces pombe	4824 (Wood *et al*., 2002)	26	22	Ca1p Ca2p Ca3p Grt1p Spa1p Spa2p Spa3p Spa4p Spa5p Spb1p Spb2p Spb3p Spb4p Spb5p Spb6p Spc1p Spc2p Spc3p Suc1p Ta1p

Table 3.1 -Continued

Fungal species	**Estimated genes (Genome sequence published?)**	**Total number of zinc binuclear proteins**	**New annotations (Novel)**	**Protein Names (Novel)**
Schizosaccharomyces pombe	4824 (Wood *et al*., 2002)	26	22	Ta2p Yhddp
Sordia brevicollis	-	1	1	Pro1p_S

The *S. pombe* genome was found to have an estimated 4824 genes, representing the smallest eukaryotic genome sequenced to date (Table 3.1). A total of 26 zinc binuclear cluster protein sequences from the *S. pombe* genome were reported in this study, 22 of these were reported novel. The *N. crassa* genome was found to have a total 8 binuclear cluster protein sequences out of an estimated 10000 genes in this study. Zinc binuclear cluster proteins found on the *S. cerevisiae* genome constitute approximately 1 % of the genes found in that genome. Zinc binuclear cluster protein sequences were found to constitute 0.5 % of the estimated genes in the *S. pombe* genome. The *N. crassa* genome had the least reported zinc binuclear cluster protein sequences relative to the genes (0.08 %).

There was no obvious correlation observed between the three published genomes when zinc binuclear cluster protein sequences were compared to the estimated number of genes (Table 3.1). The other fungal species had fewer reported zinc cluster proteins due to their partially sequenced genomes. The on-going sequencing projects on other fungal genomes and continual annotation of published genomes should unravel huge data subsets that can be used for whole genome analysis involving zinc binuclear cluster proteins.

3.2 Zinc binuclear cluster sequence alignment

Full-length alignments of retrieved motifs were done to produce zinc binuclear cluster alignments. The first residue in the zinc binuclear domain was observed to be a

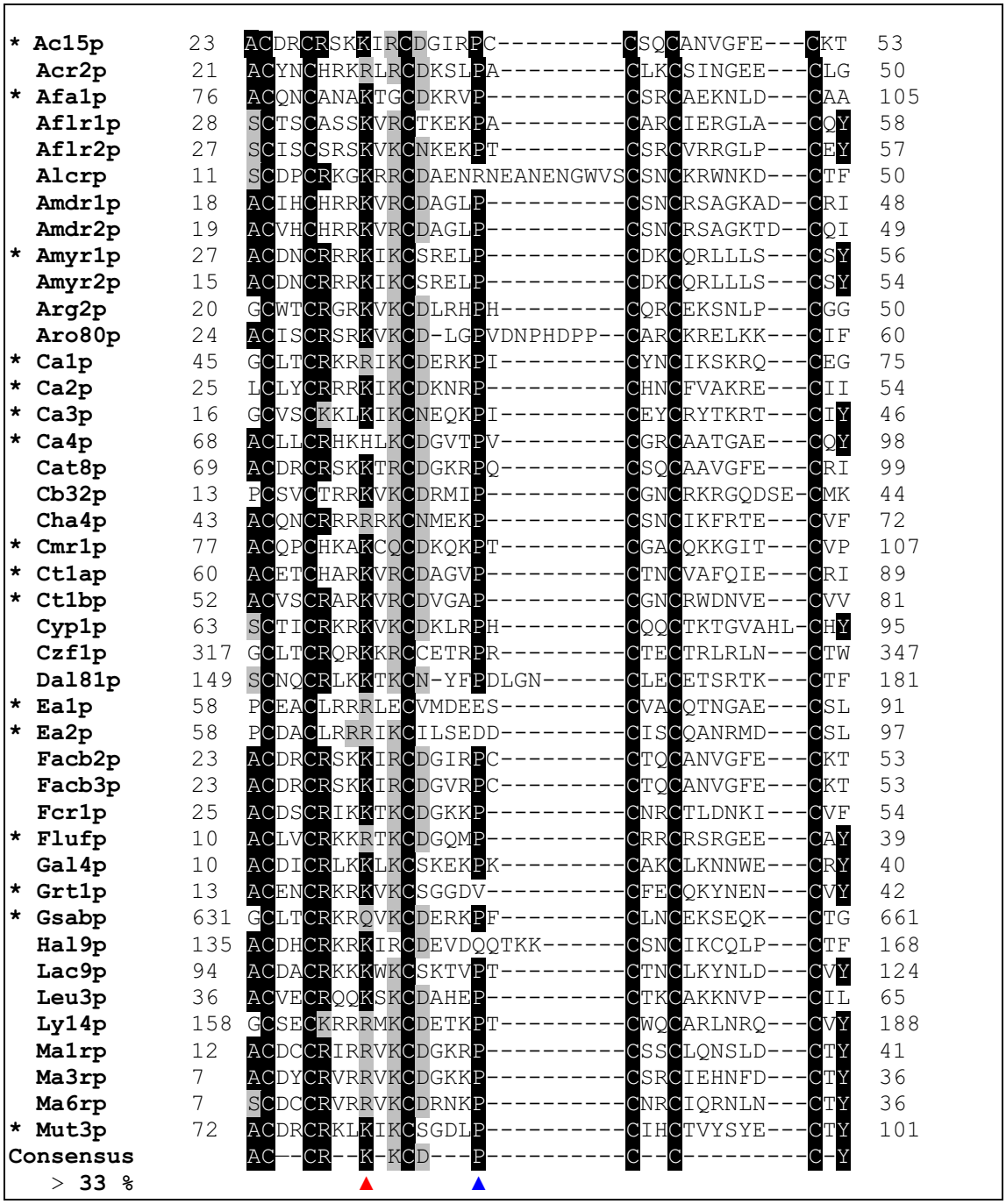

Figure 3. 1: Alignment of the C6 zinc binuclear cluster in fungal proteins of 118 proteins using ClustalW multiple sequence alignment program (Blosum62 matrix) and manual alignment techniques for optimal alignment. The red triangle shows the lysine residue responsible for specific Protein-DNA interaction with the CGG triple as determined by structural studies in Gal4p -Lys 18 (Marmorstein *et al.*, 1992) and Ppr1p-Lys 41 (Marmorstein and Harrison, 1994). The blue triangle shows the conserved proline residue. The consensus sequence is shown above. Black shaded boxes denote identical amino acid residues while grey shaded boxes denote similar residues in the alignment. New annotations are highlighted using the asterisk symbol (*).

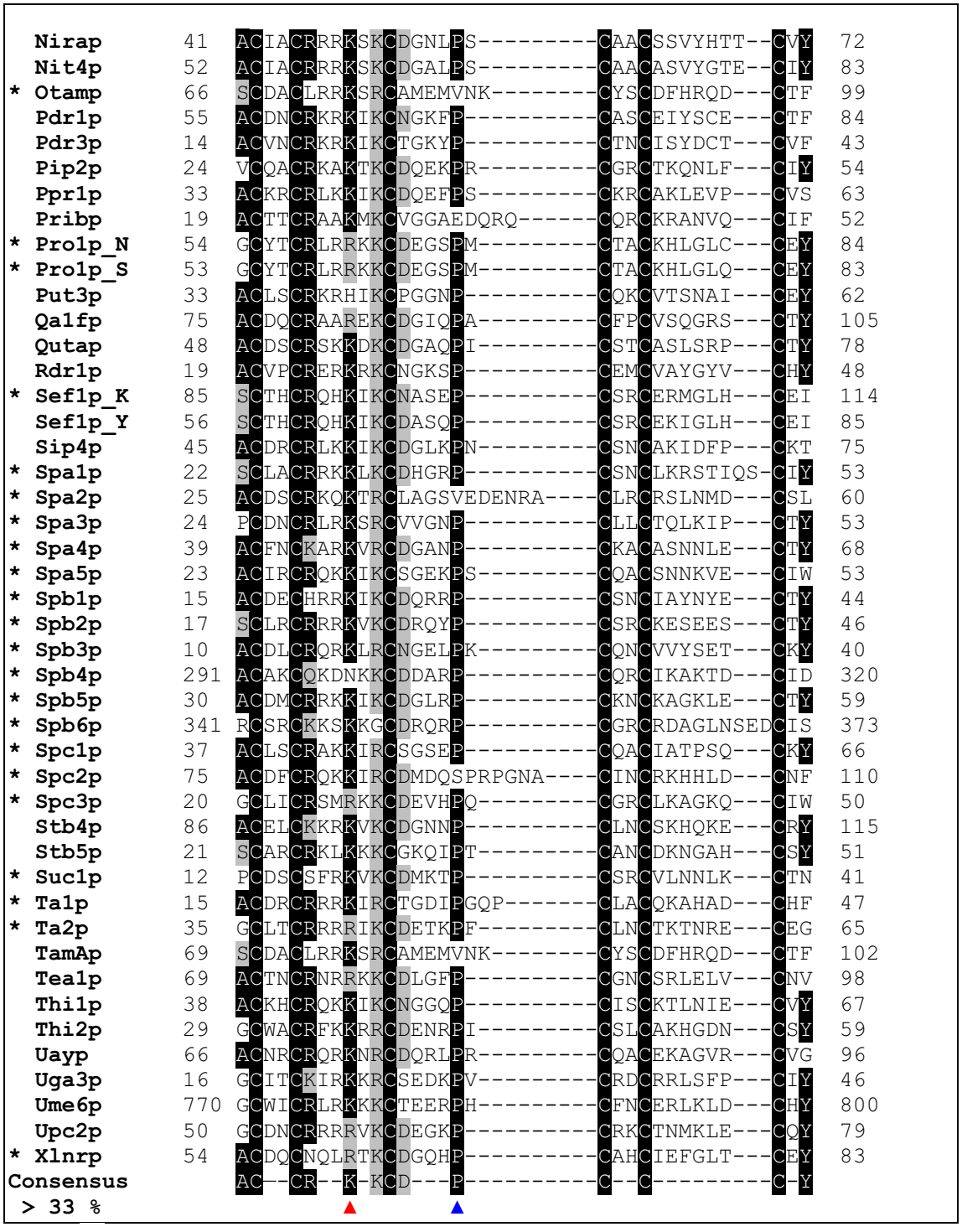

Figure 3. 1 -Continued

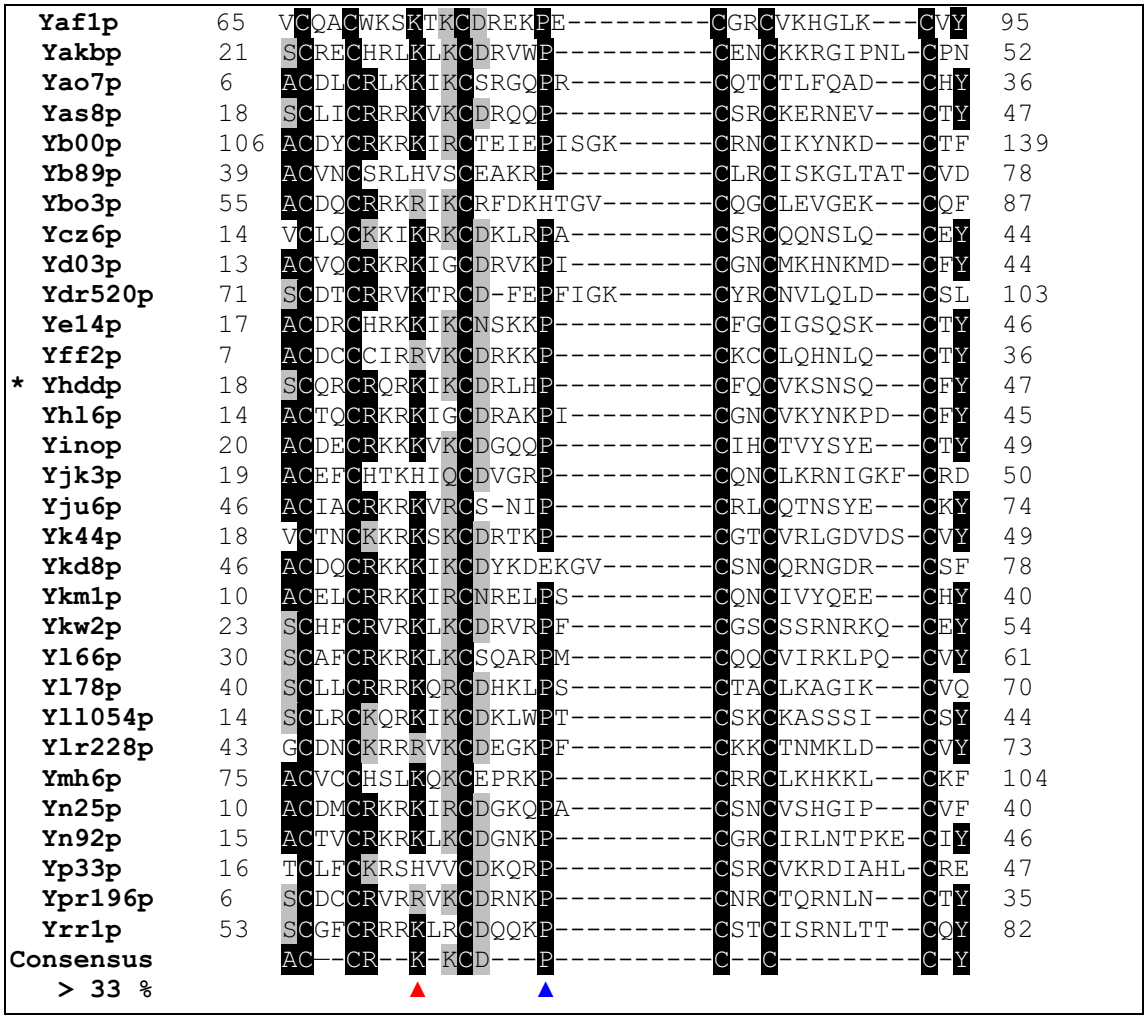

Figure 3. 1 -Continued

threonine, glycine alanine or valine residue (Figure 3.1). Alanine and glycine were found conserved in 58 % and 12 % of the 118 proteins respectively at this position. The lysine residue that has been found to be responsible for specific protein-DNA interaction with the CGG triplet as determined through structural studies on Gal4p -Lys 18 (Marmorstein *et al.*, 1992) and Ppr1p-Lys 41 (Marmorstein and Harrison, 1994) were observed conserved in the alignment. Conservation of a positive amino acid residue at this position in the 118 proteins was observed at 98 %. This supports the observations from structural data that charged or hydrophilic residues residue conserved at this position are important in DNA binding interactions (Marmorstein *et al.*, 1992, Marmorstein and Harrison, 1994). A total of 89 out of 118 (75 %) protein sequences (Figure 3.1) had a lysine residue conserved at that topologically equivalent position. The other 22 (19 %) protein sequences had arginine residues conserved at that site while 5 (5 %) protein sequences were observed to have histidine residues

conserved. The remaining two sequences had asparagine and glutamine conserved respectively at the site. Examination of the new protein sequence annotations revealed that 27 of the 39 new annotations (Figure 3.1) had a lysine residue
conserved at that topologically equivalent position. This represented conservation in 70 % of the annotations. Arginine was conserved in 23 % of the sequences while histidine, glutamine and asparagine residues were both conserved respectively conserved in 2 % of the protein sequences.

The crystal structure of Gal4p revealed that the conserved proline residue (Figure 3.1, Gal4p) was found located in the loop region of two α - helices coordinated by two zinc nuclei (Marmorstein *et al*., 1992; Baleja *et al.,* 1997). Structural analysis of Gal4p has shown that the proline residue was critical in reducing strain in the loop region (Marmorstein *et al*., 1992). The proline residue was conserved in 89 % of the protein sequences at this topologically equivalent position the fourth and fifth cysteine residues (Figures 3.1). Similarly a proline residue was conserved at this position in 82 % of the new annotations. Zinc binuclear cluster proteins lacking the proline residue at this position were observed to have a longer stretch of amino acids. Proline is found conserved at the position between the two α-helices since the rigid conformation adopted by this residue presumably allows the polypeptide chain to change direction. The net structural arrangement allows for the spatial co-ordination complex of cysteine and zinc nuclei to form (Baleja *et al*., 1997). The data supports similar observations on a comparative study of zinc binuclear cluster proteins (Schjerling and Holmberg, 1996). A protein, Aro80p was observed to have 4 proline residues in the zinc binuclear cluster, could this suggest possible novel domain architecture? Comparison of proline conservation at this position at a genomic level reveals that proline has been consistently conserved. The three published genomes: *S. cerevisiae*, *N. crassa*, *S. pombe* were used comparatively. Proline was observed at conserved 95 % of protein sequences in the *S. cerevisiae* genome, 86 % of the annotations in the *N. crassa* genome had a proline residue conserved and the *S. pombe* genome had proline conserved in 89 % of its annotations.

A hydrophobic residue was observed to occupy the last position in the zinc binuclear cluster motif (Figure 3.1) in most of the proteins. The tyrosine residue was found conserved in 61 out of 118 proteins and this represented 52 % residue conservation. The tyrosine residue was found similarly conserved in 46 % of the novel proteins. The zinc binuclear cluster domains of Gal4p and Ppr1p have been elucidated through structural studies (Figure 3.2). The first α-helix in Gal4p was found to be six residues length (Figure 3.2, Asp 12 to Lys17) while the second helix was four residues in length (Figure 3.2, Lys 30 to Lys 33).

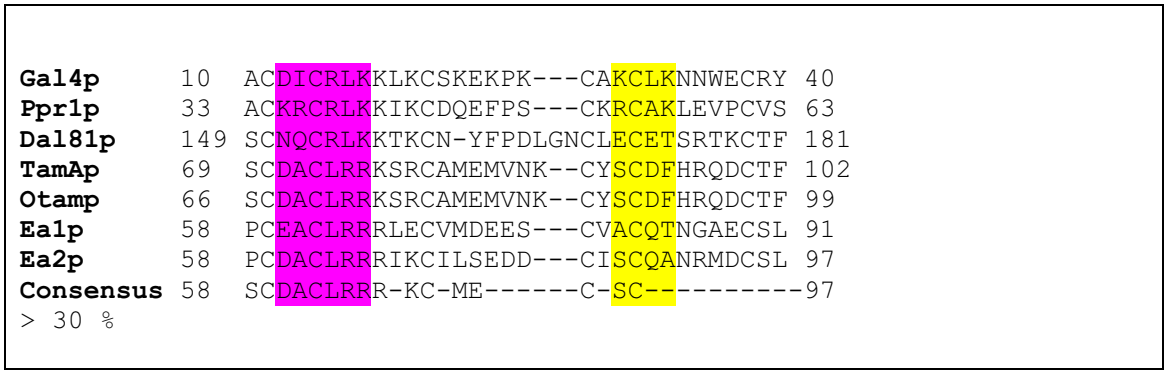

Figure 3.2: Comparative alignment showing the two α- helical regions of the zinc binuclear cluster motif (Gal4p and Ppr1p) as determined through structural data. The purple-boxed region represents first helical region. The second helical is represented the yellow-boxed region (Baleja *et al*., 1997; Marmorstein and Harrison, 1994).

The two α- helical regions in Ppr1p and Gal4p were found to align well. Analysis of the zinc binuclear cluster in Dal81p revealed that in the first α-helix Asn 151 was topologically equivalent to Asp 12 in Gal4p while Lys 156 in Dal81p was topologically equivalent to Lys 17 in Gal4p. Lys 18 in Gal4p has been shown occupy the DNA recognition position (Baleja *et al*., 1997; Marmorstein *et al*., 1992). The data suggests that there was considerable overlap between the first α-helical region of Gal4p and Dal81p.

From a topological perspective, the second α-helical region in Dal81p was found to extend from Glu 171 to Thr 174 and clearly there were insignificant amino acid sequence similarities between the second α-helical regions in both proteins. In Ppr1p

and Gal4p the second helix was characterised by two positively charged lysine residues linked by two hydrophobic residues. The second helix in Dal81p was observed to have two hydrophobic and negatively charged residues respectively. The second helical region in TamAp and Otamp was observed to possess a cluster of hydrophobic residues. Similar amino acid sequence features were observed in two putative Dal81p homologues, Ea1p and Ea2p.

The Gal4p and Ppr1p helical region was characterised by basic residues while Dal81p homologue helical region was characterised by hydrophobic residues. Gal4p and Ppr1p bind to everted DNA repeats (Marmorstein *et al*., 1992; Marmorstein and Harrison, 1994). There is no evidence for direct DNA binding of Dal81p and its homologues (TamAp and Otamp) in the literature. There was no clear pattern observed between the conserved residues in the helical regions of the zinc binuclear cluster and the DNA binding specificity of the proteins. This suggested that conserved residues in the helical regions of the zinc binuclear cluster were involved in the determination of DNA binding specificity.

3.2 Coiled coil analysis of zinc binuclear cluster family

The results of the coiled coil prediction using the COILS algorithm show that 72 of the 118 zinc binuclear proteins (61 %) have putative coiled coils (zinc binuclear cluster data-disc). The remainder 46 zinc binuclear cluster proteins (39 %) had no coiled coils detected. Some zinc binuclear cluster proteins lacking a coiled coil have been thought not to bind to repeat DNA sequences (Schjerling and Holmberg, 1996). The coiled coil motif was found predicted in 17 new annotations (44 %) and undetected in 22 new annotations (56 %).

Examination of the coiled coil data (zinc binuclear cluster data-disc) shows that the COILS algorithm accurately predicts coiled coil domain in Gal4p (Table 3.2) that was found to extend from residues 50-63 (Marmorstein *et al.*, 1992). A coiled coil domain as determined through structural studies in Cyp1p (Hap1p), residues 114-128 was also accurately predicted (King *et al*., 1999). The weakness of the COILS method (section 1.5.1) was authenticated by the false prediction of six coiled coils that were

characterised by hydrophilic residues lacking a defined heptad repeat (zinc binuclear cluster data-disc).

Predicted coiled coils in Cy1p (residues 175-188) and Dal81p (residues 86-99 and 227-240) lie on polyglutamine stretches and lack a defined heptad repeat. Dal81p was found to lack a proper coiled coil suggesting that the protein does not dimerise. TamAp, Otamp and the two putative homologues (Ea1p and Ea2p) did not have any predicted coiled coils. The data shows that these proteins lack a linker region that connects the coiled coil region to the zinc binuclear cluster domain. An interaction between the middle homology region and the linker region has been thought to reduce the affinity of binding to repeat DNA sequences (Schjerling and Holmberg, 1996). The observations support the hypothesis that states that Dal81p does not directly bind to DNA sequences. The existence of conserved novel domains in Dal81p and its homologues cannot be ruled out.

A total of 26 zinc binuclear cluster proteins had multiple coiled coils predicted. Interestingly the location of the coiled coils in the zinc binuclear cluster family was found to be variable; generally most predicted multiple coiled coils were found located towards the C-terminal of the zinc binuclear cluster motif. A novel zinc binuclear cluster protein, Gsabp was observed to possess a centrally located zinc finger with two predicted coiled coils on either side of the zinc finger, (residues 352–366 and 1788–1801 respectively). Yel14p had a predicted coiled coil region encompassing the zinc finger motif (Table 3.2). There is no evidence in literature of coiled coil regions located in the zinc binuclear cluster motif. There were no coiled coil regions detected in Prop1p_N and Prop1p_S (Table 3.2). Both proteins have been observed to exhibit high sequence similarity in their zinc binuclear cluster domains (section 3.2). The zinc binuclear cluster was located in similar position in both proteins (residues 55-82 and 54-81 respectively) and the proteins were of similar length. Similar observations were noted when two Dal81p (TamAp and Otamp) homologues were compared. There were no coiled coils detected and the location of zinc binuclear cluster differed by a few residues (residues 71-100 and 68-97 respectively). The data shows that the coiled coil motif was conserved amongst zinc binuclear cluster proteins involved in similar metabolic pathways such as maltose fermentation regulators in *S.*

Table 3.2: Summary of the sequence data and predicted coiled coils in zinc binuclear cluster proteins

Name	Sequence length	Zinc Finger (Position)	Coiled coil prediction (14)	Function	Organism
Aflr1p	437	29-56	-	Afflatoxin biosynthesis regulatory protein	*A. flavus*
*Ca4p	625	69-96	-	Hypothetical protein	*A. fumigatus*
Aflr2p	433	28-55	-	Sterigmatocystin biosynthesis regulatory protein	*A. nidulans*
Alcrp	821	12-49	-	Regulatory protein	*A. nidulans*
Amdr2p	765	20-50	587-600	Acetamidase regulatory protein	*A. nidulans*
Amyr2p	662	16-42	-	Amylase regulator	*A. nidulans*
Facb2p	867	24-51	67-96	Acetate DNA binding protein	*A. nidulans*
Nirap	892	42-70	471-484	Nitrogen assimilation transcription factor	*A. nidulans*
Qutap	825	49-76	148-161	Quinic acid utilization activator	*A. nidulans*
TamAp	739	70-100	-	Nitrogen assimilation transcription factor	*A. nidulans*
Uayp	1060	67-94	-	Positive regulator of purine utilization	*A. nidulans*
Facb3p	862	24-51	67-96	Acetate DNA binding protein	*A. niger*
*Xlnrp	875	55-81	638-651	Transcriptional activator xlnR	*A. niger*
Amdr1p	735	19-49	155-175 424-441 581-594	Acetamidase regulatory protein	*A. oryzae*
*Amyr1p	604	28-54	241-255	Amylase regulator	*A. oryzae*
*Otamp	711	67-97	-	Nitrogen assimilation transcription factor	*A. oryzae*
Czf1p	388	318-345	166-179	Zinc finger protein	*C. albicans*
Fcr1p	517	26-52	261-276	Fluconazole resistance protein 1	*C. albicans*
*Ct1ap	909	61-90	641-658	Cutinase transcription factor	*F. solani*
*Ct1bp	882	53-81	-	Cutinase transcription factor	*F. solani*
*Mut3p	929	73-99	-	Peroxisome proliferation regulator	*H. polymorpha*
Lac9p	865	95-122	135-155	Lactose metabolism regulatory protein	*K. lactis*
*Sef1p_K	1071	86-116	265-278 553-574	Supressor protein	*K. lactis*

Table 3.2 –Continued

Name	Sequence length	Zinc Finger (Position)	Coiled coil prediction (14)	Function	Organism
Pribp	565	20-50	63-85	Prib protein	*L. edodes*
*Afa1p	974	77-103	-	Putative transcription factor Pig1p	*M. grisea*
*Ea2p	805	59-95	-	Hypothetical protein	*M. grisea*
*Ac15p	865	24-51	67-98	Regulatory protein	*N. crassa*
Acr2p	595	22-49	-	Acriflavine sensitivity control protein	*N. crassa*
*Cmr1p	321	78-105	-	Regulatory protein	*N. crassa*
*Ea1p	775	59-89	-	Hypothetical protein	*N. crassa*
*Flufp	792	11-37	45-63	Conidial development protein fluffy	*N. crassa*
Nit4p	1090	53-81	96-110 324-338 491-511	Nitrogen assimilation transcription factor	*N. crassa*
*Pro1p_N	696	55-82	-	PRO1 protein	*N. crassa*
Qa1fp	816	76-103	-	Quinic acid utilization activator	*N. crassa*
*Gsabp	1862	632-659	352-366 1788-1801	Pexophagy regulatory protein	*P. pastoris*
Arg2p	880	21-48	-	Arginine metabolism regulation	*S. cerevisiae*
Aro80p	950	25-58	96-117 181-194 767-780	Putative transcription regulator	*S. cerevisiae*
Cat8p	1433	70-97	112-135 623- 641 684-698	Regulatory protein	*S. cerevisiae*
Cb32bp	608	14-42	541-554	Regulatory protein	*S. cerevisiae*
Cha4p	648	44-70	87-106	Activatory protein	*S. cerevisiae*
Cyp1p	1502	64-93	110-130 175-188	Activatory protein	*S. cerevisiae*
Dal81p	970	150-179	86-99 227-240	Transcriptional activator protein	*S. cerevisiae*
Gal4p	881	11-38	51-69	Regulatory protein	*S. cerevisiae*

Table 3.2 –Continued

Name	Sequence length	Zinc Finger (Position)	Coiled coil prediction (14)	Function	Organism
Hal9p	1030	136-166	163-182	Putative transcription regulator	*S. cerevisiae*
Leu3p	886	37-67	70-99 444-466 685-701	Regulatory protein	*S. cerevisiae*
Ly14p	790	159-186	-	Lysine biosynthesis regulator	*S. cerevisiae*
Ma1rp	473	13-39	135-149	Maltose fermentation regulator	*S. cerevisiae*
Ma3rp	468	8-34	128-141	Maltose fermentation regulator	*S. cerevisiae*
Ma6rp	473	8-34	-	Maltose fermentation regulator	*S. cerevisiae*
Pdr1p	1060	56-72	166-181	Pleiotropic drug resistance regulator	*S. cerevisiae*
Pdr3p	976	15-41	-	Pleiotropic drug resistance regulator	*S. cerevisiae*
Pip2p	996	25-52	66-79 560-573	Peroxisome proliferation regulator	*S. cerevisiae*
Ppr1p	904	34-61	80-93 197-210 479-492 699-714	Pyrimidine pathway regulatory protein	*S. cerevisiae*
Put3p	979	34-60	76-97	Proline utilization trans-activator	*S. cerevisiae*
Rdr1p	546	20-46	-	Putative transcription regulator	*S. cerevisiae*
Sef1p_Y	1057	57-87	97-118 858-871	Supressor protein	*S. cerevisiae*
Sip4p	829	46-73	745-768	SIP4 protein	*S. cerevisiae*
Stb4p	949	87-113	-	Putative transcription regulator	*S. cerevisiae*
Stb5p	743	22-49	554-572	Putative transcription regulator	*S. cerevisiae*
Tea1p	759	70-96	113-134	Enhancer activator	*S. cerevisiae*
Thi2p	450	30-57	166-182	Thiamine biosynthesis protein	*S. cerevisiae*
Uga3p	528	17-44	-	Transcriptional regulatory protein	*S. cerevisiae*
Ume6p	836	771-798	815-832	Transcriptional regulatory protein	*S. cerevisiae*

Table 3.2 –Continued

Name	Sequence length	Zinc Finger (Position)	Coiled coil prediction (14)	Function	Organism
Upc2p	913	51-78	207-225	Putative transcription regulator	*S. cerevisiae*
Yaf1p	1062	66-93	271-284 321-334 599-612	Transcriptional regulatory protein	*S. cerevisiae*
Yb00p	1094	107-137	142-158 658-671 1050-1063	Transcriptional regulatory protein	*S. cerevisiae*
Yb89p	529	40-68	-	Transcriptional regulatory protein	*S. cerevisiae*
Ybo3p	919	56-85	-	Transcriptional regulatory protein	*S. cerevisiae*
Ycz6p	832	15-42	16-29 264-281 664-678	Transcriptional regulatory protein	*S. cerevisiae*
Yd03p	885	14-42	96-117 181-194 767-780	Transcriptional regulatory protein	*S. cerevisiae*
Ydr520p	772	72-101	169-182 236-254 637-656	Putative transcription regulator	*S. cerevisiae*
Ye14p	794	18-44	9-25 126-139 194-207	Transcriptional regulatory protein	*S. cerevisiae*
Yff2p	465	8-34	-	Transcriptional regulatory protein	*S. cerevisiae*
Yhl6p	883	15-43	207-225	Transcriptional regulatory protein	*S. cerevisiae*
Yinop	964	21-47	126-139	Transcriptional regulatory protein	*S. cerevisiae*
Yjk3p	618	20-48	461-474	Transcriptional regulatory protein	*S. cerevisiae*
Yju6p	758	47-73	-	Putative transcription regulator	*S. cerevisiae*
Yk44p	863	19-47	492-505	Transcriptional regulatory protein	*S. cerevisiae*
Ykd8p	1170	47-76	-	Transcriptional regulatory protein	*S. cerevisiae*
Ykm1p	594	11-38	-	Putative transcription regulator	*S. cerevisiae*

Continued

Name	Sequence length	Zinc Finger (Position)	Coiled coil prediction (14)	Function	Organism
Ykw2p	705	24-52	53-66 75-93	Transcriptional regulatory protein	*S. cerevisiae*
Yl66p	701	31-59	82- 96	Putative transcription regulator	*S. cerevisiae*
Yl78p	1341	41-68	5-26 170-184	Putative transcription regulator	*S. cerevisiae*
Yll054p	769	15-42	103-126	Putative transcription regulator	*S. cerevisiae*
Ylr228p	814	44-71	109-122 331-352	Putative transcription regulator	*S. cerevisiae*
Ymh6p	944	76-109	606-619 780-800	Transcriptional regulatory protein	*S. cerevisiae*
Yn25p	743	11-38	56-69 246-259	Putative transcription regulator	*S. cerevisiae*
Yn92p	607	16-38	429-445	Putative transcription regulator	*S. cerevisiae*
Yp33p	446	17-45	-	Putative transcription regulator	*S. cerevisiae*
Ypr196p	470	7-34	-	Maltose fermentation regulator	*S. cerevisiae*
Yrr1p	810	54-82	150-163 255-268	Putative transcription regulator	*S. cerevisiae*
*Ca1p	419	46-73	-	Transcriptional regulation	*S. pombe*
*Ca2p	632	26-52	67-86	Transcriptional regulation	*S. pombe*
*Ca3p	510	17-42	-	Transcriptional regulation	*S. pombe*
*Grt1p	648	14-40	-	Zinc finger protein	*S. pombe*
*Spa1p	625	23-51	75-91	Probable transcriptional regulator	*S. pombe*
*Spa2p	529	26-58	81-100	Probable transcriptional regulator	*S. pombe*
*Spa3p	654	25-54	-	Probable transcriptional regulator	*S. pombe*
*Spa4p	783	40-66	489-505	Probable transcriptional regulator	*S. pombe*
*Spa5p	697	24-51	-	Probable transcriptional regulator	*S. pombe*

Table 3.2 –Continued

Fungal Zinc	Sequence	Zinc Finger	Coiled coil prediction (14)		Organism

Finger	**length**	**(Position)**		**Function**	
*Spb1p	827	16-42	-	Probable transcriptional regulator	*S. pombe*
*Spb2p	560	18-44	-	Probable transcriptional regulator	*S. pombe*
*Spb3p	594	11-38	95-108	Probable transcriptional regulator	*S. pombe*
*Spb4p	738	292-318	-	Probable transcriptional regulator	*S. pombe*
*Spb5p	815	31-57	107-121	Probable transcriptional regulator	*S. pombe*
*Spb6p	397	342-371	-	Probable transcriptional regulator	*S. pombe*
*Spc1p	857	38-64	12-26	Putative transcription regulator	*S. pombe*
*Spc2p	867	76-108	686-700	Putative transcription regulator	*S. pombe*
*Spc3p	525	21-48	-	Putative transcription regulator	*S. pombe*
*Suc1p	501	13-39	54-67 437-450	Probable sucrose utilization protein	*S. pombe*
*Ta1p	480	16-46	-	Hypothetical protein	*S. pombe*
*Ta2p	497	36-63	-	T41718 hypothetical fungal Zn(2)-Cys(6)	*S. pombe*
Thi1p	775	39- 65	9-25 126-139 194-207	Thiamine biosynthesis protein	*S. pombe*
Yakbp	782	22-48	452-470	Putative transcription regulator	*S. pombe*
Yao7p	603	7-34	464-479	Putative transcription regulator	*S. pombe*
Yas8p	563	19-45	-	Putative transcription regulator	*S. pombe*
*Yhddp	618	19-45	234-250 419-432	Putative transcription regulator	*S. pombe*
*Pro1p_S	693	54-81	-	Transcriptional regulatory protein	*S. brevicollis*

Name of protein, prefix 'p' to show protein. Protein sequence length shown, zinc finger position in protein denoted by (1st cysteine and last cysteine residues positions) in the protein sequence. Coiled coil prediction using the COILS server (Lupas *et al.*, 1991) with specified window of 14 residues for the optimal detection of coiled coils (Schjerling and Holmberg, 1996). Coiled coils not detected by the COILS algorithm are denoted (-). New annotations are highlighted using the asterisk symbol (*).

cerevisiae (Mar1p and Mar3p). The data is further supported by similar locations of the coiled coil and zinc binuclear cluster when Mar1p and Mar3p (Table 3.2) are compared. Mar6p zinc binuclear cluster domain (amino acid residues 8-34) had an identical location to the Mar3p but lacked a coiled coil. These homologues are functional paralogues derived from gene duplications that could have led to loss of the coiled coil.

3.3 Dal81p homologues

The PSI-BLAST search was carried out using the protein sequence of Dal81p (*S. cerevisiae*) (SWISS PROT accession number: P21657) against public databases. The PSI - Blast search for potential Dal81p homologues results yielded four significant hits present in the zinc binuclear cluster protein database. Otamp (*A. oryzae*) a known Dal81p homologue (Small *et al.*, 2001) had a calculated 30 % sequence identity to Dal81p (10 % gaps) with an expectation value of 2×10^{-67}. Another Dal81p homologue, TamAp (*A. nidulans*) (Small *et al.*, 2001) exhibited 28 % sequence identity (11 % gaps). Two hypothetical proteins were also identified, Ea1p (*N. crassa*) and Ea2p (*M. grisea*). Ea1p is a 775-residue long zinc binuclear cluster protein displaying 31 % sequence identity (16 % gaps) with an expectation value of 7×10^{-79} Ea2p has 805 residues with a 32 % sequence identity (18 % gaps) and an expectation value of 1×10^{-88}. The two novel proteins (Eap and Ea2p) display higher sequence identity to Dal81p when compared to the known homologues (Otamp and TamAp). The proteins were designated as putative homologues since the search result does not take evolutionary events such as gene duplication.

3.3.1 Full length Alignments: Dal81p and homologues

Full-length alignments using ClustalW of Dal81p and homologues; TamAp (*A. nidulans*), Otamp (*A. oryzae*), *Ea1p* (*N.crassa*) and Ea2p (*M. grisea*) were done. Four distinct regions with conserved blocks of amino acid residues characterised by high sequence identity and similarity amongst the homologues were defined. The regions have been designated zinc binuclear cluster domain (Figure 3.3, residues 150-180), domain one (residues 243-411), the middle homology region (residues 425-496) and domain two (residues 544-723) for the purposes of this study. Extensive functional deductions from the alignment of such a subset of proteins that have been obtained

through Psi-Blast were valid since the method accurately detected two characterised homologues. This was further supported by the existence of the

zinc binuclear cluster domain and the high sequence identity amongst these proteins. Each region was analysed and used to build a profile for further Dal81p structural analysis.

There was insignificant amino acid similarity observed in the region between the zinc binuclear cluster and domain one. This could be explained by the absence of true coiled coils and linker regions in all five homologues (section 3.2) since conserved residues imply some degree of structural conservation important for function. The evolution process of Dal81p and its homologues could have involved the loss of the coiled coils, the linker region coupled with the loss of function zinc binuclear cluster domain. The two novel domains identified amongst the homologues could be important for function.

3.3.2 Domain one

Domain one was found to encompass the region Dal81p (Figure 3.3, residues 243-411). The domain was observed to have the least pronounced amino acid sequence identity and similarity amongst the homologues. Dal81p exhibited an amino acid sequence identity similarity of 23 % and 44 % respectively to the other homologues. The alignment reveals that TamAp (36 % and 54 % sequence identity and similarity) was observed to be more similar to the putative homologues Ea1p (33 % and 50 % sequence identity and similarity respectively) and Ea2p (31 % and 52 % sequence identity and similarity respectively). Close examination at the consensus sequence residues revealed pattern of hydrophobic periodicity. There were hydrophobic residues approximately every four residues interspaced by hydrophilic residues. The pattern was indicative of amphipathic helical arrangement. There were proline residues observed at regular intervals presumably to alter polypeptide direction. The domain was also separated from the middle homology domain by approximately 12 amino acid residues.

Examination of the hydrophobicity plots of the domain one (Figure 3.4) revealed strikingly similar profiles. The hydrophobicity profiles correlated to the observed sequence similarity amongst the homologues and this suggested that this protein domain could play an important role in the function of these proteins.

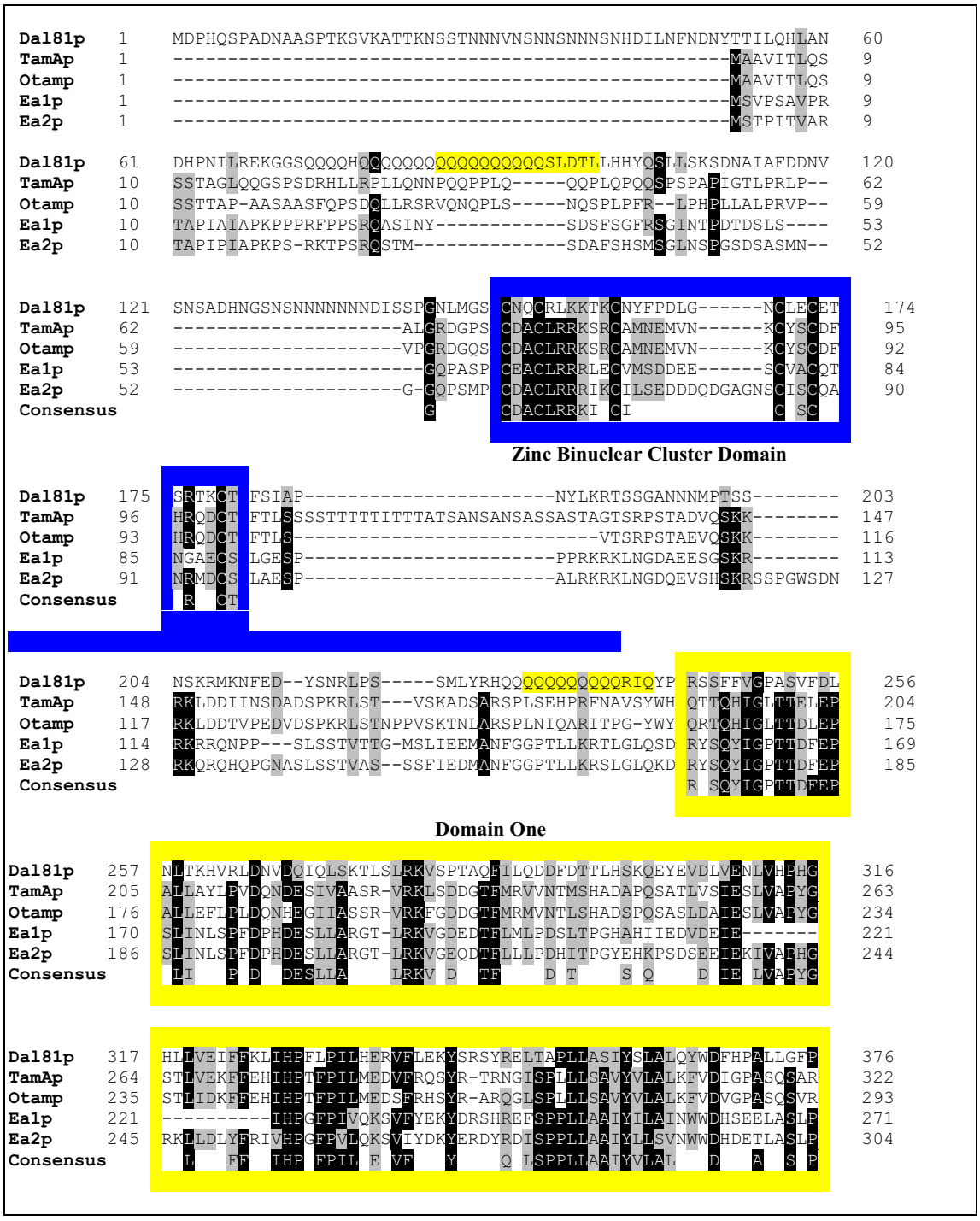

Figure 3.3: Full length alignment showing Dal81p, homologues (TamAp and Otamp) and putative homologues (Ea1p and Ea2p). Putative homologues were identified using Psi-Blast (Altschul *et. al*., 1997) at NCBI. The Blue-boxed region shows the zinc binuclear cluster domain (Dal81p, residues 149-181). The Orange–boxed region shows domain one (Dal81p, residues 243-411). Purple-boxed region shows the middle homology region (Dal81p, residues 425-496). The Lime green-boxed region represents domain two (Dal81p, amino acid residues 544-723). Predicted coiled coils regions shown by yellow-boxed residues. Identical amino acid residues are shown in black boxes while similar residues are denoted in grey boxes. The consensus sequence is shown at > 60 %.

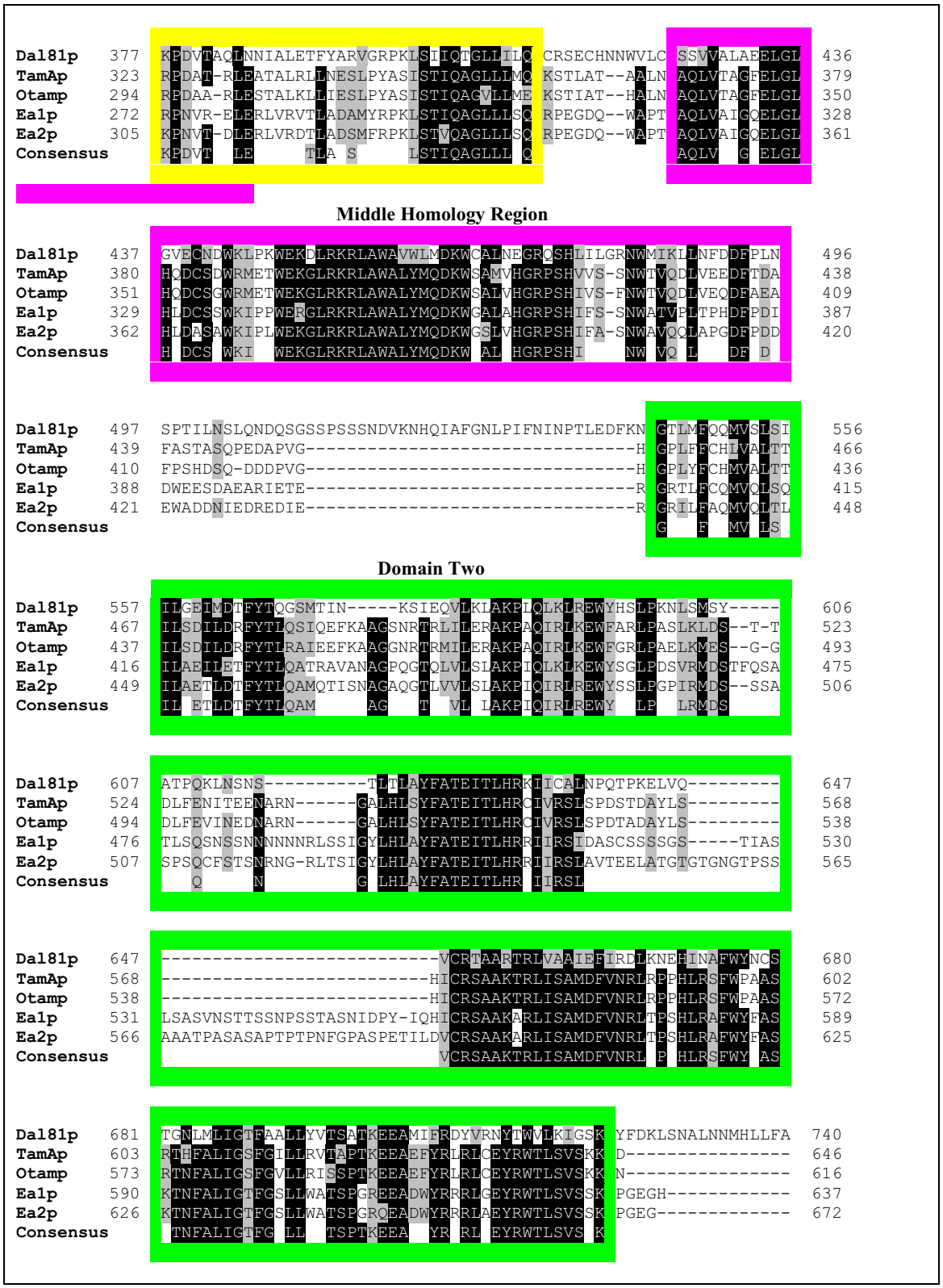

Figure 3.3- Continued

```
Dal81p    741  QIPGLLTDEPVVVSPNSNINSVNPQRSGVQSQIPIQFNVGSPAMTEQGSPLNQWKNLPQE   800
TamAp     646  --AEFLEFALESLDNATDLDHHVPAK--------------PGIDELMTSSSK------P   683
Otamp     616  --AEFMEFALDSLDNANTLDQHVPEK--------------PGIDELMTSAAK------P   653
Ea1p      637  --KGLTEFAMGMLDISTGLLKQLPEK--------------PLLSRSGSAVNVGVGVNAE   680
Ea2p      672  --RGLTEFAMGMLDISTGLLNKLPEK--------------PLVSRSGSAVDF------E   709
Consensus           EFA   LD  T  L    PEK              P

Dal81p    801  ILQQLNSFPNGTTSTTTPVNPTSRQTQLESQGSPAINSANNNSNNTPLPFAPNKSSKKTS   860
TamAp     684  YIAST--------------------S-------------------------ARSGTTQE   697
Otamp     654  TVTQP--------------------R-------------------------PGTAAQLE   667
Ea1p      681  VMRSQSLLALGTGTGS-AQRGGYGVG----------------------SPASSGFGRM   715
Ea2p      710  GMRRQ--LAMETGVSARAPPGSTGVAGS-------------------LPAGRGLGGL   745

Dal81p    861  QSSPNVTPSHMSRHPPSNTSSPRVNSSTNVNSNTQMNASPLTSINETRQESGDAADEKTA   920
TamAp     698  EAILDLDPR--SGTGGTSSVISGLASPAT-------------------SVSEESMHDAAV   736
Otamp     668  NTILAMDQGNDSGRGGTSSVISGLASPAT-------------------SVSEDSFHDTAI   708
Ea1p      716  GSMSGFNESYVRGGPDRRYQQPARGDASG--------------VQSPRSISSDSSDEGGY   761
Ea2p      746  GGGQSFSFSNLQSAYGSGVLSPRSHSGDGG-----------VGEDLDGDEEGDSSDDASG   794

Dal81p    921  GRERTANEESSTELKDDNPNSNQETSATGNQTIKMNDDKNVTINTRETPL 970
TamAp     737  APM------------------------------------------------ 739
Otamp     709  PPL------------------------------------------------ 711
Ea1p      762  GNFSVTAGMAGLAD------------------------------------- 775
Ea2p      795  DEFMYGGIPAS---------------------------------------- 805
```

Figure 3.3 –Continued

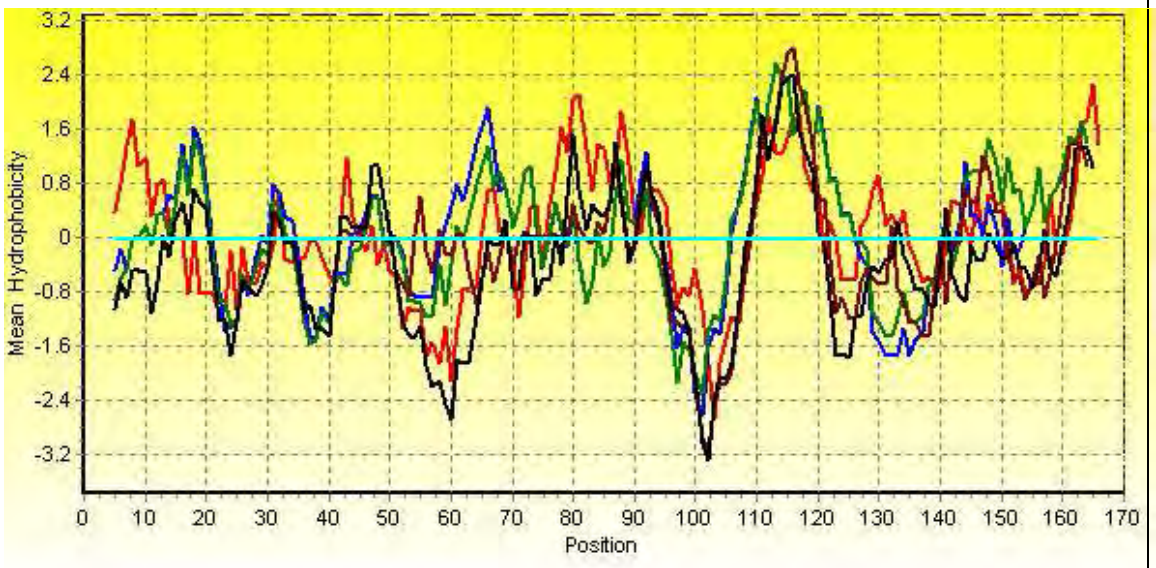

Figure 3.4: Kyte-Doolittle hydrophobicity plot of domain one. The Dal81p hydrophobicity plot shown by the red line, TamAp shown by blue line and Otamp shown by green line. The Ea1p and Ea2p hydrophobicity plots are shown by the brown and black lines respectively.

3.3.3 Middle homology Region

The middle homology region in this study (Dal81p, residues 425-496) was observed to overlap the region previously designated as the middle homology region in Dal81p

(residues 405-477). The alignment revealed highly conserved sequence features in all five homologues. Analysis of the consensus sequence shows that all five proteins share at least 60 % amino acid sequence similarity (Figure 3.3). Dal81p was found to be 44 % identical and 60 % similar to the other four homologues. The middle homology region in TamAp was (63 % and 71 % sequence identity and similarity) was also more similar to the putative homologues Ea1p (65 % and 74 % sequence identity and similarity respectively) and Ea2p (64 % and 72 % sequence identity and similarity respectively). The alignment revealed a cluster of positive amino acids -RKR- , Dal81p (Figure 3.3, residues 453-455) conserved amongst five homologues. This more prominent amino acid sequence similarity in the middle homology domain in comparison to the other domains suggested that the domain could be critical for function amongst the homologues. The middle homology domain overlapped well with the region on TamAp (section 1.3) known to be indispensable for function.

3.3.4 Domain two

Domain two in Dal81p was observed to be 38 % identical and 54 % similar to the other four homologues. TamAp (65 % and 74 % sequence identity and similarity) was found to be more similar to the putative homologues Ea1p (54 % and 66 % sequence identity and similarity) and Ea2p (53 % and 63 % sequence identity and similarity). The -YFATEITLHR- motif was observed to have a 100 % sequence identity amongst all the homologues. There were other motifs in this domain observed to 100 % identical in all the other homologues with the exception of Dal81p. The divergence of the Dal81p sequence from the other homologues could be explained by evolution. The periodicity of hydrophobic residues after approximately every 4 residues was also observed in some amino acid blocks. This suggested the presence of helical secondary structure. There was low sequence similarity observed in the extreme C-terminal portion of the homologue alignment. There was an interesting 20 amino acid consensus sequence found outside domain two (Figure 3.2, TamAp, residues 650-669). The observed consensus, -[EFA]-X (3)-[LD]-X(2)-T-X-L-X(4)-[PEK]- aligned well in four homologue sequences excluding Dal81p.

3.4 Secondary structure prediction

Domains designated were speculated to have conserved secondary structure with a functional role amongst the five homologues. The different domains were used as templates to refine and further parse data generated from the secondary structure prediction of Dal81p. There was need to answer the question of whether amino acid motifs in conserved homologues were indicative of similar secondary structure since a correlation would suggest conserved function amongst the homologues.

The polyglutamine stretch upstream of domain one (Figure 3.5, residues 78-94) was predicted to be α-helical in nature. There is no evidence for the existence of polyglutamine α-helices in literature. The region also encompassed a coiled-coil region was also predicted to lie on the polyglutamine stretch (residues 86-99) (section 1.3). The sequence was not indicative of a proper heptad repeat and can be deduced to lack the periodicity of a α-helix. Deletion studies on the polyglutamine stretch showed that the region was a critical region for Dal81p function (section 1.3). Domain one had six α -helices separated by loop regions (Figure 3.4, grey boxed area). The helices corresponded with the observed pattern of hydrophobic residue periodicity in the domain that was indicative of α -helical secondary structure (section 3.3.1). Four predicted α- helices on Dal81p (Figure 3.5, residues 316-326, 338-342, 352-363, and 400-411 respectively) aligned well with conserved amino acid residue blocks in the homologues suggesting that the secondary structure was conserved amongst the homologues.

The middle homology domain was characterised by a significant amino acid similarity amongst the homologues (section 3.2). The domain overlaps motifs IV and V as described in the eight-motif domain (Poch, 1997). The yellow shaded area (Figure 3.5) was indicative of amphipathic α- helix periodicity. The 13 residue long motif displays an alternating pattern of polar and non-polar amino acid residues. The primary sequence was conserved similarly in the homologues suggesting that the amphipathic α-helix was a potential conserved functionalsite. An α–β–α like motif (Figure 3.4, residues 448-476) was observed to have significant amino acid sequence similarity

Figure 3.4: **Predicted secondary structure of Dal81p.** Secondary structure using PHD prediction (PhD) (Rost, 1996), Predictor of non-regular secondary structure method (Nsec) (Liu and Rost, 2003) and the prediction of secondary structure using PROFsec (PROF) (Rost and Sander, 1993). Consensus of the predicted secondary structure pattern shown as Con. Helix shown as (H / h), strand (E, e), and Loop (L /l). Black boxed region, residues 73- 94 were subject to deletion studies (Bricmont *et al.*, 1991). Brown boxed region with grey background shows predicted zinc binuclear cluster domain. Grey-boxed region represents domain one. Green boxed region represents middle homology region (MHR). Yellow background in middle homology region shows a putative amphipathic α-helix. Green shaded background in the MHR shows the structure of a highly conserved amino acid residue block amongst the five Dal81p homologues. The purple-boxed region represents domain two (residues 544-723). Yellow background in domain two shows typical α-helical structure. A highly conserved motif in domain two is shown in purple. Black boxed region with brown background shows the acidic domain residues 909-936 (Schjerling and Holmberg, 1996).

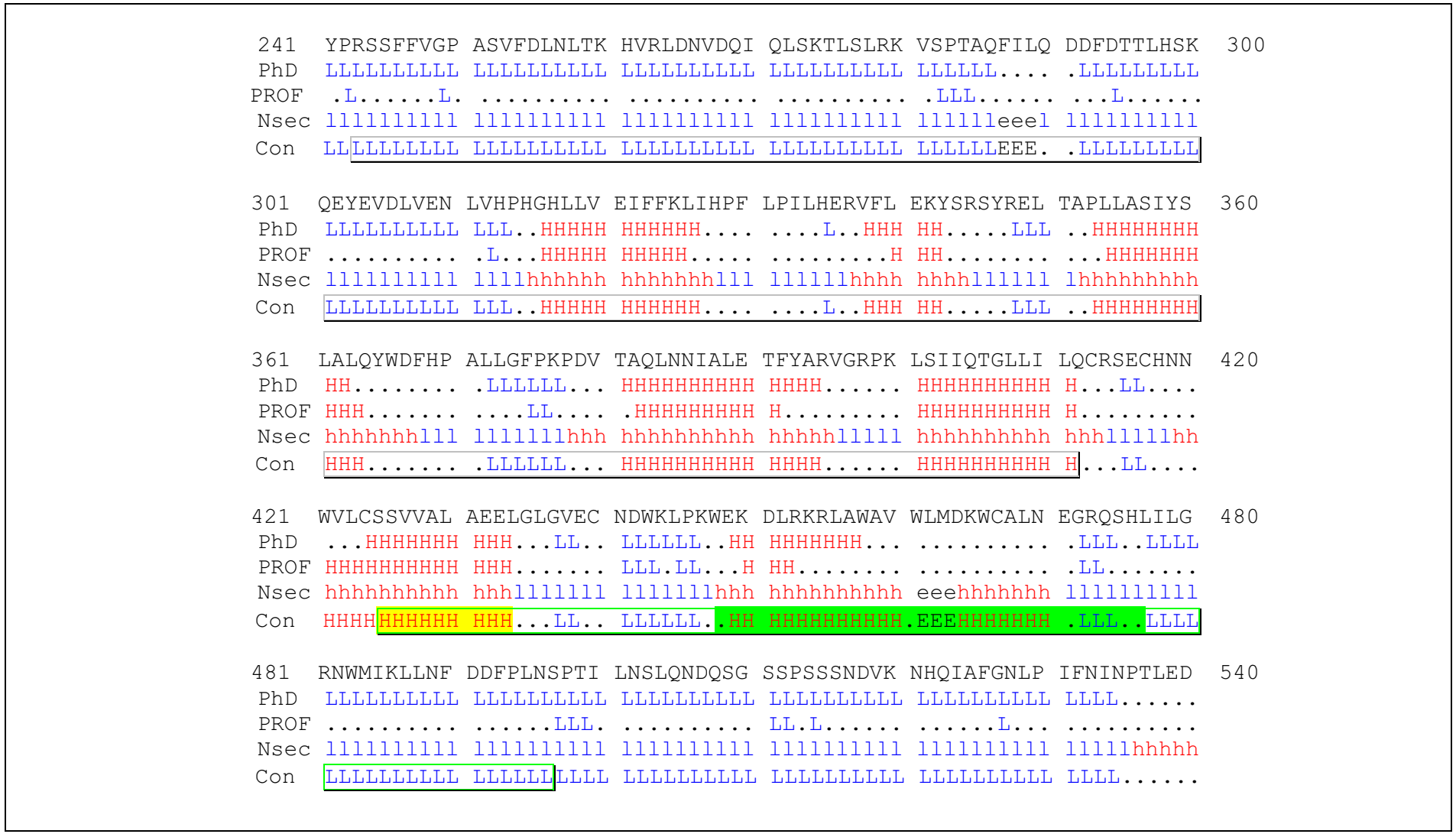

Figure 3.4 -Continued

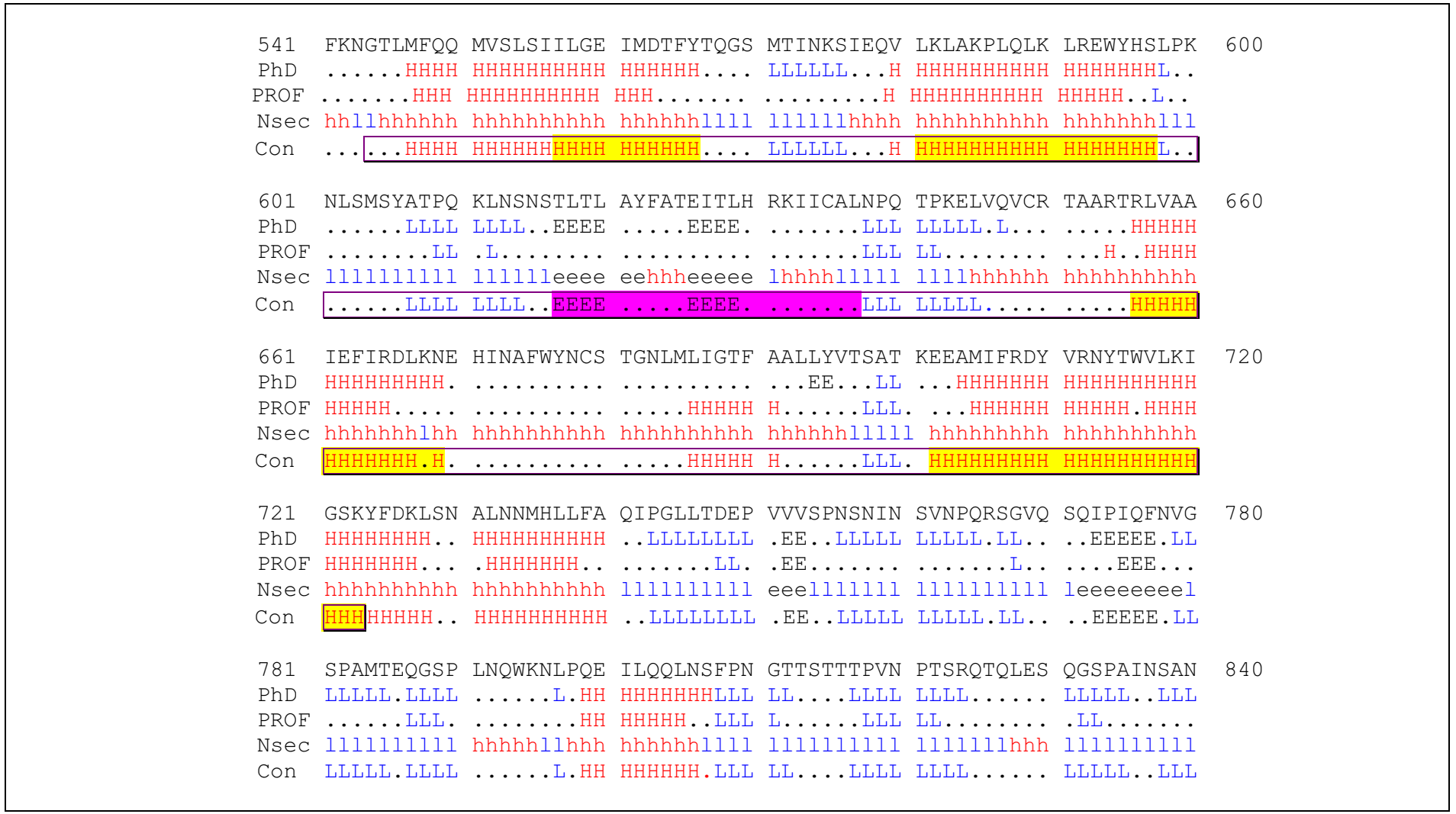

Figure 3.4 -Continued

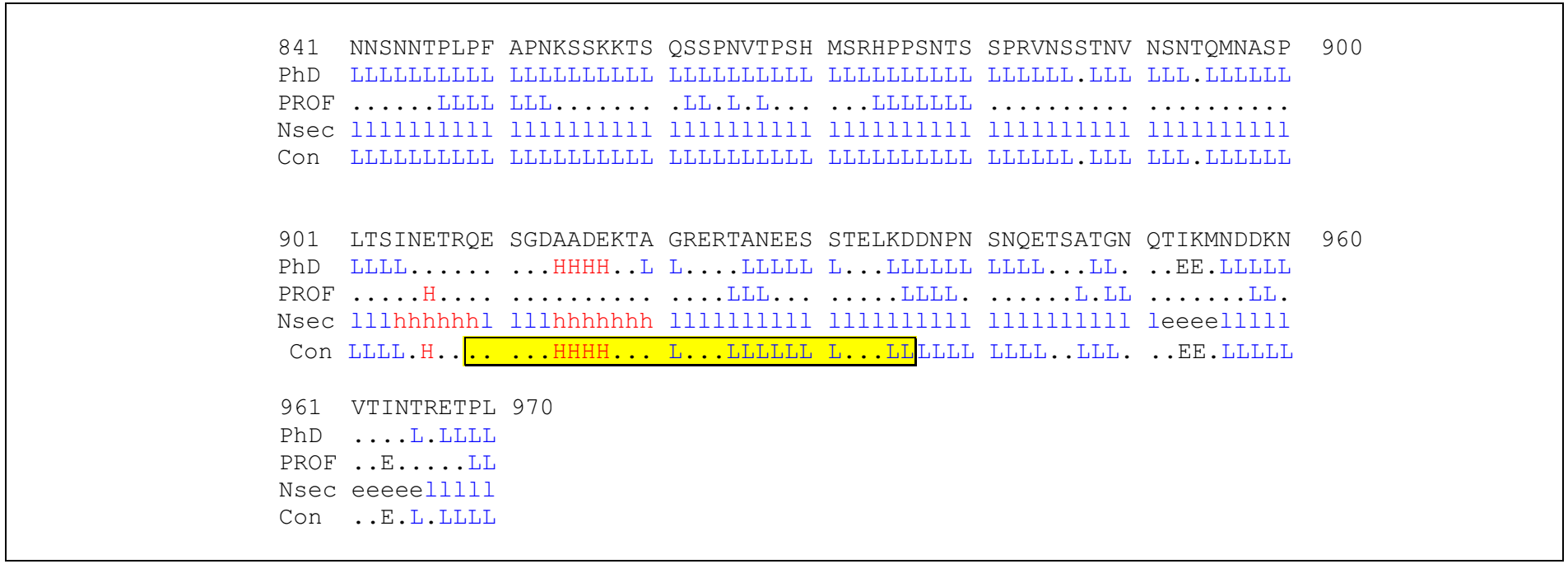

Figure 3.4 -Continued

amongst the homologues (section 3.2). Domain two was the largest domain (Figure 3.5, residues 544-723). The yellow shaded boxes showed the five α- helices with significant amino acid sequence similarity amongst the five homologues (Figures 3.2). A twenty-one amino acid motif (Figure 3.4, residues 617-637) was highly conserved in the all homologues. The motif was predicted to have a strand-loop-strand secondary structure. The motif is an interesting target for future mutational studies since the motif was highly conserved in the homologues.

3.4 Prediction of Phosphorylation sites

The phosphorylation of Gal4p has been shown to be critical for Gal4p regulation, more specifically Ser 699 must be phosphorylated for optimal transcriptional activation of the *GAL* genes in *S. cerevisiae* (Sadowski *et al.*, 1996). Is the phosphorylation of Dal81p key in the transcription regulation of *UGA* genes involved in regulation of nitrogen catabolism *in vivo*? A PROSITE motif search using the protein sequence of Dal81p revealed two different types of serine/threonine kinase phosphorylation sites. The NetPhos2.0 prediction package (neural network based) was used to approximate the probability of the phosphorylation of these PROSITE motifs. A score above 0.9 was indicative of a putative phosphorylation site.

3.4.1 cGMP / cAMP dependent protein kinase phosphorylation sites

The cGMP. / cAMP dependent protein kinase site has been observed to have the characteristic motif consensus sequence: -[RK](2)-X-[ST] - [S or T is the phosphorylation site] (Bairoch *et al.,* 1997). The kinase was observed to share a preference to phosphorylate threonine or serine residues found in the proximity of at least two positively charged residues. PROSITE motif searches (PROSITE Accession number: PDOC00004) detected three putative sites that were undetected by the NetPhos2.0 package. There was little correlation between the two datasets; this suggested that the cGMP / cAMP dependent protein kinase phosphorylation site was absent in Dal81p.

3.4.2 Protein Kinase C (PKC) phosphorylation sites

PKC displays an *in vivo* preference for serine / threonine residues located proximally to a C-terminal basic amino acid. The consensus motif is [ST]-X- [RK] [S or T is the

phosphorylation site] (Kishimoto *et al.*, 1985). The presence of additional basic residues at either the C- or N- termini enhances the Vmax and Km of the phosphorylation reaction (PROSITE Accession number: PDOC00005).

Table 3.3 Dal81p putative phosphorylation sites

Type of phosphorylation site	Designation of the phosphorylation site	Residue number	Putative phosphorylation motif	Phosphorylation score
PKC	PKC-I	Ser 17	-SVK-	0.944
PKC	PKC-II	Thr 641	-TPK-	0.989
PKC	PKC-III	Ser 855	–SSK-	0.983
PKC	PKC-IV	Ser 881	-SPR-	0.962
CK2	CK2-I	Ser 6	-SPAD-	0.985
CK2	CK2-II	Ser 108	-SKSD -	0.995
CK2	CK2-III	Ser 252	-SVFD-	0.927
CK2	CK2-IV	Ser 515	-SSND-	0.994
CK2	CK2-V	Thr 641	-TPKE-	0.994
CK2	CK2-VI	Thr 700	-TKEE-	0.744
CK2	CK2-VII	Thr 826	-TQLE	0.940
CK2	CK2-VIII	Ser 903	-SINE-	0.986
CK2	CK2-IX	Thr 925	-TANE-	0.938
CK2	CK2-X	Ser 930	–SSTE	0.974
CK2	CK2-XI	Ser 941	-SNQE-,	0.906

Four likely candidates were identified: residues Ser 17, Thr 641, Ser 855 and Ser 881 (Table 3.3). The -SVK- motif (residues 17-19) had a predicted phosphorylation score of 0.944 coupled with the presence of three basic residues, Lys 17, Lys 19 and Lys 23 on both sides of the termini (Table 3.3). The -TPK- motif (residues 641-643) had a predicted phosphorylation score of 0.989 with a single The C-terminally located basic residue, Lys 643. The –SSK- motif (residues 855-857) had a predicted score of 0.983 and had three basic residues, Lys 853, Lys 857 and Lys 858 clustered proximally on both termini. The -SPR- motif (residues 881-883) was observed to have a +2 located basic residue and a predicted phosphorylation score of 0.962. Two serine residues, Ser 17 and Ser 881 were more likely to be phosphorylated since there were observed to be proximally located positively charged residues indicating a preference for proximal PKC phosphorylation in the aqueous environment.

3.4.3 Casein Kinase II (CK2) phosphorylation sites

Casein Kinase II (CK2) phosphorylates serine / threonine residues independent of cyclic nucleotides and calcium (PROSITE Accession number: PDOC00006). The consensus motif is [ST]-x (2)-[DE] [S or T is the phosphorylation site]. Serine is preferentially phosphorylated over threonine under similar conditions. The presence of an acidic residue (Glu or Asp) must be present three residues from the phosphate acceptor site.

Phosphorylation is increased by the presence of acidic residues at the +1, +2, +3, +4 and +5 positions. Basic residues located on the N-terminal of the acceptor site decrease phosphorylation (Pinna, 1990).

Eleven potential phosphorylation sites (Table 3.3) were predicted through the PROSITE motif search correlated with data obtained from the NetPhos2.0 package from a possible sixteen sites. The CK2-I site had a predicted phosphorylation score of 0.985 and the presence of a His 6 residue located upstream this motif suggested that the phosphorylation potential at this site was lowered. The CK2-II site had predicted phosphorylation score of 0.995. The CK2-III site (-SVFD- motif) had a predicted phosphorylation score of 0.927. This motif aligned well with equivalent amino acid similar residue blocks amongst the homologues (Figure 3.5), TamAp (-TELE-, residues 200-203), Otamp (-TDLE-, residues 171-174), Ea1p (-TDFE-, residues 165-168) and Ea2p (-TDFE-, residues 181-184). Threonine was conserved at the equivalent position with serine in the alignment (Figure 3.5) and two acidic residues (+1 and +3 positions) in the homologues as compared to a serine residue (Dal81p) with a single acidic residue at the +3 position in the CK-2 motif. This raises the possibility that the conservation be evolutionary mechanism to adapt to the preferential phosphorylation of serine residues over threonine by the CK-2 phosphorylases.

The CK2–IV site (SSND-, residues 515-518) had a predicted phosphorylation score of 0.994; CK2-V (-TPKE-, residues 641-644) had a predicted score of 0.994 and interestingly was also predicted to be a protein kinase C site with an identical score.

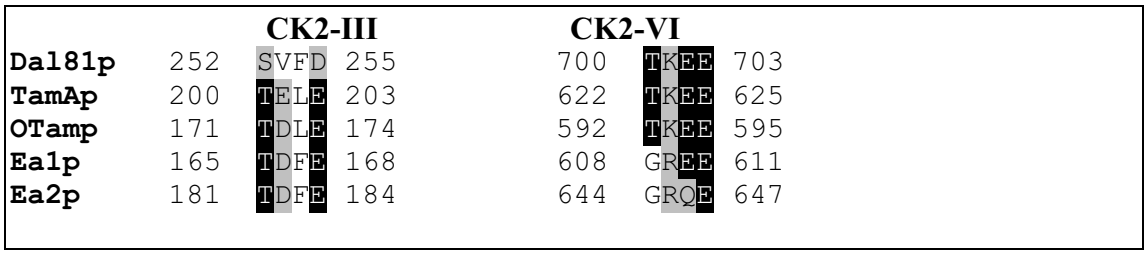

Figure 3.5: Conserved CK2-III and C2-VI phosphorylation sites amongst the five homologues. Identical amino acid residues are shown in black boxes while similar residues are denoted in grey boxes.

The next site CK2-V1 (-TKEE-, residues 700-703) had a lower predicted score of 0.744. The -TKEE- motif was also conserved at equivalent positions in the two of the four homologues, TamAp and Otamp at positions (Figure 3.5, residues 622-625 and 592-595 respectively). The other sites CK2-VII up to CK2-XI (Table 3.3) were found located on following residues: (-TQLE-, residues 826-829; -SINE-, residues 903-906; -TANE-, residues 925-928; –SSTE-, residues 930-933 and -SNQE-, residues 941-944). The other four CK2 sites (CK2-VII up to CK2XI) were observed to lie on the C-terminal region of Dal81p, a region with little sequence similarity amongst the homologues. The predicted phosphorylation scores were observed to be 0.940, 0.986, 0.938, 0.974 and 0.906 respectively. The CK2-X site (-SSTE-, residues 930-933) with a predicted phosphorylation score of 0.974 was found located in the region designated as the acidic domain (Schjerling and Holmberg, 1996) thought to act in transcriptional activation.

Eleven putative motifs were generally predicted to be likely phosphorylation sites. Two of these sites have been found conserved in homologues and this raises the possibility that the phosphorylation motif could be common sites amongst the homologues. The phosphorylation site found on the acidic domain could potentially be involved in transcriptional regulation since this domain has been previously speculated to act in transcriptional activation.

3.5 Nuclear Localisation Signals (NLSs)

There was no NLS predicted in Dal81p. The PredictNLS method at maximum accuracy detects only 43 % of all known NLS (Cokol *et al*., 2000). A 90 % overlap between the

co-existence of the NLS and the DNA binding domain of proteins has been observed (Cokol *et al*., 2000). Examples include zinc binuclear cluster proteins have NLSs spanning their respective DNA binding domains include Alcrp (Nikolaev *et al.*, 2003) and Pdr1p (Dellahodde *et al.*, 2001) The region in TamAp and Otamp (residues 146-149 and 115-118 respectively) were been identified as a putative (-KKRK-) SV40 type nuclear localization signal (Kalderon *et al*., 1984; Small *et al.*, 2001) this motif aligned well with a similar region in Ea1p (residues 112-115). A similar motif was observed with a spacer region of eleven amino acids between amino acid residues 118-129 in Ea2p. The putative SV40 type nuclear localization signal sequence was absent in Dal81p.

3.6 Solvent Accessibility (phosphorylation sites)

Putative PKC and CK2 phosphorylation sites predicted to be solvent accessible were likely to be phosphorylation sites *invivo* since reaction kinetics favour residues that are physically accessible in a solvated environment.

The Protein Kinase C (PKC) site, –SVK- (residues 17-19) were solvent exposed and predicted to lie in a loop region between (Figure 3.6), suggesting that the site was likely to be phosphorylated. The exposed residues of CK2-II site (Figure 3.6, residues 108-111) were indicative that this motif is preferentially phosphorylated. The CK2-III site (Figure 3.6, residues 252-255) was buried in domain one suggesting that the motif was not phosphorylated. The CK2-III site was observed to occupy a similar position in all five homologues in domain one (section 3.4.3). Site directed mutagenesis on this conserved motif in all five homologues would give some insight on this phosphorylation site. The CK2–IV site (Figure 3.6, residues 515-518) had all residues exposed with the secondary structure predicted to be in a loop region. This motif was located in a between the middle homology region and domain two. There was no observed amino acid sequence similarity between Dal81p and its homologues suggesting that CK2-IV site could be uniquely phosphorylated. The threonine residue located on both the PKC-II and CK2-V sites (-TPK- and TPKE- motifs respectively) was buried and this suggested that these motifs were not a putative phosphorylation sites.

Figure 3.6: **Predicted solvent accessibility on Dal81p:** .PHD prediction (Pacc) of solvent accessibility (Rost and Sander, 1994), Solvent Accessibility prediction using the predictor of non-regular secondary structure method (Nacc) (Liu and Rost, 2003). The predicted solvent accessibility consensus pattern shown as Cacc (e = exposed residue, b = buried residue) and predicted secondary structure consensus is shown as Csec. The orange and blue boxes denote the putative phosphorylation motifs for Protein kinase C phosphorylation and Casein kinase phosphorylation sites respectively. Brown boxed region shows the zinc binuclear cluster domain, Grey-boxed region represents domain one and green-boxed region shows the middle homology region. Purple-boxed region 544-723 correlates to domain two. Black boxed region with brown background shows the acidic domain, residues 909- 936 (Schjerling and Holmberg, 1996).

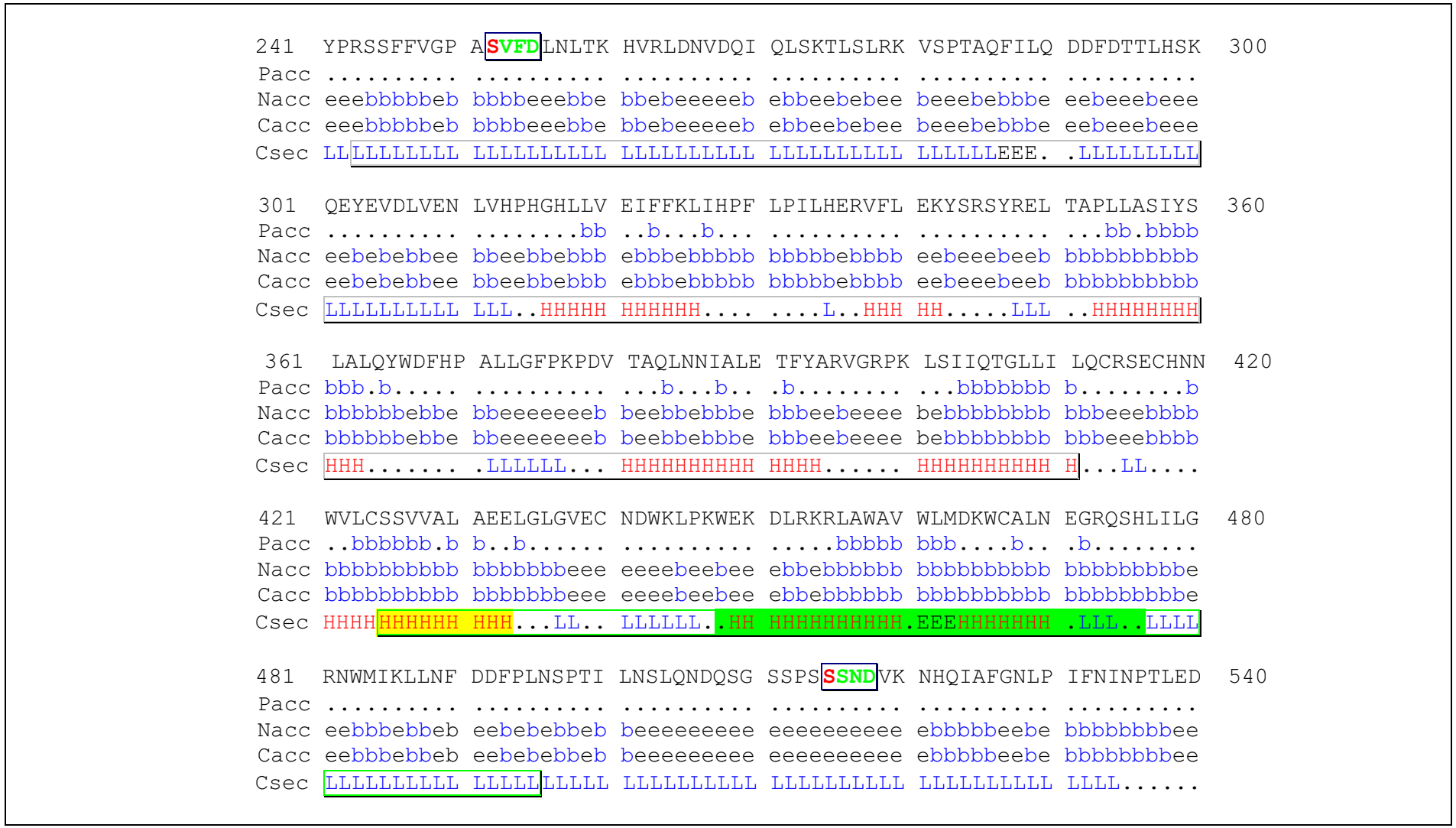

Figure 3.6 -Continued

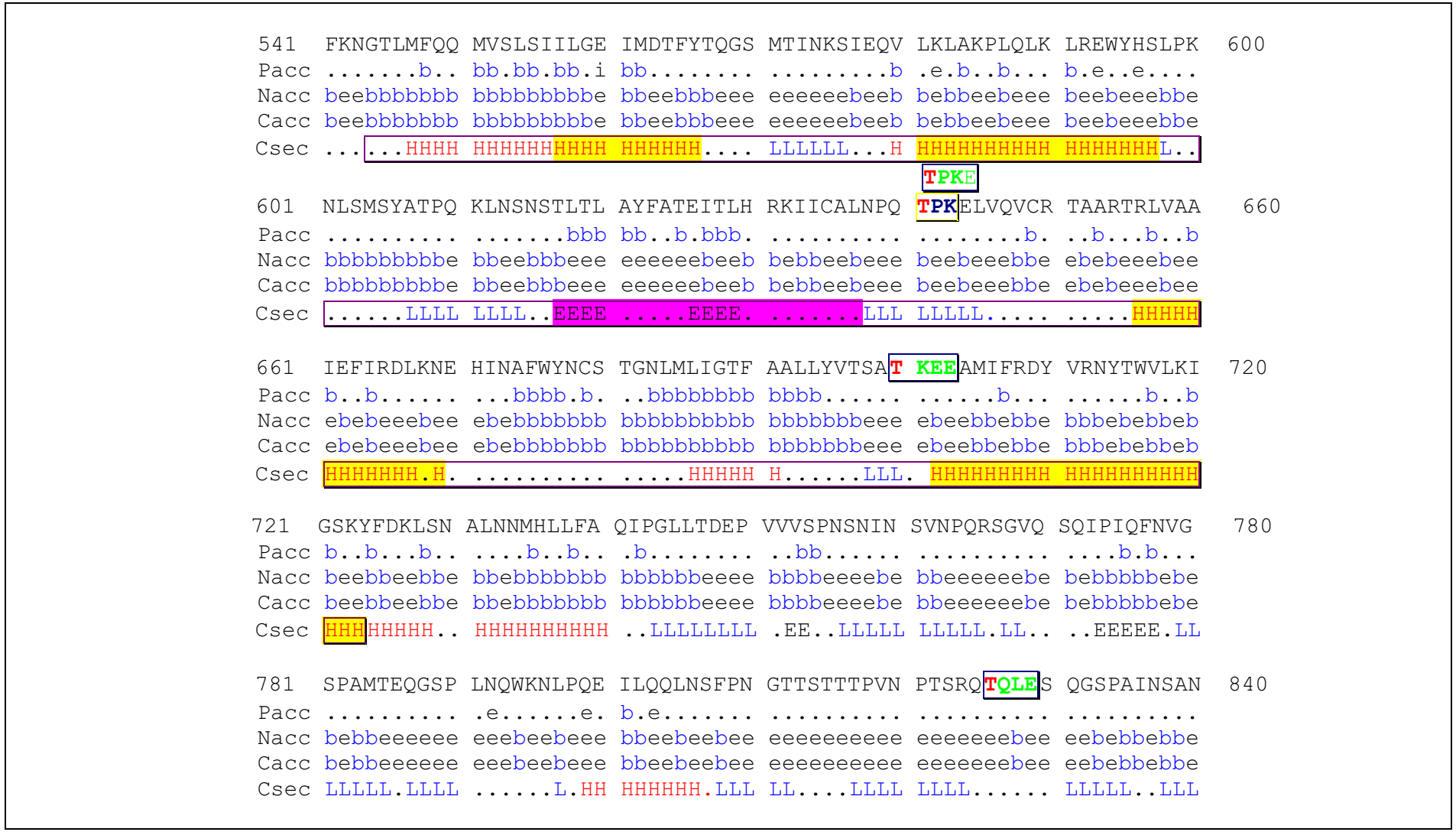

Figure 3.6 -Continued

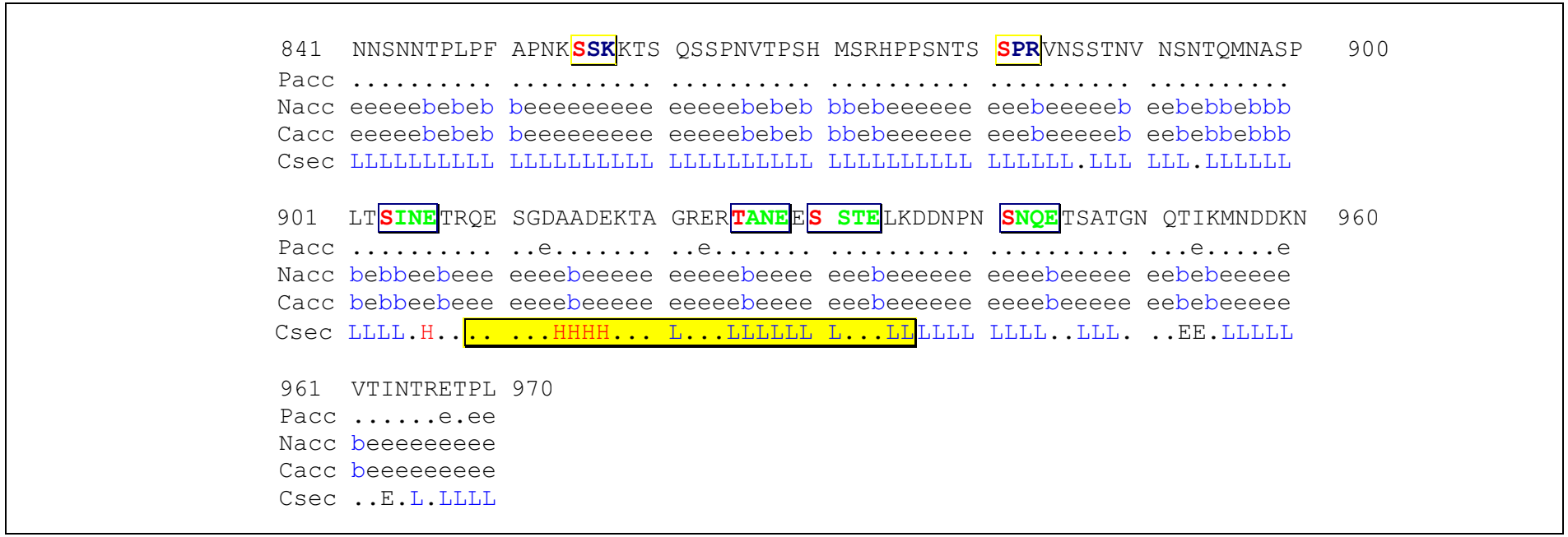

Figure 3.6 -Continued

The –SSK- motif (Figure 3.6, residues 855-857) was a putative phosphorylation site since all residues were solvent exposed. The –SPR- motif (residues 881-883) was likely to be phosphorylated since the serine residue was predicted to be solvent exposed. The -SSK- and -SPR- motifs were found located on the extreme C-terminus of the Dal81p, a region with little amino acid sequence similarity amongst the homologues. The CK2-V1 site (-TKEE-, residues 700-703) was found to be solvent exposed. This motif (Figure 3.6) was found conserved in the homologues, TamAp and Otamp (Figure 3.5, residues 622-625 and 592-595 respectively). The motif was located on domain two in Dal81p and was predicted to be α-helical. From the data above, the CK2-VI site seems to be an important site amongst the homologues. The residues found to be equivalent to the CK2-VI site in Ea1p and Ea2p were the -GREE- and -GRQE- motifs respectively (Figure 3.5). The CK2–VII motif (Figure 3.6, residues 826-829) was predicted to be solvent exposed and could be a potential phosphorylation site. The other CK2 sites, CK2-VIII up to CK2-XI were located on the extreme C-terminus region of Dal81p. The CK2-X and CK2-XI sites were solvent exposed and could be potential phosphorylation sites. CK2-X (Figure 3.6, residues 930-933) was located proximally on the acidic domain (Schjerling and Holmberg, 1996) suggesting that Dal81p activation could be mediated through phosphorylation as in the case of Gal4p.

Dal81p could be phosphorylated on multiple CK2 and PKC sites during the transcriptional activation of the *UGA* genes. Dal81p could use the multiple phosphorylation motifs to facilitate its pleotrophic role in pathway specific nitrogen catabolite repression. The conserved CK2-VI site could be a common conserved phosphorylation site between Dal81p, TamAp and Otamp.

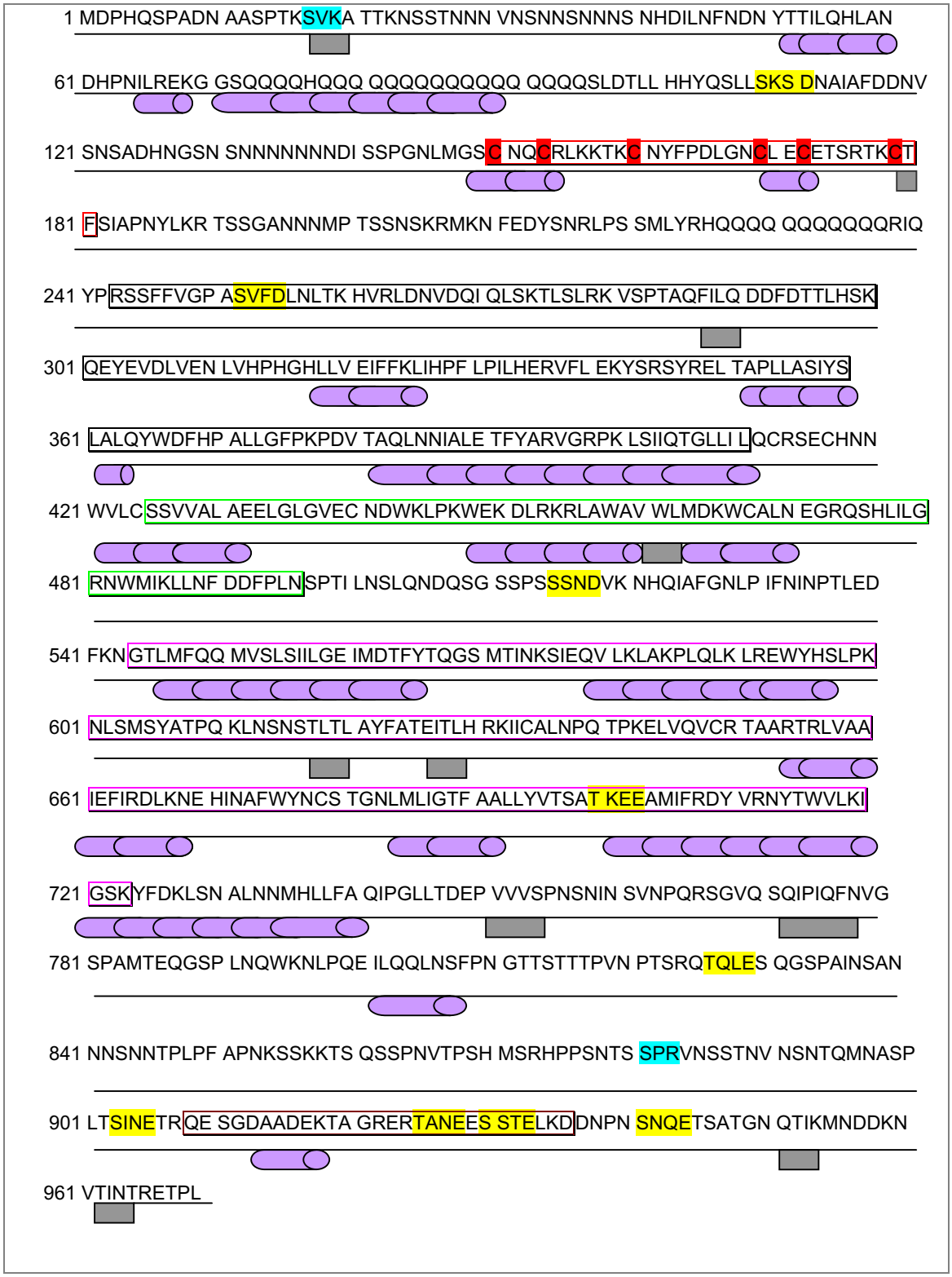

Figure 3.7: Sequence features of Dal81p. Zinc cluster region shown by the orange boxed area (residues 149-181). Domain one shown by the black boxed area (residues 243-411) and the middle homology region was shown by the lime boxed area (residues 425-496). The pink boxed area shows domain two (residues 544-723). The acidic domain was shown by the brown boxed area (Schjerling and Holmberg, 1996). The blue and yellow shaded areas show the putative phosphorylation motifs for protein kinase C and casein kinase phosphorylation sites respectively. The predicted secondary structure was represented below the amino acid residues.

Figure 3.7-Continued: The α-helical regions were shown as purple cylinders, strand regions were shown as grey rectangles. The loop regions were shown by solid black lines.

In conclusion the sequence features of Dal81p have been summarised diagrammatically as shown above (Figure 3.7).

CHAPTER 4

CONCLUSIONS AND FUTURE WORK

4.1 Bioinformatic analysis of the fungal zinc binuclear proteins

A total of one hundred and eighteen zinc binuclear cluster proteins of fungal origin were retrieved from databases by homology to the zinc binuclear cluster sequence. Thirty nine of these zinc binuclear cluster proteins were reported to be novel Alignment of these proteins revealed a consensus sequence, namely:- $ACX_2CRX_2KXKCDX_3PX_{(1-6)}CX_2CX_{(6-9)}CXY$-. Comparison of the full length protein sequences revealed that zinc binuclear cluster proteins were diverse in nature. The diversity of amino acid sequence arrangements results in different groups of structural motifs that participate in specific protein-DNA and protein-protein interactions during transcriptional activation.

4.1.1 Zinc binuclear cluster domain

The classical zinc binuclear cluster domain was generally located proximally to the N-terminus of the 118 proteins. There were large differences when zinc binuclear cluster proteins found in the *S. cerevisiae* genome were compared to the *S. pombe* and *N. crassa* genomes. The *S. cerevisiae* genome had a greater number of zinc binuclear cluster proteins relative to the *S. pombe* and *N. crassa* genomes The existence of fewer zinc binuclear cluster proteins in the *N. crassa* genome suggested that transcriptional regulation could be more complex. The transcriptional regulatory cascades could involve other classes of transcriptional activators other than the zinc binuclear cluster proteins. The conserved proline residue in the zinc binuclear cluster domain was found to be more prevalent in the *S. cerevisiae* genome relative to the *S. pombe* and *N. crassa* genomes. The data suggested that the amino acid residues in zinc binuclear cluster domains of some homologous proteins such as Prop1p_N and Prop1p_S were significantly similar.

In addition some zinc binuclear cluster protein had zinc binuclear cluster domains that were located in unusual positions along the protein sequence; one protein had a centrally located zinc binuclear cluster domain with two predicted coiled coils. Some protein annotations had zinc binuclear cluster domains in the extreme C-terminal and central regions of the protein sequence.

4.1.2 Coiled coil domain

Zinc binuclear cluster proteins were generally divided into two subgroups: those with a coiled coil and those that lacked a coiled coil. Most proteins with the predicted coiled coil motif had the following pattern, namely: a coiled coil region separated by a short linker region located to the C-terminus of the zinc binuclear cluster domain (Schjerling and Holmberg, 1996). Some of the zinc binuclear proteins had predicted coiled coils that were located in unusual positions along the protein sequence. This suggested that the coiled coil motif could also function as a protein-protein interaction module that is not specifically involved in dimerisation of zinc binuclear cluster proteins Structural studies of Gal4p, Hap1p and Ppr1p dimers have shown that these proteins all dimerise through a C-terminal located coiled coil linked to the zinc binuclear located domain via a linker region (Marmorstein *et al.,* 1992; Marmorstein and Harrison, 1994; King *et al.,* 1999). The location of the coiled coil motif in unusual positions along the protein sequence could help explain the interaction of some zinc binuclear cluster proteins in transcriptional activation cascades. The coiled coil motif was found was more to be more prevalent in the *S. cerevisiae* genome relative to the *S. pombe* and *N. crassa* genomes. This conservation of the coiled coil suggested that the coiled coil could be an important protein-protein interaction module in zinc binuclear cluster proteins in the *S. cerevisiae* genome.

4.1.3 Dal81p homologues

Four homologues of Dal81p were identified by homology searching (Dal81p, TamAp, Otamp, Ea1p and Ea2p). The alignment of Dal81p and its homologues revealed amino acid sequence identity at greater than 30 %. Four distinct domains were conserved in Dal81p and its homologues. The domains included: the zinc binuclear cluster domain, the middle homology domain and two new domains that were designated as domain one and two respectively.

There was no coiled coil motif detected in all five homologues. This suggested that these proteins did not dimerise via the coiled coil motif. The presence of common motifs in domain one and two suggested that this subgroup of zinc binuclear cluster proteins could interact with other proteins via these new domains.

The middle homology region was observed to have two structural motifs that were conserved in all the homologues. These two motifs were an amphipathic α-helix and a α–β–α like motif. The amphipathic α-helices have been implicated in protein-protein interactions, amphipathic α-helices have been shown to interact with other amphipathic α-helices by forming dimers (Lesk, 2001). Since the helix is conserved in the middle homology region of the homologues the helix could be an important functional motif. Dal81p and its homologues could potentially interact with other zinc binuclear cluster transcription factors through this motif. The possibility of Dal81p interacting with Uga3p through this domain cannot be ruled out. The presence of this structural motif further suggested that these zinc binuclear cluster proteins (in Dal81p and it homologues) could have evolved this domain to facilitate for a more specialised role in transcriptional activation. The evolution of this domain could enable Dal81p to facilitate its pleotrophic role.

The middle homology region was found to exhibit more than 60 % amino acid sequence similarity amongst all the proteins. The middle homology region in the homologues was found to lie within the eight-motif domain that was speculated to be important in transcriptional regulation of Gal4p (Poch, 1997). The conserved regions the protein alignments correlate to the regions shown to be critical for TamAp function through complementation studies (Small *et al*., 2001). In summary the zinc binuclear cluster domain seems to be an evolutionary relic while middle homology region and the two putative new domains discussed in this study could mediate a role in the regulation of these homologous proteins (Dal81p, Tamp, Otamp, Ea1p and Ea2p) during transcriptional regulation.

A putative SV40 type NLS (Kalderon *et al.,* 1984) was found conserved in four homologues but observed to be absent in Dal81p The results show that TamAp,

Otamp, Ea1p and Ea2p could be localised in the nucleus by this signal. The lack of a NLS in Dal81p suggests that the signal was undetected by current latest predictive methods. Dal81p could be nuclear localised through another NLS. The presence of a NLS would indicate the nuclear localisation of the protein since transcription factors have to enter the nucleus during transcription.

Dal81p could potentially phosphorylated like many transcription factors involved in signal transduction pathways. Evidence of multiple potential phosphorylation sites was shown on Dal81p. The CK2 sites present in domains one and two could be important since these motifs are located on conserved regions. The putative CK2-IX and CK2-X sites could be involved in transcriptional activation since the site is located in the acidic domain of Dal81p (residues 930-933). There was evidence of conserved phosphorylation sites amongst the homologues. The putative CK2-VI site was found conserved in Dal81p, TamAp and Otamp at topologically equivalent positions. These results show that multiple phosphorylation sites could be important in the transcriptional regulation of Dal81p and its homologues.

4.2 Future Work

The ongoing sequencing projects on other fungal genomes and continual annotation of published genomes will provide larger datasets for more detailed comparative bioinformatic studies. The evolutionary relationships will help explain the divergence and origin of these proteins. As more fungal genomes are elucidated, a clearer phylogenetic pattern should emerge.

Site directed mutagenesis of the serine residue through a conservative alanine substitution on the CK2-III phosphorylation could be used to determine whether the site is critical in the transcriptional regulation of Dal81p. Similar site directed mutagenic experiments performed on the CK2-VI site could determine whether this site is important for the transcriptional activation of Dal81p and its homologues. The putative CK2-IX and CK2-X sites located on the acidic domain could reveal important sites *in vivo* through site directed mutagenesis for their possible role in the transcriptional activation of Dal81p. Data generated from such experiments could be

used to describe the role of phosphorylation in the transcriptional regulation of Dal81p. The positively charged residues in putative nuclear localisation signals (NLS) present in the homologues could be mutated to alanine residues. A vector can be used to transform these mutant sequences. The resulting transformants can be then evaluated for growth on minimal media containing GABA as the sole nitrogen source. The growth of these transformants could then be used to infer whether the NLS sequences are functional *in vivo.* The availability of more structural data through x-ray crystallography will enable homology modelling of individual domains and full-length protein sequences.

REFERENCES

Adam, S.A., Lobl T.J., Mitchell, M.A., and Gerace, L. (1989) Identification of specific binding proteins for a nuclear location sequence. *Nature*. **337**: 276-279.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. (2002) Macromolecules: structure, shape and information*, Molecular Biology of the Cell*. (4th edition) Garland Publishing Inc., New York. 116-120.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z. Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389-3402.

André, B., Talibi, D., Soussi Boudekou, S.S., Hein, C., Vissers, S., and Coornaert, D. (1995) Two mutually exclusive regulatory systems inhibit UAS_{GATA} , a cluster of 5`-GAT(A/T)-3` upstream from the *UGA4* gene of *Saccharomyces cerevisiae*. *Nucleic Acids Res*. **23**: 558-564.

Bairoch, A., Bucher, P., and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res*. **25**: 217-221.

Baleja, J.D., Thanabal, V. and Wagner, G. (1997) Refined solution structure of the DNA-binding domain of GAL4 and use of 3J (113Cd, 1H) in structure determination. *J. Biomol. NMR.* **10**: 397-401.

Berger, B., Wilson, D.B., Wolf, E., Tonchev, T., Milla, M. and Kim, P.S. (1995) Predicting Coiled Coils by Use of Pairwise Residue Correlations. *Proc. Nat. Acad. Sci.* **92**: 8259-8263.

Blom, N., Gammeltoft, S. and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**: 1351-1362.

Boulikas, T. (1993) Nuclear localization signals (NLS). *Crit. Rev. Eukaryot. Gene Expr.* **3:** 193-227.

Branden, C. and Tooze, J. (1991) Introduction to Protein Structure. In: *Motifs of Protein Structure.* (1st edition), Garland Publishing Inc., New York. 12-29.

Bricmont, P.A. and Cooper, T.G. (1989) A gene product needed for induction of allantoin system genes in *Saccharomyces cerevisiae* but not for their transcriptional activation. *Mol. Cell. Biol*. **9**: 3869-3877.

Bricmont, P.A., Daugherty, J.R. and Cooper, T.G. (1991) The *DAL*81 gene product for induced expression of two differently regulated nitrogen catabolic genes is *Saccharomyces cervisiae*. *Mol. Cell. Biol*. **11**: 1161-1166.

Cardenas, M.E., Cutler, N.S., Lorenz, M.C., Di Como, C.J. and Heitman, J. (1999) The TOR signaling cascade regulates gene expression in response to nutrients. *Genes & Dev*. **13**: 3271-3279.

Coffman, J. A., Rai, R., Cunningham, T., Svetlov, V., and Cooper, T.G. (1996) Gat1p, a GATA family protein whose production is sensitive to nitrogen catabolite

repression, participates in transcriptional activation of nitrogen-catabolic genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **16:** 847-858.

Cokol, M., Nair, R., and Rost, B. (2000) Finding nuclear localization signals. *EMBO Rep.* **1**: 411-415.

Coleman, S.T., Fang, T.K., Rovinsky, S.A., Turano, F.J., and Moye-Rowley Scott, W. (2001) Expression of a glutamate decarboxylase homologue is required for normal oxidative stress tolerance in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **276**: 224-250.

Corbett, A.H., Koepp, D.M., Schlenstedt, G., Lee, M.S., Hopper, A.K., and Silver, P.A. (1995) Rna1p, a Ran/TC4 GTPase activating protein, is required for nuclear import**.** *J. Cell. Biol.* **130**: 1017-1026.

Delahodde, A., Pandajaitan, R., Corrral–Debrinski, M., and Jacq, C. (2001) Pse1/Kap 121 dependent major yeast multidrug resistance (MDR) transcription factor Pdr1p. *Mol. Microbiol*. **39**: 304-312.

Eisenberg, D. (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Nat. Acad. Sci.* **100**: 11207-11210.

Escher, D., Bodmer-Glavas, M., Aclide, B., and Schaffner, W. (2000) Conservation of Glutamine–Rich Transactivation Function Between Yeast and Humans. *American Soc. Micro.* **20**: 2774-2782.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R.L., Perkins, D.D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S., Lander, E.S., Nusbaum, C. and Birren, B. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature.* **422:** 859-868.

Glass, D., el-Maghrabi, M.R., and Pilkis, S.J. (1986) Synthetic Peptides corresponding to the site phosphorylated in 6-phosphofructokinase/fructose-2, 6-biphosphatase as substrates pf cyclic nucleotide dependent protein kinases. *J. Biol. Chem.* **261**: 2987-2993.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. (1996) Life with 6000 genes. *Science.* **274**: 563-567.

Grenson, M., Muyldermans, F., Broman, K., and Vissers, S. (1987) 4-aminobutyric acid (GABA) uptake in Baker`s yeast *Saccharomyces cerevisiae* is mediated by the general amino acid permease, the proline permease and a GABA-specific permease and a GABA-specific integrated into the GABA-catabolic pathway. *Life Sci. Adv Ser. C.* **6**: 35-39.

Hellauer, K., Rhochon, M.H., and Turcotte, B. (1996) A novel DNA binding motif for yeast zinc cluster proteins: the Leu3p and Pdr3p transcriptional activators recognise everted repeats. *Mol. Cell. Biol*. **16**: 6096-6102.

Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci.* **89**: 10915-10919.

Hodel, M.R., Corbett, A.H., and Hodel, A.E. (2001) Dissection of a nuclear localization signal. *J. Biol. Chem.* **276**: 1317-1325.

Idicula, A.M., Blatch, G.L, Copper, T.G., and Dorrington, R.A. (2002) Binding and activation by the Zinc Binuclear Cluster Transcription factors of *Saccharomyces cerevisiae*: Redefining the *UAS-GABA* and its interaction with Uga3p. *Biol. Chem.* **277**: 45977-45983.

Kalderon, D. Richardson, W.D., Markham, A.F. and Smith, A.E. (1984) Sequence requirements for nuclear localisation of SV-40 Large-T antigen. *Nature.* **311**: 33-38.

Kaffman, A., Rank, N.M., and O`Shea E.K. (1998) Phosphorylation regulates the association of the transcription factor Pho4 with its import receptor Pse1/Kap121. *Genes & Dev*. **12**: 2673-2683.

Keating, A.E., Malashkevich, V.N., Tidor, B., and Kim, P.S. (2001) Side chain repacking for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Nat. Acad. Sci.* **98**: 14825-14830.

King, D.A., Zhang, L., Guarente, L., and Marmorstein, R. (1999) Structure of a HAP1-DNA complex reveals dramatically asymmetric DNA binding by a homodimeric protein. *Nat. Struct. Biol.* **6**: 64-71.

Kishimoto, A., Nishsiyama, K., Nakanishi, H., Uratsuji, Y., Nomura, H., Takeyama, Y., and Nishizuka, Y. (1985) Studies on the phosphorylation of myelin basic protein kinase C and adenosine 3`:5`-monophosphate dependent protein kinase. *J. Biol. Chem.* **260**: 12492-12499.

Komeili, A., and O'Shea, E.K. (1999) Roles of Phosphorylation Sites in Regulating Activity of the Transcription Factor Pho4. *Science.* **284**: 977-980.

Krishna, S.S., Majumdar, I., and Grishin, N.V. (2003) Structural classification of zinc fingers. *Nucleic Acids Res.* **31**: 532-550.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**: 105-132.

Lesk, A.M. (2001) Pattern and form in Protein Structure*, Introduction to Protein Architecture*. (1st edition), Oxford University Press Inc., New York. 59-101.

Liu, J., and Rost, B. (2003) NORSp: predictions of long regions without non regular structure. *Nucleic Acids Res.* **31**: 3833-3835.

Lupas, A. (1996a) Coiled coils: new structures and functions. *Trends Biochem. Sc.* **21**: 375-382.

Lupas, A. (1996b) Prediction and analysis of coiled coils structures. *Methods Enzymol.* **266**: 513-525.

Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting Coiled Coils from Protein Sequences. *Science.* **252**:1162-1164.

Mackiewicz, P., Kowalczuk, M., Mackiewicz D., Nowicka, A., Dudkiewicz, M., Laszkiewicz, A., Dudek, M.R., and Cebrat, S. (2002) How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome? *Yeast.* **7**: 619-629.

Mamane, Y., Hellauer, K., Rochon, M., and Turcotte, B. (1998) A linker Region of the Yeast Zinc Cluster Protein Leu3p Specifies Binding to Everted repeat DNA. *Bio. Chem*. **273**: 18556-18561.

Marmorstein, R., and Harrison, S. C. (1994) Crystal structure of a Ppr1p-DNA complex: DNA recognition by proteins containing a Zn2Cys6 binuclear cluster. *Genes & Dev.* **8**: 2504-2512.

Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S.C. (1992) DNA recognition by Gal4p: structure of a protein-DNA complex. *Nature.* **356**: 408-414.

Marzluf, G.A. (1993) Regulation of sulfur and nitrogen metabolism in filamentous fungi. *Annu. Rev. Microbiol*. **47**: 31-55.

Marzluf, G.A. (1997) Genetic regulation of nitrogen metabolism in the fungi. *Microbiol. Mol. Biol. Rev.* **61**: 17-32.

Mattaj, I., and Englmeier, L. (1998) Nucleocytoplasmic transport: The soluble phase *Annu. Rev. Biochem.* **67**: 265-306.

Moore, M.S., and Blobel, G. (1994) Purification of a Ran-interacting protein that is required for protein import into the nucleus. *Proc. Natl. Acad. Sci.* **91**:10212-10226.

Nair, R., Carter, P., and Rost, B. (2003) NLSdb: database of nuclear localisation signals. *Nucleic Acids Res.* **31**: 397-399.

Neer, E.J., Schmidt, C.J., Nambudripad, R., and Smith, T.F. (1994) The ancient regulatory-protein family of WD-repeat proteins. *Nature.* **371**: 297-300.

Nikolaev, I., Cochet, M., and Felenbok, B. (2003) Nuclear Import of Zinc Binuclear Cluster Proteins Proceeds through Multiple, Overlapping Transport Pathways *American Soc. Microbiol.* **2**: 209-221.

Noel, J., and Turcotte, B. (1998) Zinc cluster proteins Leu3p and Uga3p recognise highly related but distinct DNA targets. *J. Biol. Chem*. **273**: 17463-17468.

Paschal B.M., and Gerace, L. (1995) Identification of NTF2, a cytosolic factor for nuclear import that interacts with nuclear pore complex protein p62. *J. Cell. Biol.* **129**: 925-937.

Pauling, L., and Corey, R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Nat. Acad. Sci.* **37**: 251-256. Cited in: Eisenberg, D. (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Nat. Acad. Sci.* **100**: 11207-11210.

Pauling, L., Corey, R.B., and Branson, H.R. (1951) The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Nat. Acad. Sci.* **37**: 205-211. Cited in: Eisenberg, D. (2003) The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Nat. Acad. Sci.* **100**: 11207-11210.

Pinna, L.A. (1990) Casein kinase 2: an 'eminence grise' in cellular regulation *Biochim. Biophys. Acta*. **1054**: 267-284.

Poch, O. (1997) Conservation of a putative inhibitory domain in GAL4 family members. *Gene* **184**: 229-235.

Rai, R., Daugherty, J.R., Cunningham, T.S., and Cooper T.G. (1999) Overlapping positive and negative GATA factor-binding sites mediate inducible DAL7 gene expression in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **274**: 28026-28034.

Ramos, F., El Guezzar, M., Grenson, M., and Wiame, J.M. (1985) Mutations affecting the enzymes involved in the utilization of 4-aminobutyric acid as nitrogen source by the yeast *Saccharomyces ce*revisiae. *Eur. J. Biochem*. **149**: 401-404.

Rost, B. (1996) PHD: predicting one dimensional protein structure by profile based neural networks. *Proteins*. **20**: 216-226.

Rost, B. (2001) Review: Protein secondary structure prediction continues to rise. *J. Struc. Bio.* **134**: 204-218.

Rost, B., and Liu, J. (2003) The PredictProtein server. *Nucleic Acids Res.* **31**: 3300-3304.

Rost, B., and Sander, C. (1993) Prediction of secondary structure at better than 70 % accuracy. *J. Mol. Biol*. **232**: 584-599.

Rost, B., and Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families. *J. Mol. Biol*. **232**: 584-599.

Sadowski, I., Costa, C., and Dhanawansa**,** R. (1996) Phosphorylation of Ga14p at a single C-terminal residue is necessary for galactose-inducible transcription. *Mol. Cell. Biol.* **16**: 4879-4887.

Sayle, R.A., and Milner-White, E.J. (1995) RASMOL: Biomolecular graphics for all. *Trends Biochem. Sc.* **20**: 374-376.

Schjerling, P., and Holmberg, S. (1996) Comparative amino acid sequence analysis of C_6 zinc cluster family of transcription regulators. *Nucleic Acids Res.* **23**: 550-557.

Small, A.J., Todd, R.B., Zanker, M.C., Delimitrou, S., Hynes, M.J., and Davies, M.A. (2001) Functional analysis of TamA, a coactivator of nitrogen-regulated gene expression in *Aspergillus nidulans*. *Mol. Gen. Geno*. **265**: 636-646.

Stanbrough, M. Rowen, D. W., and Magasanik, B. (1995) Role of the GATA Factors Gln3p and Nil1p of *Saccharomyces cerevisiae* in the Expression of Nitrogen-Regulated Genes. *Proc. Nat. Acad. Sci.* **92**: 9450-9454.

Talibi, D., Grenson, M., and André, B. (1995) *Cis*- and *trans* -acting elements determining induction of the genes of the γ-aminobutyrate (GABA) utilisation pathway in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **23**: 550-557.

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

Todd, R., and Andrianopoulos, A. (1997) Evolution of a fungal regulatory gene family: The Zn II)2Cys6 binuclear cluster DNA binding motif. *Fungal Gen. Biol*. **21**: 388-405.

Trainor, C.D., Ghirlando, R., and Simpson, M.A. (2002) GATA zinc finger interactions modulate DNA binding and transactivation. *J. Biol. Chem*. **275**: 28157-28166.

Vissers, S., Andre, B., Muyldermans, F., and Grenson, M. (1989) Positive and negative regulatory elements control the expression of the *UGA4* gene coding for the inducible 4-aminobuytric-acid-specific permease in *Saccharomyces cerevisiae*. *J. Biochem.* **181**: 357-361.

Vissers, S., Andre, B., Muyldermans, F., and Grenson, M. (1990) Induction of the 4-aminobutyrate and urea-catabolic pathways in *Saccharomyces cerevisiae. J. Biochem.* **187**: 611-616.

Wolf, E., Kim, P.S., and Berger, B. (1997) MultiCoil: A program for predicting two- and three-stranded coiled coils. *Protein Sci*. **6**: 1179-1189.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E.J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R.G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K.,

Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T.M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S.J., Xiang, Z., Hunt, C., Moore, K., Hurst, S.M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V.A., Garzon, A., Thode, G., Daga, R.R., Cruzado, L., Jimenez, J., Sanchez, M., Del Rey, F., Benito, J., Dominguez, A., Revuelta, J.L., Moreno, S., Armstrong, J., Forsburg, S.L., Cerutti, L., Lowe, T., McCombie, W.R., Paulsen, I., Potashkin, J., Shpakovski, G.V., Ussery, D., Barrell, B.G., Nurse, P., and Cerrutti L. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*. **415**: 871-880.

Xu, W., Harrison, S. C., and Eck, M. J. (1997) Three-dimensional structure of the tyrosine kinase c-Src. *Nature*. **385**: 595-602.

Yoo, H.K. and Cooper, T.G. (1989) The *DAL7* promoter consists of multiple elements that cooperatively mediate regulation of the gene`s expression. *Mol. Cell. Biol.* **9**: 3231-3243.

APPENDIX

Table A-1: Summary of the data on the Zinc binuclear cluster proteins used in the study

Fungal Zinc Finger	Sequence length	Database	References	Function	Organism
Aflr1p	437	sp|P41765	P41765	Afflatoxin biosynthesis regulator protein	*A. flavus*
*Ca4p	625	embl|CAD37157.1	CAD37157	Hypothetical protein	*A. fumigatus*
Aflr2p	433	sp|P52957	P52957	Sterigmatocystin biosynthesis regulator	*A. nidulans*
Alcrp	821	sp|P21228	P21228	Regulatory protein	*A. nidulans*
Amdr2p	765	sp|P15699	P15699	Acetamidase regulatory protein	*A. nidulans*
Amyr2p	662	gb|AAF17101.1	AAF17101	Amylase regulator	*A. nidulans*
Facb2p	867	gb|AAB63565.1	AAB63565	Acetate DNA binding protein	*A. nidulans*
Nirap	892	sp|P28348	P28348	Nitrogen assimilation transcription factor	*A. nidulans*
Qutap	825	sp|P10563	P10563	Quinic acid utilization activator	*A. nidulans*
TamAp	739	embl|Q00741	Q00741	Nitrogen assimilation transcription factor	*A. nidulans*
Uayp	1060	sp|P49413	P49413	Positive regulator of purine utilization	*A. nidulans*
Facb3p	862	gb|AAB63563.1	AAB63563	Acetate DNA binding protein	*A. niger*
*Xlnrp	875	sp|O42804	O42804	Transcriptional activator xlnR	*A. niger*
Amdr1p	735	sp|Q06157	Q06157	Acetamidase regulatory protein	*A. oryzae*
*Amyr1p	604	dbj|BAA25754.1	BAA25754	Amylase regulator	*A. oryzae*
*Otamp	711	embl|Q96WK9	Q96WK9	Nitrogen assimilation transcription factor	*A. oryzae*
Czf1p	388	sp|P28875	P28875	Zinc finger protein	*C. albicans*
Fcr1p	517	sp|O93870	O93870	Fluconazole resistance protein 1	*C. albicans*
*Ct1ap	909	sp|P52958	P52958	Cutinase transcription factor	*F. solani*
*Ct1bp	882	sp|P52959	P52959	Cutinase transcription factor	*F. solani*
*Mut3p	929	gb|AAK84946.1	AAK84946	Peroxisome proliferation regulator	*H. polymorpha*
Lac9p	865	sp|P08657	P08657	Lactose metabolism regulatory protein	*K. lactis*
*Sef1p_K	1071	sp|P87164	P87164	Supressor protein	*K. lactis*
Pribp	565	sp|P49412	P49412	Prib protein	*L. edodes*
*Afa1p	974	gb|AAF37291.1	AAF37291.1	Putative transcription factor Pig1p	*M. grisea*
*Ea2p	805	embl|EAA56521	EAA56521	Hypothetical protein	*M. grisea*
*Ac15p	865	sp|P87000	P8700	Regulatory protein	*N. crassa*
Acr2p	595	sp|P78704	P78704	Acriflavine sensitivity control protein	*N. crassa*

Table A-1–Continued

Fungal Zinc Finger	Sequence length	Database	References	Function	Organism
*Cmr1p	321	embl|CAD70758.1	CAD70758.1	Regulatory protein	*N. crassa*
*Ea1p	775	embl|EAA29566	Q7S2G1	Hypothetical protein	*N. crassa*
*Flufp	792	sp|O13360	O13360	Conidial development protein fluffy	*N. crassa*
Nit4p	1090	sp|P28349	P28349	Nitrogen assimilation transcription factor	*N. crassa*
*Pro1p_N	696	embl|CAB89819.1	CAB89819.1	PRO1 protein	*N. crassa*
Qa1fp	816	sp|P11638	P11638	Quinic acid utilization activator	*N. crassa*
*Gsabp	1862	sp|Q9HFR4	Q9HFR4	Pexophagy regulatory protein	*P. pastoris*
Arg2p	880	sp|P05085	P05085	Arginine metabolism regulation	*S. cerevisiae*
Aro80p	950	sp|Q04052	Q04052	Putative transcriptional regulator	*S. cerevisiae*
Cat8p	1433	sp|P39113	CAA55139	Regulatory protein	*S. cerevisiae*
Cb32bp	608	embl|CAA89804.1	CAA89804	Regulatory protein	*S. cerevisiae*
Cha4p	648	sp|P43634	P43634	Activatory protein	*S. cerevisiae*
Cyp1p	1502	sp|P12351	P12351	Activatory protein	*S. cerevisiae*
Dal81p	970	sp|P21657	P21657	Transcriptional activator protein	*S. cerevisiae*
Gal4p	881	sp|P04386	P04386	Regulatory protein	*S. cerevisiae*
Hal9p	1030	sp|Q12180	Q12180	Putative transcriptional regulator	*S. cerevisiae*
Leu3p	886	sp|P08638	P08638	Regulatory protein	*S. cerevisiae*
Ly14p	790	sp|P40971	P40971	Lysine biosynthesis regulator	*S. cerevisiae*
Ma1rp	473	sp|P53338	P53338	Maltose fermentation regulator	*S. cerevisiae*
Ma3rp	468	sp|P38157	P38157	Maltose fermentation regulator	*S. cerevisiae*
Ma6rp	473	sp|P10508	P10508	Maltose fermentation regulator	*S. cerevisiae*
Pdr1p	1060	sp|P12383	P12383	Pleiotropic drug resistance regulator	*S. cerevisiae*
Pdr3p	976	sp|P33200	P33200	Pleiotropic drug resistance regulator	*S. cerevisiae*
Pip2p	996	sp|P52960	P52960	Peroxisome proliferation regulator	*S. cerevisiae*
Ppr1p	904	sp|P07272	P07272	Pyrimidine pathway regulatory protein	*S. cerevisiae*
Put3p	979	sp|P25502	P25502	Proline utilization trans-activator	*S. cerevisiae*
Rdr1p	546	sp|Q08904	Q08904	Putative transcriptional regulator	*S. cerevisiae*
Sef1p_Y	1057	sp|P34228	P34228	Supressor protein	*S. cerevisiae*
Sip4p	829	sp|P46954	P46954	SIP4 protein	*S. cerevisiae*

Table A-1–Continued

Fungal Zinc Finger	Sequence length	Database	References	Function	Organism
Stb4p	949	sp|P50104	P50104	Probable transcriptional regulator	*S. cerevisiae*
Stb5p	743	sp|P38699	P38699	Probable transcriptional regulator	*S. cerevisiae*
Teap1	759	sp|P47988	P47988	Enhancer activator	*S. cerevisiae*
Thi2p	450	sp|P38141	P38141	Thiamine biosynthesis protein	*S. cerevisiae*
Uga3p	528	sp|P26370	P26370	Transcriptional regulatory protein	*S. cerevisiae*
Ume6p	836	sp|P39001	P39001	Transcriptional regulatory protein	*S. cerevisiae*
Upc2p	913	sp|Q1251	Q12151	Putative transcriptional regulator	*S. cerevisiae*
Yaf1p	1062	sp|P39720	P39720	Transcriptional regulatory protein	*S. cerevisiae*
Yb00p	1094	sp|P38114	P38114	Transcriptional regulatory protein	*S. cerevisiae*
Yb89p	529	sp|P38140	P38140	Transcriptional regulatory protein	*S. cerevisiae*
Ybo3p	919	sp|P38073	P38073	Transcriptional regulatory protein	*S. cerevisiae*
Ycz6p	832	sp|P25611	P25611	Transcriptional regulatory protein	*S. cerevisiae*
Yd03p	885	sp|Q06639	Q06639	Transcriptional regulatory protein	*S. cerevisiae*
Ydr520p	772	sp|Q04411	Q04411	Putative transcriptional regulator	*S. cerevisiae*
Ye14p	794	sp|P39961	P39961	Transcriptional regulatory protein	*S. cerevisiae*
Yff2p	465	sp|P43551	P43551	Transcriptional regulatory protein	*S. cerevisiae*
Yhl6p	883	sp|P38781	P38781	Transcriptional regulatory protein	*S. cerevisiae*
Yinop	964	sp|P40467	P40467	Transcriptional regulatory protein	*S. cerevisiae*
Yjk3p	618	sp|P42950	P42950	Transcriptional regulatory protein	*S. cerevisiae*
Yju6p	758	sp|P39529	P39529	Putative transcriptional regulator	*S. cerevisiae*
Yk44p	863	sp|P36023	P36023	Transcriptional regulatory protein	*S. cerevisiae*
Ykd8p	1170	sp|P32862	P32862	Transcriptional regulatory protein	*S. cerevisiae*
Ykm1p	594	sp|Q9C0Z1	Q9C0Z1	Putative transcriptional regulator	*S. cerevisiae*
Ykw2p	705	sp|P35995	P35995	Transcriptional regulatory protein	*S. cerevisiae*
Yl66p	701	sp|Q06149	Q06149	Putative transcriptional regulator	*S. cerevisiae*
Yl78p	1341	sp|Q05854	Q05854	Putative transcriptional regulator	*S. cerevisiae*
Yll054p	769	sp|Q12244	Q12244	Putative transcriptional regulator	*S. cerevisiae*
Ylr228p	814	sp|Q05958	Q05958	Putative transcriptional regulator	*S. cerevisiae*
Ymh6p	944	sp|Q03631	Q03631	Transcriptional regulatory protein	*S. cerevisiae*
Yn25p	743	sp|O59741	O59741	Putative transcriptional regulator	*S. cerevisiae*

Table A-1 –Continued

Fungal Zinc Finger	Sequence length	Database	References	Function	Organism
Yn92p	607	sp|P53749	P53749	Putative transcriptional regulator	*S. cerevisiae*
Yp33p	446	sp|P19541	P19541	Putative transcriptional regulator	*S. cerevisiae*
Ypr196p	470	sp|Q06595	Q06595	Maltose fermentation regulator	*S. cerevisiae*
Yrr1p	810	sp|Q12172	Q12172	Putative transcriptional regulator	*S. cerevisiae*
*Ca1p	419	embl|CAA20477.2	CAA20477.2	Transcriptional regulation	*S. pombe*
*Ca2p	632	embl|CAA20706.3	CAA20706.3	Transcriptional regulation	*S. pombe*
*Ca3p	510	embl|CAA21933.1	CAA21933.1	Transcriptional regulation	*S. pombe*
*Grt1p	648	sp|Q9C469	Q9C469	Zinc finger protein	*S. pombe*
*Spa1p	625	embl|CAB59617.1	CAB59617.1	Probable transcriptional regulator	*S. pombe*
*Spa2p	529	embl|CAB61777.1	CAB61777.1	Probable transcriptional regulator	*S. pombe*
*Spa3p	654	embl|CAC19742.1	CAC19742.1	Probable transcriptional regulator	*S. pombe*
*Spa4p	783	embl|CAB16735.1	CAB16735.1	Probable transcriptional regulator	*S. pombe*
*Spa5p	697	embl|CAC19729.1	CAC19729.1	Probable transcriptional regulator	*S. pombe*
*Spb1p	827	embl|CAA19035.1	CAA19035.1	Probable transcriptional regulator	*S. pombe*
*Spb2p	560	embl|CAA19036.1	CAA19036.1	Probable transcriptional regulator	*S. pombe*
*Spb3p	594	embl|CAA21917.1	CAA21917.1	Probable transcriptional regulator	*S. pombe*
*Spb4p	738	embl|CAA16906.1	CAA16906.1	Probable transcriptional regulator	*S. pombe*
*Spb5p	815	embl|CAA19174.1	CAA19174.1	Probable transcriptional regulator	*S. pombe*
*Spb6p	397	embl|CAA18884.1	CAA18884.1	Probable transcriptional regulator	*S. pombe*
*Spc1p	857	embl|CAA21815.1	CAA21815.1	Probable transcriptional regulator	*S. pombe*
*Spc2p	867	embl|CAA18305.1	CAA18305.1	Probable transcriptional regulator	*S. pombe*
*Spc3p	525	embl|CAA19070.1	CAA19070.1	Probable transcriptional regulator	*S. pombe*
*Suc1p	501	sp|P33181	P33181	Probable sucrose utilization protein	*S. pombe*
*Ta1p	480	pir|T38582	T38582	Hypothetical protein	*S. pombe*
*Ta2p	497	pir|T38582	T41718	T41718 hypothetical fungal Zn (2)-	*S. pombe*
Thi1p	775	embl|CAB62420.1	CAB62420.1	Thiamine biosynthesis protein	*S. pombe*
Yakbp	782	sp|Q09922	Q09922	Putative transcriptional regulator	*S. pombe*
Yao7p	603	sp|Q10086	Q10086	Putative transcriptional regulator	*S. pombe*

Table A-1 –Continued

Fungal Zinc Finger	Sequence length	Database	References	Function	Organism
Yas8p	563	sp|Q10144	Q10144	Putative transcriptional regulator	*S. pombe*
*Yhddp	618	sp|Q9P6I9	Q9P619	Putative transcriptional regulator	*S. pombe*
*Pro1p_S	693	embl|CAB89829.1	CAB89829.1	Transcriptional regulatory protein	*S. brevicollis*

Database abbreviations are shown with the accession number. Databases from which the sequences where retrieved included: SWISSPROT (sp), European Molecular Biology Laboratory (embl), Protein Information Resource (pir) and the DNA Databank of Japan (dbj). The references are found on the database. New annotations are highlighted using the asterisk symbol (*).