Investigating Intra- and Inter-Molecular Coevolution of Intrinsically Disordered Protein, Prothymosin-α

Brianna Biscardi

November 2015

Director of Thesis: Dr. Colin S. Burns

Major Department: Chemistry

Prothymosin-α (ProTα) is a small, highly acidic protein found in the nuclei of virtually all mammalian tissues. It belongs to a class of proteins known for their lack of a rigid three-dimensional structure called intrinsically disordered proteins (IDPs). ProTα has been shown to play essential roles in cell robustness. As an example, ProTα is involved in apoptosis, or programmed cell death by inhibiting apoptosome formation via binding Apaf1. This research focus is on detecting coevolution of ProTα and between ProTα and Apaf1 (ProTα-Apaf1 or ProTα-Apaf1 complex). Coevolution refers to correlated changes between pairs of interacting species to maintain or refine functional interaction. Coevolution can be defined at the molecular level as correlated sequence changes that occur to maintain a structural or functional interaction. Studying coevolution of ProTα and ProTα-Apaf1 may provide useful information such as structural contacts and specific residues necessary for complex formation.

In this study, a pipeline for performing molecular coevolution studies was established at East Carolina University (ECU). This pipeline was used to analyze myoglobin, ProTα, and ProTα-Apaf1. Myoglobin has been a target of previous

coevolutionary studies and was chosen to test the robustness of the pipeline developed in this study. Most of the coevolving residues that were found in myoglobin match closely with those detected in other work. ProTα, which has never been studied by way of coevolution, displays several coevolving residues involved in long range interactions or functionally important regions. These methods were also applied to ProTα- Apaf1 complex. Previous experimental studies using ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) NMR have revealed residues on ProTα necessary for interaction with Apaf1 however the residues on Apaf1 necessary for interaction with ProTα have not been resolved. Several residues of ProTα were found to have coevolution with Apaf1. Docking studies were performed to simulate binding between ProTα and Apaf1 at the sites detected in this study (ProTα: Thr8, Thr107; Apaf1: Ser1056, Asp1096). Six orientations of ProTα and Apaf1 were run for 9 nanoseconds (ns) and in each simulation, the two proteins did not drift apart from one another. This suggests that the residues detected by coevolution in this study may play a role in the interaction between ProTα and Apaf1.

Investigating Intra- and Inter-Molecular Coevolution of Intrinsically Disordered Protein, Prothymosin-α

A Thesis

Presented To the Faculty of the Department of Chemistry

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Chemistry

by

Brianna Biscardi

October, 2015

Investigating Intra- and Inter-Molecular Coevolution of Intrinsically Disordered Protein, Prothymosin-α

by

Brianna Biscardi

APPROVED BY:

DIRECTOR OF
THESIS: _____

Colin S. Burns, PhD

COMMITTEE MEMBER: _____

William E. Allen, PhD

COMMITTEE MEMBER: _____

Libero J. Bartolotti, PhD

COMMITTEE MEMBER: _____

John W. Stiller, PhD

CHAIR OF THE
DEPARTMENT OFCHEMISTRY: _____

Andrew T. Morehead, PhD

DEAN OF THE
GRADUATE SCHOOL: _____

Paul J. Gemperline, PhD

Acknowledgements

First, I would like to express my gratitude to my thesis adviser Dr. Burns for the continuous support of my thesis study and related research. His guidance helped me to better focus my research and writing and his patience and encouragement motivated me through the hard times.

Besides my adviser, I would like to thank the rest of my thesis committee, Dr. Stiller, Dr. Bartolotti, and Dr. Allen, for their insightful comments and encouragement. Their difficult questions helped me widen my research from various perspectives.

I also thank Julien Dutheil for his help implementing and understanding CoMap. His advice helped move my project forward.

My sincere thanks also goes to Dr. Kennedy who provided me the opportunity to join his research team as an undergraduate so I could develop the foundations necessary to my success in this program and in future endeavors.

I thank my lab mates and classmates who provided stimulating discussions and healthy distractions. I also thank the chemistry department - faculty, staff, and students - for fostering an enlightening academic experience.

Finally, I thank my parents for supporting me throughout the process including giving a loving “push” towards the end. Their love and support has helped me through my thesis and through life.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3D	Three-Dimensional
AMBER	Assisted Model Building with Energy Refinement
Apaf1	Apoptotic Protease-Activation Factor 1
Biscardi, 2014	Coevolution analysis on Myoglobin dataset by Biscardi executed in 2014 using methods described in this work
BLAST	Basic Local Alignment Search Tool
BLOSUM-62	Blocks Substitution Matrix
C	Compensation Index
CACS	Center for Applied Computational Sciences
CD Spectroscopy	Circular Dichroism Spectroscopy
$C_{function}$	Correlation due to function
$C_{interaction}$	Correlation due to interaction
ClustalX	Computer program for multiple sequence alignment
CoMap	Cosubstitution Mapping program for coevolution analysis
$C_{phylogeny}$	Correlation due to shared ancestry
CSD1	Cytosolic Copper/Zinc Superoxide Dismutase 1
$C_{stochastic}$	Correlation due to random mutation
$C_{structure}$	Correlation due to structure
D1	Distance Matrix 1 (in MUSCLE algorithm)
D2	Distance Matrix 2 (in MUSCLE algorithm)
Dutheil, 2007	Covolution analysis on Myoglobin dataset by Dutheil and Galtier in 2007 using methods described in previous work
ECU	East Carolina University
HSQC-NMR Spectroscopy	Heteronuclear Single Quantum Coherence Nuclear Magnetic Resonance Spectroscopy
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
kDa	Kilo Dalton
Keap1	Kelch-like ECH-associated Protein 1

MAP2c	Microtubule-associated Protein 2c
MEGA	Molecular Evolutionary Genetics Analysis
mHtt	Mutant Huntingtin
MSA	Multiple Sequence Alignment
MUSCLE	Multiple Sequence Comparison by Log Expectation
NCBI	National Center for Biotechnology Information
NLS	Nuclear Localization Signal
NMR Spectroscopy	Nuclear Magnetic Resonance
ns	Nanoseconds
PDB	Protein Data Bank
PhyML	Software the estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences
ProTα	Prothymosin-α
ProTα-Apaf1	ProTα and Apaf1 complex
p-value	Function of observed sample results used for testing a statistical hypothesis
PyMOL	Molecular visualization system
RefSeq_Protein	Database of protein sequences from reference genomes
RMSD	Root-mean-square Deviation
SP	Sum of Pairwise Alignment Score (MUSCLE)
SwissProt	Manually annotated and reviewed section of UniProt Knowledgebase
TFE	Tetrafluoroethylene
TIP3BOX	Water model used for the simulation of aqueous solution
TLR4	Toll-like Receptor 4
UniProt	Universal Protein Resource
V	Substitution Vector for correlated coevolution
V*	Weighted substitution vector
$V^{\sim}$	Weighted, signed substitution vector

1 Introduction

Proteins are essential to life. They are the workhorse of the cell responsible for defense, movements, catalysis, signaling, structure, and transport. The function of a protein is almost always correlated with its structure. Antibodies have paratopes that specifically bind harmful agents with complementary structure called antigens to provide the body with a robust defense mechanism. The head domain of myosin binds to filamentous actin and uses ATP hydrolysis to "walk" along the filament allowing muscles to flex or extend. Enzymes bind to specific substrates with a complementary shape to catalyze chemical reactions. G Protein-Coupled Receptors are a class of seven transmembrane a-helices forming a cavity within the plasma membrane that serves as a ligand-binding domain crucial to signal transduction. α-keratins are made up of two right-handed alpha helices that are cross-linked by van der Waals and ionic interactions to give animal horn, claws, and hooves a rigid structure. Finally, hemoglobin with its 4 globular subunits in a tetrahedral plane closely associated with a porphoryin ring which binds oxygen and converts Fe(II) to Fe(III) allowing oxygen to transport through the veins.

The fact that proteins adopt a specific three-dimensional structure to manifest a specific function, also known as the structure-function paradigm, has been the prevailing theory for many years and is supported by data from X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and numerous biochemical studies *(1)*. However, there exists a class of proteins that are either completely or partially unfolded in their native state. This class of proteins, known as intrinsically disordered proteins

(IDP's) or intrinsically disordered regions (IDR's), form dynamic conformational ensembles that experience large fluctuations in the average positions of their amino acids *(2, 3)*. Continued focus on IDP's suggests they are highly abundant in nature and have numerous biological activities. These activities can be largely summarized as molecular recognition, molecular assembly, transcription, replication, and chaperoning functionality can occur by remaining unbound or by transiently or permanently binding other proteins. In some cases they will undergo disorder-order transitions upon binding *(2-6)*.

Prothymosin-α (ProTα) is an IDP that has several biological functions and these depend upon cellular context and the proteins it associates with. Prothymosin-α was originally isolated from a family of peptides secreted by the thymus with proven immunomodulatory activity *(7).* It is found in the nuclei of virtually all mammalian cells and is highly expressed in cancer cells *(8).* ProTα is a small (12.5 kDa) protein with a highly unusual primary structure. With glutamic acid and aspartic acid residues making up more than half of its composition, ProTα is considered the most acidic polypeptide in the eukaryotic world. Most of the acidic residues are clustered in the central region and a nuclear localization signal (NLS) is found at the C-terminal end *(7, 9).* Additionally, cleavage of the N-terminal region by lysosomal asparaginyl endopeptidase produces thymosin-α1, another biologically active molecule *(7).* Important functional regions of prothymosin-α are shown in figure 1.1. Thymosin- α1 (orange) is a biologically active molecule when cleaved by lysosomal asparaginyl endopeptidase. The acidic region (red) is the area that is most abundant in aspartic and glutamic acids. The nuclear localization signal (blue) consists of sequence KR (87-88) & KKQK (101-104) and allows for incorporation of ProTα into the nucleus.

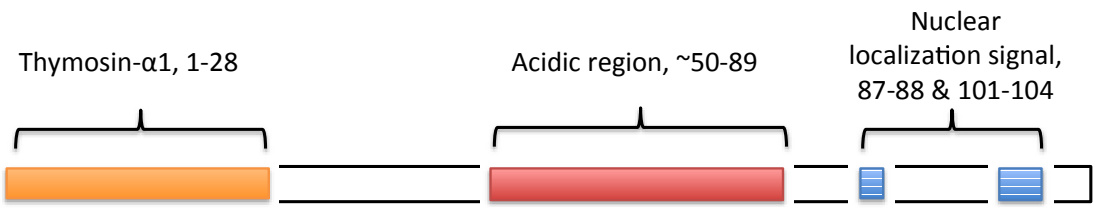

Figure 1.1: Schematic of important functional regions of prothymosin-α

New functions of this protein are constantly being discovered and existing roles are being further characterized. Some of the well-characterized roles of ProTα, such as anti-apoptotic ability, anti-oxidative stress activity, and immunomodulatory activity, are summarized below and shown in Figure 1.2 where red arrows indicate a binding interaction between proteins while black arrows and T-bars indicate turning on/off a function, respectively.

- *Oxidative stress activity:* Kelch-like ECH-associated protein 1 (Keap1), a protein responsible for inhibiting oxidative stress, interacts with prothymosin-α directly and specifically *(10).* This interaction proves a novel mechanism for anti-oxidative stress in which ProTα acts as an on/off switch *(8).* It binds Keap1; which releases transcription factor Nrf2 which upregulates anti-oxidative stress (8). It also mediates nuclear import of Keap1/Cul3/Rbx1 complex leading to degradation of Nrf2 acting as a switch off of anti-oxidative stress. Keap1 binds prothymosin-α at residues 32-52 and a pull-down assay verified that this interaction is direct *(10).*
- *Anti-apoptotic ability:* ProTα has also been reported to inhibit apoptosis, or programmed cell death. It does so by binding apoptotic protease activating factor 1 (Apaf-1) to inhibit apoptosome formation. Qi et al. (11) examined the interaction between Apaf-1 and 15N-labeled ProTα using two-dimensional ^{1}H-^{15}N heteronuclear single-quantum correlation nuclear magnetic resonance (HSQC-NMR). They identified residues 4-14, 31-43, 84-87, and 106-110 of ProTα to strongly and specifically bind Apaf-1 *(11).*

- *Immunomodulatory function:* It has recently been reported that prothymosin-α interacts with mutant Huntingtin (mHtt), the polyglutamine-expanded protein responsible for Huntington's disease. Overexpression of ProTα reduces mHtt-induced cytotoxicity while knockdown of ProTα enhances mHtt cell death. Furthermore, the central acidic region of prothymosin-α (residues 41-83) is required not only for interaction with mHtt but also to prevent mHtt caused cytotoxicity (12). It has also been reported that ProTα, which is released by CD8+ T-cells in the adaptive immune system, can suppress HIV-1 activity through the innate immune system. It does so by acting as a signaling ligand for toll-like receptor 4 (TLR4) to trigger both the MyD88 pathway for induction of proinflammatory cytokines and the TRIF-dependent pathway for IFN-β induction. This interaction is attributed in large part to ProTα residues 50-89 and suggests it may have important function in inhibiting retroviral activity *(13)*.

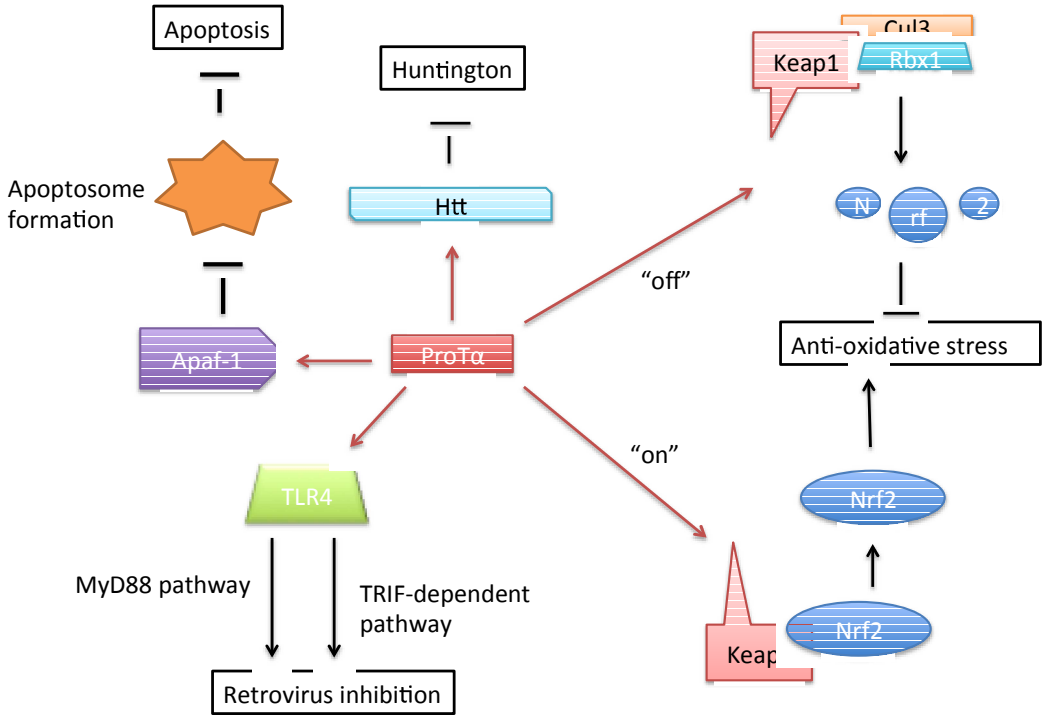

Figure 1.2: Biological functions of prothymosin-α and its protein interaction partners.

Life systems interact and thus do not evolve independently. The effect of a mutation in one system depends on the state of an interacting system. This is known as coevolution. Coevolution can be of many different types. At the molecular level, positions (amino acids or nucleotides) can be interacting because they are in contact in a 3D structure and share a common structural constraint. Distinct positions can also be involved in an active site and be under functional constraint with other molecules at higher organization levels *(14-15).* These interactions, as they affect the probability of a mutation to spread in a population and become fixed, leave a signature in the genome sequences. Comparative sequence analysis allows for these signatures to be decoded. The possibility of predicting molecule structural or functional features through sequence

comparison and coevolutionary patterns has proved successful and is still widely used for predicting RNA secondary structures *(16).* More recently, coevolution methods used to study evolution of protein sequences in relation to their structure.

This research focuses on applying methods of coevolution detection to an intrinsically disordered protein, prothymosin-α. The utility of coevolution detectors in finding potential contact sites and regions of interactions will be evaluated. This will be done for two cases; intramolecular interaction sites within ProTα and intermolecular interactions between regions of ProTα and its known or putative binding partners. A pipeline is developed in this study that is validated by successfully re-analyzing myoglobin. Myoglobin has been previously analyzed by coevolution calculations used in this study. It also has a well-established 3D structure revealed by several experimental methods; this allows for mapping of the coevolutionary findings onto the 3D structure providing a means for evaluating the soundness of the results. Since coevolution calculations have been previously performed by Dutheil et al., *(17)* myoglobin serves as a proof of concept model in this study. ProTα has experimentally been shown to be an unstructured protein and was chosen as a subject because coevolution methods have not been applied to such protein. As IDPs are highly dynamic, important, possibly transient, intramolecular interactions may be difficult to detect by standard structural characterization methodologies. Finally, ProTα-Apaf1 are known interaction partners and were chosen to answer the question as to whether coevolution is capable of locating binding regions.

2 Background

2.1 Overview

In order to study coevolution, one must look at how an amino acid has changed throughout the course of evolution with respect to another amino acid. This requires a comparative look at the protein of interest, or comparing homologues. Homologues refer to the same protein in different species *(18).* There are ways to detect homology that leverage the vast databases of fully sequenced reference genomes. Next, the homologs must be multiply aligned, meaning that individual species' sequences are displayed in stacked rows and the columns represent equivalent residue positions. In this manner one column is equivalent and represents a single amino acid and all of its mutations. Finally, in order to better account for coevolution, family history is considered to ensure correlated mutations are due to structure or function rather than a shared evolutionary trajectory. This requires constructing the best possible phylogeny for a protein *(17).* Once a sequence alignment and a tree are established, these data are parameterized to detect correlated changes by statistical methods *(19).* Finally, molecular modeling may be used to provide further evidence that correlated changes detected by coevolution are also biophysically feasible binding locations. A schematic of this study is shown in Figure 2.1.

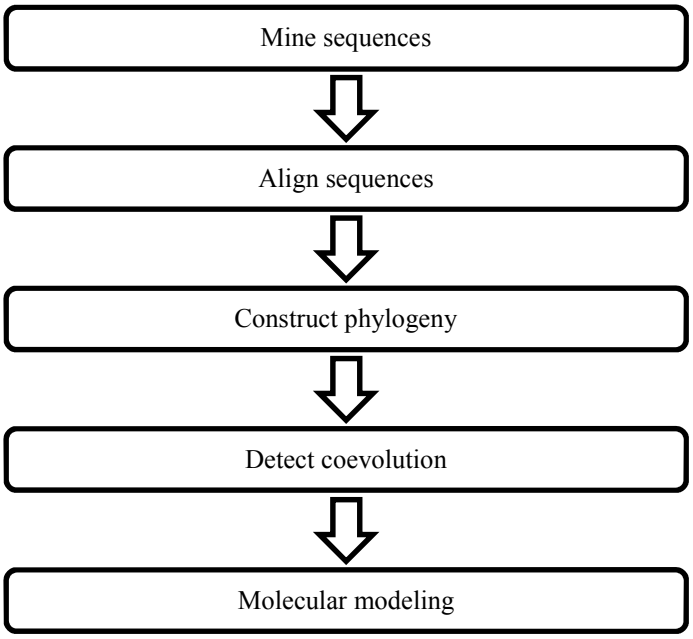

Figure 2.1: Schematic of detecting coevolution for prothymosin-α

2.2 Homology

Homology translated literally means the "study of likeness". When biologists first began to study the anatomy of humans and other vertebrates they noticed a striking similarity between their skeleton, muscles, and other structures. One example is the comparison of a human arm, a bat wing, and a porpoise paddle. They are similar structures in that they are each comprised of a humerus, radius, ulna, carpals, metacarpals, and a set of phalanges. The same bones – in relatively the same position – perform a function necessary for their respective species' survival. If an engineer were to build a grasping tool, a wing, or a paddle, they would never use the same underlying pattern. This provides logic for species descending from a common ancestor with

modified forms of the same design. Thus homology is considered a similarity that exists due to shared ancestry *(18).*

Homology also exists at the molecular level. Family resemblance is detected by comparing three-dimensional (3D) structure of proteins. Since a proteins 3D structure is manifested by its primary amino acid sequence, the underlying sequence should show that same degree of similarity *(1).* Many sequence comparison methods have been developed to examine protein homology. To show an example of homology, myoglobin, the protein that binds oxygen in the muscle, is considered. Figure 2.2 represents the crystal structure of *homo sapiens* (human) myoglobin 1-149 in blue (2MM1) and *physeter catadon* (sperm whale) myoglobin 1-153 in green (1MBN). The protein's N-terminus and C-terminus are labeled with the letters N and C, respectively. Black solid or dashed boxes indicate structural differences. Note, *homo sapiens* myoglobin is shorter by 4 residues which did not crystallize *(20, 21).* Figure 2.3 show the primary sequence of myoglobin in *homo sapiens* (blue) UniProt ID: P02144 and *physeter catadon* (green) P02185 *(22).* Carrots represent mutations. Sequences were collected from UniProt.

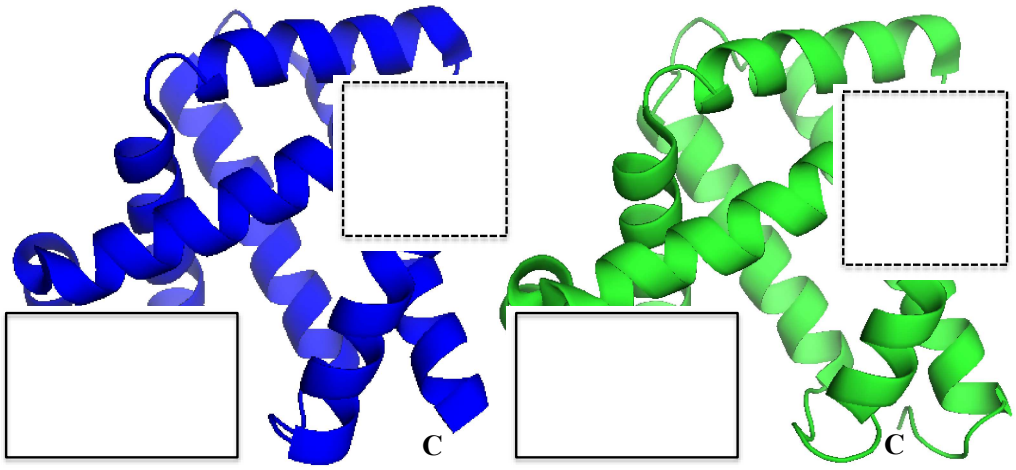

Figure 2.2: Crystal structure of myoglobin in *homo sapiens* and *physeter catadon*

```
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDE
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAE
 ^  ^       ^  ^     ^^    ^^       ^          ^     ^ ^
MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISE
MKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISE
          ^       ^            ^                ^
CIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
AIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
^  ^  ^ ^                        ^ ^^     ^
```

Figure 2.3: Primary sequence of myoglobin in *homo sapiens* (blue) compared with *physeter catadon* (green).

As seen in Figure 2.2, the 3D structure of myoglobin is similar for human and sperm whale. A few differences are present, such as slightly more alpha helical character in human myoglobin indicated with boxes, but in general their conformation is extremely similar. The underlying sequences of myoglobin in Figure 2.3 are also extremely similar. They are considered to be homologues – the same protein but in different species. So, if a newly sequenced protein is

homologous to an already sequenced one, its structure, function and evolutionary history can be inferred *(1).* Several methods exist to determine if two proteins are homologues.

Homology of protein sequences is typically determined by performing a pairwise sequence alignment in which two sequences are systematically aligned with each other to identify regions of significant overlap *(1).* This will be demonstrated keeping myoglobin as an example and comparing the protein in human to that of sperm whale. Over the course of evolution, the sequences of two proteins with a common ancestor could have diverged in many different ways. Insertions and deletions could have occurred in any functional or non-functional region of the proteins. A missense mutation could have also occurred which could change an individual amino acid. To understand how sequence alignments take into account regions of variation, one sequence (for example sperm whale myoglobin) can be slid across the other sequence (human myoglobin) one amino acid at a time and the number of matched residues will be known as sequence identities *(1).* In this case, if the two sequences are aligned from N- to C-terminus (as in Figure 2.3) they have 130 sequence identities out of 154 residues. It can be slid across as many times as necessary to find the highest sequence identity. In this case, the N- to C-terminus alignment displayed in Figure 2.3 is actually the best alignment.

Insertions and deletions are taken into consideration in sequence alignments by adding gaps *(1).* In order to demonstrate a case where gaps will be necessary, a more divergent pair of species will be demonstrated. Figure 2.4 shows the sequence of myoglobin in *homo sapiens* (blue) and in *tetraodon nigroviridis* (pufferfish) (red) aligned according to amino acid number. Sequenced were obtained from UniProt (*t. nigroviridis* ID: Q701N9). An asterick (*) represents "sequence identities", or amino acids that are the same in both sequences.

```
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDE
MGDFDMVLKFWGPVEADYSAHGGMVLTRLFTENPETQQLFPKFVGIAQSELAGNA
**   *         *      *

MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISE
AVSAHGATVLKKLGELLKAKGNHAAILQPLANSHATKHKIPIKNFKLIAEVIGKV

CIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
MAEKAGLDAAGQQALRNIMATIIADIDATYKELGFS
                                *
```

Figure 2.4: Sequence of *homo sapiens* myoglobin (blue) and *tetraodon nigroviridis* myoglobin (red) aligned according to amino acid number.

When the two sequences are aligned from N- to C-terminus, according to amino acid number, there are several residues that do not match. In this orientation, there are only 6 sequence identities (represented by an asterisk). By introducing gaps, as displayed in Figure 2.5, the number of sequence identities can be increased to 66 out of 154. This allows the alignment method to compensate for the insertions or deletions that may have taken place in one molecule but not the other.

```
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDE
MG----DFDMVLKFWGPVEADYSAHGGMVLTRLFTENPETQQLFPKFVGIAQ-SE
**         **   ** ****    **   ** ***    ***    * **        *

MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISE
LAGNAAVSAHGATVLKKLGELLKAKGNHAAILQPLANSHATKHKIPIKNFKLIAE
          ******   **   ** ** * *    *** *********  *     * *

CIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
VIGKVMAEK--AGLDAAGQQALRNIMATIIADIDATYKELGFS-
 *  *    *       *  * *           *     ******
```

Figure 2.5: Sequence comparison between *homo sapiens* myoglobin (blue) and *tetraodon nigroviridis* myoglobin (red) with gaps inserted to increase the number of sequence identities and provide a better alignment. An asterisk (*) represents a sequence identity.

Since gaps improve the number of sequence identities in an alignment, they can be used erroneously to insert an unreasonable amount of gaps. In order to avoid this, methods have been developed for the insertion of gaps in the automatic alignment of sequences. These methods use scoring systems that add a penalty if there are too many gaps inserted *(1).* As an example, an identity can be scored at +10 points and a gap (no matter what size) can be scored as -25 points. For the alignment in Figure 2.5, there are 66 identities (66 x 10 = 660) and there are 3 gaps (3 x -25 = -75). This would produce an overall score of 585 (660 – 75).

Next, the significance of this score must be found statistically to determine the level of identity. That is, to ensure the identity is not due to chance alone. Since every protein is made up of combinations of the same 20 amino acids, the alignment of any two unrelated sequences will result in some sequence identity. Likewise, if two proteins have identical composition, they may not be linked by evolution. The significance of an alignment can be assessed by randomly rearranging one of the sequences. This process can be repeated many times to yield a histogram showing, for each possible score, the number of rearrangements that received that score. If the original score (585 between *homo sapiens* and *tetraodon nigroviridis* myoglobin) is not appreciably different from the rearranged scores, then the sequence may not be due to a shared evolution but instead it is due to chance. Since *homo sapiens* myoglobin and *t. nigroviridis* myoglobin are homologs, they should have a score that lies far outside of this plot and thus can be called homologs with confidence. Thus, it is determined that they have a shared ancestry *(1).*

The method described above assigns points to positions that are occupied by identical amino acids in the two sequences being compared. It gives no credit to changes in amino acids that often have similarities to each other (i.e. polarity, charge, volume). As such, methods have

been created to compare two amino acids and assess their degree of similarity. These methods characterize amino acid substitutions as conservative substitutions and non-conservative substitutions. A conservative substitution replaces one amino acid with another that is similar in size and chemical properties. These substitutions can often be tolerated without compromising protein function. A non-conservative substitution replaces one amino acid with one that is dissimilar and usually has a greater effect on protein function *(1)*.

The type of substitution can be accounted for by creating a substitution matrix that describes a scoring system for the replacement of one amino acid with any of the other 19. In a substitution matrix, a large score can be assigned to substitutions that occur relatively frequently whereas a large negative score can be assigned to substitutions that occur rarely. A commonly used substitution matrix is Blosum-62 (blocks of amino acid substitution matrix) *(23)*. This matrix shown in Figure 2.6 is based on local alignments of over 2000 blocks in more than 500 groups of related proteins. Using systems like these, homology can be detected in less obviously related sequences with greater sensitivity *(1)*.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2.6: Blosum-62 amino acid scoring matrix *(23)*.

When a sequence of a protein is determined, it can be compared to all previously characterized sequences to understand its evolution. This can be accomplished by searching databases of known sequences. The Basic Local Alignment Search Tool (BLAST) is the most frequently used tool for finding homologs *(24)*. This resource is made available by the National Center for Biotechnology Information (NCBI). In order to use BLAST, an amino acid sequence is entered into the web browser and several vast databases (some with over 30 billion sequences) of known sequences are searched. A BLAST search yields a list of local sequence alignments and an estimate of the likelihood that the alignment occurred by chance.

The BLAST algorithm uses a heuristic search method to find matches of not the entire sequence but of short "words" between the two sequences. This method is used to save on

computational time and provide robust results quickly. It then scans the database for these hotspots and when a match is identified, a local alignment is performed of the query-match pair. This query-match pair must satisfy a certain threshold usually scored by the Blosum-62 (as described above). In order to determine whether this is a "good" alignment (that is biologically relevant) and not just due to chance, two statistical scores are given. A bit score, which gives an indication of how good the alignment is, and an E-value which gives an indication of the statistical significance of a given pairwise alignment and reflects the size of the database scoring system used. The higher the bit score the better the alignment and the lower the E-score, the better the match.

2.3 Multiple Sequence Alignment & Phylogeny Construction

Once homologs are determined and a group of evolutionarily related species is identified, the ancestry of these species can be further evaluated by constructing phylogenetic trees (or phylogenies). A phylogenetic tree is most powerful if an accurate multiple sequence alignment (MSA) is achieved. After using BLAST to identify all possible homologs of a sequence of interest, a multiple sequence alignment can be performed. Much like the pairwise alignments in Figure 2.5, an MSA represents homologs in such a way that equivalent residues are placed in the same column. So, a column in a MSA contains amino acid changes that have occurred at one position throughout evolution *(25)*. Since MSA's align more than 2 sequences, they become much more ambiguous.

One common method used to create an MSA is multiple sequence comparison by log expectation (MUSCLE) *(26)*. Although there are other more widely used methods for aligning

sequences, MUSCLE is slightly more accurate and 2-5 times faster for normal sized to large data sets *(25, 27)*

The MUSCLE algorithm, shown in Figure 2.7, uses a progressive method to optimize the sum of pairwise alignment scores (SP). It first drafts a progressive MSA with emphasis on speed rather than accuracy (stage 1). In step 1.1, Kmer distance is computed for each pair of input sequences giving matrix D1. In step 1.2, matrix D1 is clustered by UPGMA producing a binary tree, Tree1. 1.3 A progressive alignment is constructed by following Tree1. A profile is constructed at each leaf from input sequence and a pairwise alignment is constructed at each node producing MSA1. Stage 2 develops an improved progressive by reestimating guide tree using Kimura distance which is more accurate than Kmer. 2.1 Kimura distance is computed from MSA1 giving distance matrix D2. 2.2 Matrix D2 is clustered by UPGMA producing TREE2. 2.3 A progressive alignment is produced following Tree2 producing MSA2. MSA2 is optimized by computing alignments for subtrees whose branching orders change relative to Tree1. Stage 3 is the refinement stage. 3.1 An edge is chosen from Tree2. 3.2 Tree2 is divided into 2 subtrees by deleting the edge. A profile of multiple alignment of each subtree is computed. A new multiple alignment is produced by realigning the 2 profiles. If the SP score is improved, the new alignment is kept. If is not improved, the new alignment is discarded.

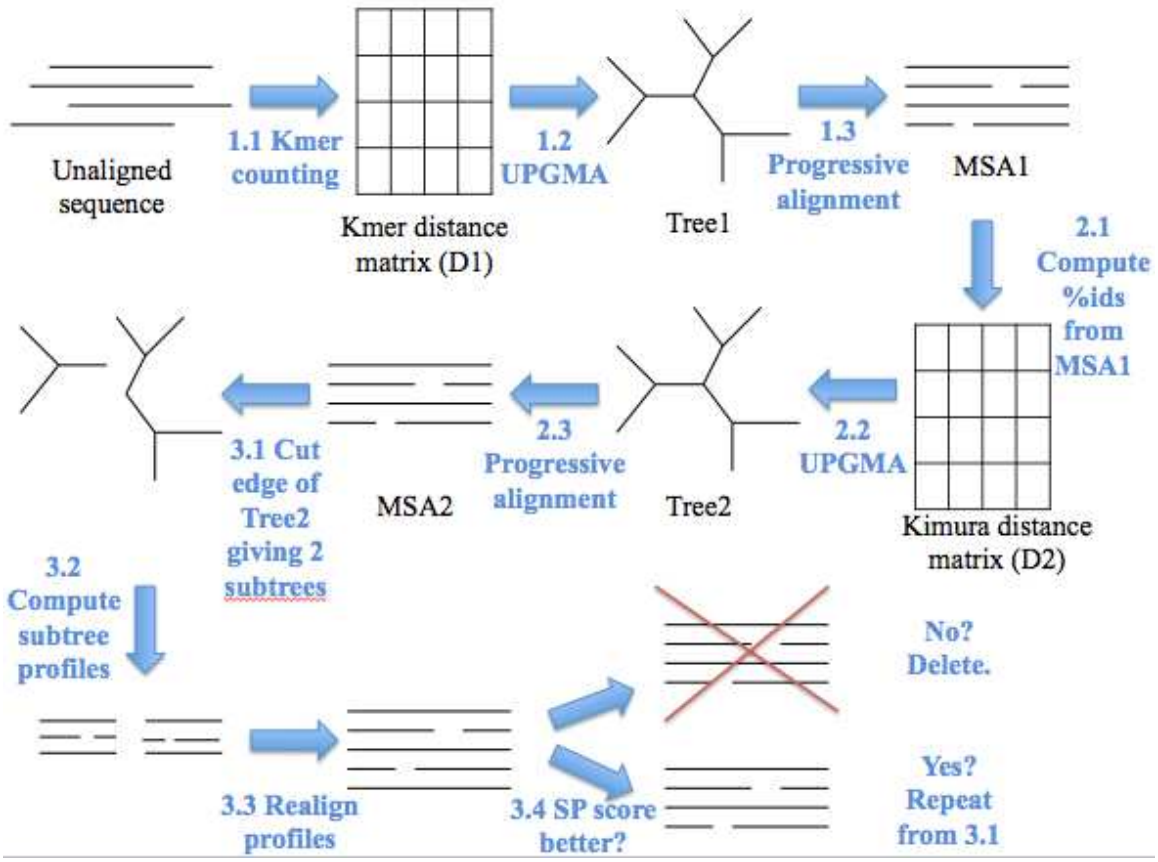

Figure 2.7: The MUSCLE algorithm.

Since homology is often manifested as sequence similarity, the relatedness of members of a family of proteins may be deduced by examining sequence similarity. Sequences that are more similar to one another have had less time to diverge than have sequences that are less similar to one another. Using this notion, MSA's are used to construct a phylogenetic tree. A phylogeny, shown in Figure 2.8, shows the evolutionary history among populations or species and clarifies who is related to whom. The tips of the tree, named X, Y, and Z, represent groups of species and the branches connecting them represent time. A node (labeled with arrows) represents a divergence or the point in time when an ancestral group split into two or more descendants. The more nodes that exist between two species, the more distantly related they are believed to be

(18). So in Figure 2.8, species Y and Z are more closely related than are species X, however all three are related because they have diverged from a common ancestor.

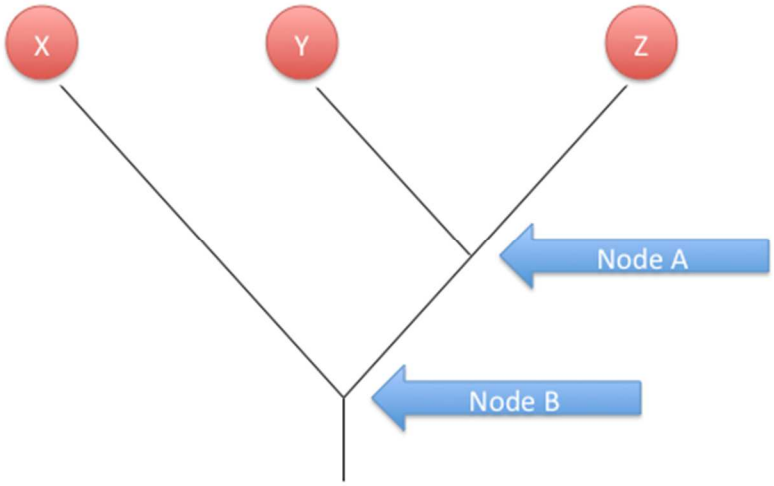

Figure 2.8: Phylogenetic tree describing relationship between species 'X', 'Y', and 'Z'.

Relationships depicted in a phylogenetic tree are estimated from morphological data, fossil record, or sequence information. When using sequence information, each terminal node represents a protein and the length of the branch connecting each pair of proteins is proportional to the number of amino acid differences between their sequences *(18)*. An example using human, sperm whale, and pufferfish myoglobin is illustrated in Figure 2.9. Sperm whale and human have the most similar myoglobin sequences and thus the smallest number of amino acid differences (as seen in Figure 2.3) and as such they are branched closest together on the phylogeny. The myoglobin sequence of pufferfish is the least homologous to both human and sperm whale, and as such it is not as closely related. The tree in Figure 2.9 is logical because human and sperm whale are both mammals and thus should be more closely related to each other than to fish. Phylogeny of myoglobin in species human, sperm whale and pufferfish. This phylogeny was constructed in Microsoft PowerPoint and does not accurately depict time of divergence.

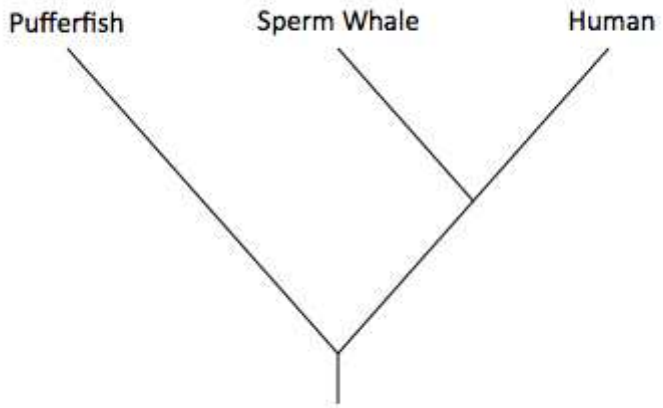

Figure 2.9: Phylogeny of myoglobin in species human, sperm whale and pufferfish.

Phylogeny construction methods can be either distance-based or character-based. Distance-based methods calculate the distance between every pair of sequences and the resulting distance matrix is used for tree construction. Character-based methods simultaneously compare all sequences in the alignment considering one character at a time to calculate a score for each tree. Neighbor joining, a distance-based method is the most commonly used method due to its speed and simplicity. This method operates by starting with a star shaped tree and successively choosing a pair of taxa to join together based on pairwise evolutionary distance until a fully resolved tree is obtained *(27)*. In Figure 2.10, the two newly joined taxa are represented by an ancestor (node Y) and the number of taxa connected to the root is reduced by one (node X).

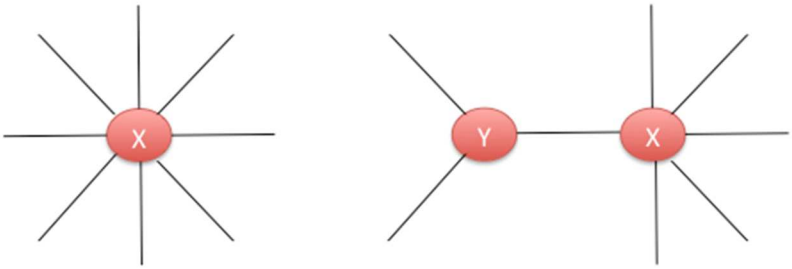

Figure 2.10: Neighbor joining method for constructing phylogenetic trees.

Although sequence information provides a systematic way of relating species to one another, other evidence is necessary to obtain a "correct" estimate of most species evolution. For this reason, the phylogeny constructed must be manually edited to match known branching orders *(17)*. This is particularly useful for mammals since they diverged much more recently than other vertebrates and thus their sequences are much more similar.

2.4 Coevolution Detection

Coevolution can be detected by determining non-independent correlation between two (or more) sites in a sequence alignment. Correlation (Corr) can be defined as:

$$Corr = C_{structure} + C_{function} + C_{phylogeny} + C_{interaction} + C_{stochastic}$$

Where $C_{structure}$ and $C_{function}$ signify correlation due to structure and function, effectively what coevolution attempts to uncover. $C_{phylogeny}$ represents correlation due to shared ancestry and $C_{interaction}$ describes the interaction between all three of these. $C_{stochastic}$ is a measure of random effects from uneven or incomplete sequence sampling. Most programs developed to detect coevolution filter stochastic noise, as this is inherent to statistics. Some programs filter phylogenetic noise, which has been proven to enhance predictions *(28)*. CoMap uses a process of probabilistic substitution mapping (described below) in order to account for phylogeny *(16-17)*.

While there are several other programs of this kind, CoMap is unique in that it detects groups of arbitrary size, accounts for biochemical properties of amino acid changes, and distinguishes compensatory evolution from other kinds of correlated evolution.

Compensatory mutations are those in which one perturbing mutation is compensated by one or several mutations at other sites to maintain a higher order structure. This may be local, i.e. when a big-to-small substitution occurs at position i it may be deleterious but may be compensated by a small-to-big substitution at position j. These sites are more likely to be in contact in 3-dimensional space. Sites involved in the recognition of an interacting partner may tend to coevolve and be compensated by changes in the interacting molecule, an example of distal compensation. Compensation is more specific and can be a more powerful approach however it requires a priori knowledge. Correlated substitution, a more common approach, is more general and less powerful however no prior knowledge is required. Correlated substitutions can be made more powerful by making hypotheses on the biochemical nature of the change. Compensation statistics require hypotheses on biochemical properties. Biochemical properties of amino acid substitutions include charge, polarity, volume, and grantham, which is a combination of volume, polarity, and atom composition. These are accounted for by employing a procedure known as "weighted probabilistic substitution mapping" which weights different types of substitutions according to the given biochemical property and further estimates the amount of biochemical change *(17)*.

The CoMap algorithm utilizes three main steps. First, given a sequence alignment and a tree, a coevolution statistic is defined as a correlation coefficient, ρ, or a compensation index, C (defined below). In this step three different vectors are computed (V, V*, and V~) which are

described in this section. Next, candidate groups are detected by clustering sites according to their coevolution statistic. Finally, a statistical test (p-value) is applied to assess the significance of candidate groups.

CoMap works by mapping substitution events onto a phylogeny. This method is known as substitution mapping and it consists of estimating the number of substitutions that occurred on each branch (b) for each site (i). This can be visualized in Figure 2.11.

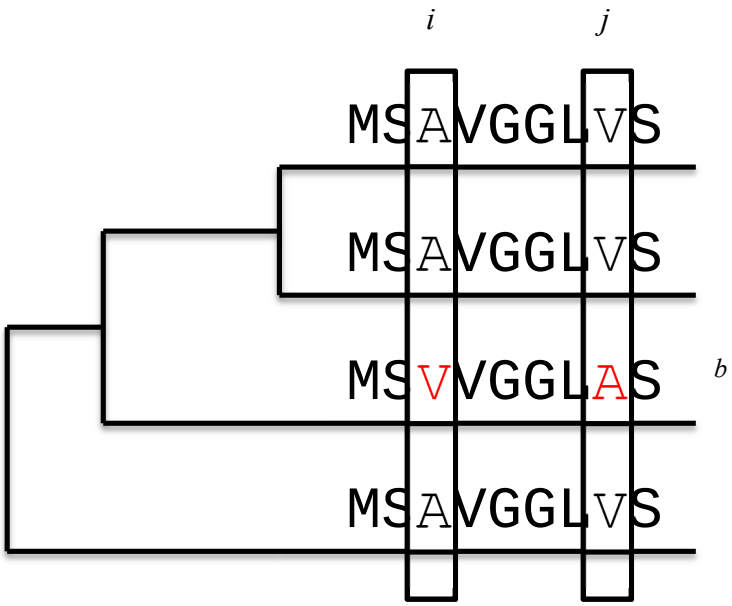

Figure 2.11: Substitution mapping to account for phylogeny in coevolution statistics.

These numbers are stored as a substitution vector (V) and can be computed by the following formula:

$$Vi,b = \Sigma xp \Sigma xq P(xp,xq|Di,\theta) \times nxp,xq(t)$$

Where Vi,b is the expected number of substitutions that occurred on branch b for site i. xp and xq are the states at the top and bottom of the node of the branch for this site. Di is the ith site of

the data set (i.e. a column of the MSA) and nxp,xq(t) is the expected number of substitutions on a branch of length t knowing its initial state (xp) and final state (xq). Exact computation of nxp,xq(t) does not improve the coevolution detection and may be approximated to 0 if xp is equal to xq and 1 if xp does not equal xq. This equation can then be generalized to account for rate across site variation:

$$Vi,b = \Sigma c \Sigma x_p \Sigma x_q P(x_p,x_q,rc|Di,\theta)x\ nx_p,x_q(t\ x\ rc)$$

Where the c^{th} rate class has a relative rate rc.

The first factor in this summation is the posterior probability (a Bayes theorem expression of conditional probability, meaning the probability of A given B, or A|B) of having state x_p at the bottom node, state x_p at the top node, and rate class c given the data and parameters. It can be computed by the following formula:

$$P(x_p,x_q\ rc|D_i,\theta) = P(x_p,x_q\ rc,D_i|\theta) / P(D_i|\theta)$$

$$= P(x_p,x_q,D_i|\theta,rc)\ x\ Pr(rc) / P(D_i|\theta)$$

Where the first factor of the numerator is the likelihood for site i conditional on states x_p and x_q at the top and bottom nodes and rate are equal to rc. This likelihood is computed after having multiplied all branch lengths by rc and summing over all possible ancestral states at each node except for the top and bottom nodes of the branch b, for which states x_p and x_q are fixed. P(rc) is the prior probability for site i of being in rate class c, and P(Di|θ) is the likelihood for site i. Then the substitution vector or V_i is solved as follows:

$$V_i = (vi,1,\ldots,\ vi,b,\ldots,\ vi,m)$$

Where m is the total number of branches in the tree. This vector, V_i, is stored as the unweighted substitution mapping.

A weighted substitution mapping (V*) is also applied which weights sites differently according to their biochemical properties. This vector, Vi*, is calculated as follows:

$$V_i,b* = = \Sigma c \Sigma x_p \Sigma x_q \; P(x_p,x_q,rc|D_i,\theta)x \; wx_p,x_q$$

Where w = {wx,y} is a matrix of substitution weights from the AAIndex database, a script included to apply a specified biochemical weight. Biochemical weights can either be attributed to correlated evolution (sites undergoing simultaneous changes in a given property) by setting wx,y = wx,y. On the other hand, setting wx,y = -wx,y allows detection of compensatory changes. Signed substitution vectors are denoted V~.

For correlated changes, the amount of coevolution (ρ) for a pair of sites (i, j) is measured by taking the Pearson correlation coefficient of the two substitution vectors:

$$\rho_{ij} = cov(V_i*,V_j*) \; / \; sd(V_i*) \; x \; sd(V_j*).$$

If the two sites tend to undergo substitution events in the same branch, ρ will be positive and tend toward 1 whereas ρ values near 0 are expected to have evolved independently. This measure is generalized to a group of arbitrary size s by defining the amount of coevolution for the group as the minimal pairwise correlation between sites in the group:

$$\rho = min_{i,j->1...s}\{\rho_{ij}\}.$$

From a geometrical standpoint, the correlation coefficient is the cosine of the angle between two substitution vectors, and the minimum correlation coefficient corresponds to the cosine of the maximum angle between the vectors of the group. This can be seen in Figure 2.12.

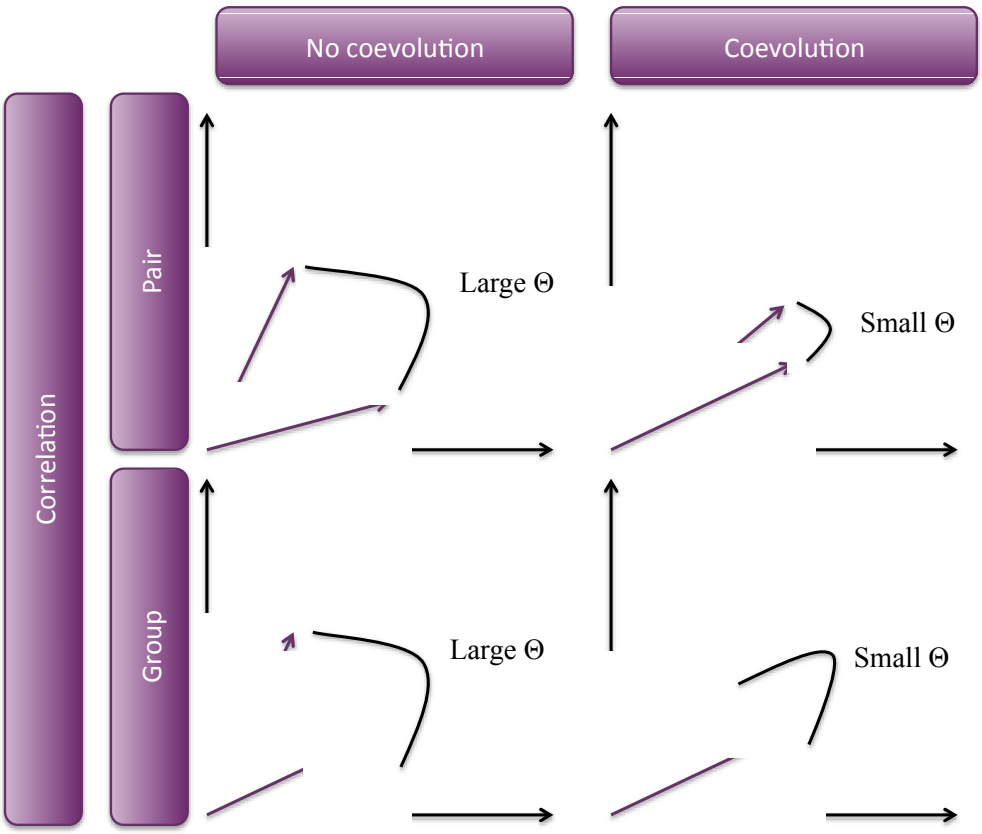

Figure 2.12: Correlation coefficient is the cosine of the angle between two substitution vectors

Weighted substitution mapping can track the direction of changed by giving opposite weight to X→ Y and Y → X. The resulting signed, weighted vectors (V~) can be used to test the compensatory nature of changes.

The tendency of sites to undergo compensatory mutations requires defining a compensation index C:

$$C_{ij} = 1\text{-}|V_i\tilde{} + V_j\tilde{}| \,/\, |V_i\tilde{}| + |V_j\tilde{}|.$$

Compensation is high when the normalized length of the sum vector of all substitution vectors in the group tends to 0. This is noted as Σ in Figure 2.13. Figure 2.13 presents a geometric interpretation of compensation. Purple arrows represent weighted vectors. Orange arrows represent the sum of vectors that are not equal to zero.

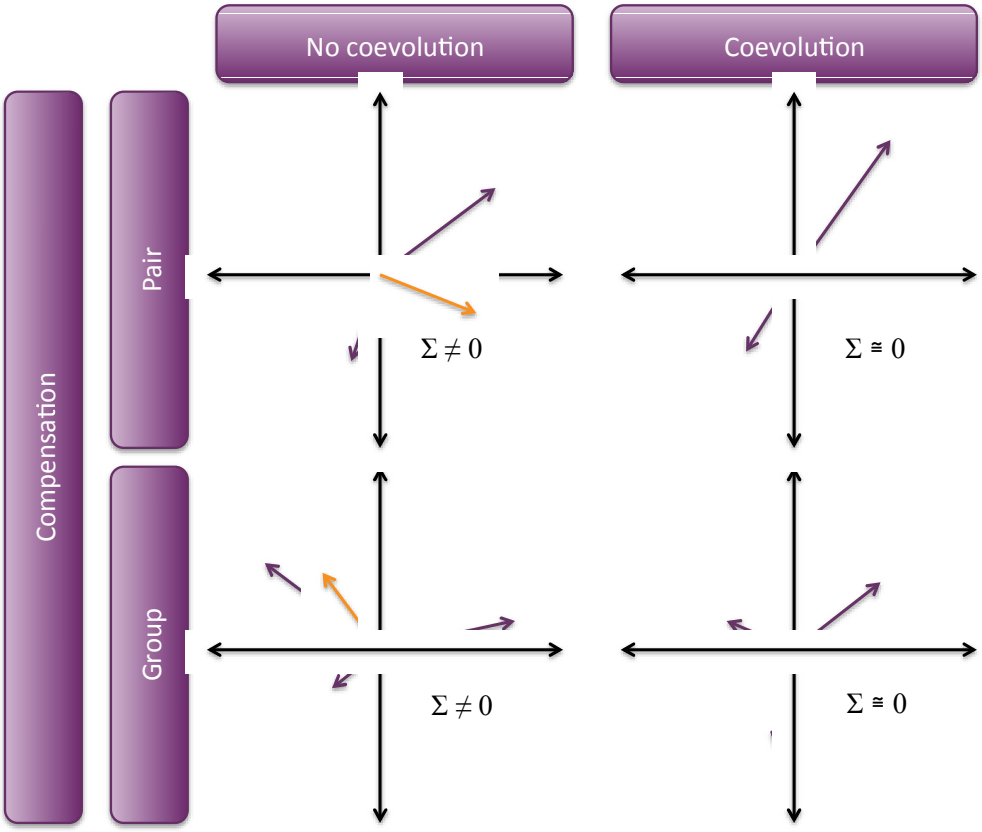

Figure 2.13: Compensation index is the sum vector of all substitution vectors.

The interpretation to be given to coevolution statistics highly depends on the evolutionary rate of the sites under consideration. The variability of site *i* is measured by taking the length N_i of its corresponding substitution vector (and N^*_i of weighted substitution vector).

$$N_i = \sqrt{\Sigma_b v^2_{ib}} \text{ or } N^*_i = \sqrt{\Sigma_b v^{*2}_{ib}}$$

The variability of a group of *s* sites is then defined at the minimum length of substitution vectors at all sites in the group, N_{min} (which coevolution statistics mechanically depend on).

$$N_{min} = min_{i\Xi 1..s}\{N_i\} \text{ or } N^*_{min} = min_{i\Xi 1..s}\{N^*_i\}$$

In order to find candidate groups of coevolving sites, a clustering analysis of substitution vectors is performed. First, the pair (u,v) in the distance matrix with the lowest distance is clustered. Next, the two selected sites are removed and a new pair (u, v) is added as a single entry. The distances between the pair and each remaining group is then found using the formula:

$$D(w,(u,v)) = max(d(w,u),d(w,v))$$

This is repeated until the matrix reaches size 1. Each step defines a new cluster and reduces the matrix size by 1. Each of these clusters is considered a candidate group of coevolving sites.

A parametric bootstrap method is used to evaluate the significance of clusters. One thousand data sets with the same number of sites as the one of interest are simulated. The substitution vectors are calculated and clustering is performed. The joint distribution of ρ and C under the null hypothesis of independence is obtained. Then the p-value is then computed for a group of sites by conditioning over N_{min}

$$p\text{-}value = Pr(\rho > \rho_{obs} | Nmin_{obs})$$

where ρ_{obs} is the measured value for statistic ρ. Since N_{min} is a continuous variable, a window centered on $Nmin_{obs}$ is used to evaluate p-values

$$p\text{-}value = N_1+1 \,/\, N_2+1$$

where N_2 is the number of simulation points with $Nmin \; \Xi \; [Nmin_{obs} - \omega/2, Nmin_{obs} + \omega/2]$, N_1 is the number of simulation points in this range with a correlation greater or equal to the observed value, and ω defines the size of the window which is 20% of the range of *Nmin* values in this work *(17)*.

2.5 Molecular Modeling

The goal of molecular modeling is to mimic the behavior of molecules and molecular systems. It is invariably associated with computer modeling and the Born-Oppenheimer approximation is assumed. This enables the electronic and nuclear motions to be separated. The much smaller mass in electrons means they can rapidly adjust to any change in nuclear positions. The energy of a molecule in its ground electronic state can be considered a function of the nuclear coordinates only. If some or all of the nuclei move then the energy will usually change. The new nuclear positions could be a result of a simple process such as a single bond rotation or it could arise from the concerted movement of a large number of atoms. The magnitude of the resulting rise or fall in energy will depend upon the type of change involved.

Molecular docking is a technique developed to attempt to predict the structure(s) of the intermolecular complex between two or more molecules. This is most widely used to suggest the binding modes of protein inhibitors *(29)*.

3 Methods

3.1 Overview

Three subjects were analyzed in this study – myoglobin, ProTα, and ProTα-Apaf1. Myoglobin serves as a proof of concept model, ProTα is the first experimental condition, and ProTα-Apaf1 is the second experimental condition. All three subjects were analyzed using the same programs and statistics. Initial protein sequences for myoglobin, ProTα, and Apaf1 were found in UniProt. Homologs for the three proteins were found in BLAST. Once a protein's homologs were established, they were imported into a program called molecular evolutionary genetic analysis (MEGA) version 6 *(30)*. Within this program, sequences were aligned with MUSCLE and a phylogeny was constructed using the neighbor-joining method. The resulting MSA and phylogeny were imported into CoMap program where coevolution statistics were applied. Apaf1, since it was analyzed within a system (ProTα-Apaf1), an MSA was constructed for both proteins individually and they were linked together in MEGA program. A tree was constructed for the entire ProTα-Apaf1 linkage. ProTα-Apaf1 was the only system analyzed with molecular docking. All MEGA analyses were performed on a Windows Desktop. CoMap and molecular modeling analyses were performed on a Quad 4 Core i7 Linux station in East Carolina University's (ECU) Center for Applied Computational Sciences (CACS).

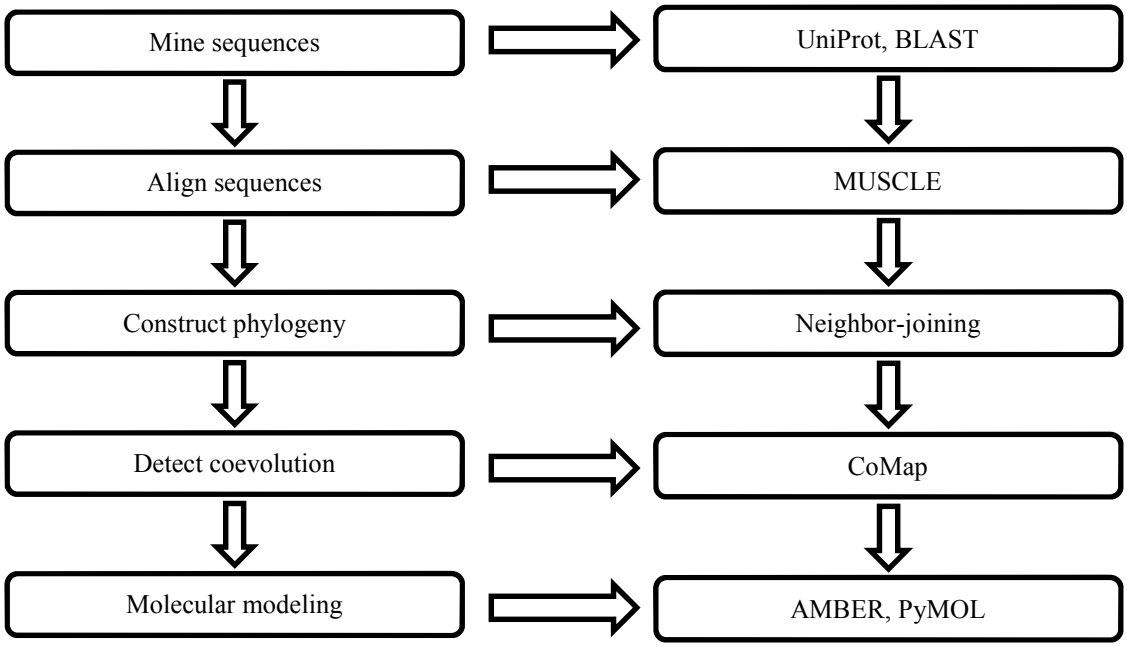

Figure 3.1. Schematic of studying coevolution of prothymosin-α and corresponding programs used.

3.2 Myoglobin

The sequence for human myoglobin was found in the UniProt database, entry P02144. This sequence (Figure 2.3) was used as a search query in the BLAST refseq_protein database (containing only highly curated, non-redundant reference sequences). 74 homologs were found in BLAST. These species, listed in Appendix I, were imported into MEGA and an MSA was performed using MUSCLE. The resulting MSA was utilized to construct a phylogenetic tree. This phylogeny was constructed using the neighbor-joining method and the optimal tree with the sum of branch length = 4.617 is shown in Appendix III. Evolutionary distances were computed using the Dayhoff model (amino acid substitutions per site) and the rates among sites was assumed to be uniform. Gaps and missing data were completely deleted and 92 positions were in the final dataset.

The resulting MSA (Appendix II) and phylogeny (Appendix III) were input into CoMap. Two CoMap analyses were performed, correlation and compensation. The correlation analyses consisted of five distance measures; grantham, polarity, volume, charge, and unweighted. The compensation analyses included four distance measures; grantham, polarity, volume, and charge. Once a statistic was defined, an R script was run to determine p-values. Only statistically significant results were considered a signal due to coevolution.

3.3 ProTα

The sequence for ProTα was also found on UniProt *(22)*. Two sequences can be found for human ProTα differing only by a glutamic acid in the 40th position. The sequence including E40 (underlined), UniProt ID P06454, was chosen as a search query in BLAST and can be seen in Figure 3.2.

MSDAAVDTSS$_{10}$ EITTKDLKEK$_{20}$ KEVVEEAENG$_{30}$ RDAPANGNAE$_{40}$ NEENGEQEAD$_{50}$

NEVDEEEEEG$_{60}$ GEEEEEEEEG$_{70}$ DGEEEDGDED$_{80}$ EEAESATGKR$_{90}$ AAEDDEDDDV$_{100}$

DTKKQKTDED$_{110}$ D$_{111}$

Figure 3.2: Primary sequence of prothymosin-α in *homo sapiens*

The resulting 65 homologs were imported into MEGA and an MSA was performed using MUSCLE. A list of species and the resulting MSA are included in Appendix IV and Appendix V, respectively. A phylogeny was constructed using the neighbor-joining method and the optimal tree had sum of branch length = 3.051. Evolutionary distances were computed using the Dayhoff model (amino acid substitutions per site) and the rates among sites was assumed to be uniform. Gaps and missing data were completely deleted from a dataset that included 116 positions. The

resulting tree, in Appendix VI, was manually edited to better represent the evolution of the species in it. This is a common practice biologists may use when sequence information alone does not give a well-resolved tree. The manually edited tree (Appendix VI) was constructed to match consensus trees in the tree of life as well as those developed by *(31-39)*.

The resulting MSA and phylogeny were input into CoMap. Two CoMap analyses were performed, correlation and compensation. The correlation analyses consisted of five distance measures; grantham, polarity, volume, charge, and unweighted. The compensation analyses included four distance measures; grantham, polarity, volume, and charge. Once a statistic was defined, an R script was run to determine p-values. Only statistically significant results were considered a signal due to coevolution.

3.4 ProTα-Apaf1

The sequence for human Apaf1, UniProt ID O14727, is shown below in Figure 3.3 *(22)*. This was used as a search query in BLAST. Homologs were selected for the same species as the ProTα dataset. These were then added to the ProTα dataset (1-111) and the two were separated by a string of 20 glycines that were manually entered (shown in Figure 3.4). An MSA was performed using MUSCLE and a neighbor-joining tree was constructed. These are included in Appendices VII and VIII, respectively.

MDAKARNCLL QHREALEKDI KTSYIMDHMI SDGFLTISEE EKVRNEPTQQ$_{50}$
QRAAMLIKMI LKKDNDSYVS FYNALLHEGY KDLAALLHDG IPVVSSSSGK$_{100}$
DSVSGITSYV RTVLCEGGVP QRPVVFVTRK KLVNAIQQKL SKLKGEPGWV$_{150}$
TIHGMAGCGK SVLAAEAVRD HSLLEGCFPG GVHWVSVGKQ DKSGLLMKLQ$_{200}$
NLCTRLDQDE SFSQRLPLNI EEAKDRLRIL MLRKHPRSLL ILDDVWDSWV$_{250}$
LKAFDSQCQI LLTTRDKSVT DSVMGPKYVV PVESSLGKEK GLEILSLFVN$_{300}$
MKKADLPEQA HSIIKECKGS PLVVSLIGAL LRDFPNRWEY YLKQLQNKQF$_{350}$
KRIRKSSSYD YEALDEAMSI SVEMLREDIK DYYTDLSILQ KDVKVPTKVL$_{400}$
CILWDMETEE VEDILQEFVN KSLLFCDRNG KSFRYYLHDL QVDFLTEKNC$_{450}$
SQLQDLHKKI ITQFQRYHQP HTLSPDQEDC MYWYNFLAYH MASAKMHKEL$_{500}$
CALMFSLDWI KAKTELVGPA HLIHEFVEYR HILDEKDCAV SENFQEFLSL$_{550}$
NGHLLGRQPF PNIVQLGLCE PETSEVYQQA KLQAKQEVDN GMLYLEWINK$_{600}$
KNITNLSRLV VRPHTDAVYH ACFSEDGQRI ASCGADKTLQ VFKAETGEKL$_{650}$
LEIKAHEDEV LCCAFSTDDR FIATCSVDKK VKIWNSMTGE LVHTYDEHSE$_{700}$
QVNCCHFTNS SHHLLLATGS SDCFLKLWDL NQKECRNTMF GHTNSVNHCR$_{750}$
FSPDDKLLAS CSADGTLKLW DATSANERKS INVKQFFLNL EDPQEDMEVI$_{800}$
VKCCSWSADG ARIMVAAKNK IFLFDIHTSG LLGEIHTGHH STIQYCDFSP$_{850}$
QNHLAVVALS QYCVELWNTD SRSKVADCRG HLSWVHGVMF SPDGSSFLTS$_{900}$
SDDQTIRLWE TKKVCKNSAV MLKQEVDVVF QENEVMVLAV DHIRRLQLIN$_{950}$
GRTGQIDYLT EAQVSCCCLS PHLQYIAFGD ENGAIEILEL VNNRIFQSRF$_{1000}$
QHKKTVWHIQ FTADEKTLIS SSDDAEIQVW NWQLDKCIFL RGHQETVKDF$_{1050}$
RLLKNSRLLS WSFDGTVKVW NIITGNKEKD FVCHQGTVLS CDISHDATKF$_{1100}$
SSTSADKTAK IWSFDLLLPL HELRGHNGCV RCSAFSVDST LLATGDDNGE$_{1150}$
IRIWNVSNGE LLHLCAPLSE EGAATHGGWV TDLCFSPDGK MLISAGGYIK$_{1200}$
WWNVVTGESS QTFYTNGTNL KKIHVSPDFK TYVTVDNLGI LYILQTLE$_{1248}$

Figure 3.3: Primary sequence of Apaf1 (1-1248).

Figure 3.4. A schematic of the ProTα-Apaf1 concatenated sequence is shown above.

The resulting MSA and phylogeny were inputs in the CoMap program. Two CoMap analyses were performed, correlation and compensation. The correlation analyses consisted of five distance measures; grantham, polarity, volume, charge, and unweighted. The compensation analyses included four distance measures; grantham, polarity, volume, and charge. Once a

statistic was defined, an R script was run to determine p-values. Only statistically significant results were considered a signal due to coevolution.

A previous experiment was performed by Qi et al *(11)* that solved residues of ProTα necessary for interaction with Apaf1. Results of the coevolution were scanned for overlap with this experiment. Any matches were further analyzed with molecular docking studies using AMBER. First, an extended structure for ProTα was built in PyMOL by sequentially adding each amino acid. Simulated annealing of ProTα took place in AMBER. This consisted of 936 heating and cooling cycles. The last 50 low energy structures were annealed further until an accepted low energy structure was achieved.

The annealed structure for ProTα and the PDB structure for Apaf1 (3sfz) were opened in PyMOL. The residues detected from coevolution calculations were manually positioned in close proximity, about 13.6 Å. Molecular dynamics were run to see if the two proteins would form a stable complex. Assuming the Born model, hydrogen atoms were removed and 64 sodium ions were added to counter the high net negative charge. The system was then solvated 12 Å using the TIP3BOX water model. Molecular dynamics were then performed for 9 nanoseconds (ns). Six different orientations were examined in order to cover a wide range of orientations of ProTα with Apaf1.

4 Results

4.1 Myoglobin

Myoglobin is the protein responsible for oxygen binding in the muscles. It has a well-defined 3D structure that enables this function. Dutheil et al. *(17),* hereafter, Dutheil, 2007, performed a coevolution study on myoglobin and found coevolving residues in functionally relevant and structurally proximal locations. They obtained their dataset of 100 vertebrate species from Uniprot. This dataset was then aligned with CLUSTALX and a phylogeny was constructed using the PhyML program. The topology was then hand corrected to match well-known branching orders. The MSA and the phylogeny were analyzed for coevolution using the CoMap package. Both the correlation and the compensation methods were utilized. Their results are summarized in Appendix IX.

Detailed examination of the groups detected to be coevolved revealed interesting patterns. For example, Gly65, Leu76, and Thr39 (orange) are close to the heme group along with Phe33 and Leu69 (yellow) (Figure 4.1). Other interesting patterns include Gly150 and Ala94 (blue) that are located close together at the end of adjacent alpha helices, Arg31 and Ser117 (fuscha) that are on the external loop of proximal helices, and Gly121 and Gly122 (pink) that are next to each other in a loop. Not every group of coevolving residues is close to one another, for example Lys96 and Asp20 (purple) exhibit a coevolution signal by charge ($p<0.05$) but are on opposite sides of the structure. These sites are highlighted in the 3D structure of myoglobin (PDB: 1MBD) in Figure 4.1 for the correlation approach and in Figure 4.2 for the compensation approach. The left image shows one orientation of myoglobin while the right image shows an

orientation obtained by rotation about the y-axis. Five different groups are presented and are indicated by different colored boxes around the residue name.

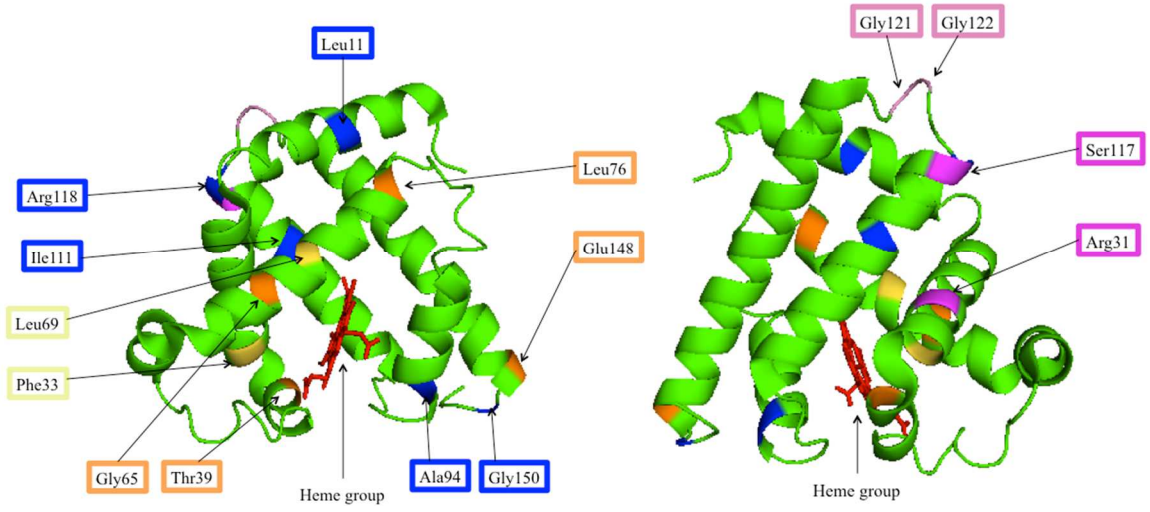

Figure 4.1: Crystal structure of myoglobin with correlation residues detected by Dutheil, 2007 highlighted.

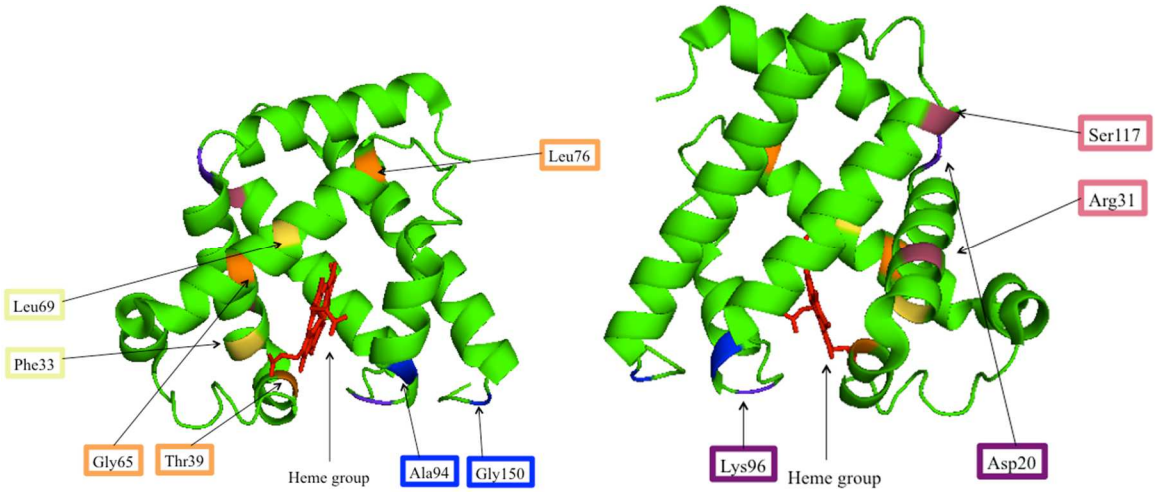

Figure 4.2: Crystal structure of myoglobin with compensation residues detected by Dutheil, 2007 highlighted.

A co-evolution analysis was performed, hereafter referred to as Biscardi, 2014, for myoglobin and results were compared to previous studies. In this dataset the sequences of 74 vertebrate species were downloaded from the BLAST database. The human myoglobin sequence (Figure 2.3) was used as a search query. These sequences were aligned with MUSCLE. The alignment of myoglobin, as seen in Appendix II, resulted in 230 sites with gaps. A phylogenetic tree was constructed from this MSA using the neighbor-joining method. This tree can be found in Appendix III. The tree and its underlying MSA were used to find coevolving groups in the CoMap program using the correlation approach and the compensation approach. Results of these calculations that are significant ($p<0.05$) are displayed in Appendix IX.

Results of Dutheil, 2007 and Biscardi, 2014 are displayed in Appendix IX. Groups of 2 to 10 residues with $p<0.05$ were considered. 38 groups of both correlation and compensation were detected by previous study (Dutheil, 2007). 17 Groups were detected in this study (Biscardi, 2014). Out of the 17 groups detected by Biscardi, 2014; 11 match the results of Dutheil, 2007; a 65% match. Groups detected by present methods include Phe33 and Leu69 ($p<0.05$) and Gly65 and Leu76 ($p<0.05$) that are close to the heme group. Ala94 and Gly150 ($p<0.05$) are at the end of adjacent alpha helices. Asp20 and Lys96 ($p<0.05$) on spatially opposite sides of the protein in both analyses. Most of the sites detected were in functional or structurally proximal regions. Given that the two sets used different pipelines, a different number of species, and different branching orders, it is assumed that the CoMap program is robust enough to detect relevant coevolving residues. For this reason, it was assumed that methods developed in this research would be successful in detecting relevant coevolving positions in other systems such as ProTα.

4.2 ProT-α

The same pipeline mentioned in section 3.2 was applied to ProTα. Sequences were retrieved from the BLAST database using the human ProTα sequence in Figure 4.1 as a search query. A total of 65 vertebrate species (named in Appendix IV) were aligned with MUSCLE and the neighbor joining algorithm was utilized to construct a phylogeny. The phylogeny was manually edited to match a well-known branching order *(31-39).* This is displayed in Appendix VI.

The phylogeny and underlying MSA were used to seek coevolving amino acids. The groups detected are displayed in Table 4.1. Ten total groups were detected; 3 by compensation and 7 by correlation. These regions are mapped on to the sequence in Figure 4.3. The positions detected by compensation, shown in black, seem to span across the entire length of the protein. Some groups detected by correlation also span the entire length of the protein while others line up with significant functional regions. Specifically, they are clustered into three discrete regions, the N-terminus, the central/acidic region, and then C-terminus. The N-terminal residues 1-28, also known as Thymosin-α, are cleaved off and this truncated protein has biological function: Ser10, Glu11, Ile12, and Thr14 ($p < 0.001$), Met1, Ser2, Asp3, Val6, and Ser15 ($p < 0.001$), Thr8 and Ser9 ($p < 0.05$), and Ser2, Asp3, Ser9, and Thr13 ($p < 0.05$). The only groups found by compensation statistics – Arg31 and Glu68; Gln44 and Lys89 ($p < 0.05$ for all groups) – flank the acidic region of ProTα on opposite sides. There are also groups that flank the acidic region that were detected by the correlation method including Gln47 and Asp78; Val23, Ala35, Ala39, Glu58, Gly61, Glu84, Ala84, Gly88, and Glu93. The nuclear localization signal, in the C-terminus, has sites Lys103, Lys104, Gln105, and Lys106 coevolving with $p<0.05$. Interestingly, the caspase cleavage site, DEDDVD, between the central region and the C-terminus was not

detected. This would cut the NLS in two probably letting ProTα leave the nucleus without inhibiting apoptosis.

Table 4.1: The results of ProTα coevolution analysis

PDB	**Analysis**	**Method**	**Statistic**	**p-value**
Arg31, Glu68	Compensation	Charge	0.9960	0.002744
Gln44, Lys89	Compensation	Charge	1	0.02954
Gln44, Lys89	Compensation	Volume	0.9618	0.04634
Ser10, Ile12, Thr14, Glu11	Correlation	Unweighted	0.6435	0.0001666
Met1, Ser2, Asp3, Ser15, Val6	Correlation	Unweighted	0.5730	0.0008277
Thr8, Ser9	Correlation	Unweighted	0.8973	0.003479
Gln47, Asp78	Correlation	Unweighted	0.8900	0.005452
Val23, Gly88, Gly61, Ala35, Ala39, Glu58, Glu93, Thr87, Glu83, Ala84	Correlation	Unweighted	0.3625	0.006240
Lys103, Lys106, Gln105, Lys104	Correlation	Unweighted	0.8127	0.01453
Ser2, Asp3, Ser9, Thr13	Correlation	Grantham	0.5397	0.03195

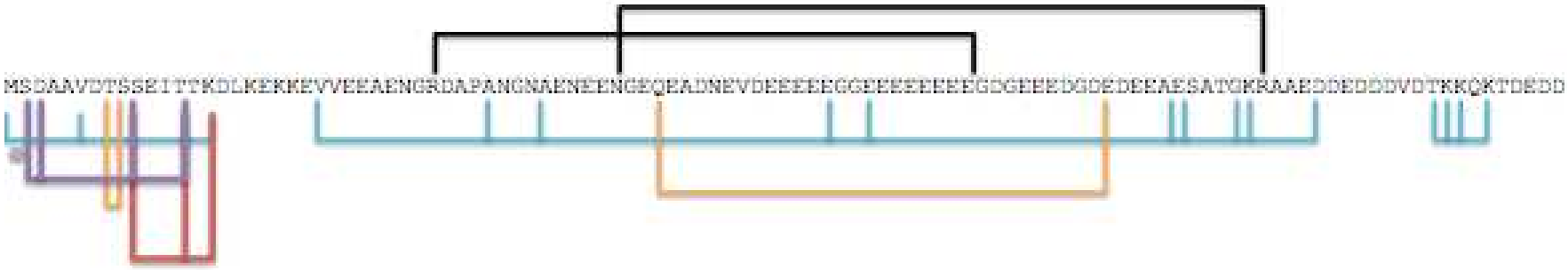

Figure 4.3: Coevolved mutations mapped onto primary sequence of ProTα. The two black brackets on the top represent groups related by compensation. The brackets on the bottom represent groups related by correlation where the colors are used only to separate overlapping groups and make them more visible.

4.3 ProTα-Apaf1

The ProTα data set was combined with Apaf1 to detect coevolving groups between the two proteins. In order to achieve this, an MSA of Apaf1 was developed independently, using the same species as in ProTα, and the two MSA's were linked together. This combined MSA and its corresponding tree were entered in the CoMap program and coevolution was detected using the correlation and the compensation approach. Results are displayed in table 4.2 for the correlation approach and the compensation approach.

Table 4.2: Residues of Apaf1 and ProTα found to have coevolution with each other. Residues highlighted with an asterisk (*) were previously determined by NMR to be necessary for interaction with Apaf1 *(11)*.

Apaf1	**ProTα**	**Method**	**Stat**	**p-value**	**Analysis**
Asp247	Glu55, Glu46	Charge	1	0.0009889	Correlation
Lys198, Val702, Lys252, Lys913	Glu57	Unweighted	0.8105	0.005257	Correlation
Ala53, Leu140, Glu410, Ser268, Leu1242	Glu26, Glu28, Glu64	Volume	0.9988	0.02941	Correlation
Glu283, Gln452, Ser1056, Gly591, Asp1096	**Thr107*, Glu26, Glu28, Thr8***	Charge	0.9979	0.03687	Correlation
Glu14, Tyr109, Ile37, Asp1115, Met55, Thr837, Val1134	Met1, Ser2	Grantham	0.8531	0.004838	Compensation
Thr112, Lys1036, Cys967, Phe1114	Val24	Grantham	0.8007	0.02558	Compensation
His183, Asp207, Gln982, Glu221, Gly880, Ser1056	Gly72, Glu84, Glu26	Volume	0.8752	0.02684	Compensation
Tyr24	Glu58	Polarity	0.9257	0.03779	Compensation
Gln219, Cys967, Leu882	Gln38, Glu68, Gly72	Polarity	0.9042	0.03894	Compensation
Asp32, Ser1210, Leu196, Gly983, Met374, Val660, Val885, Val282, Lys 1222	Glu40	Volume	0.8725	0.04176	Compensation
Glu41, Gln1044, Ser104, Ser829, Lys398	Asn41	Charge	0.7864	0.04207	Compensation
Leu1242	Glu26	Polarity	0.9677	0.04230	Compensation
Ser549	Gly70	Polarity	0.9605	0.04959	Compensation

Previous NMR experiments by Qi et al. *(11)* revealed residues on ProTα necessary for interaction with Apaf1. These residues, highlighted in Figure 4.4, include Ala4, Thr8, Ser9, Glu11, Thr13, Thr14, Arg31, Ala35, Gln43, Ser84, Thr86, Gly87, and Thr106.

MSDAAVDTSS EITTKDLKEK KEVVEEAENG RDAPANGNAN EENGEQEADN EVDEEEEEGG
EEEEEEEEGD GEEEDGDEDE EAESATGKRA AEDDEDDDVD TKKQKTDEDD

Figure 4.4: Residues on ProTα necessary for interaction with Apaf1 highlighted in red

Out of these 13 residues, 2 were found to have coevolution with Apaf1 using methods described. ProTα residues Thr8 and Thr106 combined with residues Glu26 and Glu28 resulting in coevolution with Apaf1 residues Glu283, Gln452, Ser1056, Gly591, and Asp1096 with $p<0.05$. Figure 4.5 shows residues of ProTα found to have a correlated signal with Apaf1 highlighted in orange Figure 4.6 shows the cartoon structure of Apaf1 with residues found to have a correlated signal with ProTα highlighted in blue. The crystal structure of Apaf1 (3SFZ) was downloaded from Protein Data Bank (PDB) *(40)*.

MSDAAVDTSS EITTKDLKEK KEVVEEAENG RDAPANGNAN EENGEQEADN EVDEEEEEGG
EEEEEEEEGD GEEEDGDEDE EAESATGKRA AEDDEDDDVD TKKQKTDEDD

Figure 4.5: Residues on ProTα found to have coevolution with Apaf1 highlighted in orange.

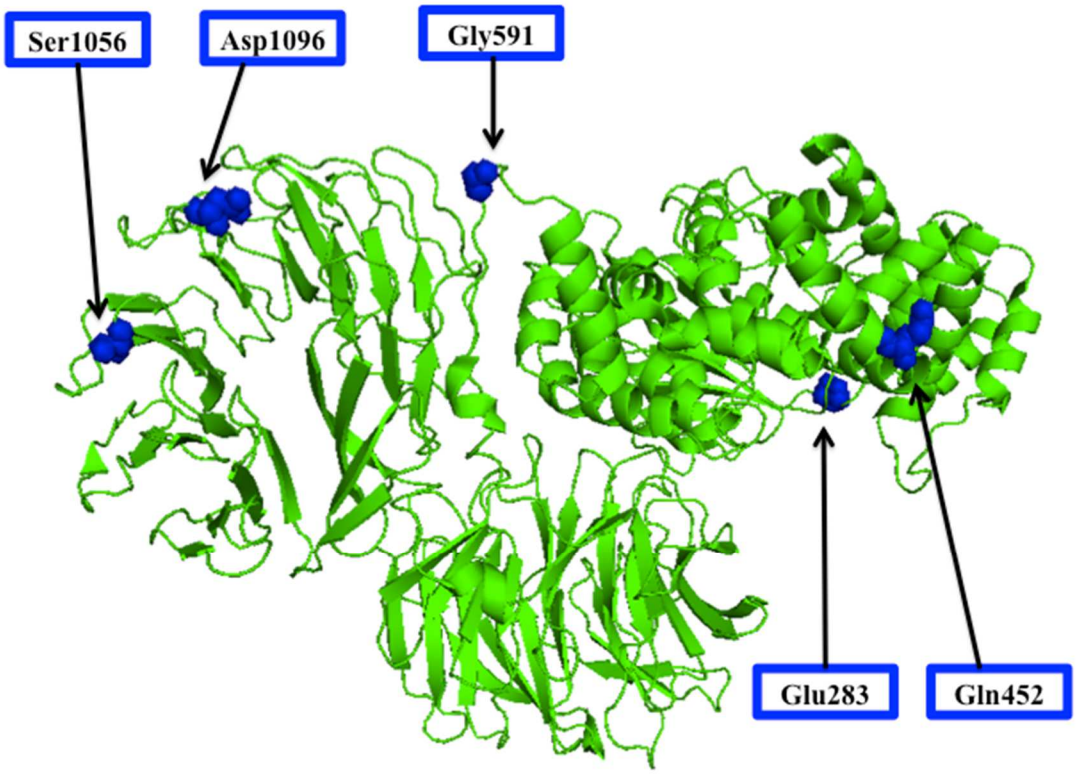

Figure 4.6: Apaf1 cartoon structure (pdb: 3sfz) with residues found to have coevolution with ProTα labeled and shown as blue spheres.

These residues are spread all across Apaf1, however they are all in solvent exposed regions. Two residues in particular, Ser1056 and Asp1096, are located on the active site of a β-propeller, a feature known for protein binding. These two Apaf1 residues and ProTα residues Thr7 and Thr106 were analyzed by molecular docking for binding activity. The starting structure highlighting Apaf1 residues Ser1056 and Asp1096 and ProTα residues Thr7 and Thr106 is shown in Figure 4.7. Several other starting structures were utilized and are summarized in Table 4.3.

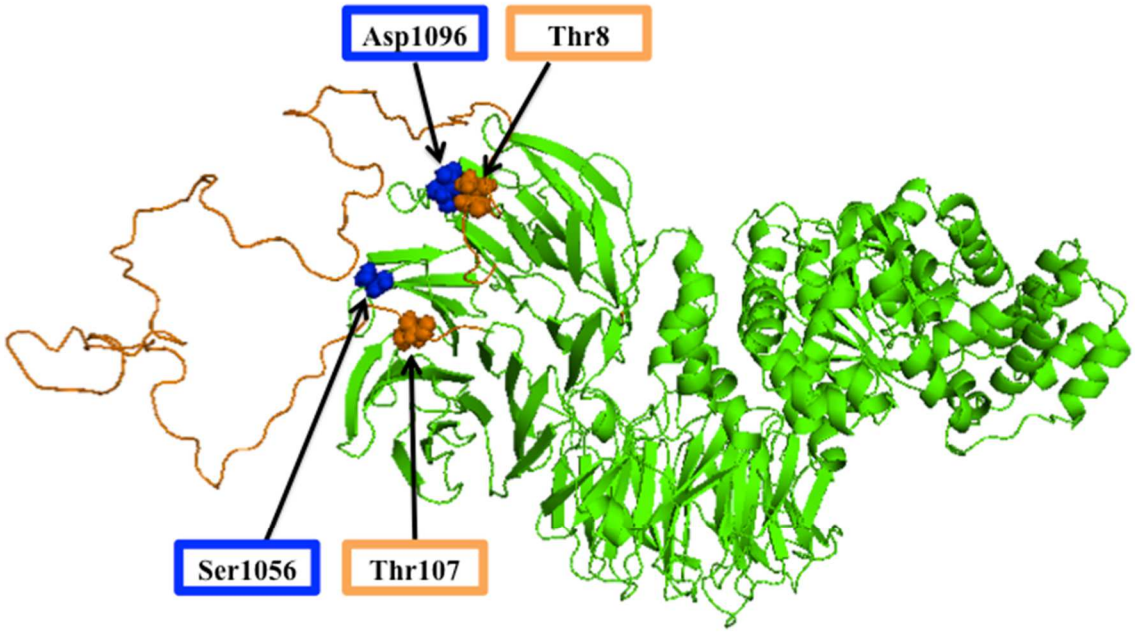

Figure 4.7: The starting structure for molecular docking studies consists of ProTα in orange and Apaf1 in green. In order to investigate binding activity between the two proteins, certain residues were brought within 13.6 Å of one another. These residues include Apaf1 Ser1056 and Asp1096 and ProTα Thr7 and Thr106.

Table 4.3: Summary of results for ProTα-Apaf1 complex. Orientations of ProTα to Apaf1 analyzed are depicted in Appendix X.

Orientation analyzed	ProTα residue	Apaf1 residue	Bonds or forces present	Distance (Å)	Figure (Appendix X)
A	Thr8	Asp1138	Polar	2.5	8.1a
A	Ser9	Leu1137	Polar	1.9	8.1a
A	Thr107	Phe1050	Van der Waals	< 4	8.1b
A	Thr108	Gln1010	Polar	2.5	8.1b
A,y	Thr8	Pro1221	Van der Waals	< 4	8.2a
A,y	Ser9	Pro1181	Van der Waals	< 4	8.2a
A,y	Asp108	Ser1056*	Polar	1.8	8.2b
A,-y	N/A	N/A	N/A	N/A	N/A*
B	Thr8	Thr1012	Van der Waals	< 4	8.3a
B	Ile12	Thr1012	Polar	1.9	8.3a
B	Asp108	Gln1054	Polar	2.1	8.3b
B,y	Thr8	Thr1012	Polar	1.9	8.4a
B,y	Thr13	Thr1012	Polar	2.9	8.4a
B,y	Thr13	Ala1011	Polar	1.9	8.4a
B,y	Thr107	Leu1137	Van der Waals		8.4b
B,y	Asp108	Ser1095	Polar	1.8	8.4b
B,-y	Thr8	Thr1012, Ala1013	Van der Waals	< 4	8.5a
B,-y	Thr107	Cys1178, Val1219, Phe1223	Van der Waals	< 4	8.5b

*Orientation did not result in any interactions within 4 Å.

5 Discussion

In this work, sites were considered coevolving when they have undergone unweighted or biochemically relevant substitutions in the same branches of a tree. These methods were applied to three different subjects: a folded protein myoglobin, an unfolded protein prothymosin-α, and the ProTα-Apaf1 complex. ProTα has not been previously analyzed by coevolution. CoMap, the program utilized in this study, has not been used previously to analyze a complex. Myoglobin has been previously analyzed for coevolution by several authors including the creators of CoMap. Further, this protein has a well-established 3D structure and has been the subject of relating coevolving residues to protein folding. For these reasons, it was analyzed to establish a pipeline in ECU's CACS and demonstrate effective ability to use this pipeline.

5.1 Myoglobin

Myoglobin was analyzed the same way as in Dutheil and Galtier's previous work *(11)*. The same correlation, biochemically weighted correlation, and compensation statistics were applied. However, the sequence alignment and tree used for these statistics were intentionally different. Homologous sequences for myoglobin were found in the BLAST refseq_protein database. Sequences from this database are constantly curated by NCBI staff and collaborators and are high quality, well annotated, and non-redundant. SwissProt, the database used by Dutheil, 2007, is also a high quality annotated and non-redundant protein sequence database. While integrity of the sequences is theoretically comparable between the two databases, they may vary in the species offered. Coevolution statistics greatly depend on sequence diversity, which may be different between the two databases. For example, in the myoglobin dataset of 75 species, 9 of those species are primates whose genetic makeup are

>90% similar. Having extra primates in a dataset does not do much for providing sequence diversity except for the fact that it increases the actual size of the dataset. Having a larger dataset would increase the ability of coevolution statistics to detect relevant positions, especially in a highly conserved protein such as myoglobin. The fact that the Dutheil dataset contained 100 species helps increase the predictive power of the statistics. These are reasons that results slightly differ between myoglobin analyzed in this study and myoglobin analyzed by Dutheil. Considering a less diverse dataset, smaller dataset (including 25 fewer species), a 65% match to previous work is enough evidence to suggest a pipeline for studying coevolution was effectively developed and utilized.

Eleven groups of coevolving positions were found that match previous work by Dutheil, 2007. A match between Dutheil, 2007 and Biscardi, 2014 was identified as having at least 2 positions in common. Many of the matching groups have a different number of positions. For example, in Dutheil's work, Lys56 and Tyr103 were determined as coevolving according to the compensation method. However, in present work, Lys56 and Tyr103 were found coevolving with 4 other positions, Glu136, Asp126, Lys78, and Asp122. Out of these 11 groups, all have matching statistic (correlation/compensation), and 10 out of 11 have matching hypotheses (biochemically weighted/unweighted).

The group that does not have the same biochemical weight is Thr39 and Gly65. This pair was detected in this work by applying the Grantham method while Dutheil's work detected this pair by applying the volume method. Interestingly, this pair is located close to the heme group suggesting a functional correlation. Another group detected by Dutheil, 2007 that is close to the heme group is Phe33 and Leu76. This pair was detected by four different methods: unweighted

correlation, Grantham compensation, volume compensation, and polarity compensation. This pair was not detected by my work at all which may be attributed to a lack of diversity in my dataset. Looking at the sequence alignment in Appendix II This pair is almost completely conserved except in *Echinops telfairi*, where there is a gap from residues 1-55, and *Alligator mississippiensis*, where Phe33 and Leu69 are switched to Leu33 and Phe69. While a switch like that is indicative of coevolution due to structural contact or functional constraint, since it only occurs in one species it is probably not a strong enough signal to detect with these statistics. Another group of residues close to the heme group was detected by both previous studies and this present study. Dutheil, 2007 detected Thr39, Gly65, Leu76, and Glu148 to be coevolving by volume correlation. Groups detected presently include Thr39, Glu6, Leu76, Lys63, Glu59, Glu148, Lys133, Asp53, His81, and Lys16. Thr39 and Leu76, two exact matches, are inside the heme-binding pocket. Glu148, another exact match appears to be in an exposed position on the C-terminus. The rest of the positions detected presently (but not by Dutheil, 2007) are in exposed regions and spread out along the entire protein. These positions are highlighted in Figure 5.1. While it is logical that several sites in a protein may be correlated, especially by volume, this doesn't help elucidate either protein structure or function. If more species were utilized in this study, a better signal associated with the heme pocket would probably result.

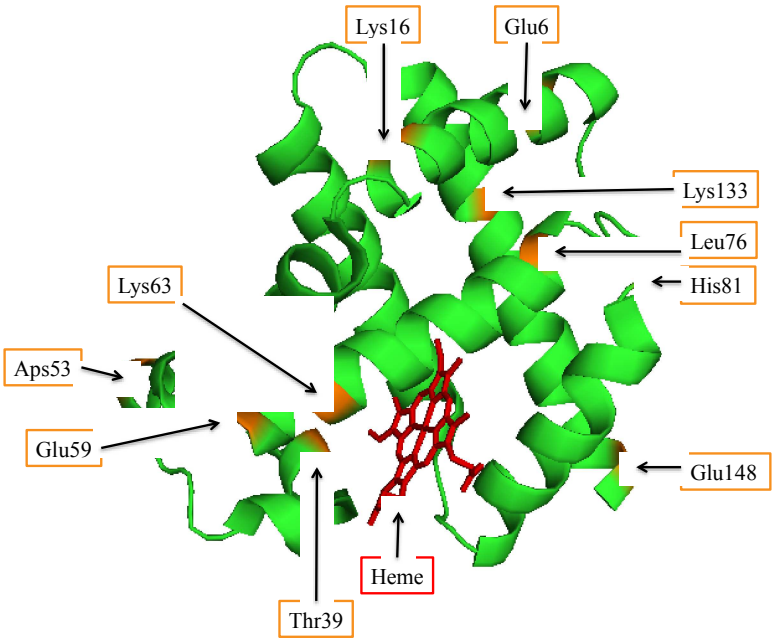

Figure 5.1: Several positions detected on myoglobin may be involved with the heme binding function.

5.2 ProTα

These methods were applied to ProTα; the results of which are displayed in Table 4.1 and Figure 4.3 Interestingly, only two groups were identified by compensation statistics. Arg31 and Glu68 were detected by applying the charge biochemical weight ($p<0.05$) and Gln44 and Lys89 were detected by charge ($p<0.05$) and volume ($p<0.05$). Since compensation methods are typically indicative of structural contacts, it is logical that an unfolded protein would have very few. The crystal structure of ProTα with Keap1 shows ProTα adopts a hairpin structure with a fold around residues 40-48 *(41)*. It is logical that residues flanking this region, like Arg31 and Glu68, are important contact sites.

Csizmók et al.*(42)* employed limited proteolysis, CD spectroscopy, and solid state 1H NMR to show short and long range structural organization in two IDP's: microtubule-associated

protein 2c (MAP2c) and the first inhibitory domain of calpastatin (CSD1). IDPs are notorious for having several cellular functions and a featureless structure, such as random coil. Because of their rapid and specific binding to their partners, residual structure may play a crucial role in their function. Binding and induced folding would take too long if the IDP had to search through all of conformational space from a fully unstructured initial state. In CSD1, it was found that long-range tertiary interactions were present. For MAP2c, regions of local structural constraints were found. If long range or transient interactions are present in ProTα, compensation statistics would be the way of detecting this.

The remaining 7 groups detected in this study were found by correlation statistics. It is logical that so many more groups were detected by correlation rather than compensation in full length IDP, ProTα. Interestingly, these groups are divided into 3 discrete regions: N-terminus, C-terminus, and central acidic.

The N-terminal region of ProTα is cleaved by asparaginyl endopeptidase at residues 1-28. The resulting truncated protein is known as Thymosin-α1. Thymosin-α1 has its own established biological function and it currently manufactured as a commercial drug, Zadaxin, with proven anti-cancer ability and utilization in treatment of hepatitis B and C. *(43)*. Literature suggests that this portion of the protein is highly conserved *(44)*. Four N-terminal groups were detected in this work by correlation; three by the unweighted calculation ($p<0.001$) and one by grantham ($p<0.05$). Restrained molecular dynamic simulations with an explicit solvent box containing 40% TFE/60% TIP3P water (v/v) were used in order to obtain 3D model of NMR structure of Thymosin-α1 (1-28). Results suggest the peptide adopts a structured conformation with two stable regions, an α-helix from residues 14-26 and two double β-turns in residues 1-12 *(43*).

Since NMR data suggest more secondary structural propensities for thymosin-α1, it makes sense that almost half of the coevolution signals detected belong to that region of full length ProTα.

The C-terminal region of ProTα is known to have extracellular immunomodulatory activity *(45)*. In cells undergoing apoptosis, cleavage by caspases generates a C-terminal truncated polypeptide, amino acids 1-99, which sequesters cytochrome c. This region is released into the cytosol upon cleavage of its C-terminal bipartite nuclear localization signal. One region was detected in this study by unweighted correlation statistics. These residues, Lys103, Lys104, Gln105, Lys106, correspond to part of the nuclear localization signal (KKQK). Further, these are part of a decapeptide, TKKQKTDEDD, which was identified as a potent lymphocyte stimulator. Attenuated total reflectance Fourier-transform infrared spectroscopy and negative staining transmission electron microscopy suggest that this decapeptide adopts an anti-parallel beta sheet conformation under various conditions.

The final region of ProTα is the central acidic region. This region spans from approximately residues 50-89 and consists of several glutamic and aspartic acid residues. This region is associated with the histone-binding and proliferative activity of ProTα. One would expect this region to have no or highly transient secondary structure due to its highly repulsive nature.

5.3 ProTα-Apaf1

The interaction of ProTα and binding partner, Apaf1, was also investigated in this study. This is a known direct interaction from prior pull-down experiments. Further, the region of ProTα necessary for interaction with Apaf-1 have been characterized using 2-dimensional ^{1}H-^{15}N HSQC-NMR with ^{15}N-labeled ProTα in the presence of Apaf1. Results show that peaks

corresponding to A4, T8, S9, E11, T13, T14, R31, A35, S84, T86, G87, and T106 in ProTα had reduced signal compared to the rest of the peaks or have been broadened beyond detection. Also, E108 and D110 display a large chemical shift change. Because of a narrow chemical shift dispersion and large peak overlap, a full analysis of the full ProTα binding interface was difficult. These data suggest that regions 4-14, 31-34, 84-87, and 106-110 are necessary for interactions with Apaf1. Since pull-down data show that ProTα and Apaf1 are interacting partners they must have coevolved. Given the NMR data which show exactly which residues of ProTα are necessary to interact, coevolution detection methods can be used to find regions of Apaf1 necessary for interaction with ProTα.

My analysis was applied to a concatenated Apaf1-ProTα sequence and 13 groups were detected, 4 by correlation and 9 by compensation. Of these, one group including Glu283, Gln452, Ser1056, Gly591, and Asp1096 from Apaf1 and Thr107, Glu26, Glu28, and Thr8 from ProTα detected by correlation ($p<0.05$) was selected for molecular docking studies. This is because residues Thr107 and Thr8 of ProTα correspond to regions previously identified as necessary for interaction with Apaf1 and because Ser1056 and Asp1096 of Apaf1 are in the active site of a Beta-propeller. These four residues (ProTα: Thr107, Thr8 and Apaf1: Ser1056, Asp1096) were utilized to determine intermolecular interactions between ProTα and Apaf1. Five starting structures were built to account for some of the different orientations by which ProTα and Apaf1 could possibly bind. Molecular dynamics were simulated and in all 5 starting structures resulted in an interaction between the two proteins with a distance of less than 4 Å. These interactions were rarely between the residues detected by coevolution calculations but were usually nearby residues. Apaf1 residue Asp1096 did not interact with any of the prothymosin-α residues in any of the orientations simulated. Apaf1 residue Ser1056 only had an

interaction with ProTα residue Asp108 in one orientation but with the strongest force and shortest distance (hydrogen bond and 1.8 Å).

Although the exact residues of Apaf1 necessary for interaction with ProTα were not precisely identified from these experiments, the two proteins did not drift apart and the root-mean-square deviation (RMSD) for each residue remained stable for 9 ns of molecular dynamics. The active site of Apaf1's β-propeller may actually be where ProTα binds to inhibit apoptosome formation for the anti-apoptotsis pathway. A similar interaction was seen between ProTα and Keap1. Studies by Cino et al. *(46)* show that ProTα competes with Neh2 to bind a 6-bladed Beta-propeller of Keap1. The Keap1-binding motif of ProTα (-NEENGE-) shares a similar sequence to that of Neh2 (-DEETGE-). Crystal structures of ProTα and Neh2 peptides bound to the Kelch domain of Keap1 reveal that these two bind to similar sites *(41)*.

6 Conclusions

1. Coevolution of myoglobin. A pipeline for performing evolutionary studies was established at ECU's CACS which allows for use of the program CoMap for detected correlated changes in protein sequences by statistical methods (Figure 2.1). The robustness of the methodology was tested using myoglobin, as it has been a target of coevolutionary studies by other research groups. Myoglobin's structure is well established allowing coevolutionary findings to be better placed in context. Although a smaller set of species was used in the coevolutionary study here, the results matched well with those from the literature.

2. Coevolution and prothymosin-α. These methods have not been applied to proteins classified as IDPs, which are proteins that typically contain little or no secondary structure under physiological conditions (at least in the absence of a binding partner). The highly dynamic nature of IDPs makes detection of long range interactions difficult although these may be functionally relevant. The use of coevolution may better direct experimental studies which may be time consuming compared to calculations performed in this work. The calculations performed, as in Figure 2.1, took under 5 seconds – a significantly shorter time than experimental techniques such as expressing mutants and testing them. For instance, the role of key residues identified by coevolution studies can be examined by mutation studies or the incorporation of specific labels for enhanced spectroscopic studies. These could be used in better exploring the ProTα-Keap1 interaction and function. For example, hypotheses can be made that mutating specific residues (give example) would lead to a decrease in binding affinity or reduced biological activity.

3. Coevolution between ProTα and Apaf1. Identification of the binding regions between interacting partners is essential for understanding how they work and for understanding how biochemical pathways function (for example ProTα in Figure 1.2). The experimental determination of which proteins interact and which regions lead to specific interactions between them is time consuming. Coevolutionary studies may provide a means of better directing experimental studies if key residues can be identified computationally.

References

1. Berg, J. M.; Tymoczko, J. L.; Lubert, S. *Biochemistry*; 7th ed.; W.H. Freeman: New York, 2012.

2. Uversky, V. N.; Gillespie, J.R.; Fink, A.L. Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins: structure, function, and bioinformatics* **2000**, *41,* 415-427.

3. Dunker, A. K.; Brown, C.J.; Lawson, J.D.; Iakoucheva, L.M.; Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **2002**, *41,* 6573-6582.

4. Uversky, V. N. Intrinsically disordered proteins from A to Z. *Int. J. Biochem. Cell Biol.* **2011**, *43*, 1090-1103.

5. Uversky, V. N. Natively unfolded proteins: A point where biology waits for physics. *Prot. Sci.* **2002**, *11,* 739-756.

6. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; Kim, P.M.; Kriwacki, R.W.; Oldfield, C.J.; Pappu, R.V.; Tompa, P.; Uversky, V.N.; Wright, P.E.; Babu, M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014,** *114,* 6589-6631.

7. Pineiro, A.; Cordero, O. J.; Nogueira, M. Fifteen years of prothymosin alpha: contradictory past and new horizons. *Peptides* **2000**, *21,* 1433-1446.

8. Ueda, H.; Matsunaga, H.; Halder, S. K. Prothymosin alpha plays multifunctional cell robustness roles in genomic, epigenetic, and nongenomic mechanisms. *Thymosins in Health and Disease* **2012,** *1269,* 34-43.

9. Mosoian, A. Intracellular and extracellular cytokine-like functions of prothymosin alpha: implications for the development of immunotherapies. *Future Med. Chem*. **2011**, *3,* 1199-1208.

10. Karapetian, R. N.; Evstafieva, A.G.; Abaeva, I.S.; Chichkova, N.V.; Filonov, G.S.; Rubtsov, Y.P.; Sukhacheva, E.A.; Melnikov, S.V.; Schneider, U.; Wanker, E.E.; Vartapetian, A.B. Nuclear Oncoprotein Prothymosin Is a Partner of Keap1: Implications for Expression of Oxidative Stress-Protecting Genes. *Mol. Cell. Biol.* **2005**, *25,* 1089-1099.

11. 1. Qi, X.; Wang, L.; Du, F. Novel Small Molecules Relieve Prothymosin α-Mediated Inhibition of Apoptosome Formation by Blocking Its Interaction with Apaf-1. *Biochemistry* **2010,** *49,* 1923-1930.

12. Dong, G.; Callegari, E.A.; Gloeckner, C.J.; Ueffing, M.; Wang, H. Prothymosin-α Interacts with Mutant Huntingtin and Suppresses Its Cytotoxicity in Cell Culture. *J. Biol. Chem.* **2012**, *287,* 1279-1289.

13. Mosoian, A. Intracellular and extracellular cytokine-like functions of prothymosin alpha: implications for the development of immunotherapies. *Future Med. Chem.* **2011**, *3,* 1199-1208.

14. Dutheil, J. Y. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief. Bioinform.* **2011**, *13,* 228-243.

15. Pazos, F.; Valencia, A. Protein co-evolution, co-adaptation and interactions. *EMBO J.* **2008***, 27*, 2648-2655.

16. Dutheil, J.; Pupko, T.; Jean-Marie, A.; Galtier, N. A Model-Based Approach for Detecting Coevolving Positions in a Molecule. *Mol. Biol. Evol.* **2005***, 22,* 1919-1928.

17. Dutheil, J. and Galtier, N. Detecting groups of co-evolving positions in a molecule: a clustering approach. *BMC Evol. Biol.* **2007***, 7*, 242.

18. Freeman, S., *Biological Science,* 3rd ed.; Pearson: San Fransisco, 2008.

19. de Juan, D.; Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **2013***, 14*, 249-261.

20. Hubbard, S. R.; Hendrickson, W. A.; Lambright, D. G.; Boxer, S. G. X-ray crystal structure of a recombinant human myoglobin mutant at 2·8 Å resolution. *J. Mol. Biol.* **1990***, 213*, 215-218.

21. Watson, H. C. The stereochemistry of the protein myoglobin. *Prog. Stereochem.* **1969***, 4*, 299.

22. Universal Protein Resource. http://www.uniprot.org (accessed October 1, 2014).

23. Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* **1992***, 89*, 10915-10919.

24. National Center for Biotechnology Information. Basic Local Alignment Search Tool. http://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed October 1, 2014).

25. Nuin, P.A.; Wang, Z.; Tiller, E.R. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* **2006***, 7*, 471.

26. Edgar, R. C. MUSCLE: mutiple sequence alignment with improved accuracy and speed. *BMC Bioinformatics.* **2004***, 32(5)*, 1792.

27. Hall, B. G., *Phylogenetic Trees Made Easy,* 4th ed.; Sinauer Associates, Inc.: Sunderland, MA, 2011

28. Horner, D. S.; Pirovano, W.; Pesole, G. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief. Bioinform.* **2008***, 9*, 46-56.

29. Leach, A. R., *Molecular Modeling Principles and Applications,* 2nd ed.; Pearson Education: Essex, 2001.

30. Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol.* **2011***, 28 (10)*, 2731.

31. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In "Atlas of Protein Sequence and Structure, vol. 5, suppl. 3." M.O. Dayhoff (ed.), pp. 345-352, *Natl. Biomed. Res. Found*., Washington, DC.

32. Song, S.; Liu, L.; Edwards, S.V.; Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A* **2012***, 109*, 14942-14947.

33. Murphy, W. J.; Eizirik, E.; O'Brien, S.J.; Madsen, O.; Scally, M.; Douady, C.J.; Teeling, E.; Ryder, O.A.; Stanhope, M.J.; de Jong, W.W.; Springer, M.S. Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science (New York, N.Y.)* **2001***, 294*, 2348-2351.

34. Amemiya, C. T.; Alföldi,J.; Lee, A.P.; Fan, S.; Phillippe, H.; MacCallum, I.; Braasch, I.; Manousaki, T.; Schneider, I.; Rohner, N.; Organ, C.; Chalopin, D.; Smith, J.J.; Robinson, M.; Dorrington, R.A.; Gerdol, M.; Aken, B.; Biscotti, M.A.; Barucca, M.; Baurain, D.; Berlin, A.M.; Blatch, G.L.; Buonocore, F.; Burmester, T.; Campbell, M.S., *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature (London)* **2013***, 496*, 311-316.

35. Alföldi, J.; Di Palma, F.; Grabherr, M.; Williams, C.; Kong, L.; Mauceli, E.; Russell, P.; Lowe, C.B.; Glor, R.E.; Jaffe, J.D.; Ray, D.A.; Boissinot, S.; Shedlock, A.M.; Botka, C.; Castoe, T.A.. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature (London)* **2011***, 477*, 587-591.

36. Smith, J. J.; Kuraku, S.; Holt, C.; Sauku-Spengler, T.; Jiang, N.; Campbell, M.S.; Yandell, M.D.; Manousaki, T.; Meyer, A.; Bloom, O.E.; Morgan, J.R.; Buxbaum, J.D.; Sachidanandam, R.; Sims, C.; Garruss, A.S.; Cook, M.; Krumlauf, R.; Wiedermann, L.M.; Sower, S.A.; Decatur, W.A.; Hall, J.A.; Amemiya, C.T.; Saha, N.R.; Buckley, K.M.; Rast, J.R.; *et al.* Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat. Genet.* **2013***, 45*, 415-421.

37. Shultz, S.; Opie, C.; Atkinson, Q.D. Stepwise evolution of stable sociality in primates. *Nature (London)* **2011***, 479*, 219-222.

38. James, D. E.; Organ, C. L.; Fujita, M. K.; Shedlock, A. M.; Edwards, S. V. Genome Evolution in Reptilia, the Sister Group of Mammals. *Annu. Rev. Genom. Human Genet.* **2010***, 11*, 239-264.

39. Blanga-Kanfi, S.; Miranda, H.; Penn, O.; Pupko, T.; DeBry, R. W.; Huchon, D. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evolutionary Biology* **2009***, 9: 71*, 1471.

40. Reubold, T. F.; Wohlgemuth, S.; Eschenburg, S. Crystal structure of full-length Apaf-1: how the death signal is relayed in the mitochondrial pathway of apoptosis. *Structure* **2011***, 19(8)*, 1074.

41. Padmanabhan, B.; Nakamura, Y.; Yokoyama, S. Structural analysis of the complex of Keap1 with a prothymosin α peptide. *Acta. Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **2008***, 64*, 233-238.

42. Csizmók, V.; Bokor, M.; Bánki, P.; Klement, E; Medzihradszky, K.F.; Friedrich, P.; Tompa, K.; Tompa, P. Primary Contact Sites in Intrinsically Unstructured Proteins: The Case of Calpastatin and Microtubule-Associated Protein 2. *Biochemistry (Easton)* **2005***, 44*, 3955-3964.

43. Elizondo-Riojas, M.A.; Chamow, S.M.; Tuthill, C.W.; Gorenstein, D.G.; Volk, D.E. NMR structure of human thymosin alpha-1. *Biochem. Biophys. Res. Commun.* **12***, 416*, 356-361.

44. Goldstein, A. L.; Goldstein, A. L. From lab to bedside: emerging clinical applications of thymosin α1. *Expert Opin. Biol. Ther.* **2009***, 9*, 593-608.

45. Skopeliti, M.; Iconomidou, V.A.; Derhovanessian, E.; Pawelec, G.; Voetler, W.; Kalbacher, H.; Hamodrakas, S.J.; Tsitsilonis, O.E. Prothymosin α immunoactive carboxyl-terminal peptide

TKKQKTDEDD stimulates lymphocyte reactions, induces dendritic cell maturation and adopts a β-sheet conformation in a sequence-specific manner. *Mol. Immunol.* **2009***, 46*, 784-792.

46. Cino, E. A.; Wong-ekkabut, J.; Karttunen, M.; Choy, W. Microsecond Molecular Dynamics Simulations of Intrinsically Disordered Proteins Involved in the Oxidative Stress Response. *PloS one* **2011***, 6*, e27371.