ARE SPERM-BINDING PROTEINS AMONG TWO CLOSELY RELATED FROG SPECIES RAPIDLY DIVERSIFYING?

Faezeh Bakhtiari-Nejad
B.S., Tehran University, Tehran, 2004

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE in

BIOLOGICAL SCIENCES at

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

FALL 2009
ARE SPERM-BINDING PROTEINS AMONG TWO CLOSELY RELATED FROG SPECIES RAPIDLY DIVERSIFYING?

A Thesis

by

Faezeh Bakhtiar Nejad

Approved by:

_________________________________, Committee Chair
Thomas R. Peavy, Ph.D.

_________________________________, Second Reader
Nicholas Ewing, Ph.D.

_________________________________, Third Reader
Brett Holland, Ph.D.

__________________
Date
Student: Faezeh Bakhtiari-Nejad

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

__________________________, Graduate Coordinator
Nickolas Ewing, Ph.D.                                    Date

Department of Biological Sciences
Abstract

of

ARE SPERM-BINDING PROTEINS AMONG TWO CLOSELY RELATED FROG SPECIES RAPIDLY DIVERSIFYING?

by

Faezeh Bakhtiari-Nejad

Vertebrate eggs have an extracellular matrix surrounding the egg, termed zona pellucida in mammals and vitelline envelope in amphibians that regulate the entry of sperm into the egg. Sperm-binding glycoprotein ZPC has been rapidly changing within mammalian species and has been thought to be involved in driving speciation events. Rapid amino acid changes at the sperm-binding sites within the protein can change ability of sperm to bind to and fertilize the egg. This can cause a prezygotic block to fertilization. We investigated the ZPC gene family in two closely related frog species *Xenopus laevis* and *Xenopus borealis* (diverged from each other about 30 million years ago) to examine how rapidly the ZPC protein is evolving compared to mammalian ZPC genes. Four and five unique cDNAs were found within ovary cDNA libraries from the frog species *X.laevis* and *X.borealis* respectively, indicating that this gene has undergone duplication event in both frog species.
In the present study, we tested the hypothesis that *X.laevis* and *X.borealis* have the same complement of ZPC cDNAs (coding for sperm binding protein) expressed within their ovaries; and amino acid sites encoded in the ZPC cDNAs show evidence of having been subjected to positive Darwinian selection toward their 3’ ends. New hypothesis could then be formulated as to where possible sperm binding sites are located given to the location of positively selected amino acid positions. The objectives of this study included 1) cloning unique ZPC genes within *X.borealis* and *X.laevis* cDNA libraries 2) determining the ZPC genes relationship (orthologus versus paralogus) 3) performing phylogenetic evolutionary analyses to further understand the evolution rate of ZPC genes in two frog species, and 4) comparing the sites under positive selection in frog ZPC with the ones identified in mammalian ZPC genes.

In order to analyze the *Xenopus* ZPC sequences, cDNAs from two were PCR amplified from ovary cDNA libraries, cloned, sequenced and assembled into partial length(middle section and 3’ ends) cDNAs. We then analyzed ZPCs genes to determine phylogenetic relationships by performing multiple alignments and generating two different tree topologies (Neigbor joining and Maximum likelihood parsimony trees). Finally the molecular evolution of the sites was statistically evaluated using PAML (Phylogenetic Analysis by Maximum Likelihood) software.
Our analysis identified 6 amino acid sites under positive selection in region close to C-terminus of *Xenopus* ZPC molecules. Although the sites under positive selection were not identical in mammals and frogs, they did fall in similar regions (within and near the region of the mammalian sperm combining site).

____________________, Committee Chair
Thomas R. Peavy, Ph.D.

____________________
Date
ACKNOWLEDGEMENTS

Several individuals have been integral in the completion of this project. First, I want to express my sincere gratitude to my professor advisor Dr Peavy for his invaluable support, guidance and much-needed encouragement throughout the duration of this process. I am indebted to him for enabling me to navigate the many obstacles of thesis completion. Without his guidance and enduring support and help, this project would not have been possible.

I express my sincere gratitude to the members of my graduate study committee Dr Ewing and Dr Holland for taking the time serving on my committee and for their valuable advice and help through my master degree. I do not have the words to express my gratitude to my parents for their endless love, care, and support. Finally, I would like to specially thank to my lovely husband whose patience and compassion during the final stages of my study was priceless.
# TABLE OF CONTENTS

Acknowledgements ........................................................................................................... vii

List of Tables .................................................................................................................... ix

List of Figures .................................................................................................................. x

Chapter

1. INTRODUCTION ........................................................................................................ 1

   Hypothesis ...................................................................................................................... 20

   Objectives ...................................................................................................................... 21

2. PCR, CLONING AND ANALYSES OF XENOPUS ZPC GENES .................. 22

   Materials and Methods ............................................................................................... 29

   Results / Discussion ..................................................................................................... 33

3. MOLECULAR AND EVOLUTIONARY ANALYSES OF XENOPUS ZPC GENES ........ 55

   Materials and Methods ............................................................................................... 58

   Results / Discussion ..................................................................................................... 59

4. SUMMARY ................................................................................................................... 66

Works Cited ..................................................................................................................... 70
LIST OF TABLES

1. Table 1 Distribution of *X. laevis* ZPC cDNAs……………………………36
2. Table 2 Distribution of *X. borealis* ZPC cDNAs……………………………37
3. Table 3 Pairwise comparison of the percent amino acid identity of the ZP domain of *X. laevis* ZPC cDNAs ………………………………………38
4. Table 4 Pairwise comparison of the percent amino acid identity of the ZP domain of *X. borealis* ZPC cDNAs……………………………38
5. Table 5 Percentage of amino acid identity between *X. laevis* and *X. borealis* middle and 3’ ends of ZPC cDNAs……………………………………54
6. Table 6 percent amino acid identity between *X.laevis*, *X.borealis*, human, mouse and zebra fish middle and 3’ ends of ZPC cDNAs………………54
7. Table 7 Likelihood ratio test of positive selection in *Xenopus* and Mammalian ZPCproteins………………………………………………………63
## LIST OF FIGURES

1. Figure 1  Model of allopatric speciation .............................................................. 4  
2. Figure 2  Structure of Sperm ................................................................. 6  
3. Figure 3  The main events in the sperm-egg interaction .......................... 7  
4. Figure 4  Transmission electron micrograph of mouse egg ZP filaments .... 10  
5. Figure 5  Structure of ZPC gene ................................................................. 11  
6. Figure 6  Illustration comparing MHC and ZPC ........................................ 18  
7. Figure 7  Phylogeny of Xenopus species ................................................ 23  
8. Figure 8  Overview of Marathon cDNA Amplification procedure ........ 24  
9. Figure 9  Sequences of the Marathon cDNA adaptor and AP1 primer ......... 25  
10. Figure 10  Nested PCR ................................................................................... 31  
11. Figure 11  Agarose gel of the PCR amplification of ZPC from a *Xenopus boreali* ovary cDNA library ................................................................. 34  
12. Figure 12  Amino acid sequence alignments of ZPC middle sections within  
               *X.borealis* and *X. laevis* cDNA libraries ........................................ 39  
13. Figure 13  Alignment of middle sections and 3' ends of ZPC glycoproteins  
               within *X. laevis* cDNA library ........................................................... 40  
14. Figure 14  Alignment of middle sections and 3' ends of ZPC glycoproteins  
               within *X. borealis* cDNA library ........................................................... 41  
15. Figure 15  Analysis of the ZPC primary sequence structure. Alignment of the  
               whole length ZPC clones .................................................................. 44  
16. Figure 16  Amino acid sequences alignment of ZPC genes from *X.laevis* and  
               *X. borealis* cDNA libraries (including the middle section and 3'ends... 45
17. Figure 17  The Neighbor-joining tree showing phylogenetic relationship of 
*X.laevis* ZPC gene ................................................................. 49
18. Figure 18  Neighbor-joining tree showing phylogenetic relationship of 
*X.borealis* ZPC genes ............................................................... 49
19. Figure 19  Neighbor-joining tree showing the ZPC gene relationships from 
to *Xenopus* species ................................................................. 50
20. Figure 20  Amino acid sequence alignment of ZPC genes from *X.laevis*, 
*X.borealis* and human, mouse and zebra fish as outgroups .......... 51
21. Figure 21  ZPC gene tree ................................................................. 52
22. Figure 22  Maximum likelihood phylogenetic tree of the ZPC genes from 
*Xenopus laevis, Xenopus borealis* and Mouse ............................ 53
23. Figure 23  A schematic representation of the ZPC molecule indicating 
functional regions ................................................................. 62
24. Figure 24  Alignment of ZPC amino acids of *X. laevis, X. borealis* and 
Swanson’s Mammalian species .............................................. 65
Chapter 1

INTRODUCTION

Fertilization is an essential step in the lifecycle for most organisms and is a highly regulated process that occurs after gamete recognition and fusion. The pathway involves relatively species-specific binding of sperm to eggs followed by fusion of a single sperm with the egg (Wassarman, Jovin and Litscher et al, 2001). Gamete recognition involves complementary proteins on sperm and egg that affect the ability of individuals to interbreed with each other. Moreover, sperm and egg molecules must be compatible in order for fertilization to occur. The inability of gametes from two individuals or a couple to recognize and bind to each other would result in infertility. Larger scale infertility problems due to incompatible gamete interactions, such as between individuals from two different populations (although originally derived from the same ancestral species), would give rise to the formation of new species (or in other words a speciation event).

Many studies have recently focused on the cause of infertility in sterile couples. Infertility affects approximately 15% of all married couples (Wassarman et al, 2002). Identifiable causes for infertility account for only 30-50% of the cases, however there is experimental evidence indicating the molecular basis for particular male or female-based infertility conditions (Bhasin and Baker et al, 2007). Assisted reproductive techniques have been used to facilitate fertilization in many couples diagnosed with infertility problems. For example, in vitro fertilization (IVF) is a technique in which eggs are removed from the woman's ovary or fallopian tubes and fertilized outside of the womb by
addition of sperm in vitro. However, if the cause of infertility is due to the inability of sperm to bind to the egg, IVF will not be able to facilitate fertilization since sperm will never be able to bind and subsequently fuse with the egg. In this case, couples may have to try other artificial reproductive techniques such as Intracytoplasmic Sperm Injection, ICSI, which is a procedure in which a single sperm is injected directly into the cytoplasm of the egg. Therefore, a better understanding of the molecular basis of fertilization is vital to improve assisted reproductive techniques and our understanding of the basis of infertility. Once the molecular basis underlying fertilization is elucidated, one can develop specific tests to diagnose the causes of reduced fertility and therapies to treat the specific cause. In addition, if we know the mechanism of sperm binding, it is possible that we can develop non-hormonal means of contraception to disrupt or prevent sperm binding and therefore, regulate the population of human and animals (Rankin and Dean et al, 2000).

On a larger scale, infertility between individuals from separate populations would be the formation of new species (a group of organisms capable of interbreeding and producing fertile offspring) since these distinct individuals can no longer reproduce with each other. Speciation is the evolutionary process whereby over time one species evolves into a new, genetically distinct species. For example, in allopatric speciation (also known as geographic speciation) biological populations are physically isolated by an extrinsic barrier and evolve intrinsic (genetic) reproductive isolation, such that if the barrier breaks down, individuals of the populations can no longer interbreed (Figure 1). If the incompatibility between individuals is due to sperm-binding, then that would be
classified as a pre-zygotic block to fertilization (versus post-zygotic mechanisms that occur due to genetic incompatibilities after zygote formation). It is thought that rapid changes in the egg's sperm-binding protein, termed ZPC or ZP3 in vertebrate species (discussed in more detail later), could drive speciation by altering the complementary portion of the molecule that binds sperm thereby resulting in the inability to bind to sperm from individuals in another population (Wassarman et al, 2004).

In animals that produce free-spawning gametes such as abalone, there is experimental evidence to suggest that prezygotic isolation (speciation) can result from the rapid changes of genes coding for sperm and egg recognition proteins during fertilization (Swanson et al, 2003). The inability of sperm to recognize and bind to the egg would create a reproductive barrier which then leads to reproductive isolation and the establishment of new species (Swanson and Vacquier et al, 2002). For example, abalone sperm has a soluble protein called lysin within their acrosomal vesicle and when it is released, binds to its receptor, VERL, on the egg vitelline envelope (Swanson and Vacquier et al, 1997; Kresge and Vacquier et al 2001). Lysin is species specific at dissolving VEs and this specificity can be attributed to at least 23 residues that evolve rapidly. Molecular analyses showed that the amino acid sequences of lysins from neighboring abalone species are remarkably divergent (Swanson and Vacquier, 2002). It is thought that the rapid changes on the VERL molecule creates a continuous pressure on lysin to adapt to the ever-changing VERL in order to maintain their interaction and provides an explanation for the adaptive evolution of lysin (Yang et al, 2000; Metz et al, 1998). In other words, this molecular conflict between the sexes seemingly causes a co-
evolutionary catch and chase scenario (Swanson et al, 2003). The result of this continuous co-evolution of the gamete recognition system could be the splitting of one population into two that are reproductively isolated (speciation).

![Diagram of Allopatric Speciation]

Figure 1. Model of allopatric speciation.

Sperm-Egg Binding Mechanism

In sexually reproducing organisms, a new individual is formed when two haploid cells, the sperm and the egg, come together to form a diploid zygote. Before fusion and zygote formation can occur in vertebrate species, the sperm must penetrate through the extracellular coat barrier that surrounds the egg (termed chorion in fish, the vitelline envelope in amphibians, and zona pellucida in mammals). However, in fish, the
chorion has a hole or entrance way through the chorion termed the micropyle which means the sperm can bypass the chorion and reach the egg's plasma membrane. Interestingly, species-specificity in fish seems to be due to the release of peptide chemoattractants (analogous to smell) by the egg. In order for amphibian and mammalian sperm to penetrate through the extracellular coat, sperm exocytotically release the contents of a large, specialized secretory vesicle termed the acrosome after binding to the egg coat proteins which enables them to, proteolytically generate a hole through the thick barrier (Figure 2). The acrosome is a membrane-bound organelle located in the head of the sperm which contains these sepecialized enzymes. The trigger for acrosomal exocytosis is a signal transduction event after binding to the egg's sperm-binding protein, ZP3 or ZPC. The binding event causes the sperm's plasma membrane to fuse with the outer acrosomal membrane and subsequently release the acrosomal contents (Bleil and Wassarman et al, 1983). Therefore, sperm binding to the extracellular coat is a critical regulatory step in the fertilization process in higher vertebrates (Figure 3).

Following fertilization, the egg's cortical granules (vesicles found at the periphery near the plasma membrane) fuse with the plasma membrane discharging their contents into the space between the extracellular coat and plasma membrane. The cortical granule contents react with the zona pellucida/vitelline envelope glycoproteins and prevent additional sperm penetration thereby establishing a block to polyspermic fertilization (Wassarman et al. 1987; Wassarman et al, 1988). Free-swimming sperm are thus unable to bind to nor penetrate through the extracellular coat of fertilized eggs.
Figure 2. The acrosome located on the head of sperm contains an inner and outer membranes that fuse during the acrosome reaction. This results in the release of enzymes that are essential for penetration of the extracellular coat by sperm (Dean, 1992).
7

Figure 3. The Main events in the sperm-egg interaction. First, sperm bind to ZP3/ZPC which induces acrosome reaction. The enzymes released from acrosome digest zona pellucida/vitelline envelope glycoproteins so that sperm can swim through the barrier to reach the egg surface. After fusion, the exocytotic release of the egg cortical granule contents modifies the extracellular coat glycoproteins resulting in a block to polyspermy (Sawnson and Vacquier, 2002).

The Structure and Function of the Zona Pellucida (ZP)

In general, the extracellular coat of eggs in mammalian and non-mammalian species serve as a barrier to fertilization and permits binding of sperm only when a sperm and egg come from the same species, termed species-specific binding (Wassarman and Jovin et al, 2002). The ZP can block interspecific fertilization by failing to permit the initial binding of sperm to eggs, induction of acrosome reaction, or penetration of bound
sperm through this extracellular coat. This indicates that the ZP possesses specific proteins that are recognized by sperm from same species (Wassarman and Jovin et al, 2001).

The mammalian ZP consists of three to four major glycoproteins encoded by genes ZPA, ZPB and ZPC that have orthologs in all mammalian species known to date (in other words, homologous gene sequences that have evolved directly from an ancestral gene and are separated by a speciation event). Unfortunately, the nomenclature has been confounded by the use of two different gene naming schemes which has named the same genes listed above as ZP2, ZP1 and ZP3, respectively. In this document, the term ZPC will be consistently used to refer to the vertebrate egg sperm-binding gene as opposed to ZP3.

In the case of the model frog species *Xenopus laevis*, five glycoproteins ZPA, ZPX1, ZPB, ZPC and ZPX2 are present in the egg's vitelline envelope. An ortholog of each of the three groups of mammalian ZP glycoproteins has been found in *X. laevis* however, two additional ZP genes (ZPX1, ZPX2), with no known mammalian orthologs, have also been characterized in *X. laevis*. Molecular phylogenetic analyses have suggested that an ancient event in ZP evolution was a gene duplication event thereby giving rise to the ZPC gene and to the precursor of the ZPA and ZPB gene paralogs (homologous gene sequences that are separated by a gene duplication event) (Spargo and Hope, 2003). In addition, the ZPC gene has been discovered in other vertebrate species and related glycoproteins are found in the vitelline envelope of virtually all non-mammalian eggs from fish to birds. In the model fish species medaka and zebrafish,
peptides released from the ZPC homologue have been shown to attract sperm to the egg (Iwamatsu T and Yoshizaki et al, 1997).

Of the three ZP glycoproteins in the egg, only purified ZPC had been found to prevent the binding of sperm to ovulated eggs in competitive binding assays and to induce the acrosome reaction thereby showing its functional importance (Bleil and Wassarman, et al, 1986; Florman and Wassarman, 1984). A variety of other glycoproteins, including other ZP glycoproteins, had no effect on binding of sperm to eggs in vitro. Moreover, in the Xenopus laevis frog, Hedrick and Yang (1997) provided evidence that the ZPC molecule is responsible for sperm binding activity. ZPC is also believed to play a structural role in assembly and maintenance of the ZP since knock-out mice for the ZPC gene (ZPC gene deleted from their genome) resulted in oocytes without a zona pellucida and were subsequently infertile (Wassarman, 2001). Therefore, ZPC serves as an essential structural component of the ZP in addition to its functional properties as a sperm receptor and acrosome reaction-inducer. ZPA has been implicated as a secondary sperm receptor that can bind sperm only after the induction of the acrosome reaction and is thought to keep the sperm lightly attached to the ZP as it penetrates through it (Bleil et al, 1988). The role of ZPB during fertilization is unclear however.

The murine ZP matrix is formed by long filaments of ZPA and ZPC dimers that are cross-linked by ZPB (Wassarman et al, 2001) (Figure 4). The three glycoproteins are held together in the ZP by non-covalent bonds (Wassarman and Greve, 2002). In mammals, the zona pellucida proteins are synthesized in the egg and secreted into the
zona matrix (Dean et al, 1992). Mouse ZPC is the smallest of the ZP proteins, with an apparent molecular weight of 83kDa (Bleil and Wassarman, 1980).

Figure 4. Transmission electron micrograph of mouse egg ZP filaments. The filaments are composed of ZPA and ZPC dimer subunits cross-linked by ZPB (Wassarman et al, 2001).

ZPC glycoproteins have conserved amino acid sequences in all vertebrate. For example mouse and human ZPC proteins are 60-70% identical, hamster and mouse ZPC are 81% identical (Dean and Chamberlin et al, 1990) and human and *Xenopus* ZPC are 40.8% identical (Yang and Hedrick, 1997). Approximately 260 amino acids of ZPC, which includes eight conserved cysteine residues, comprise a so-called “zona pellucida domain” (ZP domain; Bork and Sander et al, 1992) which is conserved between species (Figure 5). All 8 cysteine residues are invariant within the ZP domain and are located in the same place in all ZP glycoproteins (signature of ZPC molecules). A ZP domain is
present in all ZP glycoproteins, from mice to human, and is present at the C-terminus of many other glycoproteins found in both vertebrate and invertebrates (e.g., tectorin-β, uromodulin, ebnerin, and cuticlin) (Wassarman, Jovine and Litscher, 2001). Interestingly, the exact amino acid sequence of the approximately 260 amino acid ZP domain is quite divergent between family members but divergent ZP domains still preserve their overall folding conformation and nearly all of the cysteine residues. There is evidence that ZP domains plays a role in polymerization of the ZP glycoproteins and serve a structural role for to preserve its 3-D conformation (Jovine et al. 2002). In addition to the centrally-located ZP domain, ZPC has a signal peptide, a putative sperm binding site, a transmembrane domain near the carboxyl terminus, and a furin cleavage site (these features will be discussed in more depth next chapter) (Figure 5).

![Structure of ZPC Genes](image)

Figure 5. Structure of ZPC gene. Primers 1, 2, 3 and 4 are degenerate primers that were used for PCR to amplify the middle conserved ZP domain section.
Glycosylation pattern of ZP proteins

As the three proteins traffic through the cell, they are extensively glycosylated and the resultant carbohydrate side chains account for roughly half of the mass of each protein (Nagdas et al, 1994; Easto et al, 2000). Each mouse ZP protein consists of a polypeptide backbone to which asparagine-(N-) linked and serine/threonine – (O-) linked oligosaccharides are covalently attached (Florman and Wassarman, 1985). Previous studies have suggested that the sperm-binding activity of ZPC is dependent on a relatively small proportion of its O-linked oligosaccharides with potential dependence on galactose or N-acetylglucosamine at the terminal position (Florman et al, 1985; Bleil et al, 1988; Miller et al, 1992). A variety of in vitro binding assays have been performed to dissect which mouse ZPC amino acid residues and glycosylation sites are the most critical for sperm-binding which include entailed generating ZPC peptides (Rosiere et al, 1992; Litscher et al, 1996), exon swapping (Kinloch et al, 1995) and site-directed mutagenesis (Chen et al, 1988). Results from these studies suggest that sperm-binding activity of ZPC is associated with the C-terminal half of the polypeptide and within a region termed the sperm combining site. Several key O-linked glycosylation sites (serine/threonine residues) were found to be within this sperm combining site and have an effect on sperm binding activity when mutated. It is likely that both the amino acid sequence and glycosylation pattern are essential for sperm binding activity since proper presentation of key oligosaccharides chains would be critical. In support of this hypothesis, the polypeptide region within and near the sperm combining site exhibits
extensive sequence divergence in higher vertebrates and may be related to sperm-binding capabilities (Kinloch RA et al, 1995).

Sequence Divergence of ZPC

Sperm binding to the ZP is analogous to a “lock” and “key” mechanism. Molecules on the sperm surface (key) need to match up perfectly with the binding site found within the ZPC glycoprotein (lock). The specificity of this interaction serves to maintain the identity of the species and prevent hybridization with other species (species-specific fertilization). Rapid changes in reproductive genes can affect the ability of individuals to reproduce with each other since they would change the lock and key binding mechanism which can lead to infertility or speciation.

Evidence suggests that the 5’ and 3’ ends of the protein-coding portion of the ZPC gene are mutating rapidly with respect to amount of amino acid change between species (Figure 5). A molecular evolutionary analysis on eight mammalian species showed that there is statistical evidence that the sperm-binding proteins have more amino acid changes in the 5’ and 3’ ends of the molecule and particular nucleotide sites in these regions have been detected to be under the influence of positive Darwinian selection (defined in the next paragraph) (Swanson et al, 2001). A subset of the sites identified to be under positive selection fall within the sperm combining regions discussed previously.
Molecular Evolution and Selectional Forces

Rapid changes at specific amino acid sites within a protein-coding gene can be due to the influence of positive Darwinian selection (a form of natural selection) which in essence is a selectional force that acts to increase the diversity of amino acids at particular sites due to their beneficial nature. Positive Darwinian selection is thought to contribute to adaptive evolution since the beneficial nature of the mutations would likely serve to improve survival and reproductive success of the organism in their particular environment. Positive selection at the molecular level is usually indicated by an increased ratio of the number of non-synonymous substitutions (nucleotide mutations that result in amino acid changes) termed dN when compared to the number of synonymous substitutions (silent mutations in which nucleotide mutations in the protein-coding region that do not result in amino acid changes due to the degeneracy of codon sites) termed dS. In other words, an amino acid site is under positive Darwinian selection if it is undergoing a high number of amino acid changing mutations relative to silent mutations, in which case the dN/dS is > 1. Positive selection at nucleotide sites for amino acid changes is thought to be an indicator that these sites are functionally important for the protein and imparts a benefit to the organism by its diversification (Swanson et al, 2003).

The other two molecular evolutionary classifications on nucleotide sites are neutral selection and negative purifying selection. The evolution of a site is neutral when the dN/dS ratio is approximately 1 (Yang and Beilawski, et al. 2000), which means that non-synonymous changes in nucleotide bases versus synonymous changes are occurring at approximately equal rates. This means that there is no penalty or benefit to changing
the amino acid residue with respect to the resulting protein structure and function.

Fixation of this mutation in the population would follow the neutral theory of evolution which is driven by genetic drift. Purifying or negative selection is when amino acid changes are less than silent changes which in this case the \( dN/dS \) is <1. Purifying selection removes amino acid changing mutations from the genetic pool that are detrimental to the protein's structure and function. However, silent site mutations will occur at the usual rate and thus accumulate to a higher proportion when compared to amino acid changing positions within the same codon.

Most protein-coding nucleotide sites are subject to purifying and neutral selection (Swanson et al, 2003). Only a few classes of molecules have been detected to be influenced by positive selectional pressures. As mentioned above, positive selection on amino acid encoding sites within a gene occurs when amino acid substitutions are found to be beneficial to the reproduction or survival of individuals within a population. In cases where positive selection has been found, the amino acid replacements often gather on protein surfaces involved in some sort of specific recognition mechanism (e.g. receptor-ligand type interactions). Molecular evidence suggests that genes involved in the immune response, olfaction (sense of smell), and reproduction have regions that have been subjected to positive selection and are among the most rapidly evolving classes of genes (Swanson et al, 2003).
Positive Selection in MHC, Olfactory and Reproductive Genes

One of the best examples of positive selection is within the major histocompatibility complex (MHC) gene family. Immune system genes in particular show high levels of molecular diversification in order to co-evolve with pathogens that threaten organisms (Sitnikova et al, 2000). Previous work showed that the nucleotide sites in MHC genes detected to be subjected to positive selection are involved in antigen recognition sites. When the nucleotide sequences of the MHC class I genes were aligned and statistically compared, the dN/dS ratio was significantly greater than 1 for amino acid residues located in their respective antigen recognition sites as mapped by protein crystallographic studies. Positive selection on the binding portion of these MHC molecules is thought to contribute to optimization and diversification of the binding site so as to increase their ability to recognize foreign molecules. In contrast, amino acid sites that appeared to be more structurally important such as those buried within the interior of their structure often showed signs of purifying selection (Swanson et al, 2000) (Figure 6).

Another essential recognition system is the sense of smell. Olfaction is one of the most ancient and essential senses, with which animals monitor the environment enabling the recognition of mates, offspring, predators, food and chemical dangers. Nielsen et al (2005) identified 4 genes involved in olfaction which showed statistical evidence for positive selection from their pattern of nonsynonymous and synonymous substitution after the comparison of 13731 chimpanzee and human orthologous genes. The sites shown to be subjected to positive selection were found to be in the odorant binding
pocket which is analogous to MHC antigenic binding sites (Nielsen et al, 2005).

Positive selection on odorant binding sites of olfactory genes is thought to be beneficial because it enables these proteins to recognize and bind to diverse odor type molecules.

With respect to reproductive genes, comparing gene sequences within and between closely related species has shown that the genes that are involved in reproduction are more divergent than the genes that are expressed in non-reproductive tissues (Vacquire et al, 1998; Kulathinal et al, 2000). There is molecular evidence that this rapid divergence is driven by positive Darwinian selection indicating that sequence diversification is beneficial to survival (Swanson and Vacquier et al, 2002). The phenomenon of species-specific fertilization in particular, shows that the proteins that are involved in gamete recognition must have a species-specific structure and that they must bind each other with species-specific affinity. Therefore, sperm and egg proteins involved in fertilization need to co-evolve with each other in order to maintain their interaction within the population (Swanson and Vacquier et al, 2002). Thus, ZPC, MHC and olfactory genes have seemingly parallel stories. They have regions detected to be subject to positive selection and they are all involved in molecular recognition. (Figure 6)
Figure 6. Illustration comparing MHC (left) and ZPC (right) (Swanson et al., 2001). The MHC structure was resolved by X-ray crystallography whereas the structure of the ZPC protein was not theoretically modeled. The amino acid sites subject to positive selection are noted in black.

Gene Duplication

In addition to those properties stated above, all three gene families have been shown to undergo duplication events creating duplicate genes that have evolved independently. Several studies suggested that MHC molecules are subject to a birth-and-death model of evolution in which new genes are created by repeated gene duplication and some duplicate genes are maintained in the genome for a long time but others are deleted or become nonfunctional by deleterious mutations (Nei and Sitnikova et al., 1997; Klein et al., 1993). Based on this form of evolution, the number of MHC genes is assumed to be expanded depending on the need to protect the host from ever-changing groups of parasites. Evidence shows that the ZPC gene has also
undergone duplication events in a number of species. A study done by Conner and Hughes (2003) provided evidence that there are multiple copies of the ZPC gene in a number of fish species. Blast searches of the database indicated that the zebrafish (*Danio rerio*) had at least seven copies of ZPC genes in its genome. Our data has shown that there are also multiple copies of ZPC genes within ovary cDNA libraries from *X. laevis* and *X. borealis* frog species suggesting that paralogous ZPC genes exist within their genomes also.

Rationale for Current Studies

Since previous studies (Swanson et al, 2001) have been limited to mammalian species, our study is designed to examine more ancestral vertebrate species to assess whether positive selection of the ZPC gene is a phenomenon common to the evolution of vertebrate species. In addition, we are interested in determining whether the same or different amino acid sites are influenced by positive selection since these sites are likely to be involved in the binding site for sperm recognition.

We have chosen the amphibian species *Xenopus laevis* and closely related *X. borealis* for the following reasons:

- *Xenopus* frogs produce a large amount of gametes for reproductive studies, providing on the order of 3000 eggs per mating.
- Fertilization occurs externally (e.g. outside the body in pond water) which makes it easier to manipulate in the lab.
• Importantly, frogs are a transitional group of organisms that have maintained many of the original molecular adaptations from our vertebrate ancestry.

• Fish species are not good vertebrate models in this case because the chorion (homologous to the zona pellucida and contains a ZPC homologue) does not bind sperm during fertilization due to the existence of the micropyle.

• *Xenopus laevis* is a relevant model organism to study within the field of reproductive biology and *X. borealis* is a close relative so we can examine the molecular evolution of the ZPC gene family over a moderately close time period within amphibians

• ZPC genes have been identified in these frogs (*X. laevis* and *X. borealis*) and are functionally equivalent to their mammalian counterparts

Hypothesis

*X. laevis* and *X. borealis* have orthologous ZPC cDNAs expressed within their ovaries; and amino acid sites encoded in the ZPC cDNAs will show evidence of having been subjected to positive Darwinian selection toward their 3’ ends.
Objectives

1) ZPC cDNA sequences will be PCR amplified, cloned and sequenced from *X. laevis* and *X. borealis* ovary cDNA libraries.

2) ZPC cDNA sequences will be aligned and phylogenetically analyzed to assess their relationships (paralogous versus orthologous).

3) The aligned ZPC cDNA sequences will be examined for nucleotide sites that have been subjected to positive selection using molecular evolutionary analysis software.

4) The nucleotide sites subjected to positive selection in the frog ZPC cDNAs will be compared to those identified in the mammalian ZPC study.
Chapter 2

PCR, CLONING AND ANALYSES OF XENOPUS ZPC GENES

Among vertebrate genera, *Xenopus* frogs (African clawed frogs) stand out as a model vertebrate because of the degree to which they have been investigated at the molecular, cellular, and developmental levels (Cannatella et al, 1993). The *Xenopus* genus is comprised of two groups of species that diverged at least 30 million years ago (Wilson et al, 1987), the *tropicalis* (single species) with a smaller nuclear genome consisting of 20 chromosomes, and the *laevis* group (five species), which have larger genomes, most consisting of 36 chromosomes (Wilson et al, 1987). As shown by cytogenetic studies, *Xenopus* species have chromosome numbers consisting of 20, 36, 40, 72, and 108, suggesting that polyploidy has occurred frequently in this group (Hughes A and Pasquier et al, 1986). *Xenopus tropicalis* is the only species in the *Xenopus* genus that has remained as a diploid. Evidence indicates that *X. laevis* and *X. borealis* (2n=36), are thought to have arisen from a tetraploidization (4n) event.

A few studies have documented the tetraploid state of specific genes within the *X. laevis* and *X. borealis* genome. For example, comparison of serum albumins from different *Xenopus* frogs revealed that this gene is indeed present on four chromosomes within *X. laevis* and *X. borealis* genomes (Wilson and Chester, 1976). Phylogenetic and molecular evolutionary analyses of DNA sequence and cytogenetic data have lead to the hypothesis that this genome duplication occurred in the common ancestor to the laevis group approximately 30 mya (Wilson and Chester, 1976) (Figure 7). Thus, it is our
hypothesis that the *X. laevis* and *X. borealis* ZPC gene can be expressed from four chromosomes (4n) similar to the albumin gene which could all be allelic variants (if not homozygous). However, it is also possible that the ZPC gene has duplicated and the second genome location (paralogous relationship) would add to the potential total number of ZPC genes expressed.

Figure. 7. Phylogeny of *Xenopus* species. The divergence times are estimated based on immunological distance of serum albumins which assumes that amino acid sites for albumin evolve at a steady rate of about 1.7 immunological distance units per million years when pairs of species were compared. Hymenochirus (another genus of African pipid frogs) was used as an outgroup for the analysis (Wilson et al, 1976).

In order to be able to analyze expressed ZPC genes within *X. laevis* and *X. borealis* ovaries, ovary mRNAs needed to be purified and converted into cDNA (complementary DNA) by reverse transcription. Reverse transcriptase is a specialized DNA polymerase that uses single-stranded RNA as a template to synthesize complementary DNA strands. Ovary cDNA libraries for *X. laevis* and *X. borealis* were
generated using a Marathon cDNA Amplification Kit (Clonetech, MI, Figure 8) because it employs a specially designed adaptor that is added after cDNA synthesis that significantly reduces erroneous PCR when amplifying the 5' and 3' cDNA ends, also termed RACE (rapid amplification of cDNA ends). In addition, the Marathon cDNA synthesis kit reduces 3' heterogeneity at the poly A tail by using a modified lock-docking oligo(dT) primer (contains two degenerate nucleotides at the 3' end) which positions the primer at the beginning of the poly A+ tail for first strand synthesis (Figure 9). After completion of the second strand cDNA synthesis, the marathon adaptor is blunt end ligated onto the cDNA ends which contain the AP1 primer site (anchor primer) specific for RACE. A special feature of the marathon adaptor is that the AP1 site is not present initially since it is encoded by the complementary strand of the adaptor, meaning this double stranded DNA adaptor is not blunt ended but rather has a single stranded region that overhangs the AP1 site (Figure 9). Thus, PCR amplification only begins when an internal gene specific primer is used during the first round of PCR amplification which then generates the AP1 site for second round amplification (Figure 9).
Figure 8. Overview of Marathon cDNA Amplification procedure (Manual pdf. www.clonetech.com).

Marathon cDNA Adaptor:

**T7 Promoter**

5′-CTAATACGACTCACTATAGGGCTCGAGCGCCGCCGCAGGT-3′

3′-H2N-CCCCTCCA-PO4-5′

**Not I**

**Srf I/Xma I**

Adaptor Primer 1 (AP1; 27-mer):

5′-CCATCCTAATACGACTCACTATAGGGC-3′

Nested Adaptor Primer 2 (AP2; 23-mer):

5′-ACTCCTATAGGGCTCGAGCGGC-3′

Figure 9. Sequences of the Marathon cDNA adaptor and AP1 primer (http://www.clontech.com/images/pt/PT1115-1.pdf).
In order to address objective one, ovary cDNA libraries made from single individuals will be used for ZPC PCR amplification and cloning. This objective is further subdivided into two separate aims; 1) cataloging of the middle conserved portions of *Xenopus* ZPC cDNAs to determine how many different ZPC genes are expressed within an individual's ovary, and 2) completing the cDNA sequence for each unique ZPC gene found in the library by 5' and 3' RACE. In order to identify different ZPC genes, degenerate primers (designed by Dr. Peavy) were used to PCR amplify the conserved ZP domain found in all vertebrate ZPC genes by taking advantage of the highly conserved cysteine residues in their structure. Two forward (ZPC1 and ZPC2) and two reverse degenerate primers (ZPC3 and ZPC4) were used so that the various combinations could be used to PCR amplify all potential ZPC transcripts found in the libraries (i.e. ZPC1/ZPC3, ZPC2/ZPC4, ZPC1/ZPC4, and ZPC2/ZPC3) (refer to Figure 5). These degenerate primers were designed by Dr. Peavy from conserved regions of the ZP domain found within an amino acid sequence alignment of ZPC genes from vertebrate species which included mammals, *X. laevis* and the domestic chicken. The combination of two degenerate primers (e.g. ZPC1/ZPC3) in PCR reactions increased the odds of amplifying authentic ZPC cDNAs rather than using only a single degenerate primer in combination with the AP1 vector end.

After cloning and sequencing a significant number of ZPC conserved portions within individual libraries (>80 sequencing reactions), a catalog of unique clones was generated by comparing the sequences to each other. Then each unique ZPC clone was
targeted for completion of the full length cDNA sequence by designing internal gene-specific primers derived from the catalogued middle sequence so that 5' and 3' RACE could be performed using the AP1 primer sites. After cloning the PCR products into plasmid vectors, the cDNA insert will be sequenced and the full length cDNA sequence will be assembled by editing the overlapping sequences together.

All vertebrate ZPC transcripts known to date have retained specific structural features which we anticipate all authentic *Xenopus* ZPC cDNAs should also have when examined (refer to figure 5). As mentioned in chapter 1, all ZP protein families (e.g. ZPA, ZPB, and ZPC) contain a conserved ZP domain which is thought to help with the structure and folding of ZP proteins and also facilitates their assembly into the zona pellucida matrix. The ZP domain consists of about 260 residues and is characterized by eight strictly conserved cysteine residues which likely form disulfide bridges. In addition, hydrophobic and polar residues are retained at numerous positions in the ZP domain which is consistent with the formation of a conserved three-dimensional structure (Bork and Sander et al, 1992). ZP domains are also found in a variety of other gene families expressed in locations other than the egg in both vertebrates and invertebrates (e.g., tectorin-β, uromodulin, ebnerin and dumpy). Interestingly, in all these protein families the ZP domain seems to serve a structural and assembly role. For example in the uromodulin protein, the most abundant protein in human urine, the ZP domain is responsible for protein polymerization and assembly of the protein into filaments (Cattaneo and Schwend, 2008). Similarly, the ZPC protein needs to be secreted and assembled into the growing zona pellucida matrix outside of the cell.
In order to be secreted, all ZPC cDNAs encode a signal peptide at the N-terminus that directs them into a secretory pathway. Within the secretory pathway, ZPC polypeptides are glycosylated within the golgi apparatus. In mice, it is thought that sperm recognize and bind to specific serine/threonine-linked (O-linked) oligosaccharides present near the carboxy terminus of ZPC (Wassarman et al, 1995). Once ZPC proteins are glycosylated and transported to the egg surface, they initially become embedded into the plasma membrane since they contain a transmembrane domain near the carboxyl terminus but are then released into the growing zona pellucida by enzymatic cleavage by a proconvertase at a furin cleavage site adjacent to the transmembrane domain (Figure 5). The failure of the cleavage event results in ZPC accumulation in the endoplasmic reticulum (Litscher and Wassarman, 1999). The mature ZPC glycoprotein is then assembled into the ZP matrix orchestrated by the ZP domain.

After authenticating and annotating these characteristic features with each Xenopus ZPC transcript, the cDNA sequences will be analyzed phylogenetically to assess their relationships (orthologous versus paralogus). As described earlier in the introduction, orthologous genes are homologous genes in different species that are similar to each other because they originated from a common ancestor and are separated by a speciation event. By contrast, paralogous are homologous genes that were separated by a gene duplication event. If there are two different ZPC transcripts found within the same Xenopus individual, there are two options for what it can be: 1) since they are tetraploid, they could be different alleles (one chromosomal locus but there are 4 homologous chromosomes) or 2) they are paralogues and thus do not exist at the same chromosomal
locus but rather a gene duplication event occurred which created a new locus. Although
this is difficult to fully determine without knowing the entire genome sequence for each
individual or performing an indepth cytogenetic analysis (e.g. fluorescent in-situ
hybridization studies on chromosomes), a phylogenetic analysis should be able to
elucidate the most likely orthologous relationships.

Materials and Methods

PCR Amplification of middle section, 3’ and 5’ ends

PCR amplification of the middle portion of ZPC cDNAs (conserved ZP domain)
was carried out using the Advantage PCR kit (Clontech, CA). The reactions were
performed in 50μl reactions containing 39μl water, 1.5μl dNTPs mix (0.3mM ea.), 0.5μl
Advantage Polymerase (5u/μl), 5μl PCR Buffer, 1.5μl of degenerate forward primer
(10μM), 1.5μl of degenerate reverse primer and 1μ of template cDNA library dilution (1
to 100 dilution). PCR amplification was performed as follows: initial cycle at 95 °C for 5
min, 34 cycles at 94°C for 1 min, 50 °C for 1 min and 72°C for 1 min, and one final cycle
at 70°C for 7 min. PCR products were separated by electrophoresis using 1% agarose
gels and TAE buffer, and visualized by ethidium bromide staining. After purification by
a PCR Purification kit (Qiagen, MD), amplified products were ligated into the pGEM T-
easy vector (Promega, WI). Ligation reactions were performed using 2.5μl of PCR
products, 0.5μl of pGEM vector (50ng/μl), 1μl T4 DNA ligase(3u/μl), 5μl of 10X
ligation buffer and 1μl water. The reaction mixtures were incubated overnight at 4°C.
Recombinant plasmids were then transformed into E. coli bacteria (gram negative
bacterium) and plated on LB ampicilin (50μg/mL) plates. Recombinant plasmids were
purified from transformed colonies using a Qiagen Miniprep kit (Qiagen). In order to identify recombinant plasmids with appropriately sized PCR products, restriction digest reactions were carried out in 10μl reactions containing 0.5μl of Bstz1 enzyme (10u/μl)(Promega, WI), 1μl of 10 x buffers, 1.5μl of water and 6μl of each miniprep plasmid DNA. Selected plasmid preparations were then submitted for commercial DNA sequencing (Sequetech, Inc, Mountainview, CA 94043).

After the different middle portions of the ZPC cDNA sequences were analyzed and catalogued (described in next section), internal gene-specific primers were designed from the different middle portions of the ZPC cDNA sequences using the PrimerSelect program within the Lasergene software package (version 7). The AP1 primer site on the marathon adaptor was used in combination with the internal gene-specific primer for PCR amplification. PCR amplification was performed in 50μl reactions containing 39μl water, 1.5μl dNTPs (0.3 mM each), 0.5μl (5u/μl)Advantage polymerase, 5μl PCR buffer, 1.5μl AP1 (8μM.) and 1.5μl of either forward gene-specific primer or reverse gene-specific primer(10μM ea.) for 3’ RACE and 5’ RACE reactions respectively.

In many cases, a nested PCR approach was necessary to amplify the cDNA end corresponding to a specific ZPC transcript to increase the specificity of the PCR reaction and decrease amplification of non-target cDNAs (Figure 10). After the first round of PCR using the AP1 and the gene-specific primers, an aliquot was used in a second PCR reaction using the same AP1 primer but this time using a different gene-specific primer that should be internal to the first round PCR product thereby generating a shorter PCR product. The advantage of nested PCR is that if a mixture of PCR products were
amplified (e.g. multiple bands or smear on an agarose gel), the second nested PCR should result in the amplification of only the targeted ZPC cDNA since it is very unlikely that the non-specific products will also be exponentially amplified once again. After appropriately sized PCR products are observed by gel electrophoresis, they were then ligated, transformed into *E. coli* bacteria, and sequenced as earlier described.

Figure 10. Nested PCR: Step One, the DNA target template is bound by the first set of primers, Step Two, an aliquot of PCR products from the first PCR reaction are subjected to a second PCR run with a second new set of primers (www.wisconsinlab.com/images/Nested_PCR).

DNA Sequence Analyses

Nucleotide sequences from recombinant plasmids submitted to a commercial sequencing facility were analyzed and assembled using Lasergene software programs. First, the program Seqman was used to trim the vector and primer sequences and to check
the chromatogram of each sequence for discrepancies. The edited sequences were then subjected to Blast searches (http://www.ncbi.nlm.nih.gov/blast/) to test whether the query sequence matched most closely to ZPC genes in the database. The edited DNA sequences were compared to each other using the multiple sequence alignment program Megalign to determine unique ZPC clones within each library for cataloguing purposes. For example in *X. laevis*, sequences derived from recombinant clones that were >98% identical were assigned to single gene clusters. Clusters were labeled using a nomenclature system such as xlZPC1, xlZPC2 and so on. Edited full length ZPC cDNA sequences were assembled using Lasergene's Seqman program by comparing the overlapping portions of the middle section and the corresponding 5’ and 3’ ends and then splicing them together. The overlapping segments were greater than 50 bases in length (with 100% identity) to ensure that it was indeed derived from the same cDNA sequence.

**ZPC Phylogenetic Analyses**

Two different phylogenetic analyses were performed in an attempt to clarify the relationships of the ZPC cDNAs within and among two frog species. In order to perform the phylogenetic analyses, the edited *X. borealis* and *X. laevis* nucleotide sequences were translated to amino acid sequences and then aligned using the program CLUSTALW (http://www.ebi.ac.uk/clustalw/) with default settings. The BOXSHADE software program was used on the output of the CLUSTALW alignment to create a shaded alignment figure to highlight the areas that are conserved throughout the sequences. Phylogenetic analyses were performed using the CLUSTAL alignment and the Mega4(http://www.megasoftware.net/) and MrBayes
(http://mrbayes.csit.fsu.edu/index.php) software which produce consensus trees based on Neighbor Joining (distance-based) and Maximum Likelihood Parsimony (parsimony-based) methods, respectively.

Results/Discussion

I. *Xenopus* ZPC cDNA Cloning and Sequencing

A. Cataloging the Middle Conserved Portion of ZPC

Figure 11 shows the amplification of ZPC from a *X. borealis* ovary cDNA library using ZPC1, ZPC2, ZPC3 and ZPC4 degenerate primers. The forward direction primers were ZPC1 and ZPC2, whereas the reverse direction primers were ZPC3 and ZPC4. The size of DNA amplified by degenerate primers corresponds to the distance of forward and reverse primers. For example, since ZPC1 (forward primer) and ZPC3 (reverse primer) have the furthest distance from each other compared to other primer combinations (ZPC2/ZPC4, ZPC1/ZPC4 and ZPC2/ZPC3), the DNA amplified by these primers should have the biggest size (600 bp length, Figure 5 and 11). The combination of all these different degenerate primers adds specificity and increases the number of hits for ZPC and decreases the number of hits for non-ZPC genes.
Figure 11. PCR amplification of *X. borealis* ZPC cDNAs from an ovary cDNA library. PCR was performed using degenerate forward primers ZPC1 (1), ZPC2 (2) and the reverse direction primers ZPC3 (3), and ZPC4 (4). Lane A was loaded with a size standard (100 bp ladder); the 500 bp ladder band is designated by the arrow. The bands in lanes B, C, D and E are 500 bp, 600bp, 400bp, and 550bp, respectively, which were the expected sizes.

The plasmid DNA from 85 recombinant clones for *X. laevis* and 80 for *X. borealis* was sequenced using the universal primers found on the TA cloning vector. A BLAST analysis was then performed on each of the returned sequences which resulted in 34 and 40 of the sequences from *X. laevis* and *X. borealis* clearly belonging to the ZPC gene family (closest match was to the *X. laevis* ZPC sequence accession no. U44952). After removing the vector and primer sequences, the cDNA sequences were compared to each other by a combination of pairwise and multiple sequence alignments. Sequences that were >98% identical were grouped together (clustered) for further examination since they were likely to derived from the same gene. It should be noted that the Advantage Polymerase© used in the PCR reactions has a very low error rate (1 error out of 1.5 x 10⁵ bases) due to having proofreading activity (3’ to 5’ exonuclease) so it is very unlikely
that sequences with less than 98% identity (i.e. more than 2 nucleotide differences out of 100 bases) would have resulted from PCR errors.

In order to resolve the remaining sequence differences or conflicts within each cluster (e.g. G in one sequence versus an A in another), sequences alignments were analyzed in detail for DNA sequencing discrepancies. In the few cases where sequence conflicts were present, weight was given to the sequence chromatograms with the highest resolution and/or to the sequence found in the majority of the clones. As expected, most of these conflicts were noted near the ends of sequence reads where DNA chromatograms have less resolution. The sequences for the these ZPC clusters were also re-examined and confirmed after PCR amplification and sequencing of the 5’ and 3’ cDNA ends was accomplished (as noted later).

After determining the most representative sequence for each cluster, 5 different ZPC genes for *X. borealis* and 4 ZPC genes for *X. laevis* were distinguished (see Table 1 and 2). The sequences are named xlZPC.1a, xlZPC.1b, and xlZPC.2, xlZPC.3 for *X. laevis* and xbZPC.1c, xbZPC.1d, xbZPC.1e, xbZPC.2, xbZPC.3 for *X. borealis*. The *X. laevis* xlZPC.1a sequence was so named since it matched exactly (100% identity over 80 bases) to the *X. laevis* ZPC sequence deposited previously in the Genbank database (accession no. U44952) when a BLAST search was performed. The rationale for the nomenclature as listed above is that it is based on the phylogenetic relationships to the published xlZPC.1a sequence which will be discussed later in this chapter (Figure 19).

In addition, the relative abundance of the ZPC cDNAs found in the cDNA library is also tabulated in Tables 1 and 2. Although quantitative PCR would provide a much
more accurate assessment, the number of times a particular cDNA sequence was observed in the library relative to the total number of clones analyzed does provide a provisional basis to assess abundance especially since clones were picked at random for sequence analysis. For *X. laevis*, clone xlZPC.1a is seemingly the predominant cDNA found in the library (18 sequences) whereas xlZPC.3 seems to be the least represented (1 sequence). As mentioned above, the xlZPC.1a is the sequence that was described by Yang and Hedrick (1997) and is currently the only sequence reported within the sequence databases for *X. laevis*. For *X. borealis*, the xbZPC.1d cDNA seems to be the predominant transcript while the xbZPC.3 cDNA is the least represented in the library.

<table>
<thead>
<tr>
<th><em>X. laevis</em> clones</th>
<th>xlZPC.1a</th>
<th>xlZPC.1b</th>
<th>xlZPC.2</th>
<th>xlZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ZPC1/ZPC3</em></td>
<td><strong>10</strong></td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td><em>ZPC2/ZPC4</em></td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>ZPC2/ZPC3</em></td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total clones</td>
<td>18</td>
<td>10</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Distribution of *X. laevis* ZPC cDNAs. Random colony transformants from degenerate PCR were picked, plasmid DNA purified, and sequenced. Sequences were delineated into clusters identifying unique ZPC genes (e.g. xbZPC.1e). The number of times a particular ZPC gene was identified within the cDNA library is tabulated providing an incidence of relative expression. The combination of degenerate ZPC primers used for the PCR amplification is listed in column 1. More ZPC1/ZPC3 clones were submitted for sequencing because these primers resulted in the longest gene product.
Table 2. Distribution of *X. borealis* ZPC cDNAs. Random colony transformants from degenerate PCR were picked, plasmid DNA purified, and sequenced. Sequences were delineated into clusters identifying unique ZPC genes (e.g. xbZPC.1e). The number of times a particular ZPC gene was identified within the cDNA library is tabulated providing an incidence of relative expression. The combination of degenerate ZPC primers used for the PCR amplification is listed in column 1. More ZPC1/ZPC3 clones were submitted for sequencing because these primers resulted in the longest gene product.

<table>
<thead>
<tr>
<th></th>
<th>xbZPC.1e</th>
<th>xbZPC.1d</th>
<th>xbZPC.1c</th>
<th>xbZPC.2</th>
<th>xbZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ZPC1/ZPC3</em></td>
<td><strong>8</strong></td>
<td>9</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>ZPC2/ZPC4</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>ZPC2/ZPC3</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total clones obtained from degenerate primers</td>
<td>8</td>
<td>19</td>
<td>1</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

The percent identity of amino acid pairwise comparison of *X. laevis* and *X. borealis* middle section ZPC clones is tabulated in Table 3 and 4 respectively. As can be observed from the tables, the percent of identical residues for ZPC clones within their middle sections ranged from 98.1-62.6% in *X. laevis* and 96.3–42.5% amino acid identity in *X. borealis*. For *X. laevis*, clone xlZPC.1a and xlZPC.1b have the most amino acid identity whereas clone xlZPC.1a and xlZPC.3 have the least amino acid identity (Table 3). For *X. borealis*, clone xbZPC.1e and xbZPC1d share greatest amino acid identity (96.3%) while clone xbZPC.1e and xbZPC.3 share the least amino acid identity (42.5%, Table 4). Moreover, the amino acid alignment of the middle sections of *Xenopus* ZPC clones (which is the ZP domain) from both species indicates that this region is relatively conserved between species (black and gray regions in Figure 12).
consistent with the pairwise comparison scores of percent identities of ZPC middle sections.

Table 3 Pairwise comparison of the percent amino acid identity of the ZP domain of *X. laevis* ZPC cDNAs.

<table>
<thead>
<tr>
<th>Percent Identity</th>
<th>xIZPC.1a</th>
<th>xIZPC.1b</th>
<th>xIZPC.2</th>
<th>xIZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98.1</td>
<td>93.6</td>
<td>62.6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>91.6</td>
<td>93.6</td>
<td>62.6</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>62.6</td>
<td>60.6</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4. Pairwise comparison of the percent amino acid identity of the ZP domain of *X. borealis* ZPC cDNAs

<table>
<thead>
<tr>
<th>Percent Identity</th>
<th>xbZPC.1c</th>
<th>xbZPC.1d</th>
<th>xbZPC.1e</th>
<th>xbZPC.2</th>
<th>xbZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96.3</td>
<td>93.3</td>
<td>91.3</td>
<td>61.8</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>93.3</td>
<td>90.5</td>
<td>70.3</td>
<td>61.8</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>60.6</td>
<td>42.5</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

"XbZPC.1c" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XbZPC.2" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XlZPC.2" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XlZPC.1a" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XlZPC.1b" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XbZPC.1d" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XbZPC.1e" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XlZPC.3" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
"XbZPC.3" 1 SPARSNLFGDTPHEASINIANHAPMTTVDSCVATTPDVNSNPRTYIIINQCLID
Figure 12. Protein alignment of the ZP domain for *X. borealis* and *X. laevis* ZPC cDNAs. The translated cDNA sequences were aligned using ClustalW. Sequences shaded in black or gray denote position in which 7/10 sequences were either identical or conservative substitutions, respectively. The arrowheads indicate the conserved cysteine residues.

B: Cloning 3’ and 5’ ends

The 3’ ends of all *X. laevis* and *X. borealis* clones were cloned and analyzed using a nested PCR strategy. First round amplification of the 3’ cDNA ends was performed using the first sets of primer. The gene specific primers were chosen based on unique regions of sequences found in multiple alignments of all middle ZP domain sections. The dilutions of the first reaction were then used as template for a second PCR reaction using an internal (or upstream) gene-specific primer that would bind to internal sites within the target DNA.

As can be observed in the protein sequence alignments (Figures 13 and 14), the middle ZP domains of the *Xenopus* ZPC cDNAs are more conserved within each species while the 3’ ends have more sequence divergence (i.e. non-shaded regions) and gaps. In *X. laevis*, clones xlZPC.1a and xlZPC.1b seem to be the most closely related to each
other since they have less divergent amino acid regions (Figure 13) and share greater amino acid identity with each other (98.1% identical, Table 3) than other ZPC clones. As for *X. borealis*, clones xbZPC.1e and xbZPC.1d are most closely related to each other and share 96.3% amino acid identity (Figure 14 and Table 4). Clone xZPC.3 of *X. laevis* and xbZPC.3 of *X. borealis* seem to have more sequence differences and gaps than other clones within each species.

Figure 13. Alignment of middle sections and 3’ ends of ZPC glycoproteins within *X. laevis* cDNA library. The white shaded regions are the divergent regions while the sequences shaded in black and gray denote position in which 70% sequences were either identical or conservative substitutions, respectively. The arrowheads indicated conserved cysteine residues.
Figure 14. Alignment of middle sections and 3' ends of ZPC glycoproteins within *X. borealis* cDNA library. The white shaded regions are the divergent regions while the sequences shaded in black and gray denote position in which 70% sequences were either identical or conservative substitutions, respectively. The arrowheads indicated conserved cysteine residues.

The 5’ ends, however, were more difficult to obtain from cDNA libraries even though many primers were designed for PCR amplification. These gene specific primers (i.e. exact matches rather than degenerate) were chosen based on unique regions of sequences found in multiple alignments of all middle ZP domain sections. First round amplification of the 5’ cDNA ends was performed using relaxed conditions (lower
annealing temperatures) so as to ensure amplification of the ZPC 5’ end but recognizing that non-specific amplification would also occur. Thus, dilutions of the first reaction were then used as template for a second PCR reaction using an internal (or upstream) gene-specific primer and more stringent conditions (annealing temperature more closely matching the $T_m$ of the primer). Using this nested PCR strategy (which in essence was the same used to clone the 3’ ends), the 5’ cDNA ends for XbZPC.2, XbZPC.1d and partial 5’ cDNA sequences for XlZPC.1b and XbZPC.1c were accomplished (primer sequences are provided in appendix). Unfortunately, the 5’ cDNA ends for the rest of the clones were not obtained after many PCR attempts and subsequent cloning of non-ZPC PCR products (false positives). For the problematic 5’ cDNA projects, at least 10 different specific primers were designed to attempt to PCR amplify their respective 5’ cDNA ends.

In general, 5’ RACE is a more difficult endeavor for a variety of documented reasons. Firstly, the 5’ end of a particular cDNA is often not fully reverse transcribed since polymerase activity proceeds from the 3’ poly A tail up towards the 5’ end and thus is not represented equally as well in cDNA libraries. Secondly, stable secondary structures are often encountered found near the 5’ end of mRNA molecules and reverse transcriptase has difficulty reading through these regions. In addition to these typical issues, ZPC genes are very similar in their middle sections (Table 3 and 4) which it makes it difficult to design unique primers to clone 5’ ends. Taken together, these factors prevented the retrieval of those particular 5’ ends of ZPC transcripts from *Xenopus* ovary cDNA libraries. It is likely that new ovary cDNA libraries may need to be generated.
using special RNA structure denaturants such as methyl-mercuric hydroxide during cDNA synthesis to enrich for the 5’ cDNA ends (Snead et al., 1996).

As for the full length ZPC cDNAs that were cloned and sequenced, they all include a start methionine residue, a signal peptide for secretion, two potential N-linked glycosylation sites near the 5’ ends, several potential O-linked glycosylation sites (serine/threonine sites), furin cleavage sites, transmembrane domains, stop codon, and a poly A tail (not shown in the figures). These ZPC gene family features are shown in Figures 15 and 16.

The alignment of all ZPC genes from *X. laevis* and *X. borealis* shows relatively high level of conserved regions with the middle section specifically being more conserved between species (Figure 16). The C-terminal regions, however, seem to be more divergent and more discrepancy and gaps are seen in regions close to the end of molecules suggesting that this portion of the molecules is rapidly changing between two frog species. Moreover it can be observed from the alignment of all ZPC genes that, cysteine residues (found in 8 positions, located in ZP domain) are conserved among and between two *Xenopus* species (shown with arrowheads in figures). Cysteine residues play a valuable role by cross-linking proteins, which increase the protein stability in the harsh extracellular environment. Therefore cysteine residues play an essential role in structure and function of ZPC proteins and mutating any of the conserved cysteine amino acid may result in nonfunctional corresponding protein.
Figure 15. Analysis of the ZPC primary sequence structure. Alignment of the whole length ZPC clones. The start methionine is shown by ● symbol. Signal peptide amino acids are indicated by double solid underline and the putative signal peptidase cleavage site by a solid arrowhead (Ala-Glu). The two potential N-linked glycosylation sites are highlighted and also indicated with ∞ symbols. Potential furin-like cleavage residues are indicated with three black diamond suits (♦) the transmembrane domain is indicated by a dotted line; and the ZP domain is indicated by solid line. The four arrows show the degenerate primer positions (ZPC1, ZPC2, ZPC3 and ZPC4)(ZPC feature were adapted from Yang and Hedrick (et al, 1997).
Figure 16. Amino acid sequences alignment of ZPC genes from *X.laevis* and *X.borealis* cDNA libraries (including the middle section and 3' ends). The translated ZPC DNA sequences were
aligned using CLUSTALW and shaded using the BOXSHADE program. Sequences shaded in black or gray denote position in which 7/10 sequences were either identical or conservative substitutions, respectively. Arrowpoints refer to conserved cysteine positions. The furin cleavage site is indicated by three stars and the putative transmembrane domain is indicated by a dotted line. The solid underline refers to the conserved ZP domain.

III. Phylogenetic relationships of *Xenopus* ZPC cDNAs

Neighbor joining (distance-based) and Maximum Likelihood (model-based) phylogenetic analyses were performed on the sequence alignment of *Xenopus* ZPC cDNA sequences to infer their evolutionary relationships (i.e. orthologous vs paralogous). The protein alignment of the middle portion (ZP domain) and 3’ ends was used for these analyses (refer to figure 14) since all clones had this information available (i.e. complete dataset). Protein sequence alignments were used to generate these gene trees since they are phylogenetically more informative than DNA for protein-coding genes. Alignment regions with gaps (i.e. insertions or deletions) were removed as recommended by most phylogenetic algorithm software instructions so that the data analyzed is restricted to portions of the proteins that are available for all the genes being studied. As mentioned previously, the xlZPC.1a cDNA was named as such since this corresponds to the only published ZPC cDNA for *X. laevis*. The nomenclature for all of the other *Xenopus* ZPC cDNAs are thus based on their evolutionary relationships to the xlZPC.1a sequence as will be described in the following sections.

All Neighbor Joining distance-based trees constructed were consistent with the pairwise scores of percent identities between clones (refer to Tables 3, 4 and 5). For the four *X. laevis* ZPC cDNAs (Figure 17), xlZPC.1a and xlZPC.1b are most closely related to each other in the gene tree (98.7% amino acid identity, Table 5), whereas xlZPC.2 is
more divergent to this cluster (89.7% amino acid identity to xlZPC.1a) and xlZPC.3 is the most divergent (64.4% amino acid identity to xlZPC.1a). Similarly, the neighbor joining tree for the five *X. borealis* ZPC cDNAs (figure 18) had a cluster of cDNAs that were named using the ZPC.1 series nomenclature (i.e. xbZPC1c, xbZPC1d, and xbZPC1e) due to their relationships to xlZPC.1a (figure 19). These *X. borealis* ZPC.1c, d, and e clones all were within the 95-96.3% amino acid identity range to each other, and of these clones, the xbZPC.1c clone was slightly more closely related to xlZPC.1a (92.3%). The other two *X. borealis* clones followed a similar pattern to those found in *X. laevis* and were thus named in a consistent manner. The xbZPC.2 and xbZPC.3 clones were 90.4% and 65% identical to xbZPC.1c respectively. As can be seen in figure 19, the *X. laevis* and *X. borealis* ZPC.2 clones cluster together as do the ZPC.3 clones.

A Neighbor Joining tree was also constructed using the sequence alignment of human, mouse and zebrafish ZPC cDNAs as outgroups to root the tree (Figure 20 and 21). As expected, zebrafish was the most divergent sequence, ranging from 33.4-39.7% amino acid identity to all sequences examined (Table 6). As for the human and mouse sequences, they shared a slightly higher percent identity with the cluster of *Xenopus* ZPC.1 and ZPC.2 cDNAs (ranging from 48.8-52%) than with the ZPC.3 sequences (range 47.7-49%, Table 6). Another interesting result was that the *X. laevis* ZPC.1a, b sequences were more closely related to the *X. laevis* and *X. borealis* ZPC.2 sequences rather than the cluster of xbZPC1c, d, and e sequences. This was unexpected, however the bootstrap confidence value for this arrangement was very low (53%) which indicates
this is not a robust result. It should be noted though that the *Xenopus* ZPC.1 sequences are highly related to ZPC.2 sequences as was pointed out earlier.

Maximum Likelihood trees were also constructed using the same alignment datasets to assess the gene relationships using a non-distance based methodology. The Maximum Likelihood (ML) method is an advanced method and is computationally intensive. The ML approach to determining the branch topology and branch lengths takes into consideration the likelihood of an ancestral relationship for each residue in the alignment which includes some model of the amino acid substitution process. A ML tree was constructed for all *Xenopus* ZPC sequences using the mouse ZPC sequence as an outgroup. The ML gene tree is consistent with many of the findings from the neighbor joining tree (Figure 22). For example, the *Xenopus* ZPC.3 sequences are closely related and so are the xlZPC.1a and b sequences. The *Xenopus* ZPC.2 sequences are more closely related to ZPC.3 sequences than the ZPC.1 sequences. Also consistent with the neighbor joining trees is that the ZPC.1 sequences are more closely related to each other, and the xbZPC.1c sequence is more closely related to the xlZPC.1a and b sequences. The most interestingly result is that the mouse ZPC sequence was most closely related to the *Xenopus* ZPC.3 sequences. This result suggests that the ancestral *Xenopus* ZPC.3 gene was maintained in the lineage leading to mammals whereas all of the other ZPC genes were not (e.g pseudogenes or deleted).

Thus, the naming scheme proposed in this thesis is based on the consistent evolutionary relationships derived from the neighbor joining and maximum likelihood analyses. It seems likely that the ZPC.2 and ZPC.3 genes are orthologous genes from *X.*
*laevis* and *X. borealis*. It is unclear as to the orthologous relationship between ZPC.1 sequences at this point since they did not show a clear pattern. It seems probable that they were derived from a common ancestor and may have an allelic relationship (found at the same locus and incorporated evolutionary change) since there could be a possible 4 alleles found at a particular locus.

Figure 17. The Neighbor-joining tree showing phylogenetic relationship of *X. laevis* ZPC. Bootstrap re-sampling of the tree was performed using 1000 replicates.

Figure 18. Neighbor-joining tree showing phylogenetic relationship of *X. borealsi* ZPC genes. Bootstrap re-sampling of the tree was performed using 1000 replicates.
Figure 19. Neighbor-joining tree showing the ZPC gene relationships from to *Xenopus* species. Bootstrap re-sampling of the tree was performed using 1000 replicates.
Figure 20. Amino acid sequence alignment of ZPC genes from *X.laevis*, *X.borealis* and human, mouse and zebrafish as outgroups. The translated ZPC cDNA sequences from *X. laevis*, *X. borealis* zebrafish, human and mouse were aligned by CLUSTALW and shaded by BOXSHADE. Sequences shaded in black or gray denote position in which 7/10 sequences were either identical or conservative substitutions. Arrows point to conserved cysteine positions in the ZP domain. The solid underline denotes the conserved ZP domain. The double underlines indicate the putative sperm binding regions (adapted from Swanson’s et al, 2002).

Figure 21. ZPC gene tree. The Neighbor-joining tree including various ZPC genes from *X. borealis*, *X. laevis*, mouse, human and zebrafish species. The ClustalW alignment in Figure 20 was used for the analysis. Bootstrap re-sampling was performed using 1000 replicates.
Figure 22. Maximum likelihood phylogenetic tree of the ZPC genes from *Xenopus laevis*, *Xenopus borealis* and Mouse. The ClustalW alignment in Figure 20 was used for the analysis but with inclusion of only the mouse sequence as an outgroup.
Table 5. Percentage of amino acid identity between *X. laevis* and *X. borealis* middle and 3’ ends of ZPC cDNAs.

<table>
<thead>
<tr>
<th></th>
<th>XlZPC.1a</th>
<th>XlZPC.1b</th>
<th>XlZPC.2</th>
<th>XlZPC.3</th>
<th>XbZPC.1d</th>
<th>XbZPC.1e</th>
<th>XbZPC.2</th>
<th>XbZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98.7</td>
<td>98.7</td>
<td>64.4</td>
<td>90.4</td>
<td>92.0</td>
<td>92.3</td>
<td>89.8</td>
<td>65.9</td>
</tr>
<tr>
<td>1.1</td>
<td>90.3</td>
<td>63.9</td>
<td>90.4</td>
<td>92.0</td>
<td>92.3</td>
<td>90.0</td>
<td>65.5</td>
<td></td>
</tr>
<tr>
<td>10.9</td>
<td>91.1</td>
<td>62.5</td>
<td>87.1</td>
<td>88.2</td>
<td>89.0</td>
<td>89.0</td>
<td>64.3</td>
<td></td>
</tr>
<tr>
<td>48.8</td>
<td>52.1</td>
<td>62.5</td>
<td>52.2</td>
<td>95.1</td>
<td>96.3</td>
<td>90.3</td>
<td>65.3</td>
<td></td>
</tr>
<tr>
<td>10.1</td>
<td>9.7</td>
<td>13.9</td>
<td>95.1</td>
<td>96.3</td>
<td>90.3</td>
<td>65.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.3</td>
<td>8.2</td>
<td>12.6</td>
<td>50.0</td>
<td>4.7</td>
<td>95.0</td>
<td>87.2</td>
<td>64.4</td>
<td></td>
</tr>
<tr>
<td>8.2</td>
<td>8.0</td>
<td>12.9</td>
<td>52.8</td>
<td>3.6</td>
<td>5.0</td>
<td>91.8</td>
<td>65.0</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>10.7</td>
<td>9.7</td>
<td>49.6</td>
<td>10.2</td>
<td>13.9</td>
<td>8.7</td>
<td>64.6</td>
<td></td>
</tr>
<tr>
<td>46.1</td>
<td>46.7</td>
<td>48.9</td>
<td>8.7</td>
<td>46.9</td>
<td>48.8</td>
<td>47.8</td>
<td>48.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Percent amino acid identity between *X. laevis*, *X. borealis*, human, mouse and zebra fish middle and 3’ ends of ZPC cDNAs.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mouse</th>
<th>Fish</th>
<th>XbZPC.1c</th>
<th>XbZPC.1d</th>
<th>XbZPC.1e</th>
<th>XbZPC.2</th>
<th>XbZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74.4</td>
<td>34.7</td>
<td>51.8</td>
<td>51.8</td>
<td>52.0</td>
<td>51.3</td>
<td>49.0</td>
<td>51.5</td>
</tr>
<tr>
<td>31.7</td>
<td>33.4</td>
<td>50.1</td>
<td>50.3</td>
<td>49.3</td>
<td>48.7</td>
<td>47.9</td>
<td>51.0</td>
<td>50.8</td>
</tr>
<tr>
<td>170.3</td>
<td>174.4</td>
<td>39.7</td>
<td>39.3</td>
<td>39.7</td>
<td>39.8</td>
<td>39.7</td>
<td>39.5</td>
<td>39.4</td>
</tr>
<tr>
<td>78.4</td>
<td>82.9</td>
<td>129.4</td>
<td>95.1</td>
<td>96.3</td>
<td>90.3</td>
<td>65.7</td>
<td>90.4</td>
<td>90.4</td>
</tr>
<tr>
<td>78.8</td>
<td>82.5</td>
<td>134.0</td>
<td>4.7</td>
<td>95.0</td>
<td>87.2</td>
<td>64.5</td>
<td>92.0</td>
<td>92.0</td>
</tr>
<tr>
<td>78.6</td>
<td>86.5</td>
<td>130.8</td>
<td>3.6</td>
<td>5.0</td>
<td>91.8</td>
<td>65.3</td>
<td>92.3</td>
<td>92.3</td>
</tr>
<tr>
<td>81.4</td>
<td>88.7</td>
<td>137.1</td>
<td>10.2</td>
<td>13.9</td>
<td>8.7</td>
<td>64.9</td>
<td>89.8</td>
<td>90.9</td>
</tr>
<tr>
<td>88.8</td>
<td>91.5</td>
<td>148.0</td>
<td>46.2</td>
<td>48.6</td>
<td>47.3</td>
<td>48.0</td>
<td>66.0</td>
<td>65.4</td>
</tr>
<tr>
<td>80.3</td>
<td>81.0</td>
<td>140.3</td>
<td>10.1</td>
<td>8.3</td>
<td>8.2</td>
<td>11.0</td>
<td>45.9</td>
<td>98.7</td>
</tr>
<tr>
<td>79.9</td>
<td>81.3</td>
<td>134.3</td>
<td>9.9</td>
<td>8.2</td>
<td>8.0</td>
<td>10.7</td>
<td>46.9</td>
<td>1.1</td>
</tr>
<tr>
<td>80.0</td>
<td>87.6</td>
<td>147.6</td>
<td>13.9</td>
<td>12.5</td>
<td>12.9</td>
<td>9.7</td>
<td>48.1</td>
<td>10.8</td>
</tr>
<tr>
<td>90.8</td>
<td>91.7</td>
<td>153.7</td>
<td>52.4</td>
<td>49.9</td>
<td>52.8</td>
<td>49.6</td>
<td>8.7</td>
<td>49.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>XlZPC.1b</th>
<th>XlZPC.2</th>
<th>XlZPC.3</th>
<th>XlZPC.1a</th>
<th>XlZPC.1b</th>
<th>XlZPC.2</th>
<th>XlZPC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51.6</td>
<td>51.4</td>
<td>48.3</td>
<td>90.4</td>
<td>90.4</td>
<td>63.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>90.8</td>
<td>91.7</td>
<td>52.4</td>
<td>52.8</td>
<td>49.6</td>
<td>8.7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>52.4</td>
<td>51.3</td>
<td>49.9</td>
<td>52.8</td>
<td>49.6</td>
<td>8.7</td>
<td>49.0</td>
</tr>
<tr>
<td>4</td>
<td>51.8</td>
<td>51.8</td>
<td>52.0</td>
<td>51.3</td>
<td>49.0</td>
<td>51.5</td>
<td>51.6</td>
</tr>
<tr>
<td>5</td>
<td>51.6</td>
<td>51.4</td>
<td>48.3</td>
<td>90.4</td>
<td>90.4</td>
<td>63.5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>90.8</td>
<td>91.7</td>
<td>52.4</td>
<td>52.8</td>
<td>49.6</td>
<td>8.7</td>
<td>49.0</td>
</tr>
<tr>
<td>7</td>
<td>52.4</td>
<td>51.3</td>
<td>49.9</td>
<td>52.8</td>
<td>49.6</td>
<td>8.7</td>
<td>49.0</td>
</tr>
<tr>
<td>8</td>
<td>51.8</td>
<td>52.0</td>
<td>51.3</td>
<td>49.0</td>
<td>51.5</td>
<td>51.6</td>
<td>51.4</td>
</tr>
<tr>
<td>9</td>
<td>90.4</td>
<td>90.4</td>
<td>63.5</td>
<td>90.4</td>
<td>90.4</td>
<td>63.5</td>
<td>62.6</td>
</tr>
</tbody>
</table>
Chapter 3

MOLECULAR AND EVOLUTIONARY ANALYSES OF *XENOPUS* ZPC CDNAS

One of the primary goals of this study was to examine whether there is any evidence for positive Darwinian selection (adaptive evolution) to have been an influence on amino acid sites encoded by *Xenopus* ZPC (sperm-binding) cDNAs. Inevitably, orthologous or paralogous genes will accumulate mutations over time, but the real question is whether the type of selectional pressures acting on the protein-coding sites within a particular gene can be identified. In essence, negative (purifying) selection hinders the spread of alleles with destructive mutations, positive (Darwinian) selection promotes the spread of alleles with beneficial mutations, and neutral selection does not hinder nor promote the spread of alleles with neutral mutations (i.e. do not affect the protein's function). The amino acid positions that are more important for the protein's structure and/or function are generally the regions that have a positive or negative influence when changed. However, the protein can be segmented into regions that are structurally/functionally constrained which are mixed in with regions of neutrality (mutations in these regions being common). Thus, amino acid sites within a gene are not uniformly subjected to a single type of selectional pressure, but are a mixture of types. Tests to detect signals for positive selection need to take into account this issue of site variation with respect to selection.
The evolution of most protein families is due primarily to a combination of negative and neutral selection on sites. Sites subjected to negative selection tend to be found in regions that are structurally and/or functionally important since it is detrimental to change the physico-chemical properties of the amino acid found at these sites. Thus the amount of amino acid changes are very limited at sites subjected to negative selection however silent site mutations (mutations that do not result in amino acid changes due to codon degeneracy) will still occur. In other words, the rate of nonsynonymous change at an amino acid site (dN) will be less the incidence of synonymous mutations (dS). On the other hand, amino acid sites that are neutral will accumulate amino acid changes as frequently as silent site mutations (dN = dS) and will accumulate in the population in accordance with genetic drift. Positive selection does not appear to be as prevalent a selectional force within most genes. The distinguishing characteristic of positive selection is that mutations that result in amino acid changes will be in a higher ratio than silent site mutations (dN>dS) since there is a selectional force acting that preserves these amino acid substitutions due to their beneficial effects.

Phylogenetic Analysis by Maximum Likelihood (PAML) is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood (Yang et al, 1999). The package includes the following programs: BASEML, BASEMLG, CODEML, EVOLVER, PAMP, YN00, MCMVTREE and CHI2. CODEML in particular, is used to estimate synonymous and nonsynonymous substitution rate and provides statistical tests for the detection of positive Darwinian selection in protein-coding DNA sequences. A number
of likelihood models for codon substitution are implemented in CODEML for a combined analysis of data sets from multiple gene loci by measuring \( \omega \) ratios (\( dN/dS \)) across the lineage (Yang et al 1996). The CODEML program was used by Swanson's group (Swanson et al, 2001) in their foundational study to detect the positive selection at amino acid sites within eight mammalian ZPC genes so we thought it appropriate to use the same program. The same eight mammalian ZPC cDNAs included in Swanson’s study were used as a positive control data set for our studies to prove we could replicate their results.

The \( \omega \) ratio (\( dN/dS \)) calculated within CODEML measures the direction and magnitude of selection on amino acid changes, with values of \( \omega < 1, = 1, \) and \( > 1 \) indicating negative, neutral, and positive selection, respectively. For example, the \( \omega \) ratio calculated for the residues located in the MHC antigen recognition site (by crystal structure mapping) is \( > 1 \), which indicates positive selection (Hughes et al, 1988). (Figure. 5 left). CODEML statistical tests the likelihood as to whether a site can be distinguished as being influenced by negative, neutral or positive selection based on a comparison of evolutionary models and the best fit to the data. The program uses likelihood ratio tests to compare two models to the sequence alignment data; a null model without positive selection with an alternative model that incorporates positive selection. If the alternative model provides a significantly better fit to the data (\( p \) value \( \geq .05 \)), then positive selection is inferred at that particular site. Additional details as to the models used in the statistical tests are provided in the following material and methods section.
Although, it would have been preferred to analyze the entire protein coding portions of the *Xenopus* ZPC cDNA, we have a complete dataset for the latter two-thirds of the cDNA (i.e. conserved ZP domain and 3' ends) to analyze for evidence of positive selection. Since the mammalian sperm combining site is located in the 3’ end of the ZPC gene (C-terminus portion of protein) and Swanson detected positive selection in this region, it will be relevant to study the molecular evolution of this same region of amphibian ZPC genes and compare whether the same sites are under similar selectional pressures.

**Materials and Methods**

Amino acid sequences were aligned by using the CLUSTALW software program (http://www2.ebi.ac.uk/clustalw/). The PAL2NAL program (http://www.bork.embl.de/pal2nal/) was then used to convert the amino acid alignment into its corresponding DNA codon alignment (Figure 17). The resulting codon alignment was then used for the calculation of ω ratios (dN/dS) and tests of positive selection using the CODEML program within the PAML software package (Yang et al, 1999).

The CODEML analysis for the presence of sites under positive selection consists of two major steps. The first step uses likelihood ratio tests to simply detect the presence of positive selection, (i.e. presence of sites with ω>1) by comparing the fit of the data to a likelihood model that does not allow for positive selection (null model) to a model that does (alternate model). We will use the same two ratio tests that Swanson et al (2001) used in his study. The first test is uses the null model M0 which assumes one ω for all sites, and compares its likelihood to the alternate model M3 which allows for ω to vary
across sites and includes all three site classes (negative, neutral and positive). For the M3 model, the ω values are independently estimated from the data. The second likelihood ratio test for the presence of positive selection uses more computationally intensive and sophisticated models for comparison. In this case, the null model, M7, incorporates the potential for ω ratios to vary but only within the range of neutral to negative selection (ω ratio less than or equal to 1, no possibility of positive selection), whereas the alternate model, M8, includes the potential for positive selection (full range of ω ratios).

After determining that there is a signal for positive selection, the second step is to employ a more refined test to pinpoint where the particular sites under positive selection are located within the protein (ω ratios greater than 1). In order to do this, the M7 vs M8 test is used to calculate a probability for each site that it is from the positive selection class (ω ratio greater than 1) when compared to the null hypothesis. Amino acids detected to be under positive selection will have p values of .05 or below (violating the assumptions of the null M7 model). These probabilities are performed using Bayes theorem to calculate the posterior probabilities that each site, given the data at that site, are from the different ω classes (Nielsen and Yang et al, 1998).

Results/Discussion

When the dN/dS ratios (ω) for the aligned sites were averaged across the *Xenopus* codon alignment, the ratio was smaller than 1 as expected which is similar to the mammalian study (Table 7). As pointed out earlier, the majorities of amino acid sites are either neutral or are constrained by negative selection. Even if a protein did have sites subjected to positive selection, the averaged ω ratio would not been observed to be higher
than 1 since a larger number of the sites are usually constrained sites (negative selection). Therefore the average \( \omega \) is not a sensitive measure of selection.

The presence of sites under positive selection was then tested for by using likelihood ratio tests using models that account for positive selection compared to ones that do not. Both likelihood ratio tests (model M0 vs. M3 and also M7 vs M8) for ZPC genes showed evidence for variation in \( \omega \) between sites, with a class of sites having \( \omega>1 \), indicating positive selection (Table 7). The M0 (null model) versus M3 (alternative model) ratio test detected positive selection at sites within the \( X. \ laevis \) and \( X. \ borealis \) ZPC codon alignment with significance at the 1% level. The proportion of sites potentially under positive selection is 2.7% (\( p=0.027 \)) with \( \omega=3.38 \).

As a more refined likelihood ratio test, model M8 was also compared to the null model M7 which also detected the presence of positive selection (significance at the 5% level). The estimated \( \omega \) ratios for the M8 suggested the presence of sites with \( \omega>1 \). The proportions of sites under positive selection estimated by model M8 is 2.0% (Table 7). These proportions are similar to estimates from model M3 presented above. Thus, there is clear statistical evidence that \textit{Xenopus} ZPC genes have been subjected to positive Darwinian selection.

For a positive control, the eight aligned mammalian ZPC cDNAs (human, macaque, marmoset, mouse, rat, pig, dog, and cat) used in the Swanson et al (2001) study were used in the same two CODEML likelihood ratio tests. Our results were identical to those reported by Swanson (Table 7).
Identification of Amino Acid Positions under Positive Selection

To identify amino acid sites under positive selection, the Bayes theorem was used to calculate the probability that a particular amino acid site can be defined into one of the \( \omega \) classes (negative, neutral or positive) as performed in Swanson’s study. Sites with high probabilities of coming from the class with \( \omega>1 \) are likely to be under positive selection. When the analysis was performed, 6 sites were detected to be subjected to positive selection out of the total 256 codon sites aligned for the *Xenopus* ZPC cDNAs. Two sites out of these six matched with mammalian ZPC sites identified as being subjected to positive selection by Swanson (*Xenopus* genus site #13 = Mammalian site #194 and *Xenopus* #159= Mammalian #341). Since the *Xenopus* ZPCs in our study do not have the 5’ end sequences, the first amino acid aligned in *Xenopus* species is within the ZP domain and is equivalent to amino acid #187 in the mouse ZPC sequence.

Moreover, three sites identified to be under Darwinian selection in the aligned *Xenopus* ZPC sequences (*Xenopus* #s 159, 160, 166 which is equivalent to mouse #s 341, 342, 348 respectively) fall within and near the region of the mammalian sperm combining site (defined as mouse site #s 328-343)(refer to Figure 24 and Table 7). As can be seen in the alignment of *Xenopus* and mammalian sequences (Fig 24), the region within and near the sperm combining site (alignment sites figures 20 and 24) has a great deal of sequence variation and insertion/deletion events. It should also be noted that the positions of serine and threonine residues are quite different in some of the sequences which is relevant because of their potential as O-linked glycosylation sites. It is likely
that the exact position and configuration of these serine/threonine residues may influence sperm-binding as well as the surrounding amino acids.

Interestingly, three of the six sites indentified to be under positive selection for *Xenopus* ZPCs fall within the ZP domain. It is unclear as to why this may be since this is thought to be a conserved domain involved in matrix assembly. However, two sites were also identified in this region for mammalian ZPCs. It may be that several of the ZP domain residues in addition to other 5’ end residues are found within or near the sperm-binding site when the protein folds.

Figure 23. A schematic representation of the ZPC molecule indicating functional regions. Black dots represent relative positions of sites under positive selection in Swanson’s study. A cluster of sites under selection fall in the region identified as the sperm-combining site (328-343 Mouse sequence). We found three sites under positive selection in *Xenopus* ZPC genes in relatively the same region (*Xenopus* 159, 160 and 166 = *Mouse* 341, 342 and 348 respectively)
Table 7. Likelihood ratio test of positive selection in *Xenopus* and Mammalian ZPC proteins. The data has n sequences and Lc codons. The proportion of sites under positive selection or under selective constraint, and parameter p and q for the beta distribution are given under M8. Parameters indicating positive selection are in bold. Significance at 5% level indicated by *; significance at 1% level indicated by **. Sites potentially under positive selection identified under M8 are listed according to the *X. laevis* first clone (XlZPC.1a) for *Xenopus* genus and mouse sequence for mammalian analysis. Same sites were found under M3. Positively selected sites with posterior probability >0.9 are underlined, .08-.09 in bold, 0.7-0.8 in italics, and 0.5-0.7 in plain text.

<table>
<thead>
<tr>
<th>Gene</th>
<th>n</th>
<th>Lc</th>
<th>dN/ds</th>
<th>2∆M3 vs. M0</th>
<th>2∆ M8 vs. M7</th>
<th>parameters estimates positively selected Under M8</th>
<th>sites (for both tests)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZPC</td>
<td>9</td>
<td>256</td>
<td>0.375</td>
<td>67.7**</td>
<td>11.7*</td>
<td>$p_1=0.02078, \omega=3.75$</td>
<td>13, 71, 76, 159, 160, 166</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$p_0=0.97922$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$B(0.53548, 1.22501)$</td>
<td></td>
</tr>
<tr>
<td>Mammalian ZPC</td>
<td>Gene</td>
<td>n</td>
<td>Lc</td>
<td>dN/ds</td>
<td>2∆M3 vs. M0</td>
<td>2∆ M8 vs. M7</td>
<td>parameters estimates positively selected Under M8</td>
</tr>
<tr>
<td>ZPC</td>
<td>8</td>
<td>365</td>
<td>0.27</td>
<td>219.6**</td>
<td>8.6*</td>
<td>$p_1=0.076, \omega=1.7$</td>
<td>25, 28, 31, 32, 33, 34, 47, 50, 84, 185, 194, 331, 333, 340</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$p_0=0.924$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$B(0.40, 1.24)$</td>
<td>341, 345, 347, 348, 372, 373</td>
</tr>
</tbody>
</table>

**Xenopus** ZPC proteins

<table>
<thead>
<tr>
<th>Gene</th>
<th>dN/ds</th>
<th>2∆M3 vs. M0</th>
<th>2∆ M8 vs. M7</th>
<th>parameters estimates</th>
<th>positively selected Under M8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZPC</td>
<td>0.375</td>
<td>67.7**</td>
<td>11.7*</td>
<td>$p_1=0.02078, \omega=3.75$</td>
<td>13, 71, 76, 159, 160, 166</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$p_0=0.97922$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$B(0.53548, 1.22501)$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mammalian ZPC</th>
<th>Gene</th>
<th>dN/ds</th>
<th>2∆M3 vs. M0</th>
<th>2∆ M8 vs. M7</th>
<th>parameters estimates</th>
<th>positively selected Under M8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZPC</td>
<td>0.27</td>
<td>219.6**</td>
<td>8.6*</td>
<td>$p_1=0.076, \omega=1.7$</td>
<td>25, 28, 31, 32, 33, 34, 47, 50, 84, 185, 194, 331, 333, 340</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$p_0=0.924$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$B(0.40, 1.24)$</td>
<td>341, 345, 347, 348, 372, 373</td>
</tr>
</tbody>
</table>

<p>| &quot;XbZPC.3&quot;  | SPRASSVFQQLGDIFHIEAVSGVPHPVPLFLHCPVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XLZFC.3&quot;  | SPLASGVRGLDFVSIEASVTRHLGPMMLFVDRCAVTL--PELNSSPQYEIIALNGCL 58 |
| &quot;XbZFC.2&quot;  | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XLZFC.2&quot;  | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XbZFC1.e&quot; | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XbZFC1.d&quot; | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XbZC.1c&quot;  | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XLZFC.1a&quot; | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;XLZFC.1b&quot; | SPRASNLFGQLDTFRIEASIAHMPPTIYDSVATVT--PDVSNPSRYSVNGCL 58 |
| &quot;C. familiaris&quot; | SEKQSFPTQLGDIAHQLAQEVTGSHMPLRLFVDCATVT--PRDNAFPHKIVDFHGCL 58 |
| &quot;F. catus&quot;  | SEKQSFPTQLGDIAHQLAQEVTGSHMPLRLFVDCATVT--PRDNAFPHKIVDFHGCL 58 |
| &quot;S. acrofa&quot; | AERMTPTQLGDRAHQLQAQVHTGSHYPLLFLVDCATVT--PDVSNSPSHTVDFHGCL 58 |
| &quot;M. radiata&quot; | AERMTPTQLGDRAHQLQAQVHTGSHYPLLFLVDCATVT--PDVSNSPSHTVDFHGCL 58 |
| &quot;Marmoseta&quot; | AERMTPTQLGDRAHQLQAQVHTGSHYPLLFLVDCATVT--PDVSNSPSHTVDFHGCL 58 |
| &quot;H. sapiens&quot; | AERMTPTQLGDRAHQLQAQVHTGSHYPLLFLVDCATVT--PDVSNSPSHTVDFHGCL 58 |
| &quot;Mouse&quot;     | TEKSSPFHVEAVHQLAQEVTGSHLQLFLVDCATVPTSPSLPPDPNSHYVDFHGCL 60 |
| &quot;Rattus rattus&quot; | TEKSSPFHVEAVHQLAQEVTGSHLQLFLVDCATVPTSPSLPPDPNSHYVDFHGCL 60 |</p>
<table>
<thead>
<tr>
<th>Species</th>
<th>Amino Acid Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>X. laevis</td>
<td>NIHRFCK-KQSIFVLAK</td>
<td>237</td>
</tr>
<tr>
<td>X. borealis</td>
<td>NIHRFCR-KQSIFVFAK</td>
<td>237</td>
</tr>
<tr>
<td>Mouse</td>
<td>KCHSSSY-LVSLPQ</td>
<td>243</td>
</tr>
<tr>
<td>Rattus rattus</td>
<td>KCHSSSY-LVSLPQ</td>
<td>243</td>
</tr>
<tr>
<td>C. familiaris</td>
<td>RHRTASHVCPASVSQ</td>
<td>246</td>
</tr>
<tr>
<td>F. catus</td>
<td>RHHTASRFMICPVSASQ</td>
<td>244</td>
</tr>
<tr>
<td>S. scrofa</td>
<td>RRRAAAH-LVCPSASEQ</td>
<td>240</td>
</tr>
<tr>
<td>M. radiata</td>
<td>RMRTAASR----PVSASE</td>
<td>242</td>
</tr>
<tr>
<td>Marmoseta</td>
<td>RCRTASL----PVSASE</td>
<td>242</td>
</tr>
</tbody>
</table>

Figure 24. Alignment of ZPC amino acids of X. laevis, X. borealis and Swanson’s Mammalian species. The ☼ symbols indicate the mammalian sites under positive selection identified by Swanson, ∞ symbols indicate the Xenopus sites identified to be under positive selection and ♦ symbols show the sites in which both the mammalian and Xenopus are under Darwinian selection. The underline region shows the sperm binding region identified in Swanson’s study.
Chapter 4

SUMMARY

Animal fertilization is one of the most fascinating processes in biology and is defined as the process of uniting two gametes, the egg and sperm. This interaction between highly specialized cells provides a unique example of many cellular processes, including specific cell adhesion, cell signaling, regulated exocytosis, cell migration and cell fusion (Miller et al, 2007). The sperm-binding glycoprotein ZPC is a key regulator of this complex interaction between sperm and eggs during the fertilization process. Although the extracellular coat can be by-passed in vitro by micro-injection spermatozoa directly into the egg cytoplasm, the ZP components normally screen spermatozoa to ensure that species-specific, motile spermatozoa fertilize the egg (Rankin and Dean et al, 1996). Since ZPC sperm-binding proteins play such a pivotal role in the fertilization process, prezygotic reproductive speciation could result from the rapid divergence of sites within ZPC genes at the sperm-binding site. The complementary receptors on sperm will necessarily need to co-evolve with ZPC mutations for successful fertilization and will also likely be subjected to positive selection.

This study focused on the evolution of the ZPC genes between two closely related frog species (Xenopus laevis and X. borealis) to determine which amino acid sites are rapidly evolving thereby leading to predictions of the amino acid sites likely to be involved in sperm-binding. To address this issue, this research examined the cDNAs coding for the ZPC glycoprotein in individual frogs from two Xenopus species to identify how many different alleles and/or paralogs (ZPC gene duplications) are expressed within
each cDNA library and then examine the molecular evolution of the ZPC sequences to
detect whether sites within the gene have been subjected to positive selection as reported
for mammalian ZPC genes.

Our data showed that there are multiple ZPC genes within each of the *X. laevis*
and *X. borealis* ovary cDNA libraries (4 and 5 cDNAs, respectively). We conclude that
a paralogous ZPC gene exists in *X. borealis* since a tetraploid organism can only have 4
allelic forms at a particular locus. It is likely that this is the case for *X. laevis* also, but did
not find more than 4 cDNAs to support this hypothesis. ZPC gene duplication has also
been reported in fish species such as the medaka and zebrafish. A study done by Conner
et al., (2003) suggested that the availability of multiple copies of the ZPC genes in
zebrafish may allow for the generation of additional variation, which may increase the
structural complexity and diversity of the egg coat. Surprisingly, there is no evidence of
ZPC duplication in mammalian species, Mammalian genomes have been searched but
appear to have no evidence for paralogous gene locations for ZPC. It is likely that the
common ancestor leading to mammals functionally lost (deletions or pseudogenes) all
paralogous ZPC genes and maintained only one ZPC gene location that serves as the
modern sperm-binding protein. Or alternatively, ZPC gene duplications in fish and
amphibians happened solely within their lineages after the split from the common
ancestor leading to mammals.

The significance of having paralogous ZPC genes expressed in amphibians and fish
whereas mammals have only one remains obscure. This phenomenon could be related to
the fact that amphibians and fish exhibit external fertilization whereas mammalian
fertilization is internal. In external fertilization, the male deposits lots of sperm (on the order of thousands) at the moment an egg is released from the female into the medium (e.g. pond water) which creates sperm competition for entry into the egg. Most all animal species can only handle one sperm entering the egg (monospermic fertilization), and multiple sperm entry will terminate development. Prevention of polyspermic fertilization can happen at many different levels. It is possible that having multiple sperm binding ZPC genes expressed in the vitelline envelope of frog eggs serves to slow down sperm entry and selects for only the sperm with the highest affinity receptors to its ZPC glycoprotein. This would in effect reduce the incidence of polyspermy.

In addition, the existence of multiple ZPC genes would allow for many mutant forms of the sperm-binding gene to co-exist at the same time and could drive speciation events by making it necessary for sperm receptor to co-evolve to maintain binding complementarities. This sexual conflict at the gamete level could result in interspecies incompatibilities since only the co-evolving sperm will be compatible with the rapidly evolving ZPC genes. The mechanism for this is thought to be positive selection acting at the sperm binding recognition site. Evidence for positive selection at or near the putative sperm combining site exists in the mammalian ZPC gene, and we report evidence for positive selection in this same region for two *Xenopus* frog species. Although the sites under positive selection were not identical in mammals and frogs, they did fall in similar regions. It will be of interest to examine the 5’ end sequences of the *Xenopus* ZPC cDNAs to examine whether there is a similar pattern of positive selected sites in this region also. Although it is difficult to predict the tertiary or 3-D structure of the ZPC
glycoprotein within the egg coat, it is likely different portions of the ZPC polypeptide can contribute to the orchestration of the sperm binding site.

It is also possible that the amount of expression for each ZPC gene is not equal and that some genes contribute more to the extracellular coat. The existence of paralogous genes would allow for different regulatory elements to determine their relative expression. Potentially, expression differences could influence which sperm receptor within the sperm population is selected for at the time of fertilization. Different populations or species could change their relative expression of the ZPC genes within the egg coat thereby creating more gamete conflict for driving speciation or maintaining interspecies incompatibilities.

In summary, we have shown that *Xenopus* ZPC genes have duplicated in their genome at least once which is an ancestral fish-like pattern. It is uncertain as to whether all of these ZPC gene products can serve as sperm-binding proteins, but what is clear is that amino acids near the sperm combining site were detected to be under positive selection in the *Xenopus* ZPC genes examined similar to the mammalian ZPC gene. It will be of interest to examine the 5’ ends of the *Xenopus* ZPC genes to determine if there are additional sites subject to positive selection which may be involved in the sperm binding site. In addition, real time PCR experiments could be used to determine the relative expression of the *Xenopus* ZPC genes to assess their likelihood as being the primary sperm-binding proteins.
WORKS CITED


6) Bork P and Sander C. “A Large Domain Common to Sperm Receptors (Zp2 and Zp3) and TGF-β Type III Receptor.” *Mol Bio Evol.* 300(1992); 237-240.


27) Miller D, and Burkin H. “Molecular Basis of Mammalian Gamete Binding.” Review


