The Sequence and Function Relationship of Elastin: How Repetitive Sequences Can Influence the Physical Properties of Elastin

by

David Yi Ming He

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy Molecular Genetics University of Toronto

© Copyright by David Yi Ming He 2011
The Sequence and Function Relationship of Elastin: How Repetitive Sequences Can Influence the Physical Properties of Elastin

David Yi Ming He
Doctor of Philosophy
Molecular Genetics
University of Toronto
2011

Abstract

Elastin is an essential extracellular protein that is a key component of elastic fibres, providing elasticity to cardiac, dermal, and arterial tissues. During the development of the human cardiovascular system, elastin self-assembles before being integrated into fibres, undergoing no significant turnover during the human lifetime. Abnormalities in elastin can adversely affect its self-assembly, and may lead to malformed elastic fibres. Due to the longevity required of these fibres, even minor abnormalities may have a large cumulative effect over the course of a lifetime, leading to late-onset vascular diseases. This thesis project has identified important, over-represented repetitive elements in elastin which are believed to be important for the self-assembly and elastomeric properties of elastin. Initial studies of single nucleotide polymorphisms (SNPs) from the HapMap project and dbSNP resulted in a set of genetic variation sites in the elastin gene. Based on these studies, glycine to serine and lysine to arginine substitutions were introduced in elastin-like polypeptides. The self-assembly properties of the resulting elastin-like polypeptides were observed under microscope and measured using absorbance at 440nm.
Assembled polypeptides were also cross-linked to form thin membranes whose mechanical and physical properties were measured and compared. These mutations resulted in markedly different behavior than wild-type elastin-like proteins, suggesting that mutations in the repetitive elements of the elastin sequence can lead to adverse changes in the physical and functional properties of the resulting protein. Using next-generation sequencing, patients with thoracic aortic aneurysms are being genotyped to discover polymorphisms which may adversely affect the self-assembly properties of elastin, providing a link between genetic variation in elastin and cardiovascular disease.
I would like to thank the Parkinson and Keeley laboratories for their tremendous help and support in both the computational and experimental components of this project. Special thanks to Eva and Megan for teaching me the laboratory skills I needed to get most of the data found in the third chapter of this thesis. I am also grateful for the invaluable help and advice provided by my committee members, Drs. Keeley and Zhang. For his super-human patience, and never-ending support during my entire PhD at UofT, I especially want to thank Dr. John Parkinson, without whom, I would have given up years ago.

I would also like to thank my family, especially my parents who have always been my greatest fans. And lastly, I need to thank my new wife, Tiffany, who has supported me throughout my PhD study and even allowed me to mention elastin on our honeymoon.
# Table of Contents

Acknowledgments ........................................................................................................... iv

Table of Contents ............................................................................................................. v

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

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

Chapter 1 Introduction ........................................................................................................ 1

1 Elastin ............................................................................................................................ 1

1.1 Elastin and Elastic Fibres ......................................................................................... 3
1.2 The Importance of Elastin in Health and Disease ..................................................... 6
1.3 The Elastin Sequence ............................................................................................... 9
1.4 Characterization of Repetitive Elements in Elastin .................................................. 12
1.5 Elastin Structure ....................................................................................................... 14
1.6 Models of Elasticity ................................................................................................. 16
1.7 Coacervation ............................................................................................................ 20
1.8 Lysyl Oxidase and Elastin Cross-link Formation ..................................................... 22
1.9 Evolution of Elastin ................................................................................................. 25
1.10 Elastin-like Polypeptides ......................................................................................... 29
1.11 Comparison of Elastin to Other Elastic Proteins .................................................... 30

1.11.1 Abductin ............................................................................................................ 31
1.11.2 Resilin ............................................................................................................... 32
1.11.3 Wheat gluten protein ......................................................................................... 32
1.11.4 Spider Silk Proteins .......................................................................................... 33

1.12 Project Goals .......................................................................................................... 33

Chapter 2 Survey of elastin sequence .............................................................................. 36

2 Survey of elastin sequence .......................................................................................... 36
3.2.5 Coacervation behaviour .................................................................98
3.2.6 Microscopic visualization of coacervation .........................................98
3.2.7 Elastic membrane construction .......................................................99
3.2.8 Mechanical testing of membranes ...................................................100
3.2.9 Scanning electron microscopy .........................................................101

3.3 Results ...............................................................................................101
3.3.1 SNP Identification ..............................................................................101
3.3.2 SNP selection ....................................................................................104
3.3.3 Circular Dichroism Spectra...............................................................107
3.3.4 Effect of mutations on the coacervation properties of EP2 constructs ....110
3.3.5 Elastic membrane properties ............................................................117
3.3.6 Mechanical testing of membranes .....................................................119
3.3.7 Scanning electron microscopy .........................................................120

3.4 Discussion ............................................................................................122
3.4.1 Effect of mutations on EP2 structure ................................................123
3.4.2 Effects of mutation on coacervation .................................................126
3.4.3 Effect of mutations on mechanical properties ....................................130

Chapter 4 Summary and Future Directions .............................................133

4 Summary of findings and Future Directions ........................................133
4.1 Introduction .........................................................................................133
4.1.1 Summary of Findings .......................................................................134
4.2 The effects of elastin sequence variants in patients linked to late-onset cardiovascular disease on the structure and function of elastin .........................................................136
4.2.1 Identification of novel polymorphisms which are over-represented in full length human elastin gene in patients linked with cardiac disease .................137
4.2.2 Experimental and computational validation of the effects that such polymorphisms have on the structure and function of recombinant elastin or elastin-like polypeptides .......................................................... 139

References .................................................................................................................. 141
List of Tables

Table 3-1 SNPs causing non-synonymous mutations in the translated elastin protein. SNPs 1-10 were identified using dbSNP. SNPs 11-15 were identified using ESTs from dbEST. SNPs with available minor allele frequencies were sourced from the International HapMap Project........ 103

Table 3-2 The coacervation temperature, molecular weight, and average Kyte-Doolittle hydropathy values of EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, and EP2-3G3S polypeptides. Native EP2 data was collected by Judith Cirulis. ................................................................. 111

Table 3-3 Results of mechanical testing on membranes formed by cross-linking EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, EP2-3G3S polypeptides with genipin. Mechanical testing performed by Eva Sitarz. ............................................................................................................. 119
List of Figures

Figure 1-1 A schematic illustrating the model of elastic fiber assembly................................. 6

Figure 1-2 The exon arrangements of the ELN gene from different species, demonstrating the conservation of a unique style of sequence, characterized by exons alternately encoding for hydrophobic and cross-linking regions................................................................. 11

Figure 1-3 Fibrillar aggregate of tropoelastin solution at 40°C for 15 minutes. Bar 0.1µm ......... 14

Figure 1-4 Proposed models of elastin structure........................................................................ 18

Figure 1-5 Formation of elastin cross-links by lysyl oxidase (LOX) ........................................ 24

Figure 1-6 Multiple sequence alignment of elastin protein sequence from human, baboon, cat, dog, cow, pig, mouse, and rat showing gaps and generally poor alignment .................... 27

Figure 2-1 Schematic overview of the creation of a subsequence network from a protein sequence....................................................................................................................... 44

Figure 2-2 Graphs showing the relationships between G, P and V compositions for a range of proteins............................................................................................................................ 48

Figure 2-3 Incidence of 3mer, 4mer and 5mer repeats for a range of proteins....................... 50

Figure 2-4 Network maps detailing relationships between fuzzy 3mer repeats in the human tropoelastin sequence ........................................................................................................ 52

Figure 2-5 Network maps detailing relationships between fuzzy 3mer repeats in the chicken tropoelastin sequence ........................................................................................................ 53
Figure 2-6 Network maps detailing relationships between fuzzy 3mer repeats in the frog eln1 sequence. ................................................................. 54

Figure 2-7 Network maps detailing relationships between fuzzy 3mer repeats in the zebrafish eln2 sequence. ................................................................. 55

Figure 2-8 Network maps detailing relationships between fuzzy 3mer repeats from 100 random sequences with the same amino acid composition of human elastin. ................................................................. 58

Figure 2-9. Network map showing the global diversification of abundant motifs in elastin across vertebrates. ................................................................. 60

Figure 2-10 Mapping of identified motifs onto elastin sequences. ................................................................. 62

Figure 2-11. WebLogo characterization of identified motifs. ................................................................. 64

Figure 2-12. User interface of the SubSeqer web application. ................................................................. 66

Figure 2-13. Clustergram detailing sequence similarity relationships between 492 exons from 12 elastin sequences. ................................................................. 69

Figure 2-14 Clustergram detailing sequence similarity relationships between exons from mammals and chicken species. ................................................................. 71

Figure 2-15 Clustergram detailing sequence similarity relationships between exons from zebrafish, xenopus, and fugu species. ................................................................. 72

Figure 2-16 Clustergram analysis of pairs of exons and their adjoining intron. ................................................................. 75

Figure 2-17 Subseqer result for resilin. ................................................................. 82
Figure 3-1 Schematic of construct preparation.................................................................................. 94

Figure 3-2 Q-TOF mass spectrometry of EP2-based polypeptides showing the appropriate
introduction of mutations.................................................................................................................. 97

Figure 3-3 Circular dichroism spectra of EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R and EP2-2K2R
in water with polypeptide concentrations of 10uM at room temperature................................. 109

Figure 3-4 Coacervation behavior of EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R
......................................................................................................................................................... 113

Figure 3-5 Bright field microscopy images of the coacervation and maturation process of EP2,

Figure 3-6 Bright field microscopy images of the coacervation and maturation process of EP2,
EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R at 60 minutes...................................................... 116

Figure 3-7 Physical appearance of membranes made from EP2, EP2-1K1R, EP2-2K2R, EP2-
1G1S, and EP2-3G3S polypeptides ................................................................................................. 118

Figure 3-8 Scanning electron microscopy images of the surface of membranes made from EP2,

Figure 3-9 Circular dichroism spectra EP2 (solid line), EP2 with P to G mutation in one exon 24
(dotted line), and EP2 with P to G mutation in both exon 24s (dashed line) ............................... 125
Chapter 1 Introduction

1 Elastin

Elastin is a polymeric extracellular protein that is responsible for imparting elasticity to various tissues found in vertebrate organisms. Together with microfibrils, elastin forms elastic fibres, important components of the extracellular matrix (ECM). Elastin can be found in all higher order vertebrates including mammals, birds, sharks, and bony fish, but is not thought to occur in lower order vertebrate organisms such as lamprey (Debelle and Tamburro 1999). The elasticity provided by elastin can be used statically to resist long-term forces as in the dermis, or dynamically by providing a highly efficient energy storage system as in arterial tissues. The earliest appearance of elastin is estimated to date to 400 million years ago (Sage 1982).

In humans elastin is a key component of arterial tissues. For example, the thoracic aorta is composed of 50% elastin by dry weight (Uitto 1979). Elastogenesis, the formation of elastic fibres containing elastin, typically occurs only during the late fetal and early neonatal development phases in the human heart (Visconti, Barth et al. 2003). One of the most important initial steps of this process involves the self-assembly of tropoelastin, the monomeric subunit of elastin, through a process termed coacervation. This process is believed to be responsible for concentrating and aligning tropoelastin molecules for subsequent integration into the elastic fibre (Urry 1978). Since little elastin is produced after elastogenesis, it is clear that aberrations in the
formation of elastic fibres may lead to cumulative effects that weaken important vascular tissues over the course of millions of stretch and recoil cycles, potentially leading to late onset cardiovascular disease.

The polymeric form of elastin has proven difficult to study because it exists in a cross-linked and largely insoluble form. Much of the existing body of work on elastin has been done on the soluble tropoelastin monomer, which was first isolated from pigs fed a diet deficient in copper, a necessary cofactor of the cross-linking enzyme lysyl oxidase (Cox, Starcher et al. 1974). Modern elastin research is conducted using recombinant tropoelastin proteins and protein fragments (Daamen, Veerkamp et al. 2007), which are more readily available than animal-derived polypeptides. However, despite these advances, the structure-function relationship of the elastin protein has remained elusive. Although it has been suggested that the highly repetitive nature of elastin plays a role in its ability to coacervate and impart elasticity, no clear link between the two has been established.

This thesis project adopts a two-pronged approach to elucidating the sequence-function relationship of elastin by first using a novel computational method to systematically identify significant, and over-represented repetitive elements in the tropoelastin amino acid sequence from a wide taxonomic range of species. Having identified these repeats, mutations are introduced within these regions, and the resulting physical and functional changes in the protein are carefully studied using recombinant elastin-like polypeptides made up of a segment of full-length tropoelastin.
1.1 Elastin and Elastic Fibres

A major component in the extracellular matrix of elastic tissues found in the arteries, lungs, skin and ligaments, is the elastic fibre, which confers the necessary elasticity and resilience to stretch and recoil (Sandberg, Soskel et al. 1981). Elastic fibres are composed of a 10-12 nm outer sheath of microfibrils (Ramirez 2000), consisting mainly of fibrilins, surrounding a core of cross-linked elastin proteins, which make up nearly 90% of the fibre. While the composition of elastic fibres is similar in different tissues, their organization varies to fulfill the particular biological requirements of the tissue. In the arteries, which require the maintenance of uniform elastic pressure, elastic fibres form concentric rings of elastic lamellae, which alternate rings of smooth muscle around the arterial lumen (Li, Brooke et al. 1998). In lung, elastic fibres provide architectural support for areas of constant elastic stress such as the opening of the alveoli and alveolar junctions, by forming a latticework of fibres (Starcher 2000). In elastic cartilage, which require a balance between structure and flexibility, elastic fibres are organized into large three-dimensional honeycomb structures (Dietz, Ramirez et al. 1994).

Elastic fibres are formed by a process called elastogenesis (Figure 1-1), which begins with the mRNA expression of tropoelastin, the monomeric precursor to elastin. This expression is most prominent during early development, and it has been shown that there exists a strong correlation between tropoelastin mRNA levels and protein synthesis (Barrineau, Rich et al. 1981). Using the human tropoelastin promoter coupled with the chloramphenicol acetyltransferase reporter gene in mice has shown an age-dependant relationship in transcription levels (Hsu-Wong, Katchman
et al. 1994). In the chick aorta, age related decrease in tropoelastin synthesis has also been shown to be related to a destabilization of mRNA (Johnson, Robson et al. 1995).

The assembly of elastic fibres is a complex process, involving many different proteins. The process begins with the translation and secretion of tropoelastin polypeptides by smooth muscle cells and various fibroblasts. Like most extracellular proteins, translation takes place in the rough endoplasmic reticulum (RER); the protein product is released to the lumen of the RER after removal of the signal peptide; from there, tropoelastin proteins are transported to the plasma membrane via secretory vesicles (Saunders and Grant 1984). With the exception of proline hydroxylation (Uitto, Hoffmann et al. 1976), tropoelastin does not undergo significant post-translational modification.

Once translated, tropoelastin is accompanied by a 67kDa co-protein called the elastin binding protein (EBP) to prevent premature self-assembly (Hinek, Braun et al. 2004). Prior to integration into elastic fibres, but soon after it is excreted into the ECM, tropoelastin polypeptides rapidly undergo a self-assembly using the process of coacervation (Sato, Wachi et al. 2007). This process aligns and concentrates tropoelastin into small spheres of several microns in diameter. These coacervate packages remain tethered to the cell surface until they are released and become integrated onto microfibril scaffolds, which form a sort of skeletal structure for elastic fibres. The interaction between elastin and the cell-surface is not well understood, but it has been suggested that proteins such as integrins and glycosaminoglycans are involved (Broekelmann, Kozel et al. 2005).
The initial deposition of tropoelastin globules onto microfibrils scaffold is followed by a microfibril directed aggregation (Kozel, Rongish et al. 2006) requiring a large number of proteins. The final stage of elastic fibre formation begins when the enzyme lysyl oxidase (LOX) deaminates the ε-amino group of the lysines in the cross-linking regions of elastin. This process produces allysines which can condense spontaneously with other allysines or lysines to form lysinonorleucine, merodesmosine or desmosine cross-links (Reiser, McCormick et al. 1992). Elastins typically contain 6-12 cross-links per molecule, and cross-links form both inter- and intra-molecularly (Dyksterhuis, Baldock et al. 2007). Some specificity of cross-link formation has also been observed (Wise, Mithieux et al. 2005), suggesting that coacervation and the initial self-assembly of tropoelastin molecules plays a key role in aligning the molecules for specific cross-link formations to occur in assembling a proper elastic fibre. A logical extension of this observation is that changes in the sequence of tropoelastin which can negatively affect the assembly properties of elastin can lead to devastating changes in the final elastic fibre.
Figure 1-1 A schematic illustrating the model of elastic fiber assembly. Tropoelastin is secreted from the cell and coacervates into assemblies attached to the cell surface via integrins and glycosaminoglycans. These coacervates are released onto microfibrils, where lysyl oxidase (LOX) forms inter- and intra-chain cross-links using lysine residues on the tropoelastin polypeptide, forming stable and insoluble elastic fibers.

1.2 The Importance of Elastin in Health and Disease

The long half-life of elastin and elastic fibres as observed in human and mouse tissues is made possible in large part by the resistance of elastin to proteolytic degradation. In fact, elastases, proteins with the ability to degrade elastin to fulfill important physiological requirements such as growth, wound healing, and tissue remodeling, are the only proteins known to be capable of degrading insoluble elastin.

There are numerous acquired and inherited diseases which affect the assembly, structure or distribution of elastic fibres in humans. However, as there are many proteins involved in elastic
fibres, many of these diseases, such as Marfan Syndrome which is caused by mutations in the fibrillin genes, are the result of defects in other ECM genes. Only a small number of well-studied conditions have a direct genetic connection to the elastin gene. The most well-known and studied of such diseases are supravalvular aortic stenosis (SVAS) and cutis laxa (CL).

SVAS is a progressive vascular disorder that can occur on its own as an autosomal dominant disease (Eisenberg, Young et al. 1964), or together with Williams Beuren Syndrome (Bhattacharjee 2005). The disease is characterized by a spectrum of symptoms, ranging from relatively mild cases involving the development of a fibrous ring above the aortic valves, to severe cases of diffuse narrowing of the ascending aorta. In addition to the aorta, SVAS can also affect other major arteries including the pulmonary, carotid, cerebral, and renal (Milewicz, Urban et al. 2000). In general, the underlying cause of these symptoms appears to be a thickening of the arterial wall brought about by a decrease in elastin deposition coupled with a proliferation of vascular muscle cells.

Multiple studies have linked SVAS to mutations in the elastin gene, including the deletion of a large 30kb region of the gene (Olson, Michels et al. 1995). Many point mutations have also been discovered in the elastin genes of patients with SVAS. Most of these have been categorized as premature termination mutations, and these mutant alleles have been shown to result in mRNA products which are targeted for degradation by a nonsense-mediated decay mechanism (Urban, Michels et al. 2000). The net result of these mutations is the functional haploinsufficiency of the elastin gene. SVAS can also exist as a condition of a much wider and more complicated genetic
disease called Williams Beuren Syndrome, which is characterized by cardiovascular, neurobehavioral, facial, and growth abnormalities (Bhattacharjee 2005). In this context, SVAS is likely caused by a 1.5-2MB deletion at chromosomal band 7q11.2 which results in the complete deletion of one copy of the elastin gene, along with approximately 14 additional genes (Francke 1999).

Autosomal dominant cutis laxa (ADCL) is another disease that has known links to mutations in the elastin gene. ADCL is a rare genetic disease, and patients often have loose skin and may also be affected by pulmonary emphysema, and aortic and arterial dilation (Weir, Joffe et al. 1977; Tassabehji, Metcalfe et al. 1998). Known mutations in the elastin gene are all single nucleotide deletions in exons 30 and 32 near the 3’ end of the sequence (Tassabehji, Metcalfe et al. 1998; Zhang, He et al. 1999). These mutations support the hypothesis that the C-terminal region of the elastin protein is vital to its proper deposition in elastic fibre.

Polymorphisms have been identified in the human elastin gene since the early 1990s but most of these have not been linked to known phenotypes. However, a study on a polymorphism causing a serine/glycine mutation at amino acid position 422 did observe a link between this mutation and carotid artery distensibility and elasticity (Hanon, Luong et al. 2001). Recently, advanced sequencing technology has been applied to detect single nucleotide polymorphisms (SNPs) in human elastin (Cho, Ciulla et al. 2009) but so far, few links between mutations and disease phenotypes have been discovered.
External factors may also impact the durability of elastic fibres, and both excess sun exposure (Watson, Griffiths et al. 1999) and smoking (Shifren and Mechem 2006) have been shown to cause damage to elastic fibres. Regardless of the source of damage, it is clear that abnormalities in the functional and physical properties of elastin and the elastic fibres that can lead to devastating diseases, highlighting the importance of elastic fibres in maintaining healthy elastic tissues.

1.3 The Elastin Sequence

The elastin gene, coding for tropoelastin, is located on Chromosome 7q11.2 in humans within a 30 gene region known as the Williams region (Fazio, Mattei et al. 1991) due to its association with Williams Beuren Syndrome. Synteny in this region between mammalian species is generally well conserved. In every instance where the elastin gene has been identified, the gene for the protein LIM Kinase is immediately downstream of the elastin gene, and its presence has been used as an important marker when identifying the elastin gene in newly sequenced organisms (Chung, Miao et al. 2006). In addition, the style of sequence found in the elastin gene is highly conserved among elastins of different species. This distinctive style (Figure 1-2) consists of exonic sequences that code for long, hydrophobic regions of the protein interspersed by exonic sequences coding for shorter cross-linking regions of the protein (Chung, Miao et al. 2006). However, unlike other genes in this region, the nucleotide and amino acid sequence of the elastin gene is not well conserved between species (DeSilva, Elnitski et al. 2002). Together,
these exons encode a 60-72kDa product in avian and mammalian species, depending on alternative splicing. The most common isoform of elastin found in human is encoded by 34 exons, resulting in a 60 kDa product after the removal of a signal peptide.

The hydrophobic regions of the protein are composed mostly of glycines, prolines, valines, leucines, and isoleucines, which form short, 4-8 peptide long repetitive elements. Interspersed in these regions are cross-linking regions, which, in avian and mammalian tropoelastin, are predominantly composed of alanines and lysines in KAAK or KAAAK motifs. The lysine residues are oxidized by lysyl oxidase prior to the formation of intra- and intermolecular cross-links (Reiser, McCormick et al. 1992). Taken as a whole, the most abundant amino acids in elastin are valine (24%) and glycine (31%) followed by alanines (13%) and prolines (11%) (Sage 1982). Equally noteworthy is the conspicuous absence or low abundance of cysteines, histidines, and methionines. The absence of the later amino acid, with the exception of the starting codon, has been particularly useful for isolating elastin from tissues, as it allows for the use of cyanogen bromide, which hydrolyzes the C-terminus of methionine residues, leaving elastin intact while cleaving most other proteins (Indik, Abrams et al. 1990).
Figure 1-2 The exon arrangements of the ELN gene from different species, demonstrating the conservation of a unique style of sequence, characterized by exons alternately encoding for hydrophobic and cross-linking regions. Figure from (He, Chung et al. 2007), used with permission from publisher.
In addition to the alternating hydrophobic and cross-linking exons, the first exon in the elastin gene codes for a signal peptide, important for the passage of tropoelastin into the RER (Adams, Leblanc et al. 1991). The final two exons code for a well-conserved C-terminus region, containing the only two cysteine residues found in the protein, and a distinctive RKRK sequence. The absence of this C-terminus region, as well as disruptions between the disulfide bond formed between the two cysteine residues, has been shown to prevent self-assembly and deposition of elastin in elastic fibres (Kozel, Wachi et al. 2003), providing an explanation for the high degree of conservation of the C-terminus region among all elastins discovered to date.

Alternative splicing of the tropoelastin gene has been observed in humans (Indik, Yeh et al. 1987; Boyd, Christiano et al. 1991), and at least seven isoforms have been identified in humans. However, the roles of these isoforms and their tissue specificity have not been established, although it is clear that at least in other species, alternative splicing is developmentally regulated and related to the age of the organism (Barrineau, Rich et al. 1981; Yeh, Anderson et al. 1989; Heim, Pierce et al. 1991).

1.4 Characterization of Repetitive Elements in Elastin

It has been well-established since the earliest successful isolation of elastin from mammalian tissues that elastin is a low-complexity protein, rich in a small number of hydrophobic amino acids (Sage and Gray 1979). This distinct distribution of amino acids is believed to play a crucial
role in the structure of the elastin protein, providing it with the means of elasticity and self-assembly. Molecular dynamics simulations, for example, have demonstrated that there exists a minimum threshold of combined proline and glycine percentage required for providing elastin with a critical level of disorder necessary for imparting elastomeric properties (Rauscher, Baud et al. 2006). Other structural studies of elastin have identified characteristic repetitive elements consisting of three to six residue subsequences which are believed to play an important role in determining the secondary structure of the protein (Urry, Chang et al. 1989; Tatham and Shewry 2002; Yao and Hong 2004; Pepe, Guerra et al. 2005). However, no systematic identification of the most biologically significant repetitive elements in elastin had been published prior to this thesis project.

A detailed survey of existing pattern recognition tools was conducted in the Parkinson Lab by Trevis Alleyne in 2004. This survey used three state-of-the art algorithms Teiresias, MEME, and Consensus on elastin proteins from several species to identify notable repetitive elements. In this survey, patterns detected by the three algorithms were judged on the criteria of 1) being diagnostic (the ability to identify whether a given sequence belongs to the elastin family) and 2) being a motif (a description of some sequence which can be found in all members of the elastin family). However, due to the variable nature of elastin sequences, none of these methods produced satisfactory results, highlighting the need for a novel approach specifically tailored to elastin and elastin-like proteins with low complexity, highly repetitive sequences.
1.5 Elastin Structure

The earliest studies of elastin structure were on a macroscopic level, where elastin appeared to be an amorphous mass, without clear structure or organization. Later electron microscopy studies found that elastin was not amorphous, but had a ropelike structure made up of ~5nm thick filaments organized in a twisted form (Gotte, Giro et al. 1974). Freeze fracture electron microscopy on the surfaces of stretched and relaxed bovine elastin showed that elastin is organized in globules of approximately 5nm across in long filaments which would orient in the direction of stretching (Ronchetti, Fornieri et al. 1979). These observations in naturally occurring elastin were consistent with findings of 5nm diameter filaments seen using electron microscopy in samples of tropoelastin aggregates held at 40°C for 5-15 minutes (Figure 1-3) (Bressan, Pasquali-Ronchetti et al. 1986).

![Fibrillar aggregate of tropoelastin solution at 40°C for 15 minutes. Bar 0.1µm.](image)

*Figure 1-3* Fibrillar aggregate of tropoelastin solution at 40°C for 15 minutes. Bar 0.1µm. Figure from (Bressan, Pasquali-Ronchetti et al. 1986), used with permission from publisher.
At the molecular scale, X-ray crystallography and solution nuclear magnetic resonance (NMR) techniques have not been successfully applied to the study of elastin structure. One of the main reasons for this is elastin's insolubility. The availability of soluble tropoelastin monomers has not alleviated this difficulty because these monomers have a tendency to self-aggregate (Keeley, Bellingham et al. 2002). Nevertheless, some success at determining the secondary structure of elastin has been accomplished using circular dichroism (CD) analyses, and more recently, solid state NMR and molecular dynamics simulations.

Early CD studies of recombinant human tropoelastin showed that it consists primarily α-helix (3%), β-sheet (41%), and β-turn (21%) structures (Vrhovski, Jensen et al. 1997). However, further CD studies on hydrophobic exons 3, 7, and 30 in human tropoelastin have found that these regions may partly assume a polyproline II (PPII) structure (Bochicchio, Ait-Ali et al. 2004), a structure with few intramolecular hydrogen bonds, suggesting that the hydrophobic regions are likely dynamic structures capable of adopting multiple conformations. Solid state NMR methods applied to elastin support this observation, describing its backbone as largely disordered due to the presence of high proline content (Pometun, Chekmenev et al. 2004). In addition, studies using computational modeling of elastin-based sequences have also confirmed this disordered arrangement (Rauscher, Baud et al. 2006), suggesting that the earlier descriptions of extended secondary structures in tropoelastin may be questionable. However, despite these advances, there is still no conclusive knowledge of the structure of elastin, except that it contains rigid cross-linking regions interspersed with disordered hydrophobic regions.
1.6 Models of Elasticity

Several models have been proposed to explain the mechanism of elasticity in elastin, but fundamentally they all recognize that the process is largely driven by entropy in the hydrophobic regions of the protein, whereby stretching causes a decrease in entropy, driving a return to the relaxed state and maximum entropy (Weis-Fogh and Anderson 1970; Dorrington and McCrum 1977; Gosline 1978).

The earliest studies of elasticity in elastin considered the protein to behave as a classic rubber. The model proposed by these studies viewed elastin as a network of random chains, held together by random cross-links (Hoeve and Flory 1974) (Figure 1-4a). In this model, stretching causes an increase in the order of the chains, driving the system back to a relaxed state. However, circular dichroism (CD) and Fourier transform infrared spectroscopy (FTIR) studies have shown the presence of β-turns and polyproline II (PPII) conformations in elastin (Muiznieks, Jensen et al. 2003; Tamburro, Bochicchio et al. 2003; Bochicchio, Floquet et al. 2004), suggesting that the random network model of elastin is too simplistic in its view of elastin structure.

The apparent inadequacy of the random-chain model led to the introduction of a number of models, which suggested elasticity in elastin was the result of ordered structures formed by elastin monomers upon aggregation. The two main models in this group were the liquid-drop (Figure 1-4b) and oiled-coil (Figure 1-4b) models, both of which proposed that the exposure of
hydrophobic amino acids to water as elastin is stretched decreases the entropy of the surrounding water molecules, leading to an increased state of entropy, with a return to the relaxed state providing the restoring force necessary for elastin recoil. In the liquid-drop model, the hydrophobic domains of elastin form globules connected by cross-links. These globules become exposed to water upon elastin stretching (Weis-Fogh and Anderson 1970). In contrast, the oiled-coil model suggested that the hydrophobic domains adopt a coil formation composed of short repeat sequences, where the glycine residues are on the outside of the coil and more hydrophobic residues such as valine and proline are on the inside away from water. In this model, it is the Val and Pro residues which are exposed to water when elastin is stretched (Gray, Sandberg et al. 1973). However, CD and solid state NMR have shown that elastin molecules tend to adopt a more dynamic structure (Bochicchio, Ait-Ali et al. 2004; Pometun, Chekmenev et al. 2004; Rauscher, Baud et al. 2006) that is not congruent with the fixed conformation of either of these models, and today neither of these models is favoured.

An alternative, fibrillar model of elastin elasticity was proposed by Urry (Urry 1974) (Figure 1-4d). In this model, short polypeptide repeats in the hydrophobic regions of the tropoelastin sequence, notably APGVGV, VPGVG, and VPGG adopt type-II β-turns (Urry, Long et al. 1974; Urry and Long 1976) forming loose helices known as β-spirals. These β-spirals form filaments through hydrophobic interactions leading to a fibrillar network, where individual tropoelastin chains are held together by both cross-links and hydrophobic interactions. The regions between the β-spirals are thought to undergo low-amplitude, high-frequency rocking motions known as librations, which abate as the spiral is stretched, leading to a decrease in entropy, driving the system back to its relaxed state. This fibrillar β-spiral model was highly
regarded for many years. However, experimental studies of elastin using solid-state NMR (Pometun, Chekmenev et al. 2004) and $^{13}$C NMR (Elvin, Carr et al. 2005) have shown that elastin exhibits a high degree of dynamic disorder, and has significant structural mobility in its hydrated state, observations which are inconsistent with this fibrillar model of elastin elasticity.

![Proposed models of elastin structure](image)

**Figure 1-4 Proposed models of elastin structure.** (a) Random coil model (b) Liquid drop model (c) Oiled-coil model (d) Fibrillar model. Figure from (Vrhovski and Weiss 1998), used with permission from publisher.
Recent studies using molecular dynamics (MD) simulations of hydrophobic, elastin-like polypeptides have shown that these polypeptides form water-swollen aggregates, characterized by flexible polypeptide backbones which allow for the rapid inter-conversion between a large number of conformations (Rauscher, Baud et al. 2006; Glaves, Baer et al. 2008). While there are no regular secondary structures within these conformations, the polypeptides cannot be characterized as random coils because ordered structures such as polyproline II structures and hydrogen-bonded turns can be observed (Muiznieks, Jensen et al. 2003; Tamburro, Bochicchio et al. 2003).

Together with solid-state NMR and CD spectra studies, the emerging consensus is that elastin can be classified as an intrinsically disordered protein (IDP) (Dyksterhuis, Carter et al. 2009; Rauscher and Pomes 2010), which forms loose, water-swollen aggregates. Polypeptide chain entropy opposes both the folding and extension of the polypeptides in the aggregate because these events decrease the number of possible conformations, driving these aggregates to return to the relaxed state upon stretching.
1.7 Coacervation

Coacervation is a phase separation, where soluble molecules coalesce and come out of solution with an increase in temperature or ionic strength. It is believed that coacervation properties of tropoelastin are largely imparted by the hydrophobic regions of the protein (Bellingham, Woodhouse et al. 2001). The process of coacervation is integral to the proper formation of elastic fibres and, as previously mentioned, is also believed to play a large role in the elastic mechanism of elastin. For this reason, many studies have been conducted using both full-length tropoelastin as well as short, elastin-like polypeptides composed of a set of exons from the tropoelastin gene (Vrhovski, Jensen et al. 1997; Bellingham, Lillie et al. 2003; Naumann, Mithieux et al. 2009).

The process of coacervation is thermodynamically driven. It is believed that prior to coacervation water molecules in solution form a clathrate structure of hydrogen-bonded water molecules around tropoelastin monomers. As the temperature or ionic concentration of the solution increases, this clathrate structure is disrupted, and the exposed hydrophobic side-chains fold and interact with each other through hydrophobic forces to form increasingly large coacervates (Urry 1995). This process is reversible, and coacervates can be returned to solution by cooling the surrounding solution.

In the laboratory, this entire process can be monitored by measuring the turbidity of a solution of soluble tropoelastin or elastin-like polypeptides as the temperature of the solution is increased,
using a spectrometer set to 440nm. Recent work on the coacervation behavior of a well-characterized elastin-like polypeptide has shown that coacervation is a two step process, involving an initial coacervation step, marked by a rise in turbidity at the coacervation temperature, followed by a maturation step, characterized by a decrease in turbidity (Cirulis and Keeley 2010). Microscopic imaging of the coacervation droplets in these studies indicated that maturation was due to coalescence of droplets, creating larger colloidal particles which reduced effective light scattering, resulting in a decrease of turbidity.

In most studies of elastin coacervation, the temperature at which coacervation occurs is referred to as the coacervation temperature or Tc, and is taken as a measure of the protein's propensity for self-assembly. The Tc has been found to be influenced by factors such as the size and hydrophobicity of the polypeptide (Miao, Bellingham et al. 2003; Miao, Cirulis et al. 2005) as well as the ionic concentration of the solution and the concentration of soluble molecules. In general, Tc decreases with a decrease in the mean Kyte-Doolittle hydophobicity, or an increase in ionic concentration, or an increase in monomer concentration. These studies indicate that while external factors can affect Tc, and the propensity of a molecule for self-assembly, all things being equal, Tc is directly influenced by the inherent sequence properties of the molecule.
1.8 Lysyl Oxidase and Elastin Cross-link Formation

One of the most remarkable characteristics of fibrous elastin isolated from tissues is its complete insolubility in physiological buffers and detergents. The change from the soluble monomeric tropoelastin precursor to the insoluble polymeric elastin is due completely to the formation of inter- and intra-covalent cross-links between lysine residues of tropoelastin. These cross-links are vital to elastin's ability to serve as a reversible elastomer by anchoring its elastic components during stretching.

The formation of lysine cross-links is catalyzed by the enzyme lysyl oxidase (LOX), which oxidizes peptidyl lysine to α-amino-adipic δ-semialdehyde, a reactive precursor which spontaneously condenses with vicinal aldehydes or ε amino groups of unmodified lysine residues to form covalent cross-linkages, the most common of which are lysinonorleucine, allysine aldol, desmosine, and isodesmosine (Figure 1-5) (Pinnell and Martin 1968; Mithieux, Wise et al. 2005).

LOX is a 28-32kDa (dependant on originating species) extracellular enzyme which has been shown to be capable of cross-linking both elastin and collagen (Smith-Mungo and Kagan 1998). LOX purified from bovine, chicken, and human placenta all resolve into four or more bands when run through salt gradient elution chromatography on DEAE cellulose, each representing a functional variant of the protein (Robert and Hornebeck 1989). Considering the difference
between the sequences surrounding the substrate lysines in elastin and collagen, it was suggested that these variants were responsible for LOX's ability to use widely variable substrate lysines. However, in vitro studies have shown that all the variants oxidized collagen and elastin at similar rates (Sullivan and Kagan 1982).

Proper enzymatic function of the LOX protein requires a tightly bound copper as its cofactor, and removal of copper renders the enzyme inactive (Siegel, Pinnell et al. 1970). Animals fed a copper deficient diet have been found to contain numerous structural defects in their connective tissues, including fragmentation and dissolution of aortic elastic laminae resulting in aneurysms (Shields, Coulson et al. 1962). The elastin in the tissues of these animals have been found to contain increased levels of lysine and a decreased amount of desmosine cross-linkages (Miller, Martin et al. 1965). These findings highlight the vital role that LOX plays in the stability of connective tissue, as well as the importance of proper cross-links formation in collagen and elastin.

Despite the importance of LOX in vivo, in vitro studies of cross-linking in elastin and elastin-like polypeptides seldom use this enzyme as a cross-linker because it is highly difficult to purify in an active form (Mithieux, Wise et al. 2005). Instead, various other means of chemical cross-linkers, including isocyanates, HNS-esters, phosphines, aldehydes, quinones, and genipin have been used (Sallach, Cui et al. 2009).
Figure 1-5 Formation of elastin cross-links by lysyl oxidase (LOX). LOX catalyzes the conversion of lysine to allysine, initiating subsequent spontaneous condensation reactions between these molecules to form a number of lysine-based cross-links, such as desmosine and isodesmosine, the two most common cross-links found in elastin. Figure from (Vrhovski and Weiss 1998), used with permission from publisher.
1.9 Evolution of Elastin

The study of elastin evolution began in the late 1970s, when Helene Sage and William Gray used amino acid composition and histological staining to search for the presence of elastin in the arterial tissues of 43 vertebrate species and 12 non-vertebrate species. They discovered that elastin was present in all vertebrates except cyclostomes (jawless fish) (Sage and Gray 1979). This finding is still valid, as no elastin have been discovered in non-vertebrates (Davison, Wright et al. 1995), however, a protein has been found in lamprey, with similar repetitive motifs to elastin, suggesting an evolutionary relationship between the two (Robson, Wright et al. 1993).

With the increasing availability of sequence information in the 1990s and onwards, comparisons of elastin sequences from various species became possible. When the full amino acid sequence of rat tropoelastin was sequenced in 1991 using overlapping cDNA clones, comparisons of the protein sequences of rat, human, and bovine tropoelastin were made. From these studies, it was determined that the hydrophobic regions in the tropoelastin protein diverged twice as fast the cross-linking regions, and that compared to other structural proteins such as collagen, the amino acid sequence of tropoelastin is highly divergent.

Multiple sequence alignment on complete elastin sequences of eight mammalian species further revealed that unlike other genes in the region (such as RFC2 and LIMKK1 genes) which produced well-conserved non-gapped alignments, elastin was highly divergent, producing
multiple-sequence alignments with numerous gaps caused by species-specific insertions and deletions (Piontkivska, Zhang et al. 2004). However, key elements of the gene, such as splice junctions, as well as the alternation of hydrophobic and cross-linking regions were well conserved. These findings suggest that the elastin sequence contains a subset of functional important regions responsible for its key properties, as well as regions with higher tolerance for insertions and deletions, which may reflect the different biological requirement of elastin in different species.
Figure 1-6 Multiple sequence alignment of elastin protein sequence from human, baboon, cat, dog, cow, pig, mouse, and rat showing gaps and generally poor alignment. Amino acids are coloured based on similarity of chemical characteristics. Exon boundaries are indicated by the arrows. Figure from (Piontkivska, Zhang et al. 2004), used with permission from publisher.
The discovery that frog and zebrafish each had two unique copies of the elastin gene expanded the phylogenetic range of available elastin sequences to teleosts and amphibians, further providing clues to the evolution of elastin (Chung, Miao et al. 2006). Glycine content in the elastins of these organisms was significantly higher than in those found in mammalian and avian species. In addition, the lengths of the hydrophobic regions in teleosts and amphibians were found to be longer than their counterparts in mammalian and avian elastins.

Intra-sequence similarity comparisons performed on all available full-length elastin sequences revealed regions of relatively high similarity between exons 8-14, as well as regions of duplication between exons 14-26 (Chung, Miao et al. 2006). This observation is consistent with previous postulations that one possible mode of elastin evolution was the duplication of regions containing a cross-linking site followed by a hydrophobic region (Gray, Sandberg et al. 1973). This suggests that the core region between exons 14-26 in elastin may encode for functionally important properties of elastin. Outside of these regions, and the well-conserved N-terminal signal peptide and C-terminal region, the rest of the elastin gene, particularly the hydrophobic regions, have evolved with fewer constraints. This may ultimately be the explanation for the diverse range of elastin sequences, each tailored for the specific needs of the organism.
1.10 Elastin-like Polypeptides

Tropoelastin cannot be readily acquired in large quantities from natural sources. Early studies of tropoelastin using copper deficient pigs produced small amounts of the protein with a mixture of partially cross-linked and oxidized lysyl side chains (Mecham and Foster 1979). Chemical or enzymatic methods to solubilize the highly cross-linked elastin found in animal tissues are highly destructive and can cleave the peptide bonds of tropoelastin. Prior to the introduction of recombinant tropoelastin proteins, the difficult task of acquiring sufficient amounts of tropoelastin was a limiting factor in the study of elastin structure and function.

The first recombinant human tropoelastin protein was successfully produced in 1990 using a human cDNA clone derived from fetal human aorta (Indik, Abrams et al. 1990). Previous unpublished attempts to express tropoelastin as free polypeptides were unsuccessful as the proteins were rapidly degraded in the bacterial hosts. Indik et al. applied a modified procedure by expression tropoelastin as a fusion protein connected to 81 amino acids of the influenza virus NS1 protein by a methionine residue. This method took advantage of the lack of methionines in the tropoelastin sequence, which allowed efficient cleavage by cyanogen bromide (CNBr) to release the tropoelastin product. Subsequent, modifications to the cDNA sequence to account for rare codons were introduced to increase protein yield (Martin, Vrhovski et al. 1995), but the overall strategy of recombinant tropoelastin expression and purification has remained unchanged.
Extensive work has been done using recombinant elastin-like polypeptides (EPs), based on the human tropoelastin sequence. These polypeptides maintain the alternating hydrophobic/cross-linking architecture of the human sequence, but have the advantage of allowing studies focused on specific regions or patterns of interest. EPs have been shown to coacervate-like tropoelastin (Bellingham, Woodhouse et al. 2001), and cross-linked EPs can form elastic materials with similar elastomeric properties to those of native tropoelastin (Bellingham, Lillie et al. 2003). In addition, the ability to introduce mutations into these EPs make them an excellent platform to amplify the effects of potential changes caused by mutations without the complication of using large, full-length tropoelastin.

For the purposes of this thesis project, EPs based on representative hydrophobic exons 20 and 24 and cross-linking exons 21 and 23 were used. These were chosen because EPs based on these exons had similar behavior to full-length tropoelastin (Miao, Bellingham et al. 2003), as well as for the presence of potential sites of mutations which full within these exons.

1.11 Comparison of Elastin to Other Elastic Proteins

Elastin is a member of a category of structural proteins known as elastic proteins. In general, these proteins possess the ability to withstand a large degree of deformation without breakage. Depending on the biological requirement of the protein, some or nearly all of the stored energy is released with a return to the original conformation of the protein.
Like elastin, most elastic proteins are insoluble with covalent cross-links, which have made them difficult to study. As result, detailed studies of elastic proteins have been limited to a small number of well-known proteins including abductins, resilins, wheat gluten protein and spider silk proteins. With the exception of wheat gluten, which does not possess cross-linking regions, these elastic proteins share the remarkable characteristic of having highly repetitive elastomeric regions separated by shorter cross-linking regions. However, as their biological requirements are different, the sequences of these proteins are not well conserved, even though they all derive their elastomeric properties from intrinsically disordered sequences.

1.11.1 Abductin

Abductin is a protein found in the hinge ligament of mollusks, storing energy when the abductor muscle contracts and opens the shell of the mollusk when the muscle relaxes (Bowie, Layes et al. 1993). Among well-characterized elastic proteins, abductin is one of the smallest at only 136 amino acid residues in length. The protein is composed of only two domains: a short alanine-rich domain with two cross-link capable tyrosines, and a longer highly repetitive glycine-methionine-rich region responsible for elasticity (Cao, Wang et al. 1997). However, unlike most other elastic proteins, the elastomeric region in abductin also contains two lysine residues which may be capable of cross-link formation.
1.11.2 Resilin

Resilin is a protein found in the wing hinges of insects. Its main role is to dampen the vibrational energy of flight. In *Drosophila melanogaster*, resilin has been identified as a 603 amino acid residue protein with three regions: a highly repetitive N-terminal region, a non-repetitive region in the middle, and finally a repetitive C-terminal region. The repetitive region has been predicted to form β-turns and possibly loose β-spirals (Ardell and Andersen 2001). Unlike most other elastic proteins, cross-linking tyrosine residues are present throughout the protein (Ardell and Andersen 2001).

1.11.3 Wheat gluten protein

Wheat gluten proteins are a major class of storage proteins found in the starch-rich endosperm cells of grain. While these proteins need not fulfill a naturally occurring biological requirement for elasticity, humans have discovered millennia ago that when mixed with water, these proteins come together to give elasticity to dough in the production of popular forms of baked goods. Among wheat gluten proteins, the most well-studied group are the so-called High Molecular Mass (HMM) subunits, which contain a large repetitive domain which adopts β-turn and β-sheet conformations (Tatham, Marsh et al. 1990).
1.11.4 Spider Silk Proteins

Spider silk is a generic term for the silk which spiders extrude for the formation of their webs. However, within the broad category of spider silks there exist several silk proteins, each fulfilling a different biological role. The two main types of silks are flagelliform silks and dragline silks. The latter is required predominantly for tensile strength, capable of being stretched to only approximately 30% from the relaxed state (Gosline, Demont et al. 1986). Flagelliform silks, on the other hand, are used for the spirals of the spider's webs. Their main function is to provide elasticity upon prey capture. While both dragline and flagilliform spider silk proteins have a sequence style resembling those of other elastic proteins, with repetitive, hydrophobic elastomeric regions separated by spacer/cross-linking regions, the more elastic flagilliform spider silk proteins have longer elastomeric regions than dragline proteins (Hayashi and Lewis 1998).

1.12 Project Goals

Elastin production decreases significantly following elastic fibre formation. In the human cardiovascular system, elastin has a very long half-life of nearly 70 years, with a very low rate of turnover. It is likely that changes in the sequence of elastin, particularly in the regions which play a role in the coacervation and self-assembly of elastin can adversely affect the proper alignment, coacervation, and deposition of elastin, leading to incorrect formation of elastic fibres. This can lead to architectural changes in the resulting fibres which under millions of stretch and recoil
cycles in the human vascular system may show signs of premature weakening, causing late onset cardiovascular disease.

The goal of this thesis is to discover the first step in the elucidation of the relationship between the sequence and function of the elastin protein, potentially revealing novel links to diseases. In summary, my project aims are:

1) A systematic identification and characterisation of globally over-represented repetitive elements in elastin to reveal regions which may play an important role in the self-assembly and elasticity of elastin.

2) Selection of potential sites of mutation based on existing population studies like the HapMap project, and publically available databases such as dbSNP. This will be followed by the introduction of mutations in these sites to produce modified elastin-like polypeptides, whose physical properties of self-assembly and elasticity can be measured and compared.

Ultimately, it is the aim of this project to identify functionally important sequence motifs in the elastin protein, and to describe the role they play in the proteins measureable behaviour. The outcome of this work will serve as a useful first step in future investigations of the potentially damaging effects that genetic mutations in these motifs may cause, possibly leading to the discovery of links between genetics and disease. Furthermore, a better understanding of the
sequence-function relationship of elastin can be useful in the understanding of other elastic proteins found in nature, as well as improving the design of novel biomaterials.
Chapter 2 Survey of elastin sequence

Note: The text and figures in this chapter have been adapted from (He, Chung et al. 2007) and (He and Parkinson 2008), and are being reproduced with permission from the publishers.

2 Survey of elastin sequence

2.1 Introduction

The primary aim of this thesis project is to establish preliminary links between biologically important repetitive elements in the elastin sequence and the mechanical and functional properties of the protein. To accomplish this goal, an important first step was the creation of a catalogue of potentially important repetitive elements. Studies of tropoelastin sequences have revealed that the amino acid sequence of this protein is rich in penta- and hexameric repetitive elements, however no comprehensive comparisons had been completed which take advantage of the many newly sequenced tropoelastin, which have only recently become available. In depth analyses of elastin sequences have traditionally been performed on the limited number of organisms such as human, bovine, and pig, for which elastin products could be readily acquired. However, with recent advances in sequencing technology, elastin sequences have now been identified in 28 different species including zebrafish, xenopus, chicken and several mammals (Chung, Miao et al. 2006). In each example of elastin discovered to date, the general glycine,
valine, and proline rich amino acid composition is retained. In addition, the alternating pattern of exons coding for hydrophobic regions interspersed by shorter exons encoding cross-linking regions can be found in all elastins which have been identified. However, there is remarkable variability in the actual sequences (for example, alignment of the human and mouse elastin cDNA sequences shows a 64.5% identity compared with an average percent identity of 85% between human and mouse cDNAs (Piontkivska, Zhang et al. 2004)), suggesting that the assembly and functional properties of elastin require conservation of a general 'style' of sequence rather than a specific sequence (Hinek and Rabinovitch 1994; Kielty, Sherratt et al. 2002; Kozel, Wachi et al. 2003; Mithieux and Weiss 2005).

The exact relationship between this general style and the functional and mechanical properties of elastin is still unclear. It is known, however, that differences in the sequence of domains and their organization can alter both the assembly and functional properties of the protein (Miao, Bellingham et al.). These alterations may be useful for the adaptation of elastin to various biological requirements. For example, differences observed between the elastins of teleosts and mammals may be responsible for the former’s ability to assemble and function as an elastomer under poikilothermic (cold-blooded) as opposed to homeothermic (warm-blooded) conditions (Sage 1982; Miao, Bellingham et al. 2003). While there is little debate that the essential elastin properties--self-assembly, durability, and elasticity--are related to the unusual sequence and domain organization of this protein, it is still not clear how the repetitive sequences of elastin directly impacts these properties. By collating and comparing the repetitive sequences from the elastin sequences of different species, this stage of thesis project yielded a catalogue of potentially important repetitive elements as well as evolutionarily conserved regions, whose
impacts on the physical and functional properties of the elastin protein could be studied in the later parts of this project.
2.2 Materials and Methods

2.2.1 Collation of elastin sequence

A large effort has been devoted to the identification and annotation of genes from newly sequenced genomes. Due to the scale of the task, automated methods are typically applied in a first screen to identify exon boundaries. Consequently, low complexity and/or repetitive sequences such as elastin, are often masked, and the additional manual intervention required to annotate repetitive proteins is often beyond the scope of most sequence annotation projects. To perform an in depth analysis of the sequence properties of elastins from a wide spectrum of species, a more specialized database focusing on elastin, called ElastoDB (http://www.theileria.ccb.sickkids/elastin) was developed.

Elastin sequences were obtained using comprehensive BLAST and Hidden Markov Model (Genscan) searches applied to a variety of datasets, including genomic scaffolds, expressed sequence tags (ESTs) and cDNA sequences deposited in Genbank. Theses searches were performed to identify sequences sharing sequence similarity to previously identified elastin sequences derived from Genbank. Partial sequences obtained through EST datasets were manually curated using the sequence analysis tool Artemis (Rutherford, Parkhill et al. 2000). Gene models of elastin sequences were predicted from genomic scaffolds using Genscan (Burge and Karlin 1997) and also manually curated.
The resulting database contained full or partial elastin sequences from 24 different species including amphibians, birds, mammals and bony fish that have been curated in-house from data derived from public databases. Of these, nine represent full length genes from *Homo sapiens* (human), *Sus scrofa* (pig), *Bos Taurus* (cow), *Felis domesticus* (cat), *Canis familiaris* (dog), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Gallus gallus* (chicken) and *Takifugu rubripes* (fugu). In addition *Danio rerio* (zebrafish) and *Xenopus tropicalis* (frog) possess two variant copies of elastin genes (designated eln1 and eln2) (Chung, Miao et al. 2006). An additional three sequences are thought to be incomplete due to lack of available sequence data and/or errors in sequence assembly: *Papio anubis* (baboon) is missing the first exon; *Pan troglodytes* (chimpanzee) is missing a number of exons from the middle of its sequence; and *Pongo pigmaeus* (orangutan) derived from assembled expressed sequence tag (EST) data contains only the first 11 exons. For the purposes of this project, only elastin with complete sequences were used.

2.2.2 Sequence survey

The amino acid sequence of all elastins discovered to date possess a distinct style which can be characterized by a 1) high abundance of non-polar amino acids such as glycine, valine, alanine, and proline 2) distinct hydrophobic regions, made up mostly of glycine, valine, and proline, and cross-linking regions rich in alanines, and sometimes prolines, interspersed with lysines. This style of sequence results in a very low-complexity sequence. Initial surveys of the elastin amino
acid sequence were performed to compare the glycine, proline, and valine content of elastins found in different species, as well as to compare the sequence distribution of elastins versus other more conventional proteins. Total percentage of proline and valine content found in full-length elastin amino acid sequences from ten species were determined and plotted against the glycine content of the protein. Data from collagens, spider silks, and other proteins were also incorporated as a comparison to demonstrate the unique sequence composition of elastin sequences.

To demonstrate the low complexity of elastin sequences, a sliding window was used to produce a pool of 3-mers, 4-mers, and 5-mers. These were than grouped together and plotted as a histogram ranked by abundance. Similar subsequence distributions were created for a range of proteins, demonstrating that elastin sequences tended to be dominated by a small number of highly abundant short subsequences.

2.2.3 Motif identification and characterization

Several state-of-the-art pattern matching methods such as MEME (Bailey and Elkan 1995), Consensus, and TEIRESIAS (Rigoutsos and Floratos 1998) have been developed to identify motifs from highly variable sequences. The TEIRESIAS algorithm searches sequences to discover long, simple patterns using a bottom-up approach, first assembling shorter patterns before joining them. Consensus is another pattern searching algorithm, which adopts a top-down
approach allowing for the discovery of longer motifs. However, when these tools were applied to elastin by Trevis Alleyne, a former summer student, it was discovered the short, highly variable repetitive motifs in elastin were not readily identified.

In an attempt to systematically identify functionally important motifs from elastin, a novel and innovative approach was devised based on the identification of overrepresented ‘fuzzy’ motifs, using methods borrowed from protein interaction studies. This algorithm produced a network of "interactions", defined by two nodes, representing subsequences within the larger sequence connected by directional edges, representing the adjacency of the two subsequences. The end result is a visual representation of the elastin sequences which clearly showed important "hubs", subsequences which are the common building blocks of larger motifs, and these hubs could be connected by adjacency edges to build larger and more informative motifs.

A schematic of the algorithm used to derive the polypeptide neighbor maps is shown in Figure 2-1. An input sequence is first partitioned into overlapping subsequences consisting of an arbitrary number of residues. Each subsequence is then paired with the subsequence directly adjacent to it, and a list of these subsequence neighbor pairs is compiled. Wild-card characters representing any amino acid residue, are then introduced such that two subsequence pairs which are each observed once (e.g. APG->VGG and GPG->VGV) would become a twice-observed subsequence pair (*PG->VG*). This step groups subsequence pairs into classes and serves two useful purposes: 1) it allows the creation of more flexible patterns, and 2) it reduces the amount of subsequence pairs, amplifying potential signals. From the list of subsequence pairs with wild
card characters, the top 2 percent of the most frequently observed subsequence pairs were retrieved.

A network is then constructed from these subsequence pairs using the program Cytoscape (Shannon, Markiel et al. 2003). The maps consist of graphs in which nodes represent a fixed-length subsequence and edges connect neighbouring subsequences. The direction of the arrow represents the order of the subsequences, pointing from the N-terminus to the C-terminus. The thickness of the edges connecting two nodes is a function of the number of observed subsequences which were normalized for sequence length to provide the incidence of subsequence pairs per 1000 residues. This normalization allows comparisons between elastins of different length. Edges were coloured based on the observance of the subsequence pair in the top 2 percent of groups of species. For example, a subsequence pair which can be found in the top two percent of all full-length elastins was coloured red. A control network was created by picking the top 2 percent of the most frequently observed subsequence pairs generated from 100 random sequences with the same length and amino com-position as a modified human elastin sequence.

Subsequence pairs which had potential to be part of larger repetitive elements, were mapped back to the original elastin sequence. Using 20 amino acids adjacent to these chosen subsequence pairs, sequence logos, a type of graphical representation of amino acid frequency in a given location, were generated using Weblogo (http://weblogo.berkeley.edu/logo.cgi) (Crooks, et al., 2004).
Figure 2-1 Schematic overview of the creation of a subsequence network from a protein sequence. The algorithm splits sequences into user defined sizes of subsequence with the additional caveat that ‘wild cards’ may be introduced, leading to the characterization of ‘fuzzy’ repeats (e.g. AP* -- where * represents any amino acid). The algorithm then builds a network map in which common adjoining subsequences are readily identified through connecting arrows.
For the purposes of this project, subsequences of sizes 3-6 were used with wild-card numbers ranging from 1 to 3. It was determined based on the resulting network maps that for elastin sequences, which are rich in 5-mer and 6-mer repeats, that a subsequence of size 3 with 1 wild card character produced the most informative networks. However, it should be noted that this method may be readily adapted to support any arbitrary number of residues and wild cards. Indeed, a web-based version of this algorithm called SubSeqer (see Section 2.3.4) was subsequently developed to allow researchers of other low-complexity sequences to identify flexible motifs, using a web interface, in lieu of the Cytoscape program.

2.2.4 Exon Evolution

In addition to the repetitive elements found throughout the elastin sequence, there is also similarity between certain regions of the protein, which may have undergone replication. Previously it was noted that putative replicated regions in elastin include not only exon sequence but also adjoining intronic regions (Chung, Miao et al. 2006). To test this hypothesis on a wider sample size of elastin sequences from different species, individual elastin exons from each full-length elastin sequence available were extracted (492 exons in total), together with a 50 bp region from the surrounding introns. These sequences were aligned against each other using the EMBOSS (Rice, Longden et al. 2000) program, water, which uses an implementation of Smith-Waterman algorithm to find the best local alignment.
A gap penalty of 20.0 was used to reduce the introduction of gaps in the alignments. A global alignment matrix was then constructed consisting of Smith-Waterman alignment scores for each of the 492 exons aligned to all others. In addition, species-specific alignment matrices were constructed in which only exons from the same elastin sequence were included.

Single linkage hierarchical clustering based on the Spearman correlation coefficient was performed using the software tool Cluster 3.0 to group exons on the basis of their sequence similarity profiles (Gardner, Hall et al. 2002). Visualization of clusters was performed using TreeView (Saldanha 2004). On the basis of common patterns of clustering, pairs of exons and their adjoining introns were used to perform alignments using the EMBOSS program, needle, which uses a Needleman-Wunsch algorithm to find optimal global alignments. This algorithm is particularly well suited to the analysis of sequences that are deemed to be closely related, resulting in the formation of longer and more accurate alignments. The scores of each of these alignments were used to create matrices which were clustered and visualized as above.
2.3 Results

2.3.1 Distribution of simple repeats in elastin

The collation of high quality sequence data from a variety of different elastins provided a unique opportunity to examine properties of sequence conservation and identify the features that give rise to the remarkable properties of self-assembly and elastic recoil. Similar to other elastomeric proteins (Tatham and Shewry 2000), elastin possesses an atypical amino acid composition, rich in short repetitive regions made up of proline (P), valine (V) and glycine (G) in hydrophobic regions and alanines (A) and lysines (K) in the cross-linking regions (Figure 2-2). This distinct signature of amino acid composition is thought to underpin elastin’s properties of self-assembly and elastic recoil. For example, molecular dynamics simulations have previously suggested that a minimum threshold of combined P, G content may be responsible for providing elastin with a critical level of disorder required to impart elastomeric properties (Rauscher, et al., 2006). Previous structural studies of elastin have identified characteristic repetitive elements (motifs) consisting of between 3-6 residues thought to play an important role in determining the secondary structure of the protein (Pepe, et al., 2005; Tatham and Shewry, 2002; Urry, et al., 1989; Yao and Hong, 2004).
Figure 2-2 Graphs showing the relationships between G, P and V compositions for a range of proteins – (A) G-P composition. (B) V-P composition. (C) G-V composition. Each point on the graphs represents a unique protein associated with various classes of proteins. In addition to the 12 elastin sequences, sequences from actin (93 examples), tubulin (255), collagen (40) and spider silk (95) proteins were obtained from UniProt with reference to Gene Ontology annotations derived from ftp://ftp.geneontology.org/pub/go/gene-associations/ (Ashburner, Ball et al. 2000; Harris, Clark et al. 2004).
In an attempt to systematically detect commonly occurring motifs within elastin, the distribution of peptides consisting of three, four and five residues from a range of structural proteins was analyzed (Figure 2-3). Elastin, type-I collagen alpha 1 chain from human, and dragline spider silk from Nephila clavipes all showed a greater abundance of repeats compared with the other proteins examined. Compared with spider silk, both elastin and collagen show a reduction in the repetitiveness of their sequence, highlighting a greater variability in the composition of the repeats. The abundant repeats in human elastin (Figure 2-3) highlight the conserved nature of the cross-linking domains consisting of repetitive stretches of alanine interspersed with lysine. In addition, note also the high incidence of repeats composed of P, G and V reflecting the distinct sequence composition associated with elastin. However, it should be noted from the absence of GP and PV dipeptides from these repeats, that the order of the residues is not random. Indeed, from the analysis of the five-mer repeats, three of the top six most abundant repeats are VPGVG, PGVGV and GVPGV. Mapping of these motifs onto the full length human elastin sequence revealed that they represent two major discrete elements: GVPGVG and PGVGVA (Figure 2-3D). This supports the notion that these repeats represent functionally conserved units imparting elastin-specific properties. This is consistent with literature, where for example, it has previously been noted that the peptide, VPGV, is required for the formation of β-turns that are thought to contribute to the elastic mechanism of elastin (Urry and Parker, 2002).
Figure 2-3 Incidence of 3mer, 4mer and 5mer repeats for a range of proteins. (A-C) Graphs showing the relative frequency of the top 10 most abundant 3mer (A), 4mer (B) and 5mer (C) repeats for human elastin and 6 other structural proteins. Distributions of repeats within the selected sequences were determined using a sliding-window approach in which the sequence of interest was divided into overlapping subsequences of fixed size (consisting of 3, 4 and 5 residues). Inset in each graph is a table listing the incidence of each of the top ten most abundant repeats. (D) Mapping of three abundant hydrophobic 5mer repeats (VPGVG - blue, PGVGV - green and GVPVG - red) onto the sequence of human elastin.
2.3.2 A graph based approach for the detection of functionally relevant motifs in elastin

While the above analysis was able to identify several putative functional motifs composed of three, four, five or six residues, the method was unable to account for the highly variable nature of elastin thus greatly limiting the number of such elements that could be identified (witness the reduction in the numbers of abundant elements composed of 3, 4 and 5mers – Figure 2-3A-C). It was, therefore, necessary to adopt a novel, graph-based method (see Section 2.2.3). Applying this method to the twelve elastins resulted in a rich and diverse set of sequence relationships, of which four representative networks are presented in Figure 2-4. To reduce the complexity of the derived networks and extract the most informative sequence relationships for these analyses, only the top two percent of the most abundant subsequence pairs was considered. This represented 121 subsequence pairs in the human elastin network, 79 in the chicken, 88 in frog eln1 and 167 in zebrafish eln2. The discrepancy in the number of interactions was mainly a result of the length of the elastin sequences as well as the variation of motifs in the sequences, both of which will be explored later.
**Figure 2-4** Network maps detailing relationships between fuzzy 3mer repeats in the human tropoelastin sequence. The thickness of the lines represent the relative abundance of a subsequence pair, while the colour of the edges indicate taxonomic groupings of relative abundance (see legend).

<table>
<thead>
<tr>
<th>Taxon specificity</th>
<th>Normalized number of interactions*</th>
</tr>
</thead>
<tbody>
<tr>
<td>All mammals, chicken, fish, frog</td>
<td>&lt;10</td>
</tr>
<tr>
<td>3+ mammals and chicken</td>
<td>10-20</td>
</tr>
<tr>
<td>3+ mammals</td>
<td>20-30</td>
</tr>
<tr>
<td>&gt;30</td>
<td>&gt;30</td>
</tr>
</tbody>
</table>

*Interactions per 1000 residues
Figure 2-5 Network maps detailing relationships between fuzzy 3mer repeats in the chicken tropoelastin sequence. The thickness of the lines represent the relative abundance of a subsequence pair, while the colour of the edges indicate taxonomic groupings of relative abundance (see legend).
Figure 2-6 Network maps detailing relationships between fuzzy 3mer repeats in the frog eln1 sequence. The thickness of the lines represent the relative abundance of a subsequence pair, while the colour of the edges indicate taxonomic groupings of relative abundance (see legend).
Figure 2-7 Network maps detailing relationships between fuzzy 3mer repeats in the zebrafish eln2 sequence. The thickness of the lines represent the relative abundance of a subsequence pair, while the colour of the edges indicate taxonomic groupings of relative abundance (see legend).
Each interaction consisting of two 3mers may be thought of as an abundant motif. Highly connected nodes, so called hubs (e.g. PG* and *GV in the human elastin graph), represent 3mers that are found in many motifs. Such hubs may therefore represent a functionally conserved unit providing specific properties to several motifs. It should also be noted that within each network there was a degree of overlap between interactions. For example in the human network, G*V→VP* could contain instances of *GV→V*G and vice-versa. These networks thus served as a visual guide for highlighting potentially important motifs. In terms of the overall topology of the networks, the network derived from zebrafish eln2 contained many more interactions than the other three networks with an increased number of nodes and connections. Furthermore, the thickness of the connections was correspondingly reduced. This reflected both the increased length and greater diversity associated with this sequence. On the other hand, the chicken sequence contained noticeably fewer interactions than the human and frog sequences and thus the network for this species contained correspondingly thicker lines. This reflected the dominance of motifs in this sequence arising from regions of lower complexity.

For each graph, there were also nodes and connections that were both generally conserved across the four species and those that were taxon specific. For example, in all elastin sequences, the PG dipeptide, represented by two nodes – *PG and PG* – and G*G were found to be highly connected nodes (hubs) and involved in abundant pairs of 3mers. This is consistent with literature suggesting that the PG amino acid pair forms an integral part of the β-turn structures commonly found in elastin (Urry and Parker 2002). On the other hand, GV* appeared to be involved in a higher number of connections in human and chicken than zebrafish eln2, while frog
elN1 had a uniquely abundant node consisting of GL*, suggesting an increased substitution of L for V in this species.

As with the repeat analysis performed in Section 2.3.1, no nodes containing the dipeptides PV or GP were identified in these graphs, once again highlighting the non-random nature of these interactions. To further verify that the observed patterns were not simply artifacts resulting from the amino acid composition of the elastin sequence, this algorithm was applied to 100 randomized human elastin sequences (Figure 2-8). In addition to the weaker interactions identified within the random network, there was a reduced abundance of interactions involving the *PG node. Furthermore the random network also contained an abundance of interactions involving both the GP* and *GP nodes, absent in the non-random human elastin network.
Figure 2-8 Network maps detailing relationships between fuzzy 3mer repeats from 100 random sequences with the same amino acid composition of human elastin. Note the reduced abundance of interactions involving the *PG node, as well as an abundance of interactions involving GP*, *GP, PV*, and *PV nodes, absent in the non-random human elastin network (Figure 2-4A).
From Figure 2-4 to Figure 2-7, it is evident that several subsequence relationships were taxon specific. For example the *VP→GV* and related overlapping pairs (e.g. *PG→VG*) were found in relatively high abundance only in the chicken and mammalian sequences. As noted above, VPGV is thought to contribute to the elastic mechanism of elastin (Urry and Parker, 2002). Further examination revealed the presence of nine examples of VPGV in frog eln1 and only a single example in zebrafish eln2. The reduced presence of these motifs in zebrafish eln2, suggests that an alternative motif may be utilized in this species. Several abundant motifs were specific to chicken, including P*V→*GV and *VP→G*P. These were consistent with the presence of a highly repetitive stretch of PGV in exon 20 in the chicken elastin sequence.

Interestingly, for the human, chicken and frog eln1 sequences, this algorithm clearly delineated motifs associated with hydrophobic domains from those associated with the cross-linking domains. For the chicken elastin only sequence relationships between alanines were detected, while the human cross-linking region displays a much more complex set of relationships rich in both alanine and lysine. 3mers associated with cross-linking domains in frog eln1 were split into three regions rich in proline and lysine. The alternative use of KP-type cross-links in frog (and zebrafish) has been described elsewhere (Chung, et al., 2006). The split into three regions highlighted the overlap between subsequences that arise from the use of the wild card. Taken together, all three regions suggest that the cross-linking domains in this elastin are enriched for the motif GGKPPKP. The lack of any such 3mer interactions in zebrafish eln2, reflects the greater variability of the cross-linking regions compared with the other sequences.
**Figure 2-9. Network map showing the global diversification of abundant motifs in elastin across vertebrates.** This network map presents an overlaying of networks derived from the top two percent of the most abundant motifs derived from all four elastin sequence: human, chicken, frog eln1, and zebrafish eln2. This network map highlights conserved, taxon- and species-specific motifs. For example, G*V→PG* was found to be abundant in all examples of elastin presented here.
2.3.3 Expansion of sequence motifs detected in elastin

Subsequence networks demonstrated that there were many differences in the types of motifs found within the elastin sequences of different species. To observe their relative frequency and spacing, certain motifs from the network were mapped back onto their parent sequences (Figure 2-10). In general each motif is regularly spaced with few examples of motifs overlapping more than one exon. This supported the notion that these motifs may form heritable units of structure and/or function. In terms of the conserved motif G*VPG,* there did not appear to be a consistent pattern of organization between the three species. Both frog eln1 and zebrafish eln2 contained two exons rich in this motif. Frog eln1 contained five and seven copies in exons 16 and 18 respectively, while zebrafish eln2 contained seven and 13 copies in exons 8 and 45 respectively. In the zebrafish eln2 sequence, an abundant taxon specific motif (L*TGG*) was found to be most abundant in the middle of its sequence, where there was also a corresponding absence of the other two mapped motifs.
Figure 2-10 Mapping of identified motifs onto elastin sequences. For each elastin, three abundant subsequences were identified from the previously generated networks and mapped onto their respective sequences. Vertical lines within each sequence define exon boundaries.
From these interaction and mapping analyses a number of sequence motifs were identified, consisting of five or six residues that may represent cores of functional modules that impart basic properties to elastin such as its elasticity and its ability to self-assemble. Exploiting sequence logos (Crooks, et al., 2004) these ‘core’ motifs were expanded to help identify additional patterns of amino acid conservation (Figure 2-11). For the motif conserved across all four species G*VPG there was an overall increase in conserved sequence around the core motif for frog eln1 and zebrafish eln2. This may reflect the occurrence of this motif as tandem repeats in this species. For human elastin, this core motif may be expanded to G[GL]VPGV[GP] (where [GL] indicates a G or L at that position), while in chicken elastin, the expanded motif was [VG]GXVPGVG[GPV]. Surprisingly, while this unbiased approach was able to identify VPGV as a highly conserved core motif in these two species (highlighting its ability to form a β-turn important for the elastic mechanism of elastin), the motif was less well conserved in frog eln1 and not even apparent in zebrafish eln2.

Additional extended motifs could also be identified from application of sequence logos to other identified motifs. For example, the V*G*GV motif identified in chicken elastin could be associated with a highly repetitive stretch of VPGVG repeats. More interestingly, the previously identified cross-linking motif in frog eln1 – GGKPPKG – can now be extended to [YG][PA]GGKPPKPGYG, suggesting an important role for the tyrosine residues flanking the cross-linking lysines in this species. In zebrafish eln2, the L*TGG* motif could be further expanded to GGLGTGGLP[GT]. The smaller motif was found to be mainly located between exons 20-30, possibly suggesting a series of duplications in this region of the sequence.
Figure 2-11. WebLogo characterization of identified motifs. Sequence logos can illuminate patterns of amino acid conservation that are often of structural or functional importance and are otherwise difficult to perceive. In each logo, stacks of letters indicate the relative frequency of certain amino acids at each position in the sequence. The overall height provides a guide to the level of sequence conservation associated at that position. From the motifs identified from the network views and mapped onto their parental sequences (Figure 2-10), we created a sequence alignment consisting of all examples of the motifs identified within individual elastin sequences and the 10 residues immediately up and downstream of the motif. (A) Sequence logos of the abundant conserved motif G*VPG derived from the four elastins: human, chicken, frog eln1 and zebrafish eln2. (B) Sequence logos of the additional eight motifs highlighted in Figure 2-10.
2.3.4 SubSeqer: a web-based motif finding tool

The utility demonstrated by the network-based motif finding approach led to the creation of a web-based implementation of the method, creating a tool called SubSeqer (He and Parkinson 2008). SubSeqer is similar in broad concept to the method mentioned in Section 2.2.3, but has been designed to be more flexible. Users are allowed to specify the size of subsequences used for each node, and a more robust scoring mechanism was introduced to help with filtering subsequence relationships. This is accomplished by assigning a score to each subsequence pair which is a function of the number of times each subsequence is observed relative to the total number of subsequences found in the initial sequence. It is found by \( \log((s \times t) / n^2) \), where \( s \) and \( t \) are the number of times the first and second fuzzy subsequences in the pair are observed, and \( n \) is the total number of subsequences (whose sizes are specified in the 'Subsequence size' parameter) found in the sequence. The lower the Odds Score, the higher the odds that the subsequence pair did not occur randomly. Only subsequence pairs with an odds score in the top X percentile are used in the final visualization, ensuring that only significant interactions are included in the analysis. This value X is the Percentile Odds Score Cutoff, and is user modifiable.

The frontend of the application was written using the PHP programming language, and guides the users in the process of sequence analysis (Figure 2-12). Users begin by entering their sequence of choice, and are quickly presented with a subsequence distribution histogram similar to those seen in Section 2.2.2. The significance of the distributions are explained to the user, and if the resulting distribution resembles that of a low-complexity sequence similar to that of elastin,
which produces a small number of highly over-represented subsequences, the user is then shown a list of input options such as subsequence size and Percentile Odds Score Cutoff. The resulting network graph is shown using a Java canvas, and is completely interactive. Users can organize the graph using a number of popular preset layouts, zoom into regions of interest, and connected nodes can be selected to create sequence logos (Crooks et al 2004) based on regions of the entire query sequence composed of the subsequences represented by the selected nodes flanked on either side by 10 additional nucleotides or amino acids. The latter feature allows the user to expand upon the basic motif of interest to discover longer, more complete motifs.

Figure 2-12. User interface of the SubSeqer web application. (a) Histogram of subsequence distribution based on the user's input sequence. (b) Network diagram of subsequence interactions. (c) Weblogos of interactions of interest.
2.3.5 The role of domain expansions in the evolution of elastin

From the network analyses, several motifs that were relatively regularly spaced across many different exons were identified. These may represent core functional modules that are responsible for imparting properties of assembly and/or elastic recoil to the full length sequence of elastin. It is well-known that a major driving force in evolution is duplication followed by divergence (Chothia and Gerstein 1997; Chothia, Gough et al. 2003). Hence, given that many of these motifs were taxon specific, it may be possible that they have evolved via a series of duplications to impart particular physical properties. For example it has previously been reported that in all examples of elastins discovered, exons 8-14 form a relatively well-conserved core, while exons 15-30 are the result of domain expansions (Chung, Miao et al. 2006).

Due to the high variability of exons within elastin sequences, it was not possible to use conventional multiple-sequence alignment and phylogenetic algorithms to reveal the evolutionary history of individual exons. Rather, less sensitive measures such as clustering of sequence alignment scores had to be adopted to gain such insights (Bjorklund, Ekman et al. 2006; Street, Rose et al. 2006). To examine sequence relationships between exons, a systematic pairwise sequence alignment was performed between each combination of the 492 exons derived from the 12 full-length elastin sequences associated with the 11 species. Sequence alignment scores were extracted and used to cluster the exons into groups on the basis of their sequence similarity profiles. Initially, comparisons were made on individual exons from the twelve elastin sequences together.
The result of using clustering algorithms on these pair-wise alignments was a series of clustergrams, n x n matrices where the rows and columns represent exons and each entry was coloured based on the alignment score between two exons. From the derived clustergram, several key groupings were observed (Figure 2-13). Firstly, while exons from the chicken and mammalian elastins cluster together, with the exception of a handful of exons from zebrafish eln1 and eln2, frog and fish exons tend to form clusters only with other exons from the same species (e.g. exons 20, 22, 24, 26, 28 and 30 from zebrafish eln1 and exons 15-20 and 24 in fugu). In addition there was an obvious split between exons associated with cross-linking domains and those associated with hydrophobic domains. While this may appear unsurprising, it indicates that the two types of exons are composed of distinct sequences (i.e. there is very little overlap between hydrophobic regions and cross-linking regions within the same exon).

Furthermore, clustering of these exons suggests that there is a common pattern of sequence conservation, indicating that they may represent duplications – these include the hydrophobic exons: 16, 18, 20, 22, 24, 26, 28 and 30 in mammals and chicken and the corresponding cross-linking exons: 15, 17, 19, 21, 23, 25, 27 and 29.
Figure 2-13. Clustergram detailing sequence similarity relationships between 492 exons from 12 elastin sequences. A Smith-Waterman search was performed for each exon (in addition to 50bp of intronic regions up and downstream of each exon) against every other exon (and neighbouring intronic sequence) to build a profile of similarity scores for each exon. These profiles were then clustered by simple hierarchical clustering to derive the presented clustergram. Clusters of exons sharing high sequence similarity are represented as red squares. Unless otherwise indicated exons are derived from mammalian or chicken sequences.
Given that similar patterns of conservation of alternating cross-linking and hydrophobic exons were also observed in zebrafish eln1 and eln2 (exons 20-31) and frog eln1 (exons 35-51), there exists the intriguing possibility that the exons in these regions may have undergone a series of duplication events, as previously suggested (Chung, Miao et al. 2006). Interestingly, fugu elastin displayed a single cluster of exons (15-20 and 24) that might indicate an additional duplicated region in this species. Many of the other exons from these species formed distinct taxon-specific clusters with no sequence relationship to exons from outside their group, suggesting that these exons represent conserved exons that share a common ancestor. This is particularly relevant in the cluster formed by the C-terminal exon of mammals and chicken (exon 36).

To examine these exon relationships in more detail, a cluster analysis using exons only from individual species was also performed. Consistent with the findings from the larger clustergram, it was shown that for chicken and mammalian sequences, only the hydrophobic and cross-linking exons mentioned above form distinctive groupings (Figure 2-14). In the two zebrafish sequences (Figure 2-15), in addition to two clusters representing cross-linking and hydrophobic domains there was an additional cluster comprised of exons closer to the C-terminus (~43-56 in eln1 and ~39-58 in eln2). Again it is interesting to note that the fugu sequence did not display the same pattern of clustering compared with the other species, having only a single cluster composed of exons 15-20 and 24.
Figure 2-14 Clustergram detailing sequence similarity relationships between exons from mammals and chicken species. A Smith-Waterman search was performed for each exon (in addition to 50bp of intronic regions up and downstream of each exon) against every other exon (and neighbouring intronic sequence) to build a profile of similarity scores for each exon. These profiles were then clustered by simple hierarchical clustering to derive the presented clustergram. Clusters of exons sharing high sequence similarity are represented as red squares.
Figure 2-15 Clustergram detailing sequence similarity relationships between exons from zebrafish, xenopus, and fugu species. A Smith-Waterman search was performed for each exon (in addition to 50bp of intronic regions up and downstream of each exon) against every other exon (and neighbouring intronic sequence) to build a profile of similarity scores for each exon. These profiles were then clustered by simple hierarchical clustering to derive the presented clustergram. Clusters of exons sharing high sequence similarity are represented as red squares.
In these analyses it should be noted that due to the relatively short sequences involved, the derived trees associated with the species-specific clustergrams did not show a consistent set of sequence relationships. For example exon 23 and 25 appeared to be most similar in the baboon sequence, while in the cat sequence exon 23 appears to be more similar to exon 19. Given the consistency in the clustering of exons from the same region of the sequences, however, it can be postulated that the region containing exons 15-30 in the mammalian and chicken sequences may have evolved through the duplication of a neighbouring pair of exons containing a single hydrophobic and cross-linking domain. Therefore, in an attempt to derive more consistent sets of relationships from these exons, pairs of exons were combined together with their adjoining introns to undergo more sensitive analyses.

Due to the variability in the introns, it was not possible to derive convincing multiple sequence alignments and therefore we relied on the use of clustergrams generated from similar similarity profiles based on global pairwise alignments (Figure 2-16). For the mammalian sequences, only sequences from human, baboon and mouse displayed strong patterns of similarity. These patterns however were not consistent between species, although it was clear that for human and baboon, the sequences composed of the exon-intron-exon constructs from 17,18; 19,20; 21,22; 23,24; and 27,28 were most similar. From the clustergram of mouse constructs, sequences composed of the exon-intron-exon constructs from 15,16; 17,18; 23,24; 25,26; and 27,28 were most similar. Given the discrepancy in the sequence relationships between these species, the exact sequence of duplication events remain unclear. For the frog eln1 sequence, no detectable sequence relationships were observed. However, as noted above, from the clustergram of frog eln1 (Figure 2-15), the duplicated region in this sequence appeared to be closer to the C-terminus
(exons 35-51). For the chicken and frog eln1 sequences, again no clear pattern of sequence relationships was observed. However, for the two zebrafish elastins, the sequences composed of the exon-intron-exon constructs from 21,22; 23,24; 25,26; and 27,28 were found to cluster together. The lack of consistency with clustergrams from other species suggests that these exons may have recently evolved in this lineage from a pair of ancestral exons, the most likely being exons 15 and 16 since these were found to share the least similarity with any other pair.
Figure 2-16 Clustergram analysis of pairs of exons and their adjoining intron. For each pair of exons and their adjoining introns, global alignments were constructed on a species-specific basis using the Needleman-Wunsch algorithm. The scores from the alignments were extracted and used to generate sequence similarity profiles for each exon-intron-exon set. The exons and introns are labeled as ‘e’ and ‘i’ respectively, such that e15_i15_e16, represents the sequence spanning exons 15 and 16 and their adjoining intron. Note that clustering based on an alternative arrangement of exons and introns (e.g. e15_i15_e16) did not result in the formation of any meaningful clusters, suggesting that the unit of evolution involves the pairs of exons presented here.
2.4 Discussion

To perform its function, the sequence of elastin must be such that it is: a) able to self-assemble; b) able to recoil after extension; and c) robust against failure to repeated cycles of extension and recoil. However it is still unclear as to how the sequence of elastin is able to impart these unique properties. Furthermore, given the variety of contexts in which elastin is found (from aorta and skin in mammals to the swim bladder in fish (Chung, Miao et al. 2006)) – elastins in these different environments have likely adopted species or taxon specific sequence properties that allow it to fulfill specialized roles. The recent completion of genome sequence data from several species of vertebrates should in theory facilitate comparative analyses aimed at identifying these sequence properties. However, due to the considerable effort involved, sequences of low complexity such as elastin are prone to errors during automated annotation procedures. Consequently, more focused efforts were required to collate and accurately annotate such sequences, and ElastoDB was an important first step in that direction. Exploiting this resource, it has been possible to perform a systematic analysis of the sequence features that likely underlie taxon and species-specific functional differences associated with the self-assembly and elastomeric properties of elastin.

Due to their highly variable nature, uncovering evolutionary and functional relationships in elastomeric proteins such as elastin, spider silks and resilin, represents a significant computational challenge. Given the recent availability of an increasing number of genomes, it might be expected that numerous examples of such sequences would be readily available to drive
such studies (Bernal, Ear et al. 2001). However, the low complexity associated with elastomeric proteins often results in their mis-annotation or even exclusion from the final genome assembly. As a result, elastomeric sequences retrieved from various otherwise reliable sources are often incomplete. Furthermore, where examples of elastomeric proteins have been identified, comparative studies featuring traditional sequence-based approaches have been found to lack the sensitivity required for analysing such low complexity sequence.

For example, phylogenetic tools exploring patterns of evolution require the availability of sequence alignments with at least a moderate degree of conservation, while motif finding tools such as MEME (Bailey and Elkan 1995) and TEIRESIAS (Rigoutsos and Floratos 1998) fail to identify functionally important patterns of sequence. Consequently, previous attempts to understand functional and evolutionary relationships in elastin have required the adoption of alternative methods based on, for example, BLAST and dotplots (Kummerfeld, Weiss et al. 2003; Chung, Miao et al. 2006).

To perform a systematic species-wide analysis of elastin sequences, a novel graph-based approach was developed to identify functionally relevant sequence motifs based on their degree of over-representation. Using methods borrowed from protein-protein interaction studies, full-length elastin sequences from 11 species were compared, resulting in the identification of several important repetitive elements in the hydrophobic and cross-linking regions of elastin (He, Chung et al. 2007). These domains were subsequently used as starting points to propose potential sites
of mutation for the elucidation of sequence-structure-function relationships of elastin (Chapter 3).

2.4.1 Elastin motifs

It has previously been noted that elastin has an unusual composition of amino acids, rich in proline, glycine and valine, that has been correlated with an increase in disordered structure (Rauscher, et al., 2006). Furthermore a number of studies have attempted to identify sequence motifs in elastin that represent structurally distinct functional elements. For example it has been proposed that the VPGVG motif in human elastin imparts assembly and elastomeric properties (Pepe, et al., 2005; Urry, et al., 1989; Yao and Hong, 2004). Other studies on elastin-like peptides containing PGVGVA or PGVGV motifs, have also shown the capacity for self-assembly into aggregates with similar elastomeric properties to fibrils assembled from full length elastin (Keeley, et al., 2002; Miao, et al., 2003). Furthermore, mutations of PGVGVA to GGVGVA were found to result in the formation of abnormal (amyloid-like) aggregates. While these studies demonstrate the functional and structural importance of a limited number of motifs it was still unclear as to how the sequence of elastin is able to impart its unique properties of assembly and elastic recoil. Given the variety of contexts in which elastin is found (from aorta and skin in mammals to the swim bladder in fish (Chung, et al., 2006)) – elastins in these different environments have likely adopted species or taxon-specific sequence properties that allow it to fulfill specialized roles. For example, the ability of elastin to self-assemble and function as an elastomer is known to be temperature dependent (Miao, et al., 2003; Sage, 1982).
For human elastin, coacervation occurs only above 30°C (Bressan, et al., 1986). Differences observed between the elastins of teleosts and mammals may therefore be responsible for the former’s ability to assemble and function as an elastomer under poikilothermic (cold-blooded) as opposed to homeothermic (warm-blooded) conditions (Miao, et al., 2003; Sage, 1982).

This stage of my thesis project has required the development of an innovative method to identify motifs through the identification and visualization of subsequence relationships. By focusing on the discovery of subsequence relationships between pairs of ‘fuzzy’ 3mers, this method was applied to both identify important 3mers that are involved in multiple motifs, and to infer structurally and/or functionally relevant motifs. Using my method, it was noted that the node, PG*, was highly connected in all studied elastin networks, highlighting the importance of this dipeptide within a broad range of structural and/or functional elements. Furthermore, by utilizing a network approach, it was possible to observe ‘neighbourhoods’ within the resultant graphs associated with specific functionalities (e.g. cross-linking as opposed to hydrophobic domains).

Across all elastin sequences examined, only a single motif was consistently found in relatively high abundance: G*V→PG*. This motif suggests a core functional module that imparts some of the basic elastomeric properties of elastin. Mapping this motif within individual elastin sequences from different species revealed differences in the way in which the motif is distributed (e.g. in frog eln1 and zebrafish eln2, this motif is found to be present in tandem repeats). In addition a number of motifs that are conserved across a more limited phylogenetic range were also identified, highlighting differences between chicken, mammals, frog and fish. For example
unlike the other elastins, zebrafish eln2 contains an abundance of motifs containing threonine. Application of sequence logos on one of these motifs expanded the identified region of conservation resulting in the generation of a motif consisting of GGLGTGGLP[GT]. Of note, this motif was found to be mainly localized to a region of the sequence which appears to have undergone a series of exon duplication events. This strongly implicated this motif with an important functional role. In frog eln1, sequence logos were also used to expand a motif identified through the network analyses. The resulting motif [YG][PA]GGKPPKPGYG formed an integral part of several cross-linking domains.

2.4.2 Subsequence network analysis of other low-complexity proteins

Applying SubSeqer to other elastic sequences such resilin, the HMW subunit of wheat gluten, and flagelliform spider silk has demonstrated its utility in motif finding. Using a 575 amino acid resilin protein from *Drosophila melanogaster* (Accession: NP_995860.1), with a subsequence length of four with one wildcard character and a percentile odds score value of two, it was found that this elastic insect protein had two highly repetitive regions (Figure 2-17). The first repetitive region was composed of repeats with a GGRPS[SS|DT]VGAPG*G*G motif, and the second region was composed of GYSGGRPGGQDLG repeats (shown in Figure 2-17b as a sequence logo built from the nodes *GGR and PGG*). Resilin is an elastic protein found in insect cuticles, providing energy storage and vibration dampening. Of all the other elastic proteins resilin has the most similar physical properties to elastin. These motifs correspond with previous
characterization of the resilin protein (Tatham and Shewry 2002) and showcase the effectiveness of the SubSeqer software.
Figure 2-17 Subseqer result for resilin. Results were generated using a 575 amino acid resilin protein from *Drosophila melanogaster* (Accession: NP_995860.1), with a subsequence length of four with one wildcard character and a percentile odds score value of two.
Unlike resilin, elastin, and spider silk, which have evolved to provide elasticity, gluten proteins, whose primary role is energy storage in germinating cells, appear to have elastomeric properties as a coincidental result of their sequence. Applying the Subseqer algorithm to the 661 amino acid HMW subunit from *Aegilops searsii* (Accession: AAT38813.1), it was determined that this protein was composed primarily of a 14 amino acid motif QQPGQGQQG[YH]YP[TA]S.

Flagelliform spider silk forms the spirals of spider webs that are used to capture prey. Unlike the dragline spider silks which have great tensile strength, flagelliform silks are more elastic. For a 907 amino acid flagelliform protein from *Nephila clavipes* (Accession: AAC38847.1) SubSeqer found that the protein is made up of long stretches of GPGGS, GPGGY, GPGGA, and GPGGV repeats. Interestingly, in the three other elastic proteins examined, motifs containing the PG dipeptide often follow the form PG*G.

### 2.4.3 Domain evolution

In addition to the identification of sequence motifs, the availability of many examples of elastin provided an opportunity to gain insights into how the protein may have evolved. Previous comparative analyses of the exon structure of elastin have revealed a conserved pattern of alternating hydrophobic and cross-linking domains (Figure 1-2) (Bressan, Argos et al. 1987; Kummerfeld, Weiss et al. 2003; Piontkivska 2004; Chung, Miao et al. 2006). As previously reported the sequences derived from the mammals and chicken shared similar numbers of exons.
(34-36). However the frog and zebrafish sequences possess an increased number of exons (55-58) resulting in a larger gene product. Interestingly, fugu elastin was only found to possess 37 exons. It has further been suggested that exons 8-14 in mammals and chicken and their corresponding regions in frog and zebrafish eln1 (exons 9-15) and zebrafish eln2 (exons 12-17) may form an ancestral core of exons and that exons 16-32 may represent an expanded region derived from a series of duplication events (Bressan, Argos et al. 1987; Kummerfeld, Weiss et al. 2003). Interestingly Figure 2-13 shows that the fugu elastin sequence exhibited a lack of alternating exons within domains 15-22, which together with the larger size of these exons and the presence of cross-linking motifs, suggested that they may represent an equivalent region of exon duplication in which the intron between the ancestral hydrophobic/cross-linking exons had been lost. The genome size of fugu is known to have undergone compaction through the loss of repetitive elements and a shortening of introns (Aparicio, Chapman et al. 2002). The loss of introns in fugu elastin may therefore have occurred during this compaction process. However, it should be noted that, like all teleosts, fugu has two copies of the elastin gene, and only one of these appears to have undergone compaction (Keeley, personal communication).

Within mammals and chicken, global clustering of exons on the basis of sequence similarity profiles found two distinct groups: the first comprising cross-linking domains derived from odd numbered exons from 15-29, the second comprising hydrophobic domains derived from even numbered exons from 16-30. Equivalent clusters of cross-linking and hydrophobic domain pairs were observed for exons 20-31 in both copies of zebrafish elastin, and for exons 35-51 in frog eln1. Unlike the other species, only a single cluster was observed in fugu, comprising exons 15-20.
A recent investigation of proteins rich in repetitive elements revealed that, like elastin, the majority of such proteins exhibit a striking conservation of intron position and phase suggesting a simple model of repeat protein formation through local duplications (Eisenberg, Nelson et al. 2006; Street, Rose et al. 2006). Furthermore such duplication events appear to be facilitated by neighbouring introns. Together the groups of exons identified above suggest regions in the elastin sequences which have undergone more recent duplication events compared with the rest of the sequence. For all species except fugu, the fundamental unit of duplication appears to be a pair of cross-linking and hydrophobic domains. In fugu, one possible scenario is that the ancestral pair of exons may have lost their adjoining intron, effectively merging a cross-linking domain with a hydrophobic domain. Subsequent evolution of the elastin gene in this species may thus have involved the duplication of only this single merged exon. An alternative explanation would require that following the duplication of an initial pair of cross-linking and hydrophobic domains, that six independent intronic loss events occurred.

From the results of the clustering of exon-intron-exon constructs, few consistent patterns of duplication events were observed. For the mammals and chicken sequences, assuming that the events occurred prior to their speciation, this lack of consistency likely demonstrates the significant divergence of these exon-intron-exon regions. Although the baboon and human constructs demonstrated similar patterns of clustering, given the discrepancy with the mouse constructs, at this stage it is not possible to speculate on the possible sequence of duplication events in these species. This also applies to the frog and fish sequences. To help reveal the nature
of evolutionary events between these exons the application of algorithms that specifically identify and consider the different evolutionary rates of these sequences between organisms may need to be adopted.

In all, this stage of my thesis project has served as an initial investigation into the genomics of the elastin protein, and has been successful in identifying and elaborating on additional key motifs in the elastin protein as well as regions of the elastin gene which appears to have undergone common expansion in different organisms. The understanding gained in this stage of my project was useful in guiding the choice of particular regions to study in detail using recombinant elastin-like polypeptides. In addition, the approach utilized here for the study of elastin has been demonstrated to be useful in functional investigations of other proteins with sequences of low complexity, such as spider silks.
Chapter 3 Effect of Mutations in repeats

Attributions: Bright field microscopy was performed with the help of Lisa Muiznieks. Mechanical testing of membranes was performed by Eva Sitarz. SEM imaging was performed by Robert Temkin (Advanced Bioimaging Centre, Mount Sinai Hospital). Amino acid analysis of protein concentration was performed by Rey Interior (Hospital for Sick Children).

3 Effect of Mutations in Repetitive Elements on the Physical Properties of Elastin-like Polypeptides

3.1 Introduction

Over the course of a human lifetime, in which a typical artery must undergo over a billion stretch and recoil cycles, the elastin present in elastic fibres formed during the neonatal development phase are not replenished (Visconti, Barth et al. 2003). Changes in the elastin gene which can cause substitutions in the amino acid sequence of elastin may lead to cumulative effects which can render elastic tissues more prone to premature failure, leading to late onset cardiovascular diseases. To explore this hypothesis, and to examine the potential impact of sequence mutations on the functional and mechanical behaviour of elastin, I undertook a survey of publicly available human single nucleotide polymorphisms (SNPs) to reveal a number of non-synonymous mutation sites in the repetitive elements identified in the previous chapter. Biochemical
investigations were performed on mutations derived from these SNPs in elastin-like polypeptides modeled after the human elastin sequence.

Analysis of the elastin sequence from ten species covering a broad taxonomic range revealed a number of repeat elements which were well-conserved and well-represented (He, Chung et al. 2007), suggesting that they may play an important functional role in the behavior of tropoelastin and ultimately in the elastic fibres they form. If changes in the tropoelastin sequence were to cause aberrations in the formation or elastic behavior of elastic fibres, mutations in these well-conserved repeat elements would likely have the greatest effect on the functional and mechanical behaviour of the elastin protein.

As this project is a stepping stone to the association of elastin mutations in disease populations to functional changes in the elastin gene, these publicly available SNPs provided an excellent opportunity to study the effect that mutations in repetitive elements can have on the functional and behavioral properties of elastin prior to engaging in expensive and time-consuming sequencing of specific disease populations. The methods and approaches adopted for the study of these publicly available SNPs should serve as an invaluable starting point when re-sequencing and SNP detection in disease populations is completed.

The tropoelastin protein is quite large at approximately 65kDa, and the effect of a single amino acid change in such a large, disordered protein may be too small to detect within the timeframe
of a typical PhD project. For this reason, shorter polymers based on a representative region of human tropoelastin were used as a platform for the mutation studies. Previous studies using such elastin-like polypeptides (EPs) have shown that they coacervate and undergo cross-linking in a manner similar to that of the full length protein (Bellingham, Woodhouse et al. 2001; Bellingham, Lillie et al. 2003; Miao, Bellingham et al. 2003; Miao, Cirulis et al. 2005).
3.2 Materials and Methods

3.2.1 Identification of Elastin Polymorphisms

Publicly available SNPs were acquired from two main sources. The first was the NCBI SNP database dbSNP ([http://www.ncbi.nlm.nih.gov/projects/SNP/](http://www.ncbi.nlm.nih.gov/projects/SNP/)) (Sherry, Ward et al. 2001), and the second was a series of SNPs derived from publicly available elastin expressed sequence tags (ESTs) from the NCBI’s dbEST ([http://www.ncbi.nlm.nih.gov/dbEST/](http://www.ncbi.nlm.nih.gov/dbEST/)) (Boguski, Lowe et al. 1993). SNPs were retrieved from dbSNP by performing a search for the human elastin gene (ELN).

The identification of SNPs using existing EST datasets began with the collection of elastin ESTs using a BLAST search of the coding region of the human elastin gene (gi|152112965) against all publicly available human ESTs. In all, 500 ESTs were obtained, most of which were associated with the 5’ end of the coding region covering exons 1 to 15. ESTs were assembled into a consensus sequence using the software package phrap ([http://www.phrap.com/](http://www.phrap.com/)) (Green 1999). The resulting consensus sequence was 2041bp long. A dot plot comparison of this consensus sequence and the original reference elastin CDS showed that these two sequences were in 100% agreement up till the 756th base pair corresponding to the region between exons 1 and 14. The absence of an abundant collection of ESTs outside the 5’ region of this gene limited the accurate identification of SNPs using this method to the first 14 exons.
Each EST which aligned with the consensus was analyzed for the presence of potential SNPs by using the SEAN software package (Huntley, Baldo et al. 2006). SEAN takes output from a Phrap assembly and compares each base pair in every aligned EST to the consensus sequences. Base pairs which are a mismatch to consensus, but which reside in a window (by default this is 15 base pairs) of base pairs with 100% match to consensus are selected as preliminary potential SNPs. At least two ESTs must have the same base pair mismatch to be considered a presentable potential SNP.

3.2.2 EP Construct Preparation

All constructs were based on the previously characterized EP2 polypeptide (Miao, Bellingham et al. 2003) which consists of exons 20-21-23-24-21-23-24 of the elastin gene. EP2 does not have exon 22 as this exon is found to be absent from human tropoelastin cDNA. The following primers were used for the construction of the EP2 based constructs:

Primer 1: Ex-20-1G1S-BamHI
CGCGGATCCATGTTTCCCAGGCTTTGTTGTCGGAGTCGGAGGTATCCCTGGAGTCGCA
GGTGTCCCTAGTGGGTAGGGTCG

Primer 2: Ex-20-3G3S-BamHI
CGCGGATCCATGTTTCCCAGGCTTTGTTGTCGGAGTCGGAGGTATCCCTGGAGTCGCA
GGTGTCCCTAGTGGGTAGGGTCG

Primer 3: Ex-23-1K1R-PpuM
AGTGGGGACCCACAGCTGCAGCAGCTAGAGCAGGCCAGCCAAAGCC

Primer 4: Ex-21 Apa
CGATTTGGCAGCGATCGCTAGCTAGCAGGCCAGCAGCTG

Primer 5: Ex-24 EcoRI
CTGCCTAGGGGATTCTAAGGGGCAATCGGAGG

Primer 6: Ex-24 Apa
GCTTCGGGCAATCGGAGGAC

Primer 7: Ex-20 BamHI
CTGCTAGGGGGATCCATGTTTCCCGGCTT

Primer oligonucleotides were produced by Integrated DNA Technologies (Coralville, Iowa).

3.2.2.1 EP2-1G1S and EP2-3G3S

EP2-1G1S and EP2-3G3S were created using PCR amplification of an EP2 template using primer 1 and primer 2 respectively upstream, and primer 5 downstream Figure 3-1a. PCR reagents were from the Qiagen Hotstar Hi Fidelity Polymerase Kit. The resulting products were cut with the restriction enzymes BamHI and EcoRI at exons 20 and 24. As the EP2 construct contains two exon 24s, and consequently 2 EcoRI restriction sites, this digestion led to the production of a shorter byproduct containing only exons 20-21-23-24. Gel electrophoresis of the digested products allowed the retrieval and purification of the larger product using Qiagen QIAEX II beads.
3.2.2.2 EP2-1K1R and EP2-2K2R

EP2-1K1R and EP2-2K2R construction required a two step process because the lysine to arginine substitution desired was located 181 bases downstream from the start of the sequence, outside the range of a the longest available primer Figure 3-1b. The first step of this process introduced the lysine to arginine mutation into an EP1 construct which consists of exons 20-21-23-24. This was accomplished using an EP1 template, primer 3, and primer 5, which produced an intermediate construct consisting of exons 23-24 with a PpuMI site at the 3’ end of exon 23 and an EcoRI site at the 5’ end of exon 24 along with the lysine to arginine mutation in exon 23. A regular EP1 construct was then digested with BamHI and PpuMI to produce an exon 20-21 intermediate, which was ligated to the exon 23-24 intermediate to produce a EP1-1K1R construct. To generate EP2-1K1R, an unmodified EP1 template underwent PCR reactions using primer 7 and primer 6 to introduce a BamHI and an ApaI digest site at the 3’ and 5’ ends respectively. EP1-1K1R was then used as a template in a PCR reaction using primer 4 and primer 6, resulting in the creation of an exon 21-23-24 intermediate with an ApaI site in the 3’ end of exon 21 and a BamHI site in the 5’ end of exon 24. This intermediate was digested with ApaI and EcoRI and ligated with the previous EP1 intermediate to create an EP2-1K1R construct. The EP2-2K2R construct was created using a nearly identical method, with the exception that an EP1-1K1R template rather than a regular EP1 template was used to create the first intermediate.
**Figure 3-1 Schematic of construct preparation.** EP2-1G1S and EP2-3G3 constructs (a) were produced using two different primers in a single step process. EP2-1K1R and EP2-2K2R constructs (b) were produced in a two stage process.
3.2.2.3 Transformation and sequencing

All constructs were ligated into pGEX-2T GST gene fusion vectors, which contains a LAC operon allowing induction by isopropyl-β-D-thiogalactopyranoside (IPTG). Vectors with inserts were then transformed into Invitrogen DH5α (Library efficiency) cells for amplification using a standard heat shock protocol. Vectors were recovered using standard miniprep protocol, and were digested with BamHI and EcoRI, whereupon DNA gel electrophoresis confirmed the presence and size of the inserts. Vectors with inserts were then sent to the Centre of Applied Genomics at the Hospital for Sick Children and sequenced to confirm the presence of the desired mutations.

3.2.3 Protein expression and purification

Constructs were transformed into BL21 bacteria by electroporation for protein expression and grown on plates with ampicillin to select for presence of the pGEX-2T vector. A single colony was picked and inoculated in 2x YT with ampicillin (50µg/ml) at 37ºC overnight, and reamplified the following morning in five 1.2L liquid media (2X YT and 2% glucose with ampicillin) for 3.5 hours until an absorbance of A₆₀₀. Expression was induced using 1.2mL of 0.3M IPTG per 1.2L of media, followed by incubation for 4 hours. Cells were spun down and transferred to six 50mL tubes, and treated with 25mL of 70% formic acid and 1.4g of CNBr per tube overnight. The CNBr treatment takes advantage of the lack of methionines in the constructs, allowing other proteins to be cleaved while preserving the desired proteins. Treated and lysed
bacterial solutions were dialysized against water, using Pierce dialysis tubing which allows the passage of molecules smaller than 3500 Da. After a 24 hour dialysis against water, cellular debris was discarded by centrifugation for 10 minutes at 12000 rpm.

Supernatant with elastin-like polypeptides underwent purification using a Sepharose SP (Amersham Biosciences) ion-exchange column eluted with 10% NaCl in 20mM sodium acetate. The resulting solutions were desalted using Pierce 3500Da dialysis tubing against water. Further purification was performed by reverse-phase high pressure liquid chromatography using a Jupiter 10µm C4 300-Å column (Phenomenex, Torrance, CA). Q-TOF mass spectrometry, performed by the Advanced Protein Technology Centre (Hospital for Sick Children), confirmed that the molecular weights of the polypeptides were consistent with those predicted (Figure 3-2).

3.2.4 Circular Dichroism (CD) Spectroscopy

Polypeptides were prepared for CD spectrometry analysis by being dissolved in water at a concentration of 15µM, as determined by amino acid analysis (Rey Interior, Hospital for Sick Children), and stored overnight. CD spectra were obtained at 20ºC using an AVIV 62DS spectrometer.
Figure 3-2 Q-TOF mass spectrometry of EP2-based polypeptides showing the appropriate introduction of mutations. Predicted molecular weights are shown in black and actual molecular weights as determined by Q-TOF mass spectrometry are shown in red.
3.2.5 Coacervation behaviour

Coacervation experiments on all EP2-based polypeptides were performed using a Shimadzu UV-2401PC UV spectrophotometer (Mandel Scientific, Guelph, ON), together with a temperature controller (Varian, Victoria, Australia), and a stir bar controller. Polypeptide concentrations for coacervation were determined by amino acid analysis (Rey Interior, HSC Biotechnology Service Centre), using norleucine as an internal standard. Polypeptides were dissolved in 50 mM Tris, 1.5 M NaCl at pH 7.5 to a concentration of 25 µM. Solutions were placed into a quartz cuvette along with a stir bar, and inserted into the sample cell of the spectrophotometer. Samples were left to equilibrate at room temperature for 5 min. The temperature was then increased at a rate of 1°C per minute with gentle stirring. Absorbance was monitored at 440 nm, and coacervation was detected as the onset of turbidity appearing as a rapid increase in absorption. The coacervation temperature was recorded as the intersection between the tangent prior to absorbance increase and the tangent of the curve during the increase.

3.2.6 Microscopic visualization of coacervation

Images of polypeptide coacervation were obtained in real time using an Axiovert 200 epifluorescence microscope (Zeiss, Toronto, ON) under strict temperature control using an Attofluor cell chamber (Molecular Probes, Eugene, OR). Focus was set to the bottom of the chamber prior to addition of polypeptides, which were dissolved in 50mM Tirs, 1.5M NaCl, pH 7.5 to a concentration of 50µM. Polypeptides were 95% pure as determined by amino acid
analysis (Rey Interior, HSC Biotechnology Service Centre). The cell chamber was preheated to ~5°C above the coacervation temperature of the polypeptide. 1.3mL of polypeptide in buffer was then added to the chamber, and sealed with a glass cover slip to prevent evaporation. Coacervation began as soon as the sample was introduced, beginning at time zero. Bright field images were taken at regular intervals over a 2 hour period for all polypeptides.

3.2.7 Elastic membrane construction

5mg of 85% pure polypeptide elastin-like polypeptides were lyophilized and then dissolved in 370µL of 0.1M sodium borate, pH 8. This solution was allowed to incubate overnight at 4°C, before being transferred to flat-bottomed glass-containers. 80µL of 5M NaCl and 50µL of the chemical cross-linker genipin was then added to the solution to initiate coacervation and cross-linking. Genipin is an extract from the fruits of *Gardenia jasminoides*, and has been shown to react spontaneously with free amino groups such as those found in lysine, hydroxylysine, or arginine residues, forming dark blue pigments (Sung, Huang et al. 1999). The precise reaction mechanism of amino acids with genipin is still unclear, but genipin has been used as a non-toxic cross-linker for many years. Following the addition of genipin, the colloidal suspension was spun at 3200rpm for 2.5 minutes, and allowed to rest at 37°C overnight, resulting in the collection of a thin membrane of polypeptides at the bottom of the container. These membranes were removed and stored in water at 4°C until used for mechanical testing.
3.2.8 Mechanical testing of membranes

Membranes were cut in half and mounted on steel grips prior to submersion in a water bath. The grips were then installed on a Biosyntech Mach-1 test apparatus (Montreal, QC), with a custom-made strain gage transducer. The apparatus extended and relaxed membranes a rate of 6.2 mm per minute. Each set of mechanical tests began with five loading and unloading cycles. Samples were then stretched to failure, after which membrane material was removed, dried, and underwent amino acid analysis.

Stress ($\sigma$) was calculated as a function of force (or load) measured in N over the cross-sectional area of the relaxed membrane ($A_o$). As the sheets were too thin to determine thickness values experimentally, load data was normalized by dividing the load by a measure of cross-sectional area:

$$A_o = \frac{(\text{dry weight/density of elastin})+(\text{dry weight}\times \text{water content/density of water})}{L_o}$$

where the dry weight is determined after weighing the membranes after mechanical testing, water content of elastin is 0.6 g/g, density of elastin is 1.3 g/mL, and $L_o$ is the original length of the membrane (determined using the Photoshop measuring tool on a photo of the membrane).

Strain ($\varepsilon$) was calculated as $\Delta L$ (the change in length of the membrane during stretching) over $L_o$, the original length of the membrane. Elastic modulus was calculated as the slope of the stress-strain curve using linear regression. Percent energy lost between loading and unloading, was calculated from the third stretch and recoil cycle and the following formula:
Percent energy loss

\[
\text{Percent energy loss} = \frac{\text{area under the loading curve} - \text{area under the unloading curve}}{\text{area under loading curve}} \times 100\%
\]

3.2.9 Scanning electron microscopy

Membrane material was collected after mechanical testing and brought to the Advanced Bioimaging Centre at Mount Sinai Hospital. The samples were first fixed in 2% glutaraldehyde in 0.1M sodium cacodylate buffer. They were then washed with distilled water and dehydrated through a graded series of ethanol before being critical point dried. Once dried, the samples were mounted on aluminum stubs and sputter coated with gold. Visualization was performed using an FEI XL30 environmental scanning electron microscope.

3.3 Results

3.3.1 SNP Identification

From the dbSNP database, a total of 110 SNPs were identified, of which 14 were located in the coding region of the tropoelastin gene. Of these, four caused synonymous mutations in the final gene product, while the other 10 caused non-synonymous mutations which resulted in a change
in the amino acid composition of the resulting protein product. Eight potential SNPs were identified through the EST method (Section 3.2.1), five of which led to non-synonymous mutations in the protein product. In total, using both data sources, 15 potential SNPs in the coding region of the elastin gene were identified which encode for non-synonymous mutations (Table 3-1).

Several of the SNPs caused non-synonymous mutations in important repetitive motifs of the elastin protein. SNP #4 caused an amino acid change at position 422 changing a VPGVG motif into a VPSVG motif; SNP #6 at amino acid position 503 changed a PGVGV motif to PGVGM; SNP #14 at amino acid position 145 changed a PGVGL motif to PGVEL; and SNP #15 at amino acid position 191 changed the motif PGVGP to PRVGP. All four SNPs caused non-synonymous mutations in or near the PGVG motif which had previously been found to be remarkably well conserved in elastins found in almost every organism for which elastin sequence was available (Section 2.4.1). For the three SNPs, causing substitutions in the PGVG motif, neither prolines nor valines were affected.
Table 3-1 SNPs causing non-synonymous mutations in the translated elastin protein. SNPs 1-10 were identified using dbSNP. SNPs 11-15 were identified using ESTs from dbEST. SNPs with available minor allele frequencies were sourced from the International HapMap Project.

<table>
<thead>
<tr>
<th>SNP</th>
<th>dbSNP ID</th>
<th>mRNA position</th>
<th>Nucleotide</th>
<th>Amino Acid position</th>
<th>Amino acid change</th>
<th>Exon number</th>
<th>Minor allele frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs41350445</td>
<td>303</td>
<td>C</td>
<td>T</td>
<td>Ala to Val</td>
<td>5</td>
<td>1.8%</td>
</tr>
<tr>
<td>2</td>
<td>rs41526244</td>
<td>998</td>
<td>G</td>
<td>A</td>
<td>Val to Ile</td>
<td>17</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>rs41376344</td>
<td>1037</td>
<td>G</td>
<td>A</td>
<td>Ala to Thr</td>
<td>17</td>
<td>0.3%</td>
</tr>
<tr>
<td>4</td>
<td>rs2071307</td>
<td>1370</td>
<td>G</td>
<td>A</td>
<td>Gly to Ser</td>
<td>20</td>
<td>25.3%</td>
</tr>
<tr>
<td>5</td>
<td>rs34945509</td>
<td>1494</td>
<td>A</td>
<td>G</td>
<td>Lys to Arg</td>
<td>23</td>
<td>0.8%</td>
</tr>
<tr>
<td>6</td>
<td>rs41523046</td>
<td>1613</td>
<td>G</td>
<td>A</td>
<td>Val to Met</td>
<td>24</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>rs34018370</td>
<td>1737</td>
<td>C</td>
<td>A</td>
<td>Ala to Asp</td>
<td>26</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>rs17855988</td>
<td>1847</td>
<td>G</td>
<td>C</td>
<td>Gly to Arg</td>
<td>26</td>
<td>4.1%</td>
</tr>
<tr>
<td>9</td>
<td>rs34852121</td>
<td>2057</td>
<td>G</td>
<td>A</td>
<td>Gly to Arg</td>
<td>30</td>
<td>N/A</td>
</tr>
<tr>
<td>10</td>
<td>rs41511151</td>
<td>2184</td>
<td>G</td>
<td>A</td>
<td>Gly to Asp</td>
<td>33</td>
<td>1.2%</td>
</tr>
<tr>
<td>11</td>
<td>N/A</td>
<td>47</td>
<td>T</td>
<td>C</td>
<td>Leu to Pro</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>12</td>
<td>N/A</td>
<td>58</td>
<td>A</td>
<td>G</td>
<td>Ile to Val</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>13</td>
<td>N/A</td>
<td>364</td>
<td>G</td>
<td>A</td>
<td>Gly to Arg</td>
<td>7</td>
<td>1.4%</td>
</tr>
<tr>
<td>14</td>
<td>N/A</td>
<td>435</td>
<td>G</td>
<td>A</td>
<td>Gly to Glu</td>
<td>9</td>
<td>0.5%</td>
</tr>
<tr>
<td>15</td>
<td>N/A</td>
<td>571</td>
<td>G</td>
<td>A</td>
<td>Gly to Arg</td>
<td>11</td>
<td>4.0%</td>
</tr>
</tbody>
</table>
3.3.2 SNP selection

Nearly half (7/15) of identified SNPs resulted in a glycine substitution. This was unsurprising given the abundance of glycine residues in the elastin protein. However, no SNPs resulted in proline mutations, despite a similar level of abundance of proline residues. This observation is consistent with existing hypothesis that the rigid bond angles of prolines are necessary to begin β-turns, structures vital to the coacervation and elasticity of elastin proteins. Indeed, previous studies substituting prolines with glycines in a EP consisting of exons 20-21-23-24-21-23-24 showed that the resulting polypeptides were incapable of coacervation, forming amyloid fibrils in the presence of salt (Miao, Bellingham et al. 2003).

In all SNPs which resulted in glycine mutations, the substituted residues were either serine, arginine, or glutamic acid, all polar amino acids. In the entire human elastin protein, there are a total of 2 glutamic acid residues—one each in cross-linking exons 19 and 21—but none can be found in hydrophobic domains. The mutation caused by SNP #14, replacing glycine with glutamic acid in hydrophobic exon 9 would introduce the only instance of a highly hydrophilic glutamic acid residue in a hydrophobic domain. This may pose a problem for the inter-digitation of hydrophobic side chains in the relaxed state of the protein, suggesting that SNP #14 may have a detrimental effect on the function of elastin. Arginine is another highly hydrophilic amino acid, and in human elastin it is found in the signal peptide in exon 1, cross-linking domain 10, and the hydrophilic C-terminus. Unlike glutamic acid, arginine is also found in hydrophobic domains 20, 26, and 33. However, with the exception of the arginines found in exon 1 and the C-terminus,
arginines in elastin are typically found directly adjacent to alanine residues and near the boundary between hydrophobic and cross-linking domains. It may be possible that the relatively large side chain of arginine leads to a preference for smaller neighbours such as alanine to allow more degrees of movement in a flexible protein such as elastin. If this were the case, the mutation caused by SNP #15 changing a PGVGP to PRVGP in exon 11 may be unfavourable and cause a negative effect on the structure and/or function of elastin.

SNP #6 causes the substitution of a valine residue with a methionine residue in amino acid position 503 in exon 24. Although, this substitution appears unremarkable at first because both amino acids are hydrophobic and have fairly non-reactive side chains, the fact that methionine is one of two conspicuously missing amino acids in human elastin makes this a notable SNP. The presence of a methionine residue in place of a valine residue at the middle of the elastin protein may affect the assembly of the protein or render it more susceptible to enzymatic degradation.

The glycine to serine substitution caused by SNP #4 changes a VPGVG motif to create a VPSVG motif. While the relatively small size of a serine residue might reasonably accommodate the traditional VPGV beta-turn, the introduction of a more reactive polar side-chain may disrupt the preferred confirmation in the affected region. More importantly, as shown in network analysis of elastin sequences (Section 2.4.1), the proline-glycine pair is highly conserved in elastins found in every species for which we have the full elastin sequence. This suggests that the PG di-peptide is important and mutations in this pair may lead to functional changes in the protein.
SNP #5 led to an amino acid swap at position 463 switching lysine with arginine. This turned the important KAAAK cross-linking motif in exon 23 into a RAAAK motif. Elastin cross-links between two tropoelastin monomers require the presence of a pair of lysine residues from each monomer to face towards each other. Three of these lysine side chains are oxidized by lysyl oxidase, and these along with an unaltered lysine residue form the desmosine cross-links which bind monomeric tropoelastin together (Cox, Starcher et al. 1974). This mutation in the only KAAAK motif found in exon 23 would functionally remove desmosine cross-linking properties in that exon. Since exon 24 is one of the largest hydrophobic exons in the elastin protein, the absence of cross-links in exon 23 would create a large hydrophobic region in the middle of the protein that cannot be stabilized via covalently bonds to other tropoelastin monomers, possibly creating less resilient elastic fibres.

While all of the above mentioned SNPs had the potential to be informative sites of mutation, SNPs #4 and #5 were chosen. The reasons for this choice were threefold: 1) the importance of the PG dipeptide and VPG tripeptide motif in the hydrophobic regions, and the well-conserved KAAAK motif in the cross-linking regions; and 2) the location of these mutations occur within an overall well-conserved region of elastin which has undergone suspected duplication; and 3) these mutations fell within exons 20-24 which are the basis of EP-2, an extremely well-studied elastin-like polypeptide containing exons 20-21-23-24-21-23-24, providing a well-characterized platform for mutagenesis studies.
Consultation with Sarah Rauscher suggested that based on molecular dynamic simulations of short repetitive sequences, the effect of substituting a single glycine residue with a serine residue would have small effects which may not be detectable. Based on these consultations, it was determined that rather than creating an EP2 construct with one substitution, three substitutions in exon 20 of EP2 would be introduced to maximize quantifiable differences. However, subsequent results showed a surprisingly drastic effect, leading to the creation of a EP2 with only one glycine to serine substitution at the exact location determined by SNP analysis. Two constructs were also created for the lysine to arginine substitution at exon 23, one with the substitution in the second exon 23 and another with substitutions in both exon 23s. In all, the four constructs were called EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R.

3.3.3 Circular Dichroism Spectra

Circular dichroism (CD) measures the difference in absorption of left-handed and right-handed polarized light at a given wavelength. A CD spectrum plots the average difference for a molecule over a range of wavelengths. For protein analysis, the wavelengths used fall into the "far-UV" region between 190 to 250nm. Distinctive CD spectra have been derived from polypeptides with pure alpha-helices, beta sheets, and random coil secondary structural conformations, and in theory the CD spectrum of a given protein is the sum of the percentages of each component secondary structure spectrum (Rodger and Nordén 1997). Circular Dichroism (CD) spectra have been successfully obtained for elastin polypeptides by several groups, and these have shown distinctive troughs at 220nm and 200nm representing largely random coil structure with some $\alpha$-
helical regions (Mammi, Gotte et al. 1968; Urry, Starcher et al. 1969; Tamburro, Guantieri et al. 1977).

CD spectra produced by EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R polypeptides are shown in Figure 3-3. All spectra produced by the four mutated constructs displayed the same broad negative peak at 200nm, and a less pronounced peak at 220nm. These peaks are shared by spectra produced by tropoelastin as well as EP2 (Debelle, Alix et al. 1995; Vrhovski, Jensen et al. 1997).
Figure 3-3 Circular dichroism spectra of EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R and EP2-2K2R in water with polypeptide concentrations of 10uM at room temperature.
3.3.4 Effect of mutations on the coacervation properties of EP2 constructs

3.3.4.1 Effect of mutations on coacervation temperature

Elastin self-assembly has been described as a two-step process (Cirulis and Keeley 2010), beginning with a reversible phase separation above a transition (coacervation) temperature. As elastin molecules aggregate, they begin to form droplets which result in a rise in turbidity of the aqueous elastin solution (Bellingham, Woodhouse et al. 2001). This stage is reversible, and if the mixture is cooled, the droplets will dissipate back into solution (Clarke, Arnspang et al. 2006). However, if this turbid mixture is held above the transition temperature for a long period of time, an irreversible maturation stage takes place. This stage is marked by the coalescence of elastin coacervates to form larger droplets. Light microscopy observations have demonstrated that the droplet sizes increase up to 500 times their original size through coalescence (Cirulis and Keeley 2010). These larger droplets are not as effective in light scattering, reducing the absorbance of the mixture, and causing a corresponding decrease in turbidity (Cirulis, Bellingham et al. 2008).

A summary of the coacervation temperature, molecular weight, and average Kyte-Doolittle hydropathy for each polypeptide is shown in Table 3-2. Glycine to serine mutations had no effect on coacervation temperature as compared with standard EP2. However, both constructs with lysine to arginine mutations showed a marked decrease in coacervation temperature from 29°C to 24°C, suggesting a higher propensity to coacervate. This change in coacervation temperature
did not appear to be correlated with the number of lysine to arginine mutations as both EP2-
1K1R and EP2-2K2R coacervated at the same temperature.

Table 3-2 The coacervation temperature, molecular weight, and average Kyte-Doolittle hydropathy values of EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, and EP2-3G3S polypeptides. Native EP2 data was collected by Juidth Cirulis.

<table>
<thead>
<tr>
<th>Polypeptide</th>
<th>Coacervation Temperature (°C)</th>
<th>Molecular Weight (kDa)</th>
<th>Average Kyte-Doolittle hydropathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP2</td>
<td>29°C</td>
<td>16992</td>
<td>0.923</td>
</tr>
<tr>
<td>EP2-1K1R</td>
<td>23°C</td>
<td>17020</td>
<td>0.920</td>
</tr>
<tr>
<td>EP2-2K2R</td>
<td>24°C</td>
<td>17048</td>
<td>0.917</td>
</tr>
<tr>
<td>EP2-1G1S</td>
<td>28°C</td>
<td>17022</td>
<td>0.921</td>
</tr>
<tr>
<td>EP2-3G3S</td>
<td>29°C</td>
<td>17082</td>
<td>0.917</td>
</tr>
</tbody>
</table>

Coacervation and maturation behavior as observed by absorbance at 440nm is shown in Figure 3-4. For each polypeptide, measurements were taken at least three times using identical batches of polypeptides (Section 3.2.5), however only representative curves are shown. By measuring the absorbance of polypeptide solutions at 440nm, the beginning of the coacervation is defined as the initial appearance of turbidity in the solution, while maturation is defined as the gradual decrease in turbidity. Using these definitions, the coacervation and maturation behavior of lysine to arginine mutations appear to be similar to that of regular EP2, with the exception that the mutated polypeptides produced higher maximum absorbencies. However, EP2-3G3S showed a
significantly slower rate of maturation, as seen by its nearly horizontal curve after coacervation.

This effect was significantly reduced when only one glycine to serine mutation was introduced, as the EP2-1G1S polypeptide did not share this maturation slowdown.
Figure 3-4 Coacervation behavior of EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R. Coacervate formation and maturation of above polypeptides was measured at an absorbance of 440 nm using a Shimadzu UV-2401PC UV spectrophotometer (Mandel Scientific, Guelph, ON), together with a temperature controller (Varian, Victoria, Australia), and a stir bar controller. Polypeptides were dissolved in 50 mM Tris, 1.5 M NaCl at pH 7.5 to a concentration of 25 µM. Samples were left to equilibrate at room temperature for 5 min. The temperature was then increased at a rate of 1°C per minute with gentle stirring, and coacervation was detected as the onset of turbidity appearing as a rapid increase in absorption.
3.3.4.2 Effect of mutations on coacervation process at microscopic level

Bright field microscopy images representing the observed coacervation process of all four polypeptides, along with an EP2 control, are shown in Figure 3-5. For all five polypeptides the start of coacervation is marked by the formation of small spherical droplets of approximately 2-3µm in diameter. However, by the 60 minute mark, the droplet morphology of the five polypeptides can be separated into two groups, the first composed of EP2, EP2-1G1S, and EP2-3G3S; and the second composed of EP2-1K1R and EP2-2K2R. Higher magnification bright field microscopy images taken at 60 minutes are shown in Figure 3-6. These images appear to indicate that the droplets formed by EP2, EP2-1G1S, and EP2-3G3S have a higher tendency to settle to the bottom of the glass, forming large, irregular shaped droplets. In contrast, droplets formed by EP2-1K1R and EP2-2K2R polypeptides appear to form more distinct and separate droplets which are less flat than the first group which have more three-dimensional appearances. For all polypeptides, droplets were observed to increase in size through coalescence between neighbouring droplets as previously described (Cirulis and Keeley 2010).
Figure 3-5 Bright field microscopy images of the coacervation and maturation process of EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R. Polypeptides, were dissolved in 50mM Tirs, 1.5M NaCl, pH 7.5 to a concentration of 50µM. Images were taken at 3 minute intervals (those shown were taken at 3, 30, and 60 minutes). White line indicates scale of 31µm.
Figure 3-6 Bright field microscopy images of the coacervation and maturation process of EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R at 60 minutes. Polypeptides, were dissolved in 50mM Tirs, 1.5M NaCl, pH 7.5 to a concentration of 50µM. White line indicates scale of 31µm.
3.3.5 Elastic membrane properties

The physical appearance of elastic membranes made from all four polypeptides as well as regular EP2 controls are shown in Figure 3-7. All the membranes displayed a deep dark colour as a side-effect of the cross-links formed by the chemical cross-linker genipin (Sung, Huang et al. 1999). Membranes made with EP2, EP2-1K1R, EP2-2K2R, and EP2-1G1S appeared to be visually identical with uniform material distribution and dark-blue pigmentation indicating successful cross-link formation by genipin. However, membranes made with EP2-3G3S exhibited a marked loss of structural integrity compared to regular EP2. EP2-3G3S membranes contained numerous holes, and readily disintegrated when placed in water. In fact, no EP2-3G3S membranes stayed intact long enough to be mounted to the mechanical testing apparatus.

All membranes, even those made with EP2 can tear during the mounting process. However, it was noted that during the mounting process, EP2-1K1R and EP2-2K2R membranes had a higher tendency to tear compared to regular EP2 membranes (Eva Sitarz, personal communication).
Figure 3-7 Physical appearance of membranes made from EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, and EP2-3G3S polypeptides. Membranes were made using 5mg of 85% pure polypeptide in 370μL of 0.1M sodium borate, 80μL of 5M NaCl and 50μL of the chemical cross-linker genipin. Solutions were spun at 3200rpm for 2.5 minutes, and allowed to rest at 37°C overnight, resulting in the collection of the thin membranes shown above. Note the extreme fragility of membranes formed by EP2-3G3S polypeptides.
3.3.6 Mechanical testing of membranes

With the exception of EP2-3G3S membranes which could not be tested due to a lack of structural integrity, all other membranes were successfully mounted to the apparatus and tested. The results of the mechanical tests are shown in Table 3-3. Despite a greater tendency to tear, EP2-1K1R and EP2-2K2R membranes, when successfully mounted to the apparatus, had similar mechanical and elastic properties to membranes made with standard EP2, with the notable exception that the membranes formed from material containing lysine to arginine mutations appear to result in reduced energy loss.

**Table 3-3 Results of mechanical testing on membranes formed by cross-linking EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, EP2-3G3S polypeptides with genipin.**

<table>
<thead>
<tr>
<th>Membrane Type</th>
<th>Max Strain (MPa)</th>
<th>Max Load (N)</th>
<th>Modulus (Pa)</th>
<th>% Energy Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP2</td>
<td>0.7±0.3</td>
<td>799.6±412.5</td>
<td>1154.0±399</td>
<td>19.1±4.3</td>
</tr>
<tr>
<td>EP2-1K1R</td>
<td>0.8±0.4</td>
<td>1133.8±586.0</td>
<td>1375.7±42.0</td>
<td>14.2</td>
</tr>
<tr>
<td>EP2-2K2R</td>
<td>0.9±0.4</td>
<td>1013.5±435.9</td>
<td>1032.9±139</td>
<td>13.8±1.9</td>
</tr>
<tr>
<td>EP2-1G1S</td>
<td>0.7±0.4</td>
<td>695.0±727.9</td>
<td>873.1±472.9</td>
<td>18.3±6.2</td>
</tr>
<tr>
<td>EP2-3G3S</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
3.3.7 Scanning electron microscopy

Figure 3-8 shows the appearance of the surface of membranes constructed from EP2, EP2-1G1S, EP2-3G3S, EP2-1K1R, and EP2-2K2R polypeptides under scanning electron microscopy. It is important to note that SEM visualization at lower depths show a solid surface, suggesting that the circular features shown in the figure represent droplets resting on the very top of the membrane surface. These droplets most likely have not been integrated into the cross-linked membrane, and offer a representation of average droplet sizes at the instant prior to cross-link formation and integration. In general, EP2 constructs tend to form droplet sizes 3.5µm on average. However, compared to other EP2 based constructs, EP2-3G3S membrane surface contained droplets which are approximately half as large, potentially correlating to the extreme structural weakness of EP2-3G3S membranes.

This observation further lends support to the proposed explanation for the slow rate of maturation exhibited by EP2-3G3S polypeptides during the coacervation process (Figure 3-4). If EP2-3G3S polypeptides do cause a slower decrease in absorbance because the size of their coacervate droplets remain small, then these small droplets may correlate with the reduced size of droplets seen on the surfaces of the resulting membranes. However, the exact relationship between the size of EP2-3G3S droplets and their fragility is still unknown. It is possible that the small droplet size is a byproduct of structural changes which can cause both a stabilization of droplet surface area and an impedance to proper cross-link formation using the chemical cross-linker genipin.
Figure 3-8 Scanning electron microscopy images of the surface of membranes made from EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, and EP2-3G3S. Membranes were first fixed in 2% glutaraldehyde in 0.1M sodium cacodylate buffer. They were then washed with distilled water and dehydrated through a graded series of ethanol before being critical point dried. Samples were mounted on aluminum stubs and sputter coated with gold. Visualization was performed using an FEI XL30 environmental scanning electron microscope. White bars indicate scale of 10µm.
3.4 Discussion

Based on a survey of existing SNPs in publicly available databases, two SNPs resulting in amino acid substitutions in well-conserved repetitive elements in the tropoelastin gene were chosen for experimental analysis. Using the well-characterized EP2 construct consisting of exons 20-21-23-24-21-23-24, lysine to arginine mutations were introduced in exon 23 and glycine to serine mutations in exon 20. These mutations resulted in changes to the KAAK and VPG motifs in the tropoelastin protein. Both of these motifs were found to be over-represented in elastin from ten different species (He, Chung et al. 2007), suggesting that they form the structures necessary for the structure and function of the protein and of elastic fibres. Together, four new constructs were created: EP2-1K1R, EP2-2K2R, EP2-1G1S, and EP2-3G3S.

Due to the relatively large size of the EP2 polypeptide (~17kDa), the mutations did not significantly change the molecular weight or average Kyte-Doolittle hydropathy of the resulting polypeptides (Table 3-2). Compared to previous studies involving elastin-like polypeptides with more significant differences in molecular weight and hydropathy (Miao, Bellingham et al. 2003; Miao, Cirulis et al. 2005), these differences are almost negligible. As a result, the possibility existed at the outset of the experiments that these mutations would result in no observable physical and behavioral changes in the EP2 polypeptide. Indeed, this possibility was the motivation for the creation of the EP2-3G3S and EP2-2K2R constructs, which would amplify potential observable differences.
Initial predictions were that the glycine to serine mutations would result in minimal changes in the self-assembly and mechanical properties of the resulting elastin-like polypeptide, while the lysine to arginine mutations, especially the 2K2R mutation, would have a drastic impact on the mechanical integrity of the resulting polypeptide. While CD spectra showed that both the glycine to serine and lysine to arginine mutations produced polypeptides with similar secondary structure distribution as EP2, their self-assembly and mechanical properties were noticeably different. However, contrary to initial predictions, it was the 3G3S mutation which produced the largest change in the mechanical properties of polypeptide. Lysine to arginine mutations, on the other hand, exhibited the greatest effect on the coacervation temperature of the polypeptide. In all, these findings suggest that the over-represented motifs found in the elastin protein do play an important role in its functional properties, and that changes in these regions may result in disruptions to proper function.

3.4.1 Effect of mutations on EP2 structure

Circular dichroism spectra have been used extensively on elastin solutions to examine the distribution of secondary structures within the polypeptide (Tamburro, Guantieri et al. 1977; Urry 1982; Debelle, Alix et al. 1995). The human isoform of tropoelastin is believed to adopt a largely disordered conformation, whose secondary structure consists of 3% α-helix, 41% β-sheet, 21% β-turn, and 33% undetermined structure (Vrhovski, Jensen et al. 1997), with the cross-linking regions forming α-helices and the hydrophobic regions forming β-sheets and β-turns. This distribution of secondary structures gives tropoelastin a distinct CD spectrum in the "far-
UV" region between 190 to 250nm where the backbone amide transitions of the polypeptide reside. This signature curve, which is shared by the elastin-like polypeptide EP2 (Miao, Bellingham et al. 2003) contain two signature troughs at 222nm and 200nm (Mammi, Gotte et al. 1968; Urry, Starcher et al. 1969; Tamburro, Guantieri et al. 1977). The negative peak at 222nm is representative of α-helical structures and is the result of the cross-linking regions in elastin which are α-helical in nature. The peak at 200nm, on the other hand, is representative of the presence of β-turns in the hydrophobic regions of the polypeptides.

A comparison of the CD spectra produced by EP2, EP2-1K1R, EP2-2K2R, EP2-1G1S, and EP2-3G3S (Figure 3-3) shows that the overall curve, as well as the two key peaks, are largely identical. This observation suggests that the global distribution of secondary structures has been unaffected by the mutations. Previous studies of mutated forms of the EP2 polypeptide which demonstrated significant changes in CD spectra required more drastic substitutions such as the replacement of all prolines with glycines in both exon 24s (Miao, Bellingham et al. 2003). In these studies, the CD spectrum displayed a large positive peak at 195 nm, suggesting the substitution of rigid prolines with more flexible glycines reduced β-turn formation, resulting in the presence of extended β-sheet structures. It is clear that the glycine to serine mutations did not result in the same drastic changes in secondary structure distribution. Incidentally, the proline to glycine mutants were also unable to coacervate in the classical elastin sense, instead forming amyloid-like precipitate. None of the mutants examined in this thesis project exhibited this amyloid-forming behaviour, furthering supporting the finding that glycine to serine and lysine to arginine mutations left the β-turn and α-helical content undisturbed.
However, it is important to note that analyses of CD spectra provide only limited resolution, describing secondary structure in a global and qualitative manner. It is, therefore, impossible to conclude from CD alone that the secondary structures of the mutant polypeptides, especially at a local level, have been completely unaffected.

Figure 3-9 Circular dichroism spectra EP2 (solid line), EP2 with P to G mutation in one exon 24 (dotted line), and EP2 with P to G mutation in both exon 24s (dashed line). The amyloid forming polypeptide (dashed line) has a CD spectrum that is distinctively different from that of EP2 and all EP2 mutants used in this thesis (Figure 3-3) Figure from (Miao, Bellingham et al. 2003), used with permission from publisher.
3.4.2 Effects of mutation on coacervation

Despite similarities in the secondary structure distribution of the polypeptides, observations of the temperature-induced self-assembly of these polypeptides show marked differences when compared to regular EP2. The lysine to arginine mutants exhibited an increased tendency to coacervate as reflected by a lower coacervation temperature, reducing the EP2 coacervation temperature of 29°C to 24°C for both EP2-1K1R and EP2-2K2R under identical polypeptide and ionic concentrations. EP2-3G3S displayed a markedly different behaviour post-coacervation in what has been termed the "maturation" stage of elastin self-assembly.

Previous studies of the coacervation and maturation properties of elastin-like polypeptides have shown that coacervation temperature can be affected by hydropathy, concentration, or molecular weight of the polypeptide as well as the ionic concentration of the solution (Miao, Bellingham et al. 2003). In the studies performed in this project, the concentration of the polypeptides used were fixed at 25µM, while the molecular weight and average hydropathy of the polypeptides are essentially identical, with very minor differences due to the substitution of one to three amino acids. In addition, salt concentration was also consistently fixed for all trials. While it is possible that the quantification of polypeptide concentration is not consistent due to error margins of amino acid analysis, the large temperature variations observed for the EP2 construct with lysine to arginine substitutions cannot be readily explained by concentration differences. Studies using EP2 polypeptides at concentrations of 25uM and 100uM showed much higher maximum absorbance differences than those observed between EP2-1K1R, EP2-2K2R and EP2 constructs.
It is therefore unlikely that concentration differences accounted for the change in coacervation temperature.

The drop in coacervation temperature of the lysine to arginine mutants was a surprising finding. Lysine and arginine are amino acids with similar hydropathy and polarity. Initially, constructs with lysine to arginine mutations were not predicted to have significant effects on the self-assembly properties of the resulting polypeptides because previous work had established that the propensity to coacervate was related to the hydrophobicity and molecular weight of proteins (Miao, Bellingham et al. 2003). The increase in coacervation propensity found in lysine to arginine mutants cannot be readily explained by structural changes in the lysine to arginine mutants. While the inversely linear relationship between coacervation temperature and secondary structure (namely the percentage of α-helical content) have been previously explored (Muiznieks, Jensen et al. 2003), a decrease of 5°C in the coacervation temperature would require a near doubling of α-helical content, an increase not supported by the CD spectra of these mutants.

Elastin coacervation is believed to be a thermodynamically favorable, endothermic, and entropy-driven process (Urry 1988; Li, Alonso et al. 2001; Li, Alonso et al. 2001). Individual elastin monomers have solvent-exposed regions which have limited ability to hydrogen-bond with the surrounding water molecules, and instead induce the formation of clathrate water structures. These clathrate structures prevent the interaction between independent elastin monomers, but become destabilized with an increase in temperature or ionic concentration, driving the process
of coacervation (Urry, Starcher et al. 1969; Urry, Long et al. 1974; Urry, Hayes et al. 1995).

Under this model, it is possible that the replacement of the amino group found in lysine with the guanidine group in arginine may introduce a destabilizing source to the clathrate water structure. In addition, as the lysine side-chains are oriented away from the axis of the α-helix for cross-linking purposes (Foster, Bruenger et al. 1976), these residues have an even more pronounced effect on the interactions of the polypeptides and nearby water molecules. It is possible that the substitution of lysines with arginines is responsible for destabilizing the clathrate water structures surrounding the EP2 polypeptide, leading to a higher propensity to coacervate.

The glycine to serine mutants exhibited a different self-assembly behavior than the lysine to arginine mutants. The mutant with the most dramatic change of all four mutants was the EP2-3G3S construct, whose velocity of maturation is more than an order of magnitude slower than standard EP2. The difference was so marked that its behavior was similar to experiments where stirring did not occur. The phenomenon of coacervation and self-assembly has recently been studied in detail (Cirulis and Keeley 2010). Interestingly it was proposed that the decrease in absorbance subsequent to coacervation is caused by an increase in droplet size due to coalescence, causing a decreased ability to diffract light. Using this proposed model as a reference, the coacervation behavior displayed by the EP2-3G3S suggest that EP2-3G3S polypeptides possess a stabilizing effect on the surface of the droplets of a certain size, allowing for the initial coalescence of the smaller "starter" droplets, but preventing further coalescence into larger droplets as seen with standard EP2. Bright field microscopy of the coacervation process does not show any appreciable difference between the droplet morphology of EP2 and EP2-3G3S or EP2-1G1S. However, the conditions for the coacervation process during the
imaging experiments are different from those of the absorbance experiments. In the imaging experiments, the solution was not stirred as visualization of a dynamic solution is not possible. Consequently, the settling of droplets to the bottom of the glass surface may result in different droplet behavior than those seen in the absorbance experiments. When cross-linked versions of EP2 and EP2-3G3S are observed under scanning electron microscopy, however, it is clear that the droplets on the surface of the EP2-3G3S are smaller than those of EP2. Whether this change in size is enough to alter the absorbance during the maturation process is unclear, but it appears from the coacervation curve that EP2-3G3S mutations do have an effect on the maturation process of EP2 polypeptides. Based on the model proposed previously, SEM images of the surface of cross-linked EP membranes lend further support to the idea that serine substitutions play a role in limiting the growth of coacervation droplets, likely caused by a combination of local structural changes caused by the introduction of bulkier serine side chains and changes in hydrophobicity due to the introduction of more polar side chains. Indeed studies of synthetic polymers made of VPGVG repeats and VPAGV repeats have shown that the introduction of a bulkier alanine sidechain causes changes in the kinetic and thermal properties of the polypeptides (Reguera, Lagarón et al. 2003).

As the EP2-1G1S polypeptides shared identical coacervation characteristics as regular EP2, it appears that the effects of these mutations are cumulative, and must meet a certain threshold level before exhibiting marked differences, suggesting that while isolated mutations can be tolerated, cumulative mutations can have deleterious effects on the self-assembly behavior of the protein.
3.4.3 Effect of mutations on mechanical properties

The traditional view of the cross-linking and hydrophobic domains was that the latter were responsible for the elastomeric properties of the protein while the former were responsible for its strength. As a result, it was not unreasonable to predict that the glycine to serine mutants would affect the elastomeric properties of the membranes as measured by resilience, while the lysine to arginine mutants would affect the mechanical strength of the membranes as measured by maximum stress and strain.

The results, however, differed from these predictions. The lysine to arginine mutants did not display any statistically significant reduction in mechanical strength, but did offer slightly better elastomeric properties than regular EP2. The single glycine to serine mutant behaved nearly identically as regular EP2, however the mutant with three such substitutions EP2-3G3S displayed such a drastic change in mechanical strength that it could not be successfully mounted to the testing apparatus without breaking apart.

One possible explanation for the surprising durability of the lysine to arginine mutant membranes may be genipin, the chemical cross-linker used to cross-link the polypeptides. This cross-linker is different than the \textit{in vivo} cross-linker LOX. Presently, the only active form of LOX used has been derived from yeast, and it has proven to be quite difficult to use \textit{in vitro}
(Mithieux, Wise et al. 2005). As a result, suggestions to use LOX in lieu of genipin cannot be readily accommodated. However, as genipin has been shown to preferentially cross-link lysines over arginines by nearly an order of magnitude, it is unlikely that genipin is also cross-linking the substituted arginine residues (Sung, Huang et al. 1998). Amino acid analysis of the membrane materials showed that while the lysine residues could no longer be detected because they have been bound to genipin, all the arginine residues were still present, further eliminating the possibility that the arginine residues could have been involved in cross-linking. Instead, the cause of a lack of deleterious effect may be the difference between the cross-linking mechanism of genipin and LOX. Unlike LOX, which deaminates two lysine side chains to form desmosines from four deaminated lysyl sidechains, genipin requires only one pair of lysines. While the exact chemical mechanism of genipin cross-linking is still unclear, it is possible that use of only one lysine residue per polypeptide chain has masked the cross-linking deficiency of the lysine to arginine mutants.

The dramatic effect observed in EP2-3G3S raises the interesting possibility that the hydrophobic regions also play a significant role in the durability of elastin-like membranes and perhaps even in elastic fibres. The dark colour of the EP2-3G3S membranes indicates that genipin has successfully formed cross-links (Sung, Huang et al. 1998). In addition, amino acid analysis of the membrane materials did not detect lysines which were bound by genipin. It may be possible that the change in self-assembly characteristics caused by the glycine to serine mutations as seen in the absorbance experiments and SEM images of the membrane surface could have altered the alignment of the cross-linking regions such that either intra-molecular cross-linking is favored in the EP2-3G3S mutants, or else the distances between lysine residues are increased leading to a
reduction in proper cross-link formation. Under these scenarios, the EP2-3G3S membranes would be dark in appearance like regular EP2 membranes, and have the same amino acid analysis result, while having completely different mechanical properties.

This project has shown that the roles of self-assembly, elasticity, and mechanical strength cannot be rigidly assigned to cross-linking or hydrophobic regions in elastin, as traditionally thought. Rather, all the regions are involved in providing the proper function to elastin-like polypeptides. While the effects seen in these studies have been accentuated due to the relatively small size of the EPs, they nevertheless suggest that mutations in the key repetitive regions of the elastin protein can have an effect on its mechanical and elastomeric behavior. Even though the mutation sites for this study have been chosen based on a small number of polymorphisms currently available through publicly accessible databases, the methods used have demonstrated the possibility of establishing links between elastin sequence and function. Future polymorphism studies specifically focused on the elastin genes of disease populations using next generation sequencing can provide more targets for experimental validation using techniques described in this project to provide new insights into the link between elastin mutations and late-onset diseases of the cardiovascular system.
Chapter 4 Summary and Future Directions

4 Summary of findings and Future Directions

4.1 Introduction

Elastin is a critical protein for the proper functioning of elastic tissues in a variety of organisms including humans. Elastic tissues are found in the lungs, skin, and in the cardiovascular system, where for example the thoracic aorta is nearly 50% elastin by dry weight (Uitto 1979). The formation of elastic fibres containing elastin, typically occurs only during the late fetal and early neonatal development phases, and little turnover of the protein is observed (Visconti, Barth et al. 2003). Since little elastin is produced after elastogenesis, genetic mutations which alter the sequence of the elastin protein may lead to subtle defects in elastin and elastic tissues can have significant consequences over the course of a human lifetime.

This thesis project has performed crucial initial tests of this hypothesis by determining regions of importance in the elastin sequence, followed by experimental examination of the effect of alterations in these regions. The results of these studies have shown that sequence alterations do indeed affect the self-assembly and mechanical properties of the elastin protein, and has laid the foundation for future studies on the relationship between elastin sequence and function.
4.1.1 Summary of Findings

This thesis project pursued two main avenues of investigation: 1) to establish a catalogue of potentially important sequence regions and motifs by comparing full-length elastin sequences from ten different species; 2) to determine the impact that mutations in these regions would have on the self-assembly and mechanical behavior elastin-like polypeptides. The studies resulted in the following findings:

1) Elastin sequences are rich in VPG motifs in the hydrophobic elastomeric regions, and KAAK or KAAAK in the cross-linking regions. In all full-length elastins to date, the G*VPG motif is highly abundant, but the GP di-peptide motif is not found even though elastin sequences are rich in prolines and glycines.

2) Phylogenic studies of elastin sequences indicated the strong possibility that the region between exons 15-30 underwent duplication.

3) Polymorphisms exist in the human population which alter the amino acid sequence in VPG and KAAK motifs in the region between exons 15-30/4

4) The introduction of three glycine to serine mutations corresponding to known human polymorphism in elastin-like polypeptides based on a region of human elastin consisting of exons 20-24 altered the self-assembly characteristics of the polypeptide and produced elastin-like biomaterials which were significantly more fragile than unaltered polypeptides.
5) The introduction of lysine to arginine mutations in the above elastin-like polypeptide did not significantly alter the mechanical properties of the resulting biomaterial, but reduced the coacervation temperature, suggesting an increase in the propensity to begin self-assembly.

These findings suggest that two important properties of elastin-like polypeptides—its ability to self assemble, and its ability to function as a durable elastomer—can indeed be altered by mutations in over-represented repetitive motifs. However, while these elastin-like polypeptides have been shown to behave similarly to full length tropoelastin ((Bellingham, Woodhouse et al. 2001; Bellingham, Lillie et al. 2003; Miao, Bellingham et al. 2003; Miao, Cirulis et al. 2005), they are not identical to human elastin. For example, the synthetic elastin-like polymers used were significantly shorter than human elastin, and as a result the changes observed in this thesis may have been amplified. This issue is currently being addressed in the Keeley laboratory as full length human elastin with the 3 glycine to 3 serine mutation in exon 20 have already been produced and purified, and coacervation and mechanical studies are currently pending.
4.2 The effects of elastin sequence variants in patients linked to late-onset cardiovascular disease on the structure and function of elastin

While the work in this thesis has shown that known SNPs can impact elastin function, the candidate mutations sites chosen for this thesis project were determined through a survey of publicly available human single nucleotide polymorphisms (SNPs), without specific targeting for sources with known histories of cardiovascular disease. To investigate the potential of elastin mutations to impact cardiovascular disease, it would be useful to explore a more medically relevant cohort for example individuals suffering damage to their elastin-containing cardiovascular tissues such as those with late onset cardiovascular complications. To that end, an extension of this project has begun to be compile of a list of SNPs found in elastin sequences from individuals with known cardiac disease.

Genetic analyses of the major elastin related diseases such as SVAS, Williams-Beuren syndrome, and cutis laxa have revealed instances where genetic mutations have caused haploinsufficiency of the elastin gene or an inability of the protein to integrate within the extracellular matrix (Milewicz, Mikulski et al. 2000). Variants in the elastin gene has been shown to be linked to an increased risk of isolated systolic hypertension (ISH) in the Chinese Han population (Deng, Huang et al. 2009). In addition, a number of genetic association studies have linked elastin to instances of intracranial aneurysms (IA) (Onda, Kasuya et al. 2001; Ruigrok, Seitz et al. 2004; Akagawa, Tajima et al. 2006). There has also been reports of a relationship between a glycine to serine mutation in the elastin protein and the distensibility of...
elastic arteries (Hanon, Luong et al. 2001). However, some previous genome wide association studies have shown a lack of clear links between elastin and cardiac disease (Kaushal, Woo et al. 2007). These results may be caused by the use of unsuitable cohorts or a lack of statistical power resulting from the limited number of single nucleotide polymorphisms (SNPs) analysed in these studies (Mineharu, Inoue et al. 2006). By focusing on a more mechanistic approach, focusing entirely on SNPs found in the elastin gene, it may be possible to increase the number of relevant SNPs, improving the resolving power of the analyses.

The goal of future work based on this thesis project will be to establish links between SNPs found in the elastin sequence of individuals suffering from cardiac disease with two main aims:

1. Identification of novel polymorphisms which are over-represented in the exonic regions of the full length human elastin gene, which may lead to amino acid substitutions in important repetitive elements; as well as polymorphisms in the intronic regions of the elastin gene which may affect proper post-translational processing of elastin mRNAs

2. Experimental and computational validation of the effects that such polymorphisms have on the structure and function of recombinant elastin or elastin-like polypeptides

4.2.1 Identification of novel polymorphisms which are over-represented in full length human elastin gene in patients linked with cardiac disease

Through a collaboration with Dr Dianna Milewicz (University of Texas, Houston Medical School), genomic samples have been obtained from a cohort of 800 individuals diagnosed with
thoracic aortic aneurysms and dissections (TAAD). Samples were pooled into five sets to reduce sequencing costs. Illumina’s Solexa sequencing provided by the Centre for Applied Genomics (TCAG – Hospital for Sick Children) has already successfully sequenced the full length elastin gene from these pools. Currently, Mahboubeh Ghoryshi, a PhD candidate in the Parkinson laboratory is assembling these sequence reads and identifying elastin sequence variants from this pool.

However, while the pooling strategy used in the sequencing stage will allow the detection of a large number of polymorphisms in a cost-effective manner, it cannot resolve individual genotypes. To overcome this shortcoming, the list of polymorphisms identified after sequencing will be validated using Taqman genotyping, a process whereby probes designed to anneal to the SNP-containing region added to individual samples to detect for the presence of the SNP. Genotyping will be done on the 800 disease samples as well as 400 control samples obtained from the Ontario population genomics repository (OPGP) at the Centre for Advanced Genomics (TCAG - Hospital for Sick Children). As this is a costly and time-consuming process, not every SNP identified after sequencing can be genotyped. Instead, the list of SNPs will be prioritize on the basis of functional relevance for Taqman based genotyping. Initial focus will be placed on variants resulting in non-synonymous polymorphisms and indels in exons and splice junctions. Of particular interest will be those variants which have the potential to alter cross-linking capabilities or disrupt mechanical or elastomeric capabilities, including some such as the glycine to serine mutations which has already been shown to have an impact on the functional and mechanical behaviour of elastin-like polypeptides in this thesis.
4.2.2 Experimental and computational validation of the effects that such polymorphisms have on the structure and function of recombinant elastin or elastin-like polypeptides

SNPs which are found in TAAD samples will be good candidates for further experimental and/or computational analyses to determine whether the mutations caused by these SNPs play a role in altering the structural, functional, or mechanical behaviour of elastin. Experimental validation utilizing protocols and techniques similar to those described in Chapter 3 will determine whether these SNPs have direct effects on the coacervation and mechanical behavior of elastin-like polypeptides. The small size of elastin-like polypeptides will amplify potential effects, allowing the concentration of more resource-intensive studies on full-length human elastin to be performed on SNP candidates which demonstrate effects on smaller elastin-like polypeptides. While the current stretch apparatus operates at a rate which cannot simulate in a laboratory setting the millions of stretch and recoil cycles experienced by elastin over a typical human lifetime, simulations of such long term effects can be made by increasing the temperature of the mechanical testing environment above physiological levels.

In addition to experimental studies of SNPs, computational approaches will also be applied to study the effects that such perturbations can have on elastin polypeptides. The Parkinson laboratory has been using a flexible modeling platform that can simulate the formation of molecular aggregates. These models have shown the effect that domain organization and other factors such as domain size and hydrophobicity have on elastin assembly. Other computational
work using molecular simulations of monomeric and aggregated states, have been able to
determine the assembly behaviours of a number of elastomeric proteins including elastin, spider
silks, wheat gluten, and insect resilin (Rauscher, Baud et al. 2006). As these models become
more robust, they can be used to study the potential molecular consequences of sequence
variations identified from the sequencing stage.

It is the long term goal of these future studies to be able to experimentally verify the effects that
sequence variants in the elastin gene may have \textit{in vivo}. Successful validation that elastin
mutations can cause or become a factor in increasing the likelihood of cardiovascular disease
will involve the introduction of mutant elastin genes in smaller mammalian organisms such as
mice, which have relatively similar elastin sequences as humans but with much shorter life spans
and higher rates of heart-beat, allowing the long term study of the effects of elastin mutations on
cardiovascular health.

These future studies may uncover the molecular basis for cardiovascular diseases associated with
elastin polymers. Such associations are difficult to discover using simple genome-wide
association studies because they may take a significant amount of time to be established. By
targeting for mutations in regions of the elastin protein which have a higher chance of disrupting
the proper function of the protein, the chance of detecting such associations should be greatly
improved.
References


