Comparative genomic studies of prion-protein family members

Nagesh Chakka

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

April 2008

Publication from this work

Chakka N, Gready JE. **Pathway to Functional Studies: Pipeline Linking Phylogenetic Footprinting and Transcription-Factor Binding Analysis**. CRPIT 73:15-21 (2006)

Statement of originality

The contents of this thesis are the result of my original research, which has been conducted under the principal supervision of Prof. Jill Gready (John Curtin School of Medical Research, ANU), and my advisor Dr. Anneke Blackburn (John Curtin School of Medical Research, ANU).

This thesis incorporates the outcome of experimental work performed by Tatiana Vassilieva:

Xenopus project (Chapter 3): Isolation of the *Xenopus* chimeric and non-chimeric *PRNP* and *PRND* transcripts based on my sequence predictions. Tissue expression experiments for various different transcripts (Chimeric, non-chimeric *PRNP* and *PRND,* and *PRNP2*).

Monodelphis and Platypus project (Chapter 4): Isolation of *PRNP* and *PRND* transcripts based on my sequence predictions.

Dr. Lorenzo Sangargio (Visiting Post Doc Fellow, University of Milan, Italy) was involved in the implementation of the phylogenetic footprinting pipeline used to predict transcription factor-binding sites in *SPRN* (Chapter 7) and the interpretation of the resulting data.

I declare that the work presented in this thesis is to my belief original, except as acknowledged above. Any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline, and that the material has not been submitted, either in whole or in part, for a degree at this or any other university.

Nagesh Chakka

Acknowledgments

I am most grateful to my supervisor, Prof. Jill Gready, and my advisor Dr. Anneke Blackburn for their guidance and constructive discussions. Their enthusiasm for science and inquisitive nature have set a clear example for me to follow.

Many others have contributed in smaller but important ways to this work. I would like to express my gratitude to Tatiana Vassilieva for helping me with the lab work and for sharing her knowledge in the area.

My special thanks to Dr Anna Cowan (Sub-dean Biomedical Sciences) for giving me support and advice during the course of my PhD.

The support and encouragement from my wife Poornima Nayak was very crucial in motivating me through the ups and downs of the project. I would also like to say a big ‘thank you’ to Hernan Alonso, Babu Kannappan and Sue Corley for their invaluable friendship.

I would like to thank the Australian National University for supporting this work performed in the Computational Proteomics Group of the Division of Molecular Bioscience, John Curtin School of Medical Research.

Nagesh Chakka

Abbreviations

Ap3: Activating protein-3

CJD: Cruetzfeldt-Jakob disease

CPE: Cytoplasmic polyadenylation element

CWD: Chronic wasting disease

Dpl: Doppel

EST: Expressed sequence tag

FET: Fischer's Exact test

FFI: Fatal familial insomnia

GEO: Gene Expression Omnibus

GPI: Glycosyl-phosphatidylinositol

GSS: Gerstmann–Straüssler–Scheinker

HMM: Hidden Markov Model

MSA: Multiple sequence alignment

NF-Y: Nuclear factor-Y

NMR: Nuclear Magnetic Resolution

ORF: Open reading frame

PCR: Polymerase chain reaction

Prnd: Down stream to prion protein gene (Mouse)

PRND: Downstream to prion protein gene

Prnp: Prion protein gene (Mouse)

PRNP: Prion protein gene

PrP: Prion protein

PrP^{C}: Cellular prion protein

PrP^{Sc}: Scrapie prion protein

RACE: Rapid Amplification of cDNA Ends

RFLP: Restriction fragment length polymorphism

Sho: Shadoo

shRNA: Short hairpin RNA

siRNA: Small interfering RNA

SNP: Single nucleotide polymorphism

SPRN: Shadow of Prion protein gene

Spz1: Spermatogenic zip

stPrP: Similar to prion protein

TF: Transcription factor

TFBS: Transcription factor-binding site

TSS: Transcription start site

USF-1: Upstream stimulating factor-1

vCJD: variant CJD

Abstract

Prion protein (PrP), doppel (Dpl) and Shadoo (Sho) encoded by the genes *PRNP*, *PRND* and *SPRN,* respectively, are suggested to be homologues, i.e. evolved from a common ancestor and, hence are grouped as the prion-protein family. PrP is associated with a group of diseases termed as spongiform encephalopathies. Dpl plays a role in spermatogenesis and *Prnd* knockout mice have been shown to be infertile. Sho is a recently discovered gene and although very little is known about its function, the current information suggests it is important in neural tube development (in zebrafish) and has a neuro-protective effect. There is very little information about these genes and their functions in lower vertebrates. A relatively large number of studies were undertaken to understand the transcriptional regulation of *PRNP* but very few for *PRND*. There have also been no studies to analyze the association of Dpl with human male infertility.

During my PhD studies I have applied a variety of computational and experimental methods to elucidate the function and evolution of these genes, particularly, for *PRNP* and *PRND,* by identifying and characterizing them in various vertebrate lineages. Computational predictions were made by comparative sequence analysis to identify the transcription factors (TFs) involved in the regulation of *PRNP* and *PRND*. One of the key predictions for *PRND* was experimentally validated by functional studies. Single nucleotide polymorphisms (SNPs) in the *PRND* coding and promoter regions were analyzed in the human male infertile population.

Genes of interest were identified and characterized in various vertebrate species (*Xenopus* species, chicken, *Monodelphis domestica*, and platypus). The results revealed the mechanisms of gene evolution, tissue subfunctionalization and protein sequence divergence for PrP and Dpl, which are a result of gene duplication. The sequences obtained were used for a range of comparative sequence analyses which revealed conserved residues likely to play a role in functional and structural stability.

Using a comparative sequence method termed phylogenetic footprinting I identified conserved transcription factor-binding sites (TFBS), and, thus, putative mechanisms for the transcriptional regulation of these genes. These computational predictions were performed by developing an automated pipeline for high throughput analysis with an interactive web interface to analyze the results. Some of the predictions are consistent with known information about these genes, including already experimentally validated TFBSs. One of the most interesting TFBS predictions for *PRND* was that of Spz1 (spermatogenic zip) which was predicted in the core promoter of the gene. Functional studies revealed a 20-30% increase in the promoter activity in cells transfected with Spz1 expression plasmid. A 40-50% downregulation of the promoter activity was observed when the Spz1 binding site was mutated. This suggests that Spz1 is a key TF and may influence the tissue-specific expression of *PRND* which is predominantly expressed in testis along with other experimentally validated ubiquitous TFs (USF-1 and NF-Y).

SNP/mutation analysis was performed by sequencing the coding and promoter region of *PRND* among a sample of the human infertile male population (96 samples) compared with a control population sample (healthy and fertile men) (96 samples). The SNPs in the coding region corresponded with those that have already been reported. There is no published information on SNPs in the upstream region to the transcription start site of *PRND.* My studies revealed four SNPs in the -282 region. Interestingly, I found a SNP in the 3’ end of the core binding site of the USF-1 (G-171A) only in the control populations. The C521T polymorphism within the open reading frame of *PRND* showed different allele frequencies in the control and infertile populations in which the TT genotype was more common in the control than in the infertile population. However, it is difficult to draw definitive conclusions about the association of the allele frequencies with infertility due to the size of the dataset

Table of contents

Publication from this work I
Statement of originality II
Acknowledgments III
Abbreviations IV
Abstract VI
Table of contents VIII
1 Prion protein family: Introduction 1
1.1 Prion protein family 1
1.2 Prion protein 3
1.2.1 Discovery of the gene coding for PrP 3
1.2.2 Tissue expression 4
1.2.3 Structure and properties of cellular and scrapie prion protein 5
1.2.4 Functions of PrP 6
1.2.5 The prion disease process 9
1.3 Doppel 11
1.3.1 Tissue expression pattern of Dpl 13
1.3.2 Structure of Dpl 14
1.3.3 Physiological functions of Dpl 16
1.3.4 PrP and Dpl in testis 17
1.4 Shadoo 19
1.5 Prion-protein family structural features 20
1.6 Motivation of the thesis 21
1.7 Aims of the thesis 22
2 Strategy and Methods 24
2.1 Overview 24
2.1.1 Homology search 27
2.1.2 Phylogenetic footprinting and transcription factor-binding site (TFBS) analysis 28
2.1.3 Functional studies of *Prnd* promoter 31
2.1.4 Polymorphism studies of human *PRND* 31
2.1.5 PCR-RFLP assay 32
3 Findings in *Xenopus* 33
3.1 Background 33
3.2 Material and methods 33
3.2.1 Database searches 33
3.2.2 Sequence Analysis 34
3.2.3 Homology modelling 35
3.2.4 Identification of conserved regions on mouse Dpl 35
3.2.5 Experimental analysis 35
3.3 Results and Discussion 37
3.3.1 Genomic characterisation of *X. tropicalis PRNP* and *PRND* 37
3.3.2 *X. laevis PRNP* and *PRND* 43
3.3.3 Generation of chimeric trancript 45
3.3.4 Non-chimeric transcripts in *X. laevis* 51
3.3.5 Duplicate of PrP in *X. laevis* 51
3.3.6 Tissue expression of *PRNP* and *PRND* in *X. laevis* 54

3.3.7 Genomic environment of *PRNP* in *X. tropicalis* 55
3.3.8 Sequence analysis of *Xenopus* Dpl 56
3.3.9 Comparison of C-terminal domain 59
3.3.10 Comparison of Dpl sequences 62
3.3.11 *Xenopus* Dpl model 62
3.4 Conclusion 64
4 Findings in chicken 67
4.1 Background 67
4.2 Materials and Methods 68
4.2.1 Computational analysis 68
4.2.2 Isolation and cloning of *Sho-like* gene 70
4.2.3 Emu and alligator search 71
4.2.4 Tissue expression 72
4.3 Results and Discussion 73
4.3.1 Analysis of the chicken *PRNP* genomic locus 73
4.3.2 Isolation and molecular characterization of *Sho-like* cDNA 74
4.3.3 Sequence analysis of Sho-like 75
4.3.4 Expression analysis of *PRNP*, *SPRN* and *Sho-like* transcripts in chicken 77
4.3.5 Comparison with the *PRNT* gene 78
4.3.6 Comparison of chicken *PRNP* genomic loci with other vertebrates ..78
4.3.7 Comparison of chicken and fish *PrP*-related genes 79
4.3.8 Identification of *PRNP* locus genes in emu and alligator 81
4.3.9 Comparison of PrP sequence from birds and other tetrapods 81
4.4 Absence of Dpl in chicken: Its implications and insights into the evolution of acrosome reaction 84
4.5 Conclusions 85
5 Findings in other species 87
5.1 Background 87
5.2 Monodelphis domestica 88
5.2.1 Homology search 88
5.2.2 Defining exon-intron boundaries in *M. domestica* 89
5.2.3 Alternative splicing leading to two different C-terminal PrP sequences 90
5.3 Platypus 91
5.3.1 Homology search 91
5.3.2 Sequence analysis 92
5.4 Miscellaneous searches 93
5.4.1 Ciona intestinalis 93
5.4.2 Axolotl 93
5.4.3 Human 93
5.5 Summary 97
6 Comparative sequence and structure analysis of prion protein and doppel....98
6.1 Background 98
6.1.1 Methods 99
6.2 Results and discussion 101
6.2.1 Rate of evolution of amino acid sequence 101
6.2.2 Distribution of conserved residues on the protein structure 102
6.2.3 Dotplot analysis 110

6.2.4 Evolution of functional elements in 3'UTR 111
6.3 Conclusion 111
7 Phylogenetic footprinting and transcription factor binding analysis of prion-protein family genes 113
7.1 Background 113
7.2 Motivation for developing the pipeline 115
7.3 Software and Hardware 115
7.4 Development of the pipeline 116
7.4.1 Criteria for species selection 116
7.4.2 Conserved-sequence motif detection 119
7.4.3 Searching against the TRANSFAC database for TF-binding specificity 129
7.5 Visual front end for data analysis 133
7.5.1 Interface development 133
7.6 Results from application of the pipeline to the PrP family genes 139
7.6.1 Selection criteria 139
7.6.2 Prion protein gene, *PRNP* 140
7.6.3 Doppel gene, *PRND* 141
7.6.4 Applying this pipeline to *SPRN* 142
7.7 Discussion of the results 148
7.7.1 Known *vs* predicted TFs 150
7.8 Application and Conclusions 154
8 Experimental validation of predictions 155
8.1 Background 155
8.2 Materials and Methods 156
8.2.1 Database searches 156
8.2.2 Cell lines 156
8.2.3 Plasmids 157
8.2.4 Transfection and luciferase reporter assays 158
8.2.5 Endogenous gene expression: RNA extraction and PCR analysis. 159
8.2.6 Mutation of TF binding sites 159
8.3 Results 160
8.3.1 Spz1 is only expressed in testis 160
8.3.2 Endogenous *Prnd* expression analysis in GC2 and bEnd.3 cells ... 161
8.3.3 Test for *Prnd* core promoter 162
8.3.4 Effect of Spz1 and Sp1 on promoter activity 163
8.3.5 Mutation of Spz1 and NF-Y binding sites decreases promoter activity 164
8.4 Discussion 167
9 Human *Doppel* gene polymorphisms in male infertility 168
9.1 Background 168
9.2 Materials and methods 169
9.2.1 Study population 169
9.2.2 Genotyping 169
9.2.3 PCR-RFLP assay for analysis of C521T polymorphism 173
9.2.4 Statistical analysis 173
9.3 Results 175
9.3.1 Analysis of promoter and ORF for infertility-associated mutations . 175
9.3.2 Analysis of SNPs in *PRND* 175

9.3.3 Analysis of association of SNPs with infertility 178
9.3.4 Genotype analysis 179
9.3.5 Haplotype analysis 179
9.4 Discussion 180
Summary and Final Remarks 182
References 185
Appendix 202

1 Prion protein family: Introduction

1.1 Prion protein family

The term “prion” was coined by Stanley B. Prusiner - short for “proteinaceous infectious particle” (Prusiner 1982). The intriguing aspect of prions is that the infectious particle appears to lack nucleic acid. This is further supported by the fact that prions are resistant to ultraviolet and ionizing radiation (Alper et al. 1967; Adams 1991) and that prion protein is encoded by the host genome. However, this theory is still challenged by others (Manuelidis et al. 2007). The molecular basis for this disease is associated with conformational change of the normal cellular prion protein (PrP^{C}) to a disease-associated form termed as scrapie (PrP^{Sc}). The main causation of disease is unclear

Prion diseases are a group of diseases termed “transmissible spongiform encephalopathies” (Ryou 2007). As the term suggests, these diseases can be transmitted, both within the species and across species, although with more difficulty (species barrier). Prion diseases are termed differently according to the mammals affected (Table 1-1). Disease progression results in a sponge-like appearance of the affected brain (Figure 1-1) and ultimately death.

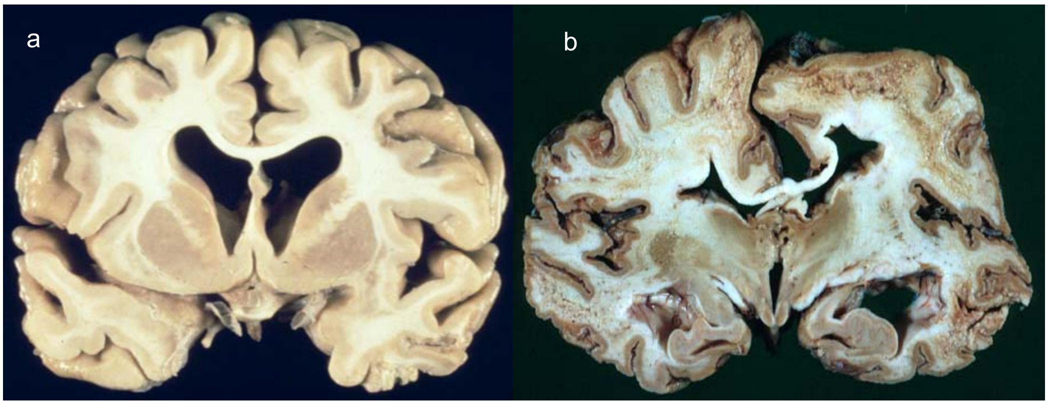


Figure 1-1 Longitudinal section of (a) normal human brain and (b) brain with spongiform encephalopathy. Image source

a: http://faculty.washington.edu/alexbert/MEDEX/Fall/adcoronalb.jpg.

b: http://www.neuropathologyweb.org/chapter5/chapter5ePrions.html

Table 1-1 Prion diseases

Species	Disease
Human	Cruetzfeld-Jakob disease (CJD), variant CJD (vCJD), Gerstmann–Straüssler–Scheinker (GSS), kuru, fatal familial insomnia (FFI)
Sheep	Scrapie
Cow	Mad cow disease or bovine spongiform encephalopathy (BSE)
Deer and elk	Chronic wasting disease (CWD)

Structural and sequence similarities to prion protein (Figure 1-2) of three different genes are used as the basis to group them as PrP family members (Moore et al. 1999; Premzl et al. 2003). These genes are *PRNP* which encodes prion protein (PrP); *PRND* which encodes doppel (Dpl); and a gene recently discovered in my group, *SPRN,* which encodes the PrP-like protein, Shadoo (Sho) (Premzl et al. 2003).

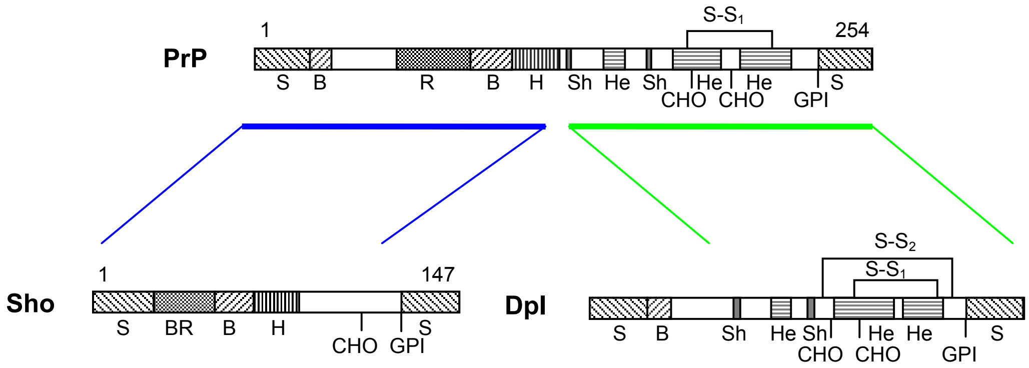


Figure 1-2 Protein sequence regions for PrP, Dpl and Sho (mouse). S, signal; B, basic; R, repeats; BR, basic repeats; H, hydrophobic; Sh, β-sheet; He, α-helix; S-S, disulphide bridge; CHO, N-glycan; GPI, GPI-anchor. The number indicates the sequence length.

1.2 Prion protein

1.2.1 Discovery of the gene coding for PrP

The identification of a protease resistant PrP$^{27\text{-}30}$ fragment which was present only in diseased individuals led to its isolation and purification (Bolton et al. 1982; Prusiner et al. 1982) The N-terminal region of purified PrP$^{27\text{-}30}$ was sequenced to reveal 17 amino acid residues (Prusiner et al. 1984). The oligonucleotide probe designed from the N-terminal sequence was used to screen a hamster cDNA brain library which revealed the gene (Oesch et al. 1985). The full length cDNA was isolated by Basler et al. (1986) who also sequenced the *PrP* gene region characterizing its gene structure. They proposed for the first time that this gene lacked a TATA box and had a potential Sp1 binding site which are characteristic features of a house keeping gene. This work established that PrP is encoded by the chromosomal gene and not by a nucleic acid within the prion. This gene was later designated as *PRNP* (Westaway et al. 1987). Despite its structural complexity and length (Figure 1-2), *PRNP* has an interesting organization in that the entire open reading frame (ORF) is contained within a (last) single exon though the entire gene comprises three exons (two exons in human and hamster).

1.2.2 Tissue expression

Human: Expression of *PRNP* is highly regulated during cellular differentiation and embryonic development (Dodelet and Cashman 1998). *PRNP* mRNA is found in neurones of the hippocampus, cortex, thalamus, cerebellum and medulla (McLennan et al. 2001). *PRNP* has also been reported in the enteric nervous system (Shmakov et al. 2000). All mononuclear leucocyte subpopulations and platelets express PrP^c, but polymorphonuclear leucocytes and red blood cells express little or none (Barclay et al. 1999). The presence of both the mRNA and the protein has been demonstrated in testicular tissue (Bendheim et al. 1992; Tanji et al. 1995). A recent study by Shaked et al. (1999) showed that PrP is also present on the membrane of sperm of different species, including humans. PrP has also been reported to be present on the surface of blood cells (Cashman et al. 1990).

Mice: PrP may have a role in neural tube development as it has been detected throughout the developing neural tube from 13.5 days of development in mouse embryos. It continues to be expressed in the adult brain, and is detected in the pyramidal and dentate granular cells of the hippocampus, Purkinje cells of the cerebellum and in large neurones of the cortex, medulla and septum (Kretzschmaret al., 1986; Manson et al., 1992). This suggests that PrP may be critical to neural function in the mammalian central nervous system. The pyramidal neurons of the hippocampus take part in the learning and memory process.

Prnp is also expressed at variable levels in different tissues (*Prnp* mRNA was detected at different levels in all tissues tested with the exception of kidney and liver) with the highest amount in brain (Ford et al. 2002; Miele et al. 2003). On a cellular level, *Prnp* was shown to be expressed in astrocytes and oligodendrocytes. PrP^{Sc} accumulates in astrocytes and other glial cells and later the misfolded isoform diffuses into other tissues (Moser et al. 1995).

Hamster: Tissue expression patterns in hamster indicate high concentrations of PrP in the brain. In non-neuronal hamster tissues, *PRNP* mRNA has been detected in circulating leukocytes, heart, skeletal muscle, lung, intestinal tract, spleen, testis, ovary, and some other organs (Bendheim et al. 1992).

Chicken: PrP is widely expressed in cholinergic and non-cholinergic neurons in the adult central nervous system and spinal cord. Using *in situ* hybridization experiments, the mRNA was detected as early as embryonic day 6 in brain, spinal chord, retina, intestine, and heart (Harris et al. 1993).

Sheep: Apart from the neuronal tissues, PrP is expressed in several non-neuronal tissues in sheep, including spleen, lymph node, lung, heart, kidney, skeletal muscle, uterus, adrenal gland, parotid gland, intestine, proventriculus, abomasum and mammary gland. However, it was not detected in the liver (Horiuchi et al. 1995).

1.2.3 Structure and properties of cellular and scrapie prion protein

PrP is a relatively short protein comprising if 254 amino acids in mouse. The characteristic features of this protein include (Figure 1-2) one disulphide bridge, two N-glycosylation sites in the C-terminal domain, an N-terminal basic region, and an N-terminal repeat region. The N-terminal repeat region varies in both repeat length (8 amino acids in mouse repeated 5 times) and sequence among various mammalian species and higher vertebrates (marsupials, birds, reptile). However, this region is totally missing in amphibian (Strumbo et al. 2001). The most remarkable feature is the highly conserved middle hydrophobic sequence. The N-terminal and C-terminal ends contain signal sequences. The N-terminal signal sequence is an endoplasmic reticulum targeting signal peptide, (Hope et al. 1986), and the C-terminal signal peptide is replaced with a GPI-anchor (Stahl et al. 1987). Both the signal peptides are cleaved in the mature protein which consists of amino acid residues 23-231. According to solution NMR structure analysis of PrP^{C}, the N-terminal domain is highly flexible and lacks a definitive secondary structure (Riek et al. 1997) and the C-terminal domain folds to form a globular domain (Riek et al. 1996) (Figure 1-3a). The latter comprises three helices and a pair of short β-strands with a single disulphide bond. This is in contrast with low-level structural information of PrP^{Sc} as determined by Fourier transform infrared spectroscopy

(Gasset et al. 1993) which shows it is rich in β-sheet (54% β-sheet and 25% α-helix). Though these two forms of protein are encoded by the same gene, their molecular properties are very different (Table 1-2).

Table 1-2 Comparison between cellular and scrapie prion protein

PrP^C	PrP^{Sc}
Secondary structural units comprising the C-terminal domain: High α-helix content (40%)	High in β-sheet (54%)
Present on the cell surface	Accumulates within the cells (lysosomes or endosomes)
Molecular weight of full length protein: 30-35 kda	Molecular weight: 30-35 kda
Digested completely when treated with protease	Incomplete digestion leading to a fragment of 27-30 kda (PrP^{27-30})
Encoded by host chromosomal gene *PRNP*	Encoded by host chromosomal gene *PRNP*
Soluble in detergent	Insoluble

1.2.4 Functions of PrP

The normal functions of PrP are not well understood despite considerable efforts to elucidate them. The studies carried out to date suggest that PrP may perform a large range of normal functions.

1.2.4.1 Knockout mouse studies

Prnp ablation studies (confined to the coding regions) which were performed to study the normal functions of the gene produced mice which appeared developmentally and behaviorally normal (Bueler et al. 1992; Manson et al. 1994). Changes in membrane localization of nitric oxide synthase were observed in some of the *Prnp* knockout mice (Ovadia et al. 1996). Other *Prnp* knockout mice showed minor phenotypic defects, however, the cerebellar cells displayed increased sensitivity for copper toxicity (Brown et al. 1998). All these features are also characteristic of scrapie-infected animals (Keshet et al. 1999). The phenotypic differences observed between different knockout studies has been attributed to the strategy used for creating the knockout mice (Weissmann and Flechsig 2003).

Loss of PrP function in knockout mice may have been compensated by adaptive mechanisms operative in embryo masking any effect (Mallucci et al. 2002). To overcome this possibility, a post-natal knockout experiment was designed (Mallucci et al. 2002). These studies concluded that post-natal loss of PrP in adult neurons has no detrimental sequelae (Mallucci et al. 2002). However, the main observation from these studies is alteration in the hippocampal CA1 (*Cornu Ammonis1*) properties (Mallucci et al. 2002).

Mice devoid of PrP did not develop prion disease (Bueler et al. 1993).

1.2.4.2 Effects on learning, memory and synaptic functions

Based on the *Prnp* knockout mice studies, it was demonstrated that PrP^C is involved in learning and memory (Nishida et al. 1997). Impairment of long-term potentiation, a form of synaptic plasticity thought to be important for memory formation, was observed in some of the knockout mice using electrophysiological studies (Collinge et al. 1994).

Some of the *Prnp* knockout mice showed minor neurophysiological effects which included defects in synaptic (Collinge et al. 1994; Manson et al. 1995) and hippocampal and normal synaptic function (Colling et al. 1995). It has been proposed that synaptic proteins play a role in Alzheimer's disease and other neurodegenerative disease (Masliah 2001). This is of some interest as PrP is one of the synaptic proteins. *PRNP* was shown to be overexpressed in neuritic plaques seen in Alzheimer's disease (Ferrer et al. 2001).

It has also been demonstrated through the knockout mice studies that PrP^C is involved in circadian rhythm (Tobler et al. 1996).

1.2.4.3 Antioxidant activity and copper binding

The octarepeats of PrP bind copper ions and may play a role in oxidative stress (Hornshaw et al. 1995; Brown et al. 1997), as evidenced by the fact that the *Prnp* knockout mice neurons were found to be more vulnerable to oxidative stress (Brown et al. 1997). Based on these observations, it was proposed that PrP may play a role in the modulation of neuronal excitability (Brown et al. 1998). PrP^c has also been found to be involved in the regulation of presynaptic copper concentration (Herms et al. 1999). This may help explain why prion diseases are connected with a dramatic loss of antioxidant defense, where the presence of the abnormal protein during prion disease causes a failure of cellular antioxidant defense (Brown 2005). It was also speculated that PrP^c, due to its multiple copper-binding sites may function by activating a copper-dependent antioxidant enzyme such as superoxide dismutase (Brown et al. 1997; Brown et al. 1999). However, the *in vivo* studies show that the role of PrP^c as a dismutase itself is not significant (Hutter et al. 2003).

A protective role against copper-induced damage has been suggested for PrP in mouse spermatozoa (Shaked et al. 1999).

1.2.4.4 Other functions

PrP^c is also involved in the regulation of intracellular calcium concentrations (Colling et al. 1996), activation of lymphocytes (Mabbott et al. 1997), signal transduction (Mouillet-Richard et al. 2000; Bounhar et al. 2001; Spielhaupter and Schatzl 2001; Chiarini et al. 2002) and has antioxidant and antiapoptotic properties (Bounhar et al. 2001).

PrP^c may play a role in skeletal muscle physiology (Massimino et al. 2006). It has been demonstrated that transgenic mice over expressing wild-type PrP develop skeletal muscle, peripheral nerve and central nervous system degeneration (Westaway et al. 1994). PrP^c may also help in the differentiation of human leukocytes (Dodelet and Cashman 1998).

Some of these functions may be disrupted by the conversion of PrP^c to PrP^{Sc}.

A number of PrP-binding proteins have been identified and a possible disruption to these interactions in prion disease may be one of the contributing factors for the pathogenesis. These binding proteins include: antiapoptotic protein Bcl2 (Kurschner and Morgan 1995), caveolin (Gorodinsky and Harris 1995; Harmey et al. 1995), the laminin receptor precursor (Rieger et al. 1997), plasminogen (Fischer et al. 2000) and N-CAMs (neural cell adhesion molecules) (Schmitt-Ulms et al. 2001).

Nerve growth factor (NGF) was shown to regulate the expression of the PrP gene in early post natal development (Mobley et al., 1988). NGF plays a role in CNS, sensory and sympathetic neuronal survival (Williams et al., 1986; Kromer, 1987).

1.2.5 The prion disease process

The uniqueness of prion diseases from other neurodegenerative diseases like Alzheimer's and Parkinson's disease is its transmissible nature. The spread of Kuru, a form of human prion disease, was attributed to a cannibalistic ritualism among a tribe in Papua New Guinea; this stopped after the ban of the ritual in 1958 (by the Australian authorities) (Mead et al. 2003). However, owing to the long incubation period of the prion disease, some fresh cases are still being reported. This was the first evidence of human-to-human spread, providing the impression of an infectious disease. However, there were also indications of a genetic basis for this disease from clusters of cases in families, the first characterized being familial Creutzfeldt - Jakob disease (CJD), which were attributed to mutations in the *PRNP* gene. Known familial prion diseases; Gerstmann-Sträussler-Scheinker (GSS), CJD, and fatal familial insomnia (FFI); have been linked to mutations on this gene (Table 1-3). Among the human prion diseases, however, the most prevalent form is sporadic CJD where the actual cause is unknown. It was proposed that this form of the disease may be caused by spontaneous somatic mutation in the *PRNP* gene (Prusiner 1989) or by a spontaneous post-translational conformational change (Aguzzi 2006). There have been reported cases of iatrogenic CJD caused by corneal transplantation, EEG electrode implantation or by contaminated surgical instruments, dura mater grafts, and pericardium grafts. Patients on human growth hormone therapy prior to 1982 are also of greater risk of CJD infection as it was derived from pituitary glands of cadavers. Variant CJD (vCJD) was first described in the United Kingdom in 1996. The disease is strongly linked to the consumption of cattle products infected with the prions that causes BSE, or mad cow disease.

Table 1-3 Familial prion disease (Ironside 1998).

Class of prion disease	Reported mutation
Creutzfeldt-Jakob disease	E200K, D178N, V210I, V180I, T183A, H208R, M232R
Gerstmann-Straussler-Scheinker	P102L, A117V, P105L, Y145STOP, F198S, Q217R
Fatal familial insomnia	D178N in combination with a polymorphism of Met at position 129

Prion disease begins with the conversion of PrP^c to PrP^{Sc} by formation of a PrP^c/PrP^{Sc} complex (Prusiner et al. 1990). There is evidence implicating the N-terminal repeats and the hydrophobic region in facilitating the conformational change (Smith et al. 1997; Prusiner 1998). Although by NMR studies it was shown that the region up to the hydrophobic region is disordered, the protease resistant core of PrP^{Sc} includes the hydrophobic region (Huang et al. 1995). The interaction between PrP^C and PrP^{Sc} and conversion to PrP^{Sc} is also facilitated by other protein molecules (Kocisko et al. 1994). The misfolded isoform, PrP^{Sc}, is an infectious form, capable of horizontal transmission of the disease, although a species barrier exists for the transmission of the disease. This barrier is associated with the amino acid sequence at a specific region, which may influence the interaction between PrP^c and PrP^{Sc} (Scott et al. 1989).

As the majority of *Prnp,* knockout mice did not show phenotypic or behavioral changes, it seems likely that the lack of PrP function is not the primary cause of the symptoms seen in the prion disease. This suggests that some cellular process or processes are being impaired by the accumulation of PrP^{Sc}. It has also been shown that the amount of PrP^{Sc} is directly related to the onset but not the severity or state of the prion disease (Bueler et al. 1994; Manson et al. 1994). Forloni et al. (1993) have shown that cerebral accumulation of PrP^{Sc} and its degradation products play a role in nerve cell degeneration in prion disease.

As PrP^{c} and PrP^{Sc} share the same epitope, there is no specific cellular or humoral immune response (Porter et al. 1973; Kingsbury et al. 1981). No antibody against prions can be detected in experimentally infected animals (Porter et al. 1973; Kingsbury et al. 1981; Berg 1994). The disease process is not associated with any inflammatory reaction, but leads to spongiform degeneration accompanied by gliosis and neuronal loss (Budka et al. 1995).

The susceptibility of an individual to prion disease based on genetic background of that individual has long been known (Dickinson and Stamp 1969). Adding to the complex nature of this invariably fatal prion disease, there is no diagnostic test available to detect prion disease in humans. The genetic forms of prion disease can be identified by screening the *PRNP* gene sequence for known mutations. However, this is not applicable to sporadic forms of CJD. Mutation/polymorphisms in the functional non-coding regions, i.e. regulatory regions, may also contribute to the disease process but this has not been characterized. The presence of amyloid plaques in the brain of the affected individual is not a common feature but, when present, is diagnostic of prion disease.

1.3 Doppel

One set of *Prnp* knockout mice experiments, conducted in anticipation that the mice deficient in PrP might develop a phenotype which could give an insight into the normal functions of PrP, produced mice which developed late-onset ataxia accompanied by Purkinje cell degeneration (Sakaguchi et al. 1996). Further experiments suggested this phenotype might be a result of upregulation of a previously unknown gene (Moore et al. 1999; Li et al. 2000). A suspicion that a protein with function overlapping that of PrP might be clustered with it led investigators to sequence downstream of the murine *Prnp* gene*.* They discovered an ORF encoding a PrP-like protein 16kb downstream to *Prnp* (Moore et al. 1999). This gene was designated as downstream prion protein-like or *Prnd* and the protein as doppel (German for ‘double’) (Moore et al. 1999). As for PrP, the entire ORF is found in a single exon although the gene is comprised of 2-3 exons. Dpl was the first PrP-like protein to be described in mammals. It shows structural similarities to the C-terminal two-thirds of PrP with about 25% sequence identity. Dpl has an N-terminal signal sequence indicating it is a secretory protein and a C-terminal signal sequence for GPI-anchor attachment, and also two N-glycosylation sites (Silverman et al. 2000) (Figure 1-2). Both the signal sequences, as in PrP, are cleaved to form the mature protein. The sequence and structural similarities between PrP and Dpl (Figure 1-3) and the organisation of these two genes are indicative of a gene duplication event (Moore et al. 1999).

Prnp knockout mouse lines produced by different laboratories had shown phenotypes ranging from minor electrophysiological and circadian rhythm disturbance to one with cerebral purkinje cell degeneration causing late-onset ataxia (Moore et al. 1999). These latter findings were interpreted in terms of the upregulation of *Prnd* from the promoter of *Prnp*. In one of the several *Prnp* knockout mice lines, it was shown that Dpl was upregulated in neuronal tissues, causing neurodegeneration (Moore et al. 1999; Li et al. 2000). This phenotype was rescued by the introduction of *Prnp* (Nishida et al. 1999), suggesting an antagonistic function of Dpl to PrP (Behrens 2003; Qin et al. 2006). This indicates that these genes, at least in higher vertebrates, would have developed regulatory control mechanisms to produce a differential expression pattern. *Prnp* knockout mice showed two different phenotypic variants, one with minor defects (Zrch *Prnp*$^{0/0}$ and Edbg *Prnp*$^{-/-}$) (Bueler et al. 1992; Manson et al. 1994) whereas the other group (Ngsk *Prnp*$^{-/-}$, Zürich II and Rcm0) developed late onset ataxia (Sakaguchi et al. 1996; Moore et al. 1999; Rossi et al. 2001). This behaviour was attributed to Purkinje cell degeneration caused not by the absence of PrP but due to over expression of Dpl as a result of intergenic splicing with its expression controlled by the *Prnp* promoter (Moore et al. 1999). This form of splice variant was also found in low abundance in wild-type mice (Moore et al. 1999).

Unlike PrP, there is no evidence of conformational changes in the folded domain of Dpl. Also, despite the structural and sequence similarities between PrP and Dpl, there is no evidence so far either directly implicating Dpl in prion disease or indicating that its regulation is altered in prion disease (Tuzi et al. 2002) even though Dpl was shown to be expressed in spleen which is a major reservoir for PrP (Li et al. 2000). However, it should be noted that the major sequence differences between PrP and Dpl is the lack of N-terminal repeats and the highly conserved hydrophobic region in Dpl. A chimeric mouse protein, composed of the N-terminal domain of PrP^C (residues 23–125) and the C-terminal part of Dpl (residues 58–157) lead to the formation of a β-sheet-rich form of Dpl with partial resistance to pepsin proteolysis *in vitro* (Erlich et al. 2008)

1.3.1 Tissue expression pattern of Dpl

The tissue distribution of Dpl has been studied in human (Peoc'h et al. 2002), mice (Li et al. 2000), and cattle and sheep (Tranulis et al. 2001). In mice, Dpl is widely expressed during embryogenesis and in the brain of newborn mice (Li et al. 2000). It is also expressed at high levels in adult testis and heart and detected at low levels in adult brain (Moore et al. 1999; Li et al. 2000). In cattle and sheep, Dpl is strongly expressed in the testes and at low levels in ovary and spleen (Tranulis et al. 2001). In humans, Dpl shows very restricted distributions limited to the male genital tract (Peoc'h et al. 2002). It is reported to be present in the seminiferous tubules, at the adluminal pole of Sertoli cells, spermatozoal extracts, seminal fluid and mature ejaculated spermatozoa (Peoc'h et al. 2002). Peoc’h et al. (2002) propose that as Sertoli cells communicate with the germ cells either directly (cell-cell interaction) or indirectly (paracrine interaction) throughout gametogenesis, it seems likely that Dpl is involved in spermatogenesis. Dpl is also thought to be involved in motility of sperm due to its presence on the flagellum of ejaculated spermatozoa. This supports the findings of Behrens et al. (2002) of poor sperm motility in mice lacking Dpl. Dpl has not been detected on testicular spermatozoa but is present on epididymal spermatozoa, suggesting that this protein is acquired after passage of the maturing spermatozoon through the epididymis (Serres et al. 2006). Dpl is also reported to be expressed during embryogenesis (Moore et al. 1999; Li et al. 2000). The expression pattern of the goat *PRND* suggest that it is involved in early testes development (Kocer et al. 2007). The studies performed on the ovine testicular tissues strongly support a role for Dpl in the later stages of spermatogenesis, in particular the final remodelling and maturation of elongated spermatids (Espenes et al. 2006).

1.3.2 Structure of Dpl

Similarly to PrP, Dpl is also a short protein comprising 179 amino acid residues in mouse. The structural topology of the Dpl folded domain (residues: 26-157) determined by solution NMR is similar to PrP (Mo et al. 2001) (Figure 1-3b). The secondary structure elements are located at the same regions in the primary sequence (Figure 1-3c).The differences between the two are:

- the presence of a kink in the second α-helix in Dpl
- the plane of the β-strands (β1 and β2) are parallel to the α2 axis in Dpl but perpendicular in PrP.
- the α3 helix is significantly shorter in Dpl.
- Dpl has two disulphide bridges compared with only one in PrP.

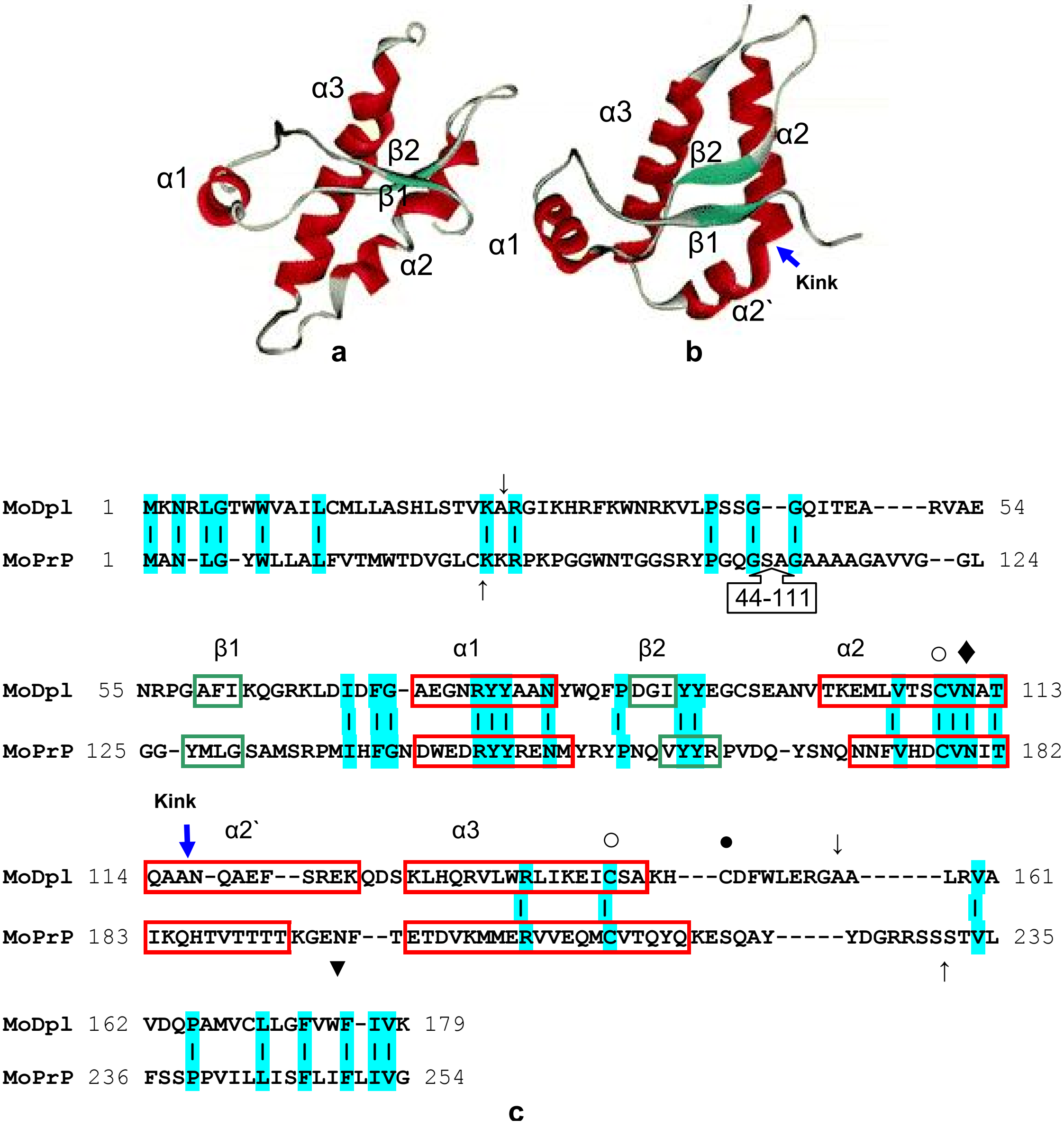

α3
α1
β2
β1
α2
a
α3
β2
α2
α1
β1
Kink
α2`
b
MoDpl 1 MKNRLGTWWVAILCMLLASHLSTVKARGIKHRFKWNRKVLPSSG--GQITEA----RVAE 54
MoPrP 1 MAN-LG-YWLLALFVTMWTDVGLCKKRPKPGGWNTGGSRYPGQGSAGAAAAGAVVG--GL 124
44-111
β1
α1
β2
α2
○ ♦
MoDpl 55 NRPGAFIKQGRKLDIDFG-AEGNRYYAANYWQFPDGIYYEGCSEANVTKEMLVTSCVNAT 113
MoPrP 125 GG-YMLGSAMSRPMIHFGNDWEDRYYRENMYRYPNQVYYRPVDQ-YSNQNNFVHDCVNIT 182
Kink
α2`
α3
○
●
MoDpl 114 QAAN-QAEF--SREKQDSKLHQRVLWRLIKEICSAKH---CDFWLERGAA------LRVA 161
MoPrP 183 IKQHTVTTTTKGENF--TETDVKMMERVVEQMCVTQYQKESQAY-----YDGRRSSSTVL 235
▼
MoDpl 162 VDQPAMVCLLGFVWF-IVK 179
MoPrP 236 FSSPPVILLISFLIFLIVG 254
c

Figure 1-3 Tertiary structure of (a) Mouse PrP (PDB ID: – 1AG2) and (b) Mouse Dpl (PDB ID: – 1I17) from solution NMR. (c) Structure based sequence alignment of mouse PrP with mouse Dpl. Note the well conserved secondary structural units of PrP and Dpl. Residues identical to mouse PrP and Dpl are highlighted in light blue. α-Helical regions (α1, α2, α2′ and α3) are indicated by red boxes and β-sheet by blue boxes. A kink in Dpl α2 helix is indicated by a blue arrow. A region containing the octarepeats that has no equivalent in Dpl has been removed from this alignment for clarity. Consensus glycosylation sites are represented by ♦; Cys involved in disulphide bridge formation are represented by ● and ○; Cleavage site of signal peptide ↑↓ (Mo et al. 2001; Mastrangelo et al. 2002).

1.3.3 Physiological functions of Dpl

Behrens et al (2002) conducted studies to elucidate the physiological role of Dpl by generating homozygous mutant mice lacking Dpl and replacing it by a neomycin resistant gene *Prnd*$^{neo/neo}$. Both male and female *Prnd*$^{neo/neo}$ mice displayed normal growth indicating no obvious effects of Dpl deficiency on growth and development. But a significant finding was that the adult male *Prnd*$^{neo/neo}$ mice were infertile, in contrast to the adult female *Prnd*$^{neo/neo}$ mice, which were fertile. The sexual behaviour of *Prnd*$^{neo/neo}$ mice was normal, as evidenced by the normal number of copulation plugs. The number of spermatozoa in the cauda epididymis of *Prnd*$^{neo/neo}$ males was shown to be reduced to 50% compared with the wild-type controls. Also the motility of mutant sperms was reported to be significantly reduced. Thus the cause of the infertility in the knockout mice is due to altered spermatogenesis rather than its sexual behaviour apparently due to lack of Dpl. The mature sperm in *Prnd*$^{neo/neo}$ mice showed several structural abnormalities ranging from disorganisation of the flagellum with respect to the sperm head to the sperm head being severely malformed and lacking a well developed acrosome (Figure 1-4). The latter defect was proposed to be the main reason for the sterility. This was supported by a follow up experiment in which partial success was achieved with *in vitro* fertilisation of wild-type oocyte whose zona pellucida was partially dissected. As Dpl is a GPI-anchored protein (Silverman et al. 2000), it was proposed to be present on the acrosomic vesicles through its GPI-anchor and possibly participates in the acrosome reaction.

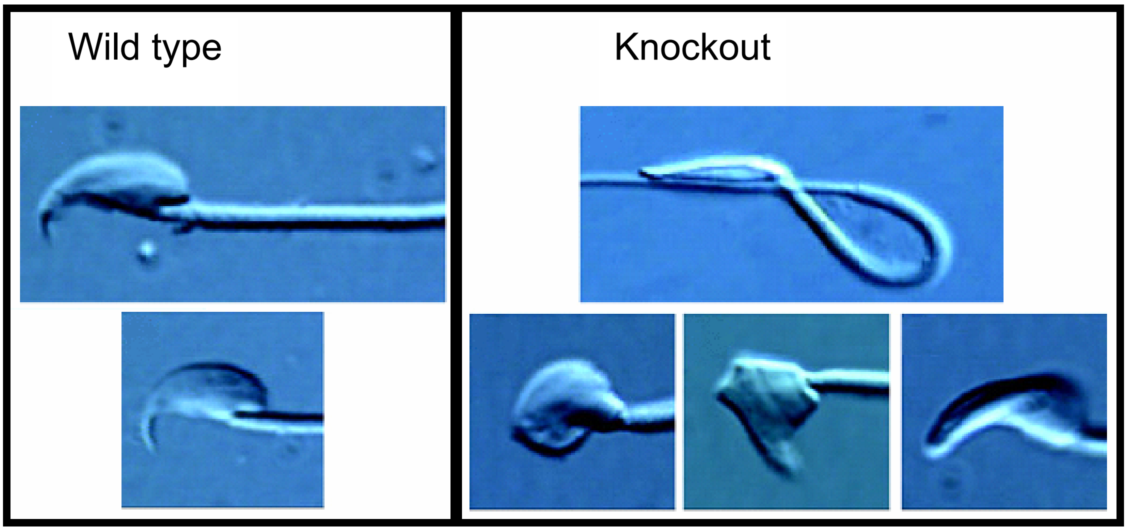


Figure 1-4 Sperm morphology as seen under microscope in wild-type mice and in *Prnd* knockout mice. Image taken from Behrens et al. 2002.

Another study by Paisley et al. (2004) generated *Prnd*$^{-/-}$ and *Prnp*$^{-/-}$ *Prnd*$^{-/-}$ mouse lines. Similar to the results of Behrens et al. (2002), the mutant mice developed normally but the male mutant mice were sterile. However, the sperm from both mutants showed normal concentration, motility and morphology and the mutant spermatozoa were able to fertilise *in vitro* but at a significantly reduced frequency compared with wild-type mice. However, most of the embryos did not reach late stage development, which was linked to observed DNA damage within the spermatozoa. Their studies supported the earlier findings that Dpl is involved in acrosome reaction, but they also hypothesised that Dpl may play a role in protecting DNA from oxidative damage in the sperm/testis.

It has been proposed that, as Dpl is a glycosylated protein, it may be involved in active protection of spermatozoa by reducing interactions between cells (Peoc'h et al. 2002).

1.3.4 PrP and Dpl in testis

Consistent with it its wide tissue distribution pattern PrP is also expressed in testis. However, unlike Dpl, PrP knockout (*Prnp*$^{-/-}$) mice are fertile (Behrens et al. 2002) indicating that Dpl function is not replaceable by PrP.

PrP found on ejaculated spermatozoa was reported to be a C-terminally truncated isoform (Shaked et al. 1999) but this could not be verified by Peoc'h et al. (2002). Interestingly the latter authors detected N-terminally truncated isoforms on ejaculated spermatozoa which were GPI-anchored and thought to result from proteolytic cleavage which begins in the epididymal fluid and greatly increases during ejaculation. They proposed that, in contrast to the full length form detected in seminal plasma, the absence of full length PrP on ejaculated spermatozoa indicates either the absence of transfer of PrP from seminal plasma to the spermatozoa or, if the transfer occurred, that they are rapidly proteolysed.

In situ hybridisation studies in mice showed *Prnp* transcripts in spermatogenic cells, but not in somatic cells such as Sertoli cells, Leydig cells, and peritubular myoid cells, suggesting that PrP plays a role in spermatogenesis (Fujisawa et al. 2004). They observed *Prnp* mRNA moderately in spermatogonia, strongly in spermatocytes and round spermatids, but not in elongated spermatids and spermatozoa. These findings may indicate a functional role for PrP in spermatogenesis.

Testis-specific prion protein (*PRNT*): Recently a testis-specific prion protein was reported in the genomic environment of *PRNP*, 3 kb downstream to *PRND* in human and this gene was referred to as *PRNT* (Figure 1-5) (Makrinou et al. 2002). *PRNT* has three alternative spliced variants (varying size of non-coding exon 1) that were detected only in adult testis (Makrinou et al. 2002). It has been proposed that this gene resulted from a part of the duplication event early during eutherian speciation (Makrinou et al. 2002). However, this gene does not share sequence similarity with any of the prion-protein family genes. Also the gene is not well conserved in mammalian species, including mouse and cow.

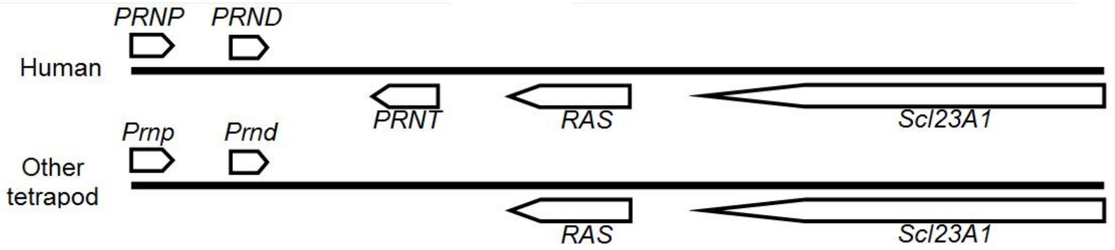


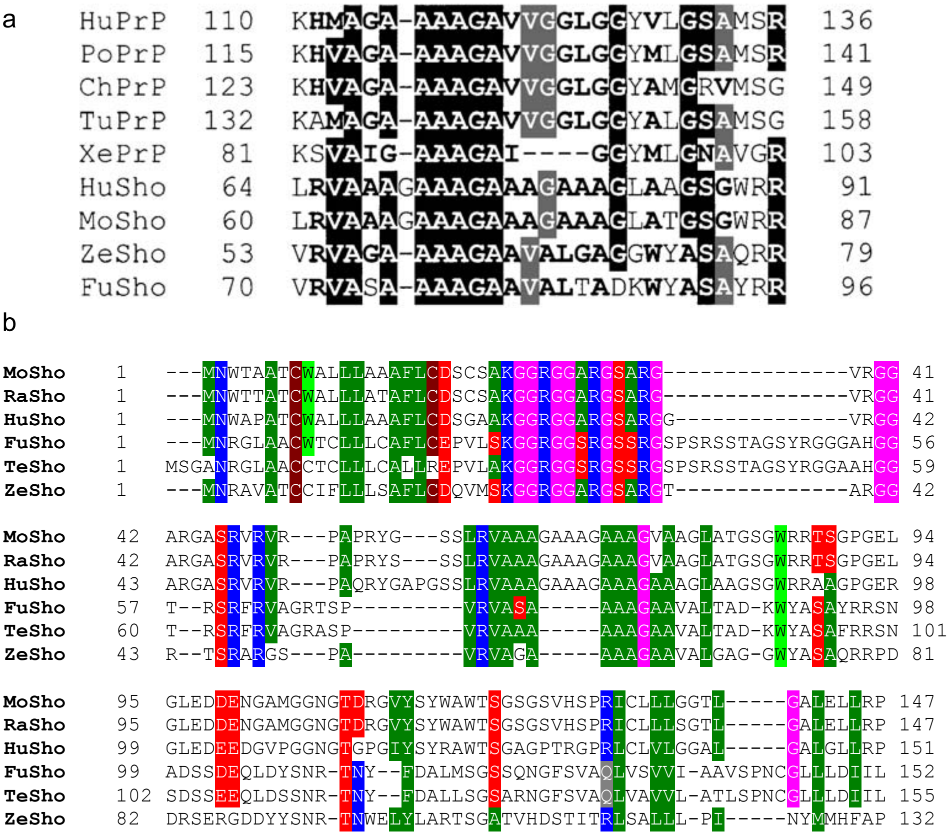
Figure 1-5 Genomic environment of *PRNP* in human and other tetrapod lineages. Note the additional gene *PRNT* in human present downstream to *PRND* and in the opposite orientation to that of *PRNP* and *PRND*.

1.4 Shadoo

Premzl et al. (2003) recently reported another PrP-like protein, Shadoo (Sho) (Japanese shadow) encoded by the gene designated as *SPRN* (shadow of prion protein). This gene has been found on a separate chromosome (No. 7 in mice) away from the *Prnp-Prnd* gene complex (present on chromosome 2 in mice). *SPRN* is expressed in embryo, whole brain, and retina. Sequence comparison shows a highly conserved N-terminal signal sequence, Arg-rich basic region, a hydrophobic region with strong homology to prion protein (Figure 1-6a), a C-terminal domain containing a conserved glycosylation motif, and a C-terminal signal sequence for GPI-anchor attachment (Figure 1-2). However, Sho lacks the disulphide bridge and has only one N-glycosylation site. It is highly conserved in mammals and also well conserved from fish to mammals (Figure 1-6b). As for the other prion-protein family genes, the *SPRN* ORF is contained within a single exon.

The functions of Sho are not well understood but it is likely to play a role in CNS development as demonstrated by gain and loss functional experiments (RNAi and overexpression) in zebrafish (Sangiorgio et al. 2007). *SPRN* has been shown to be present in CNS from early postnatal life and was proposed to have a neuroprotective effect (Watts et al. 2007). Interestingly, prion-infected mice demonstrated reduction in endogenous Sho protein (Watts et al. 2007). This suggests that Sho not only shares sequence similarity but also a number of biochemical and cell biological properties with PrP (Watts et al. 2007) and supports the initial suggestions that PrP and Sho have overlapping functions (Premzl et al. 2003).

An apparent duplicate of Sho, represented as Sho2 and encoded by the gene *SPRNB,* is found only in fish (Premzl et al. 2004; Strumbo et al. 2006). However, whole genome duplication in modern fish (ray finned) (Taylor et al. 2003) complicates analysis of the origin of the gene.



```
a   HuPrP   110   KHMAGA-AAAGAVVGGLGGYVLGSAMSR   136
    PoPrP   115   KHVAGA-AAAGAVVGGLGGYMLGSAMSR   141
    ChPrP   123   KHVAGA-AAAGAVVGGLGGYAMGRVMSG   149
    TuPrP   132   KAMAGA-AAAGAVVGGLGGYALGSAMSG   158
    XePrP    81   KSVAIG-AAAGAI----GGYMLGNAVGR   103
    HuSho    64   LRVAAAGAAAGAAAGAAAGLAAGSGWRR    91
    MoSho    60   LRVAAAGAAAGAAAGAAAGLATGSGWRR    87
    ZeSho    53   VRVAGA-AAAGAAVALGAGGWYASAQRR    79
    FuSho    70   VRVASA-AAAGAAVALTADKWYASAYRR    96
b

MoSho    1   ---MNWTAATCWALLLAAAFLCDSCSAKGGRGGARGSARG---------------VRGG 41
RaSho    1   ---MNWTTATCWALLLATAFLCDSCSAKGGRGGARGSARG---------------VRGG 41
HuSho    1   ---MNWAPATCWALLLAAAFLCDSGAAKGGRGGARGSARGG--------------VRGG 42
FuSho    1   ---MNRGLAACWTCLLLCAFLCEPVLSKGGRGGSRGSSRGSPSRSSTAGSYRGGGAHGG 56
TeSho    1   MSGANRGLAACCTCLLLCALLREPVLAKGGRGGSRGSSRGSPSRSSTAGSYRGGAAHGG 59
ZeSho    1   ---MNRAVATCCIFLLLSAFLCDQVMSKGGRGGARGSARGT--------------ARGG 42

MoSho    42  ARGASRVRVR---PAPRYG---SSLRVAAAGAAAGAAAGVAAGLATGSGWRRTSGPGEL 94
RaSho    42  ARGASRVRVR---PAPRYS---SSLRVAAAGAAAGAAAGVAAGLATGSGWRRTSGPGEL 94
HuSho    43  ARGASRVRVR---PAQRYGAPGSSLRVAAAGAAAGAAAGAAAGLAAGSGWRRAAGPGER 98
FuSho    57  T--RSRFRVAGRTSP---------VRVASA-----AAAGAAVALTAD-KWYASAYRRSN 98
TeSho    60  T--RSRFRVAGRASP---------VRVAAA-----AAAGAAVALTAD-KWYASAFRRSN 101
ZeSho    43  R--TSRARGS---PA---------VRVAGA-----AAAGAAVALGAG-GWYASAQRRPD 81

MoSho    95  GLEDDENGAMGGNGTDRGVYSYWAWTSGSGSVHSPRICLLLGGTL-----GALELLRP 147
RaSho    95  GLEDDENGAMGGNGTDRGVYSYWAWTSGSGSVHSPRICLLLSGTL-----GALELLRP 147
HuSho    99  GLEDEEDGVPGGNGTGPGIYSYRAWTSGAGPTRGPRLCLVLGGAL-----GALGLLRP 151
FuSho    99  ADSSDEQLDYSNR-TNY--FDALMSGSSQNGFSVAQLVSVVI-AAVSPNCGLLLDIIL 152
TeSho    102 SDSSEEQLDSSNR-TNY--FDALLSGSARNGFSVAQLVAVVL-ATLSPNCGLLLDIIL 155
ZeSho    82  DRSERGDDYYSNR-TNWELYLARTSGATVHDSTITRLSALLL-PI-----NYMMHFAP 132
```

Figure 1-6 (a) Alignment of PrP and Sho hydrophobic region showing reasonable sequence conservation (image taken from Premzl et al. 2003)**. (b) Multiple sequence alignment of Sho sequences in Mammals and Fish.** The conserved regions are highlighted**.** Mo, mouse; Ra, rat; Hu, human; Fu, *Fugu;* Te, *Tetraodon*; Ze, zebrafish; Fu, *Fugu*; Po, *Monodelphis*; Ch, chicken; Xe, *Xenopus*. Color coding is based on identity or similarity between the amino acid residues.

1.5 Prion-protein family structural features

All proteins in the family are secreted as indicated by presence of a well defined signal peptide at the N-terminal region, and are embedded in the extracellular membrane as indicated by C-terminal signal sequence for GPI-anchor attachment (Figure 1-2). PrP and Dpl have approximately 25% sequence identity and Dpl protein resembles an N-terminally abbreviated PrP^{C} protein lacking the octamer repeats. No region of Dpl has significant homology to the middle hydrophobic region of PrP, which is present in all known PrPs and has strong sequence conservation. PrP has two N-glycosylation sites (of the form N X T). Dpl also has two N-glycosylation sites but only one of these (the one on the second helix, Figure 1-2) is maintained in the same position as in PrP. PrP has two cysteines involved in disulphide bridge formation between helix two and helix three, which are also conserved in Dpl. In addition, Dpl has another disulphide bridge as shown in Figure 1-2. Sho has a very well conserved hydrophobic region as for PrP. The alignment of the hydrophobic segment shows strong conservation across all PrPs and Shos (Figure 1-6) (Premzl et al., 2003). The common feature in all of these proteins is that they are encoded by a single exon. It has been suggested that these genes were a result of a gene duplication event of an ancestral gene which was *SPRN*-like and have diverged sufficiently to have different functions (Premzl et al. 2004).

1.6 Motivation of the thesis

My main interest is in studying the evolution, function, tissue expression and regulation of PrP family genes. Identifying this group of proteins in newly sequenced genomes and also looking for unidentified remote homologues will give a better understanding of the evolutionary and functional characteristics of this family of proteins.

There are not many studies on the tissue expression of these genes in various lineages. Understanding the regulation of these genes in different vertebrate groups can help in understanding the evolution of regulatory mechanisms.

The Sp1 transcription factor (TF) has been shown experimentally to play a role in transcriptional regulation of *PRNP* (Saeki et al. 1996; Baybutt and Manson 1997; Inoue et al. 1997; Mahal et al. 2001). Mahal et al. (2001) also found Ap1 and Ap2 binding sites in the human *PRNP* promoter region. Nagyova et al. (2004) experimentally validated the role of USF-1 and NF-Y in *PRND* promoter activity. *PRNP* expression has been shown to be highly regulated during development (Manson et al. 1992). Although *PRNP* is expressed in a wide range of tissues, it is found at different levels in various tissues (high levels in neuronal tissues, intermediate levels in heart and lung and low levels in spleen) (Oesch et al. 1985). Similarly, Dpl is widely expressed during embryogenesis but has limited tissue expression in adult mice (high levels in adult testis and heart and detected at low levels in adult brain) (Moore et al. 1999; Li et al. 2000). Tissue-specific regulation cannot be mediated by the ubiquitous TFs like Sp1, Ap1, and Ap2 for *PRNP,* and USF-1 and NF-Y for *PRND*. It could be mediated by a cumulative regulatory control at the genetic (transcriptional regulation) level by tissue-specific TFs or at epigenetic levels (chromatin remodeling, methylation etc), or by a combination of these two mechanisms. The functions of these genes, which are not well known, can be approached by predicting the TFs involved in regulating their tissue expression.

It has been demonstrated that mutations/single nucleotide polymorphisms (SNPs) of genes involved in spermatogenesis causes human infertility (Nishimune and Tanaka 2006). As Dpl is involved in spermatogenesis, it would be interesting to study the association of Dpl in human male infertility.

1.7 Aims of the thesis

Aim 1: To identify and characterize prion-protein gene family members in newly sequenced genomes, and also to look for potential other members in already studied vertebrate groups. Based on both published members and those that have recently been found, it can be hypothesized that a variable repertoire of family members may exist in different vertebrate groups, or for particular species branches within a given vertebrate group. Chapters 3 to 6 discuss findings in various species.

Aim 2: To define the regulatory mechanisms for the major vertebrate genes (*PRNP*, *PRND* and *SPRN*) by identifying the regulatory binding sites, and correlating these with tissue and cell expression results. Chapter 7 deals with the prediction of transcription factor-binding sites (TFBSs) among the prion-protein family genes. Chapter 8 examines the experimental validation of one of the predictions in the *PRND* promoter.

Aim 3: To identify SNPs in *PRND* among the human population with male infertility. Although polymorphisms of the *PRND* gene have been studied in the context of its possible association with prion disease, there have been no reports of *PRND* polymorphisms among the male infertile population. Chapter 9 reports an initial study of SNPs observed in a sample of the human infertile and control populations.

2 Strategy and Methods

2.1 Overview

The genes, *PRNP*, *PRND*, and *SPRN* are suggested to be homologues, i.e. evolved from a common ancestor (Premzl et al. 2004). *PRNP* has been reported in various mammalian (van Rheede et al. 2003), marsupial (Windl et al. 1995; Premzl et al. 2005), reptilian (Simonic et al. 2000), avian (Harris et al. 1991; Gabriel et al. 1992; Wopfner et al. 1999) and amphibian species (Strumbo et al. 2001). *PRND* was suggested to be a result of duplication of *PRNP* (Moore et al. 1999). So far, it has been reported only in eutherian mammals. *SPRN* is well conserved between fish and mammals (Premzl et al. 2003). Fish are reported to have other genes homologous to prion-protein (PrP), coding for stPrPs (Oidtmann et al. 2003), PrP-like (Suzuki et al. 2002) and Sho2 (*SPRNB*) (Premzl et al. 2004), but none seems to be orthologous to higher vertebrate *PRNP*. Interestingly, all these genes contain a single-exon ORF.

The regulatory proteins termed transcription factors (TFs) orchestrate spatial and temporal transcriptional regulation of gene expression. TFs can be predicted by identifying the conserved sites to which they bind using a comparative genomics technique termed phylogenetic footprinting (PF) (Tagle et al. 1988). The increased amount of vertebrate genomic sequence information now available, and becoming available from genome sequencing, provides the opportunity for more sensitive (reliable, i.e. fewer false positives) searches for these conserved TF binding sites (TFBS). Previous work in my group employed a similar technique to predict TFBS for *SPRN* using sequence information from *Takifugu rubripes* (Pufferfish), mouse and human (Premzl et al. 2004) and for *PRNP* using tammar wallaby, mouse and human (Premzl et al. 2005). Subsequent genome sequencing projects have produced further sequence information from *Xenopus tropicalis*, chicken, *Monodelphis domestica* (gray short-tailed opossum), *Macropus eugenii* (tammar wallaby), and *Ornithorhynchus anatinus* (platypus), providing an opportunity to extend these studies using species across different evolutionary timelines (Figure

2-1). Sequence conservation in non-coding DNA among species separated by a wider evolutionary distance (for example, mammals and amphibians) is expected to be more significant than conservation observed across shorter timelines (for example between human and mouse). It is hypothesized that such conserved motifs have functional significance, including acting as a TFBS protein binding sites. Binding of particular regulatory proteins to these motifs could act to direct transcription in a tissue-specific fashion. Tissue specificity information on TFs can be obtained from TF databases. Thus, prediction of which TFs might bind to a particular gene can rapidly provide initial insights into potential functions of the gene, based on known modes of actions of the TFs in regulating other, better characterized, genes. Such initial predictions can greatly assist in designing focused experiments to define mechanisms for expression of the gene.

Thus, identifying and characterising (sequencing) the genes of interest in various lineages allows facilitates more robust TFBS predictions. As a major focus of this thesis (Figure 2-2) is to characterise genes of interest in different vertebrate lineages and developing an automated pipeline for TFBS analysis, only a limited amount of confirmatory experimental work is performed and only the most interesting TF predictions for *PRND* are tested by functional studies.

The possible association of polymorphisms in human *PRND* with infertility is investigated by genotyping studies taking into consideration both the ORF and part of the promoter region suggested by the results of the TFBS analysis. This work is motivated by the infertility observed in male *Prnd* knockout mice.

The following gives a brief outline of the methods used in the thesis. More detailed descriptions are given in the methods sections in the relevant chapters.

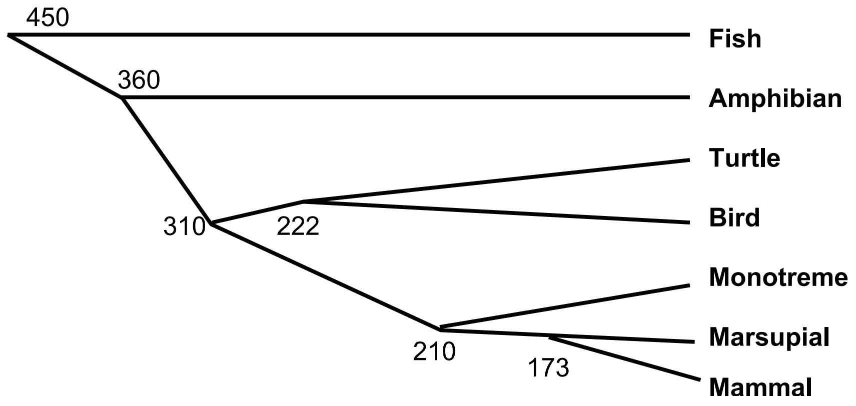


Figure 2-1 The molecular evolutionary timescale. The numbers indicate the years of separation in millions (Kumar and Hedges 1998).

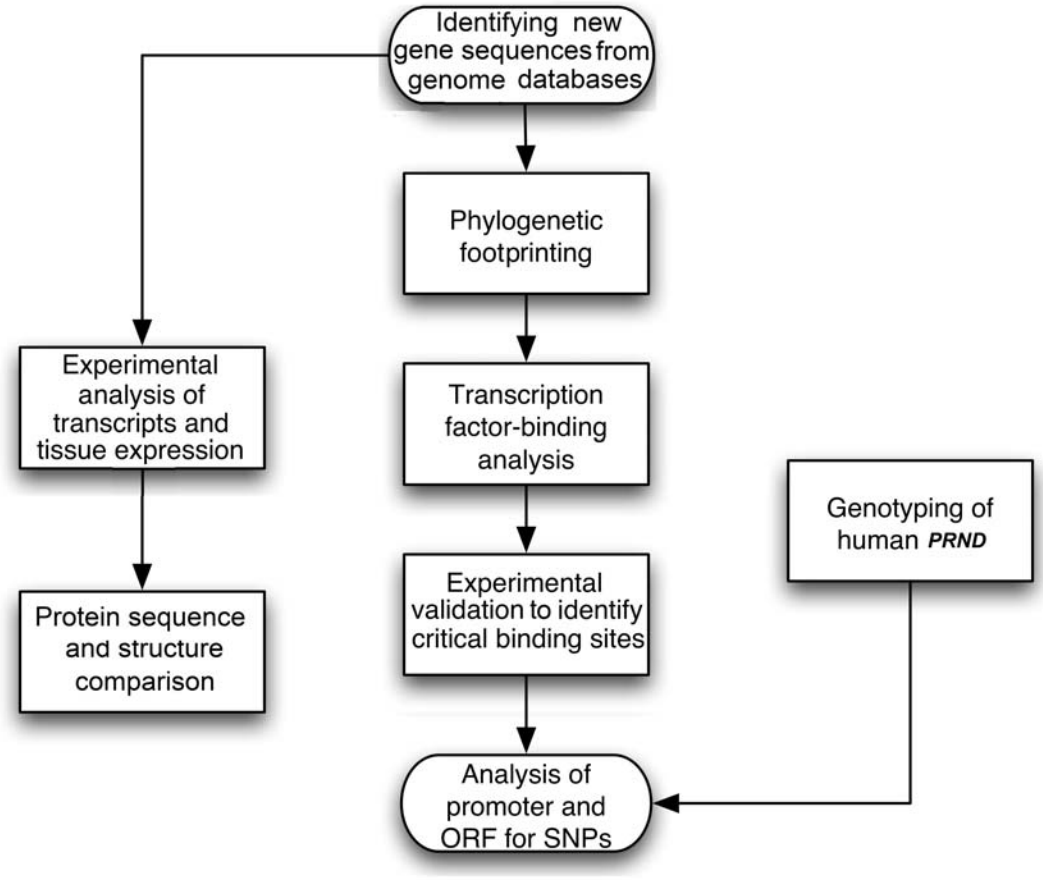

Identifying new gene sequences from genome databases
Phylogenetic footprinting
Transcription factor-binding analysis
Experimental validation to identify critical binding sites
Analysis of promoter and ORF for SNPs
Experimental analysis of transcripts and tissue expression
Protein sequence and structure comparison
Genotyping of human *PRND*

Figure 2-2 Summary of the workflow covering the main aims of the thesis.

2.1.1 Homology search

Similarity in DNA or protein sequences between individuals of the same species or among different species is termed homology. Homology search can be performed against single sequence information using methods such as BLAST (Altschul et al. 1990) or against sequence profile information using methods such as the Hidden Markov Model (HMM) (Eddy 1998).

To identify genes of interest in species selected for comparative studies, first a computational prediction of the coding region was performed followed by experimental characterization of the cDNA (Figure 2-3). The program TBLASTN (Altschul et al. 1997) was employed to scan the draft sequences to identify the coding regions of the genes. Genomic DNA sequences for both completely and partly sequenced organisms can be obtained from public repositories. Where the trace sequences were unassembled, they were downloaded and assembled into a local BLAST database. In most instances, the genomic sequence was also downloaded and a BLAST searchable local database was created for easy and fast searching. EST databases were also screened to identify transcripts of interest already deposited in the public repositories. Based on the computational findings, experiments to obtain the cDNA ends using the RACE (Rapid Amplification of cDNA Ends) technique was employed. Experimental work was carried out only for non-mammalian species (Chapters 3, 4 and 5).

The annotation of the genomic sequence to define the exon-intron boundaries, transcription start site (TSS), and ORF was performed using the cDNA sequence information which was obtained either from the experimental procedures or from the NCBI Reference Sequence (RefSeq) / non-redundant nucleotide database. The TSS, exon-intron boundaries and ORF were defined by mapping the cDNA sequence onto the genomic sequence using the program “est2genome” from the EMBOSS application (Rice et al. 2000). The ORF was annotated using the program “getORF” also from the EMBOSS application. As experiments on mammalian species were not performed, available EST database were used to annotate the genomic sequences for mammalian sequences. Also fish PrPs are not considered to be true PrP and also lack Dpl. Fish sequences were not included in the comparative analysis. A Perl module termed createBlastdb.pl was developed to handle the process of creating the BLAST database.

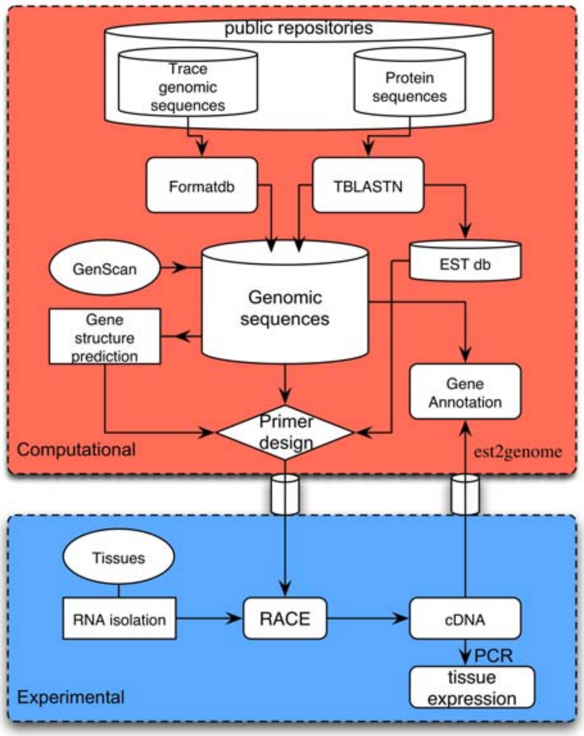


Figure 2-3 Flow diagram summarizing the steps involved in gene finding. These were divided into computational and experimental components. The end result of the analysis is the annotated gene sequence information.

2.1.2 Phylogenetic footprinting and transcription factor-binding site (TFBS) analysis

As TFBSs are under greater selective pressure than other non-protein-coding DNA, the reliability of predicting them is greatly improved by comparative genomics to filter out noise from genetic drift. Identifying such conserved sequence elements in non-coding regions of homologous genes from phylogenetic comparison is called ‘phylogenetic footprinting’ (Tagle et al. 1988). As TFBSs are short DNA motifs of 5-15 bp, analyzing a single sequence would lead to a very high percentage of false positive hits (Figure 2-4), and even more so if the considerable sequence variation between functional binding sites tolerated by most TFs is taken into account. Phylogenetic footprinting offers a solution to this problem by identifying such sequence elements that are conserved among genes that are either orthologous or co-expressed. Due to low overall similarity of non-coding regions across moderate evolutionary distances, many alignment algorithms will fail to produce biologically meaningful alignments. This can be improved by carefully manipulating the alignment parameters. There are also non-alignment methods to perform phylogenetic footprinting. Several programs implement phylogenetic footprinting but only a few combine it with TFBS analysis, for example, rVISTA (Loots et al. 2002) and ConSite (Sandelin et al. 2004). Chapter 7 deals with developing a pipeline for performing TFBS analysis using alignment and motif-discovery based methods.

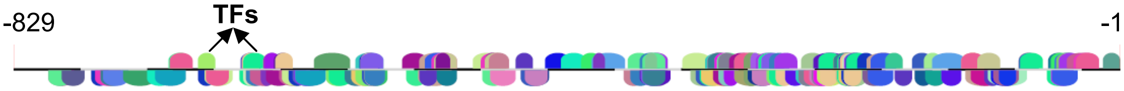


Figure 2-4 Output from MatInspector (Quandt et al. 1995) for single (human) *PRNP* promoter sequence (829 bp upstream to TSS). Note the number of large, unrealistic, TFBSs predicted for such a small region.

2.1.2.1 Phylogenetic footprinting using alignment method

DNA sequence alignment can be of two types, local and global. Local alignments are based on optimal subregion sequence similarity which would enable the detection of rearrangement events. Global alignments are based on optimal alignment over the entire length of sequences being compared. This method is better when the region compared is between orthologous genes. For initial analysis, I used global alignment methods implemented by AVID and LAGAN.

The AVID (Bray et al. 2003) alignment method is fast, memory efficient, and practical for sequence alignments of large genomic regions up to a megabase. AVID performs the pairwise alignment of two input sequences; the output comprises the alignment and additional information. The alignment files were used for downstream processing. LAGAN (Brudno et al. 2003) is a method for rapid global alignment of two homologous sequences. The algorithm is based on three main steps (Brudno et al. 2003): (1) generation of pairwise local alignments, (2) construction of a rough global map, by linking a subset of local alignments, and (3) computation of the final global alignment. LAGAN alignments were generated using the translate anchor option and binary output format was selected, which enables downstream processing.

2.1.2.2 Phylogenetic footprinting using motif-discovery method

The motif-discovery method is a non-alignment method which identifies conserved motifs based on specific word sizes that are overrepresented in the regulatory regions of orthologous genes. FootPrinter (Blanchette and Tompa 2003) uses the motif-discovery approach in identifying conserved motifs in a collection of homologous sequences. FootPrinter takes phylogeny into account and, hence, weighs the sequence based on the evolutionary relationship and implements most of the concepts of phylogenetic footprinting, in contrast to other motif-discovery methods such as MEME (Bailey and Elkan 1994). BioProspector (Liu et al. 2001) identifies motifs that are overrepresented in the input sequences and, hence, is a different approach to handling this problem.

2.1.2.3 TFBS databases

The conserved sequence motifs identified by phylogenetic footprinting need to be assessed for TF-binding specificity. TFBSs are degenerate sequence motifs that are recognized by TFs. TFBSs can be represented as a position weight matrix (PWM) or a position-specific scoring matrix (PSSM) (Stormo 2000). Databases of experimentally-validated TFBSs such as JASPAR (Sandelin et al. 2004) and TRANSFAC (Matys et al. 2003) are available. The largest and most commonly used collection of these matrices is the TRANSFAC database.

2.1.3 Functional studies of *Prnd* promoter

Reporter genes can be used to assay the activity of a particular promoter in a cell. The reporter gene is placed under the control of the target promoter and the reporter gene product's activity is quantitatively measured. The luciferase reporter gene with *Prnd* promoter attached and expressed in eukaryotic cells can used to assess the transcriptional activity of the region being tested. The most interesting predictions in the mouse *Prnd* core promoter were tested by this method (Chapter 8).

2.1.4 Polymorphism studies of human *PRND*

Nucleotide differences known as single nucleotide polymorphism (SNP) are the most common type of genetic variation (1 in every 1,200 bases). Each such variation in a chromosomal region is termed an allele, and a collection of alleles in a person's chromosomes is known as a genotype. SNPs can exist anywhere in the genome including the coding region. However, due to the degeneracy of the genetic code not all SNPs in the coding region produce an amino acid change. Other consequences of SNPs are in altering gene splicing, TF binding, or the sequence of non-coding RNA. SNPs are associated with the etiology of many human diseases or can affect susceptibility to human diseases. One of the most important applications in identifying SNPs is to compare populations with and without a particular disease.

The *PRND* ORF and promoter regions of DNA from healthy and infertile men were screened to analyze any possible association of SNPs in causing male infertility. Identification of SNPs to determine the genotype of a particular individual can be performed by a number of methods (Chapter 9).

2.1.4.1 Idaho technology LightScanner®

LightScanner® is designed to perform rapid high-throughput Hi-Res Melting^{TM} of amplified nucleic acids for mutation scanning or genotyping. This assay is performed on post-PCR amplicons. The double stranded DNA from the PCR is saturated with DNA binding dye (LCGreen®). A high-precision denaturation is performed on the amplicons by slowly heating until fully denatured and the fluorescence is recorded generating melting profiles. These melting profiles can be used to detect the sequence variants within the amplicons (Wittwer et al. 2003). This method was assessed for its suitability for analyzing the *PRND* gene to identify SNPs among the control and infertile population using this method.

2.1.4.2 Big-Dye sequencing

DNA sequencing determines the nucleotide order of a given DNA fragment. This can be performed using the chain termination method where the DNA to be sequenced acts as a template molecule. Sequencing was performed on 96-well plates to determine the genotype of the infertile and control samples. The chromatogram obtained from sequencing was analyzed using Sequencher to identify the SNPs.

2.1.5 PCR-RFLP assay

Restriction fragment length polymorphism (RFLP) analysis is used to identify SNPs that alter or create a site where a restriction enzyme cuts. By performing a digestion on a PCR product and determining fragment lengths through a gel assay it is possible to ascertain whether or not the enzymes cut the expected restriction sites. Those samples which needed further confirmation following the sequencing reaction were analyzed using this technique.

3 Findings in *Xenopus*

3.1 Background

Xenopus belongs to class amphibian. Although *X. laevis* has been favored by biologists for studying embryonic development, it is tetraploid, and also takes 1-2 years to reach sexual maturity. Hence *X. tropicalis*, which is the only diploid species in the *Xenopus* genus, was used as the model organism for genome sequencing (by the Joint Genome Institute).

PRNP has been reported in *X. laevis* (Strumbo et al. 2001). *X. laevis* PrP has all the sequence and structural motifs found in higher vertebrate PrPs except for the lack of the N-terminal repeats and exhibit some variation in residue composition in the highly conserved hydrophobic stretch. Although a true homologue of *PRNP* is not found in fish, diverged duplicated homologues and gene loci have been reported in ray-finned fish (Oidtmann et al. 2003; Premzl et al. 2004), most likely resulting from a whole-genome duplication (Taylor et al. 2003). Neither of these duplicated *PRNP* loci of fish contain the *PRND* gene. So far, *PRND* is reported only in mammals. In order to analyze the features of *PRND* in early vertebrates, I investigated the prion-protein family genes in *Xenopus* species (*X. tropicalis* and *X. laevis*). As *X. laevis* is tetraploid, it provides an opportunity to investigate whether any duplicates of PrP family genes have been retained.

3.2 Material and methods

3.2.1 Database searches

Several attempts of searching for Dpl in the *X. tropicalis* genome trace reads database (http://genome.jgi-psf.org/xenopus) using TBLASTN (Altschul et al. 1990) (default parameters) and various (full length) mammalian sequences as query did not produce any significant hits. The query sequence from human Dpl containing the N-glycosylation sites and the first three Cys residues involved in the disulphide bridges (HumDpl 57-140) picked up a hit from trace read TKS687278.g1, which corresponded to a stretch of the ORF of 159 residues but lacked the complete N-terminal region. The BLAST parameters used were a cutoff evalue of 1, default gap penalties, gapped BLAST program and low complexity regions permitted. A search with the nucleotide sequence from the 5' end of TKS687278.g1 (using BLASTN) gave another overlapping contig AAWU3059.g1 to complete the ORF for *X. tropicalis* Dpl. *X. tropicalis* PrP was identified using the *X. laevis* PrP sequence (Strumbo et al. 2001) as query. With the later release of assembled versions of the genome data, including the present version 3 (http://genome.jgi-psf.org/Xentr3/Xentr3.home.html), I was able confirm the findings obtained from trace reads and was also able to define the genes surrounding *PRNP* and *PRND*.

Screening the EST library (NCBI EST database and Sanger EST database http://www.sanger.ac.uk/cgi-bin/blast/submitblast/x_tropicalis) for Dpl in *X. laevis* and *X. tropicalis* did not produce any positive hits but showed multiple hits for PrP for both species. Both BLASTN and TBLASTN programs with default BLAST parameters were used.

3.2.2 Sequence Analysis

Prediction of Signal sequence: N-terminal signal peptide cleavage sites for *X. tropicalis* and *X. laevis* PrPs and Dpls were predicted using the SignalP program (version 3.0) (Nielsen et al. 1997) (http://www.cbs.dtu.dk/services/SignalP/). The program Big-pi (Eisenhaber et al. 1999) (http://mendel.imp.univie.ac.at/sat/gpi/gpi_server.html) was used to predict GPI modification sites; modification sites for *X. tropicalis* and *X. laevis* PrPs were identified but clear sites for *X. tropicalis* and *X. laevis* Dpls were not predicted.

Sequence alignment: Multiple sequence alignments were performed using web-based ClustalW (Higgins 1994) (http://www.ebi.ac.uk/clustalw/) using default parameters, and then manually edited using BioEdit Version 5.0.9 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), taking into account results obtained using pairwise comparison of protein structures using the DaliLite alignment program (Holm and Park 2000) (http://www.ebi.ac.uk/DaliLite/).

Calculating the percentage identity was performed using the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch 1970) provided by EMBOSS (http://www.ebi.ac.uk/emboss/align/).

3.2.3 Homology modelling

Homology modeling for *X. laevis* PrP2 and *X. tropicalis* Dpl was made using the package MODELLER (MODELLER 6v2) (Sali and Blundell 1993). A representative structure from mouse Dpl (PDB ID 1I17) and *X. laevis* PrP (PDB ID 1XU0) were used as templates for *X. tropicalis* Dpl and *X. laevis* PrP2, respectively, and iteratively refined for 4 cycles.

3.2.4 Identification of conserved regions on mouse Dpl

Dpl sequences from human, mouse, cow, sheep and *Xenopus* (*X. laevis* and *X. tropicalis*) were used to construct a multiple sequence alignment. Degree of conservation was plotted using ConSurf (Version 2.0) (Glaser et al. 2003) onto the surface of mouse Dpl 3D structure (PDB 1I17). The output was saved as a RasMol coloring script source for coloring the protein according to the conservation grades. The output was later modified to suit the requirements of PyMol (Version 0.97) (DeLano 2002), which generates high quality molecular graphic images.

3.2.5 Experimental analysis

3.2.5.1 Isolation and cloning of cDNA

RNA from adult *X. tropicalis* brain for *PRNP* and testis for *PRND* was obtained as a gift from Dr Timothy C. Grammer (Harlands lab, University of California, Berkeley). *X. laevis* tissues were salvaged from animals maintained locally in a colony in JCSMR, ANU by Dr David Tremethick's group and were stored in RNA*later*® (Ambion) until use. The Qiagen RNAeasy kit was used to extract total RNA. cDNA synthesis using 1 µg of RNA was performed using the Invitrogen Superscript™ III First-Strand cDNA Synthesis kit using random hexamers. RACE (Rapid Amplification of cDNA Ends) was undertaken in the 5’ and 3’ directions using the SMART™ PCR cDNA Synthesis Kit (BD bioscience) and using gene specific primers and 5' universal primers gene specific primers (Experiments performed by Tatiana Vassilieva. Appendix 1). Platinum®*Taq* DNA polymerase (Invitrogen) was used for PCR analysis (50 ng of DNA template and 4 pmol of primer) with PCR conditions 94°C for 30sec, 57°C for 30sec, 72°C for 1min. Only the primers I have used myself in this study are listed in Table 3-1. The PCR product was cloned into TOPO-TA Cloning® kit (Invitrogen). Sequencing reactions were performed at the Biomolecular Resource Facility, John Curtin School of Medical Research using the reagents and protocol provided by them.

Table 3-1 Summary of primers used in RT-PCR assays

Primer ID	Sequence
LP1	GCTTCTCATTTGCCTTCCTG
RP1	TGAGGGTATAGAGTGTGCCAAA
LP2	TCCTGAACCTCCCCCTGTA
LP3	CGGATCACGAGCTTCTCATT
RP2	TGAGAGGATAAGTAAGCCCAAA
LP4	TGCACATTGACTGTATCTTCCA
RP3	CTGATTACGGGGAAAAGACC
RP4	CAGTCACGACCACCCTTTG
RP5	TTCTAACCCATGGGCTGATCC
RP6	TGTTGAAATTGCTTCCTC
LD5	CACACTCCCAGACAAGAGCA
LD6	TTTTCAGGGCAAAGCAGAGT
RD4	TACGTTCTGCCTTTCCATTCTG
RD5	TGTGGTCTCTCTTGTCCGAG
RD6	GAGGACAGGAGAATGGGCCAC

3.2.5.2 Tissue expression

The following tissues were used for analyzing the tissue expression pattern: brain, testis, kidney, liver, lung, muscle, gut, spleen, skin, stomach, eye, heart, pancreas and gall bladder. These experiments were performed by Tatiana Vassilieva in my group. Primer pairs for tissue expression analysis were (see Appendix 1, Table 3-1) F4/R4 for *PRNP1-a*, LP1/R4 for *PRNP1-b*, LP3/R6 for *PRNP2-b(1)*, F7/R6 for *PRNP2-b(2)*, F4/R7 for *PRND-a*, LD6/R7 for *PRND-b(1*), F10/R7 for *PRND-b(2)* and F11/R10 for *GAPDH*.

3.3 Results and Discussion

3.3.1 Genomic characterisation of *X. tropicalis PRNP* and *PRND*

Two variant forms of *X. tropicalis PRND* which differ in their 3’ UTR length were identified (Figure 3-1, Figure 3-2). *PRND* has four exons with a total of 1272 bp for the longer variant and 762 bp for the shorter form. *X. tropicalis PRNP* also has two variant forms which differ in the 5’ UTR, with a total length of 834 and 869 bp for shorter and longer variants, respectively, distributed over three exons (Figure 3-1, Figure 3-2). The shorter exon 1 is 45 bp and the longer exon 1 is 80 bp (Figure 3-2). The significant difference in *PRND* gene organization between *Xenopus* and that of higher vertebrates is that exon 1 and exon 2 in *PRND* and *PRNP* are common for the two genes (Figure 3-2b, c). Such a transcript with shared exons is termed as chimeric. The intron separating exon 2 from exon 3 in *PRND* is 27,099 bp. This region contains the exon 3 for *PRNP (*Figure 3-2h). The entire ORF for both *PRND* and *PRNP* is contained within the last single exon. The 3’ UTRs in *X. tropicalis* are much shorter than their mammalian counterparts with 536 bp in *PRND* (human 2633 bp; mouse 1337 bp) and 48 bp in *PRNP* (human 1591bp; mouse 1233 bp).

(a) *X. tropicalis PRNP* (chimeric)

```
    1
  1 actgctagtcccactgccctcctgatctgcctggcACATCCATCCTCAGTCCCTCTGTCG 60
                       2
 61 TATCCTGCCTGGCACAGAGGCACCCACCAGAGCCTGGCACCCAGCTAGCTTCTCTTTGGG 120

    ---------|---------3---------|---------|---------|---------|
121 CATACTCACCCACACACAGGTTTGCCATGATGCTAAGAAGCCTCTGGACTTCTTTAGTCC 180
  1                               M  L  R  S  L  W  T  S  L  V  L 11

    ---------|---------|---------|---------|---------|---------|
181 TTATCTCACTTGTATGCGCACTGACTGTATCTTCCAAGAAGAGTGGTAGTGGGAAAAGCA 240
 12  I  S  L  V  C  A  L  T  V  S  S  K  K  S  G  S  G  K  S  K 31

    ---------|---------|---------|---------|---------|---------|
241 AAACCGGAGGATGGAACAGTGGGAGCAACCGGAACCCCAACTACCCAGGAGGCTATGGCT 300
 32  T  G  G  W  N  S  G  S  N  R  N  P  N  Y  P  G  G  Y  G  W 51

    ---------|---------|---------|---------|---------|---------|
301 GGAACACCGGAGGGAATACTGGAGGCAGCTGGGGCCAACCTTATAATCCCAGTGGCGGAA 360
 52  N  T  G  G  N  T  G  G  S  W  G  Q  P  Y  N  P  S  G  G  N 71

    ---------|---------|---------|---------|---------|---------|
361 ACAATTTCAACAACAAGCAATGGAAACCTCCCAAGTCAAAAACCAATATGAAGGCTGTGG 420
 72  N  F  N  N  K  Q  W  K  P  P  K  S  K  T  N  M  K  A  V  A 91

    ---------|---------|---------|---------|---------|---------|
421 CCGTAGGCGCTGCTGCAGGCGCTATCGGGGGCTACATGCTCGGTAATGCAATGGGTCGTA 480
 92  V  G  A  A  A  G  A  I  G  G  Y  M  L  G  N  A  M  G  R  M 111

    ---------|---------|---------|---------|---------|---------|
481 TGAGCTATCATTTCAGCAATCCCATGGAAGCACGTTATTATAACGACTACTACAACCAGA 540
112  S  Y  H  F  S  N  P  M  E  A  R  Y  Y  N  D  Y  Y  N  Q  M 131

    ---------|---------|---------|---------|---------|---------|
541 TGCCAGAGCGTGTTTACAGGCCAATGTACAGAGGCGAGGAGCACGTGTCAGAGGATAGGT 600
132  P  E  R  V  Y  R  P  M  Y  R  G  E  E  H  V  S  E  D  R  F 151

    ---------|---------|---------|---------|---------|---------|
601 TTGTCACGGACTGCTACAATATGTCAGTGACAGAGTACATCATCAAGCCAGCTGAAGGGA 660
152  V  T  D  C  Y  N  M  S  V  T  E  Y  I  I  K  P  A  E  G  K 171

      ---------|---------|---------|---------|---------|---------|
661 AGAACACCAGCGAGGTAAACCAGTTGGAAACCAGGGTGAAGTCCCAAATTATTCGCGAGA 720
172  N  T  S  E  V  N  Q  L  E  T  R  V  K  S  Q  I  I  R  E  M 191

    ---------|---------|---------|---------|---------|---------|
721 TGTGTATCACTGAGTACAGGAGAGGATCGGGATTTAAGGTGCTCTCTAACCCTTGGCTGA 780
192  C  I  T  E  Y  R  R  G  S  G  F  K  V  L  S  N  P  W  L  I 211

    ---------|---------|---------|---------|---------|---------|
781 TCCTCACTATCACTCTTTTTGTTTACTTTGTGATAGAGTGACCAGAGGGAAGGCCAAATG 840
212  L  T  I  T  L  F  V  Y  F  V  I  E  *                      223

841 TATGTATATAG 851
```

(b) *X. tropicalis PRND* (chimeric)

```
    1
  1 GCCCTCCTGATCTGCCTGGCACATCCATCCTCAGTCCCTCTGTCGTATCCTGCCTGGCAC 60
        2
 61 AGAGGCACCCACCAGAGCCTGGCACCCAGCTAGCTTCTCTTTGGGCATACTCACCCACAC 120
        3                                                      4
121 ACAGGGTCACCATAGATCACCAATGGGGCCATTTTAGCTCTTCCACACTTCTACCACAGG 180

    ---------|---------|---------|---------|---------|---------|
181 TGGTGACAGAATGGGAAGGCAGAATCTATTCTCCTGTCTGATTCTTCTCCTGCTCATATT 240
  1            M  G  R  Q  N  L  F  S  C  L  I  L  L  L  L  I  L 17

    ---------|---------|---------|---------|---------|---------|
241 ATATTGTAGTCTCTCTTCTCCTAGAAGAGCAGCAAGCAGCAAAAAAATTAGCAAAACCAC 300
 18 Y  C  S  L  S  S  P  R  R  A  A  S  S  K  K  I  S  K  T  T  37
```

```
     ---------|---------|---------|---------|---------|---------|
 301 AGATTTGAGCAGGGGAGCCAAAAGAAGGCCAAAAGTGACCAATTCTCCTGCCCTCGGAGA 360
  38  D  L  S  R  G  A  K  R  R  P  K  V  T  N  S  P  A  L  G  D  57

     ---------|---------|---------|---------|---------|---------|
 361 TCTGTCCTTCAGAGGCAGGGCACTCAATGTGAACTTTAACCTTACCGAGGAATCTGAGCT 420
  58  L  S  F  R  G  R  A  L  N  V  N  F  N  L  T  E  E  S  E  L  77

     ---------|---------|---------|---------|---------|---------|
 421 TTATACAGCAAACCTCTACAGCTTCCCGGATGGCCTGTACTACCCACGGCCTGCCCACCT 480
  78  Y  T  A  N  L  Y  S  F  P  D  G  L  Y  Y  P  R  P  A  H  L  97

     ---------|---------|---------|---------|---------|---------|
 481 CAGTGGTGCTGGTGGGACTGACGAGTTTATAAGTGGGTGCCTTAACACCACAATAGAAAG 540
  98  S  G  A  G  G  T  D  E  F  I  S  G  C  L  N  T  T  I  E  R  117

     ---------|---------|---------|---------|---------|---------|
 541 AAACAAGGTCTGGATCTCTCAACTGGAAGACGATGAAGAAGGGGATATTTATATGAGCGT 600
 118  N  K  V  W  I  S  Q  L  E  D  D  E  E  G  D  I  Y  M  S  V  137

     ---------|---------|---------|---------|---------|---------|
 601 GGCCACGCAGGTCCTACAGTTTCTCTGTATGGAAAATTATGTAAAGCCTACCAATGGGGC 660
 138  A  T  Q  V  L  Q  F  L  C  M  E  N  Y  V  K  P  T  N  G  A  157

     ---------|---------|---------|---------|---------|---------|
  61 AGTGACCTGCACTGGTGGATTGTGGGTCTTTATAGGTGTCATGCATTTTTTTTTTTTATT 720
 158  V  T  C  T  G  G  L  W  V  F  I  G  V  M  H  F  F  F  L  F  177

     ---------|---------|---------|---------|---------|---------|
 721 TAGGAAGGGAGACTAAAGCCTAGGAATTCTGTattttatatgaacttttaagaactaact 780
 178  R  K  G  D  *                                                181

 781 gtactagcccagaggttcagcagccctataacagcaatgatccaggccttcaaatttgtc 840
 841 cacagcagctcttggatctcatcttggatcttttgagtgtcagtgacactgcacattctc 900
 901 agtgtgcagggctgctgttaaagactacgccatctgtcatagaatgcacgtttctacaca 960
 961 ggaatatactctaatatataggaatatatgtgcatgcaccctactaatagttcagcctgc 1020
1021 cagagatcaccaaggagttgcatagcctttttttacaaacatatccttcatctaaaacttc 1080
1081 atctgtgacttctaatagtcttataaattacaacaggagcaatattatattctatattat 1140
1141 atacatataagcaggaattatggctgctgttgtgcacacagggaagttggaggtttggcc 1200
1201 acggtttgttttccccttattataattgatgaaataaaaatcaggttaaactgg       1254
```

(c) *X. laevis PRNP*1-a (chimeric)

```
    1
  1 agtccgcccccacccctctcctgcatgaAGTCTCTTCCCCATCAGCTCATCCCTAGTCTC  60

 61 ACTGCTTTCCCGATCACCCTGGAACAGCCATCCTGAATCCCCCCCTGGCACATCCATTTC 120

                                2
121 GTATTTTCCCCTTGGCACAGAGGCACAGCACCCGGACCTGACACCCACATAGCTTCTCTT 180

    ---------|---------|-----3---|---------|---------|---------|
181 TGGCACACTCTATACCCTCACCCAGGTTGTTTATGATGCCACAAAGTCTCTGGACTTGTT 240
  1                                     M  P  Q  S  L  W  T  C  L 9

    ---------|---------|---------|---------|---------|---------|
241 TAGTCCTTATCTCCCTAGTATGCACATTGACTGTATCTTCCAAGAAGAGCGGTGGTGGGA 300
 10    V  L  I  S  L  V  C  T  L  T  V  S  S  K  K  S  G  G  G  K 29

    ---------|---------|---------|---------|---------|---------|
301 AAAGTAAAACTGGAGGATGGAACACAGGGAGCAACCGGAACCCCAACTACCCAGGAGGCT 360
 30    S  K  T  G  G  W  N  T  G  S  N  R  N  P  N  Y  P  G  G  Y 49

    ---------|---------|---------|---------|---------|---------|
361 ACCCAGGGAATACTGGAGGCAGCTGGGGGCAACAACCTTATAATCCTAGCGGTTATAACA 420
 50    P  G  N  T  G  G  S  W  G  Q  Q  P  Y  N  P  S  G  Y  N  K 69

    ---------|---------|---------|---------|---------|---------|
421 AGCAATGGAAACCTCCCAAGTCCAAAACCAACATGAAGTCGGTGGCCATAGGCGCTGCTG 480
 70    Q  W  K  P  P  K  S  K  T  N  M  K  S  V  A  I  G  A  A  A 89
```

```
    ---------|---------|---------|---------|---------|---------|
481 CTGGTGCTATTGGAGGCTACATGCTCGGTAATGCAGTGGGTCGTATGAGTTATCAATTCA 540
 90  G  A  I  G  G  Y  M  L  G  N  A  V  G  R  M  S  Y  Q  F  N 109

    ---------|---------|---------|---------|---------|---------|
541 ACAATCCCATGGAGTCCCGTTATTATAACGACTACTATAACCAGATGCCAAATCGCGTTT 600
110  N  P  M  E  S  R  Y  Y  N  D  Y  Y  N  Q  M  P  N  R  V  Y 129

    ---------|---------|---------|---------|---------|---------|
601 ACAGGCCTATGTACAGAGGAGAGGAGTACGTGTCAGAGGACAGGTTCGTGAGGGACTGCT 660
130  R  P  M  Y  R  G  E  E  Y  V  S  E  D  R  F  V  R  D  C  Y 149

    ---------|---------|---------|---------|---------|---------|
661 ACAATATGTCAGTGACAGAGTACATCATAAAGCCGACTGAAGGAAAGAACAACAGCGAGC 720
150  N  M  S  V  T  E  Y  I  I  K  P  T  E  G  K  N  N  S  E  L 169

    ---------|---------|---------|---------|---------|---------|
721 TAAACCAGTTGGATACCACGGTAAAGTCCCAAATTATTCGCGAGATGTGCATCACCGAGT 780
170  N  Q  L  D  T  T  V  K  S  Q  I  I  R  E  M  C  I  T  E  Y 189

    ---------|---------|---------|---------|---------|---------|
781 ACAGGAGAGGATCGGGATTCAAAGTGCTCTCTAACCCTTGGCTGATCCTTACTATCACTC 840
190  R  R  G  S  G  F  K  V  L  S  N  P  W  L  I  L  T  I  T  L 209

    ---------|---------|---------|---------|---------|---------|
841 TCTTTGTTTACTTTGTGATAGAGTGATCAAAGGAAATATTAATAAAAAGGCCAAATGTAT 900
210  F  V  Y  F  V  I  E  *                                     216

901 GTATATATAGAGAGAGTATAAACCGATTCTGAACTGTTCCGTCTCA 946
```

(d) *X. laevis PRNP1-b* (non-chimeric)

```
    2
  1 GACCCGGAATTCCCGGGATGATGGGAGCTCTCACTGCTGTAGTGTGTCAGCCTCACATGA 60

 61 GCTTCTCATTTGCCTTCCTGTAGCACAGCACCCGGACCTGACACCCACATAGCTTCTCTT 120

    ---------|---------|-----3---|---------|---------|---------|
121 TGGCACACTCTATACCCTCACCCAGGTTGTTTATGATGCCACAAAGTCTCTGGACTTGTT 180
  1                                    M  P  Q  S  L  W  T  C  L 9

    ---------|---------|---------|---------|---------|---------|
181 TAGTCCTTATCTCCCTAGTATGCACATTGACTGTATCTTCCAAGAAGAGCGGTGGTGGGA 240
 10  V  L  I  S  L  V  C  T  L  T  V  S  S  K  K  S  G  G  G  K 29

    ----------Sequence continues as in PRNP1-a exon 3-----------
```

(e) *X. laevis PRND*-a (chimeric)

```
    1
  1 ACTGTCACACTCCCAGACAAGAGCATTCCCCTCCTCCGACACGGACGTGAGGCCCAATTA 60
 61 AGGGCAGTCCGCCCCCACCCCTCTCCTGCATGAAGTCTCTTCCCCTTCAGCTCATCCCTA 120
121 GTCTCACTGCTTTCCCGATCACCCTGGAACAGCTATCCTGAATCCCCCCCCTGGCACATC 180
                                2
181 CATTTCGTATTTTCCCCTTGGCACAGAGGCACAGCACCCGGACCTGACACCCACATACCT 240

    ---------|---------|---------|-4-------|---------|---------|
241 TCTCTTTGGCACACTCTATACCCTCACCCAGGTGGCGACAGAATGGAAAGGCAGAACGTA 300
  1                                           M  E  R  Q  N  V   6

    ---------|---------|---------|---------|---------|---------|
301 TTCTCCTGCCTGATTCTTCTTGTGCTGATATTATATTGTGGTCTCTCTTGTCCGAGAAGA 360
  7 F  S  C  L  I  L  L  V  L  I  L  Y  C  G  L  S  C  P  R  R   26

    ---------|---------|---------|---------|---------|---------|
361 TCGGGAAGTGGCATTAAAAAATATTTCAAAATCAGCGACTTGAGCAGGGGAGCCAAAAAA 420
 27 S  G  S  G  I  K  K  Y  F  K  I  S  D  L  S  R  G  A  K  K   46

    ---------|---------|---------|---------|---------|---------|
421 AGGTCAAAAGTGGCCCATTCTCCTGTCCTCGGACACCTATTCTTCAGAAGTAAGGAGCTC 480
 47 R  S  K  V  A  H  S  P  V  L  G  H  L  F  F  R  S  K  E  L   66
```

```
    ---------|---------|---------|---------|---------|---------|
481 GATGTGAACCTTAACTTCACCGAGGAATATGAGCTTTATACAGAGAATCTGTACAGATTC 540
 67 D  V  N  L  N  F  T  E  E  Y  E  L  Y  T  E  N  L  Y  R  F    86

    ---------|---------|---------|---------|---------|---------|
541 CCGGACGGACTTTACTACCCATGGCGCTCCCAGCTGAATGATGCTGCCGGCACGGAGGAG 600
 87 P  D  G  L  Y  Y  P  W  R  S  Q  L  N  D  A  A  G  T  E  E    106

    ---------|---------|---------|---------|---------|---------|
601 TTTATGAACGGGTGCCTTAACACCACCGTAGAGAGAAACAAGGTCTGGATCTCTGGACTG 660
107 F  M  N  G  C  L  N  T  T  V  E  R  N  K  V  W  I  S  G  L    126

    ---------|---------|---------|---------|---------|---------|
661 GAAGAAGAGGACGAAGGGGAAACCTATATGAGTGTAGGCATGCAGGTCCTACAGTTTCTG 720
127 E  E  E  D  E  G  E  T  Y  M  S  V  G  M  Q  V  L  Q  F  L    146

    ---------|---------|---------|---------|---------|---------|
721 TGTTATGAAAACTATGTAAAGCCTACCAATGGGGCAGTGACCTGCACAGGAGGACTGTGG 780
147 C  Y  E  N  Y  V  K  P  T  N  G  A  V  T  C  T  G  G  L  W    166

    ---------|---------|---------|---------|---------|---------|
781 GTCTTCATAGGTGTCATTCACCTCCTTTTTTTTACTCAGAAAGGGAGTTAATGGGAACTA 840
167 V  F  I  G  V  I  H  L  L  F  F  T  Q  K  G  S  *            182

841 AAGCCTGAATTCTGTATTTCTTATATTGAACTaaatgtacccgcccagaggttcagcagc 900
901 tct 903
```

(f) *X. laevis PRND*-b(1), *PRND*-b(2) (non-chimeric)

```
    3
  1 aaaggcccagatttaggcaggactcattacatctctacatgtttggcactcaggcccgga 60
 61 ctgacaatctgcccgttcggggcccgccgtctgccctgcatgacactgtcccgtctgcag 120
121 cccccctcccacctccacagagcagagtaatatcacagagcagaggtgagggcaggactgc 180
181 agaaattttgaactagtgggggaggacgcccacctgacgccatcacacactggtctattt 240
241 acaaacagtggactggagacctgtattattactgggggaagctggagagaagtatttttt 300
301 tcagggcaaagcagagttggtgCTGAAAGTGGACTGGAGAGTGGGCCAGCAGTGTCTCAT 360

    ---------|---------|---------|---------4---------|---------|
361 CAGTCTGAAGAAGGAAGAAACAGCAGCAAAGGCTCTGAGGTGGTGACAGAATGGAAAGGC 420
  1                                                   M  E  R  Q 4

    ---------|---------|---------|---------|---------|---------|
421 AGAACGTATTCTCCTGCCTGATTCTTCTTGTGCTGATATTATATTGTGGTCTCTCTTGTC 480
  5   N  V  F  S  C  L  I  L  L  V  L  I  L  Y  C  G  L  S  C  P 24

    -----------Sequence continues as in PRND-a exon 4-----------
```

(g) *X. laevis PRNP*2b-1

```
    1
  1 GGCTGGACCGGTCCGGATTCCCGGGATGATCAGCCTGGCACAGCCGTCCTGAACCTCCCC 60

                                           2
 61 CTGTACAGGCACAGTCTTTCTGAATCCCCCCTGGCACAGAGGGCCTGACAGCCAGCTAGC 120

    ---------|---------|---------|-----3---|---------|---------|
121 TTCTCTTTGGGCTTACTTATCCTCTCACACAGGTTGGCCAAGATGCCAAGAAGTCTCTGG 180
  1                                           M  P  R  S  L  W    6

    ---------|---------|---------|---------|---------|---------|
181 ACTTGTTTAGTCCTTATCTCCCTAGTGTGCACATTGACTGTATCTTCCAAGAAGAGTGGT 240
  7 T  C  L  V  L  I  S  L  V  C  T  L  T  V  S  S  K  K  S  G    26

    ---------|---------|---------|---------|---------|---------|
241 AGTGGGAAAAGCAAAACCGGAGGCTGGAACAATGGGAACACTGGGAACACCGGGAACACT 300
 27 S  G  K  S  K  T  G  G  W  N  N  G  N  T  G  N  T  G  N  T    46

    ---------|---------|---------|---------|---------|---------|
301 GGGAACAACCGGAACCCCAACTATCCAGGAGGCTATGGCTGGAACACAGGGAACACAGGG 360
 47 G  N  N  R  N  P  N  Y  P  G  G  Y  G  W  N  T  G  N  T  G    66
```

```
     ---------|---------|---------|---------|---------|---------|
 361 AACACTGGAGGCAGTTGGGGGCAACAACCTTATAATCCTAGCGGAGGAAGCAATTTCAAC 420
  67 N  T  G  G  S  W  G  Q  Q  P  Y  N  P  S  G  G  S  N  F  N   86

     ---------|---------|---------|---------|---------|---------|
 421 AACAAGCAATGGAAACCTCCCAAGTCCAAAACCAATATGAAGGCCGTGGCCGTAGGCGCT 480
  87 N  K  Q  W  K  P  P  K  S  K  T  N  M  K  A  V  A  V  G  A   106

     ---------|---------|---------|---------|---------|---------|
 481 GCTGCTGGTGCTATTGGAGGCTACATGCTTGGTAATGCAGTGGGTCGTATGAATCATCAT 540
 107 A  A  G  A  I  G  G  Y  M  L  G  N  A  V  G  R  M  N  H  H   126

     ---------|---------|---------|---------|---------|---------|
 541 TTCGACAATCCCATGGAATCCCGTTATTATAACGACTACTACAACCAGATGCCAGACCGC 600
 127 F  D  N  P  M  E  S  R  Y  Y  N  D  Y  Y  N  Q  M  P  D  R   146

     ---------|---------|---------|---------|---------|---------|
 601 GTTTACAGGCCAATGTACAAAACCGAGGAGTACGTGTCTGAAGATAGGTTCGTCACGGAT 660
 147 V  Y  R  P  M  Y  K  T  E  E  Y  V  S  E  D  R  F  V  T  D   166

     ---------|---------|---------|---------|---------|---------|
 661 TGCTACAATATGTCAGTAACAGAGTACATCATCAAGCCATCCGAAGGGAAGAATGGCAGC 720
 167 C  Y  N  M  S  V  T  E  Y  I  I  K  P  S  E  G  K  N  G  S   186

     ---------|---------|---------|---------|---------|---------|
 721 GATGTAAACCAGTTGGATACCGTGGTGAAATCCAAAATTATTCGCGAGATGTGCATCACT 780
 187 D  V  N  Q  L  D  T  V  V  K  S  K  I  I  R  E  M  C  I  T   206

     ---------|---------|---------|---------|---------|---------|
 781 GAATACAGGAGAGGATCAGGATTCAAAGTGCTTTCTAACCCATGGCTGATCCTCACTATC 840
 207 E  Y  R  R  G  S  G  F  K  V  L  S  N  P  W  L  I  L  T  I   226

     ---------|---------|---------|---------|---------|---------|
 841 ACTCTCTTTGTTTACTTTGTGATAGAGTGAccagagggaatcgaaatccaaaggccaatt 900
 227 T  L  F  V  Y  F  V  I  E  *                                 235

 901 gtatgtatatagagagtatcaaccaattctggactgtctcgtctcatgcccaatatgaca 960
 961 ctgttgggtgctatgttaatccagcccagcttcctaccatcagtaagcaaccatggatct 1020
1021 tgcttcaataccaacctgaattcccacttctgcgtcaactacaacttctgcttgacaaca 1080
1081 catgtttctgacattgcaaatgggctttatgtatcagtgcatgtatataataagagatct 1140
1141 tcatt 1145
```

(h) *X. laevis PRNP*2b-2

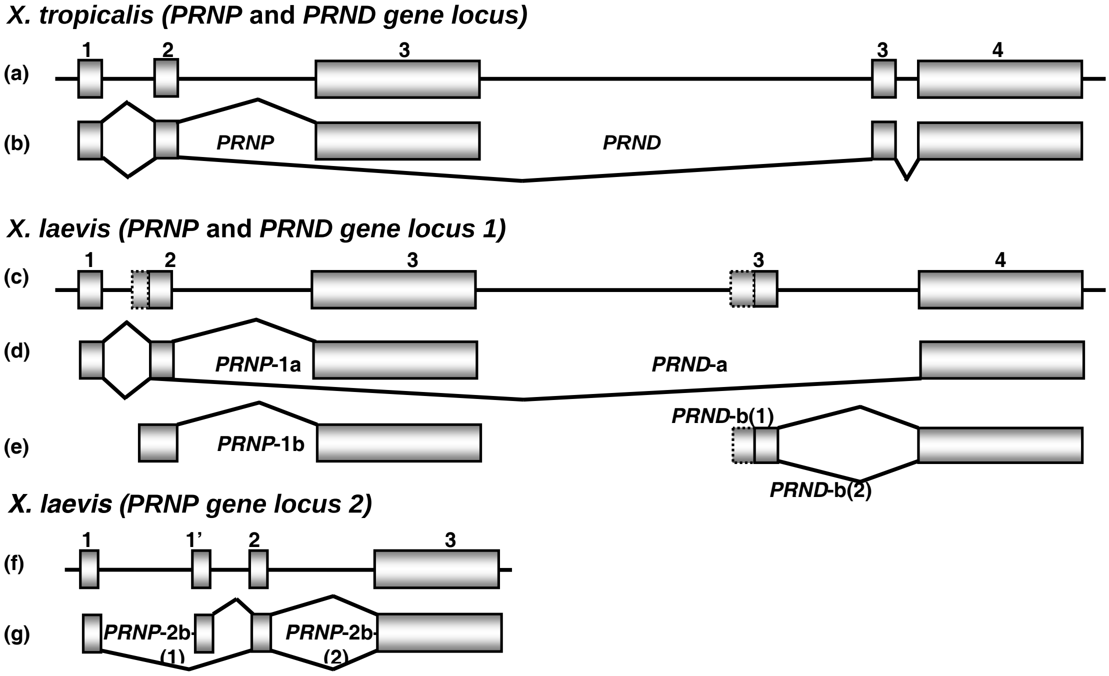
```
     1′                                          2
   1 gtccggaattctccggatcacgagcttctcatttgccttcctgtagggcctgacagccag 60

     ---------|---------|---------|---------|3--------|---------|
  61 CTAGCTTCTCTTTGGGCTTACTTATCCTCTCACACAGGTTGGCCAAGATGCCAAGAAGTC 120
   1                                                M  P  R  S  L 5

     ---------|---------|---------|---------|---------|---------|
 121 TCTGGACTTGTTTAGTCCTTATCTCCCTAGTGTGCACATTGACTGTATCTTCCAAGAAGA 180
   6   W  T  C  L  V  L  I  S  L  V  C  T  L  T  V  S  S  K  K  S 25

    --------------Sequence continues as in PRNP2b-1--------------
```

Figure 3-1 Nucleotide and deduced amino acid sequences of cDNAs for chimeric and non-chimeric transcripts. (a) *X. tropicalis PRNP* (chimeric), (b) *X. tropicalis PRND* (chimeric), (c) *X. laevis PRNP1-a* (chimeric), (d) *X. laevis PRNP1-b* (non-chimeric). (e) *X. laevis PRND-a* (chimeric), (f) *X. laevis PRND-b(1), PRND-b(2)* (non-chimeric), (g) *X. laevis PRNP2b-1, (h) X. laevis PRNP2b-2.* Exons are numbered and non-coding exons are enclosed by full, dashed or dotted boxes. In the absence of genomic sequence for *X. laevis* the exon boundaries have been predicted based on alignments of cDNA sequences with those for *X. tropicalis*. The cDNA sequence corresponding to the ORF is highlighted in grey. The 5' or 3' end variants are represented by lower case letters.


X. tropicalis (PRNP and PRND gene locus)
(a)
1
2
3
3
4
(b)
PRNP
PRND
X. laevis (PRNP and PRND gene locus 1)
(c)
1
2
3
3
4
(d)
PRNP-1a
PRND-a
(e)
PRNP-1b
PRND-b(1)
PRND-b(2)
X. laevis (PRNP gene locus 2)
(f)
1
1'
2
3
(g)
PRNP-2b-(1)
PRNP-2b-(2)

(h)

	X. tropicalis		*X. laevis*					
	PRNP	***PRND***	***PRNP*1a**	***PRND*a**	***PRNP*1b**	***PRND*-b(1,2)**	***PRNP*-2b(1)**	***PRNP*-2b(2)**
Exon1	79v1 44v2	64	143v1 115v2	208	-	-	44	Exon 1'- 99
Exon2	60	60	63	63	145	-	56	56
Exon3	711	55	740	-	649	399v1 77v2	990	990
Exon4	-	1083v1 573v2	-	633v1 569v2	-	633v1 569v2	-	-
Intron1	284	284	286	286	-	-	*	314
Intron2	7676	27099	*	*	*	-	-	-
Intron3	-	1024	-	-	-	*	*	*

Figure 3-2 Transcripts coding for *PRNP* and *PRND*. (a) *X. tropicalis* and (b) *X. laevis*. The first two non-coding exons are shared in the chimeric transcripts and the two products (*PRNP* and *PRND*) are generated by alternative splicing. The chimeric form of *X. laevis PRND* lacks an exon which is present in *X. tropicalis*. *PRND*-b(1) and *PRND*-b(2) have different start sites.

3.3.2 *X. laevis PRNP* and *PRND*

Primers constructed from *X. tropicalis PRND* (F3, R2; see Appendix 1) were used to amplify the genomic fragment corresponding to the *X. laevis PRND* ORF. Gene specific primers for *X. laevis PRNP and PRND* (see Appendix 1) were used to perform RACE on *X. laevis* brain cDNA as described in Methods (experiments performed by Tatiana Vassilieva). Sequencing of RACE products confirmed the shared 5' ends for *PRNP* and *PRND in X. laevis* (Figure 3-2d). These chimeric variants are represented as *PRNP*-1a and *PRND*-a (Figure 3-2d). It is not possible to definitively ascribe the exon-intron boundaries as the genomic sequence information is not available. Nonetheless, an attempt was made to define the boundaries based on the alignment of *PRND* and *PRNP* cDNA sequences from *X. laevis* and *X. tropicalis*. *X. laevis PRND* may have only three exons (Figure 3-2d).

Comparison of PrP sequences for *X. tropicalis* and *X. laevis* shows, as expected, very strong conservation over the complete protein with two short insertions/deletions in N-terminal region (Figure 3-3). The C-terminal end is very well conserved, including complete conservation of signal sequence. Interestingly, *X. laevis* Dpl shows a larger number of sequence variations compared with *X. tropicalis* (Figure 3-3). This observation is consistent with the behaviors of a recently duplicated gene. Following gene duplication, one of the gene becomes functionally redundant and it either evolves rapidly to perform novel functions, subject to stabilizing selection, or may become a pseudogene relatively rapidly(Ohno 1970).

```
(a)             10        20        30        40        50        60
        ....|....|....|....|....|....|....|....|....|....|....|....|
XtPrP   MLRSLWTSLVLISLVCALTVSSKKSGSGKSKTGGWNSGSNRNPNYPGGYGWNTGGNTGGS
XlPrP1  MPQSLWTCLVLISLVCTLTVSSKKSGGGKSKTGGWNTGSNRNPNYPGGY----PGNTGGS
        * :****.********:*********.*********:************     ******

                70        80        90       100       110       120
        ....|....|....|....|....|....|....|....|....|....|....|....|
XtPrP   WG-QPYNPSGGNNFNNKQWKPPKSKTNMKAVAVGAAAGAIGGYMLGNAMGRMSYHFSNPM
XlPrP1  WGQQPYNPSG----YNKQWKPPKSKTNMKSVAIGAAAGAIGGYMLGNAVGRMSYQFNNPM
        ** *******     **************:**:***************:*****:*.***

               130       140       150       160       170       180
        ....|....|....|....|....|....|....|....|....|....|....|....|
XtPrP   EARYYNDYYNQMPERVYRPMYRGEEHVSEDRFVTDCYNMSVTEYIIKPAEGKNTSEVNQL
XlPrP1  ESRYYNDYYNQMPNRVYRPMYRGEEYVSEDRFVRDCYNMSVTEYIIKPTEGKNNSELNQL
        *:***********:***********:******* **************:****.**:***

               190       200       210       220
        ....|....|....|....|....|....|....|....|....
XtPrP   ETRVKSQIIREMCITEYRRGSGFKVLSNPWLILTITLFVYFVIE
XlPrP1  DTTVKSQIIREMCITEYRRGSGFKVLSNPWLILTITLFVYFVIE
        :* *****************************************
```

```
(b)             10        20        30        40        50        60
        ....|....|....|....|....|....|....|....|....|....|....|....|
XtDpl   MGRQNLFSCLILLLLILYCSLSSPRRAAS-SKKISKTTDLSRGAKRRPKVTNSPALGDLS
XlDpl   MERQNVFSCLILLVLILYCGLSCPRRSGSGIKKYFKISDLSRGAKKRSKVAHSPVLGHLF
        * ***:*******:*****.**.***:.*  **  * :*******:*.**::**.**.*

                70        80        90       100       110       120
        ....|....|....|....|....|....|....|....|....|....|....|....|
XtDpl   FRGRALNVNFNLTEESELYTANLYSFPDGLYYPRPAHLSGAGGTDEFISGCLNTTIERNK
XlDpl   FRSKELDVNLNFTEEYELYTENLYRFPDGLYYPWRSQLNDAAGTEEFMNGCLNTTVERNK
        **.: *:**:*:*** **** *** ********  ::*..*.**:**:.******:****

               130       140       150       160       170       180
        ....|....|....|....|....|....|....|....|....|....|....|....|
XtDpl   VWISQLEDDEEGDIYMSVATQVLQFLCMENYVKPTNGAVTCTGGLWVFIGVMHFFFLFRK
XlDpl   VWISGLEEEDEGETYMSVGMQVLQFLCYENYVKPTNGAVTCTGGLWVFIGVIHLLFFTQK
        **** **:::**: ****. ******* ***********************:*::*: :*


        ..
XtDpl   GD
XlDpl   GS
        *.
```

Figure 3-3 Sequence comparisons for *X. tropicalis* and *X. laevis* (a) PrP and (b) Dpl.

3.3.3 Generation of chimeric trancript

The mapping of the cDNA sequences onto the genomic organization of the *PRNP* and *PRND* genes in *X. tropicalis* shows that these genes share the first two exons and a common promoter, and are therefore referred to as chimeric transcripts. This is evident from the fact that the 5' ends of *PRNP* and *PRND* are nearly identical. Such a process of a single promoter being responsible for transcription of more than one gene is termed polycistronic transcription, and where it involves two genes, as in this, it is termed dicistronic transcription. Other cases of polycistronic transcription have been reviewed by Blumenthal (Blumenthal 1998). The common primary dicistronic transcript is processed to give one of two alternative mature monocistronic transcripts, *PRNP* or *PRND*. Such a phenomenon of clustering and co-transcription could be a result of gene duplication (Blumenthal 1998). Interestingly, the *PRNP* coding region is entirely contained within the long second intron of *PRND*.

This unusual and interesting phenomenon can be better understood by the following analysis applied to the available *X. tropicalis* genomic sequence. Once a polycistronic pre-mRNA is transcribed, a decision has to be made as to whether to form a *PRNP* or *PRND* monocistronic transcript. *PRNP* can be made by splicing and clipping/polyadenylation at the 3' end of *PRNP* exon 3 (Figure 3-2a). Alternatively, the *PRNP* coding exon and a large non-coding intergenic sequence can be removed as an intron by a splice made at the second 5' splice site to the third *PRND* exon, resulting in a functional *PRND* mRNA (Figure 3-2b). Possible explanations for mediation of this alternative splicing are through (1) exon skipping, (2) differences in acceptor site signal sequence, or (3) actions of exonic enhancers and silencers.

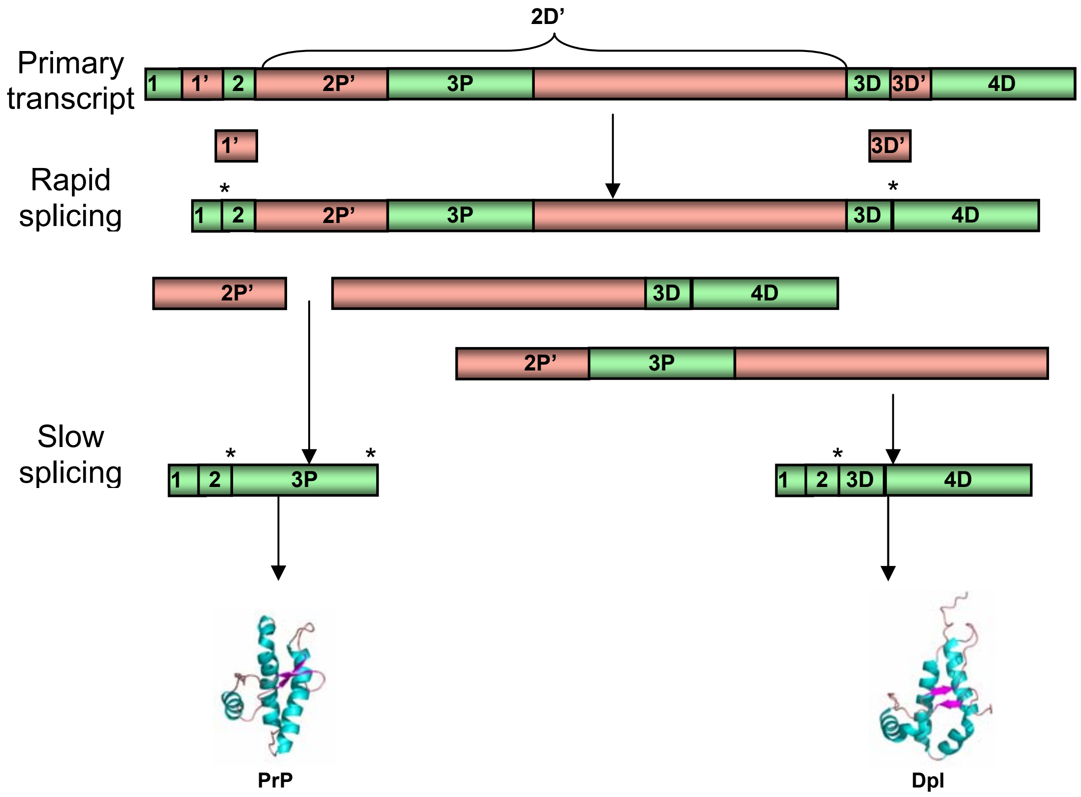


Figure 3-4 Schematic representation of dicistronic transcription leading to two different gene products, PrP and Dpl The numbering corresponds to the exon/intron number and the letter "P" and "D" represents PrP and Dpl respectively. Introns are represented by (') and Splicing events are represented by "*". See text for more detailed explanation.

3.3.3.1 Alternative splicing mediated by exon skipping

Alfonso et al., (1994) proposed an explanatory model for such a phenomenon of coordinated gene expression using the example for transcripts of *unc*-17 and *cha*-1 proteins in *Caenorhabditis elegans*; these transcripts also share the first exon and promoter. According to this model, the two proteins with related function are produced as a result of alternative splicing of a common mRNA precursor. Such a gene complex is referred to as a eukaryotic operon. They classified the rate of splicing into rapid and slow steps based on the complexity of the decision to be made for alternative splicing. Based on this hypothesis, the splicings of the common intron 1', and *PRND* intron 3D' (Figure 3-4) are rapid events as they do not affect the final outcome. This would be followed by two slow steps involving splicing at exon 2 (splice donor) and clipping/polyadenylation reactions to either exon 3P or 3D (splice acceptor) (Figure 3-4). This is a slow process as once the clipping at the 3’ end of *PRNP* (exon 3P) is made, it is no longer possible to make the *PRND* mRNA. The alternative splicing by exon skipping involves a single exon that is either included or not included with a single splice donor (3’ end of exon 2) and two alternative splice acceptors (5’ end of 3P or 3D). As this splicing is a critical step, it is important that it be a slow reaction step to make a proper exon choice. The lack of a canonical AATAAA signal in the 3’ end of the last *PRNP* exon may help in this slow reaction (Alfonso et al. 1994). The consensus sequence, TATAAA in the 3’ end of the *PRNP* gene in *X. laevis* (Strumbo et al. 2001) (Figure 3-5b) may not be a true polyadenylation signal as this might otherwise influence the alternative splicing. Evidence to support this conjecture is that although such a region is present in the genomic DNA of *X. tropicalis* 3’ to the last *PRNP* exon, it is not included in the *PRNP* mRNA transcript (Figure 3-5a) suggesting that it may not correspond to the polyadenylation signal in *X. laevis*. The 3’ end of the last *PRND* exon in *X. tropicalis* has AATAAA whereas in *X. laevis* it is ACTAAA (Figure 3-5c, d). It is not evident whether these regions act as signals for polyadenylation for *PRND*. The lack of a proper polyadenylation signal in the last exon of *PRNP* may slow down the process of splicing and help in controlling the choice of exon selection.

(a). 3’ end of *X. tropicalis* *PRNP*

AAGTCCCAAATTATTCGCGAGATGTGTATCACTGAGTACAGGAGAGGATCGGGATTTAAG
GTGCTCTCTAACCCTTGGCTGATCCTCACTATCACTCTTTTTGTTTACTTTGTGATAGAG
TGACCAGAGGGAAGGCCAAATGTATGTATATAGAGATTTAAAGAGAA**TATAAA**CCGATTC
TGAACTGTCCTGTCTCACGCCCA

(b). 3’ end of *X. laevis* *PRNP*

AAGTCCCAAATTATTCGCGAGATGTGCATCACCGAGTACAGGAGAGGATCGGGATTCAAA
GTGCTCTCTAACCCTTGGCTGATCCTTACTATCACTCTCTTTGTTTACTTTGTGATAGAG
TGATCAAAGGAAATATTAATAAAAAGGCCAAATGTATGTATATATAGAGAGAG**TATAAA**C
CGATTCTGAACTGTTCCGTCTCA

(c). 3’ end of *X. tropicalis* *PRND*

CTACAGTTTCTCTGTATGGAAAATTATGTAAAGCCTACCAATGGGGCAGTGACCTGCACT
GGTGGATTGTGGGTCTTTATAGGTGTCATGCATTTTTTTTTTTTTATTTAGGAAGGGAGAC
TAAAGCCTAGGAATTCTGTATTTTATATGAACTTTTAAGAACTAACTGTACTAGCCCAGA
GGTTCAGCAGCCCTATAACAGCAATGATCCAGGCCTTCAAATTTGTCCACAGCAGCTCTT
GGATCTCATCTTGGATCTTTTGAGTGTCAGTGACACTGCACATTCTCAGTGTGCAGGGCT
GCTGTTAAAGACTACGCCATCTGTCATAGAATGCACGTTTCTACACAGGAATATACTCTA
ATATATAGGAATATATGTGCATGCACCCTACTAATAGTTCAGCCTGCCAGAGATCACCAA
GGAGTTGCATAGCCTTTTTTACAAACATATCCTTCATCTAAAACTTCATCTGTGACTTCT
AATAGTCTTATAAATTACAACAGGAGCAATATTATATTCTATATTATATACATATAAGCA
GGAATTATGGCTGCTGTTGTGCACACAGGGAAGTTGGAGGTTTGGCCACGGTTTGTTTTC
CCCTTATTATAATTGATGA**AATAAA**AATCAGGTTAAACTGG

(d). 3’ end of *X. laevis* *PRND*

CTGGAAGAAGAGGACGAAGGGGAAACCTATATGAGTGTAGGCATGCAGGTCCTACAGTTT
CTGTGTTATGAAAACTATGTAAAGCCTACCAATGGGGCAGTGACCCGCACAGGAGGACTG
TGAGTCTTCATAAGTGTCATTCACCTCCTTTTTTTTTACTCAGAAAGGGAGTTAATGGGA
ACTAAAGCCTGAATTCTGTATTTCTTATATTGA**ACTAAA**TGTACCCGCCCAGAGGTTCAG
CAGCTCT

Figure 3-5 The 3’ ends of *PRNP* and *PRND* in *X. tropicalis* and *X. laevis*. Bases in bold correspond to the ORF. Bases highlighted in grey correspond to cDNA. Bases underlined correspond to shorter transcript. Bases which are boxed correspond to putative polyadenylation signal. (a) Bases not highlighted correspond to genomic DNA. This region is similar to the putative TATAAA site in (b) but is not included in the mRNA transcript. Also the putative polyadenylation sites in *PRND* (c) and (d) are also not conserved indicating that they may not be the true polyadenylation signal.

3.3.3.2 Alternative splicing mediated by differences in acceptor site signal sequence

Bailleul et al., (1997) reported the first mammalian example (*OB-RGRP* and *OB-R*) of genes sharing a promoter and the first two exons. They suggested that the acceptor splice site signal plays an important role in alternate splicing to produce different gene coding transcripts. The splice acceptor signal is comprised of a branch site located 20 - 50 bases upstream of the acceptor site with the consensus

"CU(A/G)**A**(C/U)" and a pyrimidine-rich region upstream to the splice acceptor site (AG) (Figure 3-6a). This splice acceptor signal is not strong in either *PRNP* and *PRND* intron 2 (2P and 2D) (Figure 3-6c,d) due to a weak branch site and/or the presence of purine bases in the required pyrimidine-rich region. Two out of twelve nucleotides upstream to the splice acceptor site in *PRNP* intron 2P are purine while four out of twelve nucleotides are purine in *PRND* intron 2D. By comparison, *PRNP* intron 2 has a stronger splicing signal than *PRND* intron 2 and the latter also lacks the branch site needed for the splice signal (Figure 3-6d). These features likely determine the ratio between the *PRNP* and *PRND* transcripts.

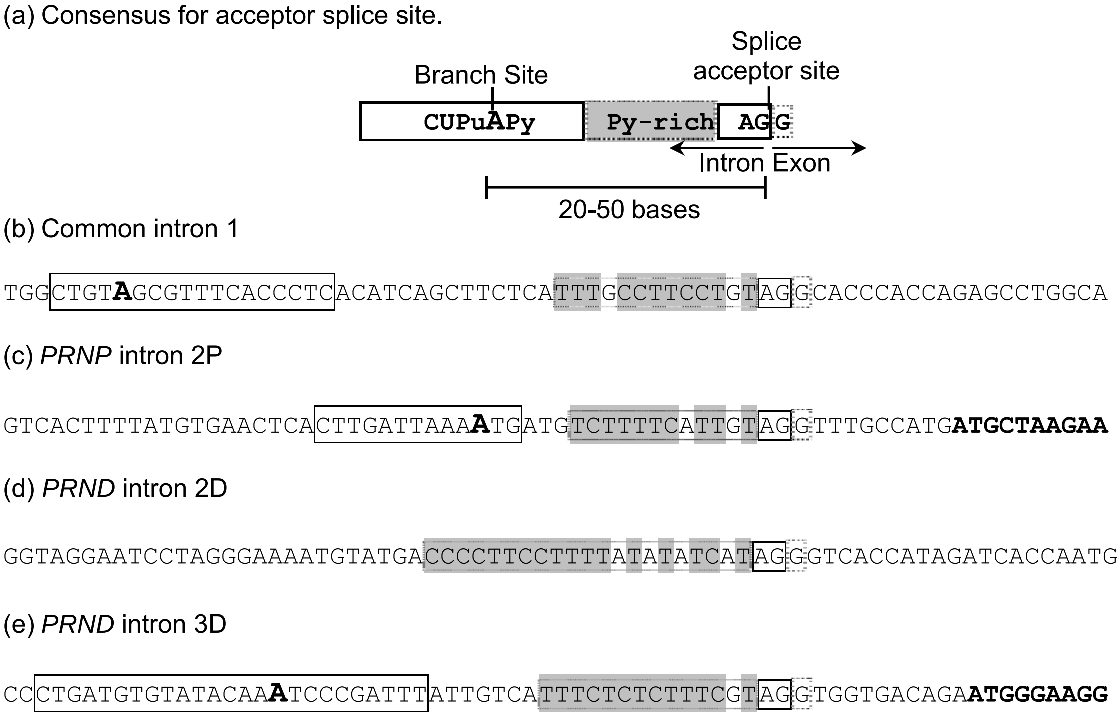

(a) Consensus for acceptor splice site.
Splice acceptor site
Branch Site
CUPuAPy
Py-rich
AG
G
Intron
Exon
20-50 bases
(b) Common intron 1
TGGCTGTAGCGTTTCACCCTCACATCAGCTTCTCATTTGCCTTCCTGTAGGCACCCACCAGAGCCTGGCA
(c) PRNP intron 2P
GTCACTTTTATGTGAACTCACTTGATTAAAATGATGTCTTTTCATTGTAGGTTTGCCATGATGCTAAGAA
(d) PRND intron 2D
GGTAGGAATCCTAGGGAAAATGTATGACCCCTTCCTTTTATATATCATAGGGTCACCATAGATCACCAATG
(e) PRND intron 3D
CCCTGATGTGTATACAAATCCCGATTTATTGTCATTTCTCTCTTTCGTAGGTGGTGACAGAATGGGAAGG

Figure 3-6 Analysis of splice signal for *PRNP* and *PRND.* (a). The consensus sequence for splicing. Purine (Pu) = A or G; Pyrimidine (Py) = C or U. The putative branch sites for *PRNP* and *PRND* are boxed in (b), (c), (d) and (e). The Py-rich region is highlighted in grey. Note that the branch site is not strong in all the introns and there are some purine bases in the Py-rich region with a maximum of such occurrences in *PRND* intron 2 which also lacks the branch site.

3.3.3.3 Role of exonic enhancers and silencers

Signals for splice-site recognition have been identified in intronic and exonic cis-elements of genes which either act by stimulating (enhancers) or repressing (silencers) splicing. Cartegni et al. (2002) proposed a model for enhancer-dependent splicing whereby a Ser/Arg-rich protein is thought to bind to the exonic DNA though the RNA Recognition Motif and assist in the splicing process by directly recruiting splicing machinery. Such a regulated splicing may occur to strengthen weak splice sites and regulate the splicing to either *PRNP* or *PRND*.

3.3.4 Non-chimeric transcripts in *X. laevis*

Sequencing of 5' RACE products (performed by Tatiana Vassilieva) indicated two additional splice variants for the *PRND* transcript with a different first exon than that found in chimeric *PRND*. Using left primers (LD5, LD6) constructed from these new exons I tried to amplify cDNA using *PRNP* right ORF primer (RP1) without success, confirming the existence of non-chimeric *PRND,* represented as *PRND*-b in Figure 3-2e. I also made an attempt to sequence the genomic DNA between these new *PRND* exons to determine their relative position but this was not successful.

A *PRNP* sequence obtained from the EST database searches showed a different exon 1 but with the same exon 2 as in the chimeric form (Figure 3-2e). Sequencing results confirmed it as a non-chimeric *PRNP* transcript (*PRNP*-1b in Figure 3-2e) as no PCR product was obtained when either the new *PRNP*-1b exon 1 left primer (LP1) or the *PRND* exon 4 (Figure 3-2c) right primer (RD5) was used on cDNA from brain and testis. Genomic DNA sequencing results indicated that the new exon 1 is a 5' extension of the common exon 2 (left primer: LP1; right primer: RP1) and, hence, the longer variant was represented as *PRND*-b(1) and the shorter variant as *PRND*-b(2).

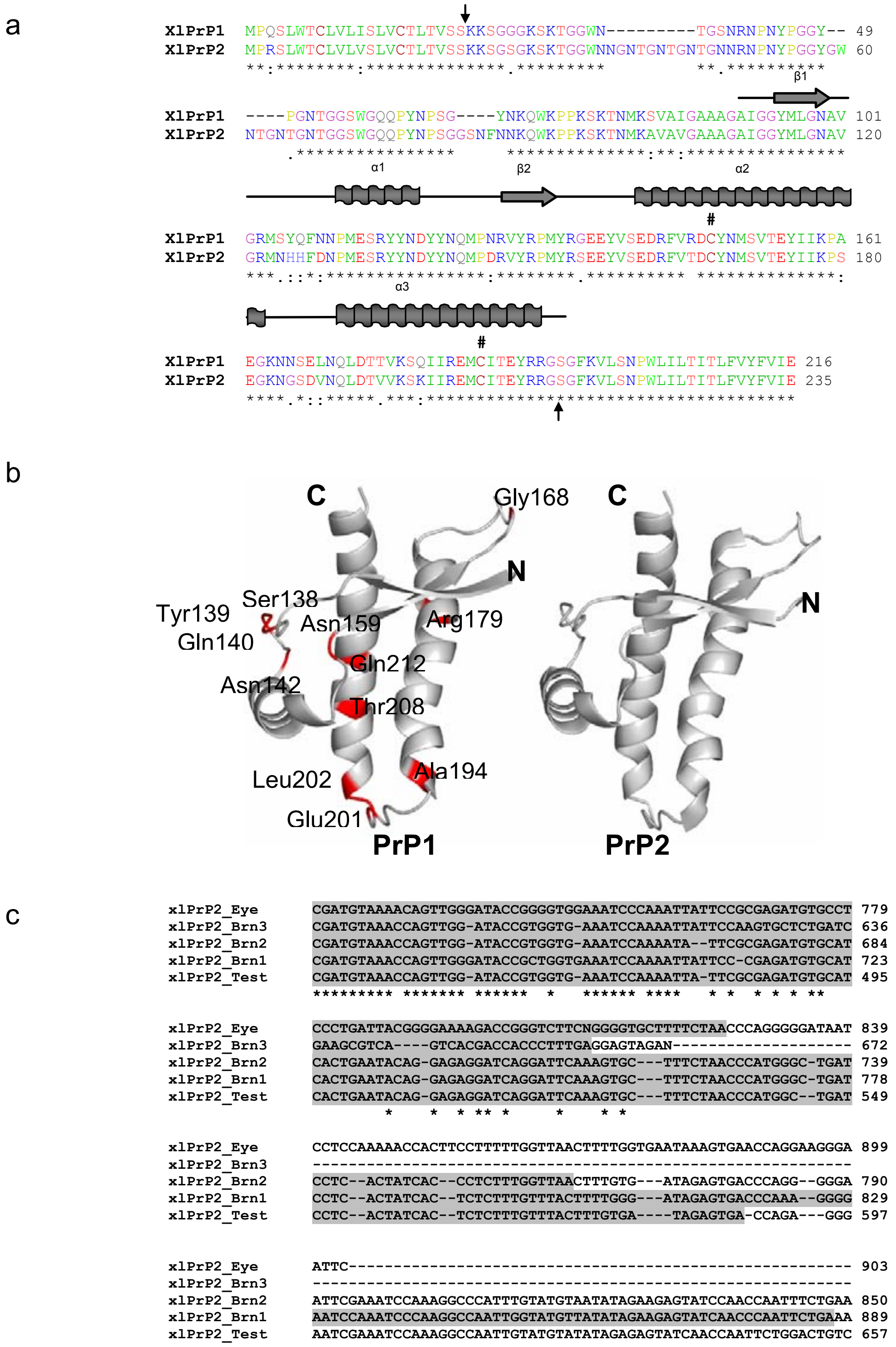
3.3.5 Duplicate of PrP in *X. laevis*

EST database searches revealed an *X. laevis* PrP sequence with some sequence variation from that which was already known (Strumbo et al. 2001). Sequence comparison of the new PrP sequence with that previously reported indicated two distinct PrP sequences, most likely encoded by two distinct genes in *X. laevis*. I designated the previously reported gene as *PRNP*1 and the new duplicate gene as

*PRNP*2, coding for PrP1 and PrP2, respectively. This represents the first discovery of retained *PRNP* duplicates in a tetrapod. The existence of duplicated genes is consistent with the genome duplication that occurred at least 30 million years ago (Hughes and Hughes 1993). *X. laevis* is a pseudo-tetraploid species which has 36 chromosomes, as compared with the diploid number (20) in *X. tropicalis.* Several retained duplicates have been characterized (Hughes and Hughes 1993). The retention of *PRNP*2 indicates that these genes have been under strong selective pressure and that subfunctionalisation or neofunctionalisation would have been the factors in retaining the duplicate form.

Sequences for *PRNP*2 from the EST database indicate two different first exons (exon 1 and exon 1') but a common second exon (Figure 3-2g). PCR experiments confirmed the existence of these variant forms. Amplifying the genomic DNA between these exons (exon 1 and 2: Primer LP2 and RP2; exon 1' and 2; primer LP3 and RP2) defined the size of only intron 1'. For the other transcript, there was no PCR product, indicating either that it may be rich in repeat regions or that the intron is large. The various PrP2 sequences obtained from the EST database showed sequence variation in the 3' end (resulting in different C-terminal amino acid sequences) suggesting the possibility of alternative terminal exon usage (Figure 3-7c). In order to test this, different right primers (RP3, RP4 and RP5) were constructed from the different sequences and tried with the common left primer (LP4). Only one of the primers showed PCR product (LP4 and RP5). This result suggests the sequence variations likely result from sequencing errors.

The two PrPs (PrP1 and PrP2) share high sequence homology (Figure 3-7a), with PrP2 showing maximum variation in the N-terminal region and the post hydrophobic region from residue 124-128 ("NHHFD"), which interestingly resembles the corresponding region of chicken and turtle PrP rather than *X. laevis* PrP1 (refer to Figure 6-5). An homology model for PrP2 based on the NMR structure of *X. laevis* PrP (Calzolai et al. 2005) was made to map the residues that differ from PrP1 (Figure 3-7b).



```
XlPrP1  MPQSLWTCLVLISLVCTLTVSSKKSGGGKSKTGGWN---------TGSNRNPNYPGGY-- 49
XlPrP2  MPRSLWTCLVLVSLVCTLTVSSKKSGSGKSKTGGWNNGNTGNTGNTGNNRNPNYPGGYGW 60
        **:********:**************.*********         **.**********

XlPrP1  ----PGNTGGSWGQQPYNPSG----YNKQWKPPKSKTNMKSVAIGAAAGAIGGYMLGNAV 101
XlPrP2  NTGNTGNTGGSWGQQPYNPSGGSNFNNKQWKPPKSKTNMKAVAVGAAAGAIGGYMLGNAV 120
            .****************     **************:**:****************

XlPrP1  GRMSYQFNNPMESRYYNDYYNQMPNRVYRPMYRGEEYVSEDRFVRDCYNMSVTEYIIKPA 161
XlPrP2  GRMNHHFDNPMESRYYNDYYNQMPDRVYRPMYRSEEYVSEDRFVTDCYNMSVTEYIIKPS 180
        ***.::*:****************:********.********** **************:

XlPrP1  EGKNNSELNQLDTTVKSQIIREMCITEYRRGSGFKVLSNPWLILTITLFVYFVIE 216
XlPrP2  EGKNGSDVNQLDTVVKSKIIREMCITEYRRGSGFKVLSNPWLILTITLFVYFVIE 235
        ****.*::*****.***:*************************************
```

```
xlPrP2_Eye     CGATGTAAAACAGTTGGGATACCGGGGTGGAAATCCCAAATTATTCCGCGAGATGTGCCT 779
xlPrP2_Brn3    CGATGTAAACCAGTTGG-ATACCGTGGTG-AAATCCAAAATTATTCCAAGTGCTCTGATC 636
xlPrP2_Brn2    CGATGTAAACCAGTTGG-ATACCGTGGTG-AAATCCAAAATA--TTCGCGAGATGTGCAT 684
xlPrP2_Brn1    CGATGTAAACCAGTTGGGATACCGCTGGTGAAATCCAAAATTATTCC-CGAGATGTGCAT 723
xlPrP2_Test    CGATGTAAACCAGTTGG-ATACCGTGGTG-AAATCCAAAATTA-TTCGCGAGATGTGCAT 495
               ********* ******* ******  *   ****** ****   * *  * * * **

xlPrP2_Eye     CCCTGATTACGGGGAAAAGACCGGGTCTTCNGGGGTGCTTTTCTAACCCAGGGGGATAAT 839
xlPrP2_Brn3    GAAGCGTCA----GTCACGACCACCCTTTGAGGAGTAGAN-------------------- 672
xlPrP2_Brn2    CACTGAATACAG-GAGAGGATCAGGATTCAAAGTGC---TTTCTAACCCATGGGC-TGAT 739
xlPrP2_Brn1    CACTGAATACAG-GAGAGGATCAGGATTCAAAGTGC---TTTCTAACCCATGGGC-TGAT 778
xlPrP2_Test    CACTGAATACAG-GAGAGGATCAGGATTCAAAGTGC---TTTCTAACCCATGGC--TGAT 549
                       *    *  * ** *     *    * *

xlPrP2_Eye     CCTCCAAAAACCACTTCCTTTTTGGTTAACTTTTGGTGAATAAAGTGAACCAGGAAGGGA 899
xlPrP2_Brn3    ------------------------------------------------------------
xlPrP2_Brn2    CCTC--ACTATCAC--CCTCTTTGGTTAACTTTGTG---ATAGAGTGACCCAGG--GGGA 790
xlPrP2_Brn1    CCTC--ACTATCAC--TCTCTTTGTTTACTTTTGGG---ATAGAGTGACCCAAA--GGGG 829
xlPrP2_Test    CCTC--ACTATCAC--TCTCTTTGTTTACTTTGTGA----TAGAGTGA-CCAGA---GGG 597

xlPrP2_Eye     ATTC-------------------------------------------------------- 903
xlPrP2_Brn3    ------------------------------------------------------------
xlPrP2_Brn2    ATTCGAAATCCAAAGGCCCATTTGTATGTAATATAGAAGAGTATCCAACCAATTTCTGAA 850
xlPrP2_Brn1    AATCCAAATCCCAAGGCCAATTGGTATGTTATATAGAAGAGTATCAACCCAATTCTGAAA 889
xlPrP2_Test    AATCGAAATCCAAAGGCCAATTGTATGTATATAGAGAGTATCAACCAATTCTGGACTGTC 657
```

Figure 3-7 Sequence analysis of *X. laevis* PrP2. (a) Pairwise alignment of *X. laevis* PrP1 and PrP2. The secondary structural components are shown above the sequence (arrows indicate cleavage site; Cys involved in disulphide bridge are represented by #). (b) The NMR structure of *X. laevis* PrP1 (PDB 1XU0) with the substituted residues between PrP1 and PrP2 labeled and highlighted in red. Model of PrP2 (modeled using Modeller with 1XU0 as template) is shown next to it for comparison. (c) 3' ends of PrP2 sequences from different tissue sources (Brn1, Brn2, Brn3 – brain; Test- testis). Bases highlighted in grey indicate sequences corresponding to ORF. Note the different stop codons.

Attempts to isolate a duplicate form of *PRND* in *X. laevis* were unsuccessful. This may be due to loss of this gene or to extreme divergence from the original sequence. The availability of genomic sequence for the *PRNP2* gene environment is necessary to resolve the question whether the duplicate form of *PRND* exists in *X. laevis*.

3.3.6 Tissue expression of *PRNP* and *PRND* in *X. laevis*

Tissue expression experiments using RT-PCR for *X. laevis PRNP*1, *PRNP*2 and *PRND* transcripts together with their variant forms were performed (Figure 3-8).

Expression of *PRND* variants: The expression of *PRND* variants is tissue specific with the chimeric form (*PRND-a*) expressed in a wide range of tissues and predominantly in brain. The non-chimeric variants (*PRND-b(1)* and *PRND-b(2)*) are restricted to testis, eye and pancreas and, significantly, show no expression in adult brain. This latter finding resembles the expression pattern of higher vertebrates (i.e. higher expression testis and none in brain). This differential expression is due to usage of two different promoters.

Expression of *PRNP1* variants: *PRNP*1 chimeric variant (*PRNP-1a*) showed expression in all the tissues analyzed suggesting a possible house-keeping function. The non-chimeric variant of *PRNP*1 (*PRNP-1b*) though widely expressed is, interestingly, absent in testis, also suggesting a different promoter usage.

Expression of *PRNP2* variants: *PRNP*2 forms also show expression in a range of tissues but are absent in a few. The coexistence of *PRNP1* and *PRNP2* in various tissues suggests that they may be performing slightly different functions and hence *PRNP2* was retained. Although there is general consistency in the expression patterns of *PRNP*2 variants (*PRNP-2b-1* and *PRNP-2b-1*), they appear to be expressed at different levels as evident by the relative contrast of the PCR bands. However, quantitative analysis using Real time PCR experiments was not performed.

Comparison of expression of *PRNP*1 with *PRNP*2: In general, expression between the two duplicate forms varies only slightly. The most significant difference is that *PRNP*2 is missing in the liver and in the gall bladder. This pattern matches that reported in mammals; expression of PrP was absent in liver of sheep (Horiuchi et al. 1995). It is tempting to speculate that *PRNP*2 functions have evolved to resemble those of *PRNP* in higher vertebrates, an idea also supported by the existence of small insertions in the N-terminal sequence of *PRNP*2.

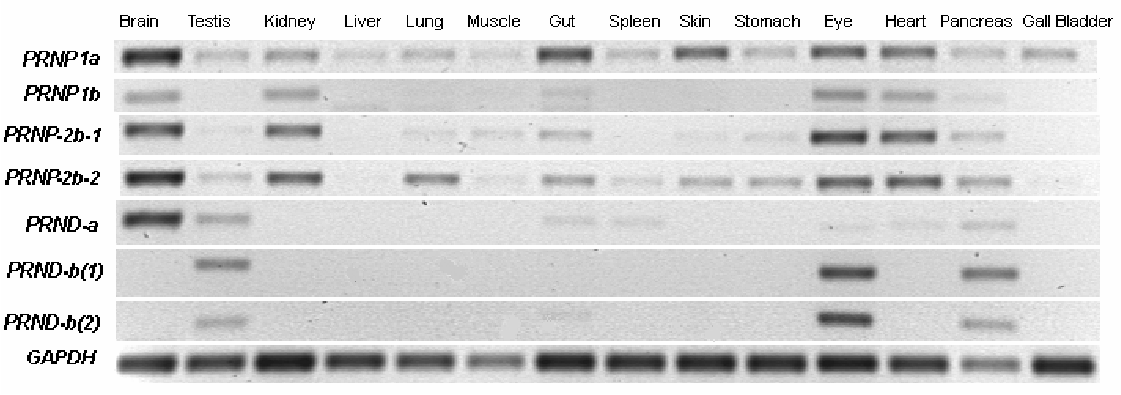


Figure 3-8 Compilation of gel picture from different experiments showing the expression pattern in different tissues along the horizontal axis and the different variants (see Figure 3-2) tested along the vertical axis. *GAPDH* was used as a positive control.

3.3.7 Genomic environment of *PRNP* in *X. tropicalis*

The release of the assembled version (version 3) of *X. tropicalis* (http://genome.jgi-psf.org/Xentr3/Xentr3.home.html) enabled me to define the complete genomic environment for *PRNP*. This version of the assembly shows *PRNP* and *PRND* sequence in Scaffold 110. The nucleotide sequence upstream (50 kb) and downstream (50 kb) to the *PRND* ORF was analysed by GenScan (Burge and Karlin, 1997). Although it failed to predict *PRND* and *PRNP*, *Rassf2* and *SCL23A1* were predicted in a position downstream to *PRND*. The organization and orientation of *PRNP*, *PRND*, *Rassf2* and *SCL23A1* is consistent with that of the mammalian *PRNP/PRND* environment indicating that this region has been conserved for 360 million years (Figure 3-9).

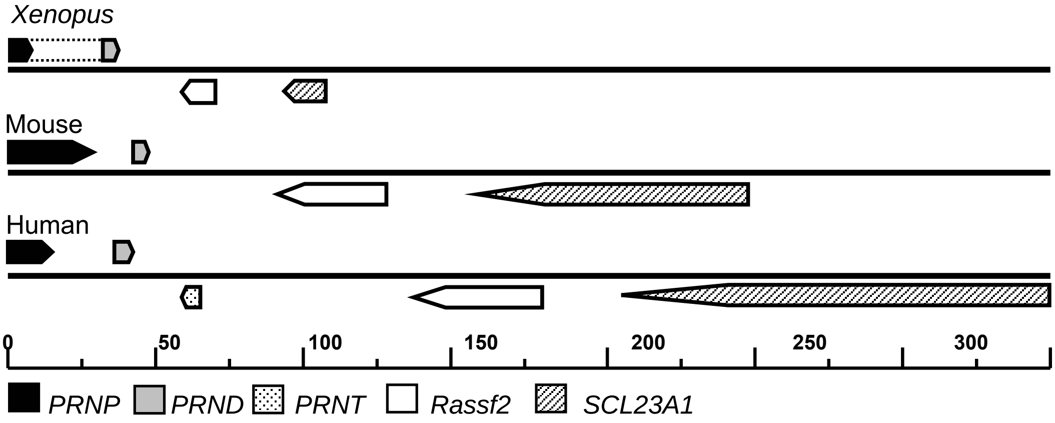


Figure 3-9 Genomic context of *PRNP* gene as in *Xenopus*, Mouse and Human. The gene organization and orientation is conserved. For *Xenopus*, the dotted lines represent the chimeric transcript.

3.3.8 Sequence analysis of *Xenopus* Dpl

Xenopus Dpl is slightly longer than the other known Dpl sequences (*X. laevis* 182 residues, *X. tropicalis* 181 residues, mouse 179 residues, human 176 residues). As in its mammalian counterpart, the Dpl sequence contains a sequence motif at the N-terminus, which encodes signals for targeting to cellular compartments like the ER for posttranslational modifications, as well as hydrophobic C-terminal region for addition of a GPI-anchor (Figure 3-10). The signal-peptide cleavage site is predicted by SignalP to be between residues 27 and 28 (RRA-AS) in *X. tropicalis,* and between residues 22 and 23 (GLS-CP) in *X. laevis* (Nielsen et al. 1997). Dpl has been reported to undergo other posttranslational modifications including addition of a GPI-anchor and N-glycosylation (Silverman et al. 2000). The Big-Pi tool was used to predict the C-terminal sites (Eisenhaber *et al*., 1999) for GPI-anchor attachment. These were Asn155 in *X. tropicalis* and Asn156 in *X. laevis* Dpls. *X. tropicalis* and *X. laevis* Dpl have two consensus N-glycosylation sites of form N-X-T. Although the second site (*X. tropicalis*: Asn 112) is conserved in position with other mammalian Dpls, the first site (*X. tropicalis*: Asn 70) is more N-terminally located than its mammalian counterpart (Figures 3-10 and 3-11). The location of the glycan attachment site can alter the behavior of glycoproteins, making them more soluble or stable, protecting them locally from proteolysis or aggregation, or masking their antigenic sites (Dwek 1996; Rudd et al. 1999;

Wormald and Dwek 1999; Helenius and Aebi 2001). Such factors explain the difference in the position of the first glycosylation site.

Qin et al. (2003) reported that His 131 binds to copper in mouse Dpl. *Xenopus* Dpl lacks this His residue. However, both bovine and ovine Dpl also lack His residue at this position. *Xenopus* Dpl has the least number of His residues (two) and none of its positions are conserved as compared with mammalian Dpls (Figure 3-11). Together, these results argue that copper binding is not an important function of Dpl.

The overall sequence conservation between *Xenopus* Dpl and mammalian Dpl is low. The well conserved mammalian sequence (AFI) of the first β-strand is not conserved in *Xenopus* Dpl (residues 58-60) which has DLS in *X. tropicalis* and HLF for *X. laevis* (Figure 3-11). The longest continuous stretch of identical residues in mammalian and *Xenopus* Dpls (residues 86-89) is "FPDG", corresponding to the region at the start of the second β strand.

The pairwise global alignment of *X. tropicalis* Dpl with other mammalian Dpl sequences showed a maximum sequence identity of 26% with ovine Dpl (45% similarity) and a minimum of 19 % in human (35% similarity). The low overall sequence conservation suggests that the functions of Dpl in these two vertebrate groups have diverged significantly since the divergence of amphibian from the common ancestor 360 million years ago (Kumar and Hedges 1998).

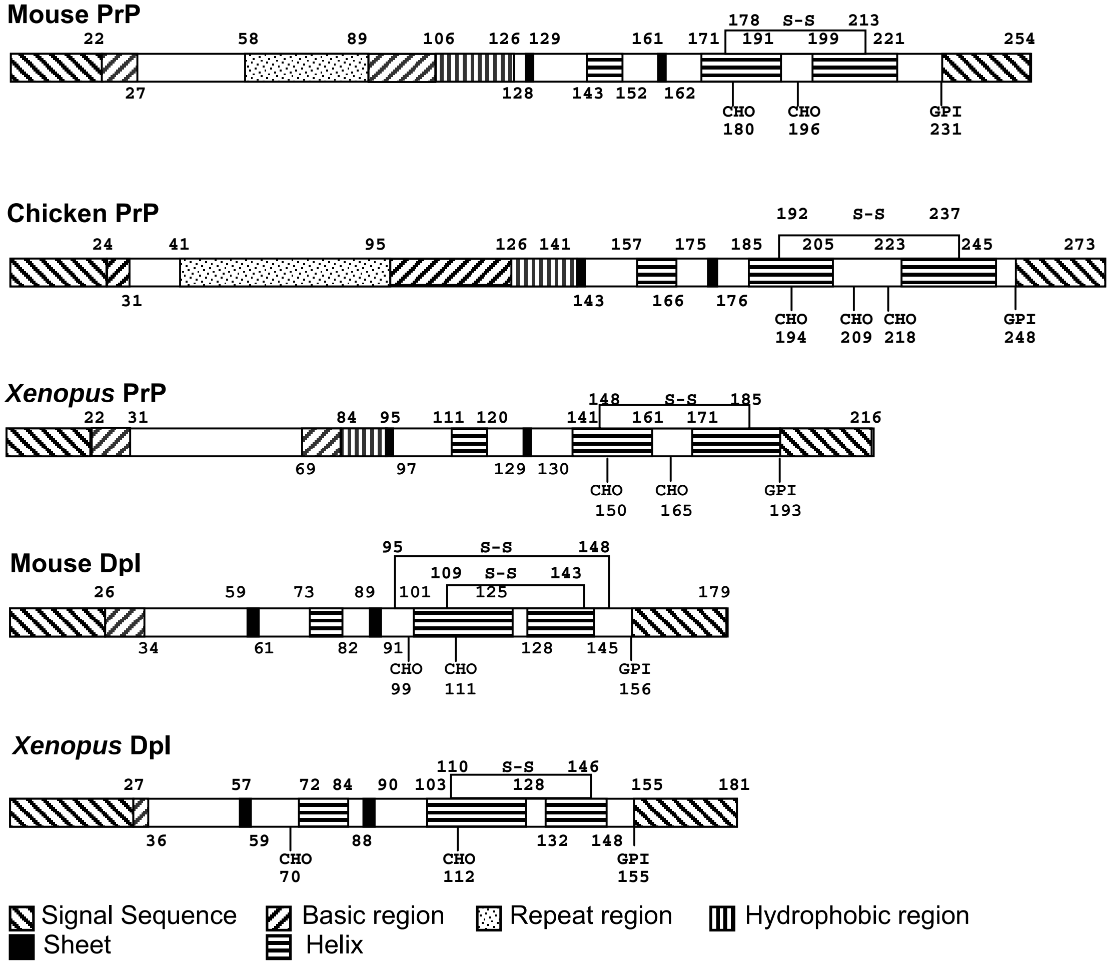
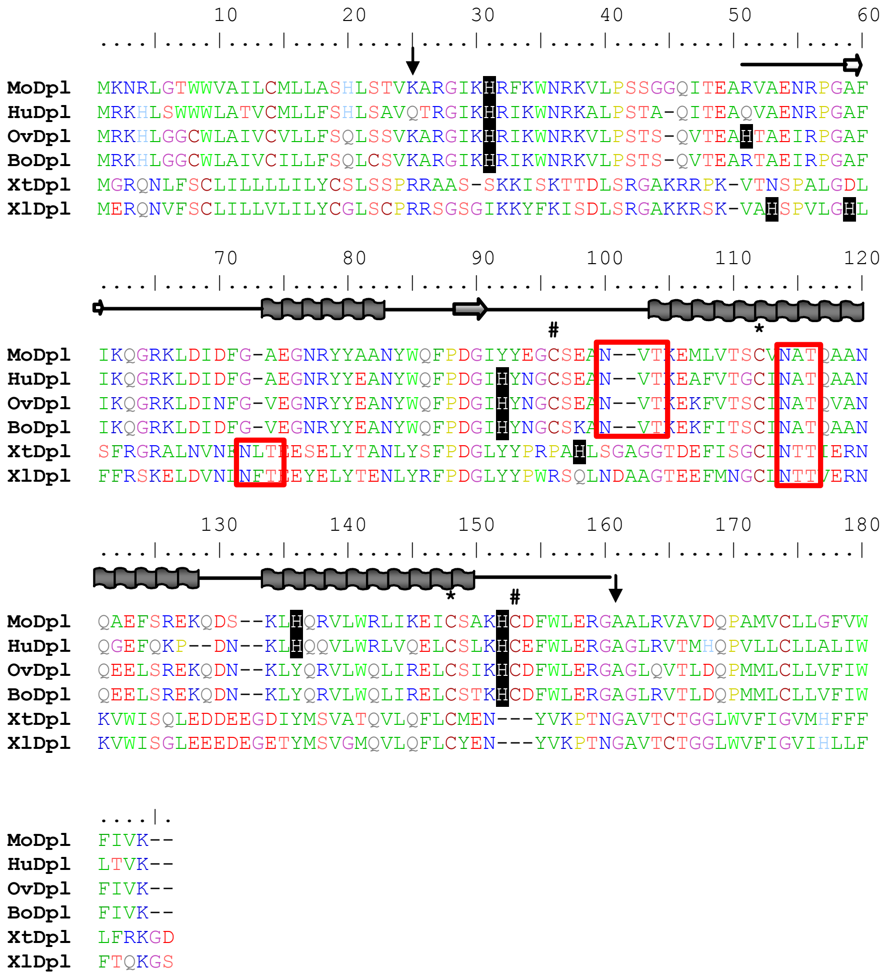


Figure 3-10 Schematic diagram of the structure of *X. tropicalis* Dpl and PrP, together with Mouse (Dpl and PrP) and Chicken (PrP) for comparison.



```
                10        20        30        40        50        60
       ....|....|....|....|....|....|....|....|....|....|....|....|
MoDpl  MKNRLGTWWVAILCMLLASHLSTVKARGIKHRFKWNRKVLPSSGGQITEARVAENRPGAF
HuDpl  MRKHLSWWWLATVCMLLFSHLSAVQTRGIKHRIKWNRKALPSTA-QITEAQVAENRPGAF
OvDpl  MRKHLGGCWLAIVCVLLFSQLSSVKARGIKHRIKWNRKVLPSTS-QVTEAHTAEIRPGAF
BoDpl  MRKHLGGCWLAIVCILLFSQLCSVKARGIKHRIKWNRKVLPSTS-QVTEARTAEIRPGAF
XtDpl  MGRQNLFSCLILLLLILYCSLSSPRRAAS-SKKISKTTDLSRGAKRRPK-VTNSPALGDL
XlDpl  MERQNVFSCLILLVLILYCGLSCPRRSGSGIKKYFKISDLSRGAKKRSK-VAHSPVLGHL

                70        80        90       100       110       120
       ....|....|....|....|....|....|....|....|....|....|....|....|
                                         #                 *
MoDpl  IKQGRKLDIDFG-AEGNRYYAANYWQFPDGIYYEGCSEAN--VTKEMLVTSCVNATQAAN
HuDpl  IKQGRKLDIDFG-AEGNRYYEANYWQFPDGIHYNGCSEAN--VTKEAFVTGCINATQAAN
OvDpl  IKQGRKLDINFG-VEGNRYYEANYWQFPDGIHYNGCSEAN--VTKEKFVTSCINATQVAN
BoDpl  IKQGRKLDIDFG-VEGNRYYEANYWQFPDGIHYNGCSKAN--VTKEKFITSCINATQAAN
XtDpl  SFRGRALNVNENLTEESELYTANLYSFPDGLYYPRPAHLSGAGGTDEFISGCINTTIERN
XlDpl  FFRSKELDVNINFTEEYELYTENLYRFPDGLYYPWRSQLNDAAGTEEFMNGCINTTVERN

               130       140       150       160       170       180
       ....|....|....|....|....|....|....|....|....|....|....|....|
                                  *    #
MoDpl  QAEFSREKQDS--KLHQRVLWRLIKEICSAKHCDFWLERGAALRVAVDQPAMVCLLGFVW
HuDpl  QGEFQKP--DN--KLHQQVLWRLVQELCSLKHCEFWLERGAGLRVTMHQPVLLCLLALIW
OvDpl  QEELSREKQDN--KLYQRVLWQLIRELCSIKHCDFWLERGAGLQVTLDQPMMLCLLVFIW
BoDpl  QEELSREKQDN--KLYQRVLWQLIRELCSTKHCDFWLERGAGLRVTLDQPMMLCLLVFIW
XtDpl  KVWISQLEDDEEGDIYMSVATQVLQFLCMEN---YVKPTNGAVTCTGGLWVFIGVMHFFF
XlDpl  KVWISGLEEEDEGETYMSVGMQVLQFLCYEN---YVKPTNGAVTCTGGLWVFIGVIHLLF

       ....|.
MoDpl  FIVK--
HuDpl  LTVK--
OvDpl  FIVK--
BoDpl  FIVK--
XtDpl  LFRKGD
XlDpl  FTQKGS
```

Figure 3-11 Multiple sequence alignment of Dpl sequences from six different species. The secondary structural units are indicated based on mouse Dpl structure (PDB 1I17). The N-glycosylation sites are indicated in red boxes. The Histidine residues are highlighted in black. Cys involved in disulphide bridge: *, #; Cleavage site: ↓. (MoDpl: Mouse Dpl, HuDpl: Human Dpl, OvDpl: Ovine Dpl, BoDpl: Bovine Dpl, XtDpl: *X. tropicalis* Dpl, XlDpl: *X. laevis* Dpl).

3.3.9 Comparison of C-terminal domain

As the putatively folded domain of *Xenopus* Dpl is more conserved than the apparently disordered N-terminal region of the protein, the C-terminal domains of PrP and Dpl were compared among *Xenopus* and mammals. A comparison of PrP with Dpl within the same species was performed to calculate the percentage identity (Figure 3-12). A higher sequence identity might be expected between *Xenopus* Dpl and PrP compared with that for mammals, considering it to be at an early stage after duplication. But the percentage similarity between the C-terminal domain of mouse PrP and Dpl was 42%, which is only slightly higher than that between *Xenopus* PrP and Dpl, at 39%. The latter finding suggests rapid mutations have occurred during 360 million years of independent evolution as the functions of amphibian Dpl have evolved.

(a)

```
                                    ↓
MoDpl    1 MKNRLGTWWVAILCMLLASHLSTVKARGIKHRFKWNRKVLPSSG--GQITEA----RVAE 54
           | | ||  |   |           | |             |  |  |
MoPrP    1 MAN-LG-YWLLALFVTMWTDVGLCKKRPKPGGWNTGGSRYPGQGSAGAAAAGAVVG--GL 124
                                   ↑
                                                     44-111

                                                    ●   ▲         ○ ♦
MoDpl   55 NRPGAFIKQGRKLDIDFG-AEGNRYYAANYWQFPDGIYYEGCSEANVTKEMLVTSCVNAT 113
                         | ||     |||  |    |   ||             |  ||| |
MoPrP  125 GG-YMLGSAMSRPMIHFGNDWEDRYYRENMYRYPNQVYYRPVDQ-YSNQNNFVHDCVNIT 182

                                           ○       ●       ↓
MoDpl  114 QAAN-QAEF--SREKQDSKLHQRVLWRLIKEICSAKH---CDFWLERGAA------LRVA 161
                                     |     |                         |
MoPrP  183 IKQHTVTTTTKGENF--TETDVKMMERVVEQMCVTQYQKESQAY-----YDGRRSSSTVL 235
                        ▼                                          ↑

MoDpl  162 VDQPAMVCLLGFVWF-IVK 179
              |    |  |  | ||
MoPrP  236 FSSPPVILLISFLIFLIVG 254
```

(b)

```
                                     ↓
XtDpl    1 MGRQNLFSCLILLLLILYCSLSSPRRAAS-SKKISKTTDLSRGAKRRPKVTNS------- 52
           | |      | |  |   | |          |        |
XlDpl    1 MERQNVFSCLILLVLILYCGLSCPRRSGSGIKKYFKISDLSRGAKKRSKVAHS------- 53
           | |      | |  |   | |          |        |
XlPrP    1 MLR-SLWTSLVLISLV--CALTVSSKKSGSGK-SKTGGWNSGSNRNPNYPGGYAVGAAAG 90
                                   ↑                         50-83

                             ▲                                       ○
XtDpl   52 -PALGDLSFRGRALNVNFNLTEESELYTANLYSFPDGLYYPRPAHLSGAGGTDEFISGCL 111
                 |          :|   |:  |       |   | |           : |   |
XlDpl   53 -PVLGHLFFRSKELDVNLNFTEEYELYTENLYRFPDGLYYPWRSQLNDAAGTEEFMNGCL 112
                 |          :|   |:  |       |   | |           : |   |
XlPrP   91 AIGGYMLGNAVGRMSYQFNNPMESRYYNDYYNQMPNRVYRP-MYRGEEYVSEDRFVRDCY 149

           ♦                                  ○         ↓
XtDpl  112 NTTIER-NKVWISQLEDDEEGDIYMSVATQVLQFLCMENYVKPT-NG-AVTCTGGLWVFI 168
           |                 :       |  |     |   |      |         |
XlDpl  113 NTTVER-NKVWISGLEEEDEGETYMSVGMQVLQFLCYENYVKPT-NG-AVTCTGGLWVFI 169
           |  |         |    :       |  |     |   |      |         |
XlPrP  150 NMSVTEYIIKPAEGKNNSELNQLDTTVKSQIIREMCITEY----RRGSGFKVLSNPWLIL 205
                           ▼                               ↑

XtDpl  169 GVMHFFFLFRKGD 181
               :
XlDpl  170 GVIHLLFFTQKGS 182
               :  |
XlPrP  206 TITLFVYFVIE-- 216
```

Glycosylation: ♦ PrP, Dpl; ▲ Dpl; ▼ PrP;

○ S-S PrP, Dpl; ● 2^{nd} S-S Mo Dpl;

↓↑ Signal sequence; ⇧ Truncated repeats.

Figure 3-12 Sequence comparison of PrP and Dpl in (a) mouse and (b) *Xenopus*. Identical residues between PrP and Dpl are shown by "|". Identical residues between XtDpl and XlPrP not present in XlDpl are represented by ":". Secondary structural units: Helix; Sheet

3.3.10 Comparison of Dpl sequences

Peoch et al. (2000) reported the possible involvement of genetic variation in (human) Dpl in the etiology of human prion diseases. Four polymorphisms in *PRND* (three protein coding changes, human- T26M, P56L and T174M and a silent polymorphism, T174T) were detected but no strong association was reported between any of these polymorphisms and human prion diseases. There was some indication for a role in other human diseases. None of these residues is conserved in *Xenopus* suggesting these residues do not play a critical functional role, or if so then it was acquired at a later stage of evolution.

3.3.11 *Xenopus* Dpl model

A homology model for *Xenopus* Dpl built by the program MODELLER using mouse Dpl as template (Figure 3-13b) exhibited a similar structural fold except that the plane of the first β-strand (β1) is perpendicular to the long axis of the helices rather than being parallel. This feature resembles that for mammalian or *Xenopus* PrPs (Calzolai et al. 2005). The notable difference is that *Xenopus* Dpl has only one disulphide bridge as opposed to two in mammalian Dpl, suggesting that mammalian Dpl has acquired the second disulphide bridge at the later stages of evolution for more structural stability. Significantly, PrP in all species has only one disulphide bridge in the position corresponding to that in *Xenopus* Dpl. One of the N-glycosylation sites is displaced from the loop region between β2 and α2 to the loop region between β1 to α1 (Figure 3-9, Figure 3-11, Figure 3-13a,b). ConSurf (Glaser et al. 2003), which uses a color coding system to represent the degree of conservation, was used to plot the most conserved residues onto the mouse Dpl structure (Figure 3-13a). Amino acid residues that are critical for structure and function are expected to be conserved throughout evolution and though these conserved residues may be widely distributed in the sequence, some usually show spatial proximity. There are two major regions of the folded domain where groups of residues (clusters) are conserved.

The first region (group-1) is associated with α1 and its preceding loop, β1 and its preceding loop, and α3 (Gly-64, Arg-65, Leu-67, Phe-71, Glu-74, Tyr-79, Ala-81, Asn-82, Phe-86, Pro-87, Asp-88, Gly-89 and Val-34) (Figure 3-13c). The conserved residues are located largely in the interior, shielded from solvent, and are mainly constituted by hydrophobic residues. In the second region (group-2), the six residues (Gly-58, Tyr-92, Cys-109, Asn-111, Thr-113 and Asn-117) are closely associated (Figure 3-13c). The other conserved residues include Ser-96, Asn-99, Glu-93, Leu-95, Ser-122, Asp-127 (Figure 3-13). These observations suggest that part of the sequence conservation is associated with the maintenance of structural integrity of the protein, and that whatever functional similarity is preserved between Dpl in higher and lower vertebrates may be associated with these regions and the remaining conserved residues. Mouse Cys-148 and Cys-95 are involved in the second disulphide bridge formation; *Xenopus* Dpl lacks both these Cys residues (Figure 3-12).

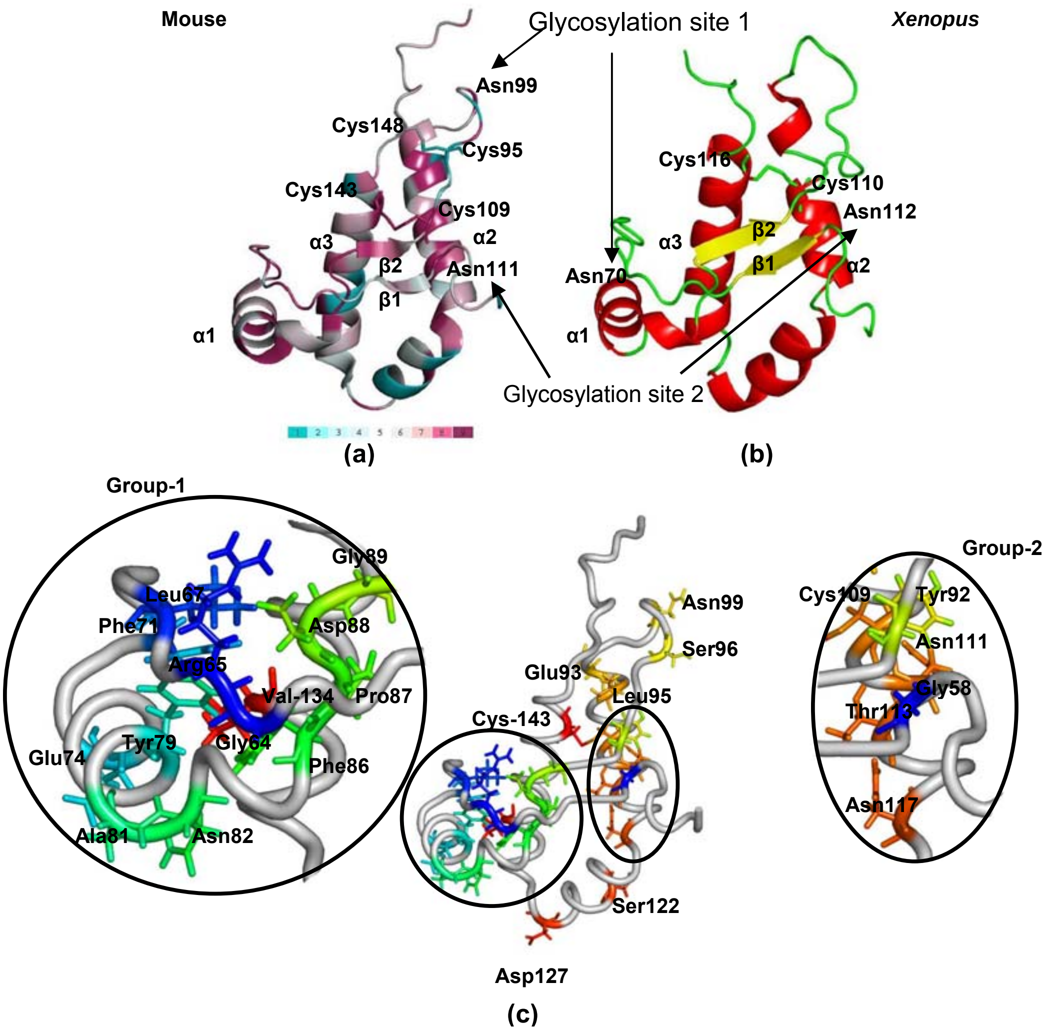


Figure 3-13 Analysis of *Xenopus* Dpl model and conserved residues. (a) Conservation of the residues projected on mouse Dpl (51- 157) (PDB id 1I17) based on sequences from human, mouse, sheep, cow and *Xenopus*. (b) A predicted model of *Xenopus* Dpl. Cys and Asn involved in disulphide bridge formation and N-glycosylation are labeled. Note the difference in the Glycosylation site 1 between mouse and *Xenopus*. (c) Highly conserved residues are grouped into clusters in three dimensional spaces though the residues are widely separated at the primary sequence level. Scale of conservation to interpret the color code in (a).

3.4 Conclusion

This is the first evidence for Dpl in early vertebrates. The organization of the *PRNP/PRND* locus and of the cDNAs strongly suggests the existence of a common primary dicistronic transcript, which is processed to give one or other of the two alternative mature transcripts. This suggests a functional requirement for coordinated expression of *PRNP* and *PRND.* In addition, non-chimeric transcripts for *PRNP* and *PRND* were found in lower abundance due to a 'juvenile' promoter.

In this scenario, the presence of non-chimeric transcripts indicates an inclination to evolve as two independent genes as is the case in higher vertebrates. The co-existence of chimeric and non-chimeric transcripts suggests an effective mechanism whereby an 'infant' gene, apparently tandemly duplicated without its promoter, can survive and begin to develop novel functions. It does this through sharing the promoter of its gene 'mother', but while also developing its own independent promoter *de novo*. This is evident by the differential expression of the dicistronic and monocistronic *PRND* transcripts. The former resembles the widely distributed *PRNP* expression, including high expression in brain, whereas the latter shows a more restricted pattern similar to that of *PRND* in adult mammals, specifically high expression in testis and none in brain. The existence of a chimeric *PRND* transcript was first reported in mouse (Moore et al. 1999). However, this transcript was expressed at extremely low levels and the predominant transcripts are that of non-chimeric form. The transformation of the predominant chimeric *PRND* transcript in *Xenopus* to a predominant non-chimeric *PRND* transcript in mouse strongly suggests a process of functional specialization with development of its own stronger promoter in higher vertebrates.

The other finding for duplicates is that both *PRNP* genes (*PRNP1* and *PRNP2*) are retained in the tetraploid *X. laevis*, but only one *PRND* gene appears to have been retained. Several studies have reported retention and diversification by subfunctionalization of genomic gene duplicates in *X. laevis* (Wu et al. 2003). The findings for *PRNP* provided an opportunity to compare expression patterns and protein sequence of the two variants during this quite short time from duplication, as well as compare these features with the single *PRNP* gene of the diploid *X. tropicalis*. The existence of a duplicate of *PRNP* in *X. laevis* raises questions about its role of either subfunctionalisation or neofunctionlisation. However, the wide expression of PrP1 and PrP2 in almost all the tissues tested may not support the subfunctionlisation theory wherein a complementary loss of tissue expression would have been expected. This may be due to slightly diverged functions being performed by both the proteins.

The discovery of the primitive form of Dpl enabled me to perform sequence comparison which led to the identification of key residues that may be critical for function or fold stability.

4 Findings in chicken

4.1 Background

Chicken PrP was the first reported avian PrP (Harris et al. 1991; Gabriel et al. 1992). The previous structural analysis performed on chicken PrP suggested that the N-terminal region is structurally different from the corresponding region in mammalian PrP (Marcotte and Eisenberg 1999). It was also suggested that the mature chicken PrP, unlike its synthetic peptide (Hornshaw et al. 1995), does not bind to copper (Marcotte and Eisenberg 1999), one of the proposed functions of mammalian PrP (Brown et al. 1997). This probably is the result of the residues involved in copper binding being buried by the rest of the protein (Marcotte and Eisenberg 1999). Wopner et al. (1999) performed a comparative analysis of mammalian and avian species. They observed that though the identity among avian PrPs was about 90%, it reduced to 30% identity compared with mammalian PrPs, with about 55% identity at the C-terminal domain (Gabriel et al. 1992). Within reported avian PrPs, chicken PrP was the most divergent species (Wopfner et al. 1999). The NMR structure of the recombinant PrP from chicken has the same molecular architecture as mammalian PrP (Calzolai et al. 2005).

In order to understand the evolutionary events that took place in the *PRNP* locus between amphibian and higher vertebrates, I analyzed the avian *PRNP* gene region, based on chicken for which the genome sequence is available (Wallis et al. 2004). Interestingly, this locus differed significantly from that in *Xenopus* (Chapter 3). It lacked the expected *PRND*, and instead I discovered a novel coding region found downstream and in opposite orientation to *PRNP* with similarities to Sho, which I named *Sho-like*.

In an attempt to understand the evolutionary events that led to an apparent loss of the *PRND* gene and the existence of the *Sho-like* gene in chicken, I investigated two species which are evolutionarily more closely related to chicken: emu (another bird) and alligator (a reptile).

4.2 Materials and Methods

4.2.1 Computational analysis

The chicken genome was screened for Dpl using the homology-based method, TBLASTN (Altschul et al. 1990). Various BLAST parameters (repeat masker turned off, gapped BLAST) were tried using a range of different query sequences. Initially, a local BLAST database was created for chicken chromosome 22 containing the *PRNP* gene locus. This was later expanded to the whole genome. Six-frame translation and the ORF analysis was performed using the EMBOSS applications (Rice et al. 2000) sixpack and getORF. *Ab-initio* gene predictions were performed using GenScan (http://genes.mit.edu/GENSCAN.html). Mapping of exons onto genomic DNA was performed using the EMBOSS application est2genome. Signal peptide sequence was predicted using SignalP (http://www.cbs.dtu.dk/services/SignalP/), GPI anchor attachment by big-Pi (http://mendel.imp.ac.at/gpi/gpi_server.html), nuclear localization signal (NLS) using PredictNLS (http://cubic.bioc.columbia.edu/cgi/var/nair/resonline.pl), and topology prediction using TopPred (http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html). Multiple sequence alignments were performed using ClustalW (http://www.ebi.ac.uk/clustalw/). Alignments were edited using BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The sequences used in the analysis are listed in Table 4-1. Phylogenetic analysis was performed using a maximum likelihood method, PROTML implemented by MOLPHY (Adachi and Hasegawa 1996) with the JTT-F model for amino acid substitution. Avian PrP sequences were obtained by BLASTing chicken PrP sequence against the NCBI nr protein sequence database (Table 4-1). Platypus PrP sequence was obtained from the NCBI trace archive by BLASTing *M. domestica* PrP sequence using TBLASTN (see Chapter 5).

Table 4-1 Sequences used for multiple sequence alignments

Sequence	Source
Mammal	
Human	ref|NP_001009093.1|
Mouse	ref|NP_035300.1|
Rat	ref|NP_036763.1|
Sheep	ref|NP_001009481.1|
Cow	ref|NP_851358.1|
Reptile	
Pelodiscus	BAC66701.1
Turtle	CAB81568.1
Bird	
Taeniopygia	*CAL59565.1*
Balearica	*AAD47046.1*
Pachyptila	*AAD47050.1*
Melopsittacus	*AAR21237.1*
Chicken	NP_990796.1
Anas	AAF82604.1
Columba	AAF73436.1
Coturnix	AAF73437.1
Pavo	AAR21236.1
Tyto	AAD47049.1
Vultur	AAD47045.1
Pelodiscus	BAC66701.1
Ostrich	AAD47048.1
Fish	
Fugu **Sho2**	CAG34292.1
Tetraodon **Sho2**	AL239301
Zebrafish Sho2	AAT72771.1
Fugu **PrP-like**	Scaffold_7 (version 4): 2849936 to 2850464
Tetraodon **PrP-like**	Chromosome 12 (version 7)
Zebrafish PrP-like	Chr:8 (version 7): 38803941 to 38804505
Monotreme	
Platypus	Trace read: 723211856; Assembly version 5: Contig 9565

Pairwise alignments for the sequences listed in Table 4-2 were performed using standalone AVID (Bray et al. 2003). The alignments were annotated using VISTA (Frazer et al. 2004).

Table 4-2 Sequences used for comparison between fish and chicken *PRNP* region. (Zebrafish assembly version 7; Fugu assembly version 4)**.**

Sequence	Source
Zebrafish *stPrP1 SPRNB*	Chromosome:ZFISH5:10:13945664-13964779
Zebrafish *stPrP2*	Chromosome:ZFISH5:8:36362134-36390056
Fugu stPrP2 PrP-like	Scaffold:*FUGU*4:scaffold_7:2849455-2854674
Fugu stPrP1 SPRNB	Scaffold:*FUGU*4:scaffold_84:78890-87871
Chicken *PRNP Sho-like*	Chromosome:WASHUC2:22:433592-444383

4.2.2 Isolation and cloning of *Sho-like* gene

Tissues were collected from a freshly dissected adult chicken and rooster (provided by Stuart Wilkinson, The University of Sydney Faculty of Veterinary Science) and were stored in RNA*later*® (Ambion) until use. The Qiagen RNAeasy kit was used to extract total RNA. RNA quality was checked using the Bioanalyser (Agilent) and quantification was performed on the NanoDrop (equipment in the BRF, JCSMR, ANU). cDNA synthesis using 1 μg of RNA was performed using the Invitrogen Superscript™ III First-Strand cDNA Synthesis kit using random hexamers. I started with an assumption that this gene may be expressed at high levels in brain based on available information from the literature for PrP and Sho transcripts and hence first used the cDNA from brain. Two primer sets (Lslike1, Rslike1; Lslike2, Rslike2) (Table 4-3) were designed from the predicted ORF. Both these primer sets were used to amplify cDNA using Platinum®*Taq* DNA polymerase (Invitrogen) (4pmol of primer and 50 ng of cDNA as template) with PCR conditions 94°C for 30sec, 57°C for 30sec, 72°C for 1min. One band of expected size was obtained. In order to obtain the 3' and 5' ends, RACE technique was employed. SMART™ PCR cDNA Synthesis Kit (BD Bioscience) was used with gene specific primers (5'slike-GSP, 5'slike-nGSP, 3'slike-GSP, 3'slike-nGSP) and 5' universal primers. An additional step of amplifying the cDNA by using single GSP primer was employed for 30 cycles and the resultant product was used for RACE PCR. This step improved amplification as the regular RACE PCR did not give any specific band. The PCR product was cloned into TOPO-TA Cloning® kit (Invitrogen). Sequencing reactions were performed at the Biomolecular Resource

Facility, John Curtin School of Medical Research using the reagents and protocol provided by them.

Table 4-3 Summary of primers used in RT-PCR assays

Primer	Sequence
Lslike1	GGTGGAGGAGGAGGGCTAC
Lslike2	GCCCTGTTCAGCAACACTG
Rslike1	CCAGAAGTAGAGTGGGGACAA
Rslike2	TCTGCTGGGTCAACCATG
Lslike-ex1	GGAAGGAGGCACAGGAAAC
5'slike-GSP	TTGCCAGAAGTAGAGTGGGGACAATGC
3'slike-GSP	ACAGGGTATGGGATGGGGCTCCT
5'slike-nGSP	AGGAGCCCCATCCCATACCCTGT
3'slike-nGSP	TGTCCCCACTCTACTTCTGGCAAA
Lgapdh	TCTGGCAAAGTCCAAGTGGT
Rgapdh	AGAACTGAGCGGTGGTGAAG
Lsho	GGGACGTGGGAGCGGAAG
Rsho	CCTCAGCGATTCAACAGTGA
Lprp	GGCTTCTTGGATCGCTCATA
Rprp	TACCACAACCAGAAGCCATG

4.2.3 Emu and alligator search

Emu and alligator genomic DNA samples were obtained as a gift from Prof Scott Edwards and Dr Dan Janes (Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University).

Homology search: Whole-genome sequence for alligator (*Alligator mississippiensis*) and emu (*Dromaius novaehollandiae)* are not available, but the NCBI trace archive has ~40,000 trace reads for both alligator and emu. A local BLAST-searchable database was created for these sequences to facilitate homology searches.

PCR-based methods: Evolutionary PCR using non-specific primers designed on the basis of known nucleotide sequences was performed using three different methods.

Method 1: Primers were designed from sequences evolutionarily close to alligator and emu. Turtle and chicken sequence was used for PrP, and *Xenopus* sequence was used for Dpl (Table 4-4).

Method 2: Vector NTI module "alignment PCR" was used to design primers from the input multiple nucleotide sequence alignment corresponding to ORF for PrP obtained from the following sequences: human, rat, mouse, cow, sheep, *M. domestica*, and turtle. For Dpl, the following sequences were used: human, mouse, rat, cow, sheep, *M. domestica*, and *Xenopus* (Table 4-4).

Method 3: Degenerate primers based on the most conserved amino acid sequence regions for PrP and Dpl were constructed (using same species as in Method 2). The degenerate primers used for isolating turtle PrP (Simonic et al. 2000) were also tried (Table 4-4). These primers were used with touch down PCR with annealing temperature ranging from 5 °C above to 5 °C below the minimum melting temperature.

Table 4-4 Degenerate primers used for evolutionary PCR. Degenerate code: M (AC), R (AG), W (AT), S (CG), Y (CT), K (GT), V (ACG), H (ACT), D (AGT), B (CGT), N (ACGT)

Name	Sequence	Comment
rdgDpl252-n	GTARTRGATBCCRTCIGGRAA	Primers based on nucleotide conservation
ldgDpl89-n	ARCAIARAAWTAAGYVRAAC	
ldgDpl187-n	GVAAGCTBGWYRTIAACTTY	
rdgDpl86-a	RTARTADATNCCRTCIGG	Primers based on amino acid sequence
rgDpl108-a	ACRAAYTTACGITGN	
ldgDpl86-a	CCIGAYGGIATHTAYTAY	
ldDpl69-a	GAYTTYGGIGARGARGGIAAN	
ldgPrP-na	TGGGRMGYGYWATGTCAGG	Based on nucleotide conservation between reptile and chicken
rdgPrP-na	TAYTGCTGCAYGCACATCTC	
ldgPrP-aa	CCIAARACIAATATGAAR	Based on amino acid conservation
LdgPrP	ATAAACCCAAAACCAACATG	
RdgPrP	CACATCTCCTGGATCACTTGC	
ldgPrP-pp	CCCAACCRIGTNTACTAC	(Simonic et al. 2000)
rdgPrP-pp	AYIGTIATRTTIANRCARTC	

4.2.4 Tissue expression

Total RNA isolation and cDNA synthesis for the following chicken tissues were performed: brain, testis, liver, heart, kidney, lung, eye, gallbladder, muscle, spleen, pancreas, skin, gut, stomach, ovary, and uterus. Tissue expression was studied for

PrP, Sho-like and Sho transcripts with GAPDH as control. Primers (Sho-like: Lslike-ex1, Rslike2; PrP: Lprp, Rprp; Sho: Lsho, Rsho; GAPDH: Lgapdh, Rgapdh) listed in Table 4-3 were used with PCR conditions 57 °C and 30 cycles using GoTAQ® green master mix (Promega, USA).

4.3 Results and Discussion

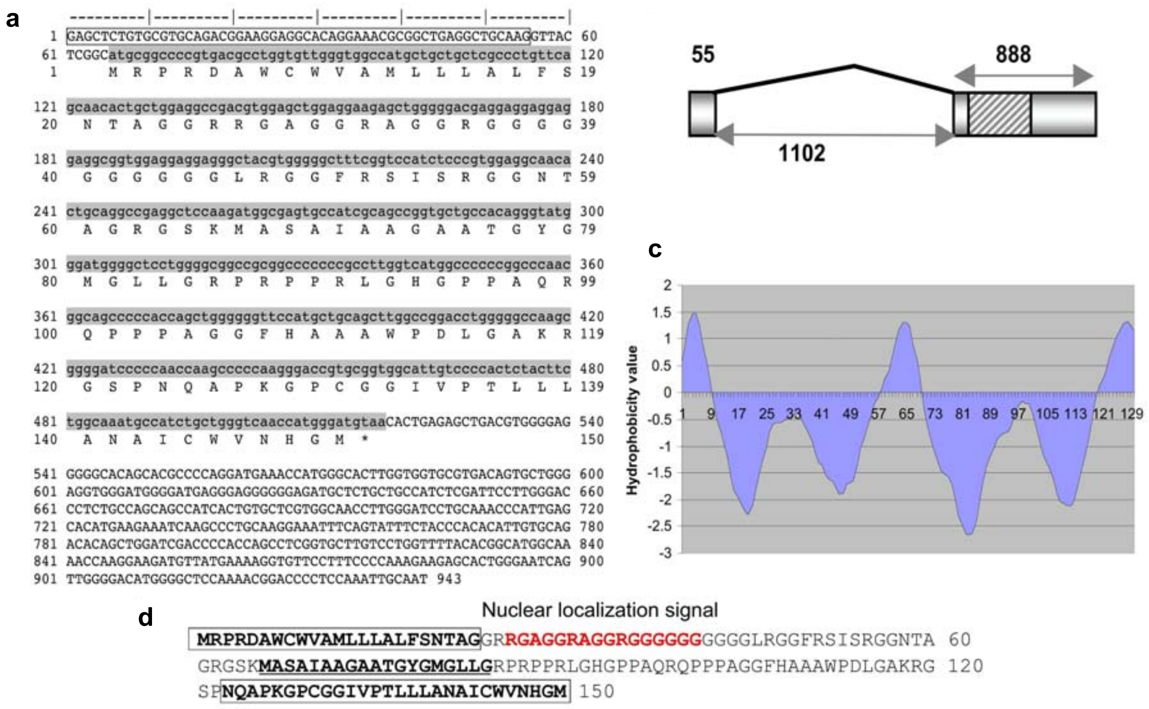
4.3.1 Analysis of the chicken *PRNP* genomic locus

Homology-based searches and gene prediction using *ab-initio* methods on chromosome 22 (using various search strategies) suggests the absence of the Dpl gene, *PRND*. While it is possible that the *PRNP* gene region has undergone rearrangements leading to the shuffling of the gene position, this seems unlikely as the homology search of the whole genome also did not produce any significant results. Lack of rearrangement is further supported by the conservation in the position and orientation of other adjacent genes, *SCL23A1* and *RASSF2,* reported in vertebrates from fish to mammals (Suzuki et al. 2002; Premzl et al. 2004). To further investigate the *PRNP* gene locus, I performed a six-frame translation of the sequence downstream from the *PRNP* gene to look for traces of *the PRND* gene or its remnants (pseudogene). Although no region similar to *PRND* was found, an ORF with sequence homology to Sho was found, which I call chicken Sho-like.

An EST database search for this new gene did not produce any significant hits. To clarify the relationship of this gene to Sho, a homology search was performed using the human Sho sequence against the chicken genome database. This confirmed that a true orthologue of Sho is present at a different genomic location (chromosome 6), where it is flanked by genes previously reported to be conserved at the Sho genomic locus (Premzl et al. 2003; Premzl et al. 2004).

4.3.2 Isolation and molecular characterization of *Sho-like* cDNA

A single transcript was isolated from chicken brain RNA. Mapping this sequence onto the genomic sequence revealed 2 exons (Figure 4-1). As for all other PrP family genes (PrP, Dpl and Sho, and the fish genes), the entire ORF is found within a single exon (exon 2) (Premzl et al. 2004), which codes for a protein of 150 residues. The gene contains a 65 bp 5' UTR and a 425 bp 3' UTR. The intron between the two exons is 1102 bp (Figure 4-1b).



```
a
        ---------|---------|---------|---------|---------|---------|
  1 GAGCTCTGTGCGTGCAGACGGAAGGAGGCACAGGAAACGCGGCTGAGGCTGCAAGGTTAC 60
 61 TCGGCatgcggccccgtgacgcctggtgttgggtggccatgctgctgctcgccctgttca 120
  1      M  R  P  R  D  A  W  C  W  V  A  M  L  L  L  A  L  F  S 19

121 gcaacactgctggaggccgacgtggagctggaggaagagctgggggacgaggaggaggag 180
 20  N  T  A  G  G  R  R  G  A  G  G  R  A  G  G  R  G  G  G  G 39

181 gaggcggtggaggaggagggctacgtgggggctttcggtccatctcccgtggaggcaaca 240
 40  G  G  G  G  G  G  L  R  G  G  F  R  S  I  S  R  G  G  N  T 59

241 ctgcaggccgaggctccaagatggcgagtgccatcgcagccggtgctgccacagggtatg 300
 60  A  G  R  G  S  K  M  A  S  A  I  A  A  G  A  A  T  G  Y  G 79

301 ggatggggctcctggggcggccgcggcccccccgccttggtcatggcccccggcccaac 360
 80  M  G  L  L  G  R  P  R  P  P  R  L  G  H  G  P  P  A  Q  R 99

361 ggcagcccccaccagctgggggttccatgctgcagcttggccggacctgggggccaagc 420
100  Q  P  P  P  A  G  G  F  H  A  A  A  W  P  D  L  G  A  K  R 119

421 ggggatcccccaaccaagcccccaagggaccgtgcggtggcattgtccccactctacttc 480
120  G  S  P  N  Q  A  P  K  G  P  C  G  G  I  V  P  T  L  L  L 139

481 tggcaaatgccatctgctgggtcaaccatgggatgtaaCACTGAGAGCTGACGTGGGGAG 540
140  A  N  A  I  C  W  V  N  H  G  M  *                          150

541 GGGGCACAGCACGCCCCAGGATGAAACCATGGGCACTTGGTGGTGCGTGACAGTGCTGGG 600
601 AGGTGGGATGGGGATGAGGGAGGGGGGAGATGCTCTGCTGCCATCTCGATTCCTTGGGAC 660
661 CCTCTGCCAGCAGCCATCACTGTGCTCGTGGCAACCTTGGGATCCTGCAAACCCATTGAG 720
721 CACATGAAGAAATCAAGCCCTGCAAGGAAATTTCAGTATTTCTACCCACACATTGTGCAG 780
781 ACACAGCTGGATCGACCCCACCAGCCTCGGTGCTTGTCCTGGTTTTACACGGCATGGCAA 840
841 AACCAAGGAAGATGTTATGAAAAGGTGTTCCTTTCCCCAAAGAAGAGCACTGGGAATCAG 900
901 TTGGGGACATGGGGCTCCAAAACGGACCCCTCCAAATTGCAAT 943
```

```
MRPRDAWCWVAMLLLALFSNTAGGRRGAGGRAGGRGGGGGGGGGLRGGFRSISRGGNTA 60
GRGSKMASAIAAGAATGYGMGLLGRPRPPRLGHGPPAQRQPPPAGGFHAAAWPDLGAKRG 120
SPNQAPKGPCGGIVPTLLLANAICWVNHGM 150
```

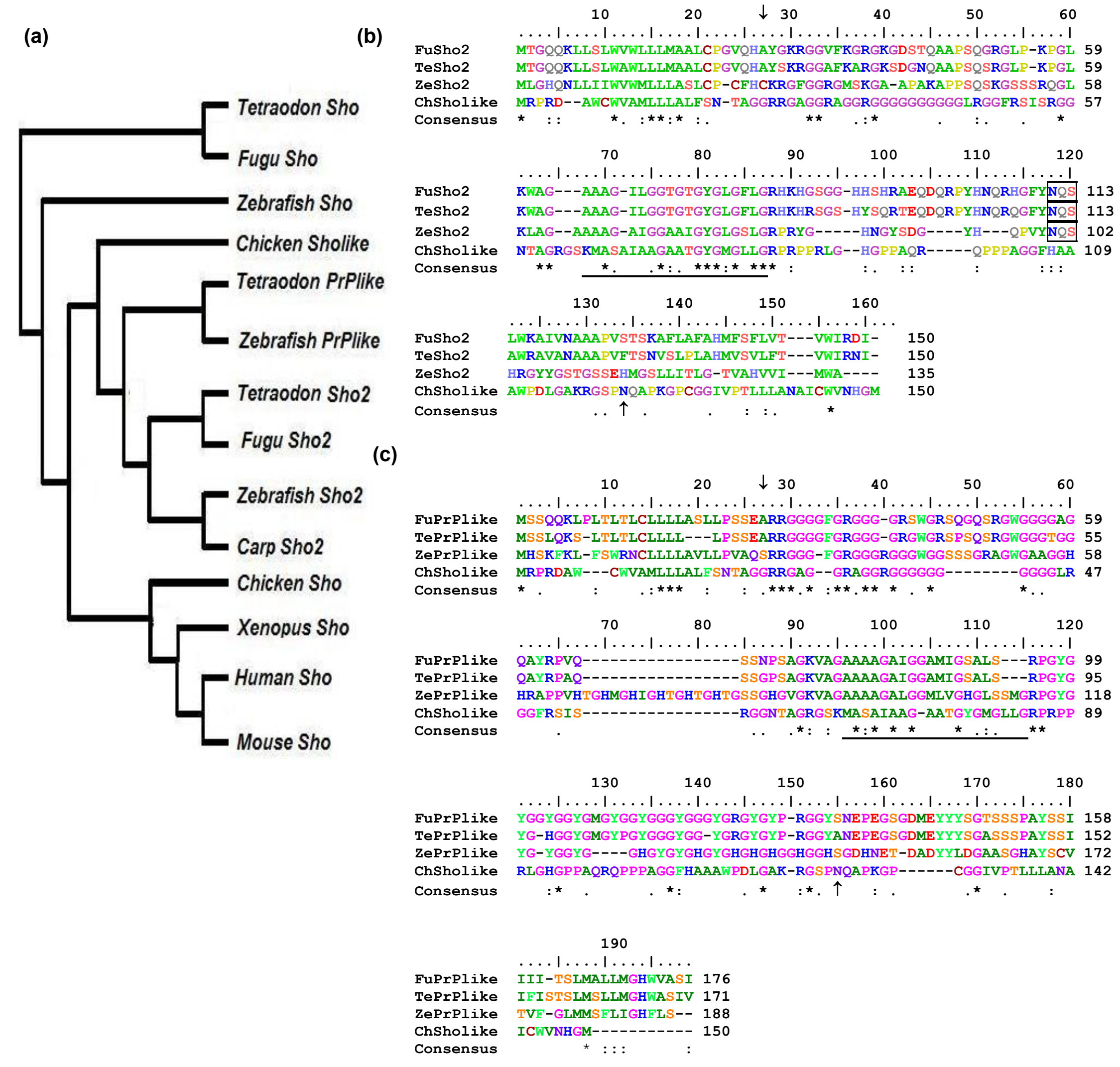
Figure 4-1 Sequence features for *Sho-like* gene. (a) cDNA sequence (brain) obtained from 5' and 3' RACE. Sequence for first exon is shown boxed and ORF in grey with translation shown below the nucleic acid sequence. (b) Schematic representation of exons with the numbers indicating the size in base pairs. (c) Hydrophobicity plot (blue line) for the Sho-like sequence indicates three distinct hydrophobic regions – the N-terminal and putative C-terminal signal sequences and a middle hydrophobic region. (d) Amino acid sequence for the 150-residue Sho-like protein. The N-terminal and putative C-terminal signal sequence are bolded and boxed, with the putative GPI attachment site indicated by an arrow. The RG-rich region corresponding to the NLS is shown in bold red, and the middle hydrophobic region of 19 residues (66-84) is shown bolded and underlined.

4.3.3 Sequence analysis of Sho-like

The search against the NCBI nr database using BLAST did not show significant similarity of the Sho-like sequence to known proteins. The hydrophobicity plot showed three well characterized regions: N-terminal, middle and C-terminal region (Figure 4-1c). The N-terminal hydrophobic region suggests the presence of a signal peptide for extracellular export of the protein; this is supported by the SignalP analysis (Figure 4-1d). The hydrophobic region at the C-terminal end could be a signal sequence for GPI-attachment; this is also supported by the big-Pi prediction of a possible GPI-modification site (Figure 4-1d; Figure 4-2b,c). Both these features are found in PrP family proteins from fish to mammals (Silverman et al. 2000; Premzl et al. 2004; Miesbauer et al. 2006; Strumbo et al. 2006) Other similarities to PrP family proteins, excepting Dpl, include the presence of an N-terminal basic region, in this case RG-rich as for the short proteins (Sho, PrP-like and Sho2), and the usual middle hydrophobic region unique to this family of proteins. As for the other short proteins, Sho-like lacks the Cys residues which form the disulphide bridge(s) in the folded C-terminal domain of PrP and Dpl, or putatively in the long-form fish proteins (stPrPs). Sho-like also lacks an N-glycosylation site, similar to PrP-like but in contrast to the conserved sites in Sho and Sho2 (Figure 4-2 b). In common with Sho, Sho-like is predicted to have a nuclear localization signal (NLS) in the N-terminal region (Figure 4-1d) which suggests a function within the nucleus.

The multiple sequence alignment with the known short PrP-related fish proteins shows some interesting features (Figure 4-2 b, c). Although the C-terminal region does not share sequence homology with these known PrP-related short proteins , the N-terminal region is most similar to fish PrP-like and the middle hydrophobic region is most similar to Sho2 (Figure 4-2 b, c; alignment with Sho not shown). Phylogenetic analysis indicates that Sho-like is more closely related to both PrP-like and Sho2 than to Sho (Figure 4-2 a), which is consistent with its presence at the *PRNP* genomic locus.

Analysis of the 3’ UTR sequence indicates a substantial amount of secondary-structure forming motifs, but no known functional UTR elements were found (data not shown).



(b)

```
                    10        20      ↓ 30        40        50        60
            ....|....|....|....|....|....|....|....|....|....|....|....|
FuSho2      MTGQQKLLSLWVWLLLMAALCPGVQHAYGKRGGVFKGRGKGDSTQAAPSQGRGLP-KPGL 59
TeSho2      MTGQQKLLSLWAWLLLMAALCPGVQHAYSKRGGAFKARGKSDGNQAAPSQSRGLP-KPGL 59
ZeSho2      MLGHQNLLIIWVWMLLLASLCP-CFHCKRGFGGRGMSKGA-APAKAPPSQSKGSSSRQGL 58
ChSholike   MRPRD--AWCWVAMLLLALFSN-TAGGRRGAGGRAGGRGGGGGGGGGLRGGFRSISRGG 57
Consensus   *  ::     *. :**:* :.          **   .:*      .   :.   .   *

                    70        80        90        100       110       120
            ....|....|....|....|....|....|....|....|....|....|....|....|
FuSho2      KWAG---AAAG-ILGGTGTGYGLGFLGRHKHGSGG-HHSHRAEQDQRPYHNQRHGFYNQS 113
TeSho2      KWAG---AAAG-ILGGTGTGYGLGFLGRHKHRSGS-HYSQRTEQDQRPYHNQRQGFYNQS 113
ZeSho2      KLAG--AAAAG-AIGGAAIGYGLGSLGRPRYG-----HNGYSDG----YH---QPVYNQS 102
ChSholike   NTAGRGSKMASAIAAGAATGYGMGLLGRPRPPRLG--HGPPAQR-----QPPPAGGFHAA 109
Consensus     **     *.   .*:. ***:* *** :       :.  ::      :      :::

                  130       140       150       160
           ...|....|....|....|....|....|....|....|...
FuSho2     LWKAIVNAAAPVSTSKAFLAFAHMFSFLVT---VWIRDI-   150
TeSho2     AWRAVANAAAPVFTSNVSLPLAHMVSVLFT---VWIRNI-   150
ZeSho2     HRGYYGSTGSSEHMGSLLITLG-TVAHVVI---MWA----   135
ChSholike  AWPDLGAKRGSPNQAPKGPCGGIVPTLLLANAICWVNHGM   150
Consensus           .. ↑ .      .   : :.     *
```

(c)

```
                    10        20      ↓ 30        40        50        60
            ....|....|....|....|....|....|....|....|....|....|....|....|
FuPrPlike   MSSQQKLPLTLTLCLLLLASLLPSSEARRGGGGFGRGGG-GRSWGRSQGQSRGWGGGGAG 59
TePrPlike   MSSLQKS-LTLTLCLLLL---LPSSEARRGGGGFGRGGG-GRGWGRSPSQSRGWGGGTGG 55
ZePrPlike   MHSKFKL-FSWRNCLLLLAVLLPVAQSRRGGG-FGRGGGRGGGWGGSSSGRAGWGAAGGH 58
ChSholike   MRPRDAW---CWVAMLLLALFSNTAGGRRGAG--GRAGGRGGGGGG--------GGGGLR 47
Consensus   * .     :    .:***  :   : .***.* :**.** * . *         *..

                    70        80        90        100       110       120
            ....|....|....|....|....|....|....|....|....|....|....|....|
FuPrPlike   QAYRPVQ-----------------SSNPSAGKVAGAAAAGAIGGAMIGSALS---RPGYG 99
TePrPlike   QAYRPAQ-----------------SSGPSAGKVAGAAAAGAIGGAMIGSALS---RPGYG 95
ZePrPlike   HRAPPVHTGHMGHIGHTGHTGHTGSSGHGVGKVAGAAAAGALGGMLVGHGLSSMGRPGYG 118
ChSholike   GGFRSIS-----------------RGGNTAGRGSKMASAIAAG-AATGYGMGLLGRPRPP 89
Consensus       .                    ..  .*: :  *:* * *    * .:.   **

                    130       140       150       160       170       180
            ....|....|....|....|....|....|....|....|....|....|....|....|
FuPrPlike   YGGYGGYGMGYGGYGGGYGGGYGRGYGYP-RGGYSNEPEGSGDMEYYYSGTSSSPAYSSI 158
TePrPlike   YG-HGGYGMGYPGYGGGYGG-YGRGYGYP-RGGYANEPEGSGDMEYYYSGASSSPAYSSI 152
ZePrPlike   YG-YGGYG----GHGYGYGHGYGHGHGHGGHGGHSGDHNET-DADYYLDGAASGHAYSCV 172
ChSholike   RLGHGPPAQRQPPPAGGFHAAAWPDLGAK-RGSPNQAPKGP------CGGIVPTLLLANA 142
Consensus      :*  .      . *:      . *   :*. ↑   : .       .*  .    :

                     190
            ....|....|....|....
FuPrPlike   III-TSLMALLMGHWVASI 176
TePrPlike   IFISTSLMSLLMGHWASIV 171
ZePrPlike   TVF-GLMMSFLIGHFLS-- 188
ChSholike   ICWVNHGM----------- 150
Consensus          * :::      :
```

Figure 4-2 Comparison of chicken Sho-like with Sho, Sho2 and PrP-like. (a) Phylogenetic tree showing the relationship of Sho-like with the other short PrP-related fish genes and with Sho. The tree was generated by the Maximum Likelihood method. (b) Multiple sequence alignment between fish Sho2 and chicken Sho-like. (c) Multiple sequence alignment between fish PrP-like and chicken Sho-like. Arrows in (b) and (c) show N-terminal cleavage and GPI-anchor attachment sites, the hydrophobic region is shown underlined, and the conserved N-glycosylation site for Sho2 is boxed. [Fu: *Fugu*, Te: *Tetraodon*, Ze: Zebrafish, Ch: Chicken]

4.3.4 Expression analysis of *PRNP*, *SPRN* and *Sho-like* transcripts in chicken

To investigate the distribution pattern of *Sho-like*, expression analysis was carried out using RT-PCR on a range of chicken tissues (testis, heart, kidney, lung, eye, gall bladder, muscle, spleen, pancreas, skin, gut, stomach, ovary and uterus). The *SPRN* expression pattern in chicken was also investigated, as there is no published data. The results are compared in Figure 4-3. Sho-like transcript was detected in most tissues at variable levels but is completely absent in muscle and is minimally present in pancreas and gut. Interestingly, brain did not show high levels of transcript; the highest levels were in internal organs, particularly gallbladder, spleen and ovary. This may explain my initial problem in amplifying the transcript using RACE on brain RNA (see Methods). On the other hand, *PRNP* and *SPRN* are highly expressed in chicken brain, consistent with reported results for mammals and other vertebrates (Cagampang et al. 1999; Premzl et al. 2003). *PRNP* is expressed strongly or relatively strongly in all chicken tissues analyzed, consistent with reported results in other vertebrates (Ford et al. 2002). The previous *in situ* hybridization studies (Harris et al. 1993) also indicated a wide spread distribution of chicken *PRNP* mRNA. *SPRN* transcript also showed widespread expression at various levels, consistent with our revised results for mammal, marsupial and frog [Vassilieva et al., unpublished results]; it is highly expressed in heart as well as brain, and is completely absent in testis and liver.

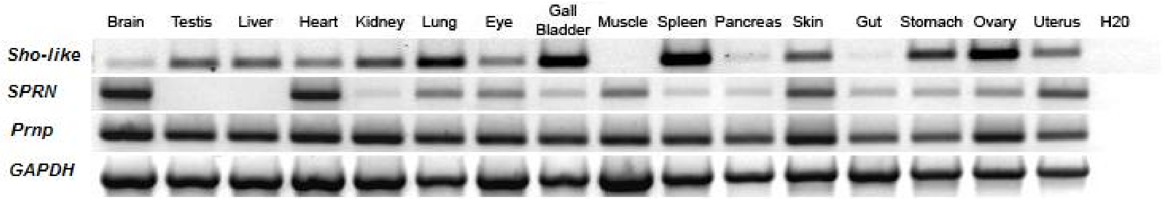


Figure 4-3 Comparison of tissue expression of transcripts of *Sho-like*, *SPRN*, *PRNP* and *GAPDH* in various chicken tissues tested with RT-PCR.

4.3.5 Comparison with the *PRNT* gene

The human *PRNP* gene locus contains a gene, *PRNT* between *PRND* and *RASSF2* on the opposite strand to *PRNP* (see Figure 4-4), which encodes a testis-specific protein (Makrinou et al. 2002). It has been suggested that *PRNT* is the result of a duplication of *PRND* (Makrinou et al. 2002). As the gene orientation and location of *PRNT* are similar to that of the chicken *Sho-like* gene, I performed a sequence comparison between the two loci. No significant sequence similarity was observed either at amino acid or nucleotide levels, excluding a possible correlation between *PRNT* and the *Sho-like* gene.

4.3.6 Comparison of chicken *PRNP* genomic loci with other vertebrates

Current knowledge of the organization of genes at the *PRNP* genomic loci in vertebrates is shown in Figure 4-4. Except for the *PRNT* gene which appears to be unique to eutherian mammals, other tetrapod lineages for which the locus has been defined (frog and marsupial) show only *PRNP* and *PRND*. The current Platypus assembly (assembly version 5) shows a part of the *PRNP* 3' UTR and the Dpl ORF. However, there are a number of ambiguous regions in this assembly and, hence, it is not possible to draw conclusions regarding the gene organization in the *PRNP* locus. Fish show two loci for the PrP-related genes, *stPrPs* (Oidtmann et al. 2003), *PrP-like* (Suzuki et al. 2002) and *SPRNB* (Premzl et al. 2004), likely resulting from a whole-genome duplication (Taylor et al. 2003). Two recent papers (Cotto et al. 2005; Rivera-Milla et al. 2006) have shown that these fish PrPs are differentially expressed, indicating that these duplicated genes have developed specialized functions. It is remarkable that the organization of the chicken *PRNP* locus correlates with the arrangement of genes at the two fish loci, and not that of the other tetrapods (Figure 4-4). This raises a very interesting question on the evolution of the locus in the intermediate-vertebrate avian and reptile branches.

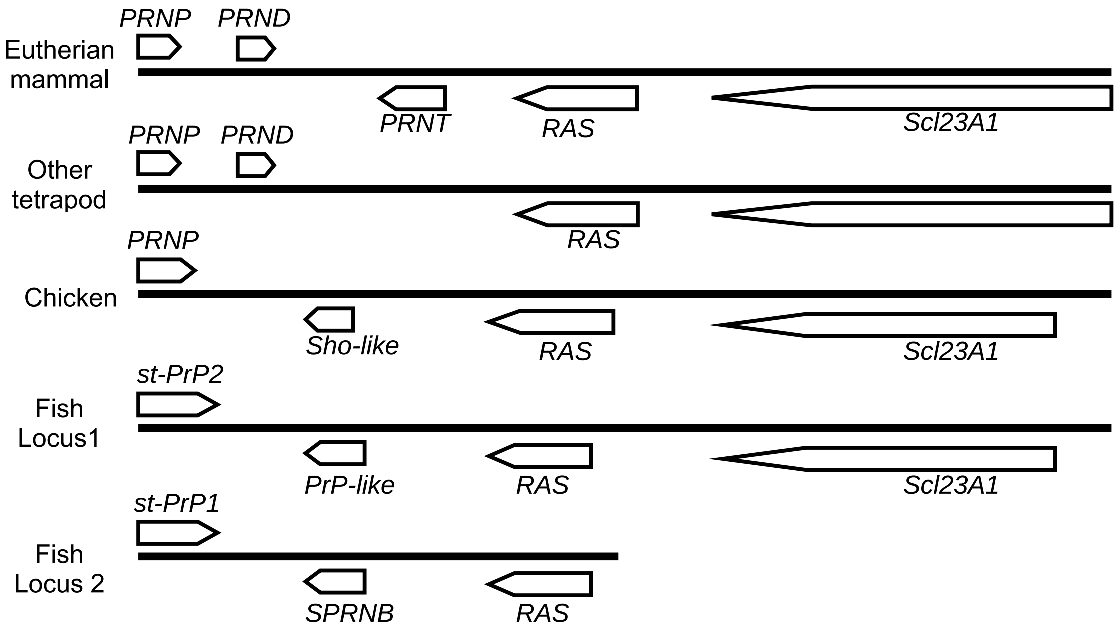


Figure 4-4 Summary of the gene organization in the *PRNP* genomic locus in vertebrate lineages elucidated to date. The order and orientation of the genes downstream from *PRNP* are shown. “Other tetrapod” includes frog (*X. tropicalis*), marsupial (*M. domestica*) and several non-primate Eutherian mammals. Figure not drawn to scale.

4.3.7 Comparison of chicken and fish *PrP*-related genes

The organization of chicken *PRNP* locus correlates with the arrangement in the fish *PrP*-related genes and *SPRNB* (Figure 4-4). To understand the origin of *Sho-like* gene, comparisons were made between the chicken *PRNP* and *Sho-like* gene region with that of the fish *PrP* genomic regions. The fish *PrP*-related gene regions included in the analysis were: *Fugu stPrP1* and *SPRNB*, *Fugu stPrP2* and *PrP-like*, zebrafish *PrP-like,* and zebrafish *stPrP1* and *SPRNB,* for which pairwise comparisons in all possible combinations were performed (Figure 4-5). The conservation patterns did not show any homology between the different fish *PrP* gene loci and the chicken *PRNP* gene locus. About 70% sequence identity was observed in the region corresponding to the ORF of orthologous fish *PrP* genes and about 50-60% among the paralogous *PrP* genes (Figure 4-5).

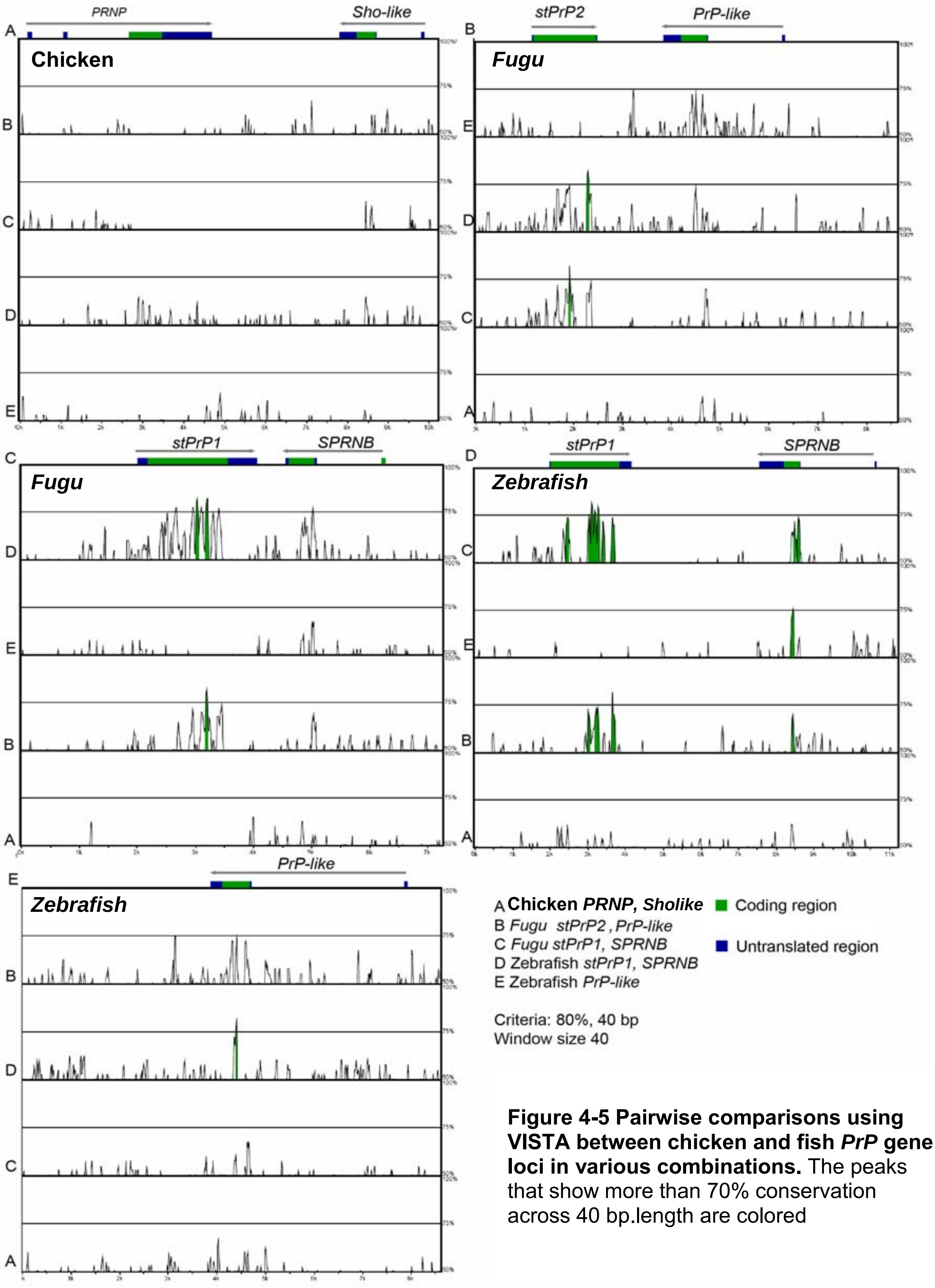


Figure 4-5 Pairwise comparisons using VISTA between chicken and fish *PrP* gene loci in various combinations. The peaks that show more than 70% conservation across 40 bp.length are colored

4.3.8 Identification of *PRNP* locus genes in emu and alligator

In the absence of whole-genome sequence data for other bird branches or any reptile, I started to address this evolutionary puzzle by homology searches for PrP family genes (PrP, Dpl, Sho-like) against a local BLAST database I constructed for emu and alligator trace sequences and EST databases. This did not produce any significant results. Using evolutionary PCR, I was able to identify PrP and Sho-like in emu. The partial Sho-like sequence obtained (residues 16-139) is identical to that of chicken, except at residues 171 (A/T) and 176 (T/A) which interestingly are in the hydrophobic region (Figure 4-1D); the emu Sho-like and PrP sequence data are given in Appendix 2. Several PCR methods were used to amplify PrP, Dpl and Sho-like genes from alligator genomic DNA, but without success. Failure may have been due, in part, to use of genomic DNA rather than RNA. Unfortunately, RNA from alligator tissues was not available.

4.3.9 Comparison of PrP sequence from birds and other tetrapods

A multiple sequence alignment of a selection of available PrP sequences from the tetrapod lineages (mammal, marsupial, monotreme, bird, reptile and frog) is shown in Figure 4-6. For the purposes of the current discussion, the selection has been highly weighted towards the bird sequences and the scant sequence data available for non-eutherian mammals. The sequences include the partial sequences for emu PrP and the platypus sequence extracted from my manual assembly of trace file data (discussed in Chapter 5); the latter contains many obvious errors and is included here solely to illustrate a point. As previously observed, PrPs from all lineages, including platypus, show a highly conserved hydrophobic region, but with variable lineage-specific repeats (Wopfner et al. 1999) which are lacking in amphibian PrP (Strumbo et al. 2001).

Although there is significant sequence variation of the post-hydrophobic (C-terminal) domain between amphibians and mammals, the PrP sequence around the conserved Cys residues is highly conserved and NMR structures for all the main branches (*Xenopus* (Calzolai et al. 2005), turtle (Calzolai et al. 2005), chicken

(Calzolai et al. 2005) and eutherian mammal (Riek et al. 1996) show a similarly folded structure. The striking difference in this region is a characteristic 9-10 residue insert in avian PrP (Figure 4-6), which structurally is located between the second and third helices and is flexibly disordered (Calzolai et al. 2005). The insert also contains a third conserved N-glycosylation site unique to birds. Homology search in databases for this sequence insert failed to identify other proteins with similar sequence.

Thus, the alignment reveals at least two features of the primary structure which differentiate chicken PrPs from both lower (frog) and higher vertebrates. These anomalies may possibly be related to the absence of Dpl in birds, perhaps through evolution of the C-terminal domain for functions which compensate for the functions of this structurally homologous domain of Dpl (Mo et al. 2001). A variant of this hypothesis is that absence of Dpl in birds has removed an evolutionary constraint on the C-terminal domain of avian PrPs, which has been suggested to interact with Dpl (Sakudo et al. 2005), allowing it to develop novel functions. A third hypothesis is that the insert is correlated with the presence of *Sho-like*, and perhaps suggests a direct interaction with Sho-like. Although the sequences of the fish stPrPs differ greatly from those of the tetrapod PrPs, it is interesting to note that much of their increased length (e.g. ~560 residues for zebrafish stPrP2 compared with ~250 residues for mammalian PrPs and ~275 residues for avian PrPs) comes from a very large insertion in the C-terminal domain predicted by homology modeling to be between the second and third helices (Rivera-Milla et al. 2003), i.e. the same as for the chicken PrP insert.

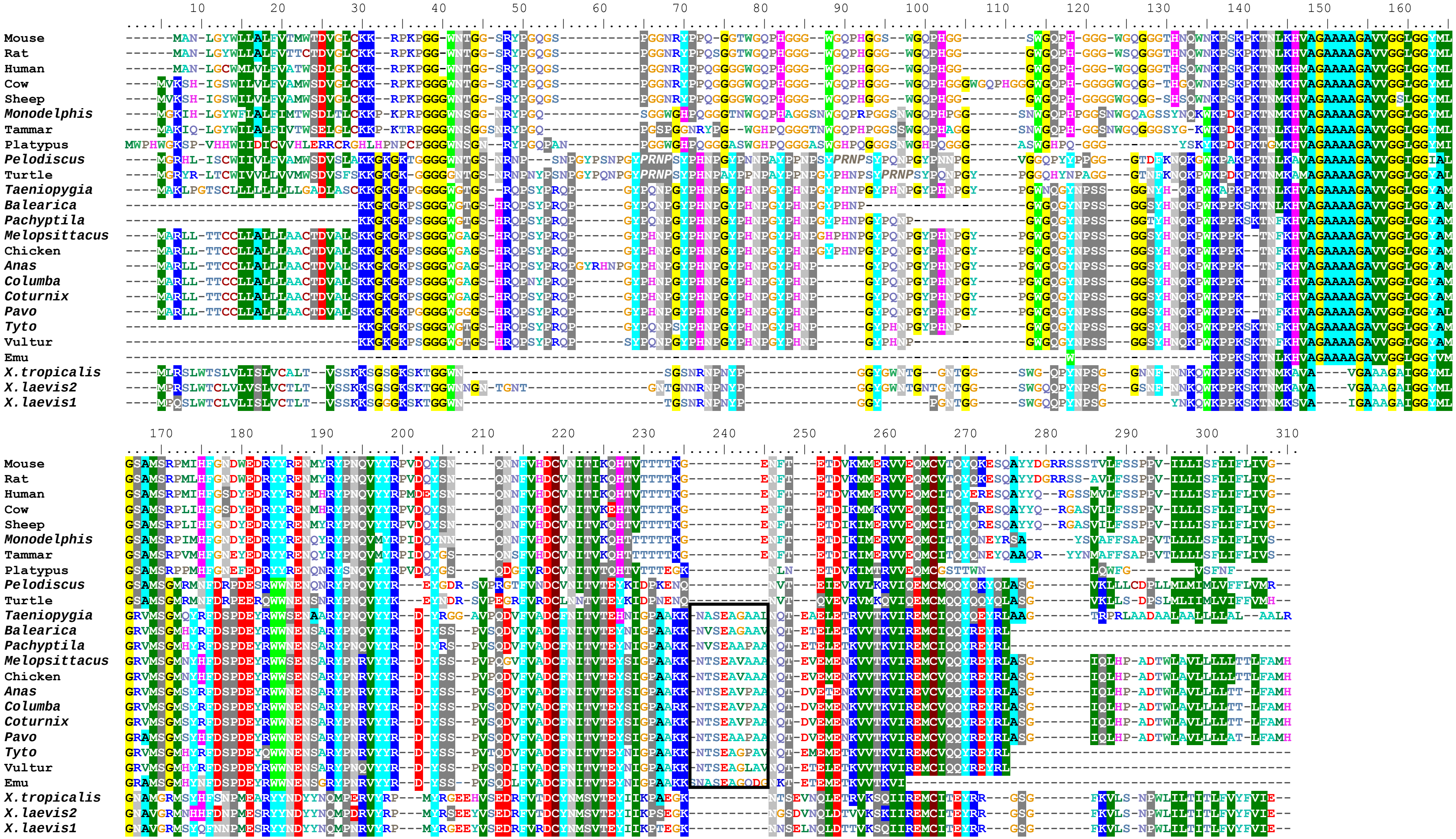


Figure 4-6 Alignment of PrP among tetrapods. Shading is based on 70% conservation. The unique region found in birds is shown in the rectangular box.

4.4 Absence of Dpl in chicken: Its implications and insights into the evolution of acrosome reaction

Dpl is associated with spermatogenesis in mice, and male *Prnd*-knockout mice were shown to be infertile with defects in morphology and number of sperm (Behrens et al. 2002; Paisley et al. 2004). The other function affected in *Prnd*-knockout mice was the inability of sperm to perform the acrosome reaction. This is a physiological process which occurs in the acrosome of the sperm as it approaches the outer layer of the egg in order to assist its fusion with the egg. The absence of *PRND* in the avian genome may indicate that this function was not acquired by Dpl in the lower vertebrates (e.g. frog), or it is compensated for by other evolutionary changes in birds. However, the acrosome reaction in lower vertebrates (Table 4-5) differs considerably from that in higher vertebrates. Bakst and Howart (1977) proposed that the cock sperm did not exhibit a typical mammalian or invertebrate type acrosome reaction. One factor underlying this may be anatomical differences between the chicken egg and that of higher vertebrates. Although the outermost perivitelline layer in the avian egg is morphologically homologous to mammalian zona pellucida (Bakst and Howarth 1977), they differ completely in their properties. It was also shown that *Xenopus* sperm can undergo the acrosome reaction (Ueda et al. 2002). Thus, though the acrosome reaction occurs in lower vertebrates, there exist considerable anatomical differences in the egg among the eutherian and subtherian groups, which suggests a different physiological mechanism. It has been proposed that in subtherian groups, including chicken (Bedford 1998), it is primarily by the lytic role of acrosomal enzymes. In eutheria, by contrast, the interaction of spermatozoa with the egg coat appears quite different, with some variation among the many genera (Bedford 1998). With the development of the thicker zona pellucida in eutherian mammals (Bedford 1998), Dpl may have acquired a new role in assisting the acrosome reaction. The role of Dpl in male fertility may, thus, be a modern function. The rapid evolution of the gene which is evident from sequence differences within the two closely related *Xenopus* species, *X. laevis* and *X. tropicalis* (Figure 3-3), and the low conservation between *Xenopus* and mammalian Dpls supports its potential to develop new roles.

Table 4-5 Comparison of the anatomy of egg and fertilization mechanisms in chicken, *Xenopus* and mammals

	Chicken	*Xenopus*	Mammal
Acrosome Reaction	Present	Present	Present
Fertilization	Internal	External	Internal
Doppel	Absent	Present	Present
Outer layer of egg	Perivitelline layer	Vitelline envelope	Zona pellucida

4.5 Conclusions

With the established relationship between PrP and Dpl in higher vertebrates, and now with the presence of both these genes in amphibian confirmed (Chapter 3), the absence of Dpl in chicken is intriguing. This may indicate that Dpl has not acquired a specialized role in lower vertebrates or that its functions are compensated by other evolutionary changes in birds. It is unlikely that the new *Sho-like* gene found in chicken downstream to *PRNP,* which has no sequence homology to Dpl, could be its functional replacement in avian lineages. The Sho-like protein has some regions of sequence similarities compared with fish PrP-like and Sho2 suggesting that this genomic locus has evolved from a common ancestor.

The absence of Dpl may have influenced the changes that are observed between avian PrPs compared with the other tetrapod lineages, the most interesting being the insertion of a 9-10 residue region in the C-terminal domain which is unique to avian species.

Finally, the chicken *PRNP* gene locus may be used as a basis for evolutionary studies of aves to establish its relationship to amphibian and reptiles once the genome sequence for reptilian *PRNP* loci is known. It is known that Dpl plays a role in the acrosome reaction in higher vertebrates. There are many anatomical differences between the egg of eutheria and subtherian species. The presence of weaker outer perivitelline layer in lower vertebrates compared with a zona pellucida in higher vertebrates, suggests the physiology of the acrosome reaction may be more complex in higher vertebrates, and that Dpl would have acquired the specialized role in testis only in mammalian lineages.

5 Findings in other species

5.1 Background

To investigate further the evolution of prion-protein family genes, examples of the most distant mammalian groups (Marsupials and Monotremes) were studied. The draft genomes of interest were those of grey short-tailed opossum (*Monodelphis domestica*) and Platypus (*Ornithorhynchus anatinus*). Sequencing for *M. domestica* and platypus are being undertaken by the Broad Institute of MIT and Harvard, and by Washington University Genome Sequencing Center, respectively. Platypus, a prototherian has both reptilian and mammalian characteristics. The mammalian features include fur bearing and mammary glands to feed the young with milk. The reptilian features include egg laying and the possession of venom. The unique evolutionary relationship of monotremes and marsupial mammals to eutherian mammals attracts particular interest. Platypus is, however, classified as a true mammal because of the milk producing ability (Griffiths 1978). Evolutionary biologists agree that monotremes diverged before the Therian (eutherian and marsupial) mammals (Figure 5-1). The platypus genome offers a unique resource for comparative genomics to identify genes and regulatory sequences.

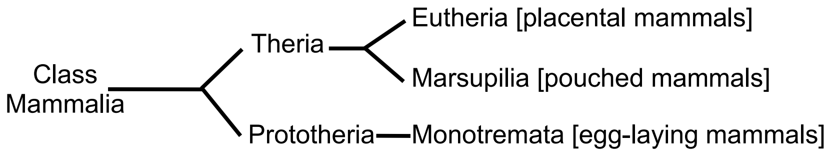

Class Mammalia
Theria
Eutheria [placental mammals]
Marsupilia [pouched mammals]
Prototheria
Monotremata [egg-laying mammals]

Figure 5-1 Evolutionary history of mammalian class.

The marsupial (metatherian) is the closest outgroup to placental mammals (eutherian) (Figure 5-1). Its genome provides an opportunity to understand the organization and evolution of the mammalian genome.

At the time of this study, the *M. domestica* draft genome assembly was released; I used this for my analysis. However, for platypus, the only available sequence information was in the form of trace sequences. To probe for the origin of the PrP family I also analyzed the *Ciona intestinalis* genome (sequenced by Joint Genome

Institute); *C. intestinalis* is a primitive and smallest chordate. To understand the PrP family genes at an intermediate stage between fish and amphibian, I also studies salamanders which belong to order caudate of class amphibia. Salamanders are distinguished from other amphibia by the presence of a tail in adults. *Ambystoma mexicanum* (Axolotl) is a model organism for this group and is used for evolutionary and regeneration studies. Finally, human sequence databases (EST and genome) were also analyzed to look for any variant forms, or new members of PrP family genes.

5.2 *Monodelphis domestica*

The first reported marsupial PrP sequence was that of *Tricosurus vulpecular* (brush-tailed opossum) (Windl et al. 1995). Premzl et al. (2005) later reported tammar wallaby PrP. The other PrP family members (Dpl and Sho) were not reported in marsupial species. My interest was to identify these genes for comparative sequence analysis.

5.2.1 Homology search

A local standalone BLAST database was created which was managed by Perl scripts (createBlastdb.pl). Initial searches were made on the assembled version 0.5, which was the latest at the time of study. Homology searches were performed using TBLASTN for the PrP family genes with the following amino acid query sequence: PrP- *T. vulpecular*; Dpl- Human; Sho- Human.

Partial sequences were found for all the genes of interested (Figure 5-2). This sequence information was used to design primers for experimental work. The protein sequences were analyzed for known sequence features.

```
>MdPrP
MGKIHLGYWFLALFIMTWSDLTLCKKPKPRPGGGWNSGGNRYPGQSGGWGHPQGGGTNWGQPHAGGSNWGQPRPGGSNWGQPHPGGSNWGQP
HPGGSNWGQAGSSYNQKWKPDKPKTNMKHVAGAAAAGAVVGGLGGYMLGSAMSRPIMHFGNDYEDRYYRENQYRYPNQVMYRPIDQYNNQNN
FVHDCVNITVKQHTTTTTTKGENFTETDIKIMERVVEQMCITQYQNEYRSAYSVAFFSAPPVTLLLLSFLIFLIVS

>MdDpl
MRRHLGICWIAIFFALLFSDLSLVKAKTTRQRNKSNRKGLQTNRTNPTTVQPSEKLQGTFIRNGRKLVIDFGEEGNSYYATHYSLFPDEIHY
AGCAESNVTKEVFISNCVNATRVINKLEPLEEQNISDIYSRILEQLIKELCALNYCEFRTGKGTGLSALFRPICYGLPGDSDLLDSEIHKHR
A

>MdSho
MNWAAVTCWTLLLLAAFFCENVTSKGGRGGARGAARGRSRSSSSSSRMRMKSAPRYSSSGSAFRVAAAASAGAAAGAAAGAVAGAAGRRMSG
EVGTSVNLERDLYYSNQTGEGIYSYRWTSGTDRGGVEPNLSLCLTLGFFQLFHP

Signal sequence
Cysteine involved in disulphide bridge formation
N-glycosylation site
GPI anchor site
```

Figure 5-2. Sequence information for *M. domestica* PrP, Dpl and Sho (obtained from genome sequence database).

5.2.2 Defining exon-intron boundaries in *M. domestica*

The full length cDNA sequences were obtained from 3’ and 5’ RACE experiments which were carried out by Tatiana Vassilieva in my group. The *M. domestica* Version 2 assembly was available at this stage of the study, and I used it for gene annotation. I mapped the cDNA sequence onto the genomic sequence to obtain the intron/exon boundaries (Figure 5-3). Two variant transcripts for *PRNP* and four different variants for *PRND* were identified.

As in *Xenopus* species where the predominant variant for *PRND* was a chimeric transcript, a similar observation was made in *M. domestica*. Different intergenic exons (2D, 3D, and 4D) were identified (Figure 5-3). The 3’ UTR also showed a long and a short variant (Figure 5-3).

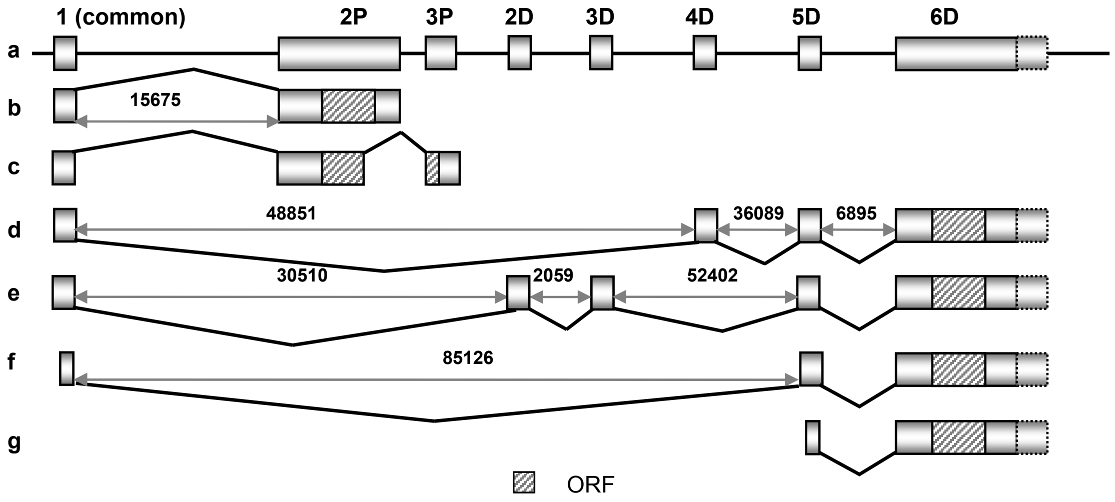


Figure 5-3 Schematic representation of the exons identified by mapping the experimental cDNA sequences onto the *M.domestica* genome assembly sequence. (a) Exons of *PRNP* and *PRND*. The number above the exon represents the exon number. The letter "P" is used to represent *PRNP* and "D" for *PRND*. (b) Non-chimeric *PRNP* variant. (c) Non-chimeric *PRNP* variant with intron in the coding region. (d), (e), (f) Chimeric transcripts with different intergenic exons. (g) Non-chimeric *PRND*. The 3' UTR of *PRND* showed two different sizes. The longer variant is represented with dotted lines. The length of the introns is shown in the transcripts in bp.

5.2.3 Alternative splicing leading to two different C-terminal PrP sequences

The structures of the two *M. domestica* PrP transcripts show different C-terminal sequence corresponding to the GPI-signal sequence. One variant is formed by alternative splicing of the 3' end of the ORF (Figure 5-3 c); this is the only case of an intron in the ORF found in a PrP gene family. This results in two different coding regions, ORF1 which does not have an intron (regular type), and ORF2 which has the intron (unique type) (Figure 5-4). Although the ORF2 variant was reported by Premzl et al. (2005) (NCBI accession number: AY659989 and BK005535) the presence of the intron was not recognized due to lack of genomic sequence information at the time of their study. The significance of the variation in the C-terminal signal sequence is unclear; it may not produce a significant functional difference as there is no change in the predicted GPI-modification site. The splicing may arise from the presence of additional splice donor and acceptor sites, resulting in some leaky splicing (see Appendix 3). Further analyses of these variants were made as part of a different project carried out by Tatiana Vassilieva and are not discussed further.

```
                               10        20        30        40        50        60
                      ....|....|....|....|....|....|....|....|....|....|....|....|
MdPrP_ORF1            MGKIHLGYWFLALFIMTWSDLTLCKKPKPRPGGGWNSGGNRYPGQSGGWGHPQGGGTNWG
MdPrP_ORF2            MGKIHLGYWFLALFIMTWSDLTLCKKPKPRPGGGWNSGGNRYPGQSGGWGHPQGGGTNWG
                      ************************************************************

                               70        80        90       100       110       120
                      ....|....|....|....|....|....|....|....|....|....|....|....|
MdPrP_ORF1            QPHAGGSNWGQPRPGGSNWGQPHPGGSNWGQPHPGGSNWGQAGSSYNQKWKPDKPKTNMK
MdPrP_ORF2            QPHAGGSNWGQPRPGGSNWGQPHPGGSNWGQPHPGGSNWGQAGSSYNQKWKPDKPKTNMK
                      ************************************************************

                              130       140       150       160       170       180
                      ....|....|....|....|....|....|....|....|....|....|....|....|
MdPrP_ORF1            HVAGAAAAGAVVGGLGGYMLGSAMSRPIMHFGNDYEDRYYRENQYRYPNQVMYRPIDQYN
MdPrP_ORF2            HVAGAAAAGAVVGGLGGYMLGSAMSRPIMHFGNDYEDRYYRENQYRYPNQVMYRPIDQYN
                      ************************************************************

                              190       200       210       220       230      ↓240
                      ....|....|....|....|....|....|....|....|....|....|....|....|
MdPrP_ORF1            NQNNFVHDCVNITVKQHTTTTTTKGENFTETDIKIMERVVEQMCITQYQNEYRSAYSVAF
MdPrP_ORF2            NQNNFVHDCVNITVKQHTTTTTTKGENFTETDIKIMERVVEQMCITQYQNEYRSAYSVAF
                      ************************************************************
                                                                            ↑

                              250       260
                      ....|....|....|....|....|....|....
MdPrP_ORF1            FSAPPVTLLLLSFLIFLIVS---------
MdPrP_ORF2            FSAPPVTLLLLSFLIFLIIPDAHSVEAIS
                      ******************:.
```

Figure 5-4 Comparison of the two *M. domestica* PrP ORF variants, ORF1 and ORF2. Note the only difference is in the C-terminal signal sequence shown in the box. GPI attachment site is indicated by arrow.

5.3 Platypus

5.3.1 Homology search

A BLAST-searchable platypus trace sequence database was created from the downloaded sequences. This was screened for the PrP family genes using the known sequence information: Dpl- *M. domestica*; PrP- Human; Sho- Human.

Partial to full length sequences of all the PrP family members were found (Figure 5-5).

```
>PrP
MWPHWGKSPVHHWIIDICVVHLERRCRGHLHPNPCPGGGWNSGNRYPGQPANPGGWGHPQGGGASWGHPQGGGASWGHPQGGGSNWGHPQGG
GASWGHPQGGGYSKYKPDKPKTGMKHVAGAAAAGAVVGGLGGYMIGSAMSRPPMHFGNEFEDRYYRENQNRYSNQVYYRPVDQYGSQDGFVR
DCVNITVTQHTVTTTEGKNLNETDVKIMTRVVEQMCGSTTWNLQWFGVSFNF

>Dpl
MMTVRRRRRSGGARWLLVFLVLLSGDLSSLQARGPRPRNKAGRKPPPVQRRALTLRAPRPPAGARGTFIRRGGRLSVDFGPEGNGYYQANYP
LLPDAIVYPDCPTANGTREAFFGDCVNATHEANRGELTAGGNASDVHVRVLLRLVEELCALRDCGPALPTGPAPRPGPPGPPAALALLTLVL
LGAQ

>Sho
MNWVAVACWTLLLLTAFLCDSVTCKGGRGGARGAARGAARGATRVRLKSVPRYSSSGSGLRVEAS

Signal sequence
Cysteine involved in disulphide bridge formation
N-glycosylation site
GPI anchor site
```

Figure 5-5 Translated ORFs obtained by homology search against the platypus trace sequence database. The Sho sequence shows a premature stop codon which may be due to poor sequence quality rather than a psuedogene; this can be confirmed only by a later release of better quality assembled sequence or by direct sequencing.

5.3.2 Sequence analysis

PrP and Dpl sequences were analyzed for known sequence elements (signal sequence, disulphide bridge and N-glycosylation). The N-terminal sequence of PrP and Dpl showed very significant variation from other known sequences. This may be erroneous and related to the low sequence coverage of the unassembled trace sequences. Interestingly, the *PRND* ORF is in a GC rich region in contrast to other species (see Appendix 4) (Table 5-1).

Because of the lack of good quality assembled sequence information required for comparative sequence analysis, no further analysis was performed using these sequences. Experimental work to characterize the cDNAs and the tissue expression were studied as a part of other project (work of Tatiana Vassilieva) and hence not discussed further.

Table 5-1 Estimating the percentage of C and G bases in *PRND* ORF among different species

	Human	Mouse	Cow	Sheep	*Monodelphis*	Platypus	*Xenopus*
C%	30.1	27.7	24.6	24.8	20.0	39.2	21.8
G%	29.9	31.8	27.6	27.2	22.6	35.3	24.0

5.4 Miscellaneous searches

5.4.1 Ciona intestinalis

The evolutionary model proposed by Premzl et al. (2004) suggested a primitive Sho-like gene to be the ancestor of all the PrP family genes. Based on this hypothesis, a homology search was performed to look for the origin of the PrP family genes in the *Ciona intestinalis* genome. Homology-based searches using BLAST revealed no regions of sequence similarity to those of PrP family genes.

5.4.2 Axolotl

Databases containing Axolotl EST sequences are reported in the literature (Habermann et al. 2004; Putta et al. 2004). Searches for Dpl, PrP and Sho using online BLAST http://salamander.uky.edu/ESTdb/blast.php (TBLASTN and BLASTN) did not yield positive results. To confirm this, a standalone BLAST database was created (using createBlastdb.pl) and was searched using various combinations for PrP, Dpl and Sho from several known sequences. However, no significant homologous sequences were obtained.

5.4.3 Human

Analysis of EST sequences: The possibility of spliced *PRNP* ORF variants as seen in MdORF2 was investigated in the human EST database. Several EST sequences corresponding to human PrP sequences were retrieved from the NCBI EST database (see Appendix 5). A multiple sequence alignment was performed to look for any variations; no variants similar to MdORF2 were discovered. However, this search lead to discovery of two interesting *PRNP* variants, one with fewer N-terminal repeats (GI: 15582753, 15581955) and the other with a frame shift mutation (GI: 52091525, 15438946) (Figure 5-6). The presence of fewer N-terminal repeats compared with wild-type PrP represents a rare polymorphism. Splice-site prediction using SpliceView software (Zhang et al. 2002) (http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html) was performed after introducing T255A (number corresponds to the nucleotide position in the wild-type human *PRNP* ORF). The introduction of this polymorphism created a new splice acceptor site at that position with the splice donor site located at position 157 (Figure 5-6). This splicing created the exact variant with fewer N-terminal repeats as that found in the EST sequences (GI: 15582753, 15581955). There is no published information on these PrP variants and hence the implication of such a variant is unknown. However, insertions in the octapeptide repeat region have been associated with prion disease (Ironside 1998) and deletions are known in other mammals (eg. cattle) (Walawski and Czarnik 2003). The variants with frame shift mutations caused by insertions/deletions in the two ESTs (GI: 52091525, 15438946) may be a result of sequencing errors.

a

```
HumPrP          MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP
15581955        MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGG-----
15582753        MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGG-----
                *******************************************************

HumPrP          HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
15581955        ---------------------------GWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
15582753        ---------------------------GWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
                                           *********************************

HumPrP          VVGGLGGYMLGSAMSRPMIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
15581955        VVGGLGGYVLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
15582753        VVGGLGGYVLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
                ********:********:******************************************

HumPrP          NITIKQRTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPV
15581955        NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQ-----------------------
15582753        NITIKQHTVTTXTKGENFTETDVKMMERVVEQMCIT------------------------
                ******:**** ************************

HumPrP          ILLISFLIFLIVG
15581955        -------------
15582753        -------------
```

b

```
ATGGCGAACCTTGGCTGCTGGATGCTGGTTCTCTTTGTGGCCACATGGAGTGACCTGGGCCTCTGCAAGAAGC
GCCCGAAGCCTGGAGGATGGAACACTGGGGGCAGCCGATACCCGGGGCAGGGCAGCCCTGGAGGCAACCGCTA
CCCACCTCAGGGCGGTGGTGGCTGGGGGCAGCCTCATGGTGGTGGCTGGGGGCAGCCTCATGGTGGT
                                                T255A
GGCTGGGGGCAGCCCCATGGTGGTGGCTGGGGACAGCCTCATGGTGGTGGCTGGGGTCAAGGAGGTGGCACCC
ACAGTCAGTGGAACAAGCCGAGTAAGCCAAAAACCAACATGAAGCACATGGCTGGTGCTGCAGCAGCTGGGGC
AGTGGTGGGGGGCCTTGGCGGCTACATGCTGGGAAGTGCCATGAGCAGGCCCATCATACATTTCGGCAGTGAC
TATGAGGACCGTTACTATCGTGAAAACATGCACCGTTACCCCAACCAAGTGTACTACAGGCCCATGGATGAGT
ACAGCAACCAGAACAACTTTGTGCACGACTGCGTCAATATCACAATCAAGCAGCACACGGTCACCACAACCAC
CAAGGGGGAGAACTTCACCGAGACCGACGTTAAGATGATGGAGCGCGTGGTTGAGCAGATGTGTATCACCC
```

Figure 5-6 Human PrP with fewer N-terminal repeats. (a) Comparison of wild-type human PrP with human PrP showing fewer N-terminal repeats. (b) Genomic region corresponding to the ORF of human PrP. One possible explanation for the presence of fewer repeats may be a rare polymorphism of T255A leading to splicing of the ORF (the splice donor and acceptor sites are shown in dotted box.

Whole genome analysis: A homology search using BLAST to look for any other undiscovered PrP-like genes was performed on the human genome sequence using PrP, Dpl and Sho sequences from different lineages.

After trying different query sequences in various combinations, only one interesting ORF (huNewORF) (Figure 5-7A) was obtained when chicken Sho was used as query (Figure 5-7B). However, it shares low sequence homology with human Sho (Figure 5-7C).

This new sequence lacks the N-terminal and C-terminal signal sequences. When a 50 KB region around this ORF was analyzed with GenScan, it predicted a gene in the opposite direction (Figure 5-8). This corresponds to Zinc finger protein (ZNF) (NP_001028895) as in the NCBI protein database. This gene is on the complementary strand to that of the new ORF which is overlapping exon no. 6 of *ZNF* gene (Figure 5-8). EST database searches with the huNewORF sequence indicated that this transcript is expressed. However, this observation is due to the exon 6 of *ZNF* gene. Interestingly, within the *ZNF* gene environment, there is another pseudogene, CDC28 (NC_000008), spanning the intron between *ZNF* exon 2 and 3 (Figure 5-8).

a

```
>HuNewORF
MQKTAWLTTNLSFPGNGLVEEQVAGLRLVDAVVWLRGAAEGLEAVSRERRRRRGRVGLVAGGAPAPVAAAAGA
GRHAAALLPGSAEALCMHTNTVSARTPACPGMVQWGNGSLGSTPVKPPQPSQSSATAGYVSFCSIYTPPPKSH
IVVYKHLIILTS
```

b

```
                      10        20        30        40        50        60
              ....|....|....|....|....|....|....|....|....|....|....|....|
HuNewORF      MQKTAWLTTNLSFPGNGLVEEQVAGLRLVDAVVWLRGAAEGLEAVSRERRRRRGRVGLVA
ChickSho      MRQRVACCWVLLLLAATFCQPAAA----KGGRGGSRGAARGMARG-AARSRHRGLP--RY
              *:: .     * : .  : :  .*     ..    ****.*:      * *:**

                      70        80        90       100       110       120
              ....|....|....|....|....|....|....|....|....|....|....|....|
HuNewORF      GGAPAPVAAAAGAGRHAAALLPGSAEALCMHTNTVSARTPACPGMVQWGNGSLGSTPVKP
ChickSho      GGALRVAAAAAAAGAAAGAALHQARAETEYHEGNGTAWTSVAPGWVEWGWAMPWLCPLAA
              ***   .****.**  *.* *  :      *  .. :* *...** *:** .     *: .

                     130       140       150
              ....|....|....|....|....|....|....|...
HuNewORF      PQPSQSSATAGYVSFCSIYTPPPKSHIVVYKHLIILTS
ChickSho      ILHHWHPP--GPLRSSAIQQRGNKAGQ-----------
                    ..  * :  .:*     *:
```

c

```
                      10        20        30        40        50        60
              ....|....|....|....|....|....|....|....|....|....|....|....|
HuNewORF      MQKTAWLTTNLSFPGNGLVEEQVAGLRLVDAVVWLRGAAEG-LEAVSRERRRRRGRVGLV
HuSho         MNWAPATCWALLLAAAFLCDSGAAKGGRGGARGSARGGVRGGARGASRVRVRPAQRY---
              *: :.     * :..  * :. .*     .*    **...*  ...** * *   *

                      70        80        90       100       110       120
              ....|....|....|....|....|....|....|....|....|....|....|....|
HuNewORF      AGGAPAP---VAAAAGAGRHAAALLPGSAEALCMHTNTVSARTPACPGMVQWGNGSLGST
HuSho         --GAPGSSLRVAAA-GAAAGAAA---GAAAGLAAGSGWRRAAGPGERGLEDEEDGVPGGN
                ***..   **** **  .***   *:* .*.  :.   *  *.  *: :  :*  *..

                     130       140       150       160
              ....|....|....|....|....|....|....|....|....|
HuNewORF      PVKPPQPSQSSATAGYVSFCSIYTPPPKSHIVVYKHLIILTS---
HuSho         GTGPGIYSYRAWTSG-----AGPTRGPRLCLVLGGALGALGLLRP
               . *   *  : *:*     :  *  *:  :*:   *  *
```

Figure 5-7 Analysis of new human ORF with sequence similarities to Sho. (a) The human sequence found by homology search using chicken Sho sequence with the possible N-glycosylation sites highlighted. (b) Alignment with Chicken Sho and (c) Human Sho

A search was performed using the huNewORF sequence in the chimp, rat and mouse genomes. Although a similar region was found, the sequence in rat and mouse had premature stop codons. In summary, the observed sequence conservation is more likely due to the *ZNF* exon rather than indicating a new gene.

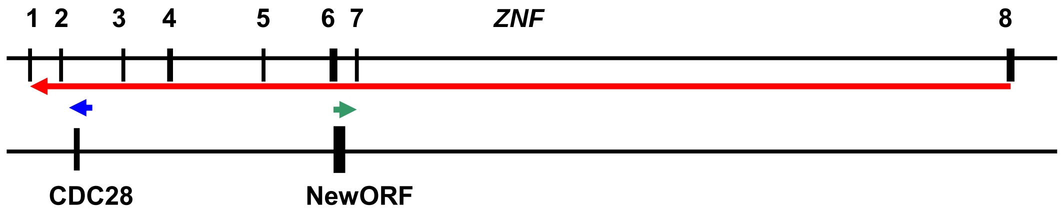


Figure 5-8 Genomic environment of the new (Sho-like) ORF in human. This ORF overlaps exon 6 of *ZNF* in the opposite direction. Also note the pseudogene CDC28 which spans the intron of ZNF and is oriented in the same direction.

5.5 Summary

M. domestica and platypus genomes were search for PrP family genes. All three genes were discovered in *M. domestica* and experimental characterization revealed chimeric *PRND* transcripts. One of the *PRNP* transcript variants had an intron in the 3' end of the ORF resulting in an altered C-terminal signal sequence. However, this did not alter the predicted cleavage site for GPI attachment. This splicing may be a result of a weak splice signal.

Searching the platypus trace sequence database for PrP family genes revealed full/partial length sequences for PrP, Dpl and Sho.

Homology-based searches for PrP family members in *C. Intestinalis* for which genome data are available, and axolotl species (*Ambystoma mexicanum*) did not reveal any homologous sequence. Human PrP EST database was searched to look for any undiscovered variants. This search revealed ESTs which had fewer repeats probably as a result of a SNP signaling as a splice acceptor site. Some other EST sequences showed frame-shift mutations which may be a result of sequencing error. These variants were not reported in the literature before. The human genome was also searched for undiscovered PrP family genes. This search revealed an ORF sequence homologous to chicken Sho. However, this region is not seen even in closely related species (Chimp, rat and mouse) suggesting that this is not a functional ORF.

6 Comparative sequence and structure analysis of prion protein and doppel

6.1 Background

Comparative sequence analyses offer powerful approaches to infer conserved amino acid motifs that are likely of general importance. Conserved motifs can then be located on the three-dimensional structure which may give additional information on binding sites, DNA binding domains and protein-protein contact regions.

The rate of evolution varies within and among lineages in that different proteins evolve at different rates. The rate of evolution varies among amino acid sites with some regions being highly conserved and other regions showing large variations. Hence, the evolutionary rate of change at an amino acid site is indicative of how conserved this site is and, in turn, allows evaluation of its importance in maintaining the structure or function of the protein.

PRNP and *PRND* are gene duplicates which have been retained over time. Apparently, this gene duplication occurred between amphibians and fish before the radiation of tetrapods (Figure 7-2). Genes that have been duplicated are normally lost over long evolutionary timescales because of lower selective pressure. More rarely, duplicate genes may be retained. In some such cases, the functions and expression patterns of the gene pairs may diverge substantially giving rise to novel functions or specialization in the organism. Force et al. (1999) have suggested that gene duplication may allow sub-functionalization to take place if a gene performing more that one function is duplicated. Lynch and Conory (2000) have suggested that following duplication, a time-dependent manner of selective pressure comes into play with the gene duplicate evolving in a nearly neutral way immediately after duplication but becoming more constrained as it becomes more divergent from its copy. The evolution of novel function may be related to changes in the coding and/or regulatory regions after gene duplication.

My main purpose in this chapter is to describe the variation in the rate of molecular evolution of *PRNP* and *PRND*. My other aim is to derive functional information from comparative sequence analysis.

6.1.1 Methods

Sequence alignment for predicting critical amino acids: The sequences included for identification of conserved residues are human, mouse, rat, sheep, cow, *M. domestica*, tammar wallaby, platypus, chicken, turtle, *X. laevis and X. tropicalis* (Table 6-1).

Table 6-1 Identifiers of known sequences used in the analysis.

Species	Source
PrP	
Human	ref|NP_001009093.1|
Mouse	ref|NP_035300.1|
Rat	ref|NP_036763.1|
Sheep	ref|NP_001009481.1|
Cow	ref|NP_851358.1|
Tammar wallaby	gb|AAT68001.1|
Possum	gb|AAA61833.1|
Turtle	emb|CAB81568.1|
Chicken	gb|AAC28970.1|
X. laevis	ref|NP_001082180.1|
Dpl	
Human	ref|NP_036541.2|
Mouse	ref|NP_075530.1|
Rat	ref|NP_001095901.1|
Sheep	ref|NP_001009261.1|
Cow	ref|NP_776583.1|
M. domestica	ref|XP_001381721.1|

The N-terminal signal sequence, C-terminal signal sequence, N-terminal repeats in PrP and the C-terminal domain regions in PrP and Dpl were aligned independently. These individual alignments were assembled to obtain the final alignment.

Alignment of the C-terminal domain: Pairwise structural alignments were generated using DaliLite (Holm and Park 2000). Mouse PrP (1AG2) and mouse Dpl (1I17) were used as templates for all the alignments, enabling generation of a multiple structure-based sequence alignment using mouse sequence as a reference.

Alignment of the signal sequence: The predicted signal sequences among different lineages in PrP and Dpl were aligned using ClustalW.

Calculation of rate of evolution: The rate of evolution of amino acid sequence was calculated by the Rate4Site algorithm (Pupko et al. 2002). Rate4Site makes use of topology and branch lengths of the phylogenetic trees constructed from multiple sequence alignments (MSA) of proteins and estimates conservation rates of amino acids based on the empirical Bayesian rule. A site-specific rate, *r*, indicates how fast a particular site evolves relative to the average evolutionary rate across all sites in the MSA and, hence, is unit less. A rate of 2.0 indicates that a site evolves two times faster than the average.

The species included for calculating the rate of evolution are: human, mouse, rat, cow, sheep, *M. domestica*, and *X. tropicalis*. As there is no information available for the existence of reptilian *PRND* and as *PRND* is absent in chicken, these two vertebrate branches were eliminated from the analysis. Platypus sequence was not included in the analysis due to its current poor sequence quality. The average values of different structural components were calculated separately to compare the rate of evolution within different segments of the protein sequence. The graphs were plotted using Microsoft Excel.

Analysis of distribution of conserved amino acids in 3D space: ConSurf is a web server (http://consurf.tau.ac.il/) for mapping the level of evolutionary conservation of each of the amino acid positions of a protein (based on multiple sequence alignment) onto its 3D structure (Glaser et al. 2003). The degree of residue conservation is translated into a coloring scale that is projected on to a known structure. Comparisons for PrP and Dpl were made as follows (reptile and avian sequence are not included for Dpl)

1. Within eutherian mammals (human, mouse, rat, cow, sheep, dog).
2. Eutherian mammals and marsupial (tammar wallaby and *M. domestica*).

3. Eutherian mammals, marsupial, avian (chicken), and reptile (turtle).
4. Eutherian mammals, marsupial, avian, reptile and amphibian (*X. tropicalis*).

All illustrations of protein structure were generated using PyMOL (DeLano 2002).

Polydot: A dotplot is a graphical representation of the regions of similarity between two sequences. Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines. It is therefore possible to see at a glance where there are local regions of similarity. Polydot (Rice et al. 2000) compares all sequences in a set of sequences and draws a dotplot for each pair of sequences by marking where words (tuples) of a specified length have an exact match in both sequences. The Dotplot for PrP was based on a window size of 4 and that for Dpl on a window size of 6.

6.2 Results and discussion

6.2.1 Rate of evolution of amino acid sequence

The rate of evolution of amino acid sequence was measured among human, mouse, cow, sheep, *M. domestica* and *X. tropicalis* based on the MSA shown in the Figures 6-5 and 6-6. The N-terminal signal sequence in PrP is evolving at a slightly faster rate compared with the other regions of the protein (Figure 6-1). The middle hydrophobic region is likely to be under selective pressure and is evolving at the slowest rate compared with the other regions. The repeat region and the C-terminal domain are evolving at a similar rate to each other. In Dpl, except for the C-terminal signal sequence which is evolving at a higher rate, other regions shows similar rates of evolution (Figure 6-1) (see Appendix 6 for rate of evolution of individual sites).

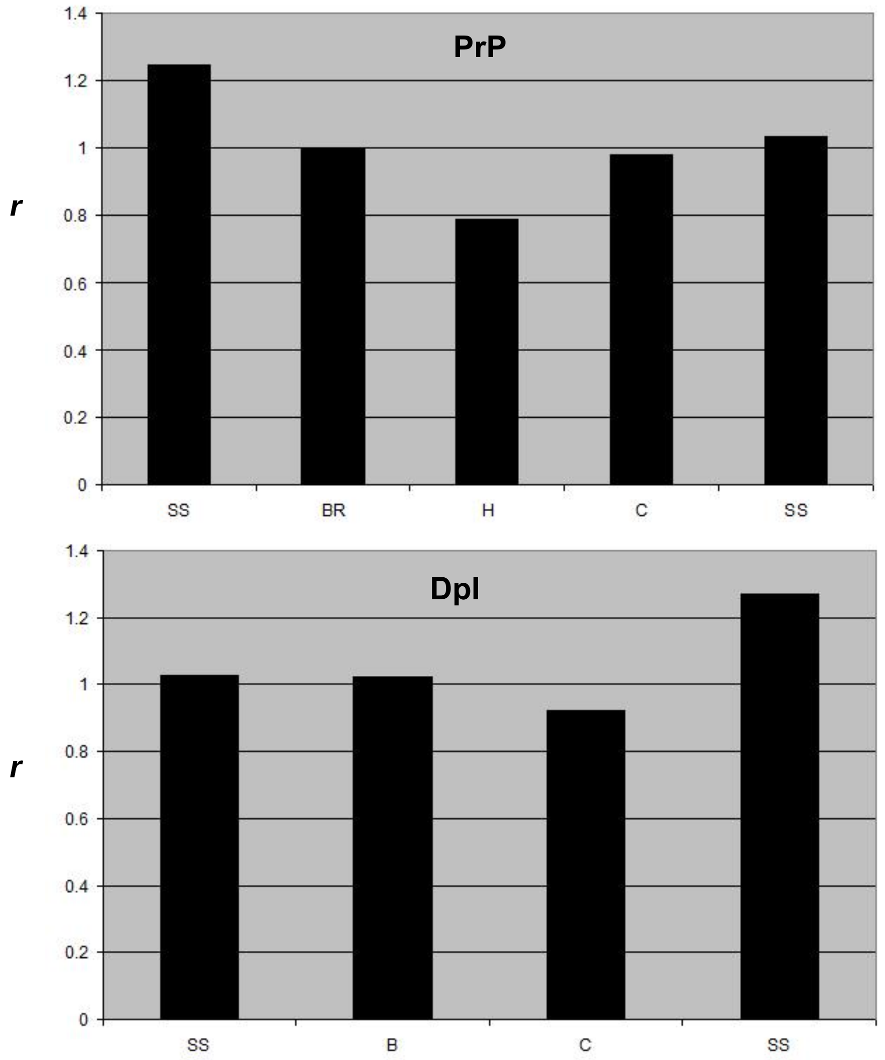


Figure 6-1 The rate of evolution (*r*) at each of the amino acid sites averaged across different sequence regions in PrP and Dpl. SS- signal sequence, BR- basic repeats, H- Hydrophobic region, C- C-terminal domain, B- basic (refer to Figure 5-3).

6.2.2 Distribution of conserved residues on the protein structure

The degree of conservation was analyzed among different lineages using ConSurf. PrP showed greater conservation compared with Dpl across different lineages (Figure 6-2). This suggests that Dpl is evolving at a faster rate compared with PrP in an attempt to acquire a specialized function in higher vertebrates. The evolutionary constraint on PrP may be related to the greater number of functions it carries out.

Predicting critical amino acid residues for structural stability: PrP and Dpl share a similar structural fold. To analyse the residues involved in the fold formation, the sequences from PrP and Dpl were compared among different lineages (Figure 6-7). A number of common conserved residues were found between PrP and Dpl which are listed below with the numbers corresponding to mouse sequence.

PrP: F141, Y149, Y150, N153, P158, Y163, F175, C179, N181, T183, C214

Dpl: F71, Y78, Y79, N82, P87, Y92, L105, C109, N111, T113, C143.

The Leu residue at position105 in mouse Dpl is uniquely different from other lineages within Dpl and PrP where Phe is dominant at that position (Figure 6-7). These residues conserved between the two proteins are distributed at similar positions in 3D space indicating significance in the fold formation and structural stability (Figure 6-3).

Predicting critical amino acid residues for function: Comparison of PrP and Dpl sequences were made among different lineages (Figure 6-5, Figure 6-6) to identify evolutionarily conserved residues. The numbering of the amino acid residues in the following section corresponds to mouse sequence.

PrP: Residues that are conserved among mammal, marsupial, avian and amphibian species (excluding those that are common between PrP and Dpl) are G126, G127, Y128, G131,F141, E146, R148,V161, R164,Y169, V176, D178, and M213. Residues those are conserved among mammal, marsupial and amphibian but missing in avian are A133, R136, P165, I182 and V209. Residues that are conserved among mammal, marsupial and avian but missing in *Xenopus* species are S135, Glu152, R156 and Y157 (Figure 6-5). The distribution of these residues in 3D space is shown in Figure 6-4. Most of the conserved residues are on the outer surface and possibly play a role in protein-protein interactions.

Dpl: Residues (constituting the mature protein) that are present in all mammals and at least in one marsupial and one of the amphibian species (excluding those that are common between PrP and Dpl) are E24, F36, D38, N49, E53, N67, and

V84. Residues that are present in all/most of the mammal and marsupial sequences compared (missing in amphibian species) are G22, G25, N26, Y33, I40, C45, T51 and L88 (Figure 6-6, Figure 6-4). One of the major structural distinctions between PrP and Dpl is the presence of a kink in helix B in Dpl. Dpl shows two different kinks. The first kink (kink1) is in helix A at position N82 (Figure 6-3) and the second (kink2) is in helix B at position N117 (Figure 6-4). Mo et al. (2001) proposed that N117 is possibly involved in the kink formation. This residue is conserved within Dpl in all species compared indicating it is critical to differentiate between PrP and Dpl (Figure 6-4). In both cases, Asn was preceded by Ala (A81 and A116). The equivalent of Dpl N82 in PrP is N153 in helix A (Figure 6-3). However, this residue did not produce a kink in PrP. This may indicate that the sequence Ala-Asn is needed to produce the kink. Interestingly, the equivalents of mouse Dpl A81 and N82 are missing in *M. domestica* and tammar wallaby (Figure 6-6) which may suggest this region is structurally different in marsupial species. Although the equivalent of mouse N117 is present in *Xenopus* species, the equivalent of mouse Dpl A116 is replaced by Arg in *Xenopus* species (Figure 6-6). This suggest that the kink in helix B may be missing in *Xenopus* species. Similar to PrP, most of the conserved residues are on the surface of the protein.

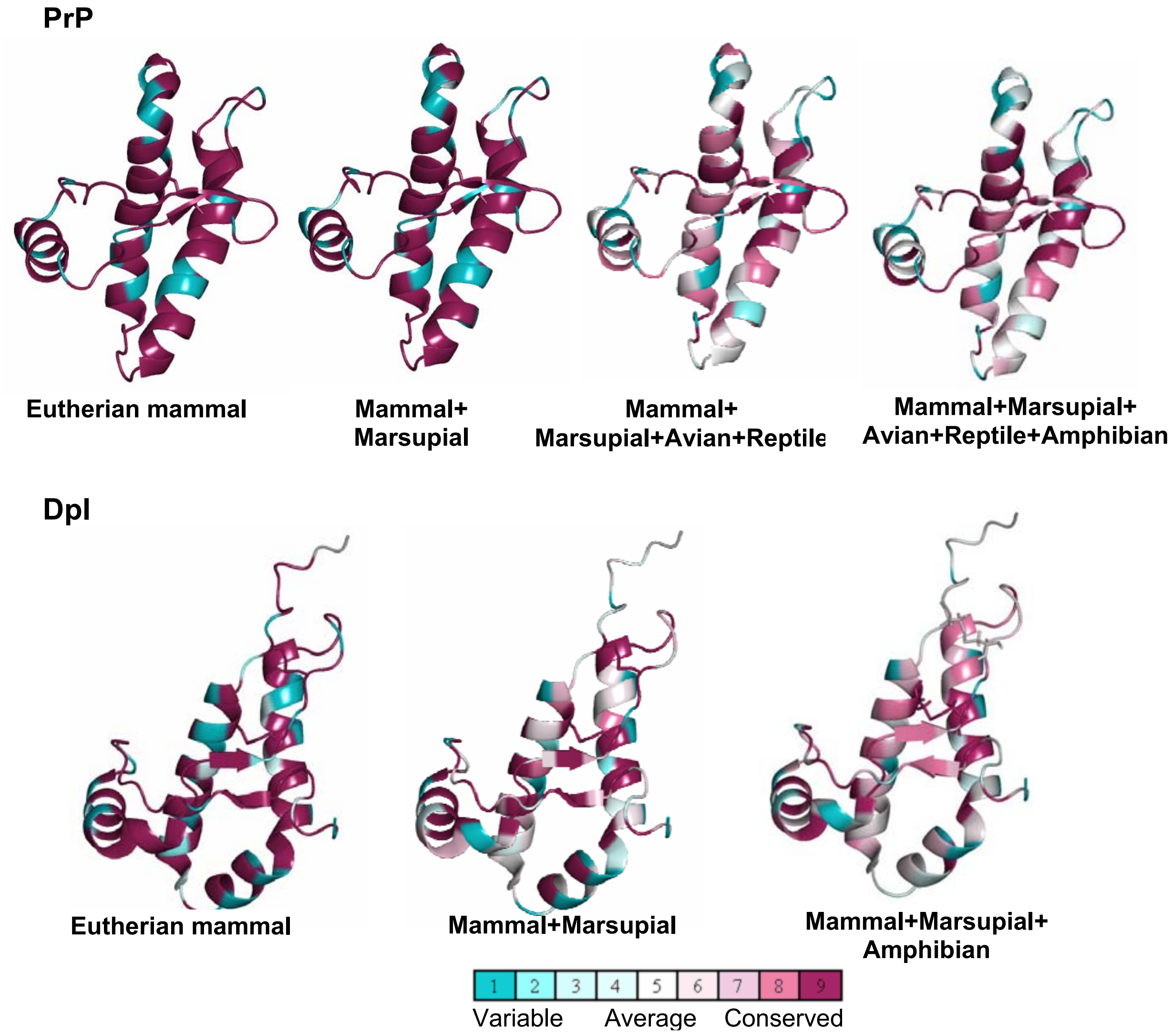


Figure 6-2 Distribution of conserved residues on mouse PrP and Dpl based on sequence conservation. Comparison was made among different lineages.

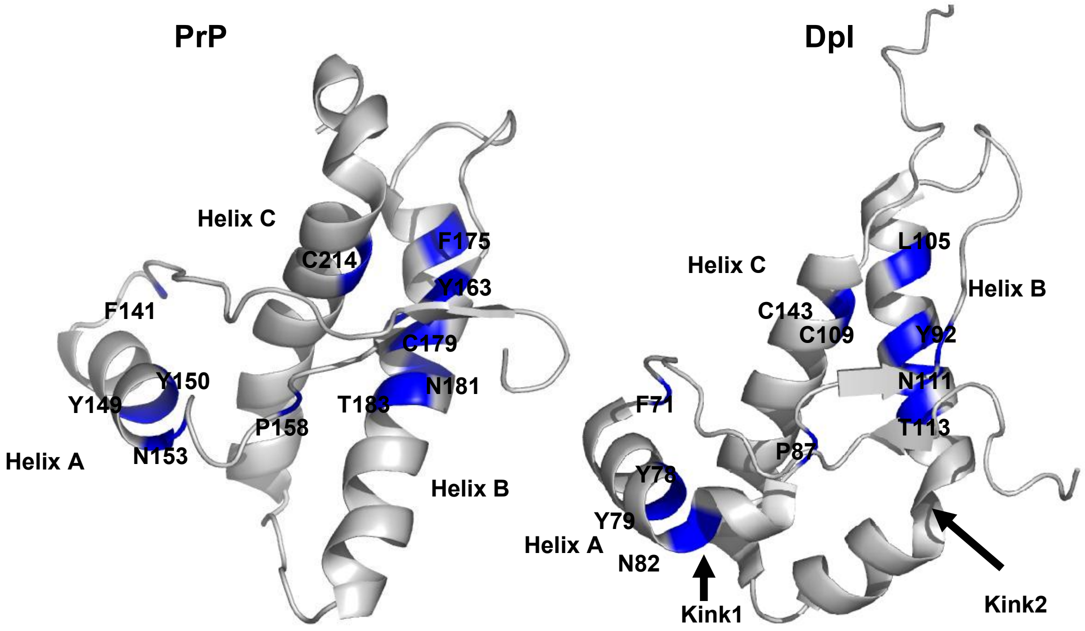


Figure 6-3 Identical residues in PrP and Dpl among mammals, marsupial and amphibian species shown in blue. The evolutionary and interprotein conservation indicates that these residues are critical for fold or structural stability.

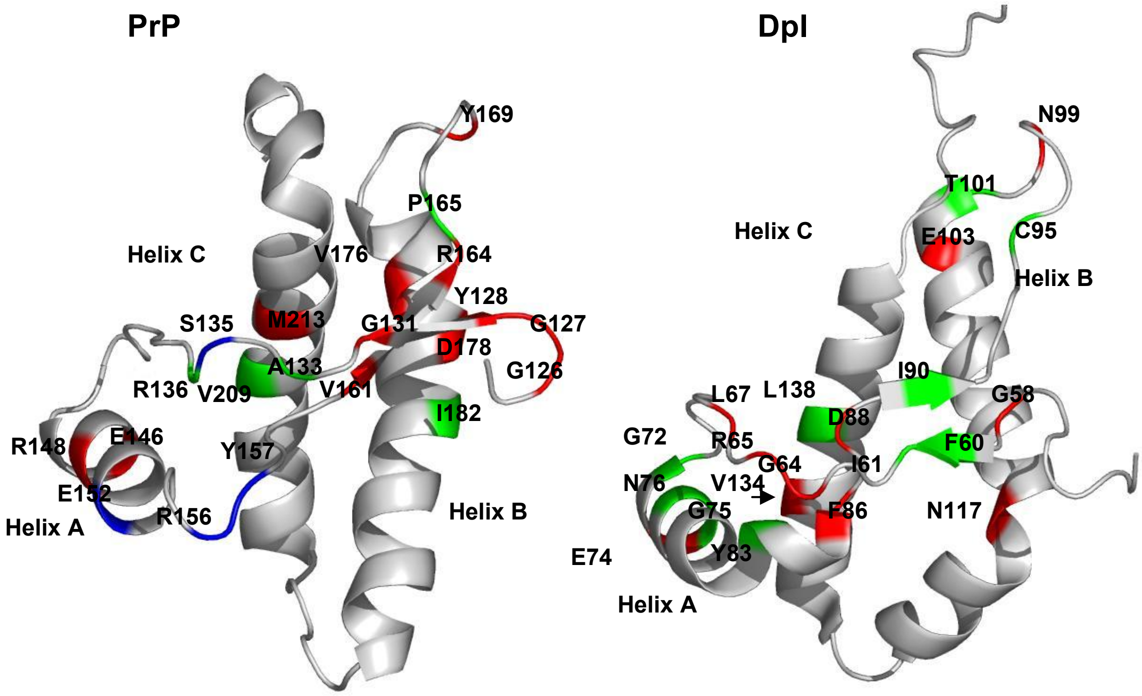


Figure 6-4 Distribution of evolutionarily conserved residues in PrP and Dpl. Sequence information from mammal, marsupial, avian and amphibian was used for PrP and mammals, marsupial and amphibian species was used for Dpl. Residues in Red are present in all lineages, Green residues are those missing in amphibian, and Blue (applicable to PrP only) missing in avian and/or reptile.

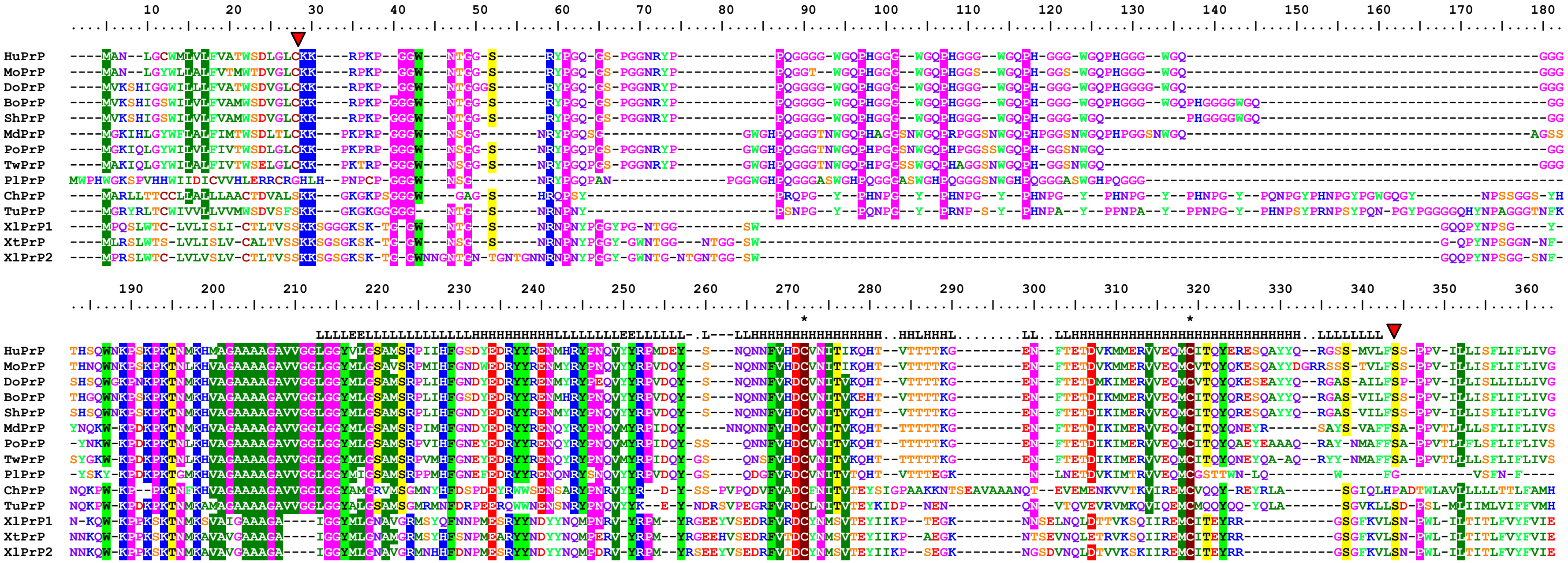


Figure 6-5 Multiple sequence alignment of PrP sequences. The secondary structural units are shown above the sequence (L- loop, E- sheet, H- helix). Cleavage sites are represented by ▼. The C-terminal domain region shows a number of conserved residues among all the species compared. The most highly conserved region corresponds to the hydrophobic region numbered between 200 and 220. The N-terminal repeats show the maximum number of variations. The Cys involved in disulphide bridge formation are indicated (*).

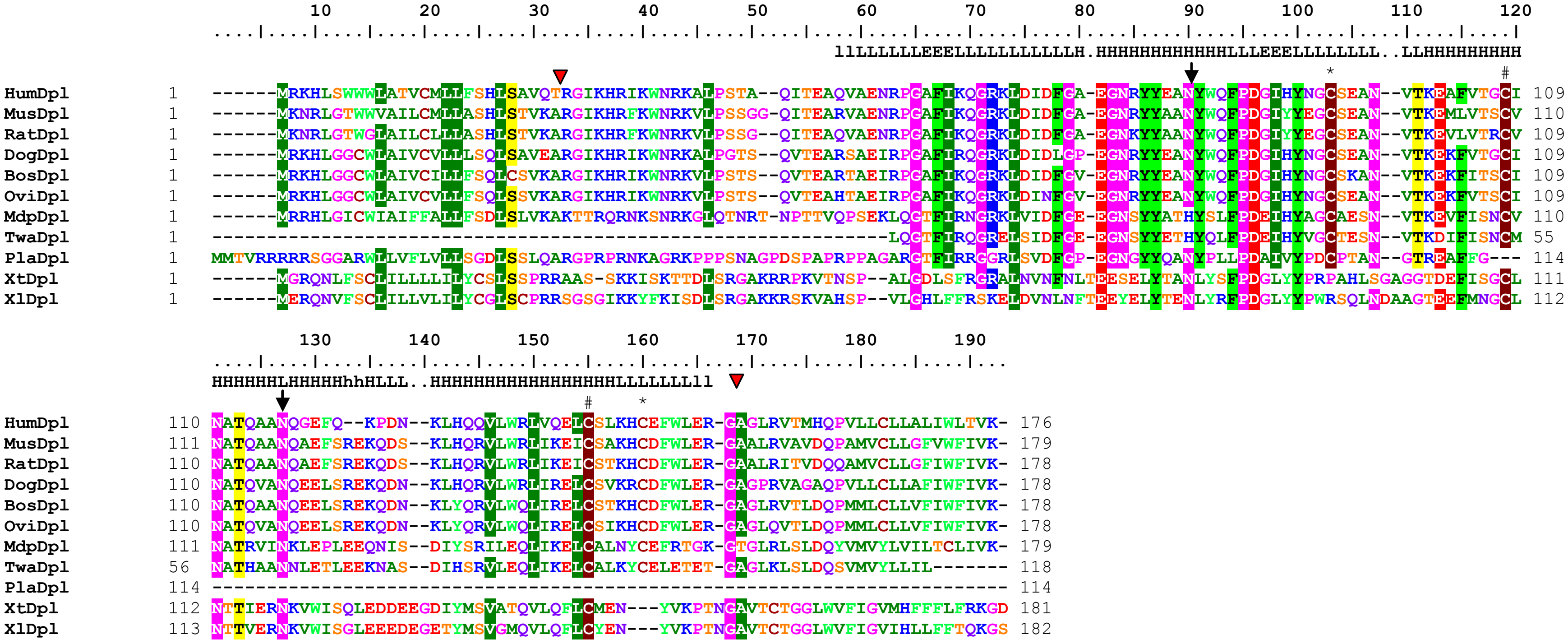
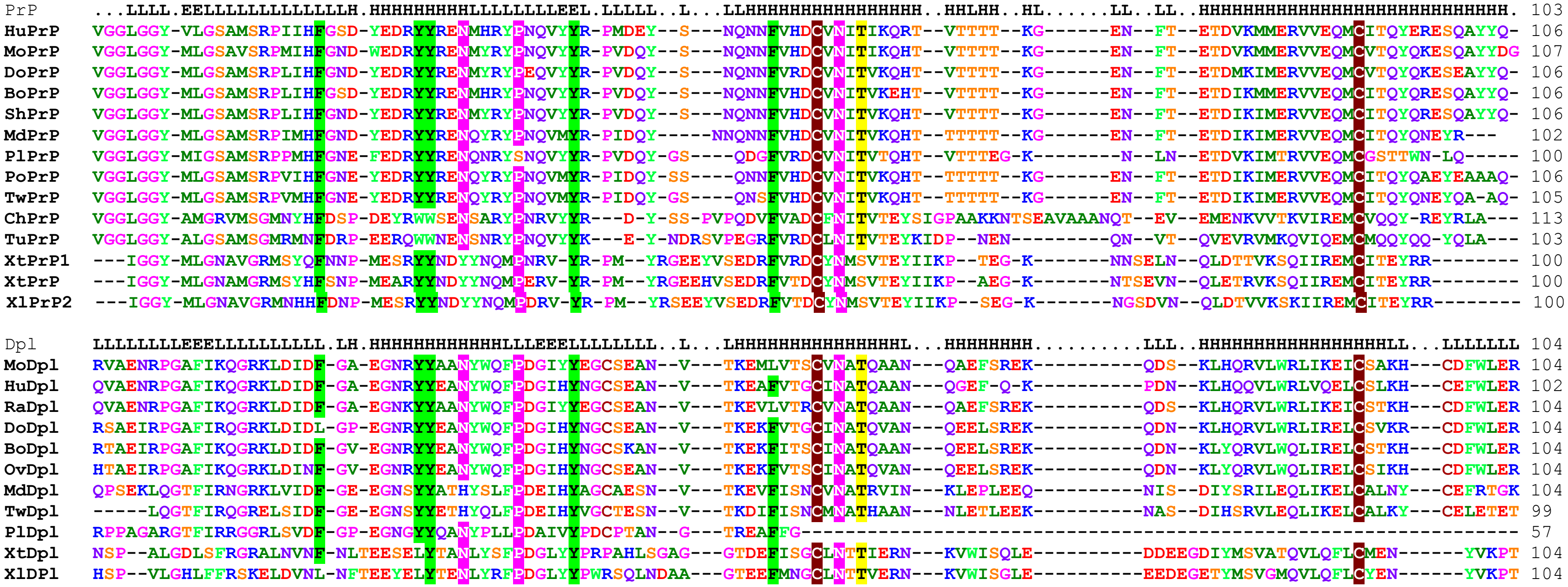


Figure 6-6 Multiple sequence alignment of Dpl sequences. The secondary structural units are shown above the sequence (L- loop, E- sheet, H- helix). Cleavage sites are represented by ▼. The C-terminal domain region shows fewer conserved residues compared with PrP. The highly conserved region corresponds to the Cysteines involved in disulphide bridge formation and the N-glycosylation sites. The Cys involved in disulphide bridge formation are indicated (*, #). Note the absence of the second disulphide bridge in *Xenopus* species. The Asn involved in Kink formation is indicated by ↓



```
PrP       ...LLLL.EELLLLLLLLLLLLLH.HHHHHHHHHLLLLLLLLEEL.LLLLL..L...LLHHHHHHHHHHHHHHHH..HHLHH..HL......LL..LL..HHHHHHHHHHHHHHHHHHHHHHHHHHHH. 103
HuPrP     VGGLGGY-VLGSAMSRPIIHFGSD-YEDRYYRENMHRYPNQVYYR-PMDEY--S---NQNNFVHDCVNITIKQRT--VTTTT--KG------EN--FT--ETDVKMMERVVEQMCITQYERESQAYYQ- 106
MoPrP     VGGLGGY-MLGSAVSRPMIHFGND-WEDRYYRENMYRYPNQVYYR-PVDQY--S---NQNNFVHDCVNITIKQHT--VTTTT--KG------EN--FT--ETDVKMMERVVEQMCVTQYQKESQAYYDG 107
DoPrP     VGGLGGY-MLGSAMSRPLIHFGND-YEDRYYRENMYRYPEQVYYR-PVDQY--S---NQNNFVRDCVNITVKQHT--VTTTT--KG------EN--FT--ETDMKIMERVVEQMCVTQYQKESEAYYQ- 106
BoPrP     VGGLGGY-MLGSAMSRPLIHFGSD-YEDRYYRENMHRYPNQVYYR-PVDQY--S---NQNNFVHDCVNITVKEHT--VTTTT--KG------EN--FT--ETDIKMMERVVEQMCITQYQRESQAYYQ- 106
ShPrP     VGGLGGY-MLGSAMSRPLIHFGND-YEDRYYRENMYRYPNQVYYR-PVDQY--S---NQNNFVHDCVNITVKQHT--VTTTT--KG------EN--FT--ETDIKIMERVVEQMCITQYQRESQAYYQ- 106
MdPrP     VGGLGGY-MLGSAMSRPIMHFGND-YEDRYYRENQYRYPNQVMYR-PIDQY-----NNQNNFVHDCVNITVKQHT--TTTTT--KG------EN--FT--ETDIKIMERVVEQMCITQYQNEYR---   102
PlPrP     VGGLGGY-MIGSAMSRPPMHFGNE-FEDRYYRENQNRYSNQVYYR-PVDQY-GS----QDGFVRDCVNITVTQHT--VTTEG-K--------N--LN--ETDVKIMTRVVEQMCGSTTWN-LQ----- 100
PoPrP     VGGLGGY-MLGSAMSRPVIHFGNE-YEDRYYRENQYRYPNQVMYR-PIDQY-SS----QNNFVHDCVNITVKQHT--TTTTT--KG------EN--FT--ETDIKIMERVVEQMCITQYQAEYEAAAQ- 106
TwPrP     VGGLGGY-MLGSAMSRPVMHFGNE-YEDRYYRENQYRYPNQVMYR-PIDQY-GS----QNSFVHDCVNITVKQHT--TTTTT--KG------EN--FT--ETDIKIMERVVEQMCITQYQNEYQA-AQ- 105
ChPrP     VGGLGGY-AMGRVMSGMNYHFDSP-DEYRWWSENSARYPNRVYYR---D-Y-SS-PVPQDVFVADCFNITVTEYSIGPAAKKNTSEAVAAANQT--EV--EMENKVVTKVIREMCVQQY-REYRLA--- 113
TuPrP     VGGLGGY-ALGSAMSGMRMNFDRP-EERQWWNENSNRYPNQVYYK---E-Y-NDRSVPEGRFVRDCLNITVTEYKIDP--NEN---------QN--VT--QVEVRVMKQVIQEMCMQQYQQ-YQLA--- 103
XtPrP1    ---IGGY-MLGNAVGRMSYQFNNP-MESRYYNDYYNQMPNRV-YR-PM--YRGEEYVSEDRFVRDCYNMSVTEYIIKP--TEG-K-------NNSELN--QLDTTVKSQIIREMCITEYRR-------- 100
XtPrP     ---IGGY-MLGNAMGRMSYHFSNP-MEARYYNDYYNQMPERV-YR-PM--YRGEEHVSEDRFVTDCYNMSVTEYIIKP--AEG-K-------NTSEVN--QLETRVKSQIIREMCITEYRR-------- 100
XlPrP2    ---IGGY-MLGNAVGRMNHHFDNP-MESRYYNDYYNQMPDRV-YR-PM--YRSEEYVSEDRFVTDCYNMSVTEYIIKP--SEG-K-------NGSDVN--QLDTVVKSKIIREMCITEYRR-------- 100

Dpl       LLLLLLLLEEELLLLLLLLLL.LH.HHHHHHHHHHHHLLLEEELLLLLLLL..L...LHHHHHHHHHHHHHHHL...HHHHHHHH..........LLL..HHHHHHHHHHHHHHHHHLL...LLLLLLL 104
MoDpl     RVAENRPGAFIKQGRKLDIDF-GA-EGNRYYAANYWQFPDGIYYEGCSEAN--V---TKEMLVTSCVNATQAAN---QAEFSREK----------QDS--KLHQRVLWRLIKEICSAKH---CDFWLER 104
HuDpl     QVAENRPGAFIKQGRKLDIDF-GA-EGNRYYEANYWQFPDGIHYNGCSEAN--V---TKEAFVTGCINATQAAN---QGEF-Q-K----------PDN--KLHQQVLWRLVQELCSLKH---CEFWLER 102
RaDpl     QVAENRPGAFIKQGRKLDIDF-GA-EGNKYYAANYWQFPDGIYYEGCSEAN--V---TKEVLVTRCVNATQAAN---QAEFSREK----------QDS--KLHQRVLWRLIKEICSTKH---CDFWLER 104
DoDpl     RSAEIRPGAFIRQGRKLDIDL-GP-EGNRYYEANYWQFPDGIHYNGCSEAN--V---TKEKFVTGCINATQVAN---QEELSREK----------QDN--KLHQRVLWRLIRELCSVKR---CDFWLER 104
BoDpl     RTAEIRPGAFIKQGRKLDIDF-GV-EGNRYYEANYWQFPDGIHYNGCSKAN--V---TKEKFITSCINATQAAN---QEELSREK----------QDN--KLYQRVLWQLIRELCSTKH---CDFWLER 104
OvDpl     HTAEIRPGAFIKQGRKLDINF-GV-EGNRYYEANYWQFPDGIHYNGCSEAN--V---TKEKFVTSCINATQVAN---QEELSREK----------QDN--KLYQRVLWQLIRELCSIKH---CDFWLER 104
MdDpl     QPSEKLQGTFIRNGRKLVIDF-GE-EGNSYYATHYSLFPDEIHYAGCAESN--V---TKEVFISNCVNATRVIN---KLEPLEEQ----------NIS--DIYSRILEQLIKELCALNY---CEFRTGK 104
TwDpl     -----LQGTFIRQGRELSIDF-GE-EGNSYYETHYQLFPDEIHYVGCTESN--V---TKDIFISNCMNATHAAN---NLETLEEK----------NAS--DIHSRVLEQLIKELCALKY---CELETET 99
PlDpl     RPPAGARGTFIRRGGRLSVDF-GP-EGNGYYQANYPLLPDAIVYPDCPTAN--G---TREAFFG----------------------------------------------------------------- 57
XtDpl     NSP--ALGDLSFRGRALNVNF-NLTEESELYTANLYSFPDGLYYPRPAHLSGAG---GTDEFISGCLNTTIERN---KVWISQLE----------DDEEGDIYMSVATQVLQFLCMEN------YVKPT 104
XlDPl     HSP--VLGHLFFRSKELDVNL-NFTEEYELYTENLYRFPDGLYYPWRSQLNDAA---GTEEFMNGCLNTTVERN---KVWISGLE----------EEDEGETYMSVGMQVLQFLCYEN------YVKPT 104
```

Figure 6-7 Comparison of C-terminal domain of PrP and Dpl among different lineages. The conserved residues among the two proteins may indicate critical residues in the fold for structural stability.

6.2.3 Dotplot analysis

The amino acid sequences from various species were analysed using dotplots. The N-terminal repeat region in PrP appears to be evolving gradually from amphibian to eutherian mammals (Figure 6-8). This region is very distinct among the mammalian species and chicken, whereas it is less pronounced in turtle and absent in amphibian. The other feature is that the repeat region is clustered among the eutherian mammals, and marsupial species indicating that the composition of the repeats is lineage specific. The middle stretch of hydrophobic region is very well conserved among the species compared.

Dpl showed lesser sequence similarity compared to PrP between eutherian mammals and marsupial species. *Xenopus* sequence in both PrP and Dpl showed little similarity compared to rest of the sequences.

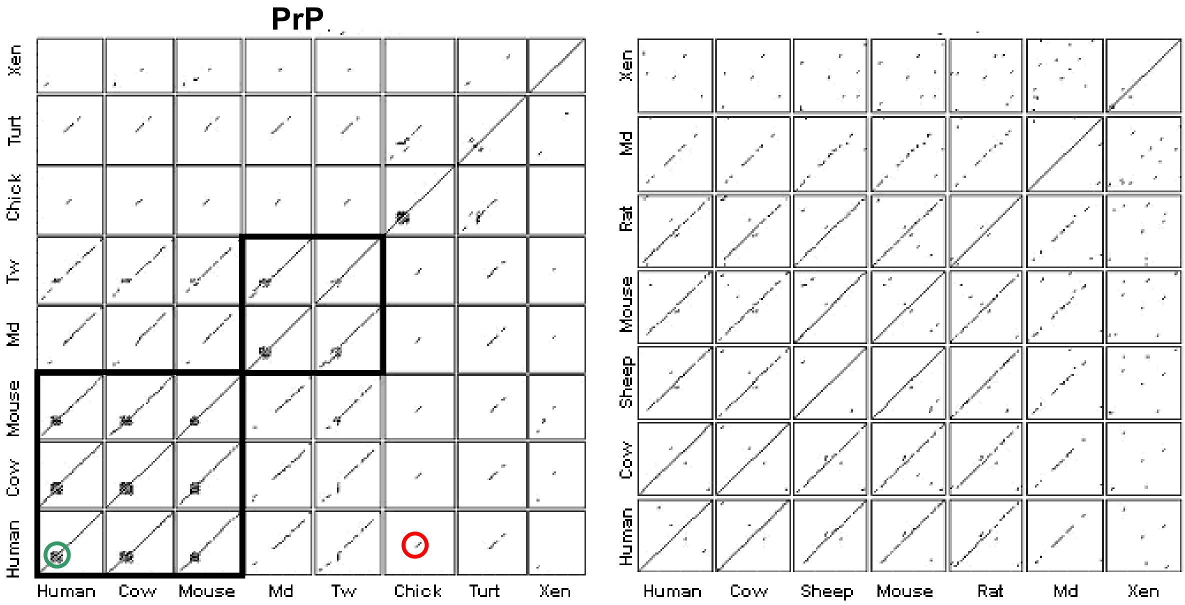


Figure 6-8 Pairwise sequence comparison for PrP and Dpl using Dotplot. PrP: Evolution of the N-terminal repeats can be seen from *Xenopus* to human. The repeat region (circled in green) is lineage specific. Note the conserved hydrophobic region (circled in red) in most of species compared. Dpl: Note the conservation of sequence among mammals and the minimal sequence match between *Xenopus* and other species. [Md=*M. domestica*; Xen =*Xenopus*]

6.2.4 Evolution of functional elements in 3'UTR

The 3'UTR in mammalian *PRNP* sequences was reported to have a number of conserved sequence motifs (Cytoplasmic polyadenylation element (CPE), nuclear-specific polyadenylation signal site) (Premzl et al. 2005). These sequence elements are also found to be conserved in the platypus sequence (Figure 6-9A). However, these motifs elements are not seen in other lower vertebrate (chicken and amphibian) indicating that they evolved in higher vertebrates. The 3'UTR regions in *Xenopus*, chicken and turtle were also compared. There were no common conserved elements in these sequences but, interestingly, separate conserved elements were detected between turtle and chicken (Figure 6-9B), and between *Xenopus* and turtle 3'UTRs (Figure 6-9C). Whether these conserved elements are just a random match or due to evolutionary selective pressure is not known.

6.3 Conclusion

Comparative sequence analysis of PrP and Dpl among different lineages indicates Dpl is evolving at a faster rate than PrP. This may be due to Dpl, a recently duplicated gene, trying to acquire a specialized function in higher vertebrates. Analysis of the rate of evolution within *PRNP* and *PRND* showed apparently different evolutionary constraints acting on different regions of the protein sequences. The signal sequences which are cleaved to form the mature protein are evolving at a higher rate compared with other regions of the mature protein. Through comparative sequence analysis, I could identify the critical residues involved in the structural stability of the C-terminal domain fold of PrP and Dpl. Comparisons made within PrP and Dpl among different species indicated a number of conserved residues. These residues may be critical for function and protein-protein interactions as most of the residues are on the surface of the protein. The unique N-terminal repeats in PrP gradually evolved, from *Xenopus* where they are almost absent to mammals, and show lineage specificity. The conserved 3'UTR elements in *PRNP,* known to play a role in the post-transcriptional regulation, appear to be a modern feature which are present in monotremes but absent in the other lower vertebrates.

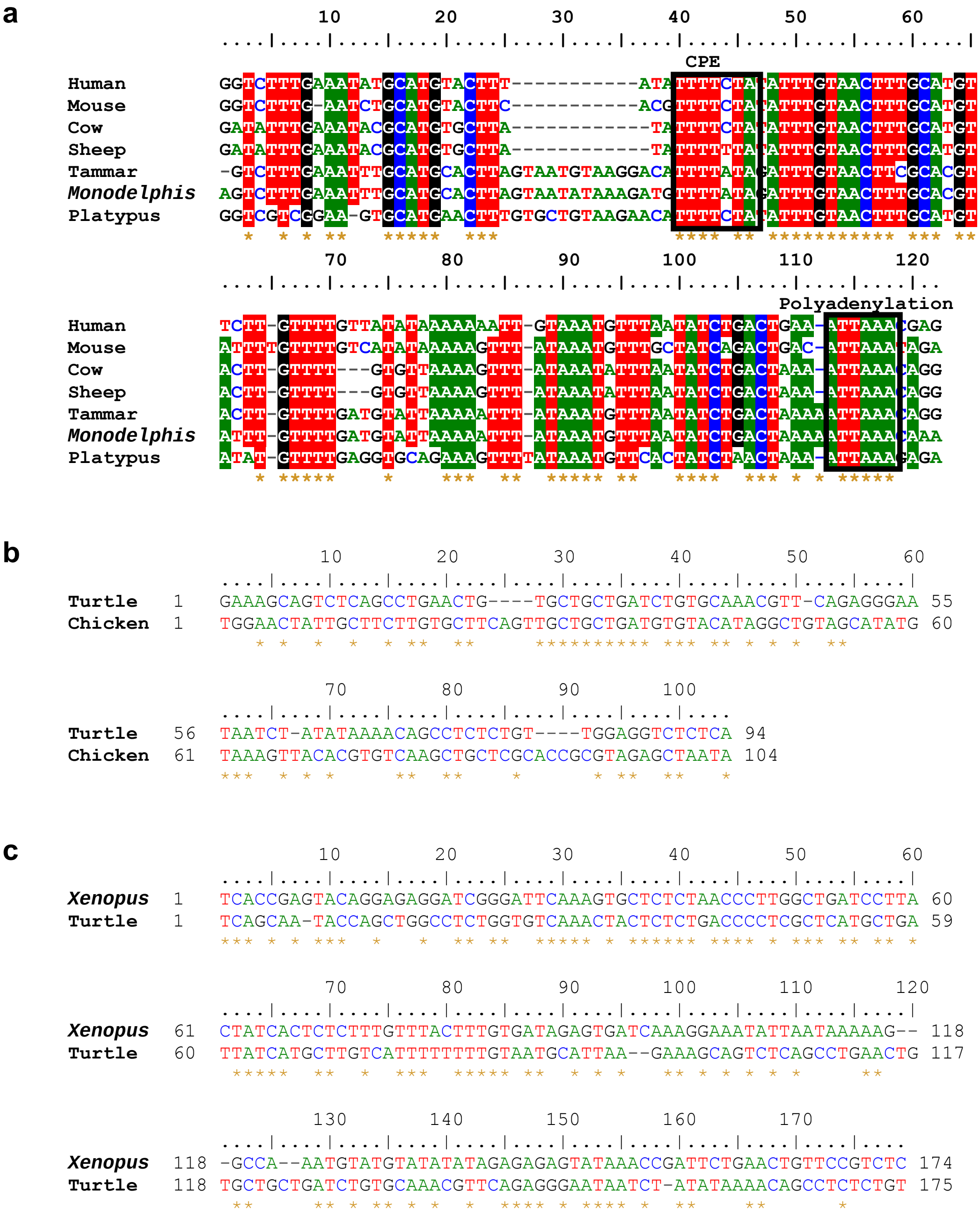


Figure 6-9 Conserved sequence elements in *PRNP* 3'UTR. (a) Demonstration of the presence of the *PRNP* 3'UTR sequence elements in platypus as reported in Premzl et al. (2005). Conserved sequence elements in the 3' UTR between (b) turtle and chicken and (c) *Xenopus* and turtle.

7 Phylogenetic footprinting and transcription factor binding analysis of prion-protein family genes

7.1 Background

The publication of draft sequence of newly sequenced genomes gives enormous potential for characterizing functional elements by using comparative genomic approaches. One of the key functional elements are proteins termed transcription factors (TFs) which play a central role in RNA-polymerase II- mediated transcriptional regulation of gene expression by binding to specific short DNA sequence motifs known as transcription factor-binding sites (TFBSs) or *cis*-regulatory elements. These regulatory sequence elements at the DNA level include: promoters; enhancers; silencers; locus control regions; and matrix attachment regions/scaffold attached regions (Figure 7-1). Predicting the binding of TFs to a particular gene gives an opportunity for deeper understanding of the potential functions of the genes regulated.

Identifying regulatory modules: When a gene is to be expressed in a number of different circumstances in a developing organism, it is usually found that separate *cis*-regulatory sub-elements carry out different parts of the overall regulatory job. These sub-elements are referred to as regulatory modules. Individual modules are always found to contain multiple TF target sites, and these contribute in various ways to overall regulatory output. The most remarkable cases are found among genes encoding TFs that are expressed in complex spatial patterns and at different times. Target sites for the set of TFs required to generate a given spatial regulatory output are often found clustered together more or less contiguously, within a given sequence region of the *cis*-regulatory DNA.

Experimental detection of TFBSs is widely done but is time consuming and expensive if prior binding information is not available. A number of reliable computational methods have been developed to predict TFBSs, which are economical both in terms of time and resources and can produce useful predictions for experimental validation. As TFBSs are under greater selective pressure than other non-protein-coding DNA, the reliability of predicting them is greatly improved by use of comparative genomics to filter out sequence noise. Identifying such conserved sequence elements in non-coding regions of homologous genes from phylogenetic comparison is called ‘phylogenetic footprinting’ (PF) (Tagle et al. 1988).

The aim of this study was to develop a systematic high throughput screening pipeline to first search for conserved motifs using two different PF methods (motif-discovery and alignment-based) , and then to rapidly evaluate the motifs as potential TFBSs for members of the PrP family. The results are displayed in an interactive graphical user interface, FactorScan, which integrates three separate complementary databases (conserved sequence motifs, TFBS motifs, TRANSFAC). This pipeline was applied to TFBS analysis of the orthologous gene regions of PrP family genes from vertebrate lineages, taking account of the gene annotations. Some initial insights into the functions of these genes, which are not well understood, were gleaned from the TFs predicted to be involved in regulating their tissue expression. (Please note that in the following sections, the term mammal refers to eutherian-mammal).

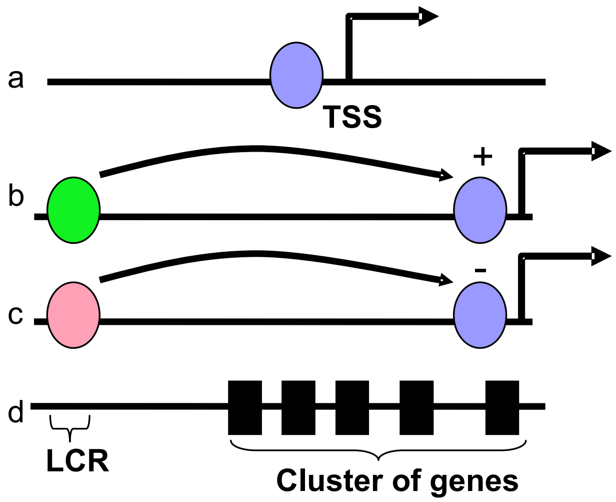


Figure 7-1 Schematic representation of (a) promoter, (b) enhancer, (c) silencer and (d) locus control region (LCR); [TSS= transcription start site]

7.2 Motivation for developing the pipeline

Several programs implement PF but only a few combine it with TFBS analysis, for example, ConSite (Sandelin et al. 2004) and rVISTA (Loots et al. 2002) that allow only pairwise comparison; there are currently no programs which perform this analysis on multiple sequences. Another restriction is that they use databases of position weight matrices (PWM), TRANSFAC public and JASPAR respectively, neither of which is as comprehensive as TRANSFAC professional. Finally, rVISTA and ConSite do not provide a facility to customize display of the results to make the maximum use of the output, for example, display of clusters of TFs.

While there are several online resources which can perform PF, none provides the flexibility for combining the conserved sequence-motif data with TFBS analysis and, at the same time, allowing the flexibility to customize the searches based on gene annotation information. To address this deficiency, I developed a two-step procedure which combines PF with TFBS analysis. This automated pipeline enables us to carry out rapid screening and evaluation of the phylogenetically conserved motifs for potential TF binding affinity. To perform the most comprehensive searches, TRANSFAC professional database was included in the pipeline.

The current approach overcomes the restrictions listed above, by providing various options for customizing searches for both pairwise and multiple sequences, for incorporating flexibility in visualizing the output, and for using databases of PWMs of choice.

7.3 Software and Hardware

Standalone versions of AVID (version 2.1), LAGAN (version 1.21) and FootPrinter (version 2.1) were used for the PF analysis. The TRANSFAC database version 9.2 and Match version 6.1 were used for TFBS analysis. The web form was implemented using HTML running on an Apache web server on a Linux operating system at valera.anu.edu.au that hosts the web page and can be accessed locally with the web address http://valera.anu.edu.au:8080/factorScan.html. The graphical package Perl GD and Common Gateway Interface package Perl CGI were used for the web interface development. Additional pipelining and analysis modules were written in Perl. All analysis was performed on a PC but some of the more memory demanding FootPrinter analyses was performed on the Dell Linux cluster at the APAC (Australian Partnership for Advanced Computing) National Facility.

7.4 Development of the pipeline

7.4.1 Criteria for species selection

An essential first step in PF is to identify orthologous genes in different species. At relatively close evolutionary distances (40-80 million years ago), it can be difficult to distinguish between functional or non-functional conserved sequences because there may not have been enough evolutionary time for accumulation of mutations. Hence, comparison with distantly related species can improve the ability to distinguish the conserved sequence that is a result of functional constraint from that retained due lack of divergence time (Frazer et al. 2003). The comparison of orthologous DNA sequences between evolutionarily distant species with greater divergence time would enable the prediction of the non-coding sequences with greater confidence. However, if the species being compared are too distant, then detecting conserved elements will be difficult because they would have diverged too much to show any conservation or they may have evolved different regulatory processes.

Duret and Bucher (1997) suggested that any sequence conservation between species that diverged 300 Myr (million years) ago indicates a strong selective pressure based on the rate of substitution of neutral bases, estimated to be around 0.5% every Myr (Li et al 1985). They also suggested that the species should be selected so that the cumulative length of branches of the phylogenetic tree uniting them to their last common ancestor represents >200 Myr. Stojanovic et al (1999) suggested that each lineage diverge independently after separation from a common ancestor which results in additive effect of their evolutionary distances.

Multiple mammalian sequences were selected as this would improve the resolving power. The timescale of evolution for these species is shown in Figure 7-2 with fish at the bottom of the timescale which is separated from human by 450 million years of evolution (Kumar and Hedges 1998). Alignment of chimp DNA sequence with human showed more than 90% sequence similarity; hence, chimp was not included for analysis (Figure 7-6).

To improve the signal-to-noise ratio, representative species for which genomic data for *PRNP* and *PRND* are available were selected. This comprises several mammalian species, and all those available for lower vertebrates; marsupial mammals *M. domestica* (South American opossum) and tammar wallaby, chicken, and the frog *X. tropicalis*. Platypus was not included in the analysis as complete genomic sequence information for the *PRNP* gene locus is not available. There are significant differences in the lengths of the intronic and intergenic regions of these genes, both among mammals and among the vertebrate lineages due to the high frequency of insertion of transposable elements (Premzl et al. 2004).

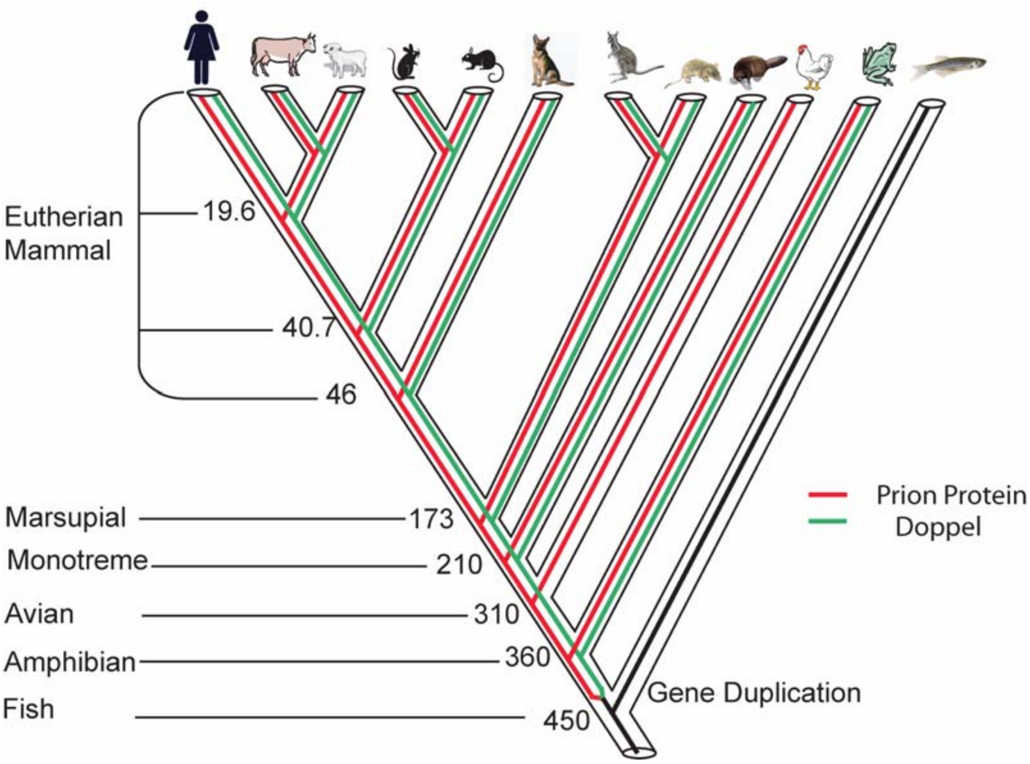


Figure 7-2 Timescale of evolution of vertebrate species of interest with the time shown on the horizontal bar in millions of years (Kumar and Hedges 1998).

A database of annotated gene sequences was created by mapping the *PRNP* and *PRND* cDNA sequence obtained from either experiment or public databases onto the genome sequence obtained from various genome sequencing projects. The EMBOSS application (Rice, Longden & Bleasby 2000) “est2genome” was used to annotate the exon-intron boundaries, and transcription start site (TSS), while “getorf” was used for detecting the coding regions which were then masked. Enhancers and silencers are reported to act at a distance from the TSS (Donoghue et al. 1988; Schachat and Briggs 2002). Hence, genomic sequence covering 2 kb upstream of the TSS, the whole of exon-intron region, and 2 kb downstream from the transcription stop site was included in the PF analysis. The sequences (other than ORF) were not masked as the aim was to look for conserved motifs by sequence comparison. The sequence dataset was divided into independent gene regions (*PRNP* and *PRND*) and one region containing *PRNP*, *PRND* together with its intergenic sequence (Figure 7-3).

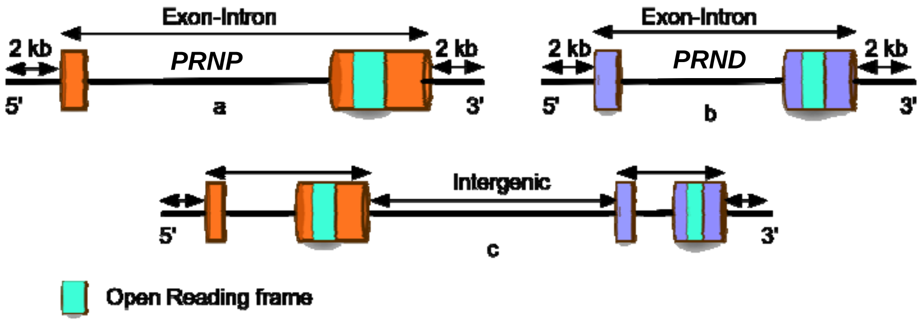


Figure 7-3 Annotated gene sequence information for (a) *PRNP*, (b) *PRND* and (c) *PRNP* plus *PRND* including intergenic region. 2kb upstream to the TSS, 2 kb downstream to the 3’ UTR and all sequence between TSS and transcription stop site except ORF was included for analysis.

7.4.2 Conserved-sequence motif detection

Conserved sequence motifs were identified by several PF methods which are categorized into two groups, alignment-based and motif-discovery-based. Separate pipelines for each, were developed.

7.4.2.1.1 Global alignment with AVID/LAGAN

To perform end-to-end comparisons, the global pairwise-alignment methods AVID (Bray et al. 2003) and LAGAN (Brudno et al. 2003) were used independently to generate pairwise alignments.

7.4.2.2 Search strategy for alignment-based methods

The input sequences are in FASTA format with the sequence identifier format as specieGene (example humDpl). Each sequence is in a separate file and the sequence identifier must match with the filename to enable processing of the sequence using the filename programmatically.

AVID performs the pairwise alignments of two input sequence files (example: file1 and file2) and generates three different outputs with filenames file1_file2.info, file1_file2.minfo, file1_file2.out. The pairwise alignment is written to the .out file which is needed for downstream processing.

LAGAN also generates pairwise alignments by taking two input sequence files. It performs local alignment first and then joins the aligned regions with gaps. The option for translate anchor was used enabling the protein coding regions to be anchored for a better alignment. Binary output format was selected which would allow downstream processing. Unlike AVID, LAGAN does not generate the output file name automatically and it needs to be specified. To maintain consistency and to make the programming step simpler, the same filename format as in AVID was specified.

Both the AVID and LAGAN alignments for all possible pairwise combinations (Figure 7-4) of sequences in the annotated gene sequence database were performed using the Perl script “doAlign.pl”.

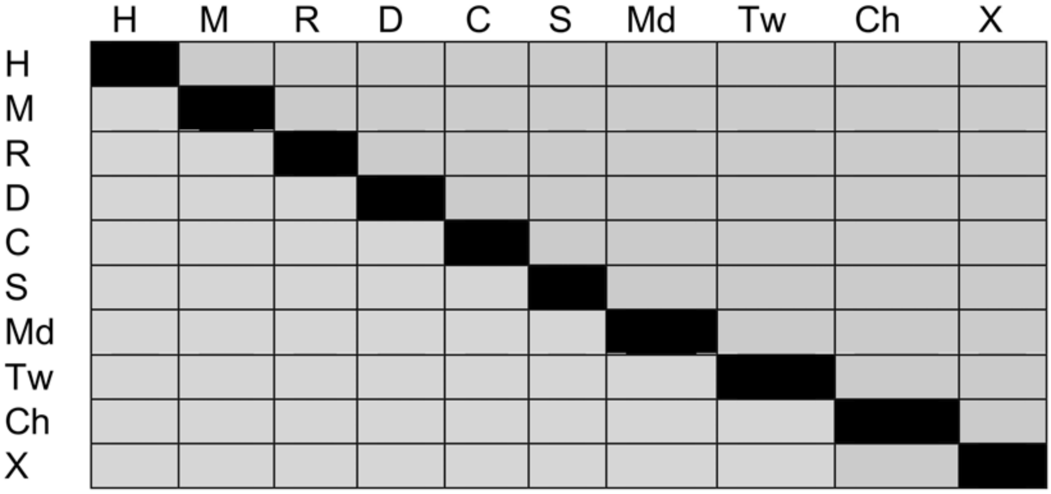


Figure 7-4 Summary of pairwise sequence comparisons (all grey cells) performed with AVID and LAGAN between the species on the X and Y axes. H, human; M, mouse; R, rat; D, dog; C, cow; S, sheep; Md, *M. domestica*; Tw, tammar wallaby; Ch, chicken; X, *Xenopus*.

7.4.2.3 Annotation with VISTA

VISTA is a program for visualization and annotation of global alignments of arbitrary length (Frazer et al. 2004). It is efficient in annotating the alignments based on user-defined parameters. VISTA plots are based on sliding a user-defined window over the entire alignment and calculating the percent identity at each base pair in the window. Conserved regions are reported based on the user defined percentage and length cutoffs. VISTA can be configured by changing several parameters (e.g. percentage identity and length), which can be defined in the input Plotfile.

Global pairwise alignments generated by AVID and LAGAN were annotated using VISTA.

7.4.2.3.1 Plotfile

VISTA executes the information specified in the Plotfile. The file path of the input alignments, percent and length cutoffs for plotting are specified in the Plotfile. To facilitate trialing of several combinations of percent identity (range: 75% to 100%) and length (range: 8 to 15 bp) values, a Perl script “runVista.pl” was developed to generate corresponding Plotfiles for percent identity and length values passed as command line arguments. In order to annotate the genes, it needs another input file containing details of the annotation (Figure 7-5). The annotation file contains information about the start and end base numbers for exon, intron, untranslated and coding regions. The annotation file for each of the sequences was made in the format suitable for VISTA.

7.4.2.3.2 Output files

VISTA generates three different output files: VISTA plot, alignment, and region files. VISTA plot contains graphical representation of the conserved regions (Figure 7-6).

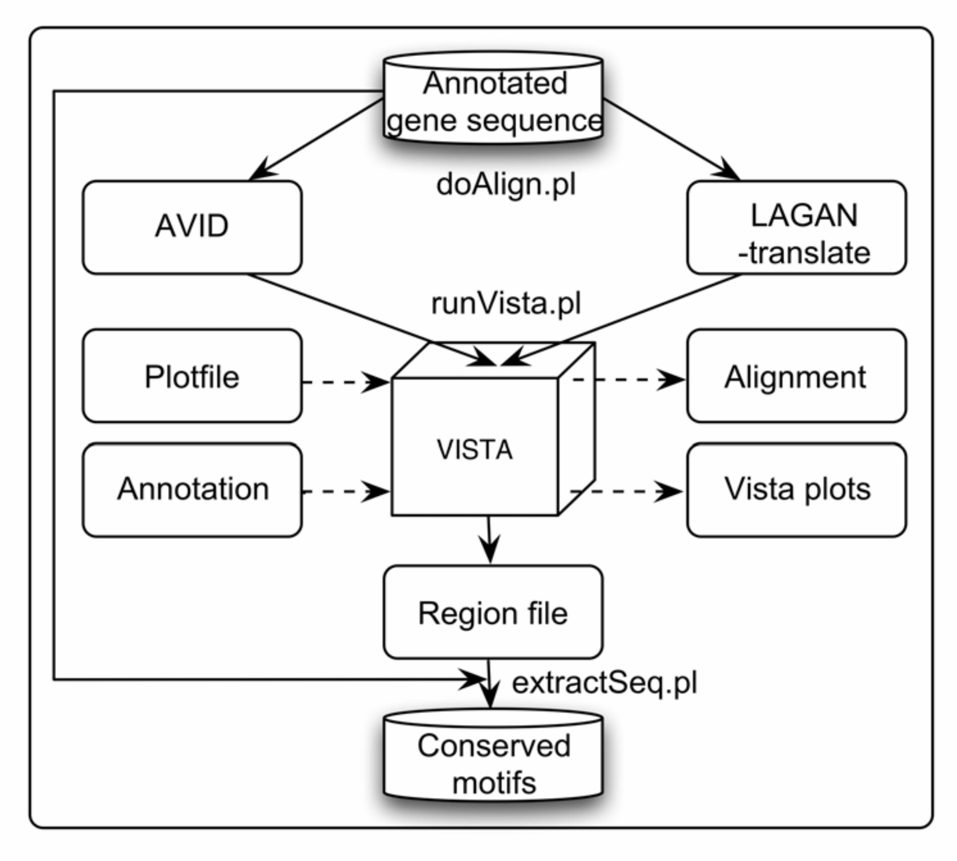


Figure 7-5 Flow chart showing the steps involved in PF using AVID/LAGAN pairwise alignments and annotation with VISTA. The final result of the analysis is the conserved sequence motif database. The background input and output files are shown with dotted arrows.

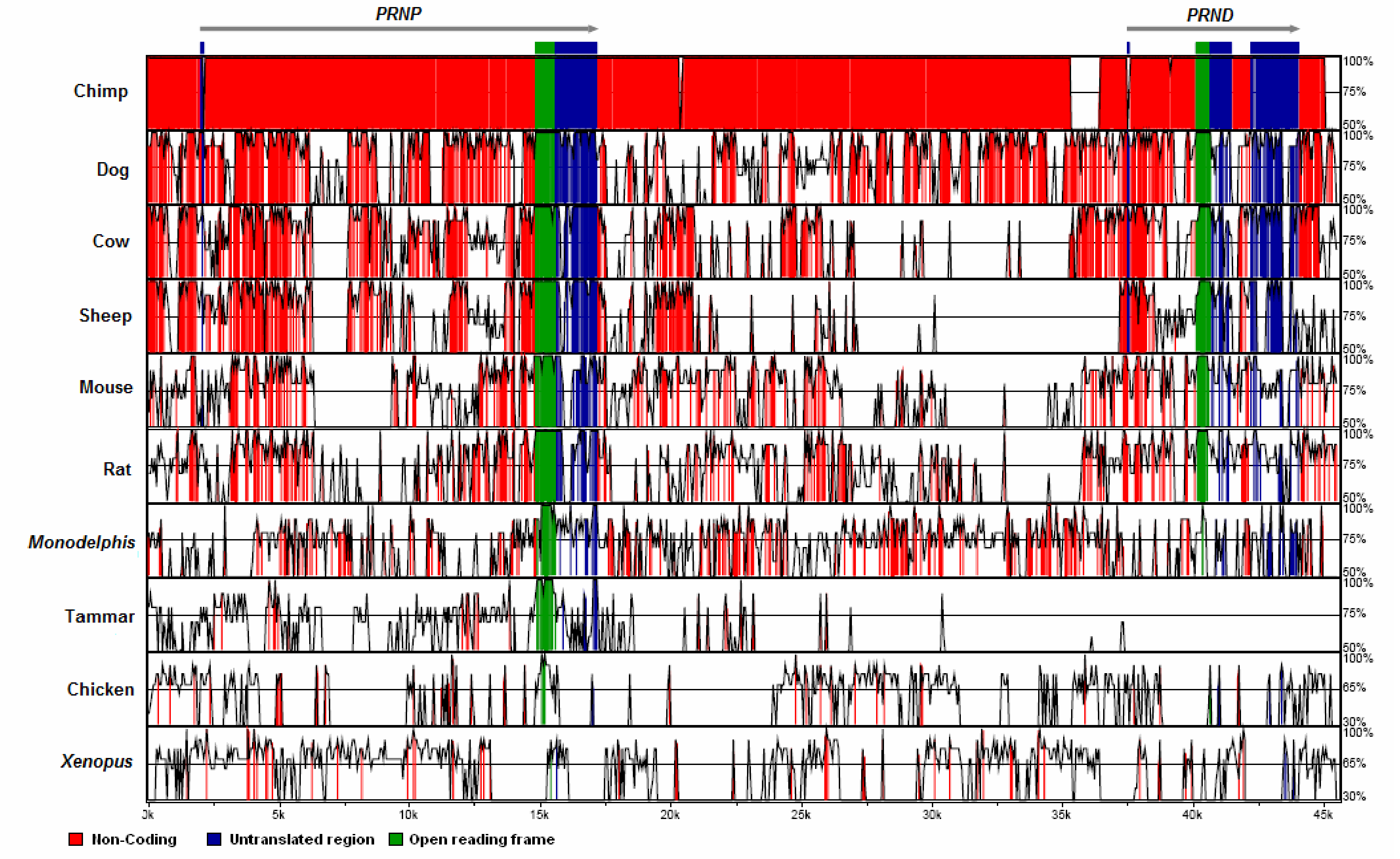
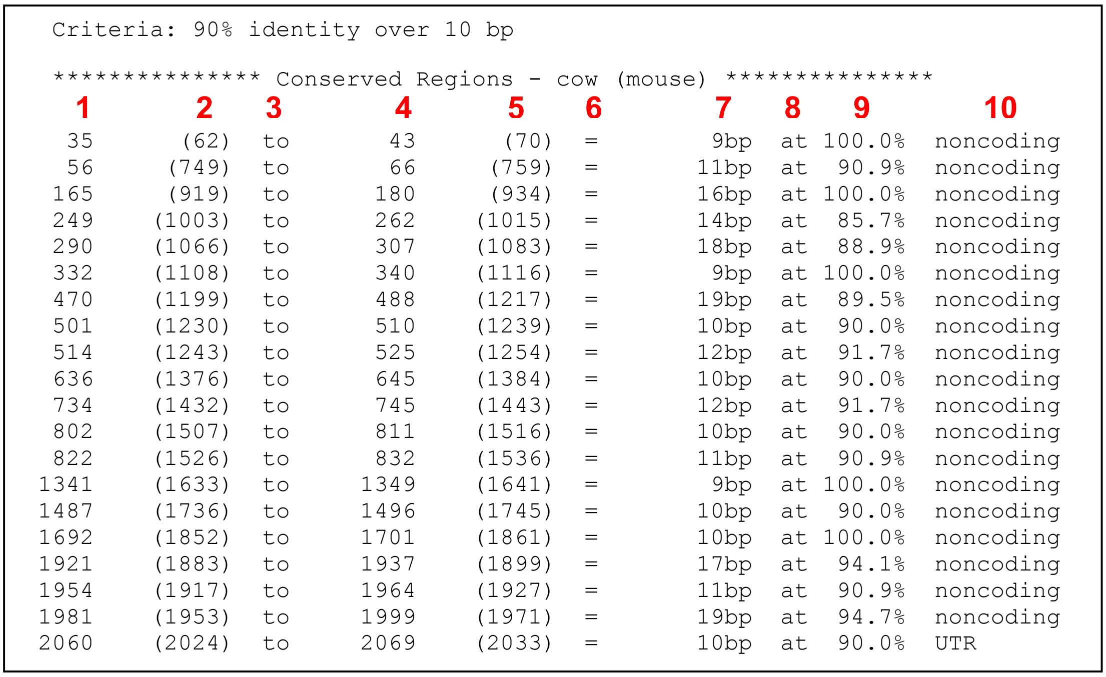


Figure 7-6 VISTA plot of pairwise alignments (AVID) between human and the remainder of the sequences on the X-axis. The alignments constituted the *PRNP* and *PRND* gene regions and the intergenic region. Tammar wallaby sequence lacked the *PRND* region. Some of the intergenic sequence information between *PRNP* and *PRND* is also missing in cow and sheep. The Y-axis represents the percentage identity using a window size of 10bp. The gene annotation information corresponds to human sequence as specified in the input annotation file.

7.4.2.3.3 Region file

This output file contains details of those regions which satisfied the user specified length and percentage cutoffs (Figure 7-7).



```
  Criteria: 90% identity over 10 bp

  *************** Conserved Regions - cow (mouse) ***************
    1        2     3       4       5      6       7    8    9        10
   35      (62)   to      43     (70)    =      9bp  at 100.0%  noncoding
   56     (749)   to      66    (759)    =     11bp  at  90.9%  noncoding
  165     (919)   to     180    (934)    =     16bp  at 100.0%  noncoding
  249    (1003)   to     262   (1015)    =     14bp  at  85.7%  noncoding
  290    (1066)   to     307   (1083)    =     18bp  at  88.9%  noncoding
  332    (1108)   to     340   (1116)    =      9bp  at 100.0%  noncoding
  470    (1199)   to     488   (1217)    =     19bp  at  89.5%  noncoding
  501    (1230)   to     510   (1239)    =     10bp  at  90.0%  noncoding
  514    (1243)   to     525   (1254)    =     12bp  at  91.7%  noncoding
  636    (1376)   to     645   (1384)    =     10bp  at  90.0%  noncoding
  734    (1432)   to     745   (1443)    =     12bp  at  91.7%  noncoding
  802    (1507)   to     811   (1516)    =     10bp  at  90.0%  noncoding
  822    (1526)   to     832   (1536)    =     11bp  at  90.9%  noncoding
 1341    (1633)   to    1349   (1641)    =      9bp  at 100.0%  noncoding
 1487    (1736)   to    1496   (1745)    =     10bp  at  90.0%  noncoding
 1692    (1852)   to    1701   (1861)    =     10bp  at 100.0%  noncoding
 1921    (1883)   to    1937   (1899)    =     17bp  at  94.1%  noncoding
 1954    (1917)   to    1964   (1927)    =     11bp  at  90.9%  noncoding
 1981    (1953)   to    1999   (1971)    =     19bp  at  94.7%  noncoding
 2060    (2024)   to    2069   (2033)    =     10bp  at  90.0%  UTR
```

Figure 7-7 VISTA region file. The first line shows the cutoffs used for the analysis. The second line contains the details of the sequences in the pairwise alignment. The rest of the report contains the annotation information (column numbers labeled in red font). Column 1 shows the start number of the region which satisfied the user specified criteria. Column 2 consists of the start number of the corresponding match in the second sequence. Column 4 and 5 corresponds to the end number of the first and second sequence respectively. Column 7 indicates the size of the matched region with the percentage identity in column 9. If the gene annotation file is supplied, it reports the region in column 10 to which the match corresponds. Query sequences values are shown within brackets.

This alignment annotation file was processed using a Perl script “extractseq.pl”. Based on the start and end numbers, the subsequences were extracted using the EMBOSS application “extractseq” integrated in “extractseq.pl”. This process was repeated for all the region files obtained for the various combinations of alignments. The Perl script finally generates a multiple FASTA file of all the conserved subsequences (Figure 7-8) which are stored in a conserved sequence database. The identifier of each sequence stored contains the information about the pair involved in the alignment, the position of the conserved sequence in both the sequence and the region to which it belongs. This enables the exact position of the conserved sequence to be tracked for further analysis. For those motifs which are shorter than 15 bp, continuous stretches of five “N” were added to both the 5’ and 3’ ends of the motif to facilitate the TFBS analysis.

```
>cowPrPDpl_align_cowPrPDpl_469_527_vs_humPrPDpl_3_61 region = noncoding
ACAATTCATGGGCATAATAAAATGGTGGTTTCTTTAAACCATTAAGTTTTGGAGTAGTT
>humPrPDpl_align_cowPrPDpl_469_527_vs_humPrPDpl_3_61 region = noncoding
ACAATCCATTGGCATAATAAAATGGTAGTTGTTTTAAACCACCTAAGTTGTGGGGTATT
>cowPrPDpl_align_cowPrPDpl_541_562_vs_humPrPDpl_76_97 region = noncoding
AATAGCCAGAATAGGACAAAAG
>humPrPDpl_align_cowPrPDpl_541_562_vs_humPrPDpl_76_97 region = noncoding
AATAACCAGAATAGGTCATAAG
>cowPrPDpl_align_cowPrPDpl_572_582_vs_humPrPDpl_107_117 region = noncoding
NNNNNTTTCGTTCCCTNNNNN
>humPrPDpl_align_cowPrPDpl_572_582_vs_humPrPDpl_107_117 region = noncoding
NNNNNTTTGGTTCCCTNNNNN
>cowPrPDpl_align_cowPrPDpl_587_596_vs_humPrPDpl_126_136 region = noncoding
NNNNNCCCTCACGAANNNNN
>humPrPDpl_align_cowPrPDpl_587_596_vs_humPrPDpl_126_136 region = noncoding
NNNNNCCCTCCAAGAANNNNN
```

Figure 7-8 Final output of “extractseq.pl” in multiple FASTA format. The identifier contains all the details of the sequence motif ; its position in relation to the main sequence, the sequence to which it is aligned together with its subsequence position, and the region to which it corresponds to such as non-coding, UTR, coding. These sequences occur in pairs corresponding to the region of the alignment which satisfies the user cutoffs of percent identity and size.

7.4.2.4 Drawbacks of alignment method

In a study conducted by Chapman et al (2004) it was shown that although some pairwise alignments perform well, success largely depends on species selected for pairwise comparison. The other factor is that the rate of sequence divergence between two species is not uniform across the genome (Li and Miller 2002). The use of multiple sequences was shown to be far superior in predicting functionally conserved regions in comparison with pairwise sequence alignment, by increasing the signal-to-noise ratio (Chapman et al. 2004). A multiple sequence alignment version of LAGAN called Multi-LAGAN (Brudno et al. 2003) which performs a progressive multiple alignment is also available and was trialled. However, theoretical considerations suggest the alignment approach may not be suited to the *PRNP* and *PRND* gene problem.

DNA has potential to undergo various rearrangement events, such as translocations (a subsegment is removed and inserted in a different location but the same orientation), inversions (a subsegment of DNA is removed from the sequence and then inserted back in the same location but in opposite orientation), duplications (a copy of a subsegment is inserted into the sequence, the original subsegment is unchanged), or a combination of the above. Global alignment algorithms are suitable with the assumption that the highly similar regions in the sequences appear in the same order and orientation. These are useful when the comparison is made within related organisms where the order and orientation is conserved across sufficient small regions. As the dataset is from different lineages, further work with multiple global alignments was not pursued.

7.4.2.5 Phylogenetic footprinting using motif-discovery approach

FootPrinter (Blanchette and Tompa 2003) uses the motif-discovery approach in identifying conserved motifs in a collection of homologous sequences (see section 2.1.2.2). FootPrinter takes as input a set of orthologous sequences and a phylogenetic tree relating the sequences used (Figure 7-9). It then reports the motifs based on the user-defined motif size and maximum number of mutations, which is represented by parsimony score. The option subregion_size divides the input sequence into subregions of defined size. This helps in eliminating those motifs whose locations vary too much. The following are some of the other options, which were used to refine the search.

Position change cost: Is the cost for a motif to change its subregion position.

Maximum number of mutations per branch: This option allows a fixed number of mutations per branch of the tree.

Triplet filtering and post filtering: This filtering eliminates those motifs that do not have a good pair of matching motifs in the other input sequence compared. This mainly reduces the memory used by the program.

Sequence type: For upstream sequence type, the 3’ ends of the sequence are assumed to be aligned and for downstream, the 5’ ends are assumed to be aligned.

Inversion cost: Is the cost for a motif to undergo inversion.

A locally installed version (FootPrinter version 2.1) was used for the analysis as it gives more options compared with the web version and also simplifies the processing of output files. FootPrinter generates a number of output files with different file formats (Figure 7-9). For programmatic processing, only html output format was used; this contains the information about the motifs which satisfied the user criteria/cutoffs.

7.4.2.5.1 Search strategy

Regulatory elements that have been acquired very recently in evolution may not be easily detectable through PF if the comparisons are made across diverse lineages. Hence, the analysis was divided into different categories. Importance was given to understanding mammalian-specific regulation; this was achieved by comparing sequences within the mammalian species and with other lineages. PF using multiple mammalian sequences has been shown to have better resolution at individual TFBSs compared with pairwise alignments (Cooper et al. 2003).

- Category 1: Mammal
- Category 2: Mammal+marsupial
- Category 3: Mammal+marsupial+amphibian
- Category 4: Mammal+marsupial+avian
- Category 5: Mammal+marsupial+avian+amphibian

Category 4 and 5 are not applicable for phylogenetic footprinting analysis related to *PRND* as chicken lacks *PRND*.

The sequences were not divided across the TSS into upstream and downstream sequence region, as the idea was to include the sequence as a whole in single analysis rather than performing two independent motif searches on upstream and downstream sequence regions. This would enable us to identification of those motifs which have undergone rearrangement across the TSS.

The output motif file (motif.html) contains the information about the motifs and their positions. A comprehensive search was performed using different FootPrinter options (Table 7-1) (subregion- 1000 to 3000bp; motif size- 6 to 10bp; parsimony score- 0 to 2). Using a Perl script "motifextract.pl", the "motif.html" output file was converted to a single multi-FASTA file. Each analysis was performed twice using upstream and downstream (FootPrinter: sequence type) option on the same input sequence. The multi-FASTA files from both analyses were combined using a Perl script, "compileTFBS.pl" to produce a non-redundant single multi-FASTA file (similar to Figure 7-8). These multi-FASTA files relating to different subregion sizes were stored in a conserved sequence database. Each sequence-motif position was registered in the sequence identifier

Table 7-1 The options used to run FootPrinter and the value range used for the analysis.

Option	Value range
subregion_size	2000-3000
position_change_cost	2
Size	8-10
max_mutations	0-1
max_mutations_per_branch	1
triplet_filtering	N/A
post_filtering	N/A
inversion_cost	1
sequence_type	upstream and downstream

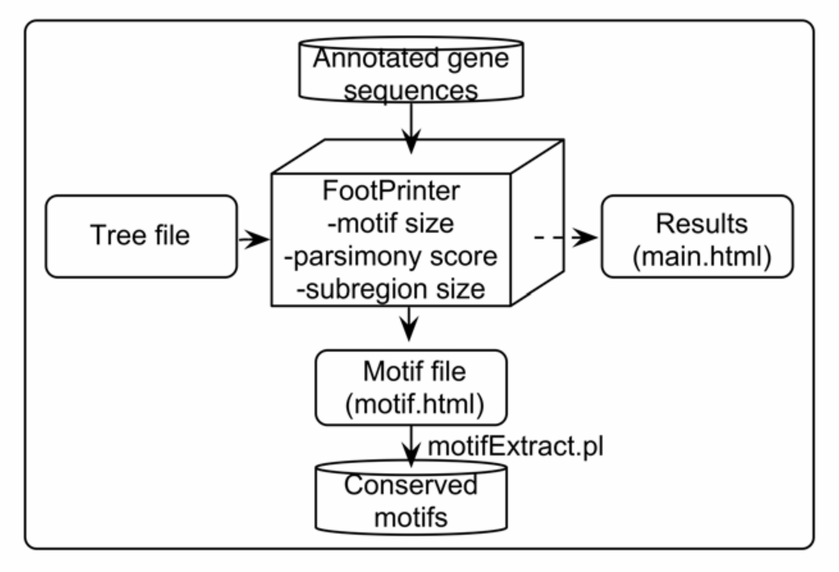


Figure 7-9 Flow diagram summarizing the steps in PF using FootPrinter. The end result of the analysis is the conserved sequence motif database.

.

7.4.3 Searching against the TRANSFAC database for TF-binding specificity

The TRANSFAC database of eukaryotic transcriptional regulation comprises data on TFs, their target genes and regulatory binding sites (Matys et al. 2003). TRANSFAC is available as a commercial version and a public version with the commercial version having more data. To enable a comprehensive analysis, the commercial version, TRANSFAC professional (version 9.2), was used. Match is a tool which uses the weight matrices in the TRANSFAC database for searching for putative TFBSs (Kel et al. 2003). The advanced version, Match professional (distributed with TRANSFAC professional), was used.

7.4.3.1 Search strategy

7.4.3.1.1 Optimization of Match search parameters

Match provides several pre-defined optimized profiles. A profile is a selected subset of matrices including default user-defined cut-off values designed for searches. Match takes DNA sequence and profile as input, scores it against the PWMs based on the profile information, and outputs a list of potential sites. The scores that are calculated are the core similarity score and the matrix similarity score. These scores measure the quality of a match between the sequence and the matrix with values ranging from 0.0 to 1.0 where 1.0 denotes the exact match. The core of each matrix is the first five most conserved consecutive positions of a matrix and the matrix similarity score is the score calculated for all the positions of the matrix. In order to find putative TFBSs, choosing the appropriate cutoffs for core and matrix similarity plays a central role. To address this task, three different pre-calculated cutoffs for each matrix is provided with TRANSFAC license

- Cut-offs minimizing false negative rate (minFN)
- Cut-offs minimizing false positive rate (minFP)
- Cut-offs minimizing the sum of both errors (minSum)

Table 7-2 Details of the test data used to define optimal search parameters for the Match tool.

TF	Binding site	Literature
ETS	ggtttcctccggggt	(Chapman et al. 2003)
GATA	tccttatcaggcgc	(Chapman et al. 2003)
Oct1	tgcatatt	(Premzl et al. 2005)
NFAT	attttcca	(Premzl et al. 2005)

Published TFBS information was used to optimize the Match search parameters, i.e. to predict maximum true positives and minimum false positives against the TRANSFAC professional database. From the literature, the binding information for ETS, GATA, Oct1, and NFAT (Table 7-2) was used to define the optimal search parameters for the Match tool. The test either showed too few or too many hits for most of the profiles but the results obtained for minSUM profile was well balanced and predicted the appropriate TF-binding to the test data. Match can report more than one TF-binding in a given position but can be made to return unique best hits by using the options -b and -u. Match can take a multiple sequence FASTA file as input and writes to the user defined output file. For each of the sequences in the input file, it gives a search report (Figure 7-11) with the first line showing the sequence identifier which corresponds to the sequence identifier in the multiple FASTA sequence file. This is followed by a five-column search result if a TFBS is found; otherwise, it is left blank. The columns are the TRANSFAC identifier of the matching matrix, position and strand where the match was found followed by core similarity and matrix similarity score, with the last column indicating the matching sequence where the core match is in capital letters. As the input file contains multiple sequences, the TF-binding information for all of this is written to a single output file. The matrix library from the TRANSFAC database includes matrices from vertebrates, plants, insects, fungi, nematodes and bacteria which have identifiers which begin with V$, P$, I$, F$, N$, B$, respectively. Using Perl scripts, “motifExtract.pl” or “extractSeq.pl”, all sequences which did not show any binding affinity were eliminated and only those sequences which showed binding to the vertebrate TFs were retained.

A systematic pipeline was developed to assess the specificity of TF binding to the conserved-sequence motifs identified by PF (Figure 7-10). The steps of the analysis were:

- Starting inputs were the motifs identified by AVID/LAGAN/FootPrinter methods.
- These motifs were scored against the TRANSFAC database using Match, which uses the information defined in the profile.
- The output file generated by Match was processed to eliminate entries for motif sequences which did not correlate with any known binding affinity; only sequences showing putative binding to the vertebrate TFs were retained.

- The Perl scripts, “motifExtract.pl” and “extractSeq.pl” contain modules that process the Match output file.
- The final output (same format as Match output) generated by these Perl scripts was stored in the TFBS database (Figure 7-11).
- When conserved motifs were obtained by non-stringent criteria, e.g. for parsimony score value > 0 for FootPrinter or percent identity value < 100% for alignment methods, it is possible that TFs predicted to bind to the same set of conserved motifs in different input sequences could differ (Figure 7-10). Such predicted TFs were eliminated. This criterion was implemented by two Perl scripts, “tfbsCons.pl” and “ultraTFBS.pl” which need to be run consecutively.
- Altogether, the resultant predicted motifs were classified as either highly conserved or less highly conserved. Both sets were stored in the TFBS database.

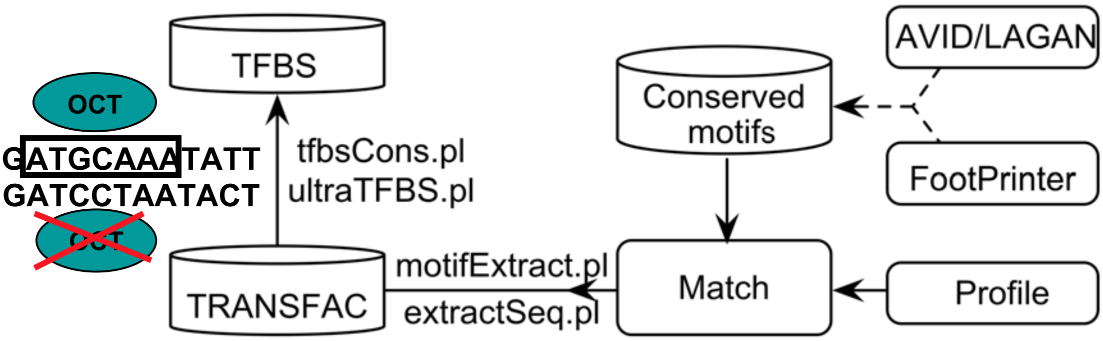
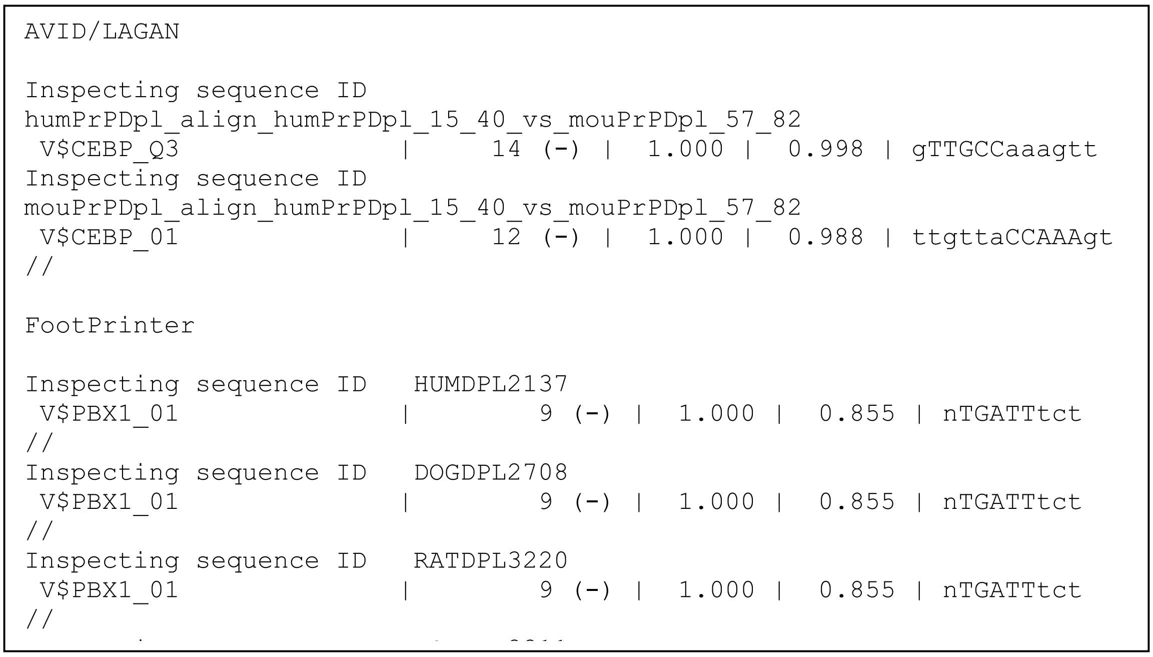


Figure 7-10 Flow diagram showing the steps in the TFBS analysis. Dotted arrows indicate processes outside the TFBS analysis. Motifs identified by non-stringent criteria (e.g. 2 mismatches among 7 bases in the core region) can result in the predicted TF binding only to one of the sequences compared and not to the others. Such motifs were not stored in the TFBS database.



```
AVID/LAGAN

Inspecting sequence ID
humPrPDpl_align_humPrPDpl_15_40_vs_mouPrPDpl_57_82
 V$CEBP_Q3              |     14 (-) |  1.000 |  0.998 | gTTGCCaaagtt
Inspecting sequence ID
mouPrPDpl_align_humPrPDpl_15_40_vs_mouPrPDpl_57_82
 V$CEBP_01              |     12 (-) |  1.000 |  0.988 | ttgttaCCAAAgt
//

FootPrinter

Inspecting sequence ID   HUMDPL2137
 V$PBX1_01              |        9 (-) |  1.000 |  0.855 | nTGATTtct
//
Inspecting sequence ID   DOGDPL2708
 V$PBX1_01              |        9 (-) |  1.000 |  0.855 | nTGATTtct
//
Inspecting sequence ID   RATDPL3220
 V$PBX1_01              |        9 (-) |  1.000 |  0.855 | nTGATTtct
//
```

Figure 7-11 Final output of the TFBS analysis for conserved sequence motifs obtained from AVID/LAGAN (top) and FootPrinter (bottom).

7.5 Visual front end for data analysis

TFBSs specific for a particular gene occur in combinations of order, distance and strand orientation. Analyzing this organization is essential for understanding transcriptional regulation. A visual front end is necessary to make this process of viewing TFBSs intuitive and easy enough to facilitate proper judgment of the results. To achieve this, an interactive user interface, FactorScan was designed to give functional access to the results obtained from the PF and TFBS analysis.

7.5.1 Interface development

FactorScan is a web-based application accessible through a web browser. It links the TFBS information, conserved-sequence motif information predicted by AVID/LAGAN/FootPrinter and the TRANSFAC database (Figure 7-12). All the databases are in flat-file format. This interface enables access to the data (conserved-sequence motifs and TFBS) generated by the various pipelines (Figure

7-7, Figure 7-9, Figure 7-10): it is not dynamically generated during the visualization process. The web interface has three main components, the web form, the results page and the report page.

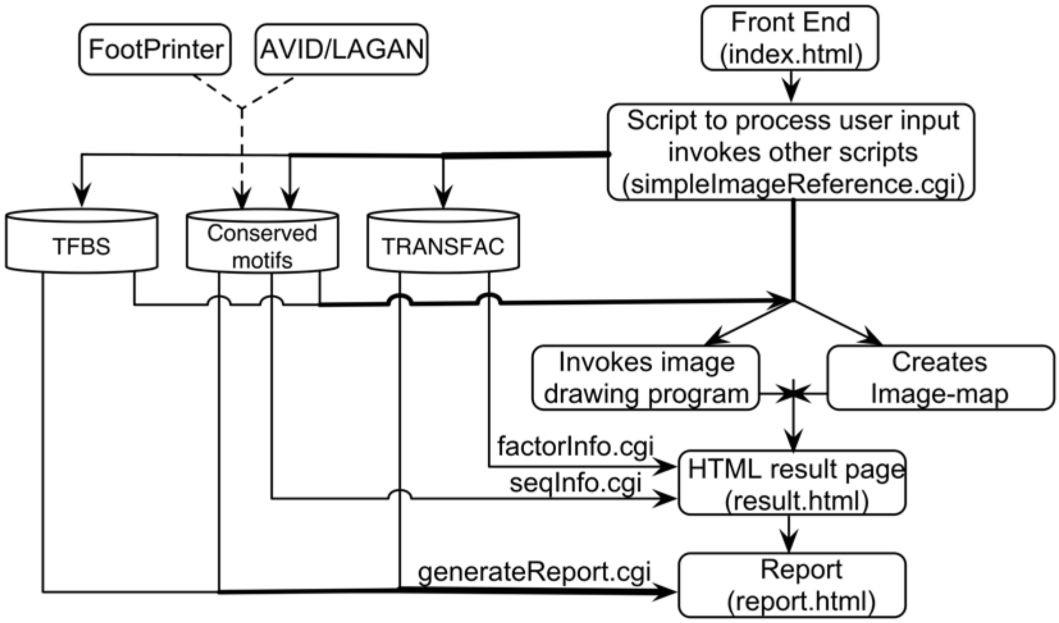


Figure 7-12 The main functional flow of information from submitting the web form to the display of results. The programs involved with each of the tasks are indicated.

7.5.1.1 Web form

Input: The user input for the web form is categorized into mandatory and optional parameters (Figure 7-13). The mandatory parameters include the gene for which the results are to be displayed and the various options used for PF to generate the data (subregion size, sequence type, sequence dataset). The optional parameters are for customizing and controlling the display of the results. Some important features are (i) Transcription Factor Search, (ii) Core Similarity Score, (iii) Title, (iv) Tissue Source and (v) Line. The Transcription Factor Search is useful to display a subset of TFs of particular interest, either individually or in combinations. The latter is particularly useful for identifying and comparing TFBS ‘modules’ (clusters of

TFBS in a defined order) (Wasserman and Sandelin 2004). The Core Similarity Score can be used to visualize TFs which satisfy criteria set by the user. This value is in the range of 0-1; by default, this is set to 1 to display the statistically most significant hits. The “Title” option can be used to visualize the name of the TF matrices for the displayed TFs. Tissue-specific TFs can be searched according to tissue, such as brain and testis. The cell-positive and cell-negative information in the TRANSFAC database is used for this purpose. The conserved-sequence motif distribution can be viewed by selecting the “Line” option. The parameters specific for alignment-based and motif-discovery based method are listed below

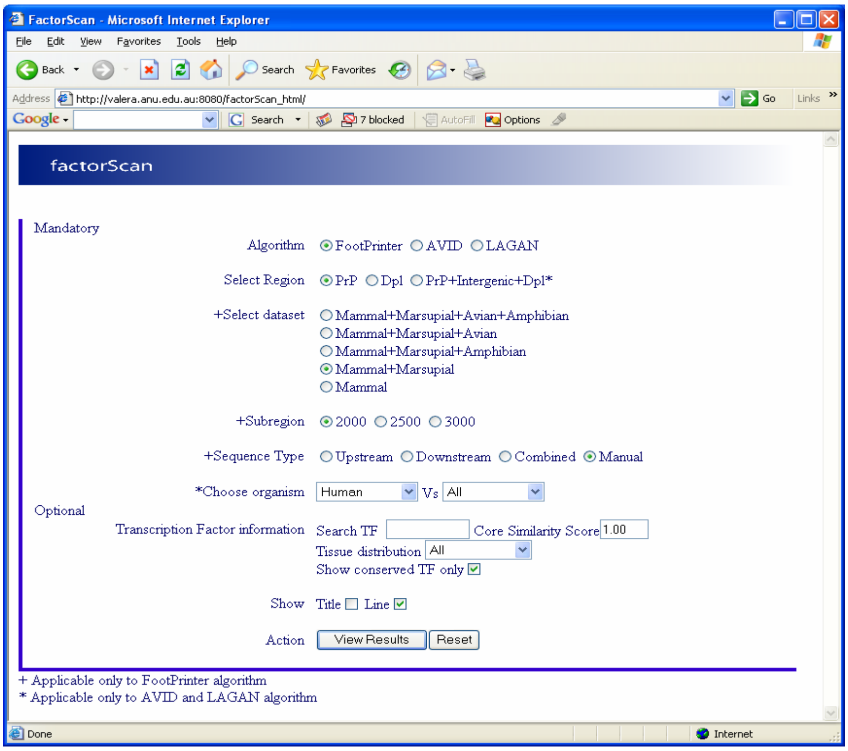
7.5.1.1.1 Options specific for alignment method

PF in the genomic region relating to *PRNP*, *PRND* and the intergenic region were performed only with alignment-based methods and, hence, the parameter “*PRNP*+Intergenic+*PRND*” is valid only if either AVID/LAGAN algorithm is selected. Also, the results obtained from alignment methods are based on pairwise sequence comparison. The option “Choose organism” is for selecting the alignment pair for viewing. Selection can be made between a pair of species or between one species with the rest of its pairwise combinations.

7.5.1.1.2 Options specific for FootPrinter

FootPrinter analysis was performed using different dataset which includes intra-mammalian species comparison and that between the mammalian and other lineages. The parameter “Select dataset” gives the combinations as analyzed in section 7.4.2.5.1 (categories 1-5) for user selection.

The options used with FootPrinter include subregion and sequence_type. The user can view the results obtained using these combinations by selecting from the “Subregion” parameter which has values 2000, 2500 and 3000 and the “Sequence Type” parameter which has values upstream, downstream and combined (non-redundant compilation of upstream and downstream results).


FactorScan - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://valera.anu.edu.au:8080/factorScan_html/
Go Links
Google Search 7 blocked AutoFill Options
Done
Internet

factorScan

Mandatory

Algorithm ● FootPrinter ○ AVID ○ LAGAN

Select Region ● PrP ○ Dpl ○ PrP+Intergenic+Dpl*

+Select dataset ○ Mammal+Marsupial+Avian+Amphibian
○ Mammal+Marsupial+Avian
○ Mammal+Marsupial+Amphibian
● Mammal+Marsupial
○ Mammal

+Subregion ● 2000 ○ 2500 ○ 3000

+Sequence Type ○ Upstream ○ Downstream ○ Combined ● Manual

*Choose organism Human Vs All

Optional

Transcription Factor information Search TF Core Similarity Score 1.00
Tissue distribution All
Show conserved TF only ☑

Show Title □ Line ☑

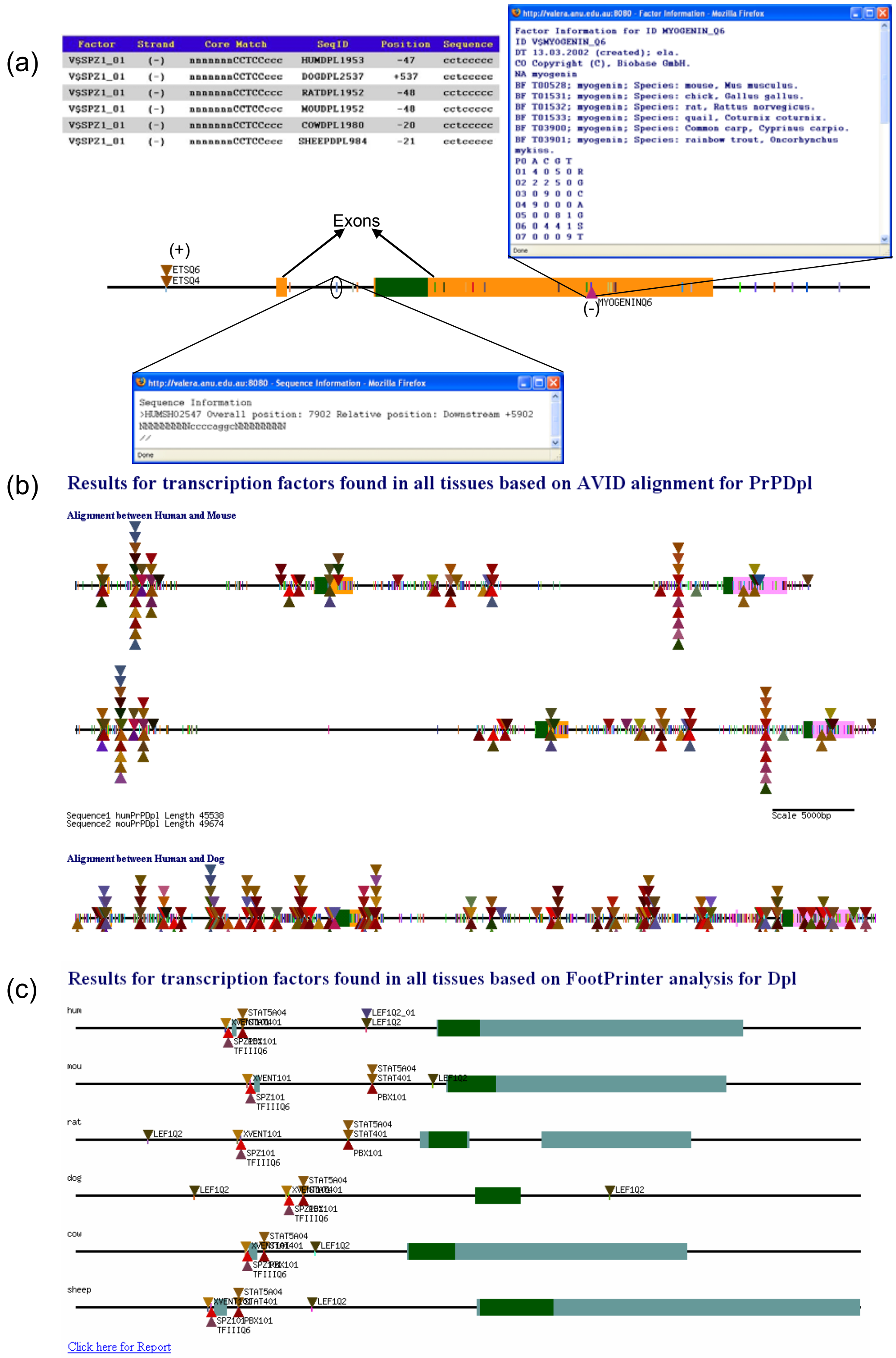
Action View Results Reset

+ Applicable only to FootPrinter algorithm
* Applicable only to AVID and LAGAN algorithm

Figure 7-13 Webform for FactorScan where the user can submit the information for viewing the results.

7.5.1.2 Results page

Output: The submitted web form is processed by a CGI script “simpleImageReference.cgi” (Figure 7-12) and the results are displayed in the same window. The results page displays a schematic of relative organization of gene annotation, TF and conserved sequence motif information. Genomic sequence is represented, conventionally, as a horizontal line with exons mapped on as rectangular boxes, and with coding and non-coding regions of exons shaded in different colors (Figure 7-14 (a)). The TFs predicted to bind are represented as triangles, inverted and upright for the forward and reverse strands, respectively (Figure 7-14 (a)). Each TF is assigned a unique color; its name is displayed if the “Title” option is selected. The conserved-sequence motifs, identified by any of the methods, are represented as vertical bars (Figure 7-14 (a)); use of color is particularly helpful to discriminate these regions when they are very close to each other. Triangles representing TFs and vertical bars representing conserved-sequence motifs are clickable areas. Clicking on the triangles invokes Perl script “factorInfo.pl” for displaying a summary of TF information, which is obtained from the TRANSFAC database. Clicking on the vertical bar invokes the Perl script “seqInfo.pl”, which displays information about the conserved sequence motif by accessing the information from the conserved sequence motif database. This is particularly useful as the conserved-sequence motif can be examined for other purposes. For FootPrinter analyses, the schematic are drawn to scale within a species, but between species the scale is not normalized (Figure 7-14 (c)). For pairwise-alignment analyses, the scale (also shown; see Figure 7-14 (b)) is normalized between the pairs, and the results can be displayed either between specific pairs or for one against all others. The latter is useful to compare the conserved TFBS distribution among various lineages. Information about species, abbreviations used and the sequence length in base pairs is provided in table form at the bottom of the schematic. The results page also has a link to view the report of the TFs and their binding sites. Clicking this link invokes a Perl script “generateReport.pl” which pops up a window (enabled by Java Script) displaying the summary of the TFs, the strand to which it binds, core match, the conserved-sequence motif identifier, the position of the TF relative to the TSS and the sequence which was used for TFBS analysis. The output is in a tabular format and includes the date and time stamp and a link to print the report (Figure 7-14 (c)).



Factor	Strand	Core Match	SeqID	Position	Sequence
V$SPZ1_01	(-)	nnnnnnnCCTCCccc	HUMDPL1953	-47	cctccccc
V$SPZ1_01	(-)	nnnnnnnCCTCCccc	DOGDPL2537	+537	cctccccc
V$SPZ1_01	(-)	nnnnnnnCCTCCccc	RATDPL1952	-48	cctccccc
V$SPZ1_01	(-)	nnnnnnnCCTCCccc	MOUDPL1952	-48	cctccccc
V$SPZ1_01	(-)	nnnnnnnCCTCCccc	COWDPL1980	-20	cctccccc
V$SPZ1_01	(-)	nnnnnnnCCTCCccc	SHEEPDPL984	-21	cctccccc

http://valera.anu.edu.au:8080 - Factor Information - Mozilla Firefox

```
Factor Information for ID MYOGENIN_Q6
ID V$MYOGENIN_Q6
DT 13.03.2002 (created); ela.
CO Copyright (C), Biobase GmbH.
NA myogenin
BF T00528; myogenin; Species: mouse, Mus musculus.
BF T01531; myogenin; Species: chick, Gallus gallus.
BF T01532; myogenin; Species: rat, Rattus norvegicus.
BF T01533; myogenin; Species: quail, Coturnix coturnix.
BF T03900; myogenin; Species: Common carp, Cyprinus carpio.
BF T03901; myogenin; Species: rainbow trout, Oncorhynchus
mykiss.
PO A C G T
01 4 0 5 0 R
02 2 2 5 0 G
03 0 9 0 0 C
04 9 0 0 0 A
05 0 0 8 1 G
06 0 4 4 1 S
07 0 0 0 9 T
```

Done

http://valera.anu.edu.au:8080 - Sequence Information - Mozilla Firefox

```
Sequence Information
>HUMSH02547 Overall position: 7902 Relative position: Downstream +5902
NNNNNNNNNccccaggcNNNNNNNNN
//
```

Done

Results for transcription factors found in all tissues based on AVID alignment for PrPDpl

Alignment between Human and Mouse

```
Sequence1 humPrPDpl Length 45538
Sequence2 mouPrPDpl Length 49674
```

Alignment between Human and Dog

Results for transcription factors found in all tissues based on FootPrinter analysis for Dpl

Click here for Report

Figure 7-14 Display of results. (a) Features of the results page. (b) Display of results page for alignment-based method and (c) FootPrinter method. Note the differences in the display pattern between (b) and (c), the titles of the TFs are seen in (c).

7.6 Results from application of the pipeline to the PrP family genes

The TF-binding predictions obtained from the alignment method did not show consistent conservation patterns when the second sequence in the pair being compared was changed. This made judgment of the significance of the predicted TFBSs difficult. FootPrinter, which uses multiple sequences, reports conserved regions found in all the input sequences. Hence, efforts were focused on analyzing the results obtained from the FootPrinter analysis, this is discussed below.

7.6.1 Selection criteria

The FootPrinter results obtained by altering the values for the subregion option were manually compiled (for mammal+marsupial and mammal comparisons) by elimination based on the following:

1) Overrepresented TFs for example TBP in “AT” and STAT in “TC” rich/repeat region.
2) Variation in the TF position between 3’ UTR in one species and intron in the other.
3) Variation in the occurrence of several of consecutive copies of the same TFBS in different species. In such cases, only the minimum number of copies found across all species compared is retained, based on their relative position with adjacent TFs.
4) Neighboring TF positions were used to eliminate TFs which show large variation in position.

The TFBS predictions based on FootPrinter results (FootPrinter search criteria; motif size 8 and parsimony score 0) are discussed below for *PRNP* and *PRND*.

7.6.2 Prion protein gene, *PRNP*

Based on the PF and TF-binding predictions, the TFs predicted by intra-mammalian species comparison and that from comparison between mammalian and other vertebrate species (Table 7-3 and Figure 7-15, Figure 7-16) are discussed below.

7.6.2.1 Mammal+Marsupial+Avian+Amphibian

The comparison made between all the four lineages resulted in very few motifs. With subregion value of 3000, only motif “gggagggg” which binds to SPZ1/MAZ/MZF (Figure 7-16 (a)) can be considered significant based on the selection criteria (section 7.6.1).

7.6.2.2 Mammal+Marsupial+Amphibian

Using a subregion value of 2000, sequence motifs binding to MEF2, CDXA/TBP, and FAC1 were predicted. Increasing the subregion value to 3000 produced another additional motif binding to MYB (Figure 7-16 (b)).

7.6.2.3 Mammal+Marsupial+Avian

Using a subregion value of 2000, only one TFBS binding to MAZ/SPZ1/MZF was predicted. With a subregion value of 3000, the additional TFBS YY1/STAT4 was found (Figure 7-16 (c)).

7.6.2.4 Mammal+Marsupial

The predictions made by comparing mammal with marsupial species are more reliable because the evolutionary distance separating them is relatively small. With a subregion value of 2000, MAF, XVENT1, CIZ, YY1 and STAT1 were predicted with good scores. Increasing the subregion value to 3000 added another TFBS corresponding to PAX2/MYB (Figure 7-16 (d)).

7.6.2.5 Mammal

A number of TFs were predicted to bind, with most of them predicted in the intron. The only TF predicted upstream to the TSS is E4BP4. These predictions were made with a subregion option of 2000. Most of the predicted TFs show conservation in position and relative order with adjacent predicted TFs (Figure 7-16 (e)).

7.6.3 Doppel gene, *PRND*

7.6.3.1 Mammal+marsupial+amphibian

The conserved sequence motifs predicted among these lineages did not produce any TFBS of statistically significant score. The only TFBS observed was TBP but it does not satisfy the selection criteria 2 (section 7.6.1).

7.6.3.2 Mammal+Marsupial

Using a subregion value of 2000, three significant conserved sites binding to LEF1, PBX1/STAT4/STAT5, and XVENT1 were predicted. XVENT1 in marsupial is predicted to be in the first exon in contrast to the upstream region in the mammal sequences (Figure 7-17 (a)).

7.6.3.3 Mammal

In addition to the above predicted TFs, two additional conserved motifs binding to TFs, SPZ1/TFIII and CDXA were predicted in the upstream and intronic region, respectively (Figure 7-17 (b)).

7.6.4 Applying this pipeline to *SPRN*

The pipeline tested on *PRNP* and *PRND* was tried on the *SPRN* gene. This analysis was performed by Dr Lorenzo Sangiorgio (Visiting Post Doc Fellow, University of Milan, Italy). *SPRN* is reported in several lineages (mammal, marsupial, avian, amphibian and fish). PF was performed in various combinations using three different fish species, zebrafish, *Fugu*, *Tetraodon*. The PF analysis was performed only with FootPrinter. Also, in addition to the combinations used for *PRNP* and *PRND* (Section 7.4.2.5.1), fish sequences were also used for the analysis. A number of TFs were predicted which are conserved in order and orientation. The results are not discussed here as this was done as a part of a separate project.

Table 7-3 Summary of (predicted) TFs, binding motifs, and position in various species. Bases downstream to the TSS begin with "+" and that upstream to the TSS begins with "-" for (a) *PRNP* and (b) *PRND*. Base underlined is either an overlapping motif or motif at a different location. Hum- Human, Mou- Mouse, Md- *M. domestica*, Tw- Tammar wallaby, Chick- Chicken, Xen- *Xenopus.** indicates not included in manual curation for mammal and marsupial results as these results are from subregion 3000.

(a) *PRNP*

TF	Motif	Hum	Dog	Rat	Mou	Cow	Sheep	Md	Tw	Chick	Xen
PAX2/MYB	TTCAGTTT	+3670	+6634	+6593	+6735	+7035	+7083				
TBP/GATA1	TTTATCAA	+2680	+7534	+2244	+2200	+2455	+2441				
HOXA4	AAAATTAG	+2168	+1773	+1725	+1674	+2071	+2057				
NF1	TAGCCAAG	+4673	+5106	+4314	+5429	+9339	+9367				
HMGIY	AGAAAATT	+2222 +3765	+782	+1723	+1672	+1354	+1350				
STAT1/HMGIY/ NFAT	ATTTTCCA	+5991	+4486	+4917	+4804	+6701	+6749				
DBP	CAGCAACC	+1808	+1339	+894	+875	+1669	+1665				
Ap3	TCTAAAAT	+2165	+1770	+4498	+2044 +4381	+2068 +4378	+2054 +4355				
MEF2	TAAAAATA	+3051	+3417	+2955	+2555	+2364	+2349				
OCT1	TGCATATT	+3247	+4419	+3701	+3503	+1702	+1698				
OCT1	CTTTGCAT	+15082	+16835	+15516	+28391 +28606	+20152 +20877	+20569 +21234				
PAX2/MYB	AAACTGAC	+2304	+1548	+1779	+1729	+1862	+1859				
PBX1	TGATTTGT	+5924	+6460	+8031	+21082	+11849	+11909				
E4BP4	TTATGTAA	-311	-396	-260	-266	-1109	-1103				
GEN_INI2	CTTCATTT	+12775	+10896	+10392	+24791	+15459	+15921				
ETS/STAT6	TTCTTCCT	+6264	+4012	+5540	+7162	+3920	+3898				
STAT1	CTTTCCTT	+10155	+9971	+8166	+6312 +7346	+8262, +11835	+11895	+6737,+11820 +11888	+9648		
MAF	TCTGACTC	+3986	+10507	+6023	+19279	+10746 +12324	+10783 +12759	+6679	+10042		
YY1	TTGCCATT	+4075	+6452	+7743	+7761	+4176	+4153	+4550	+6779		
XVENT1	ACAAATAT	+1429	+6350	+5191	+12513 +14836	+5656	+5694	+4338 +4825	+4509		
CIZ	TTTTTTCT	+7723	+7601	+8698	+19987	+12657	+11662	+11709	+12048		
*PAX2/MYB	TACAGTTT	+9953	+7154	+7510	+7522	+1765	+9324	+3517	+2821		
MEF2	TATTTTTA	+8044	+10683	+6048 +8586	+21639	+14029 +14708 +15864	+15171 +16324	+9415	+10453		+4797
CDXA/TBP	ATTTATTT	+4250	+6017	+5479	+5538	+8122	+8048	+4775	+6986		+3482
FAC1	TTTGTTGTT	+2697	+2295	+4148 +2267	+2629 +2224	+3788	+3766 +968	+3655,+6651 +6654	+7530,+6697 +6700		+5692
*MYB	TTACAGTT	+9952	+7153	+7509	+7521	+12827	+13262	+12488	+12548		+8256
*YY1/STAT4	TCCATTTC	+7730	+5354	+3542	+5645	+6499	+5475	+7474	+9020	+1955	
SPZ1/MAZ	GGGGAGGGG	+500	+204	-150	+2365 +2375	+899	+888	+1789 +2980	+610,+1722 +1816	+1450 +2407	+6333

(b) *PRND*

TF	Motif	Hum	Dog	Mou	Rat	Cow	Sheep	Md
SPZ1/ TFIII	CCTCCCCC	-47	+537	-48	-48	-20	-21	
CDXA	TTATTTAA	+386	+936	+377	+368	+413	+400	
LEF1	ACAAAGAA	+1726	-582, +4354	+1998	-1140	+759	+706	+3041
PBX1/ STAT4/ STAT5	TGATTTCT	+137	+708	+1311	+1220	+172	+172	+4584
XVENT1	TATTTGGA	-72	+512	-81	-82	-44	-45	+150

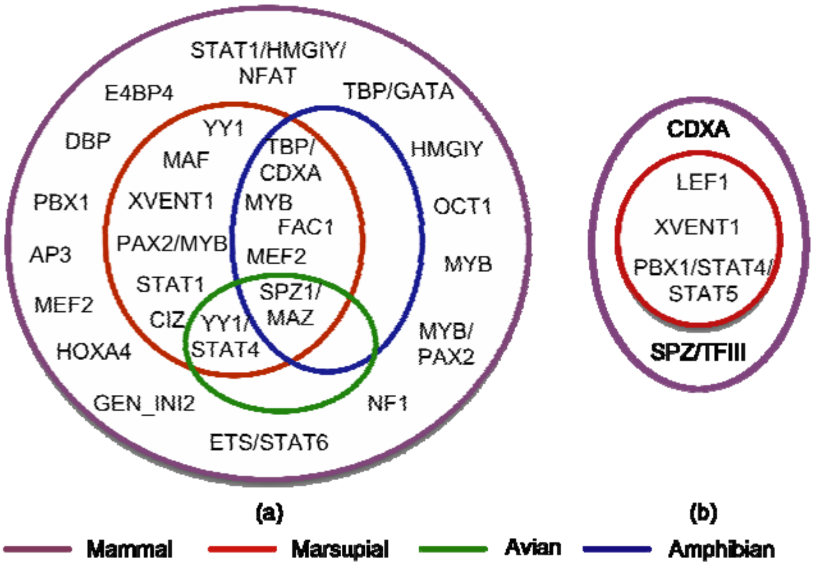
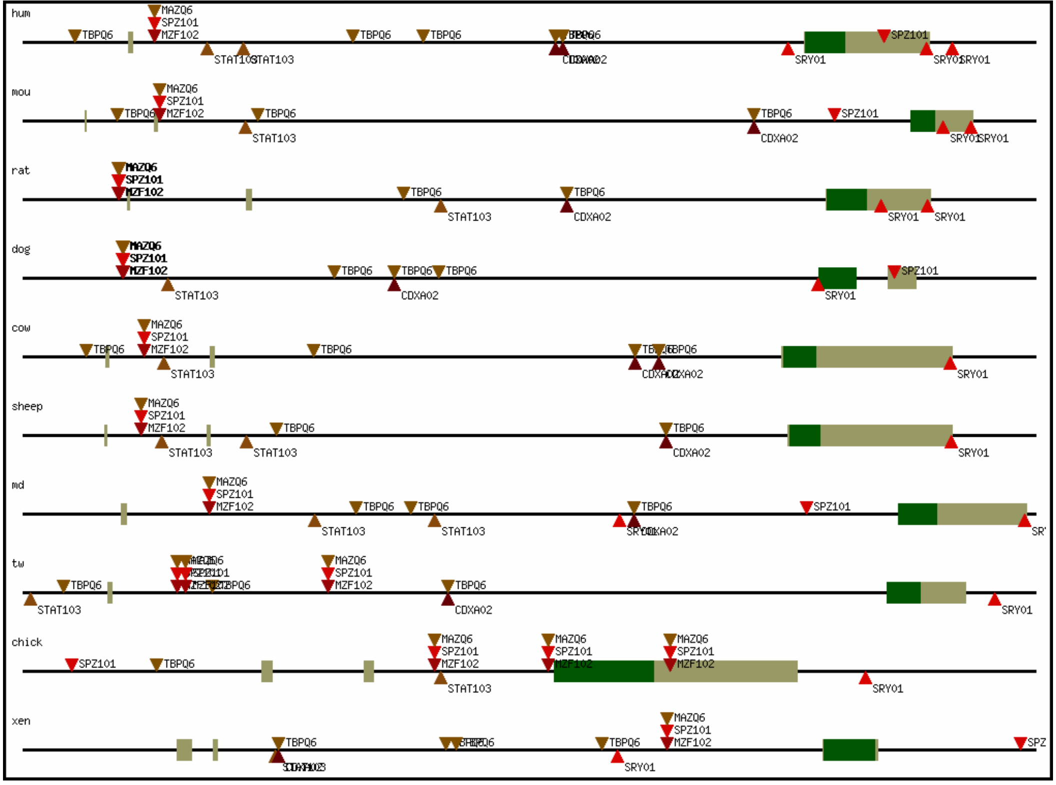
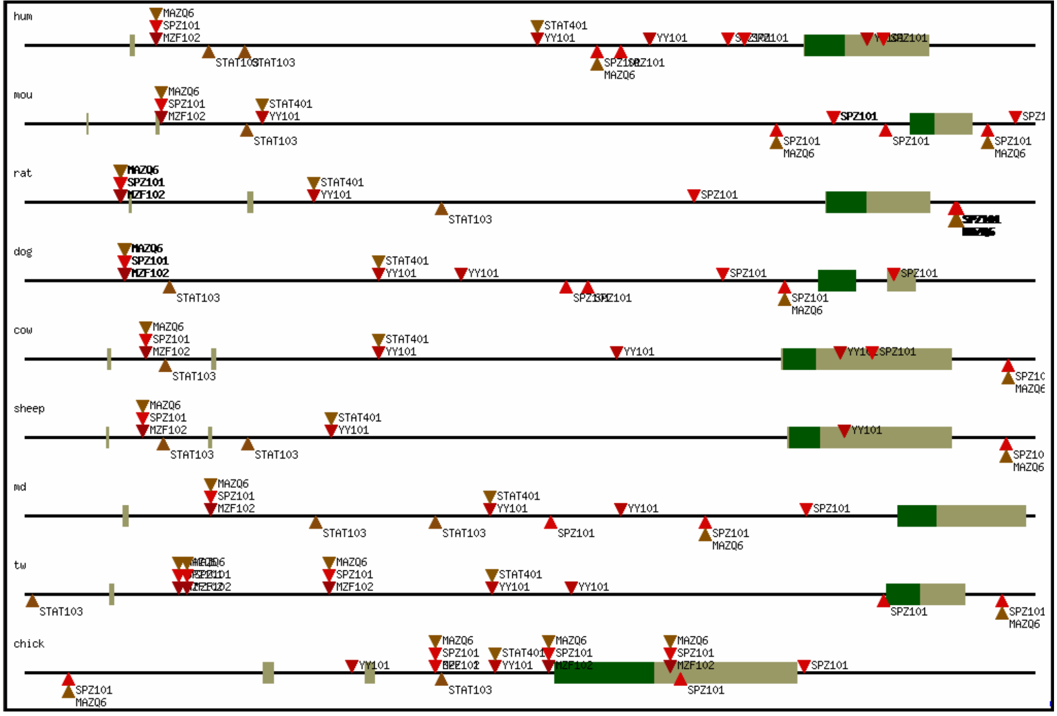


Figure 7-15 Venn diagram illustrating the TFs predicted by comparing different species for (a) *PRNP* and (b) *PRND*. For *PRND* the comparisons were only made between marsupial and mammals. As mammalian sequences were used for all the comparisons, it forms the outermost circle.

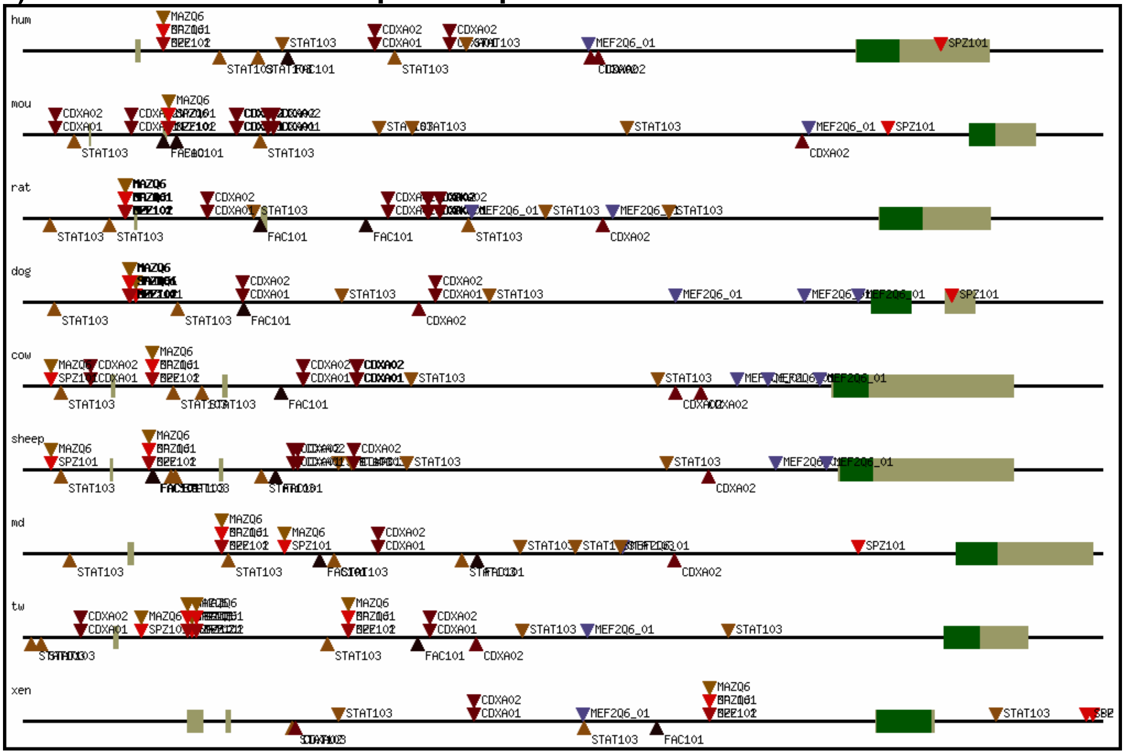
(a) *PRNP*: Mammal+Marsupial+Avian+Amphibian



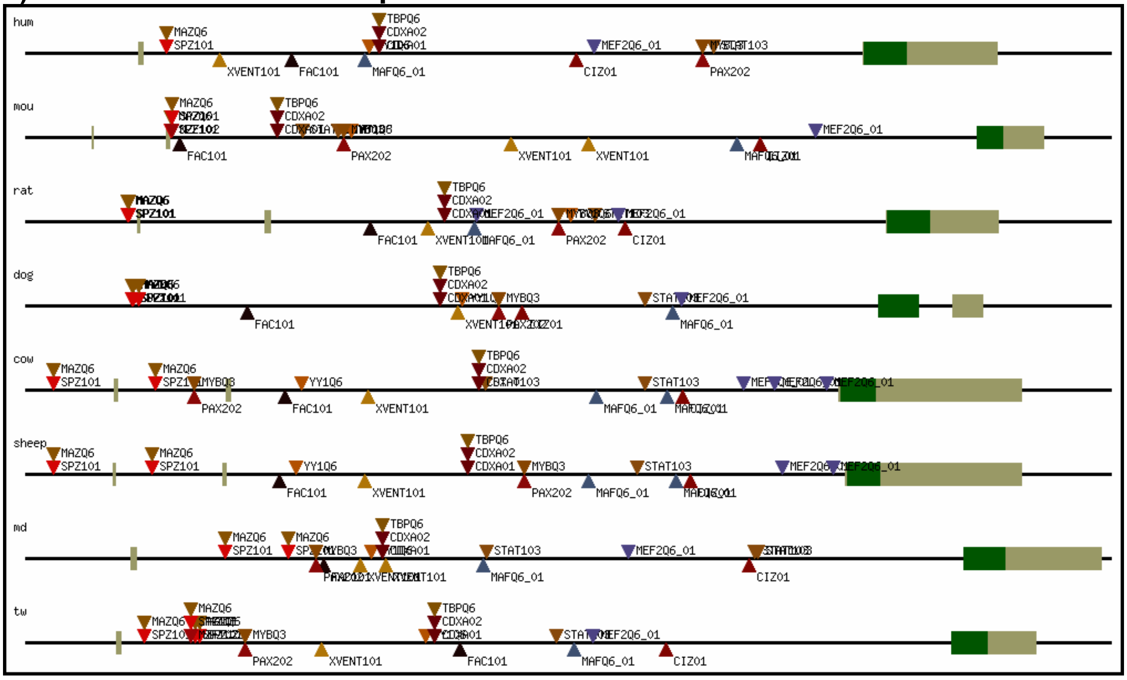
(b) *PRNP*: Mammal+Marsupial+Avian



(c) *PRNP*: Mammal+Marsupial+Amphibian



(d) *PRNP*: Mammal+Marsupial



(e) *PRNP*: Mammal

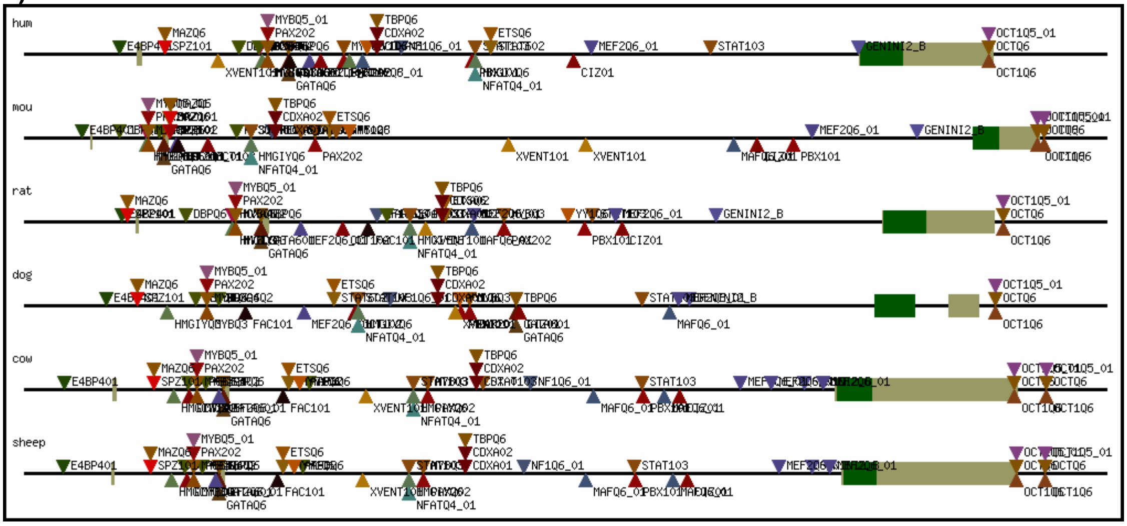
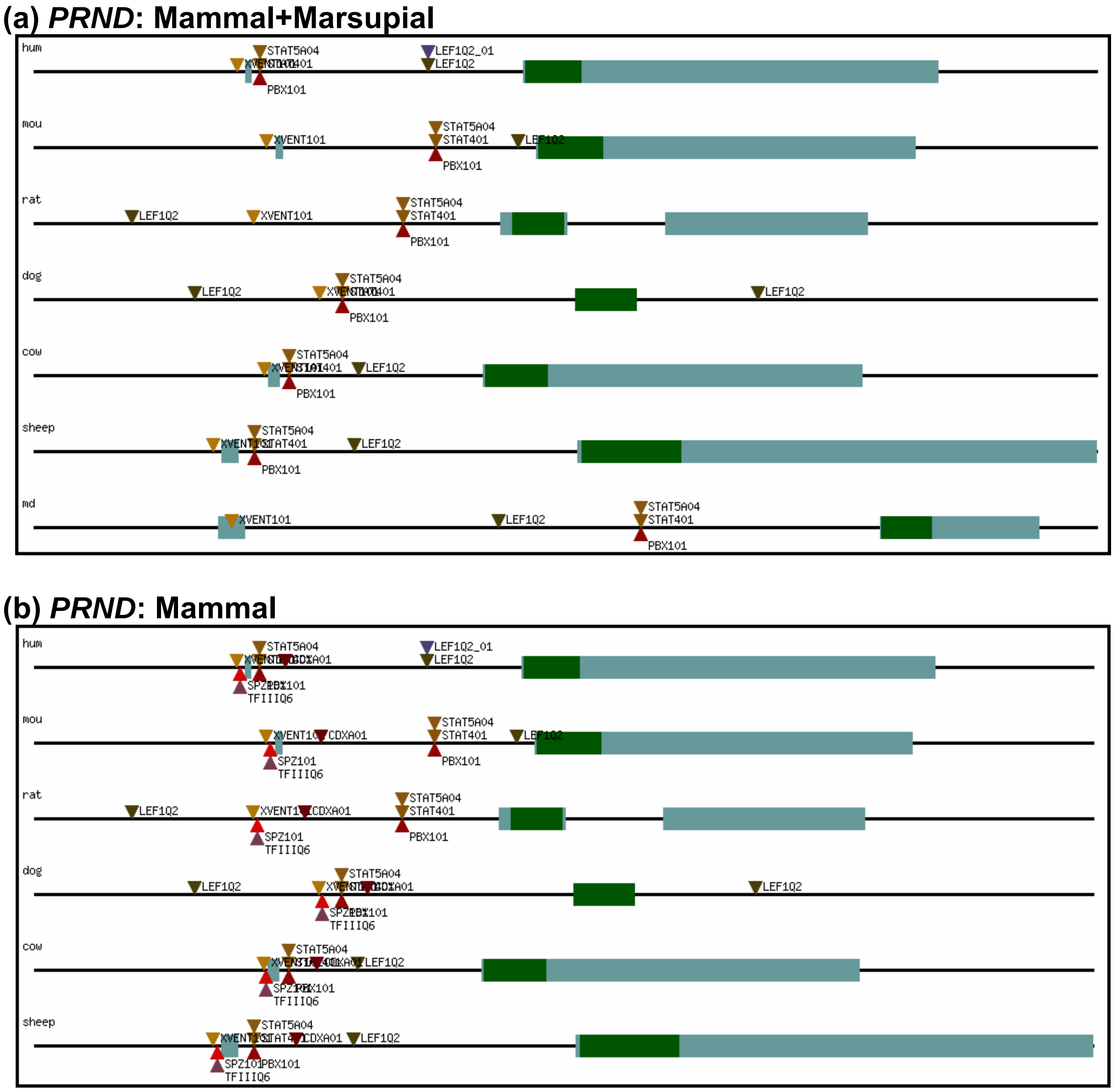


Figure 7-16 PF and TFBS predictions in *PRNP* as displayed by the image drawing program.



(a) *PRND*: Mammal+Marsupial

(b) *PRND*: Mammal

Figure 7-17 PF and TFBS predictions in *PRND.* (a) Between mammal and marsupial, and (b) within mammalian species as displayed by the image drawing program.

7.7 Discussion of the results

The number of conserved sequence motifs and TFs predicted to bind to *PRNP* is greater than that predicted for *PRND* (Figure 7-15). This likely is consistent with evolution of *PRND* at a faster rate in comparison with *PRNP*, which is likely under greater selective pressure as it has an established function in the mammalian lineage resulting it evolving at a slower rate. Some TFBSs showed variation between upstream in a few species to downstream in the rest of the species compared, for example, LEF1 in *PRND* (Figure 7-17 (a)) and MAZ/SPZ1/MAF in *PRNP* (Figure 7-16 (a)). This may be related to the (stringent) parsimony score of 0 (or no mismatches) being used which does not take into account the degenerate nature of the TFBSs (Figure 7-20). The core binding bases which are shorter than the searched motif size of 8, may be conserved in the expected position. Either the comparison of the pairwise alignments in these circumstances or repeating the FootPrinter analysis with a higher parsimony score may prove useful.

For *PRND* analysis, the occurrence of X-VENT1 in the exonic region for marsupial in contrast to the upstream region in mammalian species is interesting. This raises the issue of the validity of masking the exonic sequences for PF analysis, which is the most common practice. The rationale behind masking the exonic sequences is that they are under greater selective pressure. If that was the case, many more conserved motifs and TFBSs in the exonic region should have been predicted but this was not observed. TFBSs have been shown to be present in the 5'UTR (Calhoun et al. 2002) and even the coding exons (Neznanov et al. 1997). One possible explanation for predicting a higher number of conserved motifs in *PRNP* may be related to the use of unmasked sequences. However, masking the repeat regions which are phylogenetically conserved and statistically show binding potential to TFs may indeed have significance in binding to TFs and this information would have been lost by using masked sequences for the analysis. Some of the sites showed binding potential to multiple TFs and the judgment of the significance was based on understanding the biological function of these TFs and their tissue specificity. The input genomic sequence was not split into upstream and downstream sequence datasets allowing the possibility of rearrangement events occurring across the TSS. Comparisons made with avian and amphibian sequences did not produce many TFBSs which may be a result of different regulatory mechanisms in these lineages.

7.7.1 Known *vs* predicted TFs

The use of this combinatorial PF approach (i.e. both alignment-based and motif-discovery-based methods) predicted most of the known TFs for the *PRNP* and *PRND* genes.

PRNP: The Sp1 TF has been shown experimentally to play a role in transcriptional regulation of *PRNP* (Saeki et al. 1996; Baybutt and Manson 1997; Inoue et al. 1997; Mahal et al. 2001). Mahal et al. (2001) also found Ap1 and Ap2 binding sites in the human promoter region. Both Sp1 and Ap1/Ap2 TFBSs were predicted using the pairwise-alignment method in most pairs of sequences compared (Figure 7-18), but these TFs were not identified using FootPrinter analysis (motif absence in any sequence was not allowed). Premzl et al. (2005) reported several regulatory regions in *PRNP* using PF (FootPrinter method) with the then-available sequences (mammals and one marsupial only): most of the TFs (MEF2, Oct- 1, MyT1 and NFAT) were predicted in the intra-mammal comparison.

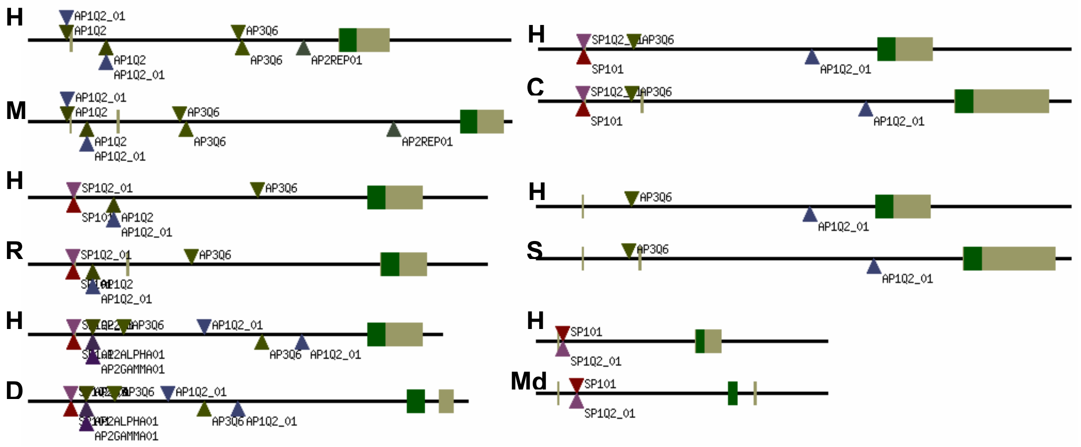


Figure 7-18 Ap1, Ap2, Ap3 and Sp1 sites identified in *PRNP* using LAGAN alignments with the search criteria of 100% identity over 10 bp. (human (H), mouse (M), rat (R), dog (D), cow (C), sheep (S), *M. domestica* (Md)).

PRND: Nagyova et al (2004) experimentally validated the role of USF-1 and NF-Y in *PRND* promoter activity. The NF-Y region was predicted using both alignment-based and FootPrinter methods (Figure 7-19, Figure 7-20). The USF-1 binding site was predicted in comparisons of some sequence pairs using alignment-based methods but not using FootPrinter: this indicates either that the USF-binding site is short and degenerate or that it is not phylogenetically conserved among the species compared.

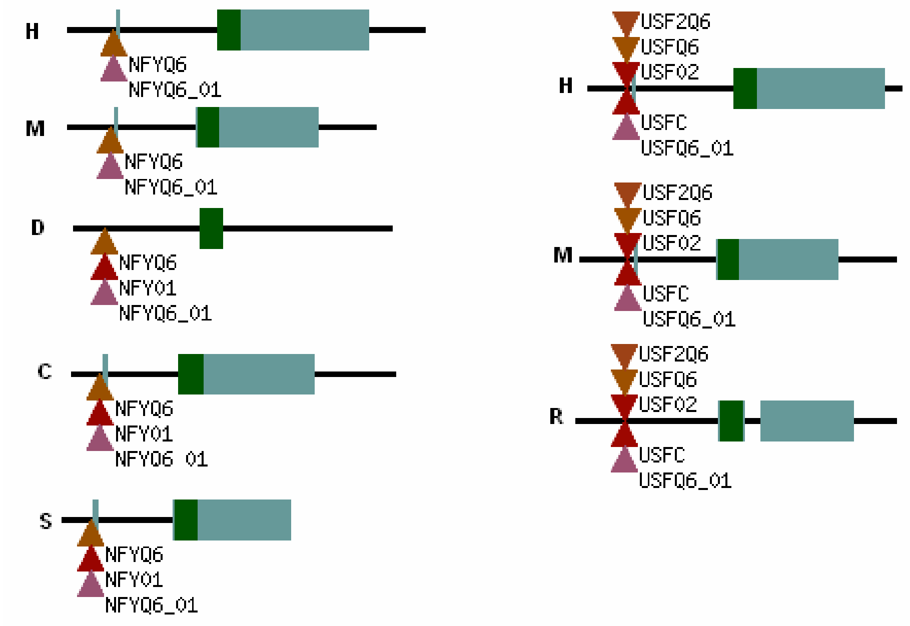


Figure 7-19 Result of a search for NF-Y and USF-1 in *PRND* in the AVID generated pairwise alignments using human (H) as a reference sequence. NF-Y was not predicted in rat (R) and USF-1 was not predicted in cow (C), sheep (S) and dog (D) with the VISTA search criteria of 100% identity over 8 bp.

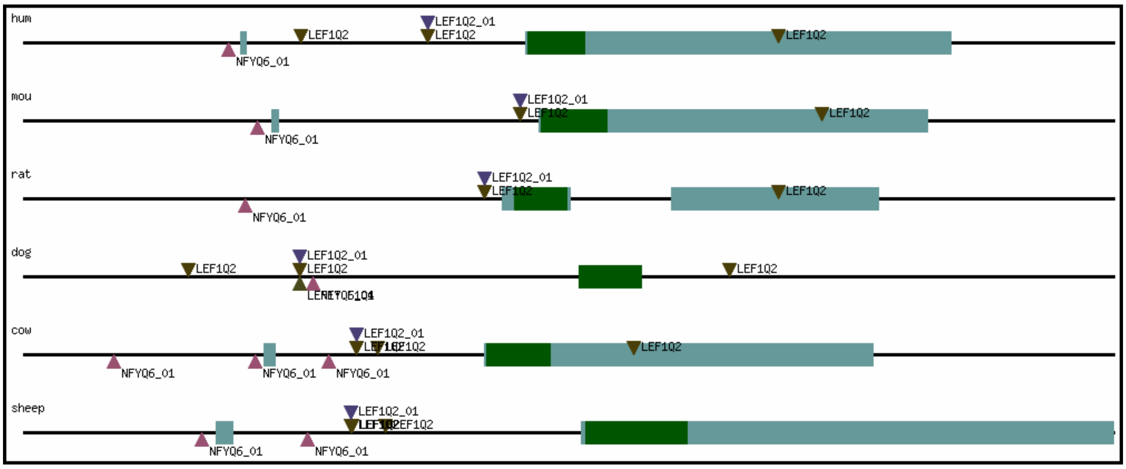


Figure 7-20 Repetition of the FootPrinter analysis of *PRND* with motif size of 8 bp, parsimony score of 1 and subregion size of 1000. Search made for NF-Y USF-1 and LEF1. Note the improvement in the positioning of LEF in rat and dog compared with the previous search criteria (**Figure 7-17**).

Several new TFBSs were predicted for *PRNP* and *PRND*, (*PRNP*: E4BP4, DBP, FAC1, MYB; *PRND*: Spz1, Ap3, CDXA, LEF1) which are phylogenetically conserved for both genes, and which correlate well with physiological behavior consistent with operation of these TFs in regulating other genes (e.g. tissue specificity, specific physiological role) (See Appendix 5).

E4BP4 and DBP binding sites in *PRNP*: The mRNA for *PRNP* is regulated in a circadian manner in the rat brain (Cagampang et al. 1999). The TFs involved in circadian regulation are referred to as "clock related" and include E4BP4 and DBP. They both have antagonistic roles in circadian oscillatory mechanisms (Mitsui et al. 2001), where the former helps in repression and the latter helps in activation of transcription. Recent studies demonstrated that fatal familial insomnia which is a condition characterized by marked changes in many physiological rhythms, is an inherited prion disease (Fiorino 1996). Based on the conserved TFBS analysis, E4BP4 and DBP are shown to be conserved in order, orientation and position in all the mammalian species compared (Figure 7-21). This association appears very significant considering the circadian regulation of *PRNP*.

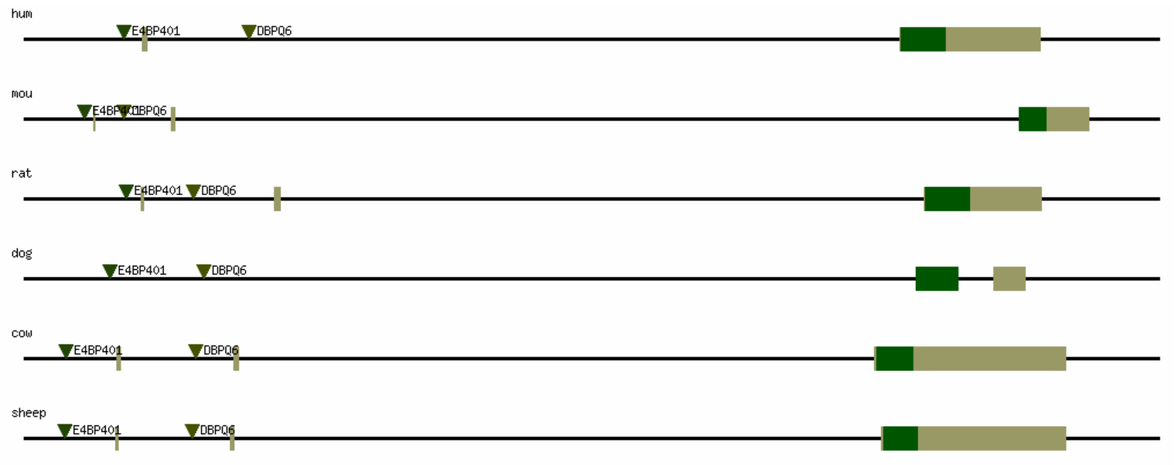


Figure 7-21 Comparison within mammals has shown two TFBS, E4BP4 and DBP, conserved in position in *PRNP*.

***PRND* core promoter**: None of the earlier studies reported the possible role of Spz1 and Ap3 in transcriptional regulation of *PRND;* my analysis shows statistically significant binding for these two TFs which are phylogenetically conserved in mammalian species in position and orientation. Their position relative to the TSS makes them strong candidates for transcriptional regulation. However, computational predictions made by Nagyova et al. (2004) indicated a possibility of the Sp1 (Cys2His2 zinc finger domain) TF binding to a motif at -48 of the *Prnd* promoter (mouse). Based on my TF-binding analysis by Match, the statistical score obtained for the -48 motif for Sp1 (core similarity =0.95) was lower than that of Spz1 (core similarity = 1). A tantalizing finding is the identification of a Spz1 binding site. Spz1 (spermatogenic zip1) plays an important role in transcriptional regulation of genes involved in spermatogenesis and is highly expressed in testis (Hsu et al. 2001). As USF-1 and NF-Y are ubiquitous TFs, the tissue specific expression of *PRND* in testis is likely controlled by some tissue specific TFs. Spz1 is strong candidate to play a critical role in transcriptional regulation of *PRND* in testis in association with other TFs. Luciferase assays in reporter constructs made from a combination of mutations to the USF-1, NF-Y, Ap-3, and Spz1 binding motifs have been used to test the functional importance of each of these predictions (Figure 7-22) (see Chapter 8).

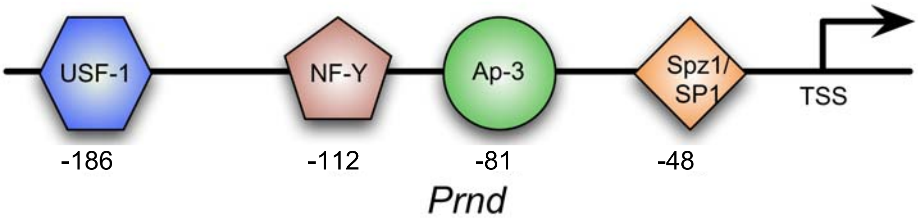


Figure 7-22 Schematic representation of mouse *Prnd* promoter. USF-1 and NF-Y are known to bind and showed transcriptional activation. The sites for Ap3 (TATTTGGA) and Spz1/Sp1 (CCTCCCC) are phylogenetically conserved and new predictions from my analysis for transcriptional regulation of *PRND*.

7.8 Application and Conclusions

The development of a graphical web interface has facilitated evaluation of results from the PF and TFBS analysis pipelines. An application of the pipeline and web interface has been illustrated by an analysis of *PRNP* and *PRND* genes. This revealed several new conserved TFBSs, in addition to detecting already published and experimentally validated TFs for regulating these genes. Detection of the latter serves as a confidence test for the pipeline analysis. Several of the newly predicted TFBSs are consistent with the known functions of these genes, providing strong starting points for follow up experimental studies. A combinatorial approach of predicting conserved motifs using FootPrinter and AVID/LAGAN methods followed by TF binding analysis significantly improved the confidence in the predicted TFBSs. This pipeline was also tested on the newly discovered PrP family gene, *SPRN*, providing us with valuable initial functional predictions of a gene whose function is not known. The development of a pipeline which incorporates both alignment-based and motif-discovery based methods with TFBS analysis is novel, and provides a powerful new tool for high throughput, robust analysis. The concurrent development of the graphical-display module to this pipeline greatly enhances its usefulness by facilitating intuitive and interactive analysis of the results.

8 Experimental validation of predictions

8.1 Background

While doppel and prion protein share similar structural properties, their tissue expression patterns differ significantly. *Prnp* is expressed widely with the highest concentration in neurological tissues whereas *Prnd* is expressed mainly in the testis of adult mice. This indicates a tighter transcriptional regulation of *Prnd*. The core *Prnd* promoter (mouse) has been identified to be in the region -185/+27 with respect to the transcription start site (TSS) (Nagyova et al. 2004). Based on the TF-binding predictions made by PF, two evolutionarily conserved regions (100% conservation in 8 bp across human, mouse, rat, cow, sheep and dog) for spermatogenic zip (Spz1) and activating protein (Ap3) were predicted to bind in the *Prnd* core promoter. It has been experimentally shown that the upstream stimulating factor (USF-1) and nuclear factor-Y (NF-Y) also play a functional role in the transcriptional regulation of *Prnd* (Nagyova et al. 2004; Sepelakova et al. 2005) which are also the part of the core promoter.

The mouse promoter was used for this study because previous published functional studies for *Prnd* were on mouse promoter. This allows the data to be compared across studies. In this work, a particular focus is placed on studying the influence of the Spz1 TF on the promoter activity. *Spz1* belongs to the bHLH-Zip family and is exclusively expressed in mouse testis and epidydimis. It plays a role in cell proliferation and differentiation and is involved in spermatogenesis (Hsu et al. 2001). The Spz1 binding motif, “CCTCCCCC” is situated at -48 bps upstream to the TSS in mouse and is conserved in position and orientation among the 6 mammalian sequences compared (Chapter 7: Table 7-3b). As the current literature supports the idea that *Prnd* is expressed in testis and that it has a role in male fertility, this predicted TF was considered as of particular interest. Nagyova et al. (2004) indicated a possibility of the Sp1 TF binding to a motif at -48 of the *Prnd* promoter (mouse). However, *in silico* binding analysis results presented in Chapter 7 (section 7.7.1) indicates that Spz1 is more likely to bind to this motif than Sp1.

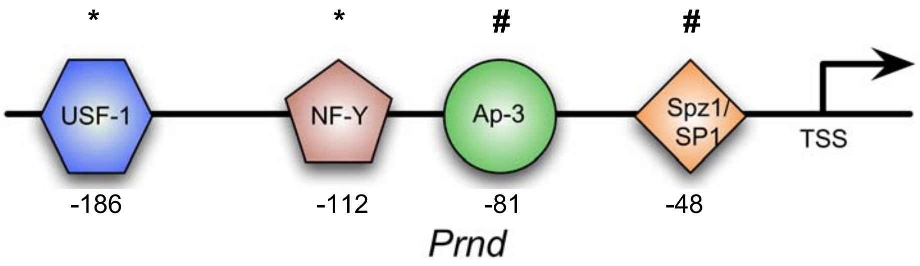


Figure 8-1 Schematic representation of the mouse *Prnd* promoter. (* Confirmed, # Predicted)

The aim of this work was to study the mouse *Prnd* promoter for transcriptional regulation encompassing the four binding sites (Figure 8-1): USF-1, NF-Y, Ap3 and Spz1 and to validate the predictions made in Chapter 7 for *PRND* core promoter. To elucidate the significance of Sp1 and Spz1 in transcriptional regulation of the mouse *Prnd*, functional studies were performed on the *Prnd* promoter. The role of Ap3 and Spz1 binding sites in *Prnd* core promoter were studied for the first time.

8.2 Materials and Methods

8.2.1 Database searches

NCBI Gene Expression Omnibus (GEO) profiles were searched for analyzing the gene expression of USF-1 (gene ID: 1426164_a_at), NF-Y (1452560_a_at), Ap3, Spz1 (1450653_at) Sp1 (1448994_at) and to compare with *Prnd* (1425681_a_at). The NCBI GEO dataset GDS565 (platform GPL339) which is designed for analyzing sex-specific transcription in somatic and reproductive tissues was used for this purpose.

8.2.2 Cell lines

The cell lines GC-2spd (ATCC: CRL-2196) (provided by Dr. Kate L. Loveland, Monash Institute of Medical Research) and bEnd.3 (ATCC: CRL-2299) (provided by Professor Carolyn Geczy, Inflammatory Diseases Research Unit School of

Medical Sciences, University of New South Wales) were used to study *Prnd* promoter activity. The GC-2spd cell line is of testicular origin that corresponds to primary spermatocytes. bEnd.3 is a mouse brain endothelial cell line. Both cell lines were maintained in DMEM with 10% fetal bovine serum, 1% glutamine and 1% penicillin. They were grown in a humidified incubator at 37 ^{o}C and 5% CO_2.

8.2.3 Plasmids

A luciferase construct containing the four TFBS of the mouse *Prnd* gene was made and is referred to as mDpl230 (-197/+27) (Figure 8-2). A second luciferase construct containing only the Spz1 binding region was made which is referred to as mDpl90 (-67/+27) (Figure 8-2). These constructs were made by amplifying the *Prnd* promoter region by PCR from mouse genomic DNA using primers lmDpl and rmDpl (Table 8-1). The PCR product was digested with *Hin*dIII and *Bgl*II restriction enzymes for the mDpl230 construct and by *Hind*III and *Sac*I for mDpl90 and cloned into the corresponding sites of the pGL4 reporter vector (Promega) upstream of the firefly luciferase gene.

Table 8-1 Primers used in this study. *Published mutant site (Nagyova et al. 2004)

Luciferase construct	
lmDpl	GAGGTTGGGTCTTGATGGTC
rmDpl	GCTGGAAGGGAAGTCACAAG
Luciferase construct with mutated binding site (underlined)	
Spz1-mut	GGTAGAGAGGCCCC**gat**CCCCTGCAGCGCCTATAT
Ap3-mut	GAAGGGCTACCCTA**gg**TG**a**AGGGTTGGAGCTCGGT
NFY-mut*	ATGCAGGAGCCCTT**tt**TTGGTCCTGCTGTGGAGGGA
USF-mut	ATCAAGATCTTCA**aga**GGTTTTATCAGTGAAG
Tissue expression	
lmex1Dpl	TCAGAGGCCACAGTAGCAGA
rmex2Dpl	GCTTGCTATCCTGCTTCTCC
lspz1	CATCTGCTCTCCCTGGACTC
rspz1	CTGGCGACTTCTACCGAAAG

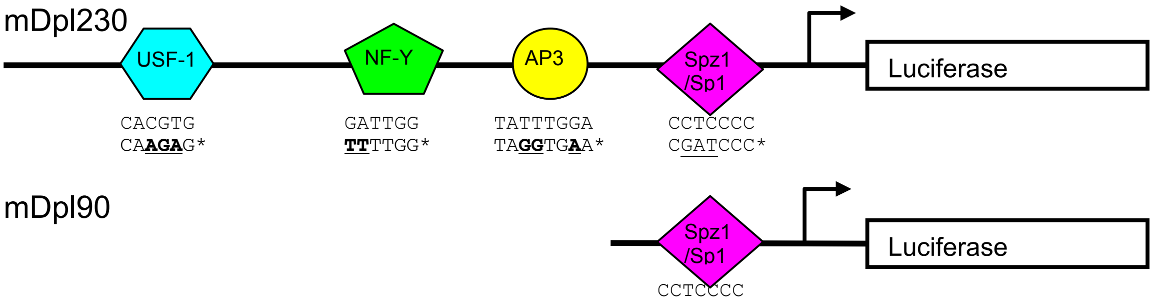


Figure 8-2 Two Luciferase constructs mDpl230 and mDpl90 were made. * indicates the mutated core binding site with the mutated bases underlined.

Plasmids for expression of TF Spz1 cloned into pCI-neo vector (provided by Dr. Jack Hsu, Kaohsiung Medical University) and Sp1 cloned into pCR3.1 vector (provided by Dr Mark Hulett, John Curtin School of Medical Research) were obtained. pcDNA was used as a control plasmid (a gift from Dr. Alison Shield JCSMR). The nucleotide sequence of all the plasmids and luciferase constructs were confirmed by automated sequencing. Plasmid DNAs for use in transfection were prepared using the Invitrogen Midiprep kit.

8.2.4 Transfection and luciferase reporter assays

1 x 10^5 cells were seeded per well of a 12-well plate, in serum free and antibiotic free media. Transfection for the luciferase reporter assay was performed using FuGENE™ 6 (data in Figure 8-5, Figure 8-6). The ratio between the DNA (1µg) and transfection reagent used was 1:3. After growing overnight to ~90% confluence, cells were transiently transfected according to the manufacturer's instructions. A constant amount of total DNA was used by adjusting with the pcDNA. The cells were trypsinised and transferred to a 96-well plate for measurement of luciferase activity. Gene expression was evaluated using the Dual-Luciferase Reporter Assay System (Promega). The luciferase activity was normalized against the Renilla expression produced by the pRL-TK renilla vector (Promega). All assays were repeated at least twice in triplicate. Luciferase assays to evaluate the effect of Sp1 and Spz1 were performed at two different concentrations of luciferase construct and TF plasmids (1:1 and 1:3) (Table 8-2).

Table 8-2 Concentrations of luciferase vector and TF plasmids used to compare the promoter activity induced by Sp1 and Spz1.

	Sp1	**Sp1**	**Spz1**	**Spz1**
Luciferase vector	800ng	500ng	800ng	500ng
TF plasmid/pcDNA	800ng	1500ng	800ng	1500ng
Renilla	300ng	300ng	300ng	300ng
Cell line	GC2	GC2	bEnd.3	bEnd.3

8.2.5 Endogenous gene expression: RNA extraction and PCR analysis

Endogenous *Prnd* expression was analyzed by semi-quantitative PCR in GC2 and bEnd.3 cells transfected with Spz1 expression plasmid and pcDNA. The test was performed with three different transfection reagents: FuGENE™ 6, Metafectamine and Lipofectamine 2000. 1 μg of Spz1 and pcDNA plasmids were transfected into GC2 and bEnd.3 cells as described in section 8.2.4.

24 hours after transfection, the cells were trypsinised and collected in 1.5 ml tubes which were briefly spun to separate cells from the media. These cells were either used directly for RNA extraction or stored in RNAlater until used. Total RNA was extracted using the RNeasy® plus Mini kit (Qiagen) as described by the manufacturer. 1 μg of total RNA was used for cDNA synthesis using the Superscript™ III First-Strand cDNA Synthesis (Invitrogen) kit. For amplification of *Prnd*, primers were designed spanning the intron (lmex1Dpl and rmex2Dpl) (Table 8-1). Success of transfection was checked using Spz1 primers (lspz1 and rspz1). β-actin was used as a control. Amplification reactions were performed using Platinum®*Taq* DNA polymerase (Invitrogen); PCR conditions were 15 min at 95 °C, followed by 30-35 cycles of 30 s at 94 °C, 30 s at 56 °C, and 30 s at 72 °C.

8.2.6 Mutation of TF binding sites

Each of the binding sites for Spz1, Ap3, NF-Y, and USF-1 were mutated using the PCR-based technique. Before the mutants were made, computational analysis was performed by substituting several different bases in the core binding region and each of these combinations was tested against the TRANSFAC database using

Match to make sure that the mutated region did not bind any other known TFs. Once the right mutations were determined, forward and reverse primers with mutations were designed (Table 8-1). High fidelity Platinum® *Pfx* DNA polymerase (Invitrogen) was utilized and the manufacturer's protocol was followed. PCR was performed for 18 cycles [94 °C for 2 min; 94 °C for 15s; 55 °C for 30s; 68 °C for 5 min.30s]. PCR product was digested with *Dpn*I restriction enzyme. 1 μl of digested PCR product was transfected into TOPO10 cells and the clones sequenced to confirm successful mutation. The successful mutants were then digested with *Bgl*II and *Hind*III and sub-cloned back into pGL4 to ensure the integrity of the vector. This generated four different luciferase vectors each with one of the four binding sites mutated (mDpl230-USFmut, mDpl230-NFYmut, mDpl230-Ap3mut and mDpl230-Spz1mut).

The mutant luciferase vectors were transfected into GC2 and bEnd.3 cells and the luciferase activity produced by endogenous TFs was measured to study the effect of the mutated site on promoter activity.

8.3 Results

8.3.1 Spz1 is only expressed in testis

The NCBI GEO database was searched for the expression profiles of TFs of interest and these were compared with those for *Prnd* (Figure 8-3). *Sp1, USF-1 and NF-Y* are widely expressed whereas *Spz1* shows expression specific to testis. The data for Ap3 was not available from GEO profiles. Consistent with the tissue expression profile of *Spz1*, *Prnd* also showed localized tissue expression in testis. The other finding supporting my prediction that Spz1 may have more affinity for the -48 binding site than Sp1 is that the consensus binding site of Spz1 ("GG(G/A)GGG(G/A)(A/T)T") (Hsu et al. 2001) is more similar to the conserved -48 motif ("CCTCCCCC") than to that of the Sp1 consensus ("CCGCCC") (Tamaki et al. 1995).

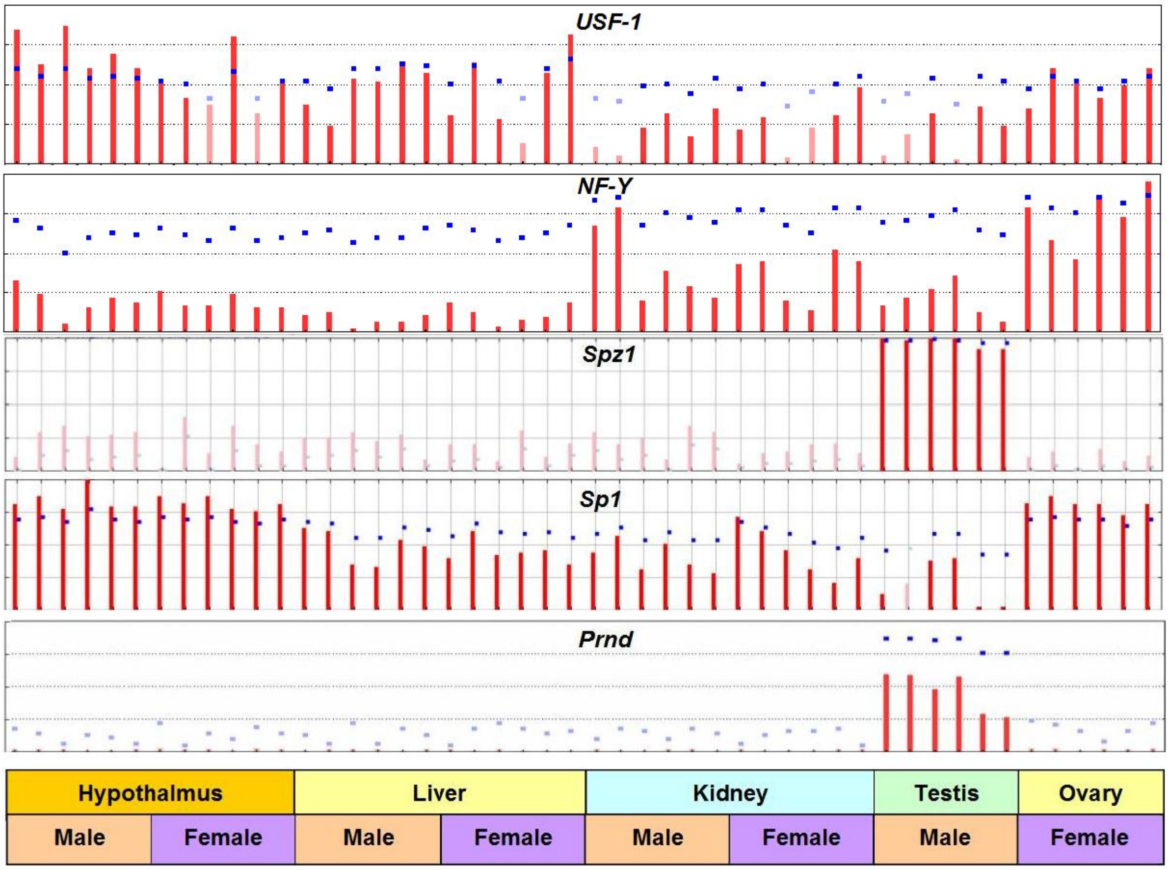


Figure 8-3 Mouse tissue expression profiles of *USF-1*, *NF-Y*, *Spz1*, *Sp1* and *Prnd* obtained from the NCBI GEO database. The red bars represent relative measure of abundance of each transcript. The blue squares represent the percentile ranked value of a spot compared to all other spots within that Sample

8.3.2 Endogenous *Prnd* expression analysis in GC2 and bEnd.3 cells

Prnd and *Spz1* transcripts are expressed endogenously in GC2 and b.END3 cells. Cells (GC2 and bEnd.3) transfected with Spz1 expression plasmid did not show any changes in endogenous *Prnd* levels compared with the untransfected or cells transfected with pcDNA (Figure 8-4). This was confirmed by transfection assays performed using Lipofectamine2000, and FuGENE™ 6. However, the endogenous *Prnd* levels varied based on the transfection reagents used as *Prnd* transcript was not detected (after 30 cycles of PCR) after transfection with Metafectamine.

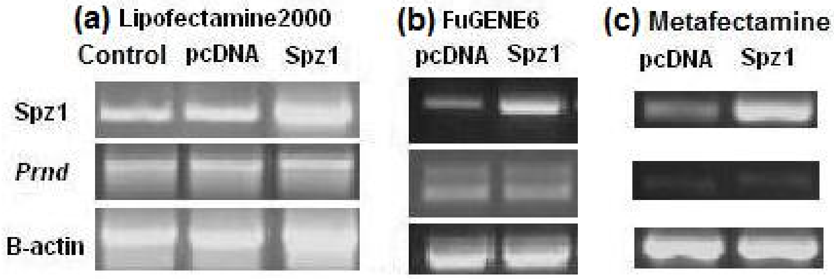


Figure 8-4 Endogenous gene expression of *Prnd* using RT-PCR. (a) Lipofectamine2000. Spz1 product has stronger intensity in the Spz1 transfected cells compared with pcDNA or untransfected cells (control) indicating successful transfection (30 PCR cycles for Spz1 and 35 PCR cycles for *Prnd*). (b) Cells transfected using FuGENE™ 6 (30 PCR cycles) and (c) cells transfected using Metafectamine (30 PCR cycles for Spz1 and 35 PCR cycles for *Prnd*). Spz1 did not have any influence on *Prnd* levels in all of the tests. The data shown are from GC2 cells (bEND.3 data not shown as it was similar to GC2 cells).

8.3.3 Test for *Prnd* core promoter

The activity of mDpl90 compared with mDpl230 (Figure 8-5) was low, in accord with the observations of Nagyova et. al. (2004). This indicates that other TFs in the -197 region are required for maximum activity, which is consistent with the core promoter region spanning -185 to +27 bp as reported previously. The significance of the Spz1 binding site (-48 region) was tested by comparing the luciferase experimental results from mDpl90 (containing only the Spz1 binding site) with mDpl230 (containing the four binding sites of interest). Co-transfecting Spz1 expression plasmid and luciferase construct (mDpl230 or mDpl90) showed that Spz1 induced expression from the *Prnd* promoter in both GC2 and bEnd.3 cells (Figure 8-5).

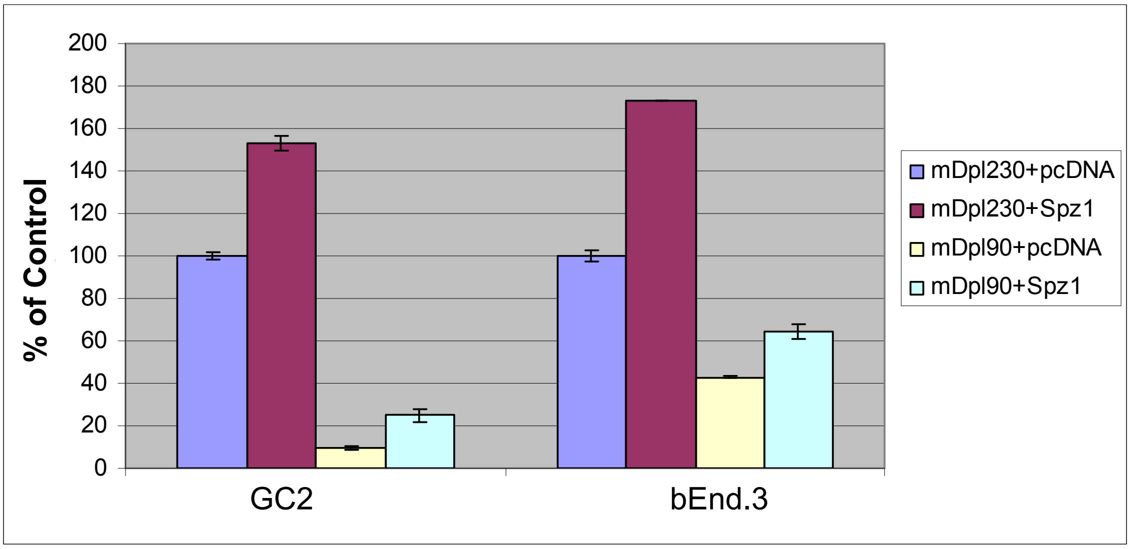


Figure 8-5 Comparison of luciferase activity between mDpl230 and mDpl90 luciferase constructs in GC2 and bEnd.3 cell lines. Note the increase in the activity of the promoter in cells transfected with Spz1 TF for both the luciferase constructs with short (mDpl90) and long (mDpl230) promoters**.**

8.3.4 Effect of Spz1 and Sp1 on promoter activity

Sp1 and Spz1 expression plasmids were co-transfected with the mDpl230 luciferase construct in GC2 and bEnd.3 cells to compare their effect on promoter activity. This was tested at two different concentrations of the mDpl230 luciferase construct and TF plasmid (1:1 and 1:3) (Figure 8-6). Both Spz1 and Sp1 plasmid had a slight positive effect (25% and 15% respectively) on luciferase promoter activity at the ratio of 1:1. Increasing the Spz1 concentration by three times (1:3) did not give any additional increase in promoter activity suggesting that the site is saturated. However, the same increase in Sp1 concentration reduced the promoter activity (from +15% to -15%), possibly by competitively binding to some other important TFBS thereby reducing the transcriptional efficiency. In both the tests, Spz1 showed more activity compared with Sp1. This data shows that Spz1 is a better TF than Sp1 for the -48 region.

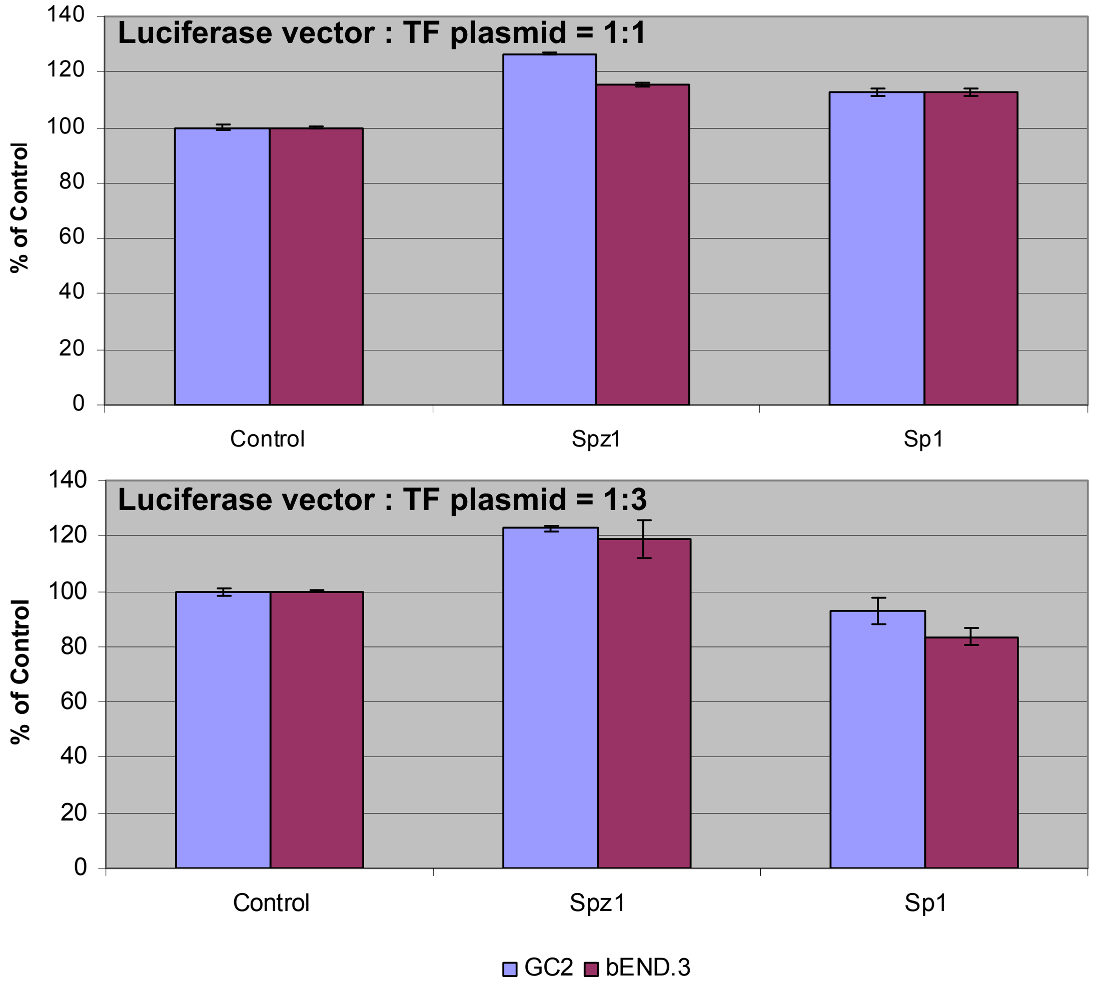


Figure 8-6 The effect of Spz1 and Sp1 TF on mDpl230 luciferase vector using different concentrations of TF plasmid in GC2 and bEnd.3 cells.

8.3.5 Mutation of Spz1 and NF-Y binding sites decreases promoter activity

Luciferase assays using mutant promoter constructs were performed in GC2 and bEnd.3 cells to evaluate the significance of each of the four binding sites under the influence of endogenous TFs. However, the endogenous expression of USF-1, NF-Y and Ap3 was not tested in these cells, but based on the data from GEO database they are expressed in a wide range of tissues (Figure 8-3). Mutation of the the Spz1 binding site (mDpl230-Spzmut) reduced the promoter activity by 40-60% (Figure 8-7 (a), (b)) compared with the cells transfected with unmutated luciferase vector (mDpl230). Co-transfecting Spz1 expression plasmid with mDpl230-Spzmut showed no additional changes in the promoter activity indicating that Spz1 does not bind to the mutated plasmid (Figure 8-7 (b)). This confirms that the increase in the promoter activity (Figure 8-6) observed in the mDpl230 luciferase vector cotransfected with the Spz1 expression plasmid was an effect produced by targeting the -48 region.

The most significant effect was seen in luciferase vector (mDpl230-NFYmut) with a mutation in the NF-Y binding site, consistent with the experimental results of Nagyova et al. (2004) (Figure 8-7). Mutation of the Ap3 binding site showed only a minor reduction of promoter activity by 15-20% (Figure 8-7). Interestingly, mutation of the USF-1 binding site (mDpl230-USFmut) showed an increase in activity in both GC2 and bEnd.3 cells indicating repression of *Prnd* expression by USF-1. This is contrary to the findings of Nagyova et. al. (2004) showing a marked reduction in the promoter activity in luciferase vector with a different mutation (Figure 8-8) on the USF-1 binding site in GC1 and bEnd.3 cells. It should be noted that a search against the TRANSFAC database using Match did not show any binding affinity to the USF-1 mutated site I used in this study. However, the USF-1 binding site mutation created by Nagyova et. al. (2004) showed binding to various TFs (Figure 8-8). This may explain the contradictory findings between the two studies. My results suggest that mutation of the USF-1 binding site has removed the repressor action by USF-1, which has been reported in other genes (Hadjiagapiou et al. 2005) thus producing the positive effect.

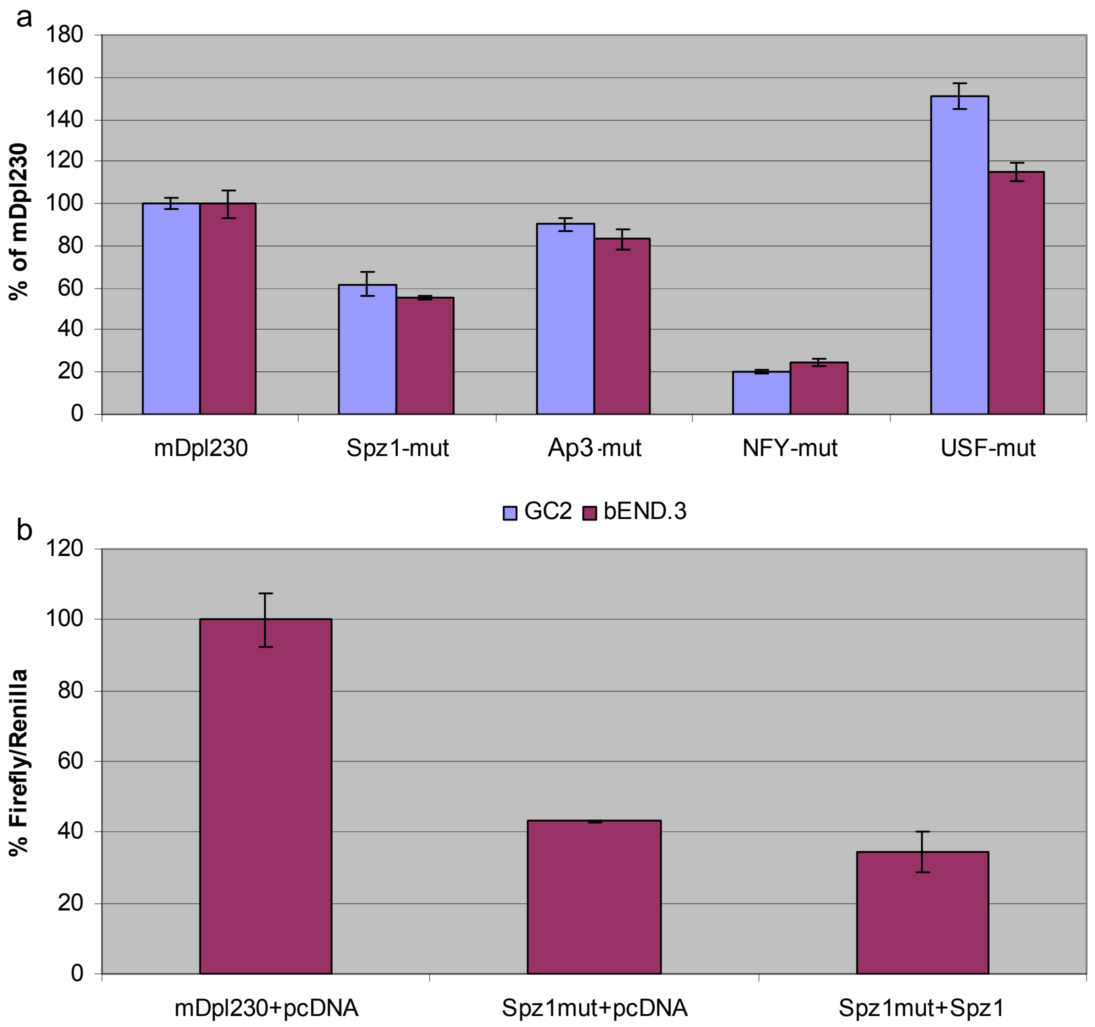


Figure 8-7 Luciferase assays using mutant promoter constructs. (a) Analysis of the promoter activity using luciferase vector with mutant TFBSs in GC2 and bEnd3 cells (no TFs were co-transfected). (b) Effect of cotransfection of Spz1 on expression from mutant luciferase vector, Spz-mut in bEnd.3 cells.

```
USF-1 mutation site: TCTTCAttaGGTTT
Search results against TRANSFAC using Match:

V$OCT1_03      |   1 (-) |  0.985 |  0.982 | tcttcATTAGgtt
V$SRY_01       |   8 (-) |  0.894 |  0.871 | taGGTTT
V$GEN_INI2_B   |   2 (+) |  0.977 |  0.968 | cttCATTA
V$GEN_INI3_B   |   2 (+) |  0.979 |  0.965 | cttCATTA
V$Ap3_Q6       |   1 (+) |  0.882 |  0.907 | tCTTCAtt
```

Figure 8-8 USF-1 mutation created by Nagyova et. al. (2004). Analysis using Match against TRANSFAC data indicated several possible TFs which might bind to the mutated region.

8.4 Discussion

The aim of this work was to validate the significance of predicted Ap3 and Spz1 binding sites in the *Prnd* core promoter with a particular emphasis on the Spz1 binding site located at -48 bp upstream to the TSS. The results suggest that the -48 region is crucial for *Prnd* promoter activity, validating the prediction of a Spz1 binding site. This is evident by the increase in the promoter activity brought about by transfecting Spz1 expression plasmid and the fact that mutating this site down regulated the promoter activity by 40-60%. Further work (electrophoretic mobility shift assay) could be done to prove whether Spz1 binds directly to this region.

Lack of a many-fold increase in promoter activity in the Spz1-transfected cell may be related to already maximal activity of promoter due to high endogenous levels of Spz1 in these cells. Further work could involve knockdown experiments with siRNA (small interfering RNAs) and shRNA (short hairpin RNA) to validate this hypothesis.

The current study supports the finding of Nagyova et al. (2004) that NF-Y is the most significant TF for *Prnd* promoter activity. The minimal (15%) changes in promoter activity for the Ap3-mutated site may be related to lack of endogenous Ap3 in this tissue. Experimental data regarding Ap3 and testis was not available from GEO (Figure 8-3) so this would need to be confirmed for these cells/tissue type. This site may be significant in other tissues, or may be that it needs some other cofactors which are not present in the test cells. NF-Y and USF-1 are ubiquitous TFs, and the testis-specific expression of *Prnd* may be due to Spz1. The experimental data for Spz1 serves to validate the computationally predicted TFBS.

Also note that while Nagyova et al (2004) predicted Sp1 binding site in the promoter, the expression level of Sp1 is much lower in testis than in other tissues, therefore more likely that Spz1 binds to the promoter *in vivo* than Spz1. I have now demonstrated a functional Spz1 element in the promoter of *Prnd*. It seems highly likely, given the level of expression of Spz1 in testis and the evolutionary conservation of this site, that this TF is responsible for the testis specific expression of *Prnd* in the mouse.

9 Human *Doppel* gene polymorphisms in male infertility

9.1 Background

PRND shows testis-specific expression in adult human (and mice). Several published studies have demonstrated its role in spermatogenesis. This which is supported by my new finding of the presence of a conserved Spz1 binding site (Chapter 7 and Chapter 8) which is known to play a role in regulating the genes involved in spermatogenesis. As *Prnd* knockout mice were shown to be infertile, *PRND* might be important for fertility in human. The purpose of this study is (a) to search for rare mutations which might cause infertility and (b) to study the possible association of polymorphisms in human *PRND* with infertility. Polymorphisms in the promoter region together with the ORF were studied, as they may affect the gene expression.

Earlier polymorphism studies focused on studying the genetic variation in *PRND* in the etiology of prion disease. This led to the detection of a few polymorphic genotypes in the *PRND* coding region: M26T, P56L and T174M (Peoc'h et al. 2000). There was considerable interest in studying the influence of these genotypes on prion disease. The implications of polymorphic codon 174 showed contradictory results (Schroder et al. 2001; Infante et al. 2002). Several studies showed no significant association between T174M with sCJD (Schroder et al. 2001; Croes et al. 2004; Jeong et al. 2005; Vollmert et al. 2006). Similar conclusions were drawn for polymorphic codon 26 (T26M) and codon 56 (P56L) (Mead et al. 2000; Schroder et al. 2001; Infante et al. 2002). The association of these polymorphisms with other neurodegenerative diseases such as Alzheimer's disease disclosed no significant difference in the frequency of *PRND* genotype betweens cases and controls was found (Golanska et al. 2004).

Although *PRND* polymorphisms have been studied to ascertain an association with prion disease, there have been no reports on *PRND* mutations or polymorphism investigating a possible association with male infertile populations. In this study, the promoter and ORF regions of the human *PRND* gene region were screened for variations in 96 control and 96 infertile male samples.

9.2 Materials and methods

9.2.1 Study population

The patients in this study were diagnosed with male infertility. The following DNA samples were obtained from Dr. Moira O'Bryan (Centre for Reproduction and Development, Monash Institute of Medical Research). The ethnic background of these samples is not known.

- 96 infertile men
- 96 controls comprising 53 fertile men (proven fathers) and 43 healthy men with normal semen analysis

The sperm concentration, motility and morphology associated with infertile samples ranged from normal to a combination of poor concentration, motility, and morphology in varying degrees. The control population contained fertile men or proven fathers and healthy individuals where the fertility factor is not known but have normal semen analysis parameters and hormone levels. All DNA samples were diluted in water and were supplied at a concentration of 50 ng/µl. A volume of 20µl of each sample was supplied. These samples were further diluted by using 10 µl from each sample and adding 30 µl of water.

9.2.2 Genotyping

Two methods were tested for detecting polymorphisms in the *PRND* promoter and ORF

- LightScanner® (using LCGreen® dye)
- Big-Dye sequencing

9.2.2.1 Idaho technology LightScanner®

The LightScanner hi-res melting technique (refer section 2.1.4.1) was assessed for its suitability for analyzing the *PRND* gene. The LCGreen dye was incorporated in the PCR mix and double the normal quantity of magnesium was used to compensate for any loss of sensitivity by incorporation of the dye. The DNA samples were amplified using primers lorfDpl and rorfDpl for ORF and lprDpl and rprDpl for promoter region (Table 9-1, Appendix 8). PCR was performed using Platinum®*Taq* DNA polymerase (Invitrogen) (4pmol of primer and 50 ng of DNA as template). The PCR conditions were optimized based on a number of test runs. The final PCR protocol was as follows: 94 °C for 5 min; denaturation at 94 °C for 30 s; annealing at 66 °C for 30 s; extension at 72 °C for 20 s; and a final extension 72 °C for 5 min; amplification for 45 cycles. The LCGreen dye does not interfere with the amplification process but is known to increase the melting temperature by 2-4 °C. The PCR amplified product was transferred to the LightScanner for the melting analysis using the LightScanner software.

Out of the 24 test samples (infertile male) on which PCR (ORF) was performed, 13 samples showed good quality product as analyzed from the melting curves. The melting curves were normalized by defining regions in the pre- and postdenaturation parts of the curve, and the value was set for the melting. The melting curve shows the decrease in the fluorescence against increasing temperature (Figure 9-1). Variations in the melting curves were observed among the 13 samples. Because of the large number of variations within the PCR amplicon, it was decided that this approach may not be appropriate for analyzing this gene and, hence, further work was not carried out by this method. As the sample set was relatively small and the number of expected SNPs large (high resolution is best applied to single SNPs with a PCR amplicon), it was decided that direct sequencing would be the best approach to search for novel mutations.

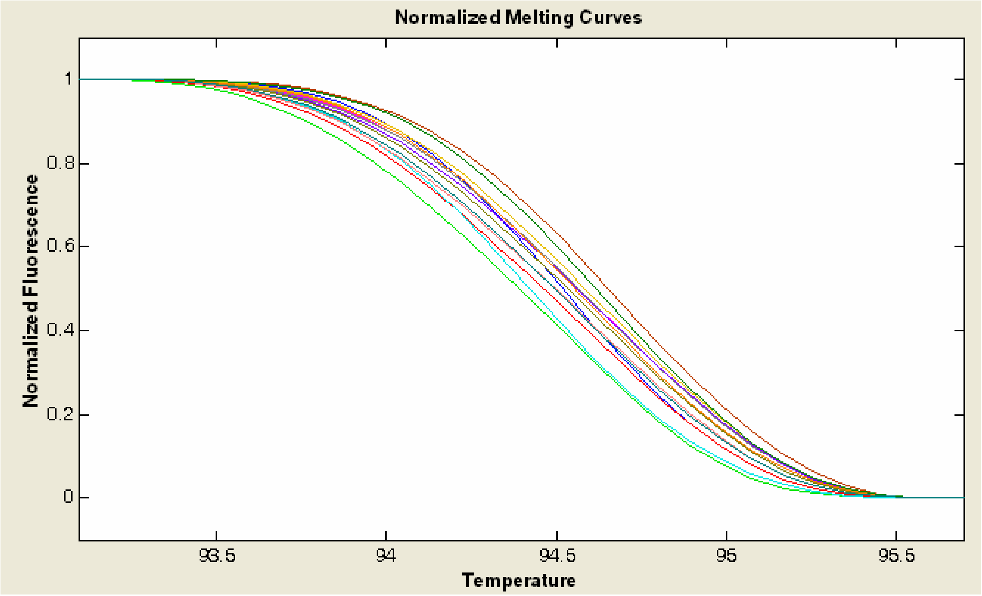


Figure 9-1 Melting curves for Dpl ORF (Infertile male population) showing a number of variations obtained using LightScanner.

9.2.2.2 Big-Dye sequencing

PCR was performed in 96-well plates using the DNA Engine Tetrad® thermocycler. Genomic DNA from infertile and control population samples was amplified using promoter primers lprDpl and rprDpl and ORF primers lorfDpl and rorfDpl. The amplified product was then sequenced using lprDpl for promoter and a nested forward primer lorfSeqDpl for ORF product.

PCR. Platinum®*Taq* DNA polymerase (Invitrogen) was used for the analysis (12 ng genomic DNA template and 4 pmols of the forward and reverse primers) with PCR conditions: 94 °C for 5 min; denaturation at 94 °C for 30 s, annealing at 66 °C for 30 s; extension at 72 °C for 20 s; and a final extension 72 °C for 5 min; amplification for 35 cycles.

Table 9-1 Primers used to amplify the promoter and ORF regions of human *PRND* gene. (See Appendix 8 for primer map).

Primer name	Primer sequence	Comment
lprDpl	CTTGCCCTCTTTTTGAGCTG	Left promoter
rprDpl	CGTACCTTGGCTCTCTCTGG	Right promoter
lorfDpl	TAGCAAAGGAGCTCGGTGTT	Left ORF
rorfDpl	GCTGCTGCACTCTGTACTGC	Right ORF
lorfSeqDpl	TAACCCTGCACAACCCAAAC	Nested primer for ORF sequencing
lDplrflp	GGGGAGTTCCAGAAGCCAGAC	Left RFLP
rDplrflp	TCAGAACGCAGGCACATACCAG	Right RFLP

SAP-Exo digest. Shrimp alkaline phosphatase (SAP) removes the phosphate groups from the excess dNTPs left over from the PCR reaction. Exonuclease I (EXO) digests the single-stranded PCR primers into dNTPs and the phosphate groups are removed by the SAP.

SAP (1 µl, 1 U/µl) and 10X SAP buffer (0.9 µl) (New England Biolabs), EXO (0.15ul, 20U/µl) and 10X EXO buffer (1.5 µl) (Roche) was added to the PCR reaction and incubated at 37°C for 1 h. SAP was denatured for 10 min at 80°C.

The PCR product was verified on an agarose gel by loading 3 µl of the PCR product from 13 random samples each from the infertile and the control population (Figure 9-2). This was done to ensure that the PCR product was of good quality for direct sequencing and that the samples were free of contamination.

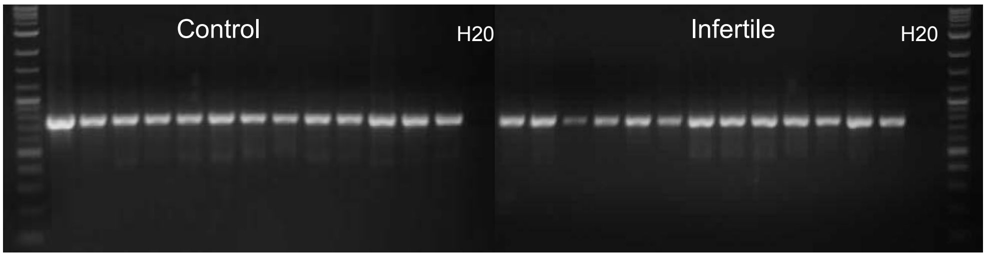


Figure 9-2 Agarose gel electrophoresis of randomly picked samples to check the quality of the ORF PCR product in control and infertile samples. The PCR product matched the expected product size of 740 bp. A negative control (water) was included in each set of PCR amplifications.

The sequencing protocol specified by the Biomolecular Resource Facility at the JCSMR was used as a starting point for sequencing. Several trial runs were made to modify the protocol to suit the needs for sequencing on 96-well plates. The final volumes of the reagents used are: BigDye 4 µl, PCR product 4 µl, 5X buffer 4 µl, water 7 µl, and primer 1 µl in a 20 µl reaction. Sequencing reaction protocol: Precipitation - for each of the samples add 100% ethanol 60 µl, 125uM EDTA 1 µl, and 3M Sodium acetate 1 µl. Spin at room temperature for 40 min at 4000 rpm, discard and then remove residue by spinning for 2 min at 400 rpm; Wash - for each of the samples add 70% ethanol 200 µl and spin for 10 min at 4000 rpm. Discard, wash and submit pellet for analysis

9.2.3 PCR-RFLP assay for analysis of C521T polymorphism

Restriction fragment length polymorphism (RFLP) was used to identify SNPs that alter or create a site where a restriction enzyme cuts. C521T genotypes were determined by means of PCR and RFLP analysis for the samples which could not be confirmed through sequencing. In total, 14 samples of known genotype were used as controls and 20 test samples whose genotype was ambiguous from the chromatogram were analyzed (see Appendix 10). The C521T polymorphism alters a *Nla* III restriction site. Genomic DNA was subjected to standard PCR (PCR conditions were 94°C for 30 s, 65°C for 30 s, 72°C for 30 s, 35 cycles) using primers lDpl-rflp, and rDpl-rflp. The PCR product was digested with *Nla* III (10 µl of PCR product, 0.5 µl *Nla* III, 1.5 µl NEB4 buffer, 0.15 µl BSA, 2.85 µl water) for 4 hours at 37°C and samples were electrophoresed on a 2.5% agarose gel.

9.2.4 Statistical analysis

Sequencher 4.7 (http://www.genecodes.com/) was used to analyze the chromatograms obtained from sequencing reactions. To reduce the risk of missing novel SNPs, especially as the dataset was relatively small, a manual approach was used for SNP screening. A reference sequence for the ORF and promoter regions was used to align the sequences (NM_012409 and NW_927317.1, respectively). The option “Matching Bases as Dashes” was used to screen for any mutations/SNPs. The regions with mutation/SNPs were further analyzed visually to rule out any sequencing errors. The criterion for identifying homozygous or heterozygous is shown in Figure 9-3. Further visual analysis of the aligned chromatograms was made to detect any mutations not reported by the first approach.

The results of the genotyping were analyzed by Chi-square test and Fischer's Exact Test (FET) (GraphPad Prism, 4.0) on each of the control and infertile groups separately to determine if each SNP was in Hardy-Weinberg Equilibrium. The allele frequencies in the control and infertile populations were compared by FET for association with infertility. Probabilities of haplotypes and their frequencies were calculated using PHASE program version 2 (Stephens and Donnelly 2003; Stephens and Scheet 2005). Chi-square test was used to determine the *p*-value of the haplotypes for promoter and ORF regions separately. Population frequencies were analyzed from the HAPMAP (http://www.hapmap.org/) project for some of the known SNPs.

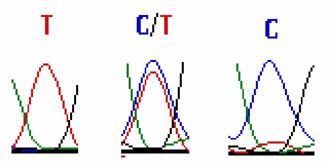


Figure 9-3 Snapshot from Sequencher showing region homozygous for T and C and heterozygous C/T at the same position from different sequences.

9.3 Results

9.3.1 Analysis of promoter and ORF for infertility-associated mutations

Direct sequencing of the PCR product identified four nucleotide substitutions in the promoter and five nucleotide substitutions in the coding region (See Appendix 9 for SNPs in each sample tested). The only sequence variant that was unique to the infertile population was C210T (rs34966363) in the ORF, which was present in only 1 individual. This SNP has a reported frequency of 0.013 in African Americans (Appendix 11) and causes no change in amino acid sequence (Table 9-2). It is therefore not likely to be a cause of infertility, but rather a low frequency polymorphism. Therefore, no mutations in *PRND* were identified which might cause infertility in these men.

9.3.2 Analysis of SNPs in *PRND*

9.3.2.1 SNPs in promoter

The region corresponding to -282 from the TSS covering the core *PRND* promoter was analyzed for SNPs. Four SNPs: G-259A, G-206A, C-202A, and G-171A were identified in this region (Figure 9-4(a)). SNPs at position -259 and -206 are reported in the ENSEMBL dbSNP (the number representing the SNPs for promoter is the relative position of the nucleotide from the TSS) (Table 9-2). SNPs at position -202 and -171 are novel and have a minor allele frequency of 1.6% and 2.6%, respectively, a likely reason for them not having been reported previously. The SNP G-171A corresponds to the 3' end of the USF-1-binding site. The G-259A SNP also shows large frequency variation in different ethnic groups with minor A allele frequency ranging from 0.07 in African Americans to 0.44 in Japanese (Appendix 11). Again, the variation in the allele frequencies between control and infertile samples may be related to the ethnic composition of the control and infertile samples.

9.3.2.2 SNPs in ORF

All five SNPs (C77T, C167T, C210T, C521T, and G522A) identified in the ORF region were previously reported (Peoc'h et al. 2000). The SNP at nucleotide position 210 was found only in one infertile sample and the SNP at nucleotide position 167 was found only in 2 control samples (Table 9-2). All of the SNPs except at nucleotide position 210 and 522 produce a change in the amino acid residue (Table 9-2).

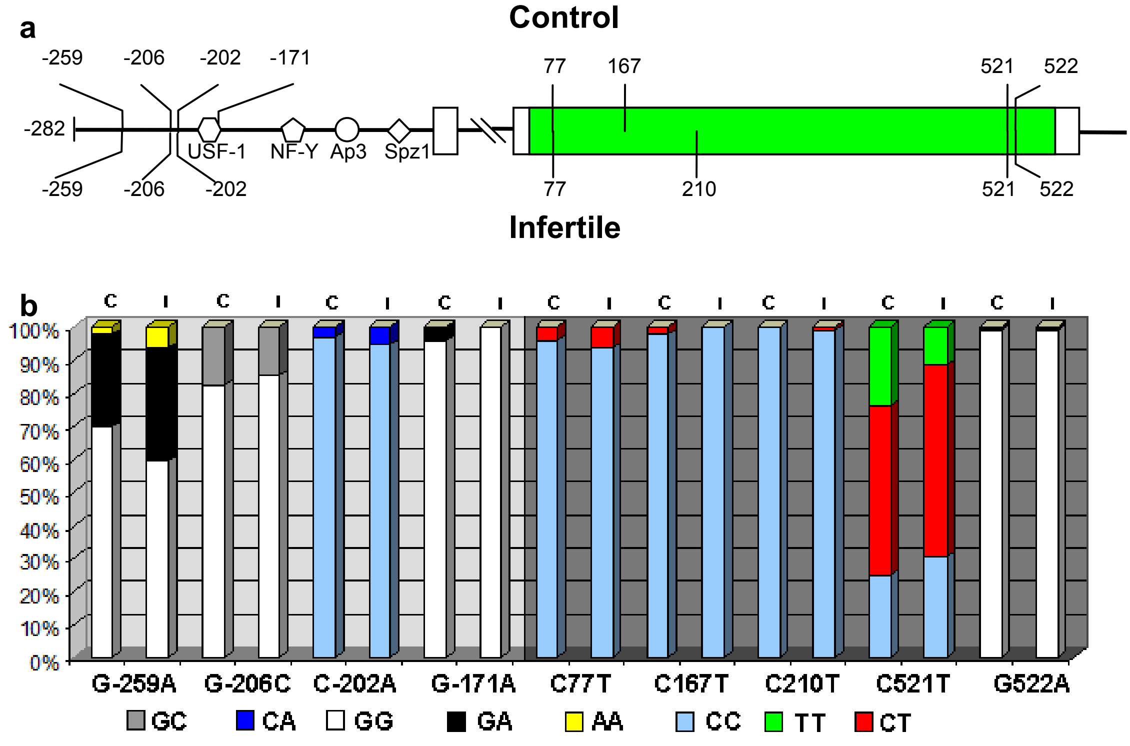


Figure 9-4 SNPs in ORF and promoter among control and infertile samples. (a) Distribution of SNPs shown on a schematic representation of human *PRND* gene. The four TFBSs in the core promoter are also shown. (Figure not drawn to scale). (b) Genotype comparison of control (C) and infertile (I) samples. The genotypes of all 9 SNPs of control and infertile were compared. Results are given in percent..

Table 9-2 Genotype, allele frequencies and Fisher's Exact Test (FET) for nine different SNPs. in control and infertile human samples. The frequency range for the reported alleles (dbSNP or published data) corresponds to the minor allele of control samples (See **Appendix 11** for detailed information about the source of data). ENSEMBL dbSNP ID is shown for some of the SNPs.

SNP	Genotype			Allele frequency		FET	Reported allele frequency range	dbSNPID/ Comment
G-259A	**GG**	**GA**	**AA**	**G**	**A**		A: 0.07-0.44	rs6133157
Control	67	27	2	0.84	0.16	0.095		
Infertile	57	33	6	0.77	0.23			
G-206C	**GG**	**GC**	**CC**	**G**	**C**		C:0-0.22	rs12481509
Control	79	17	0	0.91	0.09	0.708		
Infertile	82	14	0	0.93	0.07			
C-202A	**CC**	**CA**	**AA**	**C**	**A**			
Control	93	3	0	0.98	0.02	0.723		
Infertile	91	5	0	0.97	0.03			
G-171A	**GG**	**GA**	**AA**	**G**	**A**			3' end of USF-binding site
Control	92	4	0	0.98	0.02	0.123		
Infertile	96	0	0	1	0			
C77T	**CC**	**CT**	**TT**	**C**	**T**		T: 0-0.04	M26T
Control	92	4	0	0.98	0.02	0.751		
Infertile	90	6	0	0.97	0.03			
C167T	**CC**	**CT**	**TT**	**C**	**T**		T: 0-0.02	Rs35453518 P56L
Control	94	2	0	0.99	0.01	0.499		
Infertile	96	0	0	1	0			
C210T	**CC**	**CT**	**TT**	**C**	**T**		T: 0-0.01	Rs34966363 Synonymous
Control	96	0	0	1	0	1		
Infertile	95	1	0	0.99	0.01			
C521T	**CC**	**CT**	**TT**	**C**	**T**		T: 0.23-0.65	Rs2245220 M174T
Control	24	49	23	0.51	0.49	0.101		
Infertile	29	56	11	0.59	0.41			
G522A	**GG**	**GA**	**AA**	**G**	**A**			Synonymous
Control	95	1	0	0.99	0.01	1		
Infertile	95	1	0	0.99	0.01			

9.3.2.3 Association analysis

Comparison of the allele frequencies between control and infertile populations by FET found no significant association of any of the SNPs with infertility (Table 9-2). However, suggestive association was seen (*p*~0.1) for SNPs G-259A, G-171A and C521T. Comparison of the genotype was performed for these SNPs (Table 9-3). This analysis revealed there were significantly fewer TT individuals for the C521T SNP in the infertile group than in the control group, that is, individuals homozygous for the T allele were less likely to be infertile (relative risk=0.86; 95% confidence interval= 0.75-0.98; *p*=0.036).

Table 9-3 Genotype analysis of significant SNPs

Genotype	Analysis	*p*-val	Test
G-259A	GG/GA/AA	0.182	Chi^2
	GG/GA+AA	0.174	FET
G-171A	GG/GA+AA	0.121	FET
C521T	CC/CT/TT	0.075	Chi^2
	CC+CT/TT	0.0365	FET,

9.3.3 Analysis of association of SNPs with infertility

The frequency of SNPs in the control and infertile populations were compared (Table 9-2) to look for association between any of these SNPs and infertility. Initial sequencing results found that all SNPs were in Hardy-Weinberg equilibrium, except for SNP C521T. This prompted closer examination of the sequencing traces in this position and it was found that 20 samples had very ambiguous sequencing results for this SNP only. These samples were re-genotyped by PCR-RFLP (Figure 9-5, Appendix 10), and all were unambiguously genotyped; 4 as CC, 9 as CT and 7 as TT. After correctly assigning these genotypes, this SNP was also in Hardy-Weinberg equilibrium. This SNP is also reported in the dbSNP (rs2245220) and showed varied distribution of alleles in different ethnic groups analyzed. The frequency range for the T allele varied from 0.25% in Chinese population to 0.65% in Youroba population (Appendix 11). Earlier studies indicated that there is no association of this polymorphism with Prion or Alzheimer’s disease. The SNPs at nucleotide positions 521 and 522 lead to 3 different combinations for the polymorphic codon 174 which codes for either Threonine (T) or Methionine (M): ACA (T), ACG (T), and ATG (M).

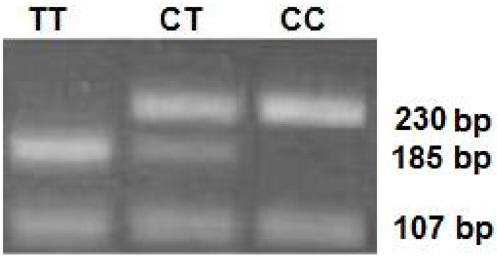


Figure 9-5 PCR-RFLP analysis for the C521T polymorphism. Restriction enzyme *Nla*III was used to digest the PCR product. Homozygosity for thymine is revealed by a 185-bp, 107-bp and 40 (not shown here) bp fragment. Heterozygosity for cytosine and thymine generates 230-bp, 185-bp, 107-bp and 40-bp fragments. Homozygosity for Cytosine generates 185-bp, 107-bp and 40-bp fragments.

9.3.4 Genotype analysis

No significant difference in genotype distribution or allele frequency was observed between the control and infertile populations (Figure 9-4 (b) and Table 9-2). A minor distribution in the allele frequency was observed for the C521T polymorphism with the infertile population showing a lesser allele frequency for T. The allele frequencies of all the identified SNPs in the control and infertile samples are within the range of the population frequencies observed in various ethnic groups (Table 9-2, Appendix 11).

9.3.5 Haplotype analysis

Independent haplotype analysis was performed on promoter and ORF regions. A summary of the different possible haplotypes for control and infertile populations is listed in Table 9-4. No significant association of infertility with the promoter haplotypes was found when considering all 5 promoter haplotypes (*p*=0.11), or when pooling the rare (< 5%) haplotypes (i.e. combining Hap-pr4 and 5, *p*=0.3). Similarly, no significant association with haplotypes of the ORF was found, considering all 6 haplotypes (*p*=0.2) or when pooling the rare haplotypes (i.e. Hap-orf3 – 6, *p*=0.2).

Table 9-4 Distribution of major haplotypes in control and infertile humans. Haplotype analysis was performed independently for ORF and promoter region.

Name	Haplotype	Control	Frequency	Infertile	Frequency
Hap-pr1	GGCG	144	0.750	133	0.693
Hap-pr2	AGCG	31	0.161	45	0.234
Hap-pr3	GCCG	10	0.052	9	0.047
Hap-pr4	GCCA	4	0.021	0	0.000
Hap-pr5	GCAG	3	0.016	5	0.026
Hap-orf1	CCCTG	95	0.495	77	0.401
Hap-orf2	CCCCG	90	0.469	107	0.557
Hap-orf3	TCCCG	4	0.021	6	0.031
Hap-orf4	CTCTG	2	0.010	0	0.000
Hap-orf5	CCCCA	1	0.005	1	0.005
Hap-orf6	CCTTG	0	0.000	1	0.005

9.4 Discussion

The *PRND* promoter and ORF was studied for polymorphisms in the infertile population for the first time. Nucleotide substitutions at four sites in the promoter and five sites in the ORF were identified. Out of the 192 samples screened, only one polymorphism G-171A was found within 185 bp upstream to the TSS which was proposed to be the core promoter region (Nagyova et al. 2004). This observation supports the significance of this region by showing only one polymorphic site. The effect of the SNP G-171A at the 3'end of the USF-1 binding site was analyzed using Match against the TRANSFAC database. The polymorphism did not alter the statistical score for USF-1 and may not affect it binding potential. All of the ORF SNPs found in this study were reported either in the literature or in the HAPMAP project. The C521T polymorphism showed different allele frequencies in the control and infertile population in which the TT genotype was more common in the control than in the infertile population. As this SNP showed a large variation in the allele frequency among different ethnic groups, it cannot be established whether the observed variation has any functional significance, i.e. individuals homozygous for the T allele are less likely to be fertile, or is due to the mixed ethnic population in Australia (ethnic information for the samples is unknown). The C521T polymorphism produces an amino acid change

M174T, which is a part of the C-terminal signal peptide that is cleaved to form a mature protein. However, this amino acid change does not alter the potential of the C-terminal end to act as a signal (as predicted by big-PI). No significant differences between infertile and control were found in the frequency of genotype distribution of other polymorphisms.

In summary, no mutations in the *PRND* gene which are specific to human infertile population were found. The slight variation in the genotype frequency observed between control and infertile population may be related to size of the dataset screened and the ethnic composition of the two test groups.

Summary and Final Remarks

The main objectives of my PhD were to understand the function and evolution of PrP family genes, predicting TFBS involved in the regulation of these genes and to explore a possible association of Dpl with human male infertility. To this end, I applied several computational and experimental methods to identify and characterize the genes of interest in various vertebrate lineages. The results revealed a plethora of interesting information on the evolution of these genes. A dominant chimeric *PRND* transcript encoded by a dicistronic transcript in *Xenopus* species supports the hypothesis of a *PRNP* gene duplication event. The co-existence of monocistronic *PRND* and *PRNP* transcripts, although at a lower tissue concentration, indicates a drive towards evolution of an independent promoter which would enable *PRNP* to evolve a separate specialized function. I also discovered a retained duplicate of *PRNP* designated as *PRNP2* in *X. laevis* which has a tetraploid genome. The pattern of co-existence of dicistronic and monocistronic transcripts was also observed in a marsupial species (*M. domestica)*. Although the chimeric transcript has been reported in mammalian species (mice), it is found at a very minimal level in adult mice brain and is possibly due to a leaky ancestral promoter. In this scenario, it is interesting to note that *PRND* is missing in the chicken genome. Instead, a novel gene with sequence similarities to Sho was discovered downstream and in opposite orientation to *PRNP,* in a similar organization to that in fish. The findings from the various species indicate that the *PRNP* locus is dynamic and rapidly evolving. This is consistent with the observation that when genes are duplicated, one of the genes evolves at a higher rate in an attempt to find a novel function.

The comparative sequence analysis identified conserved residues within PrP and Dpl, and those that are conserved between PrP and Dpl. These residues may be critical in function and maintaining the structural fold. PF revealed conserved TFBSs for *PRNP* and *PRND*. The number of conserved TFs predicted for *PRNP* was much larger compared with that for *PRND*. This may account for the wide range of functions performed by PrP and also to its widespread tissue distribution. One of the most interesting TF-binding predictions for *PRNP* was that of E4BP4 and DBP which are both clock-related proteins and may be involved in the regulation of *PRNP*, which is associated with fatal familial insomnia, in a circadian manner. These predictions provide a good starting point for planning experimental validations and may help in better understand this prion disease. The testis-specific TF Spz1 was predicted in the core promoter of *Prnd*. Spz1 in association with the other experimentally validated ubiquitous TFs, USF-1 and NF-Y, may play a role in testis-specific expression of *Prnd.* Functional studies were performed to validate the significance of the Spz1 binding site. Co-transfection of Spz1 expression plasmid with the luciferase vector containing the core *Prnd* promoter showed a 20-25% increase in the promoter activity. Mutating the same site reduced the activity by about 40%. However, I did not directly test the binding of Spz1 to the predicted binding site. Another conserved TFBS in the core promoter region is that of Ap3. Mutating the Ap3 binding site produced a very minor reduction in promoter activity by about 15%. This binding site may play a more significant role in other cell types.

Prnd knockout mice have been shown to be infertile. Motivated by this observation, I investigated the possible association of SNPs/mutations in *PRND* among the human male infertile population. Genotyping of the core promoter and ORF were performed by the direct sequencing method on 96 infertile male and 96 controls. This identified five already known SNPs in the ORF. None of the four SNPs in the promoter region is in the core binding sites of the known TFBS. Interestingly, some of the control samples had a SNP at the 3'end of the USF-1 binding site (G-171A) which was not detected in the infertile samples, indicating that it may not have a significant effect on the TF binding. However, one of the allele frequencies (C521T) was different among the control and infertile population. Further sequencing of more samples needs to be done to validate the significance of this finding.

Although Sho was predicted to be present in the different species studied, experiments to characterize this gene and comparative sequence analysis were not performed (mainly due to time factors).

During my PhD studies, I have explored various species for prion-protein family genes and performed comparative sequence analysis using a series of computational tools. I hope that the results I have presented highlight the importance of comparative genomics in the analysis of function and evolution of genes. The computational predictions significantly helped in planning the experimental studies. Ever increasing computational power and constant development of new algorithms in computational biology and bioinformatics will open up exciting new capabilities to address complex biological systems, such I have studied.

References

Adachi, J. and M. Hasegawa (1996). "MOLPHY Version 2.3: Programs for molecular phylogenetics based on maximum likelihood." The Institute of Statistical Mathematics, Tokyo,

Japan.

Adams, D. H. (1991). "Does the infective agent of scrapie replicate without nucleic acid? An assessment." Med Hypotheses **35**(3): 253-64.

Aguzzi, A. (2006). Prions of humans and animals.

Alfonso, A., K. Grundahl, et al. (1994). "Alternative splicing leads to two cholinergic proteins in Caenorhabditis elegans." J Mol Biol **241**(4): 627-30.

Alper, T., W. A. Cramp, et al. (1967). "Does the agent of scrapie replicate without nucleic acid?" Nature **214**(5090): 764-6.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bakst, M. R. and B. Howarth, Jr. (1977). "Hydrolysis of the hen's perivitelline layer by cock sperm in vitro." Biol Reprod **17**(3): 370-9.

Barclay, G. R., J. Hope, et al. (1999). "Distribution of cell-associated prion protein in normal adult blood determined by flow cytometry.[see comment]." British Journal of Haematology **107**(4): 804-14.

Baybutt, H. and J. Manson (1997). "Characterisation of two promoters for prion protein (PrP) gene expression in neuronal cells." Gene **184**(1): 125-31.

Bedford, J. M. (1998). "Mammalian fertilization misread? Sperm penetration of the eutherian zona pellucida is unlikely to be a lytic event." Biol Reprod **59**(6): 1275-87.

Behrens, A. (2003). "Physiological and pathological functions of the prion protein homologue Dpl." Br Med Bull **66**: 35-42.

Behrens, A., N. Genoud, et al. (2002). "Absence of the prion protein homologue Doppel causes male sterility." Embo J **21**(14): 3652-8.

Bendheim, P. E., H. R. Brown, et al. (1992). "Nearly ubiquitous tissue distribution of the scrapie agent precursor protein." Neurology **42**(1): 149-56.

Berg, L. J. (1994). "Insights into the role of the immune system in prion diseases." Proc Natl Acad Sci U S A **91**(2): 429-32.

Blanchette, M. and M. Tompa (2003). "FootPrinter: A program designed for phylogenetic footprinting." Nucleic Acids Res **31**(13): 3840-2.

Blumenthal, T. (1998). "Gene clusters and polycistronic transcription in eukaryotes." Bioessays **20**(6): 480-7.

Bolton, D. C., M. P. McKinley, et al. (1982). "Identification of a protein that purifies with the scrapie prion." Science **218**(4579): 1309-11.

Bounhar, Y., Y. Zhang, et al. (2001). "Prion protein protects human neurons against Bax-mediated apoptosis." J Biol Chem **276**(42): 39145-9.

Bray, N., I. Dubchak, et al. (2003). "AVID: A global alignment program." Genome Res **13**(1): 97-102.

Brown, D. R. (2005). "Neurodegeneration and oxidative stress: prion disease results from loss of antioxidant defence." Folia Neuropathol **43**(4): 229-43.

Brown, D. R., K. Qin, et al. (1997). "The cellular prion protein binds copper in vivo." Nature **390**(6661): 684-7.

Brown, D. R., B. Schmidt, et al. (1998). "Effects of copper on survival of prion protein knockout neurons and glia." Journal of Neurochemistry **70**(4): 1686-93.

Brown, D. R., W. J. Schulz-Schaeffer, et al. (1997). "Prion protein-deficient cells show altered response to oxidative stress due to decreased SOD-1 activity." Experimental Neurology **146**(1): 104-12.

Brown, D. R., B. S. Wong, et al. (1999). "Normal prion protein has an activity like that of superoxide dismutase." Biochem J **344 Pt 1**: 1-5.

Brudno, M., C. B. Do, et al. (2003). "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA." Genome Res **13**(4): 721-31.

Budka, H., A. Aguzzi, et al. (1995). "Neuropathological diagnostic criteria for Creutzfeldt-Jakob disease (CJD) and other human spongiform encephalopathies (prion diseases)." Brain Pathol **5**(4): 459-66.

Bueler, H., A. Aguzzi, et al. (1993). "Mice devoid of PrP are resistant to scrapie." Cell **73**(7): 1339-47.

Bueler, H., M. Fischer, et al. (1992). "Normal development and behaviour of mice lacking the neuronal cell-surface PrP protein." Nature **356**(6370): 577-82.

Bueler, H., M. Fischer, et al. (1992). "Normal development and behaviour of mice lacking the neuronal cell-surface PrP protein.[see comment]." Nature **356**(6370): 577-82.

Bueler, H., A. Raeber, et al. (1994). "High prion and PrPSc levels but delayed onset of disease in scrapie-inoculated mice heterozygous for a disrupted PrP gene." Mol Med **1**(1): 19-30.

Cagampang, F. R., S. A. Whatley, et al. (1999). "Circadian regulation of prion protein messenger RNA in the rat forebrain: a widespread and synchronous rhythm." Neuroscience **91**(4): 1201-4.

Calhoun, V. C., A. Stathopoulos, et al. (2002). "Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex." Proc Natl Acad Sci U S A **99**(14): 9243-7.

Calzolai, L., D. A. Lysek, et al. (2005). "Prion protein NMR structures of chickens, turtles, and frogs." Proceedings of the National Academy of Sciences of the United States of America **102**(3): 651-5.

Cashman, N. R., R. Loertscher, et al. (1990). "Cellular isoform of the scrapie agent protein participates in lymphocyte activation." Cell **61**(1): 185-92.

Chapman, M. A., F. J. Charchar, et al. (2003). "Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting." Genomics **81**(3): 249-59.

Chapman, M. A., I. J. Donaldson, et al. (2004). "Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci." Genome Res **14**(2): 313-8.

Chiarini, L. B., A. R. Freitas, et al. (2002). "Cellular prion protein transduces neuroprotective signals." Embo J **21**(13): 3317-26.

Colling, S. B., J. Collinge, et al. (1996). "Hippocampal slices from prion protein null mice: disrupted Ca(2+)-activated K+ currents." Neurosci Lett **209**(1): 49-52.

Colling, S. B., T. M. King, et al. (1995). "Prion protein null mice: abnormal intrinsic properties of hippocampal CA1 pyramidal cells." Brain Res. Assoc. Abstr. **12**((Abstract)).

Collinge, J., M. A. Whittington, et al. (1994). "Prion protein is necessary for normal synaptic function." Nature **370**(6487): 295-7.

Cooper, G. M., M. Brudno, et al. (2003). "Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes." Genome Res **13**(5): 813-20.

Cotto, E., M. Andre, et al. (2005). "Molecular characterization, phylogenetic relationships, and developmental expression patterns of prion genes in zebrafish (Danio rerio)." Febs J **272**(2): 500-13.

Croes, E. A., B. Z. Alizadeh, et al. (2004). "Polymorphisms in the prion protein gene and in the doppel gene increase susceptibility for Creutzfeldt-Jakob disease." Eur J Hum Genet **12**(5): 389-94.

DeLano, W. L. (2002). "The PyMOL Molecular Graphics System." from on World Wide Web http://www.pymol.org.

Dickinson, A. G. and J. T. Stamp (1969). "Experimental scrapie in Cheviot and Suffolk sheep." J Comp Pathol **79**(1): 23-6.

Dodelet, V. C. and N. R. Cashman (1998). "Prion protein expression in human leukocyte differentiation." Blood **91**(5): 1556-61.

Donoghue, M., H. Ernst, et al. (1988). "A muscle-specific enhancer is located at the 3' end of the myosin light-chain 1/3 gene locus." Genes & Development **2**(12B): 1779-90.

Dwek, R. A. (1996). "Glycobiology: Toward Understanding the Function of Sugars." Chem Rev **96**(2): 683-720.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-63.

Eisenhaber, B., P. Bork, et al. (1999). "Prediction of potential GPI-modification sites in proprotein sequences." J Mol Biol **292**(3): 741-58.

Erlich, P., J. Y. Cesbron, et al. (2008). "PrP N-terminal domain triggers PrP(Sc)-like aggregation of Dpl." Biochem Biophys Res Commun **365**(3): 478-83.

Espenes, A., I. Harbitz, et al. (2006). "Dynamic expression of the prion-like protein Doppel in ovine testicular tissue." Int J Androl **29**(3): 400-8.

Ferrer, I., R. Blanco, et al. (2001). "Prion protein expression in senile plaques in Alzheimer's disease." Acta Neuropathol **101**(1): 49-56.

Fiorino, A. S. (1996). "Sleep, genes and death: fatal familial insomnia." Brain Res Brain Res Rev **22**(3): 258-64.

Fischer, M. B., C. Roeckl, et al. (2000). "Binding of disease-associated prion protein to plasminogen." Nature **408**(6811): 479-83.

Ford, M. J., L. J. Burton, et al. (2002). "Selective expression of prion protein in peripheral tissues of the adult mouse." Neuroscience **113**(1): 177-92.

Frazer, K. A., L. Elnitski, et al. (2003). "Cross-species sequence comparisons: a review of methods and available resources." Genome Res **13**(1): 1-12.

Frazer, K. A., L. Pachter, et al. (2004). "VISTA: computational tools for comparative genomics." Nucleic Acids Res **32**(Web Server issue): W273-9.

Frazer, K. A., L. Pachter, et al. (2004) "VISTA: computational tools for comparative genomics." Nucleic Acids Res **Volume**, W273-9 DOI:

Fujisawa, M., Y. Kanai, et al. (2004). "Expression of Prnp mRNA (prion protein gene) in mouse spermatogenic cells." J Reprod Dev **50**(5): 565-70.

Gabriel, J. M., B. Oesch, et al. (1992). "Molecular cloning of a candidate chicken prion protein." Proc Natl Acad Sci U S A **89**(19): 9097-101.

Gasset, M., M. A. Baldwin, et al. (1993). "Perturbation of the secondary structure of the scrapie prion protein under conditions that alter infectivity." Proc Natl Acad Sci U S A **90**(1): 1-5.

Glaser, F., T. Pupko, et al. (2003). "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information." Bioinformatics **19**(1): 163-4.

Golanska, E., K. Hulas-Bigoszewska, et al. (2004). "Polymorphisms within the prion (PrP) and prion-like protein (Doppel) genes in AD." Neurology **62**(2): 313-5.

Gorodinsky, A. and D. A. Harris (1995). "Glycolipid-anchored proteins in neuroblastoma cells form detergent-resistant complexes without caveolin." J Cell Biol **129**(3): 619-27.

Griffiths, M. (1978). "The Biology of the Monotremes." Academic Press: 367.

Habermann, B., A. G. Bebin, et al. (2004). "An Ambystoma mexicanum EST sequencing project: analysis of 17,352 expressed sequence tags from embryonic and regenerating blastema cDNA libraries." Genome Biol **5**(9): R67.

Hadjiagapiou, C., A. Borthakur, et al. (2005). "Role of USF1 and USF2 as potential repressor proteins for human intestinal monocarboxylate transporter 1 promoter.[see comment]." American Journal of Physiology - Gastrointestinal & Liver Physiology **288**(6): G1118-26.

Harmey, J. H., D. Doyle, et al. (1995). "The cellular isoform of the prion protein, PrPc, is associated with caveolae in mouse neuroblastoma (N2a) cells." Biochem Biophys Res Commun **210**(3): 753-9.

Harris, D. A., D. L. Falls, et al. (1991). "A prion-like protein from chicken brain copurifies with an acetylcholine receptor-inducing activity." Proc Natl Acad Sci U S A **88**(17): 7664-8.

Harris, D. A., P. Lele, et al. (1993). "Localization of the mRNA for a chicken prion protein by in situ hybridization." Proc Natl Acad Sci U S A **90**(9): 4309-13.

Helenius, A. and M. Aebi (2001). "Intracellular functions of N-linked glycans." Science **291**(5512): 2364-9.

Herms, J., T. Tings, et al. (1999). "Evidence of presynaptic location and function of the prion protein." J Neurosci **19**(20): 8866-75.

Higgins, D. G. (1994). "CLUSTAL V: multiple alignment of DNA and protein sequences." Methods Mol Biol **25**: 307-18.

Holm, L. and J. Park (2000). "DaliLite workbench for protein structure comparison." Bioinformatics **16**(6): 566-7.

Hope, J., L. J. Morton, et al. (1986). "The major polypeptide of scrapie-associated fibrils (SAF) has the same size, charge distribution and N-terminal protein sequence as predicted for the normal brain protein (PrP)." Embo J **5**(10): 2591-7.

Horiuchi, M., N. Yamazaki, et al. (1995). "A cellular form of prion protein (PrPC) exists in many non-neuronal tissues of sheep." J Gen Virol **76 (Pt 10)**: 2583-7.

Hornshaw, M. P., J. R. McDermott, et al. (1995). "Copper binding to the N-terminal tandem repeat regions of mammalian and avian prion protein." Biochem Biophys Res Commun **207**(2): 621-9.

Hsu, S. H., H. W. Shyu, et al. (2001). "Spz1, a novel bHLH-Zip protein, is specifically expressed in testis." Mechanisms of Development **100**(2): 177-87.

Hsu, S. H., H. W. Shyu, et al. (2001). "Spz1, a novel bHLH-Zip protein, is specifically expressed in testis." Mech Dev **100**(2): 177-87.

Huang, Z., S. B. Prusiner, et al. (1995). "Scrapie prions: a three-dimensional model of an infectious fragment." Fold Des **1**(1): 13-9.

Hughes, M. K. and A. L. Hughes (1993). "Evolution of duplicate genes in a tetraploid animal, Xenopus laevis." Mol Biol Evol **10**(6): 1360-9.

Hutter, G., F. L. Heppner, et al. (2003). "No superoxide dismutase activity of cellular prion protein in vivo." Biol Chem **384**(9): 1279-85.

Infante, J., J. Llorca, et al. (2002). "Polymorphism at codon 174 of the prion-like protein gene is not associated with sporadic Alzheimer's disease." Neurosci Lett **332**(3): 213-5.

Inoue, S., M. Tanaka, et al. (1997). "Characterization of the bovine prion protein gene: the expression requires interaction between the promoter and intron." Journal of Veterinary Medical Science **59**(3): 175-83.

Ironside, J. W. (1998). "Prion diseases in man." J Pathol **186**(3): 227-34.

Jeong, B. H., N. H. Kim, et al. (2005). "Polymorphisms at codons 56 and 174 of the prion-like protein gene (PRND) are not associated with sporadic Creutzfeldt-Jakob disease." Journal of Human Genetics **50**(6): 311-4.

Kel, A. E., E. Gossling, et al. (2003). "MATCH: A tool for searching transcription factor binding sites in DNA sequences." Nucleic Acids Res **31**(13): 3576-9.

Keshet, G. I., H. Ovadia, et al. (1999). "Scrapie-infected mice and PrP knockout mice share abnormal localization and activity of neuronal nitric oxide synthase." J Neurochem **72**(3): 1224-31.

Kingsbury, D. T., D. A. Smeltzer, et al. (1981). "Evidence for normal cell-mediated immunity in scrapie-infected mice." Infect Immun **32**(3): 1176-80.

Kocer, A., M. Gallozzi, et al. (2007). "Goat PRND expression pattern suggests its involvement in early sex differentiation." Dev Dyn **236**(3): 836-42.

Kocisko, D. A., J. H. Come, et al. (1994). "Cell-free formation of protease-resistant prion protein." Nature **370**(6489): 471-4.

Kumar, S. and S. B. Hedges (1998). "A molecular timescale for vertebrate evolution." Nature **392**(6679): 917-20.

Kurschner, C. and J. I. Morgan (1995). "The cellular prion protein (PrP) selectively binds to Bcl-2 in the yeast two-hybrid system." Brain Res Mol Brain Res **30**(1): 165-8.

Li, A., S. Sakaguchi, et al. (2000). "Physiological expression of the gene for PrP-like protein, PrPLP/Dpl, by brain endothelial cells and its ectopic expression in neurons of PrP-deficient mice ataxic due to Purkinje cell degeneration." Am J Pathol **157**(5): 1447-52.

Li, A., S. Sakaguchi, et al. (2000). "Physiological expression of the gene for PrP-like protein, PrPLP/Dpl, by brain endothelial cells and its ectopic expression in neurons of PrP-deficient mice ataxic due to Purkinje cell degeneration." American Journal of Pathology **157**(5): 1447-52.

Li, J. and W. Miller (2002). "Significance Of inter-species matches when evolutionary rate varies." Proceedings of the sixth annual international conference on Computational biology 216 - 224.

Liu, X., D. L. Brutlag, et al. (2001). "Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127–138.

Loots, G. G., I. Ovcharenko, et al. (2002). "rVista for comparative sequence-based discovery of functional transcription factor binding sites." Genome Res **12**(5): 832-9.

Mabbott, N. A., K. L. Brown, et al. (1997). "T-lymphocyte activation and the cellular form of the prion protein." Immunology **92**(2): 161-5.

Mahal, S. P., E. A. Asante, et al. (2001). "Isolation and functional characterisation of the promoter region of the human prion protein gene." Gene **268**(1-2): 105-14.

Makrinou, E., J. Collinge, et al. (2002). "Genomic characterization of the human prion protein (PrP) gene locus." Mamm Genome **13**(12): 696-703.

Mallucci, G. R., S. Ratte, et al. (2002). "Post-natal knockout of prion protein alters hippocampal CA1 properties, but does not result in neurodegeneration." Embo J **21**(3): 202-10.

Manson, J., J. D. West, et al. (1992). "The prion protein gene: a role in mouse embryogenesis?" Development **115**(1): 117-22.

Manson, J. C., A. R. Clarke, et al. (1994). "129/Ola mice carrying a null mutation in PrP that abolishes mRNA production are developmentally normal." Molecular Neurobiology **8**(2-3): 121-7.

Manson, J. C., A. R. Clarke, et al. (1994). "PrP gene dosage determines the timing but not the final intensity or distribution of lesions in scrapie pathology." Neurodegeneration **3**(4): 331-40.

Manson, J. C., J. Hope, et al. (1995). "PrP gene dosage and long term potentiation." Neurodegeneration **4**(1): 113-4.

Manuelidis, L., Z. X. Yu, et al. (2007). "Cells infected with scrapie and Creutzfeldt-Jakob disease agents produce intracellular 25-nm virus-like particles." Proc Natl Acad Sci U S A **104**(6): 1965-70.

Marcotte, E. M. and D. Eisenberg (1999). "Chicken prion tandem repeats form a stable, protease-resistant domain." Biochemistry **38**(2): 667-76.

Masliah, E. (2001). "Recent advances in the understanding of the role of synaptic proteins in Alzheimer's Disease and other neurodegenerative disorders." J Alzheimers Dis **3**(1): 121-129.

Massimino, M. L., J. Ferrari, et al. (2006). "Heterogeneous PrPC metabolism in skeletal muscle cells." FEBS Lett **580**(3): 878-84.

Mastrangelo, P., L. Serpell, et al. (2002). "A cluster of familial Creutzfeldt-Jakob disease mutations recapitulate conserved residues in Doppel: a case of molecular mimicry?" FEBS Lett **532**(1-2): 21-6.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Research **31**(1): 374-8.

McLennan, N. F., K. A. Rennison, et al. (2001). "In situ hybridization analysis of PrP mRNA in human CNS tissues." Neuropathol Appl Neurobiol **27**(5): 373-83.

Mead, S., J. Beck, et al. (2000). "Examination of the human prion protein-like gene doppel for genetic susceptibility to sporadic and variant Creutzfeldt-Jakob disease." Neurosci Lett **290**(2): 117-20.

Mead, S., M. P. Stumpf, et al. (2003). "Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics." Science **300**(5619): 640-3.

Miele, G., A. R. Alejo Blanco, et al. (2003). "Embryonic activation and developmental expression of the murine prion protein gene." Gene Expr **11**(1): 1-12.

Miesbauer, M., T. Bamme, et al. (2006). "Prion protein-related proteins from zebrafish are complex glycosylated and contain a glycosylphosphatidylinositol anchor." Biochem Biophys Res Commun **341**(1): 218-24.

Mitsui, S., S. Yamaguchi, et al. (2001). "Antagonistic role of E4BP4 and PAR proteins in the circadian oscillatory mechanism." Genes Dev **15**(8): 995-1006.

Mo, H., R. C. Moore, et al. (2001). "Two different neurodegenerative diseases caused by proteins with similar structures." Proc Natl Acad Sci U S A **98**(5): 2352-7.

Moore, R. C., I. Y. Lee, et al. (1999). "Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel." J Mol Biol **292**(4): 797-817.

Moore, R. C., I. Y. Lee, et al. (1999). "Ataxia in prion protein (PrP)-deficient mice is associated with upregulation of the novel PrP-like protein doppel." Journal of Molecular Biology **292**(4): 797-817.

Moser, M., R. J. Colello, et al. (1995). "Developmental expression of the prion protein gene in glial cells." Neuron **14**(3): 509-17.

Mouillet-Richard, S., M. Ermonval, et al. (2000). "Signal transduction through prion protein." Science **289**(5486): 1925-8.

Nagyova, J., J. Pastorek, et al. (2004). "Identification of the critical cis-acting elements in the promoter of the mouse Prnd gene coding for Doppel protein." Biochim Biophys Acta **1679**(3): 288-93.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.

Neznanov, N., A. Umezawa, et al. (1997). "A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice." J Biol Chem **272**(44): 27549-57.

Nielsen, H., J. Engelbrecht, et al. (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Protein Eng **10**(1): 1-6.

Nishida, N., S. Katamine, et al. (1997). "Prion protein is necessary for latent learning and long-term memory retention." Cell Mol Neurobiol **17**(5): 537-45.

Nishida, N., P. Tremblay, et al. (1999). "A mouse prion protein transgene rescues mice deficient for the prion protein gene from purkinje cell degeneration and demyelination." Lab Invest **79**(6): 689-97.

Nishimune, Y. and H. Tanaka (2006). "Infertility caused by polymorphisms or mutations in spermatogenesis-specific genes." J Androl **27**(3): 326-34.

Oesch, B., D. Westaway, et al. (1985). "A cellular gene encodes scrapie PrP 27-30 protein." Cell **40**(4): 735-46.

Ohno, S. (1970). Evolution by gene duplication, Springer-Verlag.

Oidtmann, B., D. Simon, et al. (2003). "Identification of cDNAs from Japanese pufferfish (Fugu rubripes) and Atlantic salmon (Salmo salar) coding for homologues to tetrapod prion proteins." FEBS Lett **538**(1-3): 96-100.

Ovadia, H., H. Rosenmann, et al. (1996). "Effect of scrapie infection on the activity of neuronal nitric-oxide synthase in brain and neuroblastoma cells." J Biol Chem **271**(28): 16856-61.

Paisley, D., S. Banks, et al. (2004). "Male infertility and DNA damage in Doppel knockout and prion protein/Doppel double-knockout mice." Am J Pathol **164**(6): 2279-88.

Peoc'h, K., C. Guerin, et al. (2000). "First report of polymorphisms in the prion-like protein gene (PRND): implications for human prion diseases." Neurosci Lett **286**(2): 144-8.

Peoc'h, K., C. Serres, et al. (2002). "The human "prion-like" protein Doppel is expressed in both Sertoli cells and spermatozoa." J Biol Chem **277**(45): 43071-8.

Peoc'h, K., C. Serres, et al. (2002). The human "prion-like" protein Doppel is expressed in both Sertoli cells and spermatozoa. Journal of Biological Chemistry. **277:** 43071-8.

Porter, D. D., H. G. Porter, et al. (1973). "Failure to demonstrate a humoral immune response to scrapie infection in mice." J Immunol **111**(5): 1407-10.

Premzl, M., M. Delbridge, et al. (2005). "The prion protein gene: identifying regulatory signals using marsupial sequence." Gene **349**: 121-34.

Premzl, M., J. E. Gready, et al. (2004). "Evolution of vertebrate genes related to prion and Shadoo proteins--clues from comparative genomic analysis." Mol Biol Evol **21**(12): 2210-31.

Premzl, M., L. Sangiorgio, et al. (2003). "Shadoo, a new protein highly conserved from fish to mammals and with similarity to prion protein." Gene **314**: 89-102.

Prusiner, S. B. (1982). "Novel proteinaceous infectious particles cause scrapie." Science **216**(4542): 136-44.

Prusiner, S. B. (1989). "Creutzfeldt-Jakob disease and scrapie prions." Alzheimer Dis Assoc Disord **3**(1-2): 52-78.

Prusiner, S. B. (1998). "Prions." Proc Natl Acad Sci U S A **95**(23): 13363-83.

Prusiner, S. B., D. C. Bolton, et al. (1982). "Further purification and characterization of scrapie prions." Biochemistry **21**(26): 6942-50.

Prusiner, S. B., D. F. Groth, et al. (1984). "Purification and structural studies of a major scrapie prion protein." Cell **38**(1): 127-34.

Prusiner, S. B., M. Scott, et al. (1990). "Transgenetic studies implicate interactions between homologous PrP isoforms in scrapie prion replication." Cell **63**(4): 673-86.

Pupko, T., R. E. Bell, et al. (2002). "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues." Bioinformatics **18 Suppl 1**: S71-7.

Putta, S., J. J. Smith, et al. (2004). "From biomedicine to natural history research: EST resources for ambystomatid salamanders." BMC Genomics **5**(1): 54.

Qin, K., M. O'Donnell, et al. (2006). "Doppel: more rival than double to prion." Neuroscience **141**(1): 1-8.

Quandt, K., K. Frech, et al. (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." Nucleic Acids Res **23**(23): 4878-84.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.

Rieger, R., F. Edenhofer, et al. (1997). "The human 37-kDa laminin receptor precursor interacts with the prion protein in eukaryotic cells." Nat Med **3**(12): 1383-8.

Riek, R., S. Hornemann, et al. (1996). "NMR structure of the mouse prion protein domain PrP(121-321)." Nature **382**(6587): 180-2.

Riek, R., S. Hornemann, et al. (1997). "NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231)." FEBS Lett **413**(2): 282-8.

Rivera-Milla, E., B. Oidtmann, et al. (2006). "Disparate evolution of prion protein domains and the distinct origin of Doppel- and prion-related loci revealed by fish-to-mammal comparisons." Faseb J **20**(2): 317-9.

Rivera-Milla, E., C. A. Stuermer, et al. (2003). "An evolutionary basis for scrapie disease: identification of a fish prion mRNA." Trends Genet **19**(2): 72-5.

Rossi, D., A. Cozzio, et al. (2001). "Onset of ataxia and Purkinje cell loss in PrP null mice inversely correlated with Dpl level in brain." Embo J **20**(4): 694-702.

Rudd, P. M., M. R. Wormald, et al. (1999). "Roles for glycosylation of cell surface receptors involved in cellular immune recognition." Journal of Molecular Biology **293**(2): 351-66.

Ryou, C. (2007). "Prions and prion diseases: fundamentals and mechanistic details." J Microbiol Biotechnol **17**(7): 1059-70.

Saeki, K., Y. Matsumoto, et al. (1996). "Identification of a promoter region in the rat prion protein gene." Biochem Biophys Res Commun **219**(1): 47-52.

Sakaguchi, S., S. Katamine, et al. (1996). "Loss of cerebellar Purkinje cells in aged mice homozygous for a disrupted PrP gene." Nature **380**(6574): 528-31.

Sakudo, A., D. C. Lee, et al. (2005). "Cell-autonomous PrP-Doppel interaction regulates apoptosis in PrP gene-deficient neuronal cells." Biochemical & Biophysical Research Communications **333**(2): 448-54.

Sali, A. and T. L. Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints." Journal of Molecular Biology **234**(3): 779-815.

Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res **32**(Database issue): D91-4.

Sandelin, A., W. W. Wasserman, et al. (2004). "ConSite: web-based prediction of regulatory elements using cross-species comparison." Nucleic Acids Research **32**(Web Server issue): W249-52.

Sangiorgio, L., G. Gaudenzi, et al. (2007). Analysis of the expression and functional characterization of Shadoo coding gene (sprn) during zebrafish development.

Schachat, F. and M. M. Briggs (2002). "Phylogenetic implications of the superfast myosin in extraocular muscles." Journal of Experimental Biology **205**(Pt 15): 2189-201.

Schmitt-Ulms, G., G. Legname, et al. (2001). "Binding of neural cell adhesion molecules (N-CAMs) to the cellular prion protein." J Mol Biol **314**(5): 1209-25.

Schroder, B., B. Franz, et al. (2001). "Polymorphisms within the prion-like protein gene (Prnd) and their implications in human prion diseases, Alzheimer's disease and other neurological disorders." Hum Genet **109**(3): 319-25.

Scott, M., D. Foster, et al. (1989). "Transgenic mice expressing hamster prion protein produce species-specific scrapie infectivity and amyloid plaques." Cell **59**(5): 847-57.

Sepelakova, J., M. Takacova, et al. (2005). "Involvement of upstream stimulatory factor in regulation of the mouse Prnd gene coding for Doppel protein." Biochim Biophys Acta **1731**(3): 209-14.

Serres, C., K. Peoc'h, et al. (2006). "Spatio-developmental distribution of the prion-like protein doppel in Mammalian testis: a comparative analysis focusing on its presence in the acrosome of spermatids." Biol Reprod **74**(5): 816-23.

Shaked, Y., H. Rosenmann, et al. (1999). "A C-terminal-truncated PrP isoform is present in mature sperm." J Biol Chem **274**(45): 32153-8.

Shmakov, A. N., N. F. McLennan, et al. (2000). "Cellular prion protein is expressed in the human enteric nervous system." Nature Medicine **6**(8): 840-1.

Silverman, G. L., K. Qin, et al. (2000). "Doppel is an N-glycosylated, glycosylphosphatidylinositol-anchored protein. Expression in testis and ectopic production in the brains of Prnp(0/0) mice predisposed to Purkinje cell loss." J Biol Chem **275**(35): 26834-41.

Simonic, T., S. Duga, et al. (2000). "cDNA cloning of turtle prion protein." FEBS Lett **469**(1): 33-8.

Smith, C. J., A. F. Drake, et al. (1997). "Conformational properties of the prion octa-repeat and hydrophobic sequences." FEBS Lett **405**(3): 378-84.

Spielhaupter, C. and H. M. Schatzl (2001). "PrPC directly interacts with proteins involved in signaling pathways." J Biol Chem **276**(48): 44604-12.

Stahl, N., D. R. Borchelt, et al. (1987). "Scrapie prion protein contains a phosphatidylinositol glycolipid." Cell **51**(2): 229-40.

Stephens, M. and P. Donnelly (2003). "A comparison of bayesian methods for haplotype reconstruction from population genotype data." Am J Hum Genet **73**(5): 1162-9.

Stephens, M. and P. Scheet (2005). "Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation." Am J Hum Genet **76**(3): 449-62.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Strumbo, B., S. Ronchi, et al. (2001). "Molecular cloning of the cDNA coding for Xenopus laevis prion protein." FEBS Lett **508**(2): 170-4.

Strumbo, B., L. Sangiorgio, et al. (2006). "Cloning and analysis of transcripts and genes encoding fish-specific proteins related to PrP." Fish Physiology & Biochemistry **32**(4): 339-353.

Suzuki, T., T. Kurokawa, et al. (2002). "cDNA sequence and tissue expression of Fugu rubripes prion protein-like: a candidate for the teleost orthologue of tetrapod PrPs." Biochem Biophys Res Commun **294**(4): 912-7.

Tagle, D. A., B. F. Koop, et al. (1988). "Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." J Mol Biol **203**(2): 439-55.

Tamaki, T., K. Ohnishi, et al. (1995). "Characterization of a GC-rich region containing Sp1 binding site(s) as a constitutive responsive element of the alpha 2(I) collagen gene in human fibroblasts." J Biol Chem **270**(9): 4299-304.

Tanji, K., K. Saeki, et al. (1995). "Analysis of PrPc mRNA by in situ hybridization in brain, placenta, uterus and testis of rats." Intervirology **38**(6): 309-15.

Taylor, J. S., I. Braasch, et al. (2003). "Genome duplication, a trait shared by 22000 species of ray-finned fish." Genome Res **13**(3): 382-90.

Tobler, I., S. E. Gaus, et al. (1996). "Altered circadian activity rhythms and sleep in mice devoid of prion protein." Nature **380**(6575): 639-42.

Tranulis, M. A., A. Espenes, et al. (2001). "The PrP-like protein Doppel gene in sheep and cattle: cDNA sequence and expression." Mamm Genome **12**(5): 376-9.

Tuzi, N. L., E. Gall, et al. (2002). "Expression of doppel in the CNS of mice does not modulate transmissible spongiform encephalopathy disease." J Gen Virol **83**(Pt 3): 705-11.

Ueda, Y., N. Yoshizaki, et al. (2002). "Acrosome reaction in sperm of the frog, Xenopus laevis: its detection and induction by oviductal pars recta secretion." Dev Biol **243**(1): 55-64.

van Rheede, T., M. M. Smolenaars, et al. (2003). "Molecular evolution of the mammalian prion protein." Mol Biol Evol **20**(1): 111-21.

Vollmert, C., O. Windl, et al. (2006). "Significant association of a M129V independent polymorphism in the 5' UTR of the PRNP gene with sporadic Creutzfeldt-Jakob disease in a large German case-control study." J Med Genet **43**(10): e53.

Walawski, K. and U. Czarnik (2003). "Prion octapeptide-repeat polymorphism in Polish Black-and-White cattle." J Appl Genet **44**(2): 191-5.

Wallis, J. W., J. Aerts, et al. (2004). "A physical map of the chicken genome." Nature **432**(7018): 761-4.

Watts, J. C., B. Drisaldi, et al. (2007). "The CNS glycoprotein Shadoo has PrP(C)-like protective properties and displays reduced levels in prion infections." Embo J **26**(17): 4038-50.

Weissmann, C. and E. Flechsig (2003). "PrP knock-out and PrP transgenic mice in prion research." Br Med Bull **66**: 43-60.

Westaway, D., S. J. DeArmond, et al. (1994). "Degeneration of skeletal muscle, peripheral nerves, and the central nervous system in transgenic mice overexpressing wild-type prion proteins." Cell **76**(1): 117-29.

Westaway, D., P. A. Goodman, et al. (1987). "Distinct prion proteins in short and long scrapie incubation period mice." Cell **51**(4): 651-62.

Windl, O., M. Dempster, et al. (1995). "A candidate marsupial PrP gene reveals two domains conserved in mammalian PrP proteins." Gene **159**(2): 181-6.

Wittwer, C. T., G. H. Reed, et al. (2003). "High-resolution genotyping by amplicon melting analysis using LCGreen." Clinical Chemistry **49**(6 Pt 1): 853-60.

Wopfner, F., G. Weidenhofer, et al. (1999). "Analysis of 27 mammalian and 9 avian PrPs reveals high conservation of flexible regions of the prion protein." J Mol Biol **289**(5): 1163-78.

Wormald, M. R. and R. A. Dwek (1999). "Glycoproteins: glycan presentation and protein-fold stability." Structure **7**(7): R155-60.

Wu, K. H., M. L. Tobias, et al. (2003). "Estrogen receptors in Xenopus: duplicate genes, splice variants, and tissue-specific expression." Gen Comp Endocrinol **133**(1): 38-49.

Zhang, X., H. Yang, et al. (2002). "Genomic organization, transcript variants and comparative analysis of the human nucleoporin 155 (NUP155) gene." Gene **288**(1-2): 9-18.

Appendix

Appendix 1: **Sequence information of primers used by Tatiana Vassilieva for the *Xenopus* project.** GSP indicates gene specific primers.

	PRNP, X. tropicalis
5'GSP-R	ACCGAGCATGTAGCCCCCGATAG
5'nested GSP-R	TGGGAGGTTTCCATTGCTTGTTGTTG
3'GSP-F	CCAGATGCCAGAGCGTGTTTACAG
3'nested GSP-F	AGGGAAGAACACCAGCGAGGTAAAC
	PRND, X. tropicalis
5'GSP-R	GGTCACTGCCCCATTGGTAGGCTT
5'nested GSP-R	AGCTCAGATTCCTCGGTAAGGTTAAAG
3'GSP-F	CTTCAGAGGCAGGGCACTCAATGTG
3'nested GSP-F	GGTGCTGGTGGGACTGACGAGTTT
	PRNP1, X. laevis
5'GSP-R	CGACCCACTGCATTACCGAGCAT
5'nested GSP-R	TACTTTTCCCACCACCGCTCTTCTTG
	PRND, X. laevis
5'GSP-R	CTCCTCCGTGCCGGCAGTATCATT
5'nested GSP-R	GCTGGGAGCGCCATGGGTAGTAAA
3'GSP-F	TCCTGTCCTCGGACACCTATTCTTCA
3'nested GSP-F	CTGTACAGATTCCCGGACGGACTTT
	PRNP1, X. laevis
F4	CTCACTGCTTTCCCGATCAC
R4	GTTGCTCCCTGTGTTCCATC
	PRNP2, X. laevis
R6	CCAAGCATGTAGCCTCCAAT
	PRND, X. laevis
F4	CTCACTGCTTTCCCGATCAC
F10	GGAAGAAACAGCAGCAAAGG
R7	GTGCCGGCAGTATCATTCA
	GAPDH, X. laevis
F11	AACTGTCTGGCTCCTCTTGC
R10	ACTTTGCAGGCTTCTTCAGG

Appendix 2: Emu Sholike

>EmuSlike

```
ALFSNTAGGRRGAGGRAGGRGGGGGGGGGGLRGGFRSISRGGNTAGRGSKMASAITAGAAAGYGMGLLGRPRP
PRLGHGPPAQRQPPPAGGFHAAAWPDLGAKRGSPNQAPKGPCGGIVPTLLL
```

Amino acid sequence alignment

```
chickSLike      MRPRDAWCWVAMLLLALFSNTAGGRRGAGGRAGGRGGGGGGGGGGLRGGFRSISRGGNTA 60
emuSlike        ---------------ALFSNTAGGRRGAGGRAGGRGGGGGGGGGGLRGGFRSISRGGNTA 45
                               *********************************************

chickSLike      GRGSKMASAIAAGAATGYGMGLLGRPRPPRLGHGPPAQRQPPPAGGFHAAAWPDLGAKRG 120
emuSlike        GRGSKMASAITAGAAAGYGMGLLGRPRPPRLGHGPPAQRQPPPAGGFHAAAWPDLGAKRG 105
                **********:****:********************************************

chickSLike      SPNQAPKGPCGGIVPTLLLANAICWVNHGM 150
emuSlike        SPNQAPKGPCGGIVPTLLL----------- 124
                *******************
```

Appendix 3: ***M. domestica*** **alternatively spliced 3'end. Grey portion corresponds to the ORF. ORF2 variant underlined. Note the possible splice donors and acceptors boxed. Bases in red background corresponds to exon**

```
GTACCAGAACGAGTACCGCAGTGCTTACAGCGTGGCGTTCTTCTCTGCCCCACCTGTGACCCTCCTCCTCCTC
AGTTTCCTTATTTTCCTGATTGTGAGCTAAGAAGCCTACCAATGTTTACTCTCTTCATGTTTCTTCTCTTAAT
CTTTGCAGAGAAGGAGGTCCTTCTGTCTGCAAGGGCAGCCCAAATAGCAGCAATTTCTCATTTCTATGTTTAT
CTGTCCCCCATAGGTTAAGGCACTAATGAGTACTGGTGAATGTACAGTAGACCCTAGATGCCAGGCCACCACT
CTTCCCCCGAACCATTTTGATCATGCATCCATCAGGGCAATGCCATACTTGTCAGTATCCTTTACAAGAGAGG
AGACCATTAAGTAACTTCTGGTCCATCAAGACACTTCTATAGTATAGCAGATTAAGGCCAAAACAGAAATGAT
TTCAAACTACATTTTCCAAATAGACACAACCATGGGCCCTTTTGCTTCCTGAAATGCCACCTAAATCTTTCTC
CCTGCTTGTATAGTCAATTAATGAGTAGATAAAGAATTAGCTAATTTAGAGCCCCATCTCTTCTTTGGCTTTA
CCAGCTGTGATATCATAGCCAGTTAAATATCTTTAGGAAACTCATTCATACATTTCAATACATCCTGATGCAT
TTTCCTCTTCAAATAGAAATTTTCATCATTAGGAAAGAAAGACTAGAGACCATCTAGGCAAACTATAGAAATC
ATCCGACCATGTGGGTCCAGGCTCGAGCTGCTGAATCAAGCGTAGCTACACCACAGAGAACTGGGACTACTGT
CTTGAATTTTATAAATGGGACACACTCAGTAGAGGCTGCACCAAAAATATATCCGGCCCAGGGGCTGGGTCTA
TTTGTCAACAGTGGAATGAGCATCGAGCATCATTGCGTGAGCTTTATCCTGTGCTACAGAGACTGTTCTTCTG
GGTAACCTACAATTTGGAAAGTAGTATTACTATGGGTATGATTTGTCATCTCAGACCATTTTGGTGTGTACAT
ATTGTTCAATACCAGTGTAAACTATTTCTACCAGAGCATTCCACCCAGGAATATGAAAATGCAAAGCAGATAC
TCCCGGATAACGAAGGAATCCCTCTCCAATGTTAGCAAATAATCCAAGACTGGCAGGCAGCACTTGGCTAGGC
CTCTTTGCCTGGAAGCCCACAAAGGTGCCAGCATTGTTAGAGTTACAGCAGGAGGCTGGTTAGCTCAGAGGAA
TCAGCAACTGGTTAAATTGCAATTTTAGGAATTAATGAGATAGTTTGGTGAGTTTATAAGCAAAAGAAGCACT
TCTTTTTTTCATGTCTCAAGATAAAGCAGTACCTAACTATACTGGATGAGCCCTTAAAGTCAGATTAGAAGAA
GAATCATTCAGATGGAGGAAGCCAGCCTGTTGATTTTTGGATTTAATCCAATCTTCAAAAACAATGTCTCAGA
AAAGGAGGAGAGGAGGGGAGGAGAAGACGGGGAGATGATCTGTCTGTATAAGTGATTTTTGTATTGAGTTGTG
AAAAAGTTGTTCTCATAGAATAAATTATTACTTTTTTCTTGTCACTATTTTGCATTGTGATGGGCTTGCAGTC
TTTGTAATCAGTAAATATCAGATATGATAAACTAGAAATTCTGCTCAGAACAATTCACCCAGAGTAGTATCCA
ATGTATTCTGTGTCTTTCCAAGATCCCAGATGCCCATTCCGTGGAAGCAATATCCTGAGGCTTACGGCACCAC
ATGCTGAATGGACTCCTGCTTGAGTCCCCAGTGCACCAGGGAAAGTAGTCTTTGAAATTTGCATGCACTTAGT
AATATAAAGATGTTTTATAGATTTGTAACTTTGCACGTATTTGTTTTGATGTATTAAAAATTTATAAATGTTT
AATATC
```

Appendix 4: Platypus *PRNP* and *PRND* ORF as obtained from trace sequence archive. Note the GC rich region in *PRND* ORF

```
>PlatypusPrP_ORF 723211856
ATGTGGCCACATTGGGGAAAATCCCCTGTACATCACTGGATAATAGACATCTGTGTGGTACACCTGGAGCGCA
GATGCCGTGGACACCTACACCCAAATCCCTGCCCCGGCGGGAGGTGTGTGCAACAGCAGCCAAACAGATACCC
AGGCCAGCCGGACCACCCCCGGCGGATAGGGTCACCCCCAGAGCGGGGGGGTCCAGATGGGGCCACCCCCAGG
GCGGGGGAGCCAGCTGGGGTCACCCCCAGGGGCGGGGGCTCCAACTGGGGTCATCCGCAGGGCGGGGGGGCCA
GCTGGGGTCACCCCCAGGGCGGGGGCTATAGCAAGTACAAGCCGGACAAGCCCAAGACCGGCATGAAGCACGT
GGCCGGGGCGGCGGCGGCCGGGGCGGTGGTGGGGGGCCTGGGGGGCTACATGATCGGCAGCGCCATGAGCCGG
CCCCCCATGCACTTCGGCAACGAGTTCGAGGACCGCTACTATCGGGAGAACCAGAACCGCTATTCCAACCAGG
TTTACTACAGGCCCGTGGACCAGTACGGCAGCCAGGACGGCTTCGTCCGCGACTGCGTCAACATCACCGTCAC
CCAGCACACCGTCACCACCACCGAGGGGAAGAACCTCAACGAGACCGACGTCAAGATCATGACCCGCGTCGTG
GAGCAGATGTGCGGATCCACCACATGGAACCTCCAGTGGTTCGGAGTAAGTTTCAATTTC
```

```
>PlatypusPrP
MWPHWGKSPVHHWIIDICVVHLERRCRGHLHPNPCPGGRCVQQQPNRYPGQPDHPRRIGSPPERGGPDGATPR
AGEPAGVTPRGGGSNWGHPQGGGASWGHPQGGGYSKYKPDKPKTGMKHVAGAAAAGAVVGGLGGYMIGSAMSR
PPMHFGNEFEDRYYRENQNRYSNQVYYRPVDQYGSQDGFVRDCVNITVTQHTVTTTEGKNLNETDVKIMTRVV
EQMCGSTTWNLQWFGVSFNF
```

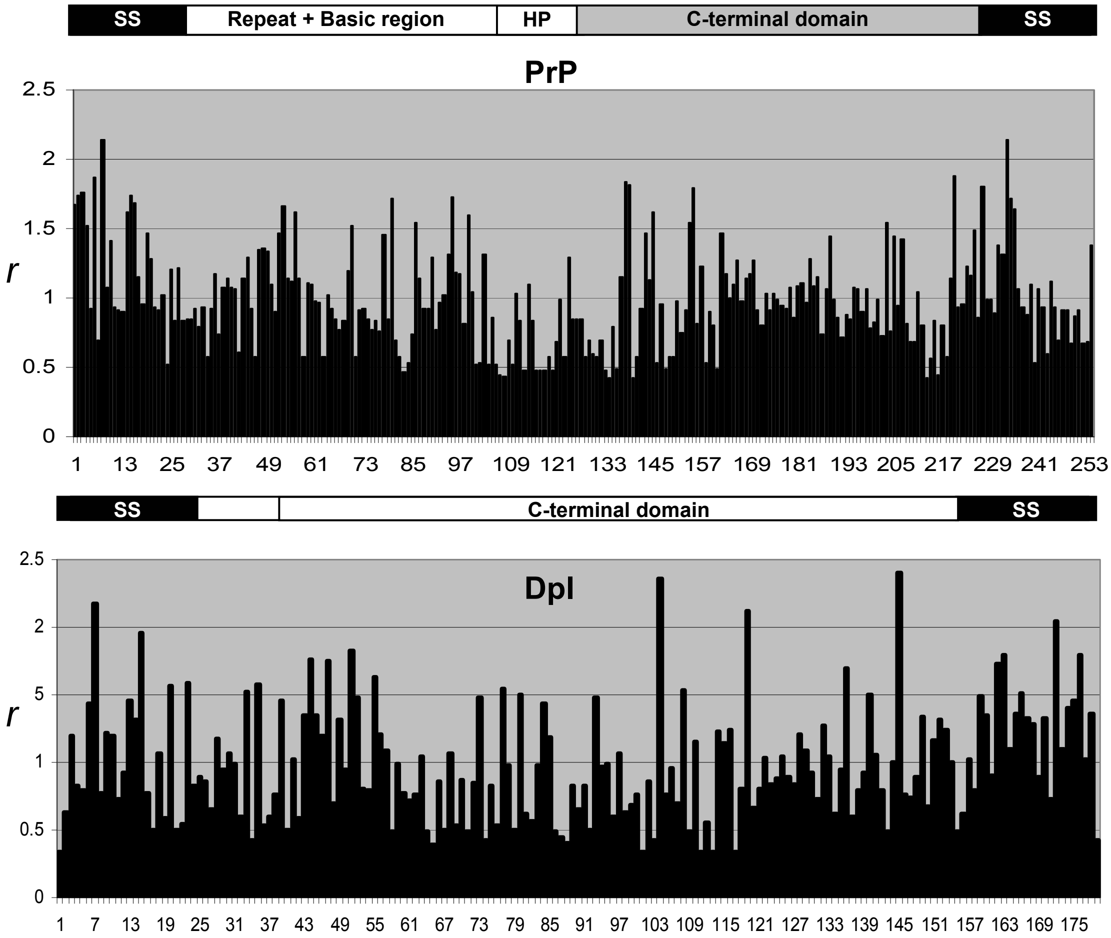
```
>PlatypusDpl
ATGATGACGGTGAGGAGGAGGAGGAGGAGCGGAGGAGCCCGGTGGCTCCTGGTCTTCCTGGTCCTGCTGAGCG
GCGACCTGTCCTCCCTCCAGGCTCGGGGGCCGAGGCCGAGGAACAAGGCCGGCCGGAAACCCCCCCCGTCCAA
CGCCGGGCCCGACTCTCCGGCCCCCCGGCCCCCGGCGGAGCCCGGGGGACTTTCATCCGGCGAGGCGGGAGG
CTTTCCGTCGATTTCGGGCCCGAGGGCAACGGCTACTACCAGGCCAACTACCCGCTCTTGCCCGACGCCATCG
TCTACCCGGACTGCCCGACGGCCAACGGGACCAGAGAGGCCTTCTTCGGGGACTGCGTCAACGCCACCCACGA
GGCCAACCGGGGCGAGCTGACGGCCGGCGGGAACGCCAGCGACGTCCACGTCCGGGTGCTCCTCAGGCTGGTC
GAAGAACTCTGCGCCCTCCGGGACTGCGGCCCGGCGCTCCCGACGGGGCCGGCGCCGCGGCCCGGACCGCCGG
GCCCGCCCGCCGCGCTCGCCCTGCTGACCCTCGTCCTCCTCGGGGCCCAGTGA
```

```
>PlatypusDpl
MMTVRRRRRSGGARWLLVFLVLLSGDLSSLQARGPRPRNKAGRKPPPSNAGPDSPAPRPPAGARGTFIRRGGR
LSVDFGPEGNGYYQANYPLLPDAIVYPDCPTANGTREAFFGDCVNATHEANRGELTAGGNASDVHVRVLLRLV
EELCALRDCGPALPTGPAPRPGPPGPPAALALLTLVLLGAQ
```

Appendix 5: Human *PRNP* EST sequences (GenBank gi) used for the analysis

45857930, 34889317, 15490243,13980625, 13994109, 15494325, 13976124, 15495677, 19370296, 34479361, 15440431, 13967872, 15492735, 15440704, 5433432, 15490584, 15580878, 13976743, 15493048, 15434125, 15583207, 15438103, 15440700, 15490764, 15496140, 13976202, 66264383, 15583400, 11002886, 15494802, 22705803, 47372420, 21857720, 46921643, 46925984, 51481408, 10202561, 14001854,45749230, 22285044, 22271655, 22659989, 31446754, 45751517, 22697249, 45703565, 21855704, 15433349, 18520023, 20405989

Appendix 6: **The rate of evolution (*r*) at each of the amino acid site in PrP and Dpl. The different region of the protein sequence is shown on top of the graph. (SS- signal sequence, HP- Hydrophobic region).**



Appendix 7: **Known information about some of the TFs of interest**

Transcription factor	Important function	Information
CDXA1: caudal type homeobox transcription factor 1	Early expression in developing organs derived from the three germ layers, and a late expression confined to organs of endodermal origin. Cdx-1 and Cdx-2 homeobox genes belong to the regulatory network that controls intestinal development and homeostasis. They participate in the definition of positional information along the intestinal A–P axis, they are involved in the regulation of the continuous renewal of the digestive epithelium, and they are key actors of the reciprocal epithelial–mesenchymal cell interactions.	
CIZ: Cas-associated zinc finger protein	Potential role in the regulation of neurodevelopment or neuroplasticity	
DBP: D site albumin promoter binding protein **Classification:** bZIP (proline and acidic amino acid-rich basic leucine zipper)	Show high-amplitude circadian expression in the suprachiasmatic nucleus, the master circadian pacemaker in mammals. They are expressed at nearly invariable levels in most brain regions, in which clock gene expression only cycles with low amplitude. Show higher amplitude circadian cycles of expression in liver than in brain.	
E4BP4: Protein factor encoded by lambda-P4 which binds to E4 promoter **Classification:** bZIP	Diverse range of processes including commitment to cell survival versus apoptosis, the anti-inflammatory response and, most recently, in the mammalian circadian oscillatory mechanism (central circadian clocks reside in several neuronal tissues such as the suprachiasmatic nucleus (SCN), pineal gland and retina). E4BP4 appears to act antagonistically with members of the related PAR family of transcription factors with which it shares DNA-binding specificity.	**Interaction:** Competitive binding site with PAR **Transcriptional activity:** Repression
ETS: E26 transformation-specific	Ets-1 is transcriptionally up-regulated by H2O2 via an antioxidant response element in tumor cell lines. Results suggest that Ets-1 might play an important role in carcinogenesis and/or the progression of human prostatic carcinomas. Ets-1 expression increases the transformed phenotype of HeLa cells, by promoting cell migration, invasion and anchorage-independent growth, while Ets-1 downregulation reduces cell attachment. Ets-1 regulated angiogenesis through the induction of angiogenic growth factors	
FAC1: fetal Alzheimer antigen / fetal Alz-50 reactive clone 1 **Classification:** Zinc	Gene identified in brain homogenates from patients with Alzheimer's disease. High levels of FAC1 were detected in fetal brain and in patients with neurodegenerative diseases.	

finger domain		
GATA1: Globin binding protein 1 **Classification:** Diverse Cys4 zinc fingers	The protein plays an important role in erythroid development by regulating the switch of fetal hemoglobin to adult hemoglobin. Mutations in this gene have been associated with X-linked dyserythropoietic anemia and thrombocytopenia. Results are consistent with GATA1 regulating some but not all pathways of platelet activation. A multiprotein complex containing GATA-1, Oct-1, and other protein factors may contribute to the formation of a repressive chromatin structure that silences gamma-globin gene expression. GATA-1 has a role in erythropoiesis and megakaryocytopoiesis. GATA1 is likely to play a critical role in the etiology of myeloproliferative disorder and Down syndrome acute megakaryoblastic leukemia, and mutagenesis of GATA1 represents a very early event in DS myeloid leukemogenesis. Roles of hematopoietic transcription factors GATA-1 and GATA-2 in the development of red blood cell lineage.	**Interaction:** OCT1, Sp1
GEN_INI2: General Initiator2		
HMGIY: High mobility group AT-hook 1	Encoded protein preferentially binds to the minor groove of A+T-rich regions in double-stranded DNA. Might regulate lymphoid differentiation. Loss of Hmga1 expression, induced in mice by disrupting the Hmga1 gene, largely impaired insulin signaling and severely reduced insulin secretion, causing a phenotype characteristic of human type 2 diabetes. HMGA1 functions as a transcriptional enhancer co-activator in B cells through indirect association with DNA. HMGA1 as one of the first mediators in the development of human atherosclerotic plaques. Regulated dynamic properties of HMGA1a fusion proteins indicate that HMGA1 proteins are mechanistically involved in local and global changes in chromatin structure. HMGI-Y physically interacts with Sp1 and C/EBP beta and facilitates the binding of both factors to the insulin receptor promoter.	
HOXA4 Homeo box A4 **Classification:** Homeo domain	Expression of these proteins is spatially and temporally regulated during embryonic development. Involved in developmental and organogenesis.	
LEF1: lymphoid enhancer-binding factor 1	Expressed in pre-B and T cells. LEF-1 is abundantly expressed in human tumors.	
MAF: v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)	High levels of c-maf mRNA is associated with multiple myeloma	
MAZ: MYC-associated zinc finger protein		

MEF2: myocyte enhancer factor 2 **Classification:** MADS box (MCM1, AGAMOUS, DEFICIENS and SRF)	Play a key role in the differentiation of muscle tissues and are important in the muscle-specific expression of a number of genes. Plays an important role in cardiac muscle development is the MEF2 protein family. MEF2 regulated neuronal survival by stimulating MEF2-dependent gene transcription. The protein, myocyte enhancer factor 2 (MEF2), turns on and off genes that control dendritic remodeling. That MEF2 activation leads to the inhibition of synapse formation, makes sense in light of what is known about the nervous system. In memory and learning, as well as development, activity leads to a sculpting, or cutting away, of synapses. What may be more surprising is the way activity causes MEF2 to switch from repressor to activator.	
MYB: v-myb myeloblastosis viral oncogene homolog **Classification:** Tryptophan clusters	The c-myb proto-oncogene product (c-Myb) regulates both the proliferation and apoptosis of hematopoietic cells by inducing the transcription of a group of target genes. c-Myb activity is regulated during the cell cycle in hematopoietic cells. Involvement of c-myb in the regulation of intestinal nutrient absorption. Activation of c-MYC and c-MYB proto-oncogenes is associated with decreased apoptosis in tumor colon progression. c-Myb activity is regulated during the cell cycle in hematopoietic cells. It is generally believed that Myb proteins, including MybA and MybB, two additional vertebrate Myb proteins that are related to c-*myb*, play roles in the cell division cycle. c-*myb* and MybB have been implicated in the G1/S transition, whereas MybA is more likely to be involved in cellular differentiaton.	**Interaction:** GATA1 **Transcriptional activity:**
NF1: Nuclear factor 1 **Classification:** Basic domain		
OCT1 Octamer-binding transcription factor 1 **Classification:** POU domain factors	Oct-1 modulates the activity of genes important for the cellular response to stress. Oct-1 is widely expressed in adult tissues and is the only known POU family member not expressed in a specific temporal or spatial pattern. The Oct-1 transcription factor regulates a variety of tissue-specific and general housekeeping genes by recruiting specialized coactivators of transcription.	
PAX2: Paired box Gene2 **Classification:** Paired box domain	PAX2 is believed to be a target of transcriptional suppression by the tumor supressor gene WT1. Mutations within PAX2 have been shown to result in optic nerve colobomas and renal hypoplasia. Over expression of Pax2 is associated with apoptosis resistance and angiogenesis favoring renal tumor growth. Pax2 protein regulates expression of secreted frizzled related protein 2. The PAX2 gene was frequently expressed in a panel of 406 common primary tumor tissues and endogenous PAX gene expression is often required for the growth and survival of cancer cells. PAX2 has a role in urogenital tract development and disease	
PBX1: pre B-cell leukemia transcription factor 1	Data suggest that Pbx1 acts together with multiple Hox proteins in the development of the caudal pharyngeal region, but that some functions of Hox proteins in this region are Pbx1-independent. Early requirement for Pbx1 in urogenital development. is an essential regulator of mesenchymal function during renal morphogenesis. Development of pancreas.	**Interaction:** Hox proteins

Spz1: spermatogenic Zip 1	Highly expressed in adult testis; may play an important role in spermatogenesis and fertility in males in the humans. Spz1 has a regulatory role during spermatogenesis.	**Interaction:** PP1cgamma2
STAT1: Signal transducer and activator of transcription 1	This protein mediates the expression of a variety of genes, which is thought to be important for cell viability in response to different cell stimuli and pathogens. These results imply that STAT-1 plays a crucial role in the DNA-damage-response by regulating the expression of 53BP1 and MDC1, factors known to be important for mediating ATM-dependent checkpoint pathways. STAT1 plays an important role in the regulation of erythropoiesis. Cells lacking STAT-1 show reduced apoptosis in response to heat or ischaemia. Expression of STAT-1 in these cells does not enhance cell death but restores sensitivity to stress-induced death.	
STAT4: signal transducer and activator of transcription 4	Expression of Stat4 in connective tissue-type mast cellss plays an important role on Th1 immune responses. STAT4 is required for the generation of an effective innate host defense against bacterial pathogens of the lung. Play a role in Immunity and defense. STAT4 appears to be a critical transcription factor in defence against mycobacterial infection	**Interaction:** Ap-1
STAT6: signal transducer and activator of transcription 6	STAT6 plays a protective role against hemodynamic stress in hearts. Role of STAT6 in late phase of allergic responses of mast cells	
YY1 **Classification:** Kruppel class of zinc finger proteins	The protein is involved in repressing and activating a diverse number of promoters. Contributes to vascular smooth muscle proliferation and differentiation in normal pulmonary artery development. YY1 may play a role in prostate cancer development. YY1 is involved in a positive feedback loop during apoptosis.	**Interaction:** Ap2 Sp1 GATA4

Appendix 8: Primer map for mouse promoter and ORF. Bases highlighted in green correspond to UTR and bases in bold and underlined correspond to ORF.

lprDpl
CTTGCCCTCTTTTTGAGCTGAACCTCCCCACAGAGGCTGTCAGCAAAGACTGCTTTGCTGTCCAAGGCCGGCC
TTTAGGTACAGGACACTATGCATGTCCAGCGAGGCTTGTTGAAGCCACATCTGTCATAGATATTGATGGGAAA
AGCATTTTCCTTCTAACACATCTCATTCCAATTCTAAAAGGCACCTCTGAAGCCTTGCTGAACTTCATCAAGA
TTTTCACGTGGTTTCCTTAGTAAAGTGTGATGAGAAGGTCCATCCTTCTCAGGATGAAGGAGTGGTCCAGGAA
GCCCTGATTGGTCTGCCGGGGAGGGAAGGGCTGCCTTATTTGGAGACCTGCAGGAATGCCACCTCCCCCGGCA
Exon1
GCTCCTATATAGCTGGGCGGACCTGGCTGCCAAGAGGGTGTGCTGGGGGACTGTGCAGCTCGAGGCTCCAGAG
rprDpl **Intron1**
GCACACT**CCAGAGAGAGCCAAGGTACG**TGGGGG

---------------------intronic region removed for brevity -----------------
----

lorfDpl **lorfseqDpl**
TAGCAAAGGAGCTCGGTGTTTGAGTTAACCCTGCACAACCCAAACATGGGGAAACAATTATGCTTTTGAGACC
ACATAAATAGCACAAGGATGCGATTCCTTCCTTAAAATCTCCTGCACTTGGGAGGGGGCAGGGGAGCCCAGGC
Exon2 ORF
AGGCCTGGTGGGGAGCTGACCCACCGCCGTTTCTCTGGCAGGTTCTGACGCG**ATGAGGAAGCACCTGAGCTGG**
TGGTGGCTGGCCACTGTCTGCATGCTGCTCTTCAGCCACCTCTCTGCGGTCCAGACGAGGGGCATCAAGCACA
GAATCAAGTGGAACCGGAAGGCCCTGCCCAGCACTGCCCAGATCACTGAGGCCCAGGTGGCTGAGAACCGCCC
GGGAGCCTTCATCAAGCAAGGCCGCAAGCTCGACATTGACTTCGGAGCCGAGGGCAACAGGTACTACGAGGCC
AACTACTGGCAGTTCCCCGATGGCATCCACTACAACGGCTGCTCTGAGGCTAATGTGACCAAGGAGGCATTTG
lDplrflp
TCACCGGCTGCATCAATGCCACCCAGGCGGCGAACCAGGGGGAGTTCCAGAAGCCAGACAACAAGCTCCACCA
GCAGGTGCTCTGGCGGCTGGTCCAGGAGCTCTGCTCCCTCAAGCATTGCGAGTTTTGGTTGGAGAGGGGCGCA
GGACTTCGGGTCACCATGCACCAGCCAGTGCTCCTCTGCCTTCTGGCTTTGATCTGGCTCACGGTGAAATAAG
rorfDpl
CTTGCCAGGAGGCTG**GCAGTACAGAGTGCAGCAGC**GAGCAAATCCTGGCAAGTGACCCAGCTCTTCTCCCCCA
AACCCACGCGTGTTCTGAAGGTGCCCAGGAGCGGCGATGCACTCGCACTGCAAATGCCGCTCCCACGTATGCG
rDplrflp
CC**CTGGTATGTGCCTGCGTTCTGA**

Appendix 9: Detailed information of SNPs in the promoter and ORF region for human (a) infertile and (b) control samples. The sequence identifiers that are color coded show more than one SNP.

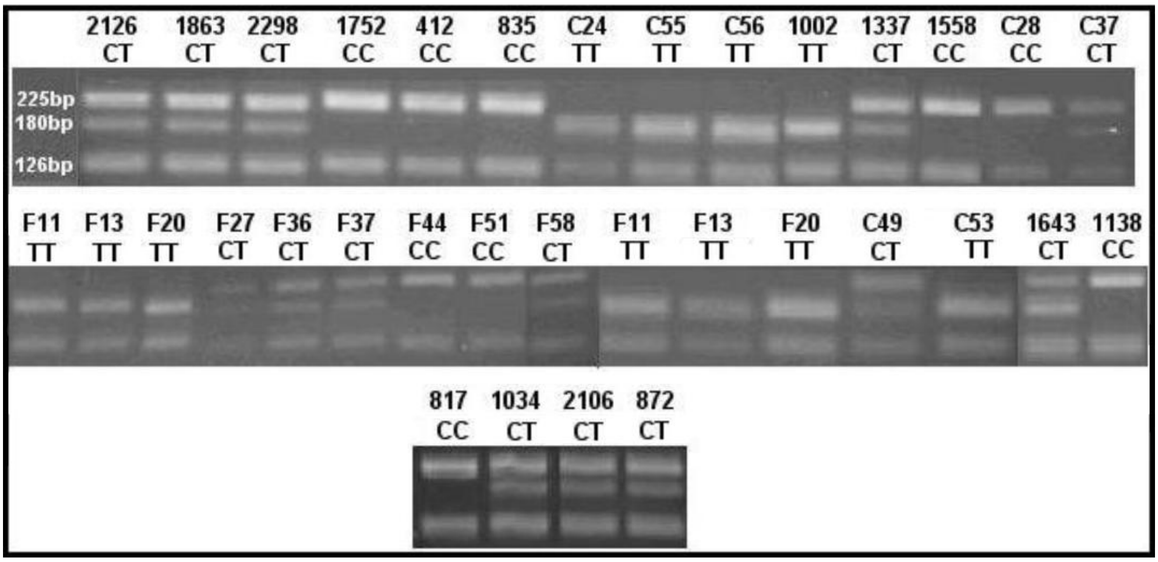
a

Infertile	Promoter				ORF			
Majority	G/G	G/G	C/C		C/C	C/C	C/T	G/G
SNP	**G/A**	**G/C**	**C/A**		**C/T**	**C/T**	**C/T**	**G/A**
Position	-259	-206	-202		77	210	521	522
	678						$678^{C/C}$	
	2362						$2362^{T/T}$	
	751						$751^{C/C}$	
	2245						$2245^{T/T}$	
	835						$835^{C/C}$	
	$2039^{A/A}$						$2039^{C/C}$	
	$866^{A/A}$						$866^{C/C}$	
	1643	1643						
	1466						$1466^{C/C}$	
	1616	1616						
	1467						$1467^{T/T}$	
	1002						$1002^{T/T}$	
	1658						$1658^{C/C}$	
	2016					2016		
	1558						$1558^{C/C}$	
	1559	1559	1559		1559		$1559^{C/C}$	
	2070	2070						
	2107						$2107^{T/T}$	
	1406						$1406^{C/C}$	1406
		2184	2184		2184			
		412	412		412		$412^{C/C}$	
		2064			2064		$2064^{C/C}$	
		920	920		920		$920^{C/C}$	
		1827	1827		1827		$1827^{C/C}$	
		1251					$1251^{T/T}$	
		1752					$1752^{C/C}$	
	2331	2184						
	817	648						
	756	1019						
	2080	1102						
	1378							
	2297							
	2329							
	975							
	1270							
	806							
	$1183^{A/A}$							
	821							
	1647							
	1783							
	1529							
	1138							
	2034							
	1672							
	1806							
	$1631^{A/A}$							

b

Control	Promoter					ORF			
Majority	G/G	G/G	C/C	G/G		C/C	C/C	C/T	G/G
SNP	**G/A**	**G/C**	**C/A**	**G/A**		**C/T**	**C/T**	**C/T**	**A/G**
Position	-259	-206	-202	-171		77	167	521	522
	fmc53							fmc53$^{T/T}$	
	cc36							cc36$^{T/T}$	
	cc46							cc46$^{C/C}$	
	FMC54$^{A/A}$							fmc54$^{C/C}$	
	cc42							cc42$^{C/C}$	
	cc6							cc6$^{C/C}$	
	cc28	cc28		cc28				cc28$^{C/C}$	
	cc55							cc55$^{T/T}$	
	cc59$^{A/A}$							cc59$^{T/T}$	
	fmc47							fmc47$^{T/T}$	
	fmc1	fmc1						fmc1$^{C/C}$	
	cc24							cc24$^{T/T}$	
	fmc8							fmc8$^{T/T}$	
	cc19	cc19						fmc57$^{C/C}$	
		fmc57							
		fmc39		fmc39					
		fmc13		fmc13					
		cc30						cc30$^{T/T}$	
		fmc27	fmc27			fmc27			
		fmc14	fmc14			fmc14			
		fmc60						fmc60$^{T/T}$	
		fmc10						fmc10$^{C/C}$	
		fmc18	fmc18			fmc18			
						cc21		cc21$^{C/C}$	
							cc9	cc9$^{T/T}$	
								cc12$^{C/C}$	cc12
	fmc61	fmc9				cc44	fmc42	cc17$^{T/T}$	
	fmc37	cc8						cc56$^{T/T}$	
	cc37	cc15						cc33$^{T/T}$	
	cc16	fmc23						cc29$^{C/C}$	
	fmc32							cc23$^{C/C}$	
	fmc56							cc50$^{C/C}$	
	cc58							cc7$^{T/T}$	
	fmc15							cc14$^{T/T}$	
	cc51							fmc24$^{C/C}$	
	fmc28							fmc17$^{T/T}$	
	cc39							fmc18$^{C/C}$	
	fmc48							fmc34$^{T/T}$	
	fmc49							fmc31$^{C/C}$	
	fmc41							fmc52$^{T/T}$	

Appendix 10 Compiled gel picture of the results obtained from RFLP analysis for C521T polymorphism. The sample identifiers and genotype for the known samples are indicated (2126 to C37). The samples F11 through to 872 are genotyped using this method.



Appendix 11 The population frequency from the SNPs found in dbSNP. Minor allele for control is used as reference. Con: Control, Inf: infertile, AA: African American, Euro: Europeans, Chi: Chinese, Jap: Japanese, You: Youroba, Cau: Caucasian, sCJD: sporadic Creutzfeldt Jakob disease, eAD: early-onset Alzheimer's disease, IAD: late-onset Alzheimer's disease, Neu: Neurological disease. [1] Croes et al. (2004) [2] Golanska et al. (2004) [3] Jeong et al. (2005) [4] Peoc'h et al. (2000) [5] Mead et al. (2000) [6] Schroder et al. (2001) [7] Infante et al. (2002).

Ethnicity			AA	Euro	Chi	Jap	You	Cau	Dut	Dut	Pol	Pol	Kor	Kor	Fre	Fre	Bri	Bri	Bri	Ger	Ger	Spa	Spa
n	96	96							250	52	33	43	102	110	106	95	175	76	41	111	58	283	288
Source					HAPMAP				[1]		[2]		[3]		[4]		[5]			[6]		[7]	
	Con	Inf							Con	sCJD	eAD	IAD	Con	sCJD	Con	sCJD	Con	sCJD	vCJD	Con	Neu	Con	AD
	A	A	A		A	A		A															
G-259A	0.16	0.23	0.07		0.24	0.44		0.23															
	C	C		C	C	C	C																
G-206C	0.09	0.07		0.092	0	0	0.22																
	T	T							T	T			T	T	T	T	T	T	T				
C77T	0.02	0.03							0	0.02			0.01	0	0.02	0.04	0	0	0				
	T	T	T						T	T					T	T							
C167T	0.011	0.00	0.013						0	0.02					0.01	0.02							
	T	T	T																				
C210T	0.00	0.005	0.013																				
	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
C521T	0.49	0.41	0.541	0.533	0.25	0.32	0.658	0.58	0.43	0.51	0.63	0.63	0.29	0.23	0.49	0.46	0.48	0.45	0.44	0.37	0.46	0.51	0.51
			T	T	T																		
			0.587	0.458	0.208																		
			T																				
			0.67																				

PERLEGEN ABI CSHL APPLERA