2012

Evolution of Interactomes

Monica Micek
Loyola University Chicago

LOYOLA UNIVERSITY CHICAGO

EVOLUTION OF INTERACTOMES

A THESIS SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

MASTER OF SCIENCE

PROGRAM IN BIOLOGY

BY

MONICA MICEK

CHICAGO, IL

DECEMBER 2012

ACKNOWLEDGEMENTS

I am indebted in the preparation of this thesis to my advisor, Dr. Catherine Putonti, whose patience and kindnesses, as well as her academic experience, have been invaluable to me. She has been tireless in her willingness to read and critique my writing. I am also extremely grateful to my committee, Dr. Terry Grande and Dr. Eric Schroeter for agreeing to serve on my thesis committee and for their support and efforts in reviewing my work. I am also grateful to both Dr. Richard Fay and Dr. Domenic Castignetti for allowing me to work and gain research experience in their laboratories. It is an honor to have worked and learned from them. I would also like to thank Paula Martin for her support and kindness. She has been and continues to be a constant source of inspiration and motivation. Finally, I would like to thank all the faculty, staff, students and friends I have met at Loyola University Chicago. The informal support and encouragement of many friends has been indispensable. Special thanks are due to my husband, David Korsak, for his understanding, endless patience and encouragement when it was most required.

For my parents, Jolanta and Krzysztof Micek.

I have no dress except the one I wear every day. If you are going to be kind enough to give me one, please let it be practical and dark so that I can put it on afterwards to go to the laboratory.

Marie Sklodowski Curie

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Protein-protein interactions are part of all biological processes and are responsible for directing the development and maintenance of all systems in a species. Identifying such interactions provides insight into molecular processes in addition to their importance in understanding disease. Identifying protein-protein interactions experimentally is expensive, both in terms of cost and effort, and can generate erroneous results. Thus computational methods are key in reducing the scope of experimental assays, providing predictions for subsequent verification. Herein I present a new computational tool for the prediction of protein-protein interactions which, looking at sequence data alone, can identify putative interacting proteins as a result of their coordinated evolution. This new approach builds on previous molecular evolutionary methods and combines evolutionary information from individual proteins. As a proof of concept, the new approach was tested on the well-studied interaction networks of the visual and auditory systems. From this analysis, several protein clusters were identified warranting further experimental investigation. Furthermore, this effort also identified areas for future refinement of the software tool.

CHAPTER ONE

INTRODUCTION

Protein-Protein Interactions (PPI) are part of all biological processes and are responsible for orchestrating and directing the development and maintenance of all systems in a species. A PPI network is the collection of the interactions between proteins in a single organism, also called an interactome (Figure 1).
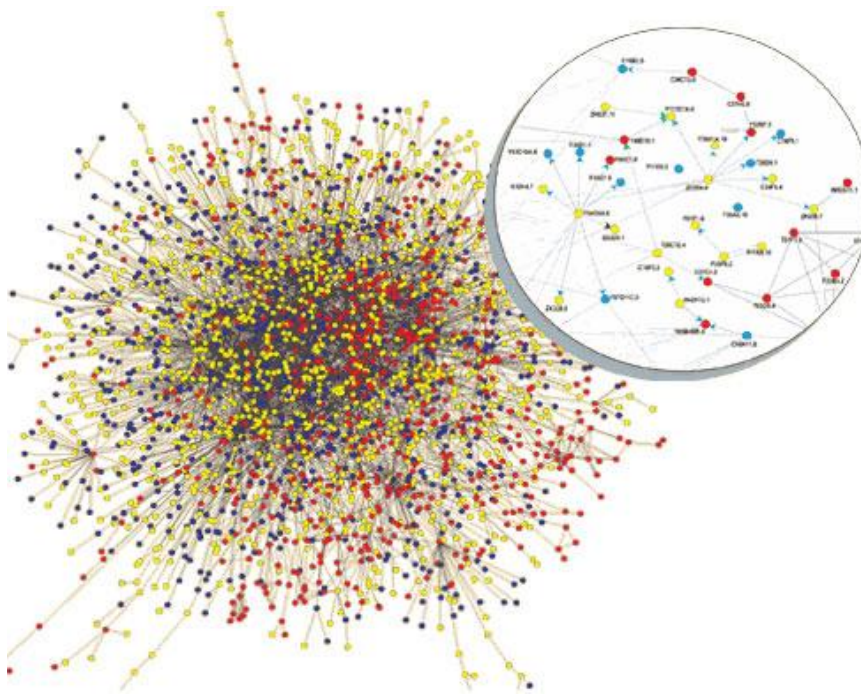
Figure 1. Current *Caenorhabditis elegans* interactome. [Reproduced from Perkel, 2004]

For the vast majority of species, the complex dynamics of the interactome have yet to be fully actualized as the dynamic network of PPI changes under different stimuli as well as periods of development. Moreover, the PPI networks have evolved over the

1

course of time, specifically fine-tuned for an individual species. Investigating the evolution of such networks is complex, particularly since a complete record of all of the PPI does not exist for any extant organism, although some such as *Saccharomyces cerevisiae* are much more complete than others. There are many different experimental as well as computational methods for inferring PPI; however, both accuracy as well as the number of proteins these methods can investigate is limited. A means to integrate these two approaches and build reliable PPI networks of different biological systems that can be combined into one extensive interactome is greatly needed.

The primary aim of this thesis is to develop a tool to elucidate the evolution of protein interactions making it possible to more readily identify putative interactions computationally for further experimental investigation. This new method uses the premise that PPI networks can be visualized as undirected graphs or diagrams that can clearly and quickly show the overall view of the biological system. Consequently, the method for inferring PPI presented here builds upon existing phylogenetic methods to visualize relationships between proteins that belong to the same network. As a proof-of-concept, the well-studied proteins involved in the development and function of the visual and auditory system will be investigated.

Protein-Protein Interactions

PPI are complex, dynamic systems controlling all biological functions. Proteins are built of amino acids whose chemical properties such as polarity and side chain charge determine the biological activity of the protein (Baldwin & Lapointe, 2003). As follows,

proteins are involved in all cell functions and biological systems. Proteins have many diverse functions, e.g. catalysts, antibodies, hormones, transport, structure, storage and contractile proteins, etc., and carry out many important tasks such as cell growth, gene expression, intracellular communication, proliferation and apoptosis (Bailey, 2008). Although a particular function within the cell can be achieved by a singular protein, the majority of proteins must interact with other proteins in order to achieve their functions. These protein interactions can be stable or transient. Stable protein interactions are between multi-subunit complexes with quaternary structure of folding such as hemoglobin. Transient protein interactions are temporary in nature and require specific conditions for the interaction such as a conformational change in the shape of a protein as a result of, for instance, temperature, pH, phosphorylation or a binding of a ligand. Transiently interacting proteins are involved in cellular processes such as protein folding, signaling and transport. Most cellular processes are transient interactions and for that reason difficult to capture by experimental methods. As a result, studies on PPI have primarily been conducted within the context of static interaction networks, providing only a glimpse into these complex interactions at a specific point in time (Thermo Scientific Pierce, 2010).

Protein interactions are very specific and require the interacting proteins to have complementary geometric shapes that fit like a "lock and key" (Figure 2). PPI is a dynamic process in which one protein, the ligand, finds and fits into another protein's binding site, a cavity in the protein surface (Alberts *et al.,* 2002).
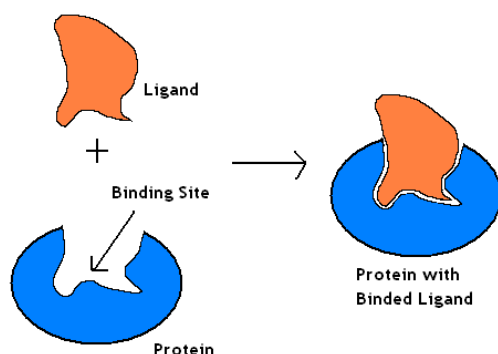
Figure 2. Schematic representation of the "lock and key" model. [Reproduced from Berg *et al*., 2006]

The strength of the bond depends on the formation of many weak, non-covalent bonds such as hydrogen bonds, ionic bonds and van der Waals attractions. A change in the three dimensional shape of either protein has the potential to destroy the binding. The leucine zipper is an example of a three dimensional structural motif in proteins that stabilizes the intricate folding pattern of proteins (Figure 3).
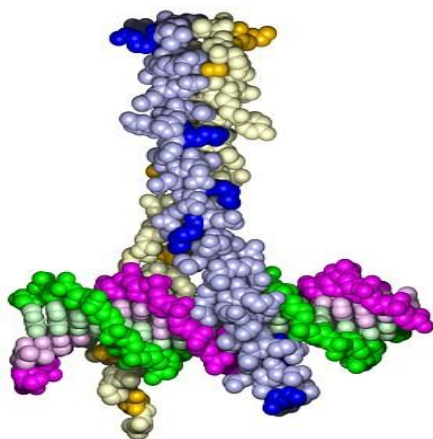
Figure 3. The leucine zipper structure. [Reproduced from Zeably Images, 2012]

The leucine zipper is a part of a DNA-binding domain and consists of two α-helices held together by hydrophobic interactions between leucine residues (Malacinski, 2003). This tight molecular packing of leucine zippers provides stable binding for multi-protein complexes and facilitates stable interaction.

PPI have a wide range of biological effects on a protein ranging from changes in the binding rate to completely inactivating the protein. For example, PPI often change the kinetic properties of the enzyme, consequently effecting the allosteric regulation of an enzyme by either increasing or decreasing the protein's activity at sites other than the active site (Alberts *et al.,* 2002). PPI can also change the binding specificity of a protein through interaction with other proteins leading to a new protein function or creating totally new binding sites. In addition, protein interactions also have a regulatory role in either upstream or downstream regulation as well as product formation.

The Coevolution of Interacting Proteins

When studying PPI it is essential to consider the evolutionary rate of all the proteins involved in a particular network as the evolution of a protein can impact the proteins with which it interacts. There are five determinants of evolutionary rates of proteins, namely: protein dispensability, protein structure, stability and the presence of functional sites, stage of development, and range of expression in different tissues and expression level (Lovell *et al.,* 2010). All five determinants of the evolutionary rate of proteins depend on the protein's function, number and type of binding partners and location in the network. The coevolution of proteins involves not only a shared evolutionary history but also a reciprocal nature. When an interacting protein evolves it may incite changes in its partner protein driven by selection pressures to maintain the interaction (Lovell *et al.,* 2010). When one protein mutates, it can hinder its interaction with its partner(s) by disrupting their "lock and key" fit. In turn, biological functions can

be compromised because of the binding interruption. Reciprocal evolutionary change allows for the interaction between the proteins to continue as amino acid changes at one site give rise to selective pressures at another site accounting for the conservation of binding surfaces. Thus, a change in one protein will be mitigated by a compensatory change(s) in its binding partner, maintaining interaction and function in the face of evolutionary change. As a result, physical PPI can lead to linked evolutionary change between the binding partners (Lovell *et al.,* 2010).

One example of PPI constraints on amino acid residue substitution is illustrated by the spatial relationship between amino acids in one of the most studied examples of self-aggregating protein system, the Tobacco Mosaic Virus (TMV) (Altschuh *et al.,* 1987). The intricate mosaic structure of the virus consists of individual proteins called capsomeres that are closely packed and arranged in the form of a regular spiral or a helix (Figure 4). The capsomeres form the capsid, which encloses and protects the genetic material of the virus.
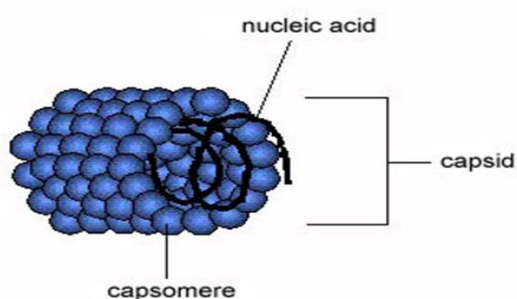
Figure 4. Schematic model of Tobacco Mosaic Virus (TMV). [Reproduced from Splettstoesser, 2012]

The TMV capsomere illustrates strong conservation of amino acids residues that are in contact with other residues in order to maintain the PPI, which contribute to the mosaic shape of the virus. TMV sequence analysis shows that there is a conservation of

residues or a complementary mutation occurring at adjacent residues to conserve the PPI otherwise mutations could cause structural changes that would disrupt the folding or assembly of the virus and compromise the fitness of the virus. The PPI complexity is further emphasized by the hydrophobicity of protein cores (Altschuh *et al.,* 1987). Any mutations in the hydrophobic amino acids residues can change the geometry of the protein, disrupt PPI and compromise fitness. In the case of the TMV, the protein core sequences and amino acid residues that make up the tobacco mosaic structure are highly conserved in all of the seven serogroups to the extent that some of them are structurally identical and can co-assemble despite divergence of overall sequnce (Altschuh *et al.,* 1987).

TMV is a clear example of coordinated coevolution between interacting proteins, thus illustrating the intricate relationship between their protein structures. The capsomers have an important function to arrange in such way that the genetic material of the virus is protected. As a result, the capsomers and network structure determine the evolutionary rate of the interacting proteins by placing selective pressures to constrain mutations in order to conserve interactions that make up the capsid. Considering that capsomers in the TMV interact to form a structural network and share the same function, they must also share similar rates of evolution. The rate of evolution of these capsomers is slower because they have to maintain bonds with many interacting partners.

Changes in amino acid residues can also affect the structure of a protein's interface with another protein in a network. In turn, structural changes in the protein

interface can lead to: (1) reduced binding affinity with other proteins, (2) loss of the ability to interact with a specific protein, or (3) the capability to interact with new protein partners. In the case of the former, modifications within one protein of a complex can incite adaptations within the protein partner(s) to select for increased binding affinity. As a result, structural analysis of a number of protein families show that residues in protein interfaces are more conserved than average amino acid residues (He *et al.,* 2006).

The coevolution of interacting proteins is also influenced by the individual proteins' functional role and location in the network. Generally, proteins sharing the same functional class occupy the same specific part of the PPI network (are closely related within the PPI network) and create a cluster of proteins called the functional module. In turn, these highly connected functional modules are often referred to as the hub proteins that share well defined functions and a large number of interactions (Tillier *et al.,* 2009). Hub proteins evolve slowly as the number of PPI partners for proteins is negatively correlated with their evolutionary rates (Makino *et al.,* 2007). Moreover, hub proteins are often referred to as the central-lethal proteins because mutations in these centrally located proteins are essential to the network architecture as well as species survival (He *et al.,* 2006).

Furthermore, location and function as well as number and nature of interacting partners in a PPI network affects the evolutionary rates of individual member proteins. Proteins interacting with proteins belonging to different functional classes and being in a sparse part of the PPI network are under the strongest functional constraint (Makino *et*

al., 2007). That is because proteins in a sparse part have less suitable substitutable PPI partners. For example, the mitogen-activated protein kinase (MAPK) is involved in various cellular functions, and interacts with proteins having different functions such as ribosomal biogenesis, cytoskeleton and directional cell growth (Makino *et al.,* 2007). For proteins such as the MAPK, the gene expression patterns do not correlate with those of their PPI partner proteins, suggesting that they interact with the PPI partners at different subcellular localizations or different time points (Makino *et al.,* 2007). As a result, the number of interactions among the PPI partners for proteins such as MAPK is expected to be smaller than that of proteins in the functional module. In addition, since proteins in the PPI networks evolve under the influence of their partners, the number of PPI partners is significantly correlated to their evolutionary rates. As a result, proteins that interact with a different functional class and are in a sparse part of the network evolve slower than proteins which bind with others sharing the same functionality as itself because they are under stronger functional constraints.

S. cerevisiae is one of the best studied organisms in terms of its genetics and biological processes and is often used as a eukaryotic model organism in molecular and cellular biology. Of all the eukaryotes, the PPI network of *S. cerevisiae* is the best understood and most complete and for that reason yeast is often used as a model organism in terms of PPI. Furthermore, approximately 383 proteins in *S. cerevisiae* have been suggested as direct orthologs for proteins associated with or responsible for human disease (Hsu *et al.,* 2007). Considering the degree of evolutionary distance between yeast

and human, this is very striking and suggests that many of these proteins are involved in fundamental cellular processes in the eukaryotic cell. As such, protein interactions of *S. cerevisiae* offer valuable information about the evolution of PPI. In addition, protein interactions of *S. cerevisiae* provide examples of the factors that influence PPI such as protein function, indispensability, connectivity as well as reciprocity of amino acid mutations among interacting proteins. Interacting proteins in *S. cerevisiae* that are essential to fitness and survival of the organism have been evolutionarily conserved across the Ascomycota species (Pagel *et al.,* 2004). Hence, the study of interacting proteins for model systems is informative given the assumption that orthologous protein pairs in other species also engage in similar interactions. So far, studies of yeast orthologs to human proteins have already generated significant insight into human disease as mutations in these orthologs correspond to either a loss or gain of PPI which translate to certain endocrine, immune and metabolic disorders (Hsu *et al.,* 2007).

Experimental Methods for Identifying PPI

The most widely used experimental method to infer PPI is the Yeast Two Hybrid System (Y2H), which uses the "bait-prey" model to screen for the physical binding of two proteins (Figure 5). Y2H uses a genetically engineered strain of yeast that lacks a selectable marker such as an essential amino acid. The engineered yeast will not grow in a medium that lacks the selected amino acid. The Y2H method takes advantage of the ability to introduce new foreign DNA via plasmid into the mutant yeast strain (Thermo Scientific Pierce, 2010).
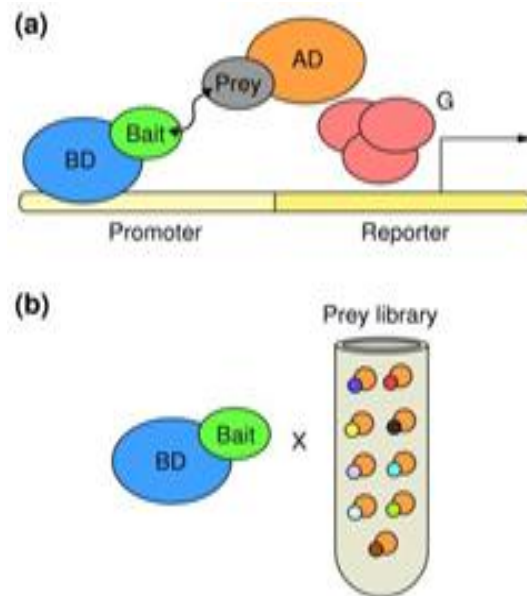
Figure 5. Schematic representation of the Yeast Two Hybrid System (Y2H), which uses the "bait-prey" model. [Reproduced from Giorgini & Bukowski, 2005]

Consequently, in the Y2H method, two plasmids are introduced simultaneously to the mutant yeast strain. One plasmid, the "bait" plasmid has the protein of interest fused to the DNA-binding domain. The protein of interest is usually a known protein used as bait to identify new binding partners. The other plasmid, the "prey" plasmid is made of up a known protein or a library of known or unknown proteins fused to a transcription activation domain. When the bait and prey proteins interact via binding, the bait binding domain and the prey transcription activation domain are brought into close proximity and create a functional transcription factor. This incites the transcription of the selectable markers. Only the transformed mutant yeast strain in which the bait and prey proteins interact will grow in a selectable amino acid deficient medium. As a result, an interaction between the bait and prey proteins rescues the original phenotype of the mutant yeast strain.

Since the development of the Y2H assay over a decade ago, this technique has detected numerous PPI. It is, however, not without its limitations. The biggest drawback for using the Y2H assay is its high identification of false positive protein interactions (Thermo Scientific Pierce, 2010); the bait can often activate transcription on its own without interaction with the prey protein. This is called auto-activation of transcription, and it is present in about 5% of proteins and even more in randomly generated libraryfragments (Thermo Scientific Pierce, 2010). As a result, preliminary tests must be done to test for auto-activation of transcription in the protein of interest. In addition, another drawback of the Y2H assay is the extensive use of fusion proteins, such as the bait and prey proteins which may change the conformation of the bait protein because of the modular nature of proteins. As previously mentioned, changes in protein folding can alter the activity and binding of a protein. Lastly, using yeast as a model for PPI may not reveal true interaction because yeast does not normally produce the experimentally implanted proteins on its own. True PPI may depend on certain post transcriptional modifications such as disulfide bridge formations, glycosylation and phosphorylation, which may not occur in the yeast system (Thermo Scientific Pierce, 2010). These post transcriptional modifications play an important role in the folding and stability of the proteins and can affect the binding and thus interaction of proteins.

Another popular experimental method for identifying PPI is Co-Immunoprecipitation (Co-IP), which finds interacting proteins by using protein specific antibodies and indirectly captures proteins that are bound to a specific target protein

(Figure 6). Co-IP is a powerful technique for identifying PPI but may require several rounds of precipitation with different antibodies to pull down each member of a protein complex, and antibody contamination is one of the most encountered problems in the co-IP approach (Thermo Scientific Pierce, 2010).
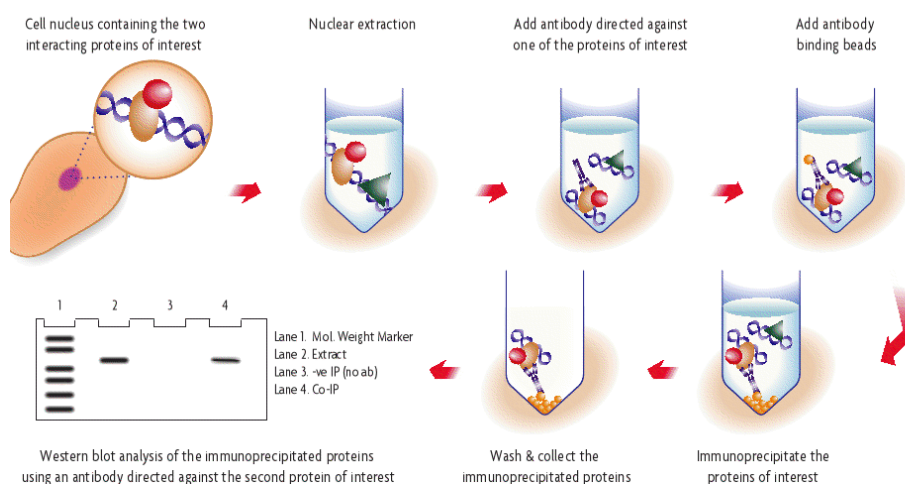
Figure 6. Protein Co-Immunoprecipitation (Co-IP). [Reproduced from Nuclear Complex Co-IP Kit, Active Motif North America]

The Far-Western Blot is another experimental method for detecting PPI that uses a targeted bait protein rather than a target protein specific antibody as in Western Blot. Nonetheless, all these techniques are time consuming and require many reagents and can only be performed on proteins of known structure. In addition, the three-dimensional protein conformation makes PPI analysis cumbersome to study because of its folding capabilities (Thermo Scientific Pierce, 2010).

Computational Methods for Identifying PPI

Since the publication of the human genome and the onset of proteomics, inferring PPI from sequence and genomic information has been one of the main focuses of

computational biology. Protein sequences capture information at the amino acid level, which cannot be resolved using standard experimental methods mentioned above. As a result, computational methods seek to determine PPI based on high-throughput protein sequences. Current computational methods aim to integrate and improve information from various available approaches that predict PPI. This in turn, expands the database and improves the accuracy of current PPI networks. Existing computational methods that predict PPI include approaches such as the Rosetta stone, conservation of gene neighbor and phylogenetic profiling predict and analyze genomic structure and pattern changes associated with PPI using the genomic context (Hirose, 2012) (Table 1).

Method	Interaction Type	Interaction
Rosetta stone	Indirect functional association	Protein
Conservation of gene neighborhood	Indirect functional association	Protein
Phylogenetic profiles	Indirect functional association	Protein/domain
Mirror tree	Indirect functional association	Protein/domain

Table 1. Summary of computational methods for predicting PPI.

The Rosetta stone method is a computational method that infers PPI by comparing different genomes. The Rosetta stone algorithm searches for proteins that are conserved between different organisms and attempts to find proteins that may interact based on a fused form in another species (Hirose, 2012). Thus, as shown in Figure 7, if two proteins in one genome (here genome i) appear as a fused protein in another genome (here genome j), they are predicted to be interacting proteins.
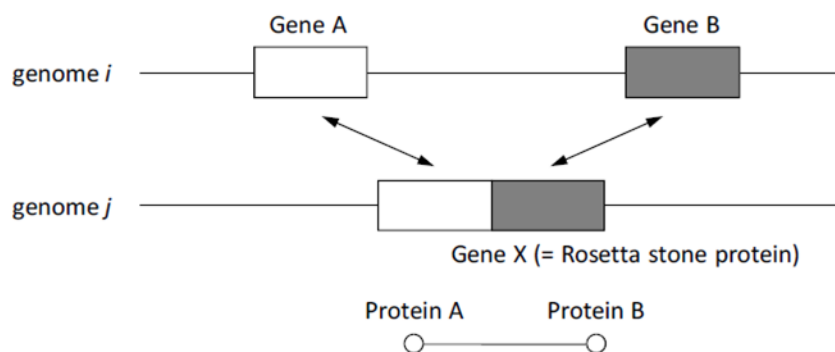
Gene X is the Rosetta stone protein, indicating that protein A and protein B are functionally related.

Figure 7. Rosetta stone approach for PPI. [Reproduced from Hirose, 2012]

Thus, the two proteins that interact are functionally related and will have significant sequence similarity to a single fused protein in another genome (Hirose, 2012). The fused protein is called the Rosetta stone as it unlocks the functional relationship between two genes that are encoded independently in a genome (Hirose, 2012). However, the Rosetta stone method cannot be applied to promiscuous domains such as SH3 domains and ATP-binding cassettes that are found in many otherwise unrelated proteins. In addition, the method is prone to false negative predictions as some proteins may interact without there being a Rosetta stone to indicate an interaction.

Another computational method for inferring PPI predicts that proteins encoded by conserved neighboring gene pairs interact. This method suggests that gene order on a genome is conserved if the gene products physically interact with each other either by complex formation or if the proteins are transcribed as a single unit (Dandekar *et al.,* 1998). Furthermore, the genes encoding proteins that either form a complex via physical interaction or work together in the same pathway are encoded in the same operon in

different genomes. However, this approach is limited to bacteria and archaea that have

operon structures (Hirose, 2012). The conservation of gene neighborhood approach is not

quite as direct an association in eukaryotes because only some eukaryotes, including *C.*

elegans and the *Drosophila melanogaster*, have operons. In addition, the conservation of

gene order can incorrectly infer PPI in closely related organisms as the lack of time for

genome rearrangements after divergence of the two organisms from their last common

ancestor could be a reason for the observed gene order conservation. Consequently,

onlyorganisms with relatively long evolutionary distances should be considered for such

type of analysis (Jothi & Przytycka, 2008).

Finally, the phylogenetic profile approach infers PPI from genome comparisons

by examining the patterns of presence or absence of proteins across multiple genomes

(Figure 8). In this approach the functional relationship between two proteins is detected

by comparing their phylogenetic profiles. The premise behind the phylogenetic profile

approach is that proteins with identical or similar profiles are inferred to be functionally

interacting under the assumption that proteins involved in the same pathway or functional

system are likely to have been co-inherited during evolution. Thus, proteins that interact

with each other co-occur in different genomes. A phylogenetic profile is constructed for

each protein as a vector of N elements, where N is the number of genomes. Each position

of the profile represents whether the protein that is homologous to the target protein is

absent (signified by 0) or present (signified by 1) in each genome (Hirose, 2012).

Consequently, the phylogenetic distribution is shown by a long binary number along each

genome. A functionally related protein pair is detected by searching for the same

phylogenetic distribution patterns (Hirose, 2012).

	genome *i*	genome *j*	genome *k*	genome *l*	genome *m*
gene A	1	0	1	0	1
gene B	1	1	0	1	1
gene C	1	0	1	0	1
gene D	1	0	1	1	1
gene E	0	1	1	0	1

Protein A   Protein C

○————————————○

Phylogenetic profile (10101)

Figure 8. An example of the phylogenetic profile approach. [Reproduced from Hirose, 2012]

One advantage of using the phylogenetic profiles approach to infer PPI is that it is

applicable to eukaryotes, since it is not necessary to consider gene order and operon

structure (Hirose, 2012). The disadvantage of the phylogenetic profile approach is that

the analysis is limited to the organisms with completely sequenced and annotated

genomes, because whether a certain gene or protein is actually encoded in the

genomemust be known (Hirose, 2012). In addition, another limitation of the phylogenetic

profile approach is the lineage-specific gains and losses of genes, which are prevalent in

microbial evolution but not in higher-ordered organisms and as a result could artificially

decrease the similarity between functionally interacting genes.

Furthermore, computational methods take advantage of the assumption that interacting proteins most likely coevolved with parallel or reciprocal mutations between amino acids. Thus, predicting interactions from evolutionary distances or similarities between putative coevolving proteins is yet another computational approach that has been developed. Methods based on coevolution, such as the mirror tree approach, predict PPI based on the assumption that phylogenetic trees of interacting proteins are highly likely to be similar due to the inherent need for coordinated evolution. For this purpose, similarity matrices are constructed from alignments of orthologous sequences taken from a common set of species. The degree of correlated evolution between families of orthologs is assessed by computing the correlation coefficient between the corresponding similarity matrices (Figure 9). The mirror tree method measures the correlation between evolutionary distances and thus, indirectly, the correlation between evolutionary rates along individual branches of evolutionary trees of two families (Pazos & Valencia, 2001).

One advantage for using the mirror tree approach to infer PPI is that this method does not require fully sequenced genomes. However, inferring PPI solely on the similarity of phylogenetic trees many result in false positive interactions. For example, the method reported false positives for metallothionein and cytochrome c proteins. The metallothionein protein is Cys rich and many of its quaternary structures are the result of dimerization and disulfide bridge formation through metals bound by Cys residues. More interestingly, dimeric metallothionein proteins often gain new properties. In spite of that, the authors attributed the false positives to the Cys rich composition bias that could affect

the sequence alignment and prediction of interaction (Pazos & Valencia, 2001).

Nonetheless, the mirror tree approach is a robust method as it uses coevolution to link all

functionally interacting proteins to predict PPI based on phylogenetic distance between

the proteins.
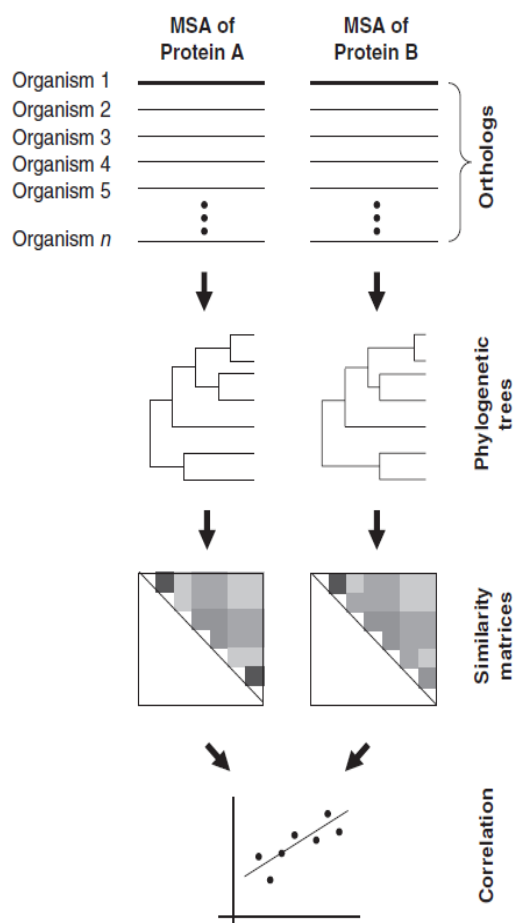
Figure 9. Schema of the mirror tree method. [Reproduced from Pazos & Valencia, 2001]

In yet another approach, Bayesian methods integrate data from a wide variety of

sources, including both experimental results and prior computational predictions, and use

this data to assess the likelihood that a particular potential protein interaction is a true

positive result (Kim *et al.,* 2007). This is not a solely computational approach, rather a hybrid approach. Bayesian methods are useful because experimental procedures, such as the Y2H assays produce many false positives, while the previously mentioned computational methods can only provide circumstantial evidence that a particular pair of proteins might interact. To evaluate the probability of a hypothesis, the Bayesian model relies on some prior probability, which is then updated in the light of new, relevant data. As a result, Bayesian methods relay probabilistic graphical models and integrate heterogeneous datasets to infer PPI. However, the drawbacks for Bayesian methods include generalization of assumptions and extraction of biological knowledge from multiple datasets and classification problems (Kim *et al.,* 2007).

Single Gene Phylogenies

Single gene phylogenies seek to reconstruct the evolutionary relationships among species using only one gene. There are two main approaches for reconstructing single gene phylogenies, namely, the algorithmic approach and the tree searching method (Hall, 2011). The algorithmic approach is based on a distance method that uses an algorithm to calculate from multiple alignments a distance matrix of pairwise differences between the sequences. Distance methods use this matrix as the data from which branching order and branch lengths are computed to construct a phylogenetic tree. The most popular distance methods are UPGMA and Neighbor Joining. One large drawback to distance methods is that because only two samples are used for each distance calculation the distance is

underestimated. Overall, distance methods are avoided because the relationship between

individual characters and the tree is lost in the process of reducing characters to distances.

Alternatively, tree searching methods are character-based, meaning that they use multiple

alignment directly by comparing characters within each column in the alignment (Hall,

2011). The most widely used character-based approaches are maximum parsimony (MP)

and maximum likelihood (ML) methods. Maximum parsimony is based on the

assumption that the most likely tree is the one that requires the fewest number of changes

to explain the data in the alignment. The basic premise of MP is that taxa sharing a

common characteristic do so because they inherited that characteristic from a common

ancestor. MP can give several trees that differ slightly from one another. These trees are

consistent with the same number of events, and are equally parsimonious. MP or

minimum change is the criterion for choosing the best tree. One drawback of using MP is

that sometimes extra steps, referred to as homoplasy, exemplified by characteristic

reversal, convergence or parallelism, are needed to explain the data (Hall, 2011).

Nevertheless, MP operates by selecting the tree or trees that minimize the number of

evolutionary steps, including homoplasy (Steel, 2001).

Another character-based method is maximum likelihood, which directly uses the

aligned characters, such as DNA or protein sequences during tree inference, and

evaluates all possible trees for the one that most likely fits the evolution. ML is a

powerful statistical method that seeks the tree that makes the data most likely. For a

particular data set, ML applies the log-likelihood of that tree to compare various models

of nucleotide substitutions (Steel, 2001). The ML method is the slowest, yet arguably the most rigorous and computationally intensive method, producing the most informative tree (Saitou, 1988). Some recognized advantages of ML are that it uses all of the sequence information and it provides a robust evolutionary model. For example, the Protein ML (ProML) program in the PHYLogeny Inference (Phylip) package (Felsenstein, 1993) implements three important assumptions about change between amino acids. First it ascertains that each position in the sequence as well as different lineages evolve independently. Secondly, it assumes that each amino acid undergoes substitution at a specified rate, one possibility being a constant rate of change. Most importantly, ProML assures that all relevant positions are included in the sequence, not just those that have changed or those that are presumed to be "phylogenetically informative." Finally, ProML implements either the Taylor-Thornton or the Dayhoff probability model of change between amino acids. ML outperforms alternative methods such as MP or distance methods and produces one tree with known likelihood.

The methodology for single gene phylogenies is well established; however single-gene analyses often do not provide sufficient resolution and sometimes give conflicting results. For instance, single gene phylogenies produce incongruent tree topologies between gene trees and species trees (Figure 10). That is because single genes have different evolutionary rates from that of the organism. A gene tree contains the evolutionary history of genes and a species tree depicts the descent of the taxa (Maddison, 1997). Some genes will evolve both individually and along with the organism but ultimately,

genes and species are different entities. 16S rDNA phylogenies are known to produce

trees with puzzling topology but with the growing sequence data, multiple gene

phylogenies can resolve these inconsistencies (Wolf *et al.,* 2002).
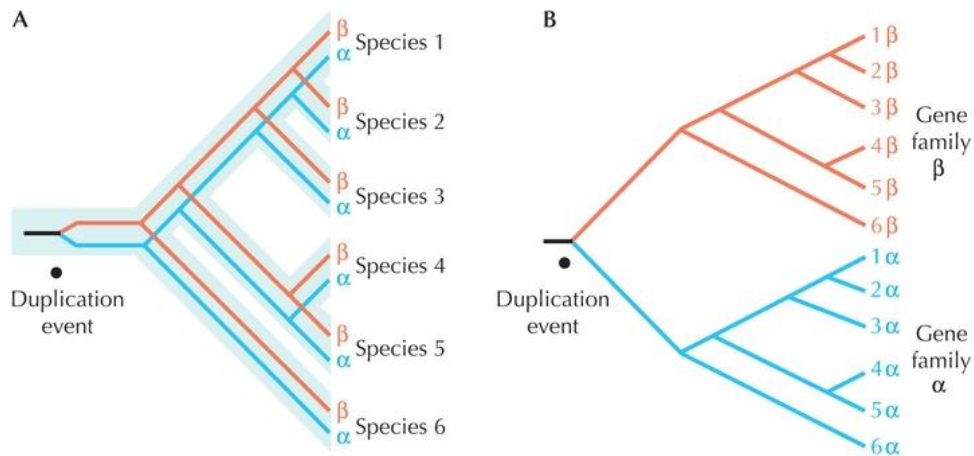
Figure 10. Species Tree vs. Gene Tree. [Reproduced from Barton *et al.,* 2008]

Multiple Gene Phylogenies

The goal of multiple gene phylogenies is to reconstruct more comprehensive

evolutionary relationships. Concatenation and consensus are two different methods for

deriving multiple gene phylogenies. Concatenated phylogenies, are also called total

evidence trees, and attempt to include all available information into the phylogenetic

analyses. This method links together all aligned genes for the same set of species to give

one single alignment (Fitzpatrick *et al.,* 2006). The different genes are combined into a

single supermatrix in which all genes must conform to the same network topology. Total

evidence trees not only use sequence congruence to find the best phylogeny but also

weigh all characters equally. As a result, the branch lengths are comparable across the entire tree because they are derived from the same set of orthologous genes or proteins (Gadagkar *et al.,* 2005). It is expected that concatenation will resolve nodes as well as basal branching by the root of the tree and as a result improve phylogenetic resolution (Sanderson *et al.,* 2003).

The premise of total evidence trees is that homoplastic traits that are similar in species of different ancestry (which are the result of convergent evolution), are randomly distributed and will be overshadowed if enough data are collected (Edwards *et al.,* 2008). Unfortunately, studies suggest that homoplasy is not randomly distributed and adding more sequence data may produce incorrect phylogenetic inferences with false confidence (Brown *et al.,* 2001). Furthermore, concatenation or total evidence trees can be misleading when comparing sequences from distantly related taxa because the historic signal of DNA sequence data from these species may be greatly attenuated and unreliable due to convergent evolution (Naylor *et al.,* 1998). Another drawback for using concatenation is that the method forces a single sequence model on all of the genes under study, suggesting that they all evolved at the same rate, which is known not to be true (Wolf *et al.,* 2002). Concatenated approaches also struggle to deal with missing or incomplete data sets that devalue phylogenetic inference by increasing the uncertainty of taxa grouping and the number of possible solutions. However, the amount of missing data that can still produce an acceptable result remains debatable (Sanderson *et al.,* 2003).

The concatenation approach can be improved by including enough sequence data thus increasing the resolution of the phylogenetic signal (Rokas *et al.,* 2003). Increasing sampling can also help resolve the long branch length artifacts such as long branch length attraction (LBA), which mistakenly groups two or more long branch lengths as sister groups (Bergsten, 2005). Moreover, sampling more genes might also resolve deep and shallow nodes and provide a clearer phylogenetic picture (Maddison, 1997). Deciding how many sequences are needed for phylogenetic inference is fundamentally important because concatenating numerous sequences will, however, magnify any biases that were introduced with the model or the evolutionary process. Moreover, concatenating numerous sequences comes at a computational cost, both in terms of memory usage and runtime. Nevertheless, large phylogenies can be reconstructed using concatenation by taking advantage of biclique enumeration (Sanderson *et al.,* 2003). Quasi-bicliques are a new method that compensates for different rates of substitution (Sanderson *et al.,* 2003).

The other multiple gene phylogeny method for reconstructing evolutionary relationships is the consensus approach. Consensus phylogenies are inferred from individual gene trees that are then added in agreement to generate the consensus tree, taking a "divide and conquer" strategy (Sanderson & Driskell, 2003). In general, the consensus method takes a set of phylogenetic trees as input and produces a single summary tree output (Fitzpatrick *et al.,* 2006). There are numerous types of consensus tree approaches, including strict consensus, semi-strict consensus, majority-rule consensus and Adams consensus. The two most widely used consensus approaches are

the strict consensus and majority-rule consensus methods. To illustrate, the strict

consensus tree includes only those groups (clades, subtrees) that are present in all primary

trees. Whereas, the majority- rule consensus has a more relaxed criterion and includes

clades that are present in more than 50% of the primary trees. In turn, Adams consensus

can be applied to trees that have different topology by moving the branches that did not

exist in any of the trees from the original set to the root (Sul & Williams, 2011). Overall,

the consensus method tries to conserve the historical signal from the individual trees

before combining them into a single tree. As a result, the consensus approach accounts

for the differences in evolutionary rates and substitution patterns among individual trees

(Gadagkar *et al.,* 2005). The consensus approach produces a conservative estimate of

phylogeny as it produces high resolution in the branching pattern only when there is at

least a majority consensus among the different data sets (Gadagkar *et al.,* 2005).

Nonetheless, one criticism of the consensus method is that it concentrates too much on

the individual tree topology instead of analyzing the information contained in the

sequence data. Another criticism points out that not all sequences are usually available

for some species and thus not all trees will be identical with respect to the terminal taxa

they contain.

The consensus approach is very similar to the supertree method, which also

reconstructs multiple gene phylogenies by combining tree topologies. However, the

supertree approach differs from the traditional consensus methods in that the input trees

do not have to be identical but only overlapping. As a result, the supertree method is

more forgiving to missing data sets. One of the most used supertree methods is the Matrix

Representation Parsimony (MRP), which converts each input tree to a representation of

matrix binary characters (Fitzpatrick *et al.,* 2006). This collection of input trees is then

represented by a single matrix derived from combined individual matrices. In the

supertree method, missing data is assigned a question mark in its respective matrix cell

(Sanderson & Driskell, 2003). However, even though supertree methods such MRP

combine trees rather than data, the similarities between the two methods are superficial

and supertrees are not analogous to consensus methods (Pisani &Wilkinson, 2002).

In order to reconstruct robust and reliable multiple gene phylogenies, both the

concatenation and consensus approaches have to consider the unique behavior of

individual genes as well as the entire organism. For instance, different regions of a

genome undergo varying evolutionary pressures allowing for particular genes to evolve at

different rates during the history of life. The evolutionary change is effected by both

mutation and natural selection. Mutations produce new genetic variations, whereas

natural selection determines the fate of the new genetic variant (Lenski, 2001). In

addition to the problem of unequal evolutionary rates of different genes and species,

horizontal gene transfer, gene duplication, deep coalescence and branch length all pose

additional challenges to phylogenetic reconstruction.

Branch lengths are especially important in phylogenetic reconstruction as they

reflect the proposed evolutionary history. Branch lengths represent the passage of time

measured as character difference. Branch lengths, however, are easily affected by

heterotachy, within-site rate variations, as the evolutionary rate of a given position is not always constant throughout time. It has been proposed that the positions that show switches in substitution rates over time (that is, heterotachous sites) are good indicators of functional divergence (Lopez *et al.,* 2002). Sites that have different rates in different parts of the phylogeny are said to be heterotachous. In turn, heterotachous sites can be used to estimate the functional divergence of proteins (Maddison, 1997). As a result, identification of heterotachous sites can be very informative in phylogenetic reconstruction. However, branch length heterogeneity can cause long branch length artifacts and produce mistaken inferences of relationship between taxa. Branch length heterogeneity also generates unexpected phylogenies by incorrectly quantifying the evolutionary rate at each site. Branch length heterogeneity will not detect the fact that the evolutionary rate of a particular site can vary even though the function remains the same (Bergsten, 2005). In addition, heterotachy was found among homologous sequences of distantly related organisms, often with different functions (Lopez *et al.,* 2002). Thus, branch length heterogeneity plays an important part in improving the reconstruction of phylogenetic methods.

Both the concatenation and consensus approaches have been implemented to resolve phylogenetic uncertainties in the tree of life. For instance, the concatenation method was tested on 6,591 protein sequences to reconstruct the monophyly of the three domains: Bacteria, Archaea and Eukarya (Brown *et al.,* 2001). Another successful example where both concatenation and consensus approaches were implemented to

resolve phylogenetic uncertainties was in the fungal phylogeny including 153 orthologs from 42 different fungal species to infer fungal phylogenies (Fitzpatrick *et al.,* 2006). Furthermore, the concatenation method has also been applied to examine the phylogeny of complex eukaryotic organisms (e.g., Giribet *et al.,* 2001; Wolf *et al.,* 2002).

In conclusion, both the concatenation and consensus approaches reconstruct comprehensive multiple gene phylogenies that are supported by well-established studies. When genetic processes such as gene duplication, gene transfer and creation of LBA are taken into consideration during analysis, both the concatenation and consensus methods can infer robust phylogenies. Unequal evolutionary rates of different genes and species are a major challenge to phylogenetic reconstruction, for which the concatenation method is especially vulnerable. In summation, even though concatenation and consensus approaches use different methods, they generate congruent phylogenetic results. Moreover as the amount of readily available genomic sequences has increased, the computational ability of both approaches to handle large data sets has also improved.

Representing and Visualizing Trees

In mathematics, a tree is a diagram or graph that visually illustrates the connectivity and relatedness of entities. A tree is a set of points, called nodes and straight line segments called branches or edges that connect two distinct nodes (Harvey &Wright, 1999). A tree is then a collection of nodes and edges that connect the nodes without assigning direction and referred to in mathematics and computer science as an undirected graph (Figure 11).
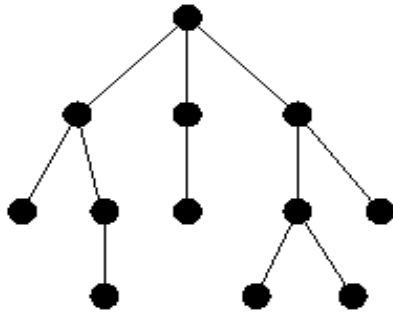
Figure 11. Undirected tree.

In addition, every node in a tree is connected in a hierarchical structure with the parent node above the child node. The parent-child relationship can be extended to ancestors and descendants. Nodes that have the same parent are called siblings (Harvey &Wright, 1999).This hierarchical structure makes the tree the most fundamental data model used in computer science. Thus, a tree is a mathematical structure that can be either viewed as a graph or a data structure. This is because a tree data structure contains a set of elements, as well as connections between elements which produce a tree in graphical form. Accordingly, a graph or data structure is analogous (Weisstein, 2012).

In computer science, a data structure is a particular way of storing and organizing data in a computer's memory so that it can be used efficiently. The tree data structure is a powerful tool for organizing multiple data objects in terms of a hierarchical relationship. Each node in a tree is contained within the data structure and consists of a value(s) or attribute(s). There are many data structures that can be used to represent trees. Generally, the nodes of a tree can be represented by structures in which the fields link the nodes together in a manner similar to the way in which the nodes are connected in the abstract tree. Thus, the tree structure focuses on how nodes are represented. The basic data

structures used to represent trees in programs include arrays of pointers, leftmost-child-right sibling and parent pointers (Garcia-Molina *et al.,* 2008).

The simplest form of a binary tree consists of a node and all of its proper descendants called a subtree. In turn, subtrees themselves represent binary trees. A binary tree is a recursively defined data structure in which the function being defined is applied within its own definition (Weisstein, 2012). Such data structures can be easily represented in several functional programming languages including Lisp, which has distinctive parenthesized Polish prefix notation for nested list (tree-structured) data (Walker & Blum, 1985). Tree data structures in Lisp and several other functional programming languages use parenthesized lists, which consist of balanced strings of parentheses (Figure 12). The nested sequences of parentheses imply structure by controlling sequence order and grouping.

```
(:= ("S" + ("A" * ("B" "C") ) ) )
```

Figure 12. The parenthesized form of tree.

Trees have also been used extensively in biology to graphically represent various types of hierarchical relationships, eg. Figure 13.
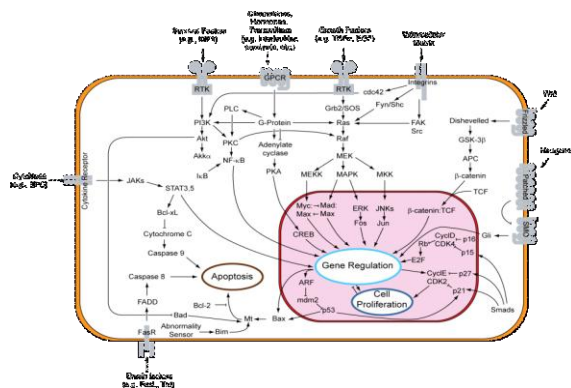
Figure 13. Mitogen-activated protein (MAPK) kinases pathway. [Reproduced from Alberts *et al*., 2002]

Just as in computer science and mathematics, in biology trees are drawn pictorially as a set of points called nodes, connected by line segments called edges that connect two distinct nodes. Any tree data can be represented linearly. The Newick format is an example of such linear tree representation. Simply, the Newick tree format writes out trees in a text form (Figure 14). As shown in this figure, the linear representation utilizes parentheses similar to the implementation programmatically (Figure 12).
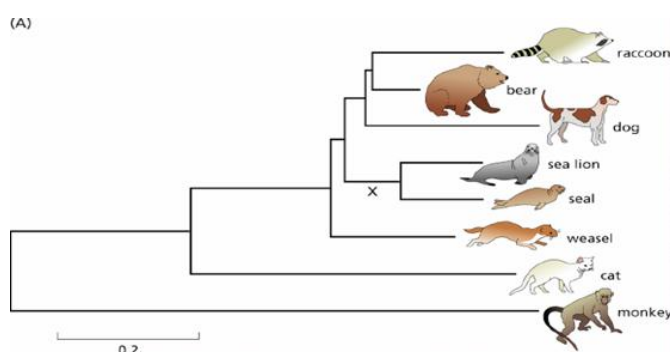
Figure 14. Linear tree representation in Newick format. [Reproduced from Nakhleh, 2010]

(((raccoon, bear),((sea_lion,seal),((monkey,cat), weasel))),dog);

Furthermore, the Newick tree file format incorporates information about edge lengths that connect the nodes of the undirected graph and thus provides branch length numbers for the tree that indicate the amount of character change (Felsenstein, 1993) (Figure 15).

(((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99700, seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):20.59201, weasel:18.87953):2.09460)):3.87382,dog:25.46154);

Figure 15. Linear tree representation in Newick format with branch lengths.

The Newick tree format has been adopted by many programs for sequence data analyses including the previously mentioned Phylip package, and it is very useful for exchanging trees between different types of software (Felsenstein, 1993). Consequently,

the Newick tree format can be graphically visualized on many tree drawing programs

such as TreeView, MATLAB, PhyloDraw, PhyloWidget an NJPlot. These tree drawing

programs open or save Newick-formatted files as well as view, edit, format, and explore

tree data for further analysis.

Visualizing Interactomes

Experimentally confirmed interacting proteins are documented and published in

databases of interacting proteins. Most of PPI knowledge has been accumulated using

biochemical experiments, including Y2H assays and Co-IP. The Biological General

Repository for Interaction Datasets (BioGRID) is an example of a database that curates

experimentally established interactions for all major model organism species (Stark *et al.,*

2006). Aside from BioGRID there are many other PPI databases that curate protein pair

interactions based on kinetics, geometric 3D analysis, signal transduction pathway and

disease association. Other PPI databases are organism specific. For example, Protein

Interaction Maps (PIMs) explore protein pathways of *Helicobacter pylori,* and *D.*

melanogaster, whereas other databases focus on yeast or bacterial PPI. However,

BioGRID as well as the other experimental databases are not frequently curated and as a

result not all supported organisms show the most recently established protein interactions.

Consequently, computational databases not only aggregate validated PPI from literature,

orthology and high-throughput experiments, but also use the compilations to predict new

PPI. One example is the ProLinks Database, which holds a collection of computational

methods such as Rosetta stone, Gene Neighbor and Phylogenetic Profile that infer PPI. In

addition, ProLinks uses a simple algorithm to identify proteins that are found together in scientific abstracts. ProLinks contains 83 fully sequenced microbial genomes to which the PPI interaction computational methods can be applied to in order to generate predicted protein associations for each organism (Bowers *et al.,* 2004). ProLinks displays the PPI network in graphical form using nodes and edges. The layout of the graph is determined using a spring minimization algorithm, which assumes that each node pair is connected by a spring whose stiffness is proportional to the graph-theoretic distance between nodes (Bowers *et al.,* 2004). Thus, the ProLinks database takes advantage of visualization, data mining and integrative approaches to produce new insights into otherwise scattered information.

The Search Tool for the Retrieval of Interacting Genes (STRING) is the largest database of known and predicted PPI for a large number of organisms. The database currently covers 5,214,234 proteins from 1,133 organisms. The PPI in the STRING database are derived from genomic context, high-throughput experiments, coexpression and published literature (Szklarczyk *et al.,* 2011). STRING also includes both direct and indirect associations and combines reported information from databases centered on specific model organisms. However, one limitation of STRING is that to infer PPI the database uses clusters of orthologous groups (COG), that have been previously determined using sequence alignment techniques rather than individual genes (Bowers *et al.,* 2004). Thus, a STRING-generated COG network might include some COGs that are not present in the organism of interest. STRING visually displays the PPI network using

nodes to represent proteins and edges to represent predicted functional links. All of the

PPI databases take advantage of visualization tools to assist in understanding the complex

data. In the prior discussion of phylogenies, phylogenetic analyses result in the creation

of phylogenetic trees, which visualize the evolutionary relationship among species, genes

or proteins. Similarly, interactomes are often represented as a tree and such visual

representation of interactomes can clearly illustrate and organize relationships.

CHAPTER TWO

METHODS

Deriving tree of trees

For our proof-of-concept work individual proteins with known functionality were selected within two systems – the auditory system and the visual system. Some of the interactions between the selected proteins have been experimentally verified while others are hypothesized. Furthermore, mutations in this set of proteins can cause malfunctions and disorders in the vertebrate systems. This further confirms that these proteins are fundamental to the development and function of the auditory and visual system. Consequently, all of the chosen proteins had orthologs in different vertebrate species. In total, eleven auditory ortholog proteins, CDH23, EYA1, FGF3, MYO7A, OTX1, IRF6, SOX10, PAX3, SIX1, UGDH, and SPT6, and sixteen visual ortholog proteins, RHO, OPN1SW, RPE, GNA14, GNGT A, GNGT Γ, GNB1, RIBEYE, PAX6, BSN, ATOH7, SIX3, SIX6, and PDE6A-D, were chosen to be examined (See Appendices A and B for the complete list of protein accession numbers and protein description).

<u>Sequences</u>

The desired protein sequences were found through the Gene database curated by National Center for Biotechnology Information (NCBI). Consequently, for each protein, all sequences were pulled for vertebrate species having complete, annotated genomes, including *H. sapiens, P. troglodytes, M. musculus, R. norvegicus, B.taurus, C. familiaris,*

G. gallus, M. domestica, *X. laevis, D. rerio.* Two invertebrates, *D. melanogaster* and *C. elegans* both fall outside of the clade being studied, were selected to serve as outgroups for comparison. In addition, for each protein, Basic Local Alignment Search Tool (BLAST) was used to find additional homologous sequences using a protein query (blastp) to return the most similar protein sequences from the NCBI Protein database (Altschul *et al.,* 1997). Query results with E-values closest to zero were considered a significant match and further referenced. In total, 587 protein sequences were evaluated and used in this study. The breakdown for the analysis was as follows: 251 protein sequences for the auditory system and 336 protein sequences for the visual system.

Individual Trees

The retrieved protein sequences were saved in FASTA format, in which amino acids are represented using single-letter codes. The FASTA files were then imported into the BioEdit Sequence Alignment Editor v7.1.3. ClustalW (Thompson *et al.,* 1994*),* a general purpose multiple sequence alignment program for DNA or proteins was run using full multiple alignment and 1000 Bootstraps option, which samples about 2/3 of the multiple sequence alignment (MSA) with noise replacement to ensure that only homologous residues are aligned in every column of the MSA. The protein maximum likelihood program (ProML) (Felsenstein, 1993) was then used to infer phylogenies. For each protein ProML generated a Newick format tree file which was saved as a text file and then opened in NJ plot (Perrière & Gouy, 1996), a general purpose tree drawing

program to analyze the individual protein phylogenies. In total 27 Newick format files

were generated and saved for later use in the tree of trees method.

Tree of Trees

In order to visualize how all of the individual proteins interact and evolved in a

particular system, the branch lengths were recalculated to include all of the proteins in

one tree. In essence, individual protein Newick format files were pooled together and

each protein branch length was evaluated against all other protein branch lengths. The

formula used to recalculate the branch lengths was:

$$D_{i,j} = \sqrt{\sum \left(\left|\text{branch length } A - \text{branch length } B\right|\right)^2}$$

where A and B represent two different nodes in the tree. In the case where one protein's

tree included a species that was not represented in all other protein trees, this species was

removed from consideration. Next, the recalculated branch lengths were put into a two-

dimensional array (matrix) the size of $N \times N$ where N is the number of species in the

individual protein tree. (In the event that more than one sequence was considered for a

species, the average branch length between species was calculated.) In this way, the

obtained matrix, D, includes the branch lengths. This varies from a distance matrix

directly obtained from distance-methods such as Neighbor Joining which directly

measure pairwise distance between two genes and construct the tree from a resultant

distance matrix.

Next, when combining phylogenetic signals from several proteins, the topological

variations among the individual protein trees needed to be taken into account. Even

though all branch lengths were recalculated for the proteins involved in the same system,

the proteins might have varying evolutionary rates among them. As a result, in order to

account for possible unequal evolutionary rates the obtained branch lengths were scaled

using the K-score (Soria-Carrasco *et al.,* 2007). The K-score scales any two compared

trees so that both have comparable evolutionary rates. The formula to calculate the K-

score was:

$$K = \frac{\sum_{i=1}^{N}(b_i b_i')}{\sum_{i=1}^{N} b_i'^2}$$

Two trees with very different topologies and different relative branch lengths

have a high K-score whereas two trees that have comparable topologies and comparable

relative branch lengths, and thus have a similar among lineage rate of variation, has a low

K-score. Consequently, the K-score is a type of statistical tool that evaluates the trees in

terms of topological accuracy as well as reproduction of branch length variation.

Moreover, the K-score selects for orthologous proteins and genes by choosing proteins

and genes that allow for the global divergence as similar as possible to the other tree

(Soria-Carrasco *et al.,* 2007). Accordingly, the branch length matrix was scaled using the

K-score to correct for the differences in topology and branch lengths of the various

protein trees. The formula that takes into account the scaling factor to recalculate the

branch lengths was:

$$BLD(K) = \sqrt{\sum_{i=1}^{N}(b_i - Kb'_i)^2}$$

Visualization of Trees

Finally, the scaled recalculated branch lengths were put back into a two-

dimensional array, matrix, which was saved as an input file and opened on a tree drawing

program, producing an undirected tree graph. The undirected tree graph shows the

relationship between all the selected proteins in a particular system. Thus, the derived

undirected tree graphs for the auditory and visual systems illustrate proteins which

evolved at similar rates amongst vertebrate species and thus putative interacting proteins.

Program Implementation

The goal of this algorithm is to develop an automated method for detecting PPI

and assess confidence in the inferred relationships. Hence, a source code which takes in

and reads individual gene tree files in Newick format and stores the pairwise branch

lengths between nodes in the tree was created. The ultimate goal of this program is to

generate a single tree file which represents the relative rates of evolution of the individual

genes under consideration.

The program to derive PPI and relationships was written in C++, a higher-level

programming language which allows the programmer to concentrate on the logic of the

problem to be solved rather than the intricacies of the machine architecture (Heller,

1996). There are several advantages of high-level programming languages. First, the high-level languages are easier to learn than low-level languages. The statements written for the program are similar to English statements. In high-level language, a new program can easily be written in a very short time and the errors in a program can be easily detected and removed (Arnold & Gosling, 1996). Programs written in high-level language are machine independent, i.e. the program can be executed on any computer (Arnold & Gosling, 1996). For the past 30 years, C++ has remained one of the most widely used languages in the software industry included but not limited to device drivers, video games and high-performance servers (Bhattacharya & Neamtiu, 2011).

To describe the functionality implemented in creating the code to infer PPI, the pseudocode (an abstraction of the actual code written) is presented. This pseudocode describes the entire logic of the algorithm so that implementation becomes a rote mechanical task of translating line by line into source code.

Pseudocode

In this pseudocode, there are several calls to functions made. The implementation is not included here. These functions were developed to transform the Newick file (NEWICK) into a matrix of pairwise branch lengths (indicated in pseudocode as "newick_to_matrix"). This required reading the Newick format file created by phylogenetic software and determining the structure of the tree such the branch length between all pairs of nodes (species) in the tree could be calculated. The second function alluded to in the pseudocode, "calculate_k_score", calculates the K-score for each pair of

trees (now represented as matrices of branch lengths). In other words, it calculates a

scaling value between two protein trees, following the formula discussed above (Soria-

Carrasco *et al.,* 2007). The last function call indicated in the pseudocode is

"calculate_scaled_branch_lengths". This function takes the K-value calculated and scales

each matrix *m* such that two protein trees can be compared. This produces a matrix, *p*,

which quantifies the relatedness of each pair of proteins considered. The resulting matrix,

p, can then be transformed into a tree and visualized using any available software.

```
variables:
x= number of protein trees;
n= array of NEWICK format protein trees;
m= array of protein tree matrices;
k= array of k-scores;
p= PPI matrix;

for (i=0; i<x; i++)
{
        //Transform NEWICK file into matrix of pairwise branch lengths
        m[i]=newick_to_matrix(n[i]);
}
for (i=0; i<x; i++)
{
        for (j=0; j<x; j++)
        {
                //Calculate k-score (scaling factor) for each pair of protein trees
                k[i][j]=calculate_k_score(m[i],m[j]);
        }
}
for (i=0; i<x; i++)
{
        for (j=0; j<x; j++)
        {
                //Calculate scaled branch lengths
                p[i][j]=calculate_scaled_branch_lengths(m[i],m[j],k[i][j]);
        }
}
```

Figure 16. Pseudocode of the TOT software developed.

CHAPTER THREE

RESULTS AND DISCUSSION

As a proof-of-concept for the proposed approach, proteins involved in the

development of the auditory system and the visual system were chosen to validate the

results and test the capacity of the program to elucidate the evolution of unknown PPI.

Visual Protein	Literature Source
Atonal homolog 7 (ATOH7)	Brown *et al.,* 2002
Bassoon (BSN)	tom Dieck *et al.,* 1998
Guanine nucleotide binding protein (G protein), rod alpha transducing activity (GNAT1)	Morhardt *et al.,* 2009
Guanine nucleotide binding protein (G protein), rod beta (GNB1)	Morhardt *et al.,* 2009
Guanine nucleotide binding protein (G protein), rod gamma transducing activity (GNGT1)	Morhardt *et al.,* 2009
Short Wavelength Sensitive Opsin, cone (OPN1SW)	Larhammar *et al.,* 2009
Paired box 6 (PAX6)	Khan *et al.,* 2012
Phosphodiesterase 6A, cGMP-specific, rod, alpha (pDE6A)	Morhardt *et al.,* 2009
Phosphodiesterase 6B, cGMP-specific, rod, beta (pDE6B)	Morhardt *et al.,* 2009
Phosphodiesterase 6C, cGMP-specific, cone, alpha prime (pDE6C)	Morhardt *et al.,* 2009
Phosphodiesterase 6D, cGMP-specific, rod, delta (pDE6D)	Nancy *et al.,* 2002
Rhodopsin (RHO)	Larhammar *et al.,* 2009
C-terminal binding protein 2 (RIBEYE)	Zenisek *et al.,* 2004
Retinal G-protein receptor (RPE)	Radu *et al.,* 2008
Sine oculis homeobox homolog 3 (SIX3)	Anderson *et al.,* 2012
Sine oculis homeobox homolog 6 (SIX6)	Anderson *et al.,* 2012

Table 2. Sixteen visual proteins involved in the development of the visual system chosen
as a proof-of-concept for the proposed method.

Auditory Proteins	Literature Source
Cadherin-related 23 (CDH23)	Muller, 2008
Eyes absent homolog 1 (EYA1)	Zheng *et al.,* 2003
Fibroblast growth factor 3 (FGF3)	Chung *et al.,* 2011
Interferon regulatory factor 6 (IRF6)	Restivo *et al.,* 2011
Myosin VIIA (MYO7A)	Gillespie *et al.,* 2009
Orthodenticle homeobox 1 (OTX1)	Chatterjee *et al.,* 2010
Paired box 3 (PAX3)	Zhang *et al.,* 2012
Sine oculis homeobox homolog 1 (SIX1)	Zheng *et al.,* 2003
SRY (sex determining region Y)-box 10 (SOX10)	Dutton *et al.,* 2009
Suppressor of Ty 6 homolog (SPT6)	Keegan *et al.,* 2002
UDP-glucose 6-dehydrogenase (UGDH)	Busch-Nentwitch *et al.,* 2003

Table 3. Eleven auditory proteins involved in the development of the auditory system chosen as a proof-of-concept for the proposed method.

VISUAL SYSTEM RESULTS
INDIVIDUAL TREES OF THE VISUAL SYSTEM GNAT1 GNB1 GNGT1   VISUAL SYSTEM TREE OF TREES (TOT) G-PROTEIN (α), (β) (γ) subunit TOT PDE6 (α), (β) (γ) (δ) subunits TOT VISUAL TOT (16 proteins)
VISUAL SYSTEM DISCUSSION
G-PROTEIN (α), (β) (γ) subunit TOT PDE6 (α), (β) (γ) (δ) subunits TOT VISUAL TOT
AUDITORY SYSTEM RESULTS
INDIVIDUAL TREES OF THE AUDITORY SYSTEM MYO7A FGF3 SOX10   AUDITORY SYSTEM TREE OF TREES (TOT) AUDITORY TOT (11 proteins)
AUDITORY SYSTEM DISCUSSION
AUDITORY TOT

Table 4. Overview of Results and Discussion.

Visual System Results

Individaul Trees of the Visual System

Sixteen individual visual phylogenies were inferred using Maximum Likelihood from the visual proteins listed in Table 2. Each of the inferred individual visual phylogenies depicts the protein's evolutionary history across different vertebrate species. All of the inferred individual visual phylogenies had similar tree topology and were drawn on the same scale. In addition, the individual visual phylogenies produced trees that paralleled phylogenies inferred using other molecular markers such as ribosomal, mitochondrial or random amplified polymorphic DNA. Three individual visual trees of the signal transducing G-proteins, namely, GNAT1, GNB1 and GNGT1 will be discussed here (See Appendix C for additional phylogenies of proteins associated with the visual system).

The GNAT1, GNB1 and GNGT1 individual phylogenies describe the evolutionary history of extant vertebrate species based on G-proteins (Figure 17-19). Representative of the three G-protein trees, the branch lengths among all of the mammalian sequences are short and closely patterned, which indicates recent shared evolutionary history among all of the extant mammals as well as minimal protein sequence divergence. An area graph was generated to further illustrate the percent of sequence conservation of the GNAT1, which is also characteristic of the other G-proteins (Figure 20). Evolutionary relatedness can be estimated by the amount of sites conserved in a sequence. It can be observed from the graph that the GNAT1 has very few variable

sites, suggesting that the conserved amino acids sites are functionally important.

Moreover, the inferred individual G-protein trees paralleled phylogenies derived using

other molecular markers such as 18S rRNA, which was used to reconstruct the metazoan

tree of life (Meyer *et al.,* 2010).
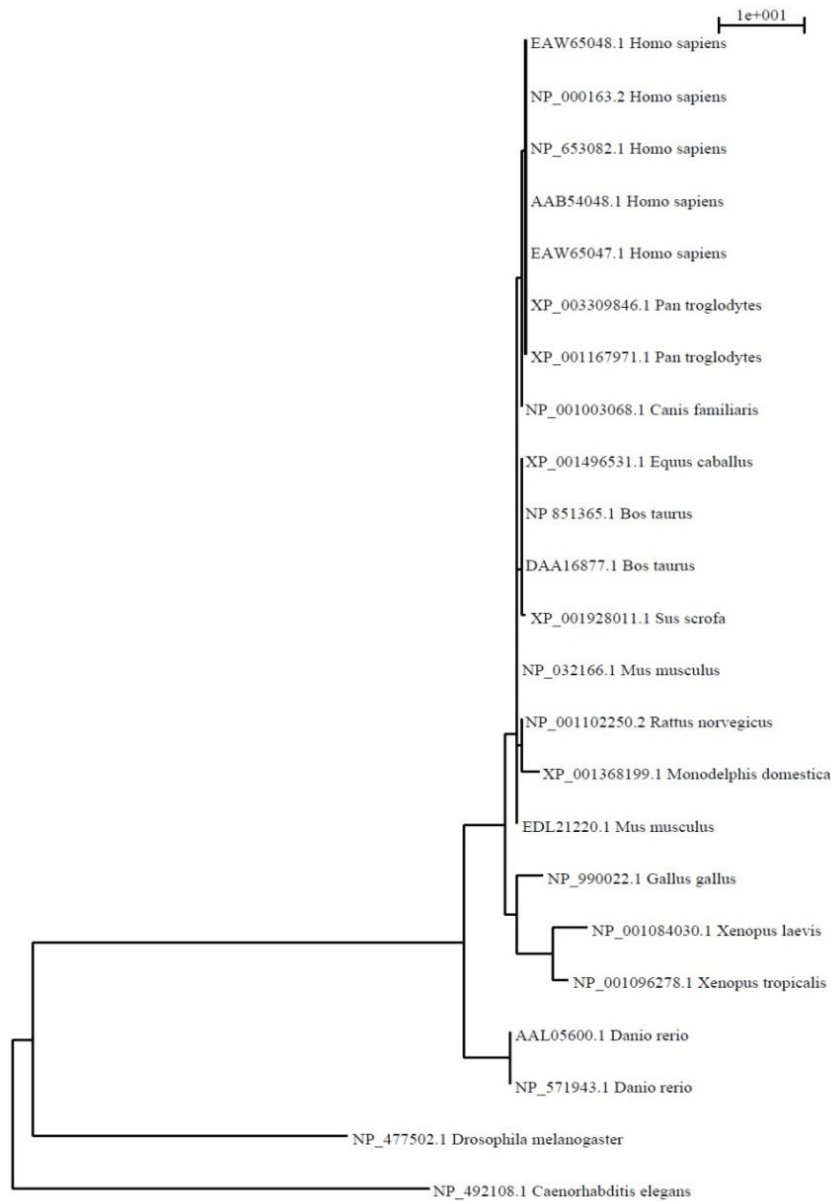
.
Figure 17. GNAT1 Tree derived using ProML from aligned protein sequnces.
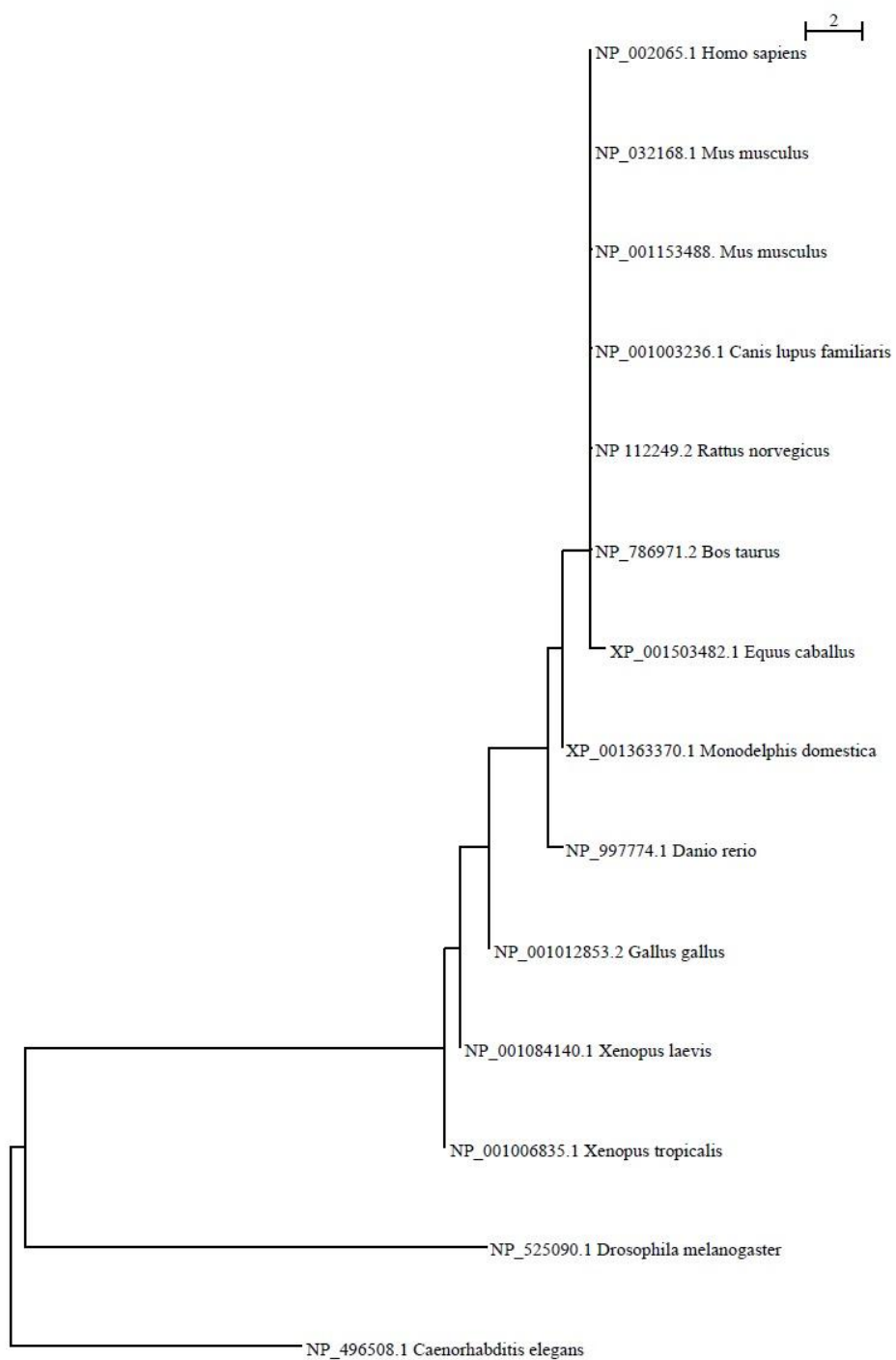
Figure 18. GNB1 Tree derived using ProML from aligned protein sequnces.
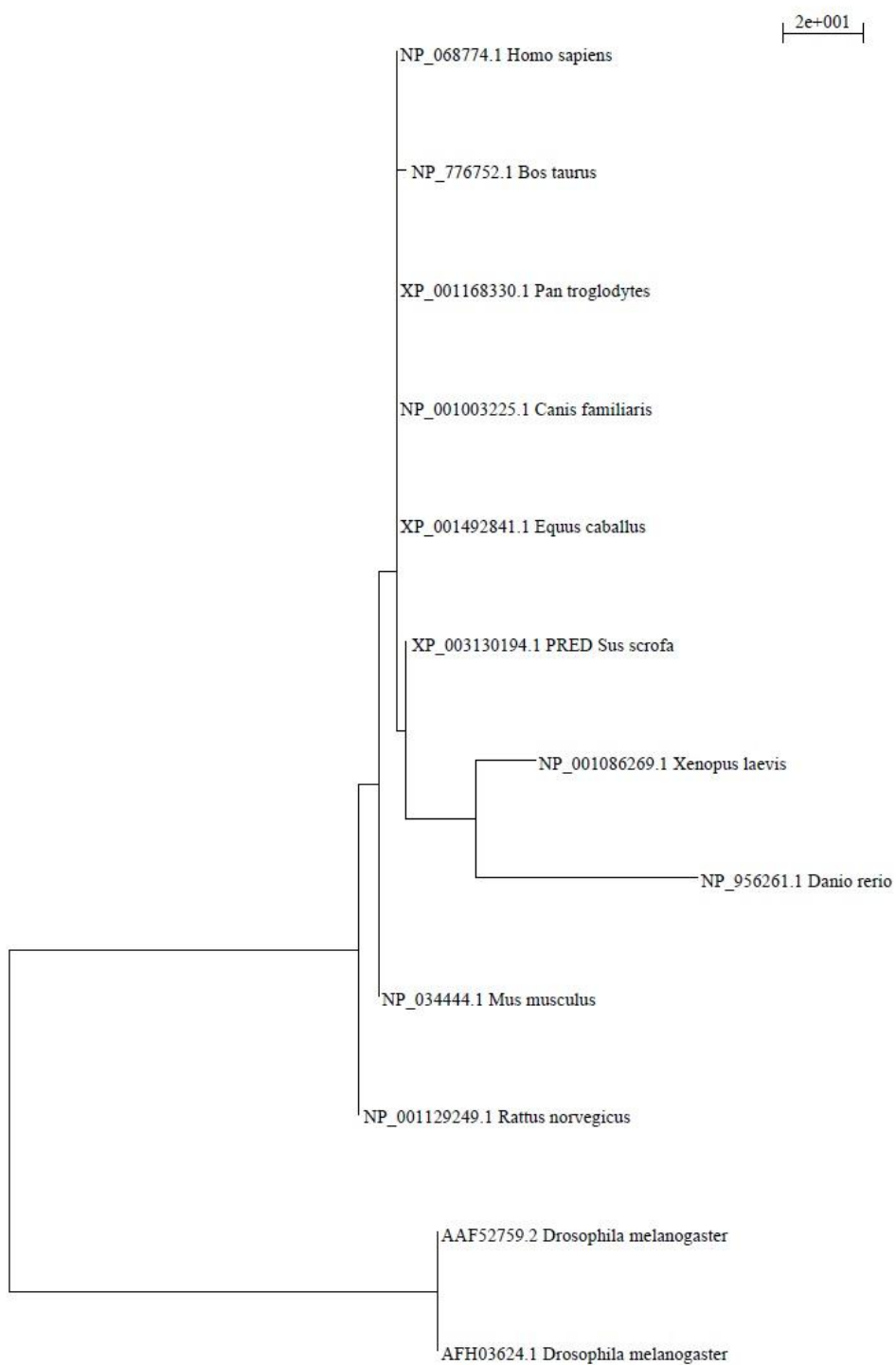
Figure 19. GNGT1 Tree derived using ProML from aligned protein sequnces.
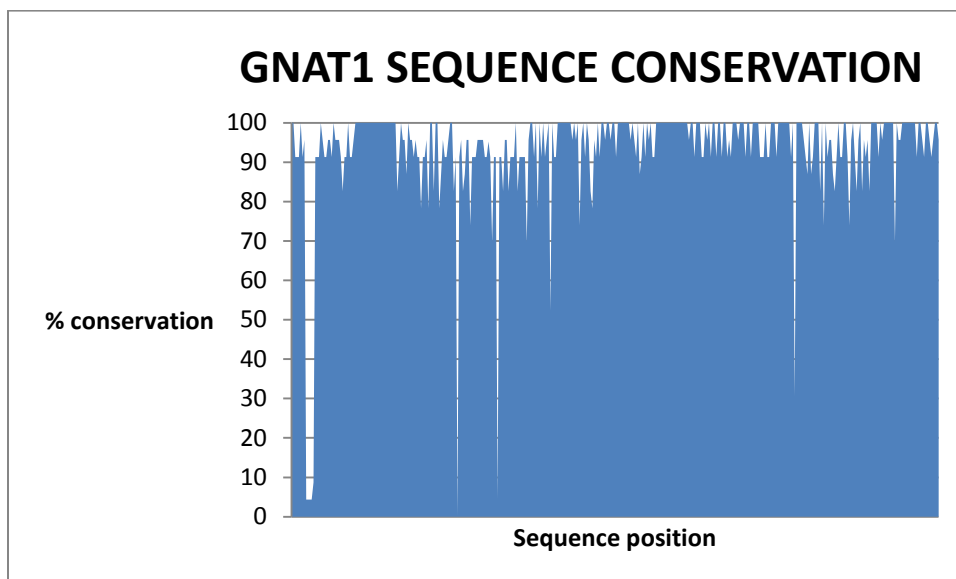
Figure 20. GNAT1 sequence conservation based on percent of identical amino acids within the alignment length.

Visual System Tree of Trees

G-PROTEIN (α), (β), (γ) subunits

In order to first test and validate the new method for predicting PPI based on coevolution, a visual Tree of Trees (TOT) was derived using only the three G-proteins (Figure 21). A TOT, as its name suggests, combines the information captured in the individual protein trees to identify proteins that have had a similar evolutionary trajectory which may be the result of coordinated evolution between interacting proteins. Consequently, the G-protein TOT was formed after scaling all of the individual protein trees with a K-score factor so that all of the trees had similar global divergence and then recalculating all of the branch lengths among different trees.

In the visual system, signal transducing G-proteins exist as heterotrimeric molecules that are activated in response to stimulation from sensory rhodopsin, this

stimulation triggers enzymatic activity and leads to dissociation of the hetrotrimeric unit.

Experimental studies show that the β and γ subunits form a tightly associated complex

that activates different signaling cascades (Morhardt *et al.,* 2009). As expected, the

derived G-protein TOT using GNAT1, GNB1 and GNGT1 correctly predicts that the

GNB1 and GNGT1 coevolved together (Figure 21).
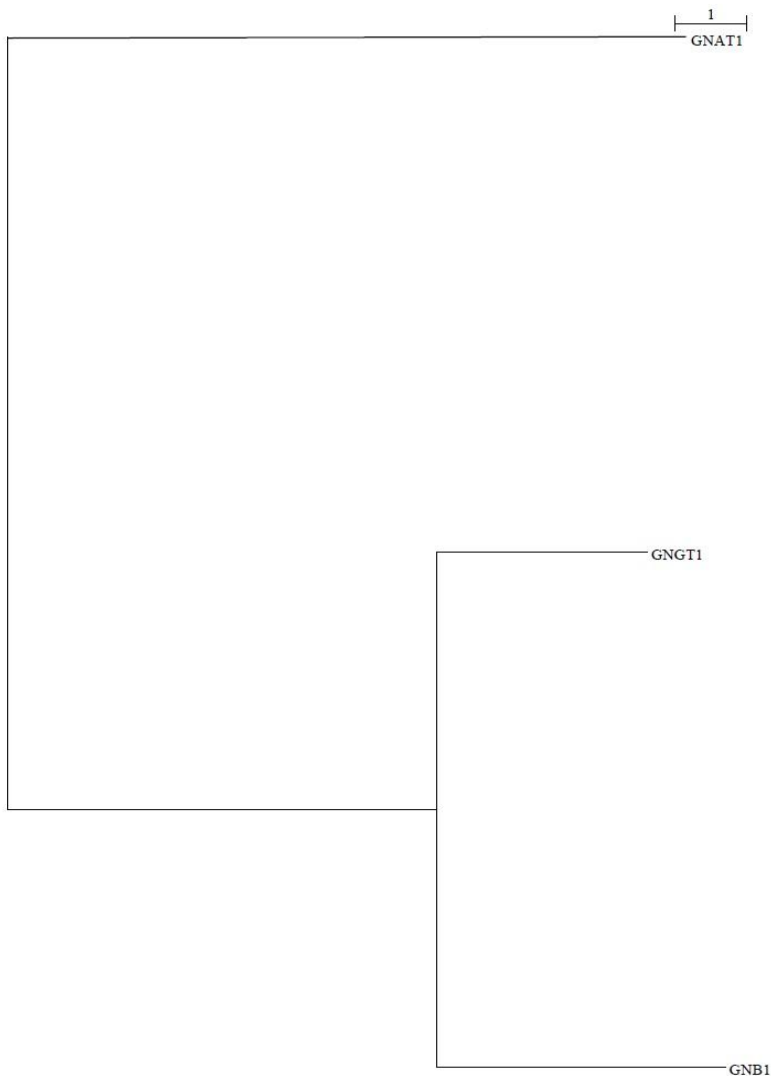
Figure 21. G-protein TOT derived using GNAT1, GNB1 and GNGT1.

The G-proteins of the visual system have been extensively studied and it is known that GNB1 and GNGT1 form a βγ complex that acts as a signaling molecule in visual phototransduction. Accordingly, a change in one protein will be mitigated by a compensatory change(s) in its binding partner, maintaining interaction and function in the face of evolutionary change. Thus, the obtained G-protein TOT shows that the TOT approach can correctly predict PPI, based on shared evolution.

PDE6 (α), (β), (γ), (δ) subunits

To further test and validate the TOT method, four phosphodiesterase-6 (PDE6A-D) subunits were used to derive the PDE6 TOT (Figure 22). In the visual system, rod and cone receptors use PDE6 enzymes that activate in response to light as part of the phototransduction cascade. PDE6 is composed of catalytic αβ dimers that activate intracellular protein kinases via the hydrolysis of secondary messengers in phototransduction in the eye (Kolandaivelu *et al.,* 2011). The inferred PDE6 TOT correctly predicted that the catalytic αβ subunits form a complex. In addition, the PDE6 TOT was able to differentiate between the cone and rod subunits and correctly predict that the α cone subunit coevolved with the β rod subunit. Moreover, the PDE6 TOT identified the δ subunit as being evolutionarily district from the other three PDE6 subunits. Thus, the PDE6 TOT also confirmed that the TOT method is successful in capturing PPI based on coevolution of protein partners.
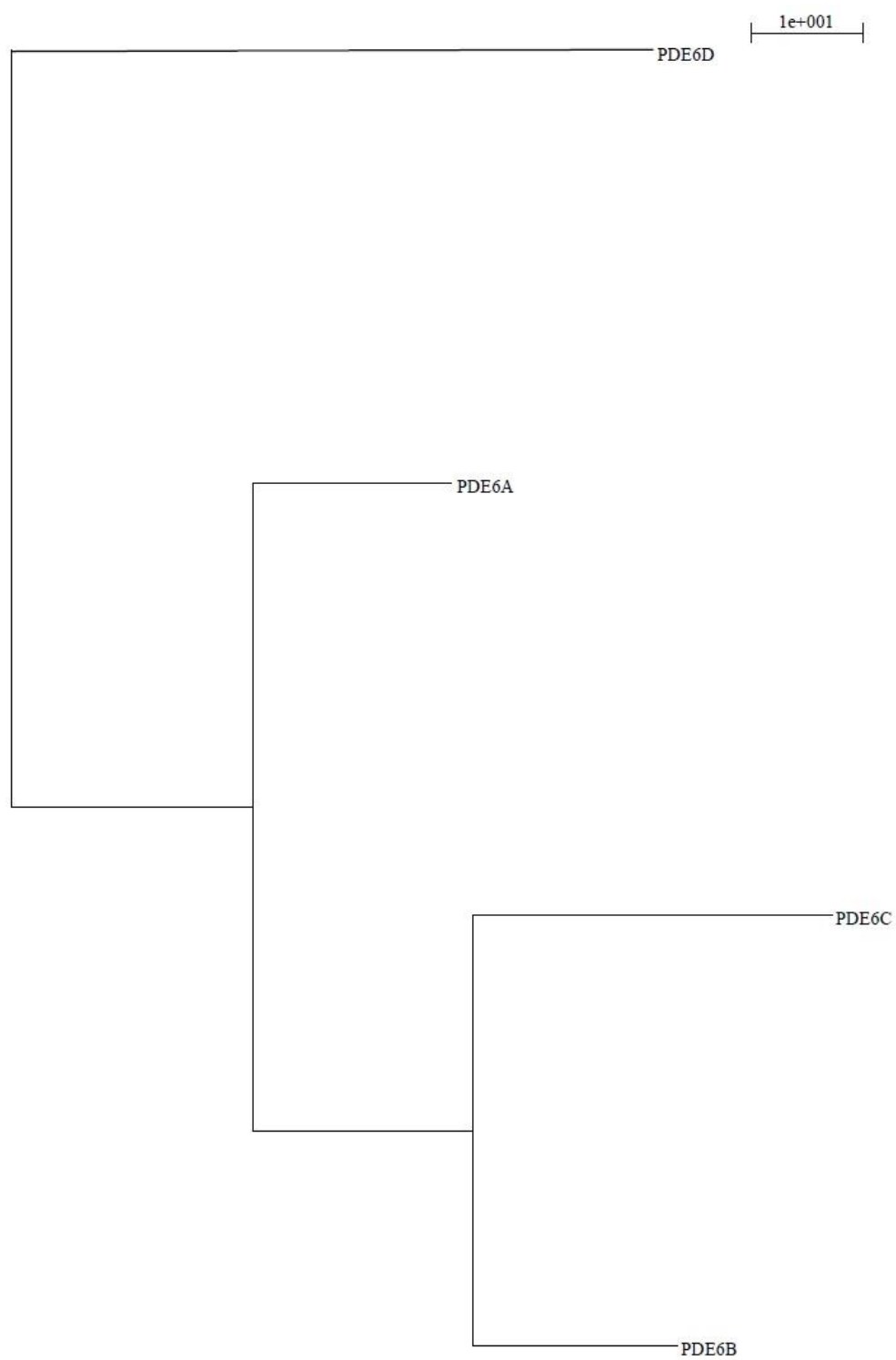
Figure 22. Phosphodiesterase-6 TOT inferred using four (PDE6A-D) subunits.

Visual TOT

The visual system TOT was deriving using sixteen different proteins involved in the development and function of the visual system listed in Table 2. The resulting visual TOT indicates that the interacting proteins evolved in clusters highly correlated with the proteins' tissue specificity, expression level during development and their direct and indirect interaction (Figure 23). The three clusters observed in the visual system TOT are: (PDE6C, RIBEYE, RPE, PDE6D, PAX6, GNAT1), (OPN1SW, PDE6A, RHO, GNGT1, SIX6, GNB1) and (ATOH7, BSN). PDE6B and SIX3 fall outside of the clusters pointing that they did not coevolve directly with the included visual proteins. The inferred TOT only represents a snapshot of all the PPI involved in the visual system. As such PDE6B and SIX3 may have shared similar evolutionary histories with proteins for which they interact, but those proteins are not included in the set of proteins which were considered here. There are many more proteins that participate in the development and function of the eye, however including all of them is beyond the range of this project. The three clusters in the visual system TOT indicate that the grouped proteins coevolved together by undergoing reciprocal changes in order to maintain interaction and their function.
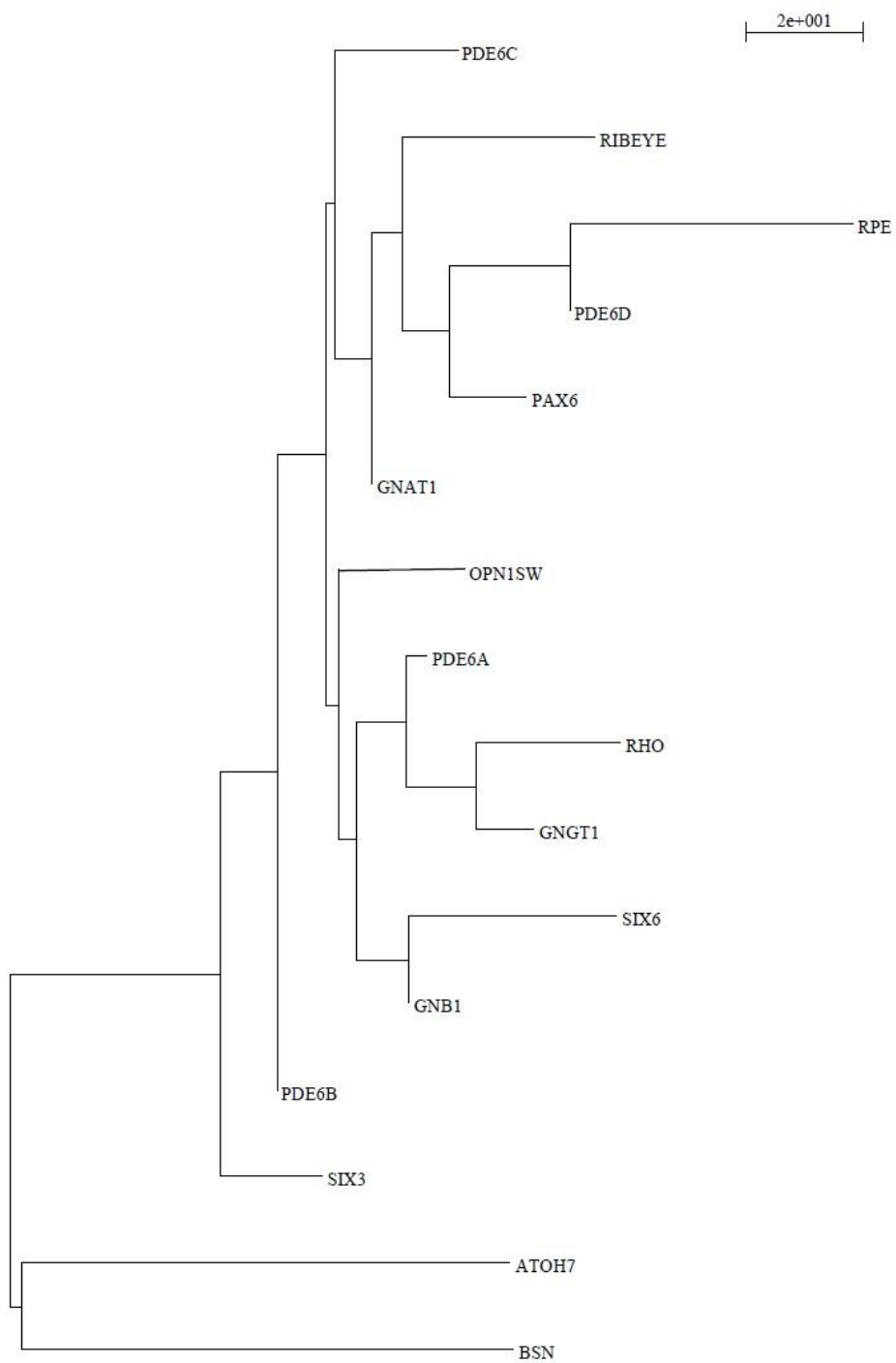
Figure 23. Visual System TOT inferred using sixteen proteins involved in the development and function of the visual system.

Discussion of Visual System Tree of Trees

G-PROTEIN (α), (β), (γ) subunits TOT

The G-proteins are important transducins in the vertebrate phototransduction that are naturally expressed in the rod and cone photoreceptor cells. The G-proteins are heterotrimeric macromolecules composed of the alpha (α), beta (β) and gamma (γ) subunits. Homologous G-proteins such as GNAT1, GNB1, GNGT1 have distinct *α, β, γ* subunits for both rod and cone photoreceptors in various vertebrates. The G-proteins of the visual system have been extensively studied and it is known that GNB1 and GNGT1 form a βγ complex, which when released from the α subunit acts as a signaling molecule that activates secondary massagers. The βγ complex is only released from the α subunit when the G-protein is activated catalyzing a GDP-GTP exchange, thereby releasing the βγ complex from the α subunit (Figure 24). In turn, the α subunit activates phosphodiesterase (PDE6A), a third enzyme in the phototransduction cascade capable of hydrolyzing cyclic guanosine monophosphate (cGMP) (Morhardt *et al.,* 2009).

Figure 24. G-protein signal transduction cascade. [Reproduced from Burns *et al.,* 2005]

The G-protein TOT correctly predicts the GNB1 and GNGT1 coevolved together as a protein pair (Figure 21). In order for the GNB1 and GNGT1 to continue interacting in vivo, the two proteins had to coevolve together by not only participating in the same

system but also undergoing similar evolutionary pressures and selections. For GNB1 and GNGT1 to coevolve together in the visual system, any changes in one protein had to be reciprocated in its partner. In the case of the GNB1 and GNGT1, the βγ complex has been conserved throughout the vertebrate evolutionary history, which suggests that these proteins coevolved together so that they can continue interacting and playing their roles in the visual phototranduction.

PDE6 (α), (β), (γ), (δ) subunits TOT

PDE6, as its name implies, is an enzyme that breaks phosphodiester bonds of cyclic guanosine monophosphate (cGMP). In the visual system, the role of cGMP is to keep the sodium channels in the rod's membrane open (Figure 25). PDE6A hydrolyzes cGMP to 5'-GMP, which causes the sodium ion channels in photoreceptors to close, leading to signal amplification. This hyperpolarization of the rod's membrane potential generates electrical response that is propagated to other retinal neurons and ultimately to the brain (Morhardt *et al.,* 2009). Therefore, PDE6A is an important regulator of signal transduction mediated by these second messenger molecules.

PDE6 is formed by two large catalytic α and β subunits and two smaller γ subunits as well as a delta δ subunit. The catalytic α and β subunits form a dimer that binds and hydrolyzes cGMP. The γ subunits inhibit the hydrolytic activity of the α and β dimer. The role of the delta subunit is unknown as it does not affect the catalytic activity of PDE6 (Nancy *et al.,* 2002).
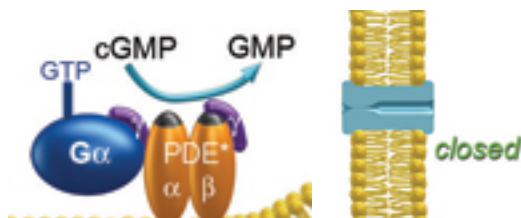
Figure 25. Phosphodiesterase (PDE6) activation cascade. [Reproduced from Burns *et al.,* 2005]

PDE6 TOT correctly predicts that the catalytic α and β subunits form a complex. In addition, the PDE6 TOT is capable of differentiating between the cone and rod subunits and correctly predicts that the α cone subunit coevolved with the β rod subunit. This is because either rod or cone PDE6 can successfully bind to the cone phototransduction pathway to mediate visual signaling (Kolandaivelu *et al.,* 2011). In a similar manner, rod (PDE6) catalytic α,β subunits restore cone function in a mouse model lacking cone PDE6 catalytic subunit. On the other hand, PDE6 falls outside of the formed αβ complexes. This observation agrees with the current finding that PDE6D does not affect the catalytic activity of the other subunits (Nancy *et al.,* 2002).

Thus, the clustering of the α,β subunits in the PDE6 TOT points out that in order for the subunits to continue their PPI the α,β subunits had to coevolve together as a pair throughout the vertebrate evolutionary history. Moreover, mutations in the PDE6 catalytic subunits cause retinal degeneration resulting in vision impairment and blindness (Burns *et al.,* 2005). As a result, reciprocal changes in the PDE6 catalytic subunits that maintained the PPI were selected for in order to conserve the important role of the subunits in the phototransduction of the visual system.

Visual TOT

The visual system TOT forms three clusters (PDE6C, RIBEYE, RPE, PDE6D, PAX6, GNAT1), (OPN1SW, PDE6A, RHO, GNGT1, SIX6, GNB1) and (ATOH7, BSN), which predict the visual PPI network based on coevolution of protein pairs. The inferred three clusters of the visual system TOT contain protein interactions that have been identified by previous experimnets as well as suggest novel and unexpected interactions.

PDE6C, RIBEYE, RPE, PDE6D, PAX6 and GNAT1 cluster

The first cluster in the visual TOT includes the PDE6C, RIBEYE, RPE, PDE6D, PAX6 and GNAT1 proteins (Figure 26). This group contains some proteins involved in the development and some in the function of the visual system. For instance, PDE6C, PDE6D, RPE and GNAT1 can be traced to the visual phototransduction pathway, whereas, PAX6 and RIBEYE play different roles discussed in more detail in the following section.
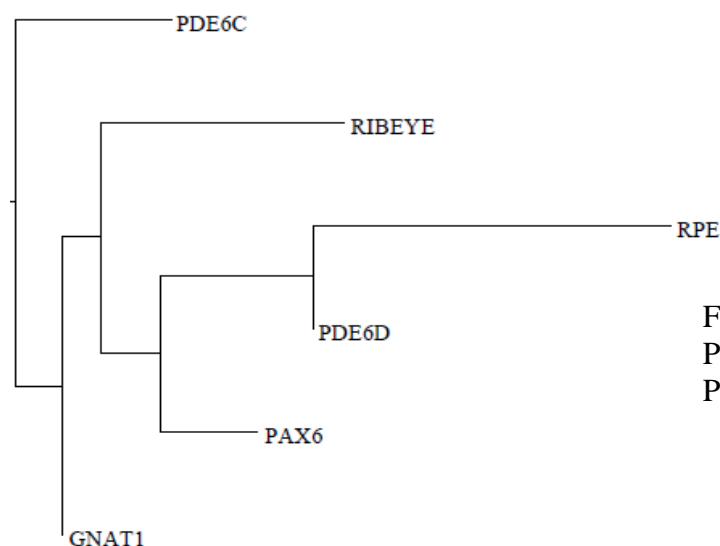
Figure 26. PPI cluster including PDE6C, RIBEYE, RPE, PDE6D, PAX6 and GNAT1.

The main protein in the first cluster of the visual TOT is PAX6, which is the master eye controller of vertebrate eye development (Brown *et al.,* 2002). PAX6 is a highly conserved transcription factor in both vertebrates and invertebrates. PAX6 is universally expressed in the developing eye tissues of various species including all vertebrates as well as the squid, flatworm, sea urchin and nematode. An ancestral PAX6 might have been involved in the development of a primitive eye. In addition, PAX6 directs the fate of multipotent embryonic and fetal cells, predetermined to develop into the adult retina, also known as retinal progenitor cells (RPCs). Considering their ontogenetic role in the early eye morphogenesis, the developmental population of RPCs must have the potential to differentiate into each of the six neuronal cell types (ganglion, horizontal, amacrine, cone, rod, or bipolar cells) or one glial cell type (Müller cell) present in the mature neural retina (Schmitt *et al*., 2009). Experimental evidence suggests that PAX6 is necessary for the multipotent state of RPCs (Marquardt *et al*., 2001).

The eye develops from the out pocketing of the neural tube called the optic vesicles, the epidermis, and the periocular mesenchyme which receive support from both the neural crest and mesoderm lineages. Once the otic vesicle comes into contact with the epithelium it thickens and forms the lens placode which signals back to the otic vesicle and allows it to transform into the otic cup that differentiates into the neural retina and the retinal pigment epithelium (Gilbert, 2000). PAX6 is expressed during the early eye development as well as regulates the migration of neural crest cells to form the lens placode (Khan *et al.,* 2012).

Mutations in PAX6 are associated with similar eye anomalies in human, rat and mouse. In humans, disruptions of PAX6 are responsible for different congenital disorders, e.g. aniridia and Peter's anomaly, which lead to variable eye malformations. Aniridia is characterized by a reduction or complete absence of the iris and is often accompanied by further defects in the cornea, lens, retina and optic nerve. Peter's anomaly is most often associated with malformations of the anterior chamber of the eye (Halder *et al.,* 1995). PAX6 is well-known for its retina-promoting activity and also plays a crucial role in early pigment epithelium development. As a result, PAX6 interacting partners must also play important ancillary roles in the eye development.

The visual TOT suggests that PAX6 shares similar evolutionary history with RPE, the retinal G protein-coupled receptor is part of the phototransduction regeneration pathway that binds all-trans- and 11-cis-retinal (Figure 27). Its biological function is essential for phototransduction, protein-chromophore linkage and visual perception. RPE acts as a photoisomerase and catalyzes the conversion of all-trans-retinal to 11-cis-retinal following light absorption (Radu *et al.,* 2008). RPE is expressed in tissue adjacent to retinal photoreceptor cells, the retinal pigment epithelium and Müller cells of the neural retina (Radu *et al.,* 2008).
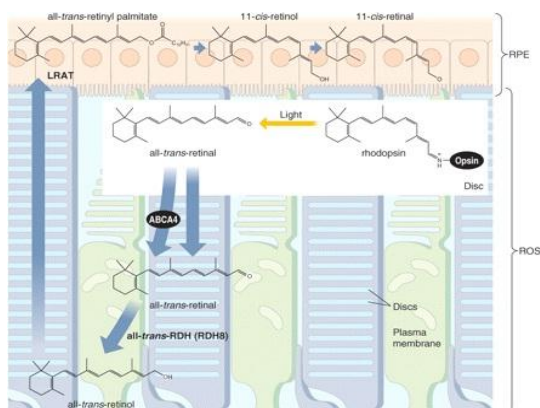
Figure 27. RPE visual pigment regeneration via all-trans-retinol to 11-cis-retinal conversion. [Reproduced from Maeda *et al.,* 2009]

Mutations in RPE are associated with Leber's congenital amaurosis type 2 (LCA2), which is characterized by severe loss of vision or blindness. In addition, mutations in RPE can also lead to retinitis pigmentosa, a progressive degeneration of the retinal cells. This is because RPE also plays an essential role in maintaining the viability of the neighboring photoreceptor cells (Radu *et al.,* 2008).

The visual TOT also predicts that RIBEYE coevolved in the same PPI cluster as PAX6 and RPE. RIBEYE is a major component of the synaptic ribbons of the retina that transmit sensory signal via the release of neurotransmitters (Figure 28). Vertebrates have ribbon synapses in the retina and in other sensory structures that are specialized for rapid, tonic release of synaptic vesicles at the synapse between cells (Zenisek *et al.,* 2004).
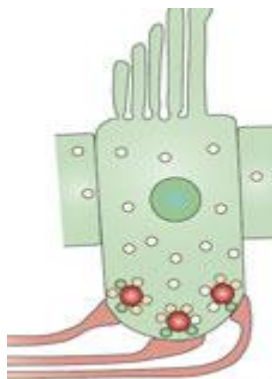
Figure 28. Synaptic ribbons are shown in dark red. Vesicles attached to ribbons are shown by yellow circles and docked vesicles by green circles. [Reproduced from Matthews & Fuchs, 2010]

Ribbon synapses release neurotransmitters tonically, with a high release rate made possible by continuous docking of synaptic vesicles on presynaptic ribbons. In the vertebrate visual system, synaptic ribbons serve to increase the rate of synaptic vesicle docking via an ongoing cycle of exocytosis and endocytosis in response to graded changes of membrane potential and thereby bringing a continuous flow of synaptic vesicles to the active zones (Zenisek *et al.,* 2004). In addition, synaptic ribbons act like a conveyor belt capturing synaptic vesicles from the cytosol and transporting them to the active zone, a site of synaptic vesicle docking and neurotransmitter release. Similar to PAX6, RIBEYE has also been highly conserved among all vertebrate and invertebrate species including *D. melanogaster* and *C. elegans.*

Mutations in RIBEYE lead to disorders that affect both vision and hearing, including the various forms of Usher syndrome that are characterized by deafness and gradual loss of vision (Schmitz *et al.,* 2000). There are three clinical types of Usher syndrome: type 1, type 2, and type 3. The major symptoms of all the forms of Usher syndrome are hearing loss and progressive degeneration of the retinal cells called retinitis pigmentosa (Yang *et al.,* 2012). As the disease progresses, it leads to night blindness and the loss of peripheral vision. In addition, many people with Usher syndrome also have severe balance problems which points out the intricate inter-connection among the vestibular, auditory and visual proteins.

Thus, the inferred clustering of the PDE6C, RIBEYE, RPE, PDE6D, PAX6 and GNAT1 proteins in the visual TOT provides an interesting picture as it clusters proteins

that are expressed in various tissues at different times. For instance, PDE6C, PDE6D, RPE and GNAT1 are phototransducion proteins and their clustering has experimental support (Burns *et al*., 2005; Kolandaivelu *et al*., 2011; Morhardt *et al.,* 2009). In contrast, PAX6 is a transcription factor that localizes to the nucleus and RIBEYE is a component of the synaptic ribbons which is expressed at the synapse. Yet, RIBEYE is composed of two functional domains, a unique A domain specific for ribbons, and a B domain identical with CtBP2, a transcriptional repressor complex (Schmitz *et al.,* 2000). The CtBP2, domain could possibly regulate the transcription of PAX6, however, further experimental work is needed to test if this interaction really exists. The unexpected clustering of phototransducion proteins such as PDE6C, PDE6D, RPE and GNAT1 with transcription factor PAX6 may reflect the fact that: 1) these proteins evolved at the same rate but do not interact; 2) the proteins indirectly interact, as there are numerous other proteins involved in these pathways; 3) there exists high "background noise" within the sequence data.

OPN1SW, PDE6A, RHO, GNGT1, SIX6 and GNB1 cluster

The visual TOT also suggests a cluster consisting of the OPN1SW, PDE6A, RHO, GNGT1, SIX6 and GNB1 proteins (Figure 29), all of which have been experimentally identified to be involved in the visual phototransduction cascade (Burns *et al*., 2005; Kolandaivelu *et al*., 2011; Morhardt *et al.,* 2009) and are further discussed below.
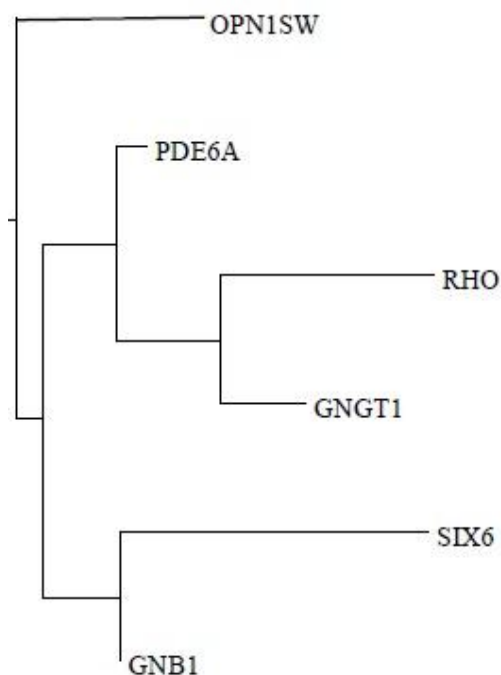
Figure 29. Visual TOT cluster comprised of the OPN1SW, PDE6A, RHO, GNGT1 and SIX6, and GNB1 proteins.

Vision begins when a visual pigment molecule such as RHODOPSIN (RHO) that is expressed in the rods cells absorbs a photon of light (Figure 30). The photon activated RHO begins a cascade of changes, including the isomerization of the chromophore, a vitamin A derivative molecule that absorbs certain wavelengths of visible light and transmits or reflects others (Larhammar *et al.,* 2009). Similarly, OPN1SW is a visual pigment protein responsible for short wavelength perception of approximately 475 nm, which is the blue region of the electromagnetic spectrum. As a result, it is not surprising that the visual TOT, clusters OPN1SW with the other proteins involved in the visual phototransduction cascade as opsins are the universal photoreceptors of light of all vertebrate and invertebrate visual systems. Experimental data also confirm that molecules that capture light photons such as OPN1SW and RHO initiate the signaling cascade that produces physiological responses (Shichida & Matsuyama, 2009).
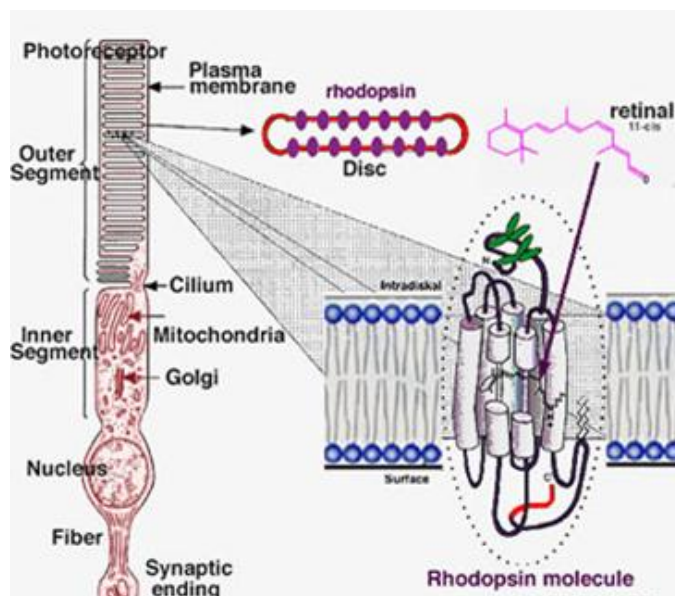
Figure 30. Schematic diagram of the rod photoreceptor and its rhodopsin pigment molecule. [Reproduced from Hargrave, 1992]

Following isomerization, RHO is transformed into a metarhodopsin, which activates a second membrane-bound protein in the rod called transducin, the G-protein, to exchange GDP for GTP. The heterotrimeric G-proteins then dissociate in the presence of GTP and form a stable dimeric βγ complex. The dissociated rod α subunit activates rod-resident phosphodiesterase (PDE6A), a third enzyme in the cascade, which is capable of hydrolyzing cyclic guanosine monophosphate (cGMP) (Morhardt *et al.,* 2009). The hydrolysis of cGMP to 5'-GMP amplifies the signal by closing the sodium ion channels in photoreceptor cells. As a result, sodium ions can no longer enter the cell, and the photoreceptor outer segment membrane becomes hyperpolarized due to the charge inside the membrane becoming more negative. This change in the cell's membrane potential causes voltage-gated calcium channels to close causing a drop in the amount of neurotransmitter released. Reduction in the neurotransmitter release means one

population of bipolar cells will be depolarized and a separate population of bipolar cells will be hyperpolarized, depending on the nature of receptors. Hyperpolarization of bipolar cells signals the ganglion cells which send the visual signal to the brain. Eventually the α subunit will begin hydrolyzing GTP to GDP and re-associate with the βγ complex and await RHO activation.

The visual TOT also inferred that SIX6 is included in the same cluster as OPN1SW, PDE6A, RHO, GNGT1, and GNB1 proteins. SIX6 is a transcription factor and repressor during eye development. For instance, SIX6 represses transcription of cyclin dependent kinase inhibitor (Cdkn1b), and thus promotes proliferation of retinal precursor cells. Interactions with specific co-factors allow SIX family proteins to either activate or repress transcription of downstream targets (Anderson *et al.,* 2012). The most extensive evidence for SIX6 proteins functioning as transcriptional repressors within the retina comes from studies of OPTIX and its mammalian orthologs SIX3 and SIX6. SIX3 and SIX6 appear to bind physically to Groucho/transducin-like Enhancer (GRG/TLE) family members that repress transcription and in turn are crucial for promoting retinal growth and differentiation (Chen & Courey, 2000). In addition, experimental studies also confirm that SIX6 is expressed during the early stages of visual system development and is expressed in the lens placode, lens epithelium, and the retina (Zhu *et. al.,* 2002). Thus, mutants of SIX6 in mice, humans and zebrafish display similar developmental deficits affecting multiple placodal derivatives (Schlosser *et al.,* 2008).

As a result, clustering of phototransduction proteins such as OPN1SW, PDE6A, RHO, GNGT1, and GNB1 with a transcription factor like SIX6 that attaches itself to specific sequences of DNA adjacent to the genes it regulates is highly unlikely. Such a result is possible when proteins involved in various functions and times during development are compared. One possibility is that the phototransduction proteins evolved at the same rate as the transcription factors, or as previously discussed PAX6, these proteins interacted indirectly and/or were subject to data interference as a result of background noise.

ATOH7 and BSN cluster

The final cluster formed by the visual TOT groups the ATHO7 and BSN proteins together (Figure 31). Experimental studies indicate that both ATOH7 and BSN play a role in the delivery of visual signal to the brain (Brown *et al.,* 2002; tom Dieck *et al*., 1998), and are further described in this section. To date, a direct interaction between ATHO7 and BSN has not been proven.
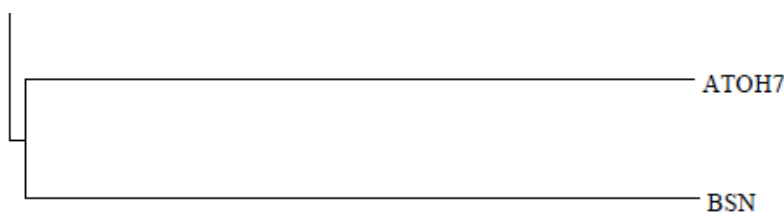
Figure 31. Visual TOT cluster of ATOH7 and BSN proteins.

ATHO7 regulates the genesis of retinal ganglion cells, through which all visual and photosensory information is transmitted to the human brain. ATOH7 is a

transcription factor with a basic helix-loop-helix (bHLH) protein structural motif. bHLH

transcription factors are central to retinal neurogenesis and regulate multiple aspects of

retinal neuron formation in vertebrates as well as invertebrates. Not only is ATOH7

required for RGC development but it is one of the first transcription factors to be

expressed in retinal progenitor cells and this expression coincides with the onset of

neurogenesis (Khan *et al.,* 2012). ATHO7 is also responsible for the development of the

otic nerve and optic nerve as well as the fate of multipotent cells (Brown *et al.,* 2001).

In addition, ATOH7 is involved in the development of the anterior structures of

the eye such as the cornea, iris and the lens. As a result, ATOH7 mutations in humans

lead to bilateral retinal detachments and microcornea, an abnormally thin (less than 11

mm horizontal diameter) and flat cornea also known as "small eye". Lastly, ATOH7

mutations in humans also cause vitreo-retinal dysplasia, also known as Peters' anomaly,

which is characterized by a congenital bilateral nonattachment of the retina (Brown *et al.,*

2002). Furthermore, murine ATOH7 mutations manifest in a reduced number of retinal

ganglion cells, lack of optic nerve and increased number of cones through a switch in cell

specification. Consequently, ATOH7 plays an important role in the formation of retinal

ganglion cells and its mutations in humans cause congenital malformations and

degenerative diseases of the optic nerve (Brown *et al.,* 2002).

BSN (BASSOON) is a protein involved in presynaptic cytomatrix organization at

the site of neurotransmitter release (Figure 32). Unlike RIBEYE that localizes to ribbon

synapses, BSN is a major component of the active zones (Schmitz *et al.,* 2000). BSN is

localized between the two sub-compartments of the ribbon synapses linking them together. Consequently, BSN is detected in axon terminals of hippocampal neurons where it is highly concentrated at the vicinity of the active zone. The lack of functional BSN significantly impacts synaptic functioning and structure. When BSN is functionally disrupted, the ribbons float free in the cytoplasm of the photoreceptor terminal (Matthews & Fuchs, 2010).
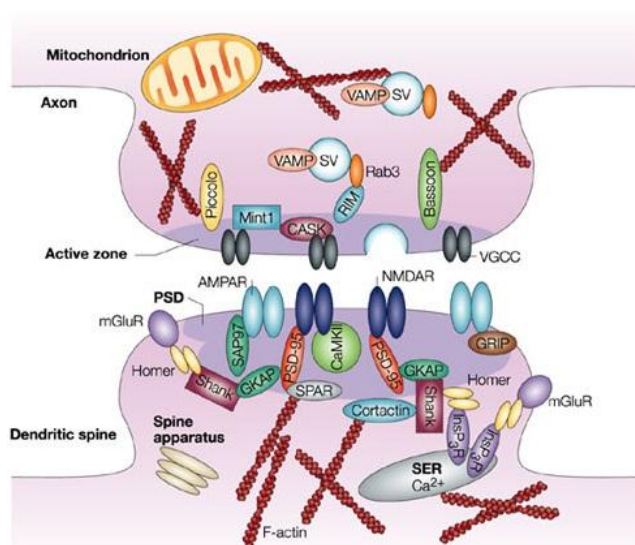
Figure 32. Schematic diagram of the presynaptic terminal containing synaptic vesicles. [Reproduced from Li & Sheng, 2003]

The visual TOT reveals that BSN and ATOH7 evolved at similar rates over the course of vertebrate evolution. The visual TOT suggests that ATOH7 and BSN interact directly on the protein-protein level because the two are depicted as a protein pair on the third cluster. However, experimental work suggests the expected interaction is between BSN and RIBEYE, which is a central portion of ribbon synapses (Schmitz *et al.,* 2000). In addition, BSN is predicted via computational analysis, to have three coiled-coil-

forming domains of different lengths that may play a role in the interaction of BSN with

other presynaptic proteins such as RIBEYE (tom Dieck *et al*.,1998). The unexpected

predicted interaction of ATOH7 and BSN by the visual TOT may be due to these proteins

having shared evolutionary rates without interactions because they are expressed in

different locations within the cell. However, it could in fact be that BSN and RIBEYE do

not interact as previous work has only suggested or predicted this interaction.

Auditory System Results

Individual Trees of the Auditory System

Eleven individual auditory phylogenies were inferred using Maximum Likelihood

from the hearing proteins listed in Table 3. Each of the inferred individual auditory

phylogenies depicts the protein's evolutionary history across different species. All of the

inferred individual auditory phylogenies had similar tree topologies, agreeing with

phylogenies derived using other molecular markers such as ribosomal, mitochondrial or

random amplified polymorphic DNA. Three individual auditory phylogenies of FGF3,

MYO7A and SOX10 are described here to illustrate the evolutionary relatedness of the

visual proteins (See Appendix D for all of the individual auditory phylogenies).

FGF3 Tree

FGF3 plays a critical role in the induction of otic tissues in all vertebrates and its

signaling is required for normal otic placode formation, maintenance and inner ear

patterning. The FGF3 phylogeny illustrates the evolutionary relationship between extant

species based on FGF3 protein (Figure 33). The FGF3 tree has a scale bar of 5e 001,

representing 5 substitutions per 100 amino acids. In the FGF3 phylogeny, *C. elegans*

represents the outgroup, as it is the most distantly related to all of the examined

vertebrates. The FGF3 tree topology groups the species into five major clades (*H.*

sapiens, P. troglodytes) (M. musculus, R. norvegicus) (Bos taurus, C. familiaris) (G.

gallus, M. domestica, *X. laevis, Danio rerio)* and *(D. melanogaster* and *C. elegans).* The

branch lengths between the (*H. sapiens, P. troglodytes)* and *(M. musculus, R. norvegicus)*

clades are relatively short, pointing out that these species share a recent common

ancestor. Consequently, the branch patterning of the FGF3 tree also illustrates the gradual

vertebrate emergence.
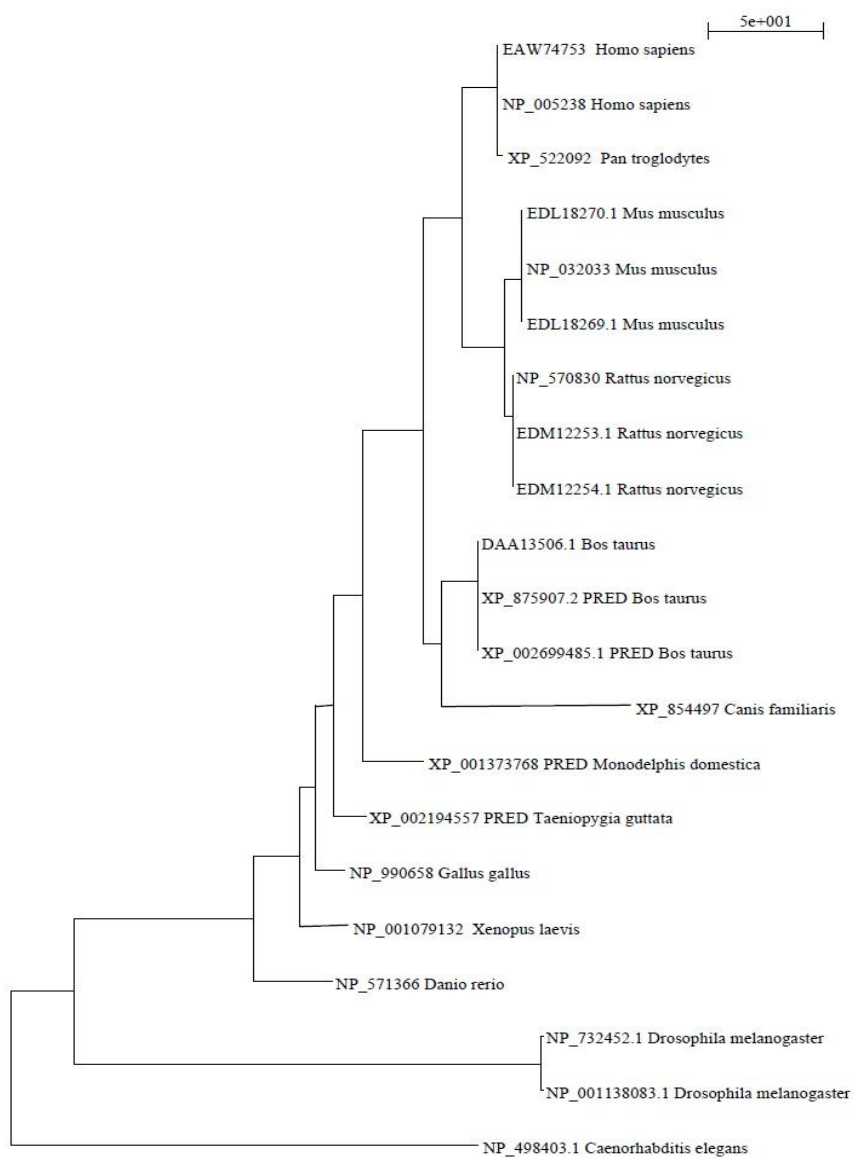
Figure 33. FGF3 Tree derived using ProML from aligned protein sequnces.

In addition, the FGF3 topology depicts *C. familiaris* as a sister taxa to *B. taurus* (Figure 34). The longer branch length of *C. familiaris* captures the lineage splitting event that took place over 15,000 years ago and produced *the C. familiaris*. As a result, the FGF3 phylogeny reflects the close genetic relationships among species modeling on the tree of life.
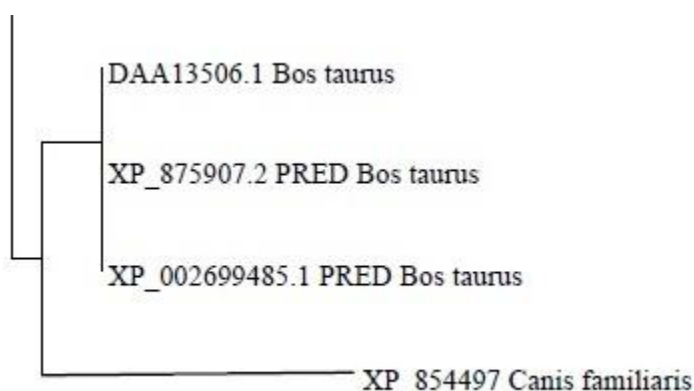
Figure 34. Speciation event of *C. familiaris*.

MYO7A Tree

MYO7A is a mechanochemical protein important for the proper structure and function of stereocilia, which convert sound waves to nerve impulses (Figure 35). The MYO7A phylogeny is drawn on the same scale as the FGF3 tree. Here again, *C. elegans* represents the outgroup as it is most distantly related to all of the examined vertebrates. *C. elegans* shares the most recent common ancestor *D. melanogaster* which is also an invertebrate. Similar to the FGF3 phylogeny, the MYO7a tree topology also groups the species into five major clades and the branch lengths between the most recently derived mammalian clades are relatively short pointing that they all share a recent common

ancestor. The branch lengths within the clade of aves, amphibians and fish are also short; however, the branch length that connects the aves, amphibians and fish clade to the mammals is, as expected, much longer reflecting their distant relation. Consequently, this tree captures the conserved evolutionary history of MYO7a across different vertebrates.

SOX10 Tree

The SOX10 phylogeny also illustrates the evolution of this protein within vertebrates (Figure 36). SOX10 plays an important role in development of the neural crest, which gives rise to neurons and glia of the peripheral nervous system. In addition, SOX10 encodes transcription factors that are required for the patterning of otic epithelium and normal development of the otic vesicle and are thus essential for proper auditory development. The SOX10 tree includes short branch lengths for all of the mammals examined. This close branch patterning points out that all of the mammals depicted in SOX10 tree share a recent common ancestor. The branch lengths also show a delineation between the placental mammals and vertebrates that lay eggs outside of their body such as the *(G. gallus, M. domestica*, *X. laevis, D. rerio)* clade. The SOX10 tree topology is similar to the topology of MYO7A and FGF3 phylogenies with the same scale. The similarity in scale suggests that FGF3, MYO7a and SOX10 evolved at a similar rate.
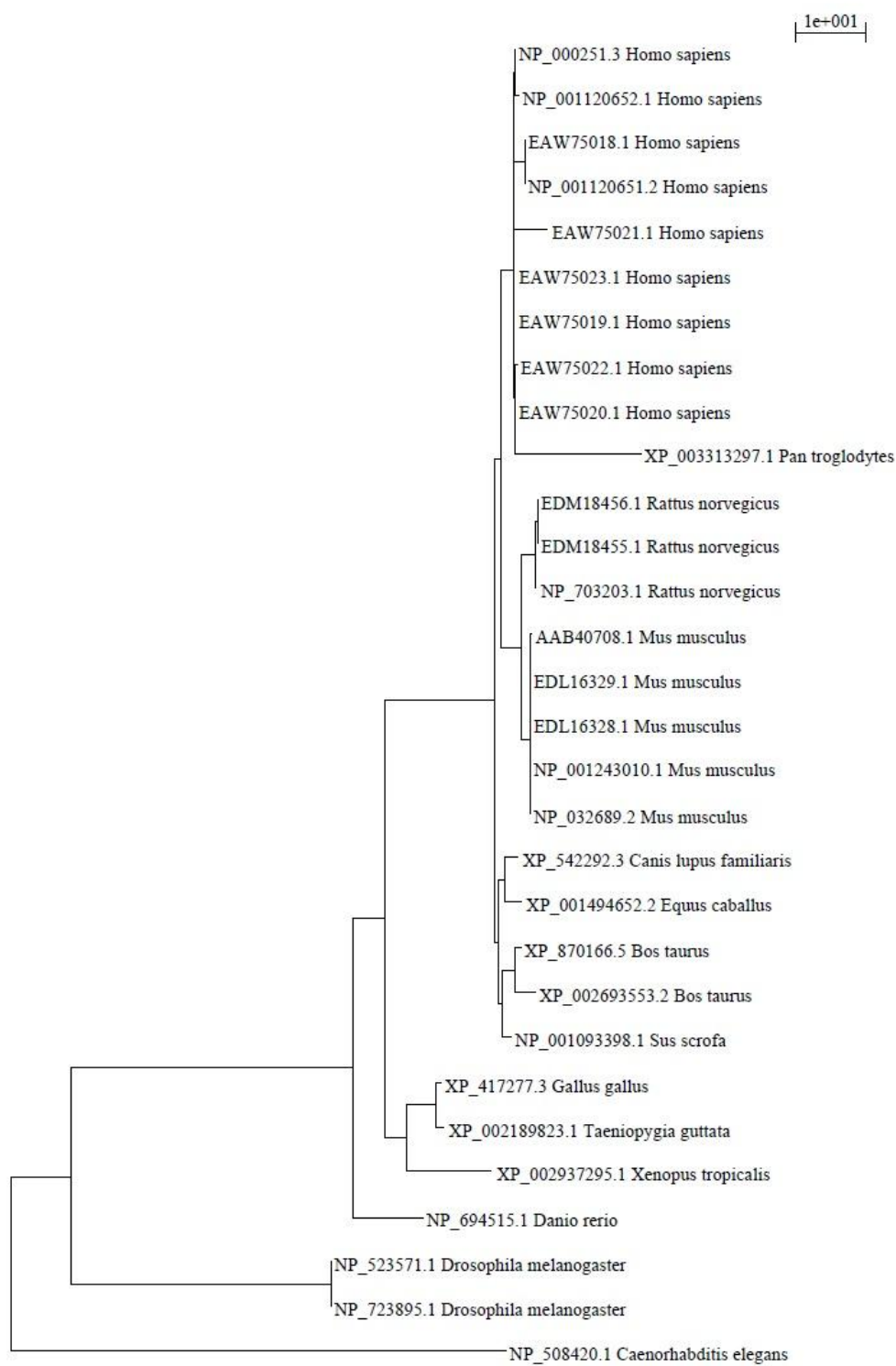
Figure 35. MYO7A Tree derived using ProML from aligned protein sequnces.
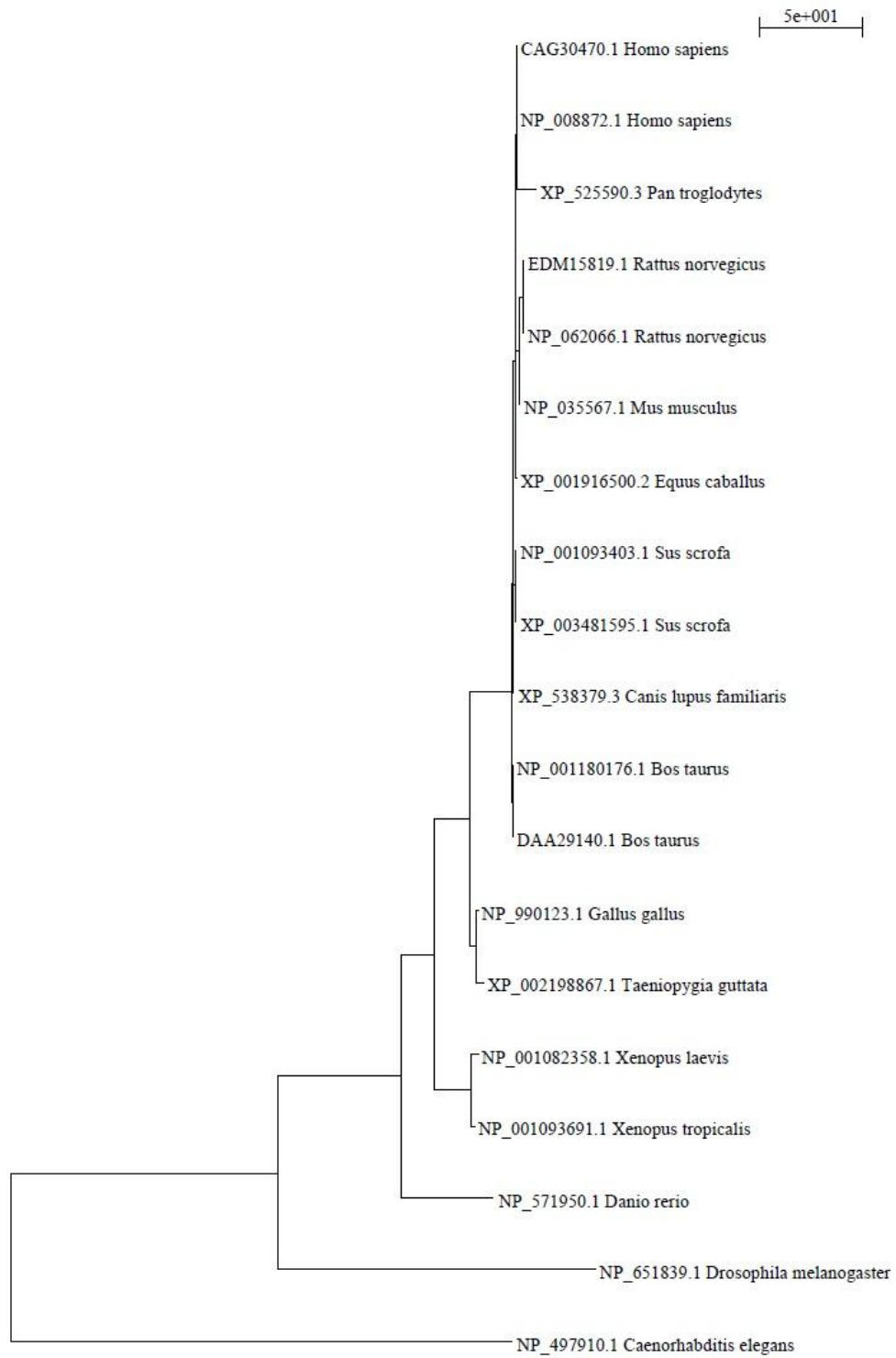
Figure 36. SOX10 Tree derived using ProML from aligned protein sequnces.

Auditory System Tree of Trees

The auditory TOT was derived in the same manner as the visual TOT by combining, recalculating and scaling the branch lengths among the eleven well-studied proteins involved in the development and function of the auditory system (See Table 3 for the list of auditory proteins). The inferred auditory TOT illustrates the coevolution of the interacting proteins as well as a part of the auditory interactome involved in the development and function of the vertebrate hearing system (Figure 37). It is important to remember that the inferred auditory TOT only represents a snapshot of all the PPI involved in the hearing system as there are many more proteins that participate in the development and function of the ear.

The auditory TOT also indicates that interacting proteins evolved in clusters highly correlated with the proteins' tissue specificity, expression level during development and their direct and indirect interaction. The three PPI clusters observed in the auditory TOT are: (MYO7A, PAX3, SPT6), (SIX1, SOX10, UGDH), and (CDH23, OTX1, EYA1). Overall these three PPI clusters display a developmental time line starting from early otic placode formation from which the ear develops to the mechano-signal transduction by the hair cells to the brain. However, these three clusters also suggest some unexpected PPI interactions that need to be further investigated experimentally.

In the auditory TOT both FGF3 and IRF6 fall outside of the three identified clusters showing that they are evolutionarily distinct However, according to the auditory TOT, the evolutionary histories of FGF3 and IRF6 are more similar to the (MYO7A,

PAX3, SPT6) and (SIX1, SOX10, UGDH) clusters than the (CDH23, OTX1, EYA1)

cluster (Figure 37). The distinct evolutionary histories of FGF3 and IRF6 can also be

observed by looking at the recalculated branch lengths in the TOT, which represent the

amount of divergence between two nodes. The recalculated branch lengths in the auditory

TOT for FGF3 and IRF6 are much longer, representing a greater amount of divergence

between FGF3 and IRF6 and the three clusters. As a result, FGF3 and IRF6 proteins do

not belonging to the aforementioned clusters but rather link the three PPI clusters

together.

The reconstructed auditory TOT has a scale bar of $2e + 001$, which indicates the

number of amino acid changes per site, suggesting that a minute amount of evolutionary

changes occurred between the proteins. Consequently, the auditory TOT is drawn on the

same tree scale as the previously discussed individual auditory phylogenies. These

comparable individual tree scales suggest that all of the proteins coevolved on the same

molecular scale and thus share similar evolutionary histories and constraints. In turn, this

allows us to infer that the PPI clusters observed in TOT most likely coevolved to

maintain their overall function and structural integrity of the auditory system.
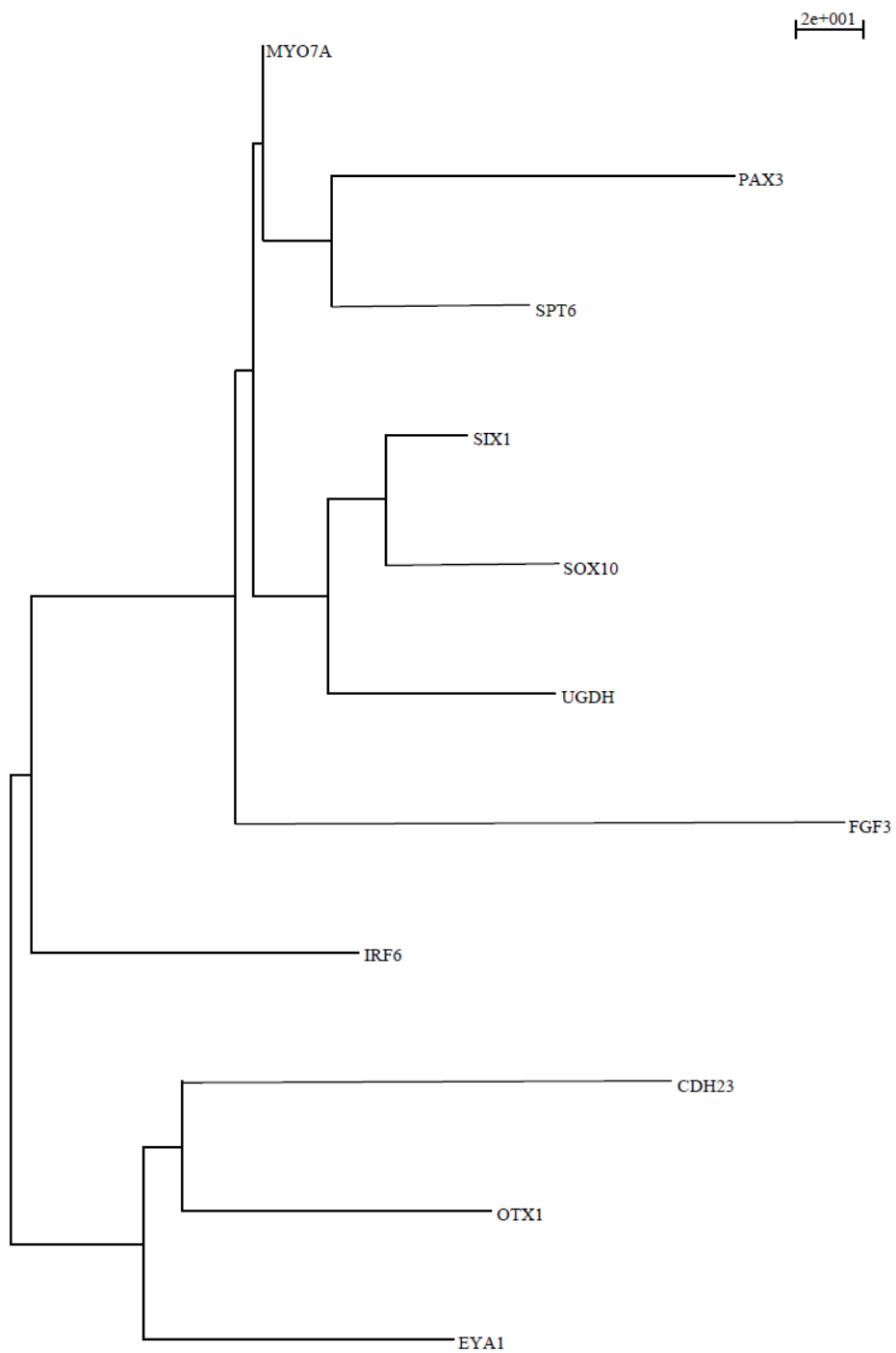
Figure 37. Auditory TOT illustrating the PPI of the auditory system.

Discussion of Auditory Tree of Trees

The vertebrate inner ear is composed of two parts: the auditory portion dedicated to hearing and the vestibular system dedicated to balance. In mammals, the cochlea is a coiled bony labyrinth that is lined with sensitive hair cells that move to the sound vibrations. Non-mammalian vertebrates also have an auditory organ similar to the cochlea; however, it is not coiled up, but still contains sensory hair cells for hearing. The ability to perceive sounds and maintain balance is dependent on the process of mechanotransduction, which is the conversion of mechanical stimulus that is evoked by sound waves and head movements into an electrical signal that can be processed by the central nervous system. The ears of all vertebrates use mechanosensory hair cells to convert mechanical energy to electrical signals compatible with the nervous system. The basic structure of hair cells is ubiquitous among the vertebrates and hair cells are also found in the lateral line of fishes and aquatic amphibians.

The mechanically sensitive organelle of the hair cell is the hair bundle, a highly elaborated structure of actin based stereocilia arranged in precise rows of increasing height (Vollrath *et al.,* 2007). Hair cells respond to deflections of their hair bundles by opening and closing transduction channels and respond best to stimuli directed toward the gradient of stereocilia height. Deflections that tilt the bundle toward the tallest stereocilia induce transduction channels to open, whereas deflections toward the shortest stereocilia close channels (Gillespie *et al.,* 2009). Therefore, hair cells of the inner ear execute the fundamental process by which mechanical stimulus originating from head movement or

acoustic waves is converted to neural signal and a release of neurotransmitter onto

afferent fibers of the eight nerves, which encodes and carries the auditory and vestibular

information to the brain (Vollrath *et al.,* 2007).

The auditory TOT indicates that the proteins involved in the development of the

auditory system form clusters and the clustering of the proteins is highly correlated with

the proteins' tissue specificity, expression level during development and their direct and

indirect interaction. For instance, some of the proteins are only found in specific areas

such as the forebrain, or the hair cell bundles, while others are expressed only during

certain periods of development. Conversely, some of the proteins require direct

interactions with each other to elicit certain functions, while for others indirect

interactions are enough to produce a desired effect. In order for the PPI to cluster

together, the interacting proteins had to undergo organized reciprocal evolutionary

changes to conserve the interplay among them otherwise not compensated changes could

lead to reduced or lost binding with other proteins or a gain of interaction with new

protein partners. The three clusters inferred by the auditory TOT include the following

proteins: (MYO7A, PAX3, SPT6), (SIX1, SOX10, UGDH), and (CDH23, OTIX1,

EYA1) and will be discussed in more detail below.

MYO7A, PAX3, and SPT6 cluster

The first cluster formed in auditory TOT includes the MYO7A, PAX3, and SPT6

proteins (Figure 38), which are essential for signal detection and transduction and proper functioning of the auditory system (Gillespie *et al.,* 2009; Zhang *et al.,* 2012;.Keegan *et al.,* 2002) and are further discussed below.
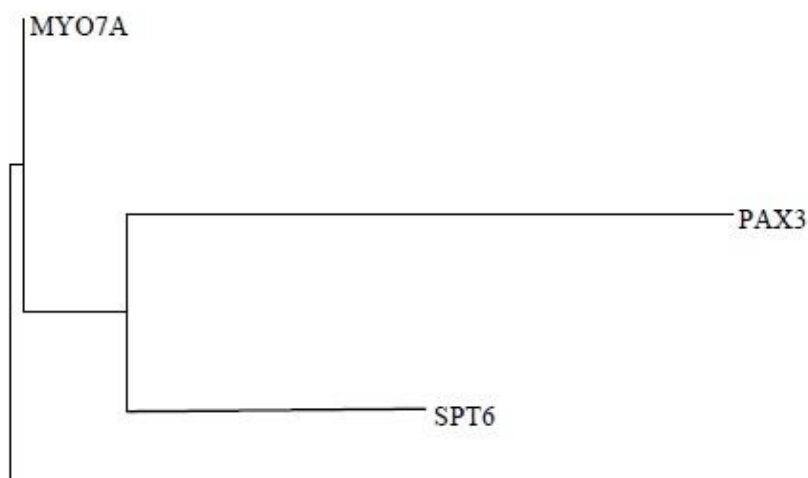
Figure 38. Auditory TOT cluster of the MYO7A, PAX3, and SPT6 proteins.

The first protein grouped in this cluster is MYO7A, which belongs to a class of unconventional myosins, characterized by a very short tail domain, that greatly differ from the conventional myosins required for processes such as muscle contraction. Unconventional myosins, such as MYO7A, power many forms of actin-based motility and organelle trafficking. MYO7A is a scaffolding protein that is closely tied to transduction. For instance, MYO7A is involved in harmonin transport, a protein important for the development and maintenance of stereocilia (Gillespie *et al.,* 2009). Consequently, MYO7A is expressed in mechanosensory hair cells and the actin based molecular motor MYO7A is critical for proper development of the hair bundle and hair cell signal mechanotransduction. MYO7A also contains the FERM domain (F for 4.1

protein, E for ezrin, R for radixin and M for moesin), which is often found in many

cytoskeletal-associated proteins that interact with various proteins at the interface

between the plasma membrane and the cytoskeleton (Wu *et al.,* 2011). Many disease-

causing mutations occur in the MyTH4-FERM domain of the myosin tail. For example,

truncation of the FERM domain affects myosin7a motor function and cargo transport.

Mutations in the FERM domain also destabilize the MYO7A transcript in the inner ear,

thereby causing defects in hair bundle development (Schwander *et al.,* 2009). In addition,

disruptions of the FERM domain affect interactions with melanosomes, disturbing their

transport in retinal pigment cells. Furthermore, Usher Syndrome 1B, which is

characterized in zebrafish, mice and humans by deaf-blindness is also caused by

mutations in the FERM domain of MYO7A (Schwander *et al.,* 2009).

The inferred auditory TOT groups PAX3 in the same PPI cluster as MYO7A.

This is surprising as one would not expect the two proteins to interact with each other as

they have very different roles in the auditory system. MYO7A is mainly expressed in the

mechanosensory hair cells, whereas PAX3 is a member of paired box family of

transcription factors. It could be that this is a very indirect interaction as there are

numerous proteins in between that are involved in the function and development of the

auditory system.

PAX3 is characterized by the presence of two DNA-binding domains, the paired

domain (PD) and the homeodomain (HD) (Figure 39). The N- and C- terminus of the PD

contain the classical helix-loop-helix (HLH) subdomains. The N terminal HLH

subdomain contains: a β-hairpin structure, a type II β-turn and an HLH motif which

contribute to DNA binding and facilitate interactions with other proteins (Zhang *et al.,*

2012).
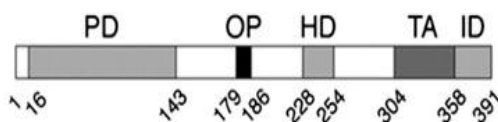
Figure 39. PAX3 functional domains PD,
DNA binding paired domain and HD.
[Reproduced from Underwood *et al*., 2007]

PAX3 is first expressed in the dorsal neural tube and then can be detected in the

developing brain, neural crest and their derivatives such as the melanocytes (Zhang *et al.,*

2012). Furthermore, PAX3 plays a role in the induction of melanoblasts, the melanocyte

precursor, by activating the expression of MITF, which is critical for survival,

proliferation and differentiation of melanocytes, melanin-producing cells that localize to

the eyes, ears and skin. Mutations in PAX3 lead to the Waardenburg syndrome (WS), an

auditory-pigmentary disorder resulting from abnormal proliferation, survival, migration,

differentiation of neural crest cell derived melanocytes.

The protein SPT6 is also in the same PPI cluster as MYO7A and PAX3. This is a

positive inference as both SPT6 and PAX3 are part of the transcription machinery. SPT6

is a transcription elongation factor. During transcriptional elongation, RNA polymerase II

(Pol II) and positive transcription elongation (P-TEFb) complexes catalyze the

transcription of DNA to synthesize precursors of mRNA (Keegan *et al.,* 2002). Several

classes of transcription elongation factors have been identified in prokaryotes and

eukaryotes. The SPT4, SPT5 and SPT6 are conserved factors and have been grouped

together because of shared developmental defects in pigmentation and ears (Kaplan *et al.,*

2000). SPT4 and SPT5 function as a complex, whereas SPT6 often functions along with SPT5 in active transcription without forming a complex. Transcription elongation factor SPT6 is one of the many classes of elongation factors that have been very well conserved throughout the animal kingdom. For example, SPT6 in the zebrafish is similar to the SPT6 of human, mouse, fly, nematode and yeast. Furthermore, SPT6 zebrafish mutants show several developmental defects including reduced pigmentation and problems in ear formation (Keegan *et al.,* 2002). SPT6 mutants also have disrupted transcriptional efficiency. Thus, genetic analysis indicates that zebrafish SPT6 is a conserved transcription elongation factor and plays an essential role during embryogenesis by controlling multiple aspects of differentiation through stimulation of gene expression (Keegan *et al.,* 2002).

Thus, the first PPI cluster inferred by the auditory TOT consists of MYO7A, PAX3, and SPT6. The proposition that these three proteins cluster together based on shared evolutionary history within a PPI network is inferred based on individually verifiable results. Consequently experimental studies show that MYO7A, PAX3, and SPT6 are essential for signal detection and transduction as well as proper functioning of the auditory system (Gillespie *et al.,* 2009; Zhang *et al.,* 2012; Keegan *et al.,* 2002). In depth examination of MYO7A, PAX3, and SPT6 revealed that the three proteins involve neural crest derived inner ear melanocytes, which reside in the stria vascularis of the cochlear duct (Dutton *et al.,* 2009). The neural crest derived melanocytes are essential for both the maintenance of endolymph fluid contained in the membranous labyrinth of the

inner ear and generation of the endocochlear potential, which drives current through hair cells when they move in response to a sound stimulus (Dutton *et al.,* 2009). Reduction of neural crest derived melanocytes in the stria vascularis of the cochlea, leads to a reduction or collapse of endolymph volume, and a loss of the endocochlear potential and subsequent hair cell degeneration (Dutton *et al.,* 2009). Thus, MYO7A transport of melanosomes depends on melanocytes, which in turn are essential for the proper mechno-signal transduction. However, melanocytes are depended on PAX3 induction of melanoblasts, the melanocyte precursors. In turn, PAX3 mutations lead to varying combinations of sensorineural hearing loss and abnormal pigmentation of the hair, skin and inner ear. Likewise, SPT6 mutants also show several developmental defects including reduced pigmentation in the inner ear.

<u>SIX1, SOX10, and UGDH cluster</u>

The second PPI cluster in the auditory TOT consists of the SIX1, SOX10, and UGDH proteins (Figure 40), which play an important role in formation of the inner ear structures. (Zheng *et al.,* 2003; Dutton *et al.,* 2009; Busch-Nentwitch *et al.,* 2003) and are described below.
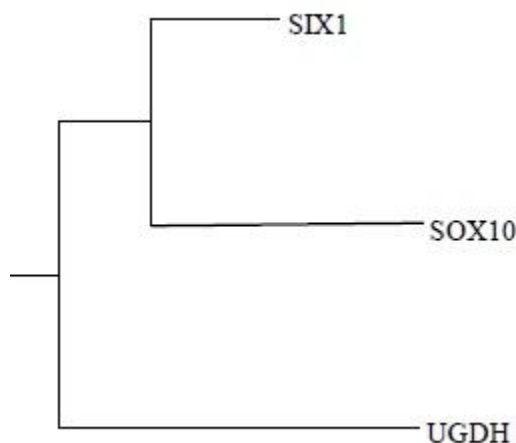
Figure 40. Auditory TOT cluster including the SIX1, SOX10, and UGDH proteins.

Inner ear development begins with the induction of the otic placode, a thickened area of surface ectoderm on each side of the hindbrain. Once the otic placode forms the otic vesicle, differentiation, proliferation and apoptosis take place to ensure the formation of highly organized structures of the inner ear (Zheng *et al.,* 2003). During inner ear development, SIX1 expression is first detected in the ventral region of the otic cup and later is restricted to the middle and ventral otic vesicle, within which the vestibular and the auditory epithelia form (Zheng *et al.,* 2003). Consequently, SIX1 expression is turned on during the invagination of the otic placode where it begins to control inner ear morphogenesis by regulating the programmed cell death and proliferative growth of the otic epithelium. Thus, SIX1 is required for regional specification and patterning of the otic vesicle and it is probably an early regulator for the specification of all sensory epithelia of the inner ear (Zheng *et al.,* 2003). Inactivation of the SIX1 leads to malformation of the auditory system involving the outer, middle and inner ear. Moreover, SIX1 is also expressed in the 8th vestibuloacoustic ganglion, which develops from the

otic placode and is responsible for transmitting sound and equilibrium information from the inner ear to the brain.

In the second cluster of the auditory TOT, SIX1 is grouped with SOX10, a universally occurring transcription factor SRY (sex determining region Y)-box 10. Both SIX1 and SOX10 are transcription factors expressed mainly in the nucleus. SOX10 is also required for the patterning of otic epithelium and normal development of the otic vesicle. In addition, SOX10 belongs to the SOX group E, which includes SOX8, SOX9A, SOX9B proteins, which together are important in otic induction pathways and subsequent otic patterning (Dutton *et al.,* 2009). In turn, SOX10 participates in a feedback loop with SOX9 to establish otic specific patterns of gene expression. As a result, SOX10 plays a direct role in the maintenance of otic epithelium, patterning of the otic vesicle and thus inner ear development. Mutations in SOX10 are attributed to loss or reduction of neural crest and abnormalities in neural crest development cause neurocristopathies, which refer to a diverse class of pathologies that arise from defects in the development of tissues containing cells commonly derived from the embryonic neural crest cell lineage and include conditions like the Waardenburg-syndrome (Dutton *et al.,* 2009).

One would expect SOX10 and PAX3 to belong to the same cluster as both proteins play an important role in development of the neural crest, which gives rise to neurons melanoblasts and glia of the peripheral nervous system. Furthermore, experimental results confirm that SOX10 and PAX3 direcetly interact (Underwood *et al.,*

2007). However, the auditory TOT shows the two proteins on different clusters,

suggesting that there are numerous other proteins in between that play a role in the

development and function of the auditory system.

The auditory TOT also groups UGDH in the same cluster as SIX1 and SOX10.

UGDH (uridine 5' diphosphate – (UDP)-glucose dehydrogenase) is an important protein

required for proper semicircular canal formation and function. UGDH is an enzyme that

is necessary for the production of proteoglycans including hyaluronic acid (HA). In turn,

HA is an integral part of the extracellular matrix, which plays a role in cell migration,

differentiation and morphogenesis of the ear and jaw (Busch-Nentwitch *et al.,* 2003).

This is another result supported by the literature as neural crest cells migrate extensively

through prescribed regions of the embryos, where they differentiate into most of the

peripheral nervous system as well as the facial skeleton and pigment cells (Bronner-

Fraser, 1994).

Moreover, HA is essential for differentiating cartilage as well as the outgrowth of

the epithelial projections forming the semicircular canals (Figure 41). Secretion of HA

within the projections provides the driving force for growth in particular direction.
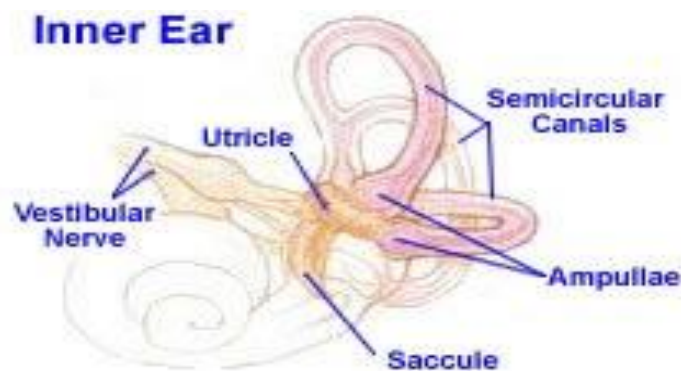
Figure 41. Projections of the Semicircular Canals. [Reproduced from Evolve Media, 2012]

Reduction of HA levels results in an uncoordinated outgrowth of the epithelial

projections and disorganization of the epithelium (Busch-Nentwitch *et al.,* 2003). In

addition to the uncoordinated outgrowth of the canal columns, reduction in HA levels

leads to impaired facial cartilage differentiation. It should also be mentioned that

reduction of HA biosynthesis affects the extracellular matrix and its disruption causes

mechanical stress on hair cells, which leads to premature aging of hair cells and hearing

loss that first compromises the high frequency range. Such symptoms are indicative of

non-syndromic hearing loss.

Thus, the predicted second cluster of SIX1, SOX10, and UGDH in the auditory

TOT consists of proteins whose function within the auditory network is based on

validated results (Zheng *et al.,* 2003; Dutton *et al.,* 2009; Busch-Nentwitch *et al.,* 2003).

Consequently the second cluster focuses on proteins necessary for the development of

auditory precursors required for the formation of the inner ear. For example, in this

cluster, SIX1 specifies the region for the otic placode induction and SOX10 regulates the

formation of the otic vesicle from the otic placode. Once the otic vesicle is formed,

SOX10 plays a role in the otic epithelium patterning and formation of highly organized

structures of the inner ear. UGDH also plays a role in cell migration, differentiation and

morphogenesis of the ear and jaw by regulating the production of HA in the extracellular

matrix (Busch-Nentwitch *et al.,* 2003). Thus, the SIX1, SOX10, and UGDH cluster

begins the inner ear morphogenesis.

The auditory TOT also shows that FGF3 and IRF6 fall outside of the three

identified clusters, however, the evolutionary histories of FGF3 and IRF6 are more similar to the (MYO7A, PAX3, SPT6) and (SIX1, SOX10, UGDH) clusters than the (CDH23, OTX1, EYA1) cluster (Figure 42).
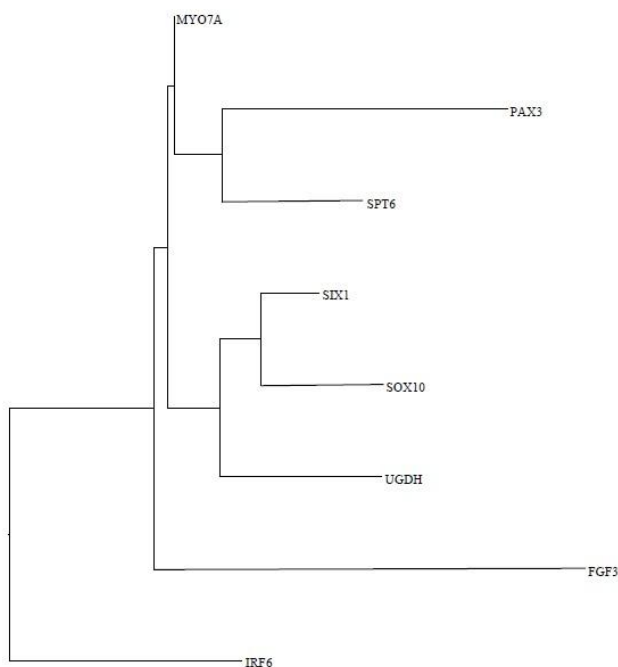
Figure 42. The auditory TOT shows that FGF3 and IRF6 do not cluster with (MYO7A, PAX3, SPT6) and (SIX1, SOX10, UGDH).

As previously mentioned, the otic placode induction is also largely dependent upon fibroblast growth factors like FGF3, FGF8 and FGF10. Growth factors typically act as signaling molecules between cells and stimulate cellular growth, differentiation and maturation. In addition, FGFs are key players in the processes of proliferation and differentiation of wide variety of cells and tissues. FGF3 plays a critical role in the induction of otic tissues in all vertebrates as its signaling is required for normal otic placode formation, maintenance and inner ear patterning. FGF3 mutants lack ears and vestibular structures. Similar to SIX1, FGF3 also plays a role in the formation of the 8th

vestibuloacoustic ganglion, and mutations in FGF3 lead to severe hypomorphic development of the 8th ganglion.

In addition, the previously described melanocytes in the (MYO7A, PAX3, SPT6) cluster also depend on FGF3 to stimulate and regulate their proliferation and differentiation (Chung *et al.,* 2011). In short, melanocytes stretch out to connect with neighboring keratinocytes (Figure 43). The main function of keratinocytes is to produce keratin protein and form an epidermal barrier against environmental damage. The relationship between melanocytes and keratinocytes is bidirectional. Melanocytes drive the vertebrate pigmentation by transferring from the tips of their dendrites melanin containing vesicles called melanosomes. As a result, melanocytes protect keratinocytes from ultraviolet damage. Keratinocytes mediate melanocyte function via cell-cell adhesion, cell-matrix adhesion and paracrine signaling of nearby targets. In turn, melanocytes depend on specific growth factors secreted by keratinocytes to be able to survive, proliferate and migrate. Consequently, melanocytes maintain cell-cell adhesion with keratinocytes via expressing proteins such as E-cadherin and connexins. Consecutively, keratinocytes secrete melanocyte-stimulating hormones and factors such as basic fibroblast growth factors which stimulate and regulate the proliferation and differentiation of melanocytes (Chung *et al.,* 2011).
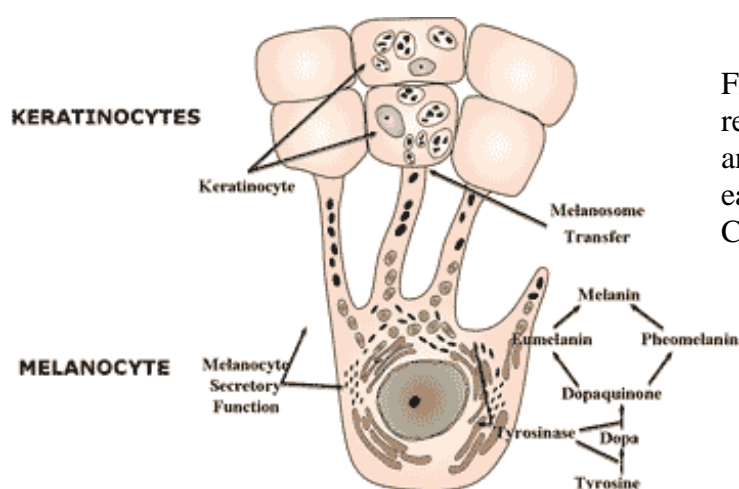
Figure 43. Schematic representation of keratynocytes and melanocytes of the inner ear. [Reproduced from Copper Cup Images, 2012]

IRF6 also falls outside of the three identified clusters but according to the auditory TOT (Figure 42). This shared evolutionary history may be the result of IRF6's involvement in the Notch signaling pathway. IRF6 belongs to a family of interferon regulatory factors that regulate transcription of interferons, which are cell defense proteins. IRFs also engage the JAK-STAT signaling pathway, which uses chemical signal receptors to transmit information from signals outside the cell to activate DNA transcription in the cell. In addition, IRF6 also plays a critical role in keratinocyte development (Restivo *et al.,* 2011). For example, self- renewing epithelia such as keratinocytes depend on finding the right balance between cell growth and differentiation. Many signals and factors including direct cell-cell communication, cell adhesion and contact influence the right balance between keratinocytes cell growth and differentiation. Notch signaling is one of the pathways coordinating keratinocyte differentiation, growth and arrest though modulation of IRF6 expression (Restivo *et al.,* 2011). As a result, self- renewing epithelia such as keratinocytes can take advantage of IRFs and the JAK-STAT signaling pathway to monitor and regulate keratinocytes.

IRF6 is expressed in the pharyngeal arches, olfactory and otic placodes, as well as in the epithelial cells of endoderm derived tissues. The function of IRF6 is related to the formation of connective tissues such as the palate that separates the oral cavity from the nasal cavity. Accordingly, IRF6 mutations are involved in common forms of cleft lip and cleft palate. Mutations of the IRF6 lead to Van Der Woude syndrome (VDWS), which is characterized by congenital facial malformations such as cleft lip and palate and hypodontia. However, various forms of VDWS exhibit other abnormalities in addition to the cleft lip/palate such as limb anomalies, popliteal webs, accessory nipples, congenital heart defects, and Hirschsprung disease (Wang *et al.,* 2011). In addition, some patients with VDWS also report sensorineural hearing loss. Interestingly, mutations in other proteins in addition to IRF6 including FGFs, SOXE group proteins such as SOX10, and EYA1 also play a role in oro -facial malformations. Thus, from these examples it follws that the auditory TOT depicts IRF6 as sharing evolutionary history with (MYO7A, PAX3, SPT6) and (SIX1, SOX10, UGDH) clusters.

CDH23, OTX1, and EYA1 cluster

The third cluster in the auditory TOT consists of the CDH23, OTX1, and EYA1 proteins, which are essential for further development of the inner ear precursors (Figure 44). This cluster sequentially follows the formation of the otic vesicle.
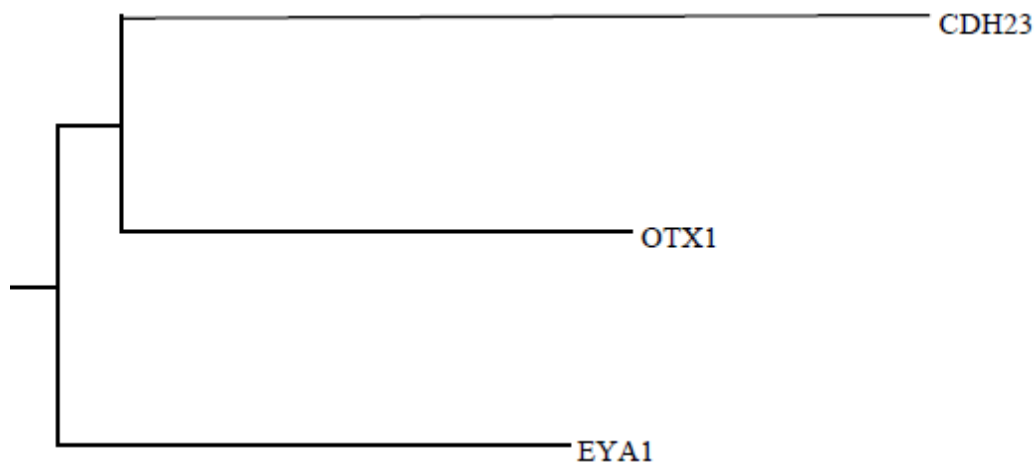
Figure 44. Auditory TOT cluster formed by (CDH23, OTX1, EYA1) proteins.

The first protein in this developmental cluster is Cadherin 23 (CDH23), which

belongs to the cadherin superfamily that consists of about 100 members that play a

variety of roles in tissue development, maintenance and function. The defining feature of

the cadherin superfamily is extracellular cadherin (EC) domain that binds Ca2+. The EC

domain occurs in varying repetitions in all cadherins (Muller, 2008). Cadherins are

named for "calcium-dependent adhesion" since the stability of cadherins depends on the

presence of Ca2+, whose binding with the (EC) portion of the polypeptide chain is a

prerequisite for cadherin cell-cell adhesion. Tip link cadherins differ from classical

cadherins in that instead of the 5 extracellular cadherin (EC) repeats in classical

cadherins, the extracellular domain of tip link cadherins contain 27 EC repeats (Gillespie

et al., 2009). CDH23 localizes to hair bundles and it is an essential component of the tip

links as it acts as an intracellular adhesion molecule connecting stereocilia to each other

and to kinocilia (Figure 45). Consequently, CDH23 forms several of the extracellular

filaments in vertebrates and one of these filaments is the tip link that has been proposed

to gate the mechanotransduction channel in hair cells (Stollner *et al.,* 2004). Therefore,

CDH23 not only functions in maintaining bundle integrity but also has a direct role in

mechanotransduction.
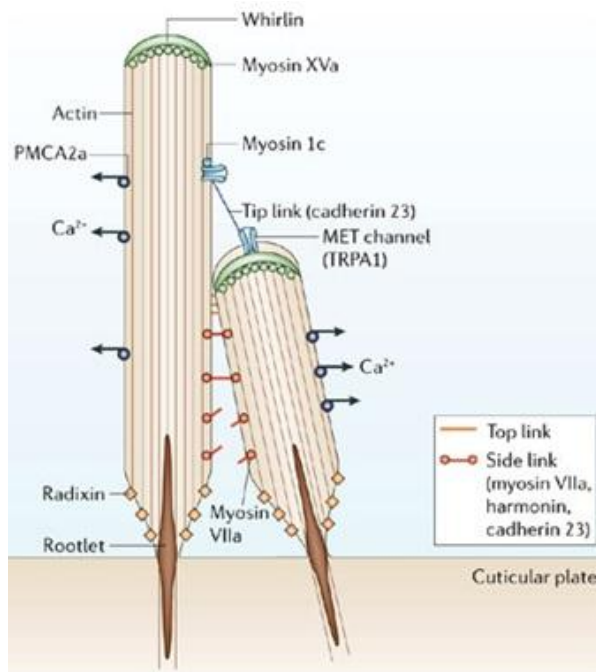
Figure 45. Schema of tip link CDH23. [Reproduced from Fettiplace & Hackney, 2006]

Mutations in CDH23 lead to deafness and vestibular defects in zebrafish, mice

and humans. Zebrafish CDH23 encodes a 3,366 amino acid protein containing 27 EC

repeats followed by a single transmembrane domain and a carboxy-terminal tail. The

zebrafish CDH23 is 68% identical and shares 81% similarity with the human and mouse

CDH23 (Muller, 2008). Zebrafish CDH23 is concentrated near the tips of hair bundles

and mutations in CDH23 lead to tip link loss and thus affect mechanotransduction.

Zebrafish with mutations in CDH23 are referred as sputnik mutants and show various

degrees of splaying in their sterociliary bundles, which leads to reduced

mechanotransduction (Stollner *et al.,* 2004). In addition, zebrafish contain hair cells in their lateral line organ which is used to detect movement and vibration in the surrounding water. Consequently, zebrafish sputnik mutants have defects both in hearing and balance. Hence, zebrafish studies conclude that CDH23 is an essential tip link component required for hair cell mechanotransduction (Stollner *et al.,* 2004).

In the inferred auditory TOT, CDH23 is grouped together with OTX1, which plays a role in the inner ear as well as forebrain development. OTX1, a transcription factor is the homolog of the *Drosophila* orthodenticle homeobox 1 which shows limited amino acid sequence divergence among vertebrate and invertebrate species. OTX1 is expressed during the morphogenesis of the murine brain and CNS and is required for the normal development of the inner ear (Simeone *et al.,* 1992). Following formation of the otic vesicle, OTX1 is expressed very early in the posteroventrolateral and ventral apex of the otic epithelium. The ventral cells of the otic epithelium are believed to be fated to give rise to the cochlear duct and organ of Corti. OTX1 mutants not only have cochlear and saccular defects that are consistent with its ventral expression domain but also lack both the lateral semicircular duct and the lateral sensory patches of the ear (Chatterjee *et al.,* 2010). In addition, OTX1 interacts with other regulatory proteins to develop the caudal forebrain and structures of the inner ear.

The final protein in this third cluster of the auditory TOT is EYA1. The eye absent (EYA1) is a transcription co-activator that is also expressed early in the otic epithelium. Mice knockout studies of EYA1 show arrest of the inner ear development at

the otic vesicle stage. As a result, EYA1 is crucial for normal growth of the otic vesicle. Furthermore, EYA1 is also important for the development of the cochlea and the posterior ampulla (Zheng *et al.,* 2003). Therefore, the third PPI cluster formed in auditory TOT consists of proteins that further control the differentiation of the inner ear.

The three PPI clusters inferred by the auditory TOT illustrate a developmental time line starting from early otic placode formation from which the ear develops to the mechano-signal transduction by the hair cells to the brain. The three PPI clusters illustrate the cascade of PPI necessary for the development and function of the auditory system. There are some proteins that form unexpected groups that have different expression times and locations. Further investigation of such anamolies can aid in identifying the cause. Thus, the auditory TOT not only predicts PPI based on coevolution of protein pairs but also provides a developmental roadmap from simple to more complex structures and functions.

CHAPTER FOUR

CONCLUSIONS

The developed TOT program offers potential advantages over existing computational methods for inferring PPI. Depending on a user's computational power, TOT can theoretically accommodate an infinite number of genes or proteins under study. In addition, once new sequences are known and available they can be easily added and integrated into the analysis. TOT also gives flexibility in using preferred tree derivation methods which then can be synthesized and used to create universal Newick files from which information can be derived to build the TOT. The multifaceted power of TOT is that it can take not only sequence information but also morphological input in future development. In the case of proteins, such morphological input could comprise of protein structure information. This information could then also be used to aid in predicting interacting proteins.

The results for both visual and auditory TOT illustrate that this method can be successfully applied to infer PPI based on coevolution of protein partners. The derived TOT for the G-proteins of the visual system suggests that the visual TOT is not only able to infer subunit complexes but also a full cascade of PPI responsible for visual phototransduction as in the (OPN1SW, PDE6A, RHO, GNGT1 and SIX6, GNB1) cluster. In addition, the auditory TOT results illustrate the PPI as well as predict their time line in development. For instance, the auditory TOT suggests that first cluster

consisting of (MYO7A, PAX3 and SPT6) proteins is essential for signal detection and transduction. The second PPI cluster in the auditory TOT consists of (SIX1, SOX10, and UGDH) proteins and plays an important role in the formation of the inner ear structures. The third cluster formed in auditory TOT consists of (CDH23, OTX1 and EYA1) proteins and sequentially follows the formation of the otic vesicle and is essential for further development of the inner ear precursors of the second cluster.

The developed TOT program also predicted some unexpected protein pairs between functionally different proteins that are expressed at different times and locations during development. Nevertheless, the TOT program inferred some interesting results that need to be further analyzed using experimental methods. For instance, the RIBEYE and PAX6 interaction is unexpected; however, RIBEYE is composed of two functional domains, including a B domain, which could possibly regulate the transcription of PAX6. Similarly, the ATOH7 and BSN putative interaction needs to be further analyzed as the two proteins have parallel functions and are both expressed in the nucleus.

A concern of any bioinformatics solution for detecting PPI is the rate of false positives and false negative results.Within the software solution presented here, the occurrence of false positives and false negatives can be addressed by refining the analysis to proteins that are expressed in the same tissues and times during development. In addition, to increase the signal to noise ratio, instead of using full length protein sequences, analyses could be confined to structural domains and active sites to amplify the signal.

Moreover, the proteins included in the validation of the TOT are specific to a particular system; however, it seems that functional systems may be developmentally connected by sharing activation of transcription factors as well as other proteins. Since the visual and auditory systems develop in a great proximity; the inductive signals from the neighboring tissues of the developing ear and eye intermix. As a result, some proteins have dual functions in two different developing systems. This can cause difficulty in isolating proteins to their specific system.

Although TOT illustrates only a snapshot of the PPI in a particular system, the number of proteins used to derive the TOT was sufficient to successfully test the program, validate results and identified areas for future improvement. Adding more protein sequences to the analysis is dependent on Protein GenBank availability as well as the ability of the computer memory to execute all the runs. At this point, this program can be a very useful tool for researchers to identify possible PPI based on the coevolution model. TOT is a novel approach that differs from current online PPI databases that are manually curated. The TOT approach not only allows researchers to identify proteins that might have evolved with conserved roles in the same functional or developmental network but also allows the researchers to use specific proteins of interest.Thus, the TOT approach should be used as preliminary tool for predicting PPI, which then can be tested using experimental methods.

APPENDIX A:

VISUAL PROTEIN ACCESSION NUMBERS,

IN ALPHABETICAL ORDER

ATOH7

Accession	Description
NP_571707.1 [Danio rerio]. 134 aa chromosome 13	protein atonal homolog 7
CAD52125.1 [Danio rerio]. 134 aa chromosome 13	atonal homolog 7
NP_001079289.1 [Xenopus laevis]. 138 aa	protein atonal homolog 7-A
XP_002936902.1 [Xenopus (Silurana) tropicalis] 139 aa chromosome unknown	PREDICTED: protein atonal homolog 7-B-like
NP_660161.1 [Homo sapiens]. 152 aa chromosome 10	protein atonal homolog 7
AAL11911.1 same as AF418922.1 [Homo sapiens]. 152 aa chromosome 10	ATOH7
XP_521492.2 [Pan troglodytes]. 152 aa chromosome 10	PREDICTED: protein atonal homolog 7
ABM85298.1 [Drosophila melanogaster]. 152 aa	atonal homolog 7 (Drosophila), partial [synthetic construct].
ABM82115.1 [Drosophila melanogaster]. 152 aa	atonal homolog 7 (Drosophila) [synthetic construct].
XP_546132.2 [Canis lupus familiaris]. 152 aa chromosome 4	PREDICTED: protein atonal homolog
XP_590240.2 [Bos taurus]. 152 aa chromosome 28	PREDICTED: protein atonal homolog 7
XP_002698904.1 [Bos taurus]. 152 aa chromosome 28	PREDICTED: protein atonal homolog
DAA14284.1 [Bos taurus]. 152 aa chromosome 28	atonal homolog 7-like
NP_058560.1 [Mus musculus]. 149 aa chromosome 10	protein atonal homolog
AAL11912.1 [Mus musculus]. 149 aa chromosome 10	ATOH7
NP_001163953.1 [Rattus norvegicus] 149 aa chromosome 20	atonal homolog 7
NP_989999.1 [Gallus gallus]. 151 aa chromosome 6	protein atonal homolog 7
NP_508410.2 [Caenorhabditis elegans]. 142 aa chromosome 10	abnormal cell Lineage family member (lin-32)

| NP_508725.1 [Caenorhabditis elegans]. 147 aa chromosome 10 | Helix Loop Helix family member (hlh-13) |

BASSOON

Accession	Description
CAA76598.1 [Mus musculus]. 3942 aa chromosome 9F	Bassoon
EDL21270.1 [Mus musculus]. 3942 aa chromosome 9	bassoon, partial
NP_031593.2 [Mus musculus]. 3942 aa chromosome 9	protein bassoon
NP_062019.2 [Rattus norvegicus]. 3933 aa chromosome 8	protein bassoon
NP_003449.2 [Homo sapiens] 3926 aa chromosome 3	protein bassoon
CAA77176.1 [Homo sapiens] 3851 aa chromosome 3	Bassoon protein, partial
EAW64993.1 [Homo sapiens]. 3926 aa chromosome 3	bassoon (presynaptic cytomatrix protein), isoform CRA_a
EAW64992.1 [Homo sapiens]. 3926 aa chromosome 3	bassoon (presynaptic cytomatrix protein), isoform CRA_a
XP_002697122.1 [Bos taurus]. 3529 aa chromosome 22	PREDICTED: protein bassoon
XP_601010.4 [Bos taurus]. 3532 aa chromosome 22	PREDICTED: protein bassoon
XP_541885.2 [Canis lupus familiaris]. 3921 aa chromosome 20	PREDICTED: protein bassoon
XP_003361436.2 [Sus scrofa]. 3495 aa chromosome unknown	PREDICTED: protein bassoon, partial
XP_516463.3 [Pan troglodytes]. 3947 aa chromosome 3	PREDICTED: LOW QUALITY PROTEIN: protein bassoon
XP_001494301.2 [Equus caballus]. 4007 aa chromosome 16	PREDICTED: protein bassoon
XP_001378098.2 [Monodelphis domestica]. 3960 aa chromosome 6	PREDICTED: protein bassoon
XP_003642021.1 [Gallus gallus]. 3771 aa chromosome 12	PREDICTED: protein bassoon-like
XP_002936530.1 [Xenopus (Silurana) tropicalis].	PREDICTED: protein bassoon-like

4498 aa chromosome unknown	
XP_001920788.3 [Danio rerio]. 4097 aa chromosome 8	PREDICTED: protein bassoon-like
NP_001096896.1 [Drosophila melanogaster]. 4979 aa chromosome 10	CG34417, isoform H
NP_001188549.1 [Drosophila melanogaster]. 4862 aa chromosome 10	CG34417, isoform L
CCD63778.1 [Caenorhabditis elegans]. 1231 aa	Protein F45E4.3, isoform a
CCD63779.1 [Caenorhabditis elegans]. 836 aa	Protein F45E4.3, isoform b

GNAT1 – Alpha subunit

Accession	Description
EDL21220.1 [Mus musculus]. 350 aa chromosome 9	guanine nucleotide binding protein, alpha transducing 1
NP_032166.1 [Mus musculus]. 350 aa chromosome 9	guanine nucleotide-binding protein G(t) subunit alpha-1
NP_001102250.2 [Rattus norvegicus]. 350 aa chromosome 8	rod-type transducin alpha subunit
XP_001167971.1 [Pan troglodytes]. 350 aa chromosome 3	PREDICTED: guanine nucleotide-binding protein G(t) subunit alpha-1 isoform 1
XP_003309846.1 [Pan troglodytes]. 350 aa chromosome 3	PREDICTED: guanine nucleotide-binding protein G(t) subunit alpha-1 isoform 2
AAB54048.1 [Homo sapiens]. 350 aa chromosome 3	Rod transducin (alpha-1 subunit)
NP_000163.2 [Homo sapiens]. 350 aa chromosome 3	guanine nucleotide-binding protein G(t) subunit alpha-1
NP_653082.1 [Homo sapiens]. 350 aa chromosome 3	guanine nucleotide-binding protein G(t) subunit alpha-1
EAW65047.1 [Homo sapiens]. 350 aa chromosome 3	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1, isoform CRA_a
EAW65048.1 [Homo sapiens]. 350 aa chromosome 3	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1, isoform CRA_a
NP_001003068.1 [Canis lupus familiaris]. 350 aa chromosome 20	guanine nucleotide-binding protein G(t) subunit alpha-1

XP_001496531.1 [Equus caballus]. 350 aa chromosome 16	PREDICTED: guanine nucleotide-binding protein G(t) subunit alpha-1-like
NP_851365.1 [Bos taurus] 350 aa chromosome 22	guanine nucleotide-binding protein G(t) subunit alpha-1
DAA16877.1 [Bos taurus] 350 aa chromosome 22	guanine nucleotide-binding protein G(t) subunit alpha-1
NP_990022.1 [Gallus gallus]. 350 aa chromosome 12	guanine nucleotide-binding protein G(t) subunit alpha-1
NP_571943.1 [Danio rerio]. 350 aa chromosome 6	guanine nucleotide-binding protein G(t) subunit alpha-1
AAL05600.1 [Danio rerio]. 350 aa chromosome LG6	rod transducin alpha subunit
NP_001084030.1 [Xenopus laevis]. 350 aa	guanine nucleotide-binding protein G(t) subunit alpha
NP_001096278.1 [Xenopus (Silurana) tropicalis]. 350 aa	guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1
XP_001928011.1 [Sus scrofa]. 350 aa chromosome 13	PREDICTED: guanine nucleotide-binding protein G(t) subunit alpha-1-like
XP_001368199.1 [Monodelphis domestica]. 350 aa chromosome 6	PREDICTED: guanine nucleotide-binding protein G(t) subunit alpha-1-like
NP_477502.1 [Drosophila melanogaster]. 355 aa chromosome 3L	G protein alphai subunit 65A
NP_492108.1 [Caenorhabditis elegans]. 354 aa chromosome 1	G protein,O, Alpha subunit family member (goa-1)

GNB1 – Beta subunit

Accession	Description
NP_032168.1 [Mus musculus]. 340 aa chromosome 4	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_001153488.1 [Mus musculus]. 340 aa chromosome 4	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
XP_001503482.1 [Equus caballus]. 340 aa chromosome 2	PREDICTED: guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_001003236.1 [Canis lupus familiaris] 340aa chromosome 5	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
XP_001363370.1 [Monodelphis domestica]. 340 aa chromosome 2	PREDICTED: guanine nucleotide-binding protein G(I)/G(S)/G(T)subunit beta-1-like

NP_786971.2 [Bos taurus]. 340 aa chromosome 2	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_997774.1 [Danio rerio]. 340 aa chromosome 8	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_001012853.2 [Gallus gallus] 340 aa chromosome 21	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_112249.2 [Rattus norvegicus]. 340 aa chromosome 5	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_001084140.1 [Xenopus laevis]. 340 aa	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_001006835.1 [Xenopus (Silurana) 340 aa	guanine nucleotide binding protein (G protein), beta polypeptide 1
NP_002065.1 [Homo sapiens]. 340 aa chromosome 1	guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1
NP_525090.1 [Drosophila melanogaster]. 340 aa	G protein beta-subunit 13F, isoform A
NP_496508.1 [Caenorhabditis elegans]. 340 aa	G Protein, Beta subunit family member (gpb-1)

GNGT1 – Gamma subunit

Accession	Description
NP_068774.1 [Homo sapiens]. 74 aa chromosome 7	guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-T1
XP_001168330.1 [Pan troglodytes] 74 aa chromosome 7	PREDICTED: guanine nucleotide-binding protein G(T) subunit gamma-T1 isoform 2
NP_034444.1 [Mus musculus]. 74 aa chromosome 7	guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-T1 precursor
NP_776752.1 [Bos taurus]. 74 aa chromosome 4	guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-T1 precursor
XP_003130194.1 [Sus scrofa]. 74 aa chromosome 9	PREDICTED: guanine nucleotide-binding protein G(T) subunit gamma-T1-like
NP_001129249.1 [Rattus norvegicus]. 74 aa	guanine nucleotide-binding protein G(T) subunit gamma-T1
NP_001086269.1 [Xenopus laevis]. 73 aa	guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 1

NP_956261.1 [Danio rerio]. 73 aa chromosome 4	PREDICTED: guanine nucleotide-binding protein G(I)/G(S)/G(O)subunit gamma-T1
NP_001003225.1 [Canis lupus familiaris]. 74 aa chromosome 14	guanine nucleotide-binding protein G(T) subunit gamma-T1 precursor
XP_001492841.1 [Equus caballus]. 74 aa chromosome 4	PREDICTED: guanine nucleotide-binding protein G(T) subunit gamma-T1-like
AAF52759.2 [Drosophila melanogaster]. 72 aa	G protein gamma30A, isoform A
NP_491935.1 [Caenorhabditis elegans]. 62 aa	G Protein, Gamma subunit family member (gpc-2)

OPN1SW

Accession	Description
DAA30407.1 [Bos taurus]. 349 aa protein chromosome 4	blue-sensitive opsin
NP_776992.1 [Bos taurus]. 349 aa protein chromosome 4	short-wave-sensitive opsin 1
NP_571394.1 [Danio rerio]. 336 aa chromosome 4	opsin-1, short-wave-sensitive 1
NP_990769.1 [Gallus gallus] 347 aa	violet-sensitive opsin
NP_001079121.1 [Xenopus laevis]. 347 aa	violet-sensitive opsin
NP_001119548.1 [Xenopus (Silurana) tropicalis]. 349 aa	blue-sensitive opsin
NP_112277.1 [Rattus norvegicus]. 346 aa	short-wave-sensitive opsin 1
NP_031564.1 [Mus musculus]. 346 aa chromosome 6	short-wave-sensitive opsin 1
XP_001502785.1 [Equus caballus]. 347 aa	PREDICTED: short-wave-sensitive opsin 1-like
XP_539386.2 [Canis lupus familiaris]. 348 aa	PREDICTED: short-wave-sensitive opsin 1
NP_001009127.1 [Pan troglodytes]. 348 aa chromosome 7	short-wave-sensitive opsin 1
NP_001699.1 [Homo sapiens]. 348 aa chromosome 7	short-wave-sensitive opsin 1
EAL24112.1 [Homo sapiens].	opsin 1 (cone pigments), short-

348 aa chromosome 7	wave-sensitive (color blindness, tritan)
NP_999255.1 [Sus scrofa]. 351 aa	blue-sensitive opsin
NP_001138556.1 [Monodelphis domestica]. 346 aa protein	blue-sensitive opsin
NP_001070172.1 [Taeniopygia guttata]. 346 aa protein	blue-sensitive opsin
NP_524368.3 [Drosophila melanogaster]. 369 aa	rhodopsin 6, partial
NP_509725.2 [Caenorhabditis elegans]. 402 aa	hypothetical protein F41E7.3

PAX6

Accession	Description
NP_001231107.1 [Sus scrofa]. 355 aa chromosome 2	paired box protein Pax-6 isoform 2
NP_001231101.1 [Sus scrofa]. 422 aa chromosome 2	paired box protein Pax-6 isoform 1
XP_001368528.2 [Monodelphis domestica] 553 aa	PREDICTED: paired box protein Pax-6 isoform 1
EDL79721.1 [Rattus norvegicus]. 456 aa chromosome 3	paired box gene 6, isoform CRA_a
NP_037133.1 [Rattus norvegicus]. 422 aa chromosome 3	paired box protein Pax-6
EDL79723.1 [Rattus norvegicus]. 470 aa chromosome 3	paired box gene 6, isoform CRA_c
AAS48919.1 [Rattus norvegicus]. 436 aa chromosome 3	paired box 6 isoform 5a
NP_001231127.1 [Mus musculus]. 436 aa chromosome 2	paired box protein Pax-6 isoform 1
NP_038655.1 [Mus musculus]. 436 aa chromosome 2	paired box protein Pax-6 isoform 1
EDL27749.1 [Mus musculus] 436 aa chromosome 2	paired box gene 6, isoform CRA_b
EDL27748.1 [Mus musculus] 499 aa chromosome 2	paired box gene 6, isoform CRA_d, partial
EAW68233.1 [Homo sapiens]. 456 aa chromosome 11	paired box gene 6 (aniridia, keratitis), isoform CRA_a
NP_000271. 1 [Homo sapiens].	paired box protein Pax-6 isoform a

422 aa chromosome 11	
NP_001595.2 [Homo sapiens]. 436 aa chromosome 11	paired box protein Pax-6 isoform b
EAW68236.1 [Homo sapiens]. 470 aa chromosome 11	paired box gene 6 (aniridia, keratitis), isoform CRA_c
ACZ28705.1 [Homo sapiens]. 401 aa chromosome 11	paired box protein 6 isoform c
NP_001091013.1 [Canis lupus familiaris]. 436 aa	paired box gene 6
XP_001918200.2 [Equus caballus]. 395 aa	PREDICTED: paired box protein Pax-6
NP_990397.1 [Gallus gallus]. 436 aa	paired box protein Pax-6
NP_001035735.1 [Bos taurus]. 422 aa	paired box protein Pax-6
NP_001079413.1 [Xenopus laevis]. 453 aa	paired box 6
NP_001006763.1 [Xenopus tropicalis]. 424 aa	paired box 6
NP_571379.1 [Danio rerio]. 451 aa chromosome 25	paired box protein Pax-6
NP_524638.3 [Drosophila melanogaster]. 543 aa chromosome 4	twin of eyeless
AAF59395.4 [Drosophila melanogaster]. 543 aa chromosome 4	twin of eyeless
NP_001024570.1 [Caenorhabditis elegans]. 455 aa	Variable ABnormal morphology family member (vab-3)
AAA82991.1 [Caenorhabditis elegans]. 455 aa	variable abnormal-3

PDE6A

Accession	Description
NP_001003073.1 [Canis lupus familiaris]. 861 aa chromosome 4	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
XP_001503809.1 [Equus caballus]. 861 aa	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha-like
NP_001001526.2 [Bos taurus]. 859 aa	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
NP_000431.2 [Homo sapiens]. 860 aa chromosome 5	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha

XP_518030.2 [Pan troglodytes]. 860 aa chromosome 5	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha isoform 2
XP_003310946.1 [Pan troglodytes]. 779 aa chromosome 5	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
NP_666198.1 [Mus musculus]. 860 aa chromosome 18	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit
NP_001100856.1 [Rattus norvegicus]. 860 aa	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
XP_003124142.3 [Sus scrofa]. 1229 aa	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha-like
XP_003339581.1 [Monodelphis domestica]. 872 aa	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
NP_001007161.1 [Danio rerio]. 858 aa chromosome 14	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
NP_491544.2 [Caenorhabditis elegans]. 710 aa	PhosphoDiEsterase family member (pde-5)
NP_650369.3 [Drosophila melanogaster]. 1118 aa	phosphodiesterase 6

PDE6B

Accession	Description
NP_001002934.1 [Canis lupus familiaris]. 856 aa chromosome 3	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta precursor
XP_001487932.2 [Equus caballus]. 838 aa chromosome 3	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta-like
NP_776843.1 [Bos taurus]. 853 aa chromosome 6	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta precursor
DAA28431.1 [Bos taurus]. 784 aa chromosome 6	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta precursor
EAW82661.1 [Homo sapiens]. 803 aa chromosome 4	phosphodiesterase 6B, cGMP-specific, rod, beta (congenital stationary night blindness 3, autosomal dominant), isoform CRA_a
NP_000274.2 [Homo sapiens]. 854 aa chromosome 4	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta isoform 1
NP_001138763.1 [Homo sapiens]. 853 aa chromosome 4	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta isoform 2
EDL20120.1 [Mus musculus].	phosphodiesterase 6B, cGMP, rod receptor,

857 aa chromosome 5	beta polypeptide, isoform CRA_a
NP_032832.2 [Mus musculus]. 856 aa chromosome 5	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta precursor
NP_001099494.1 [Rattus norvegicus]. 856 aa chromosome 14	rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta
XP_001366327.1 [Monodelphis domestica]. 856 aa chromosome 5	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta-like
XP_002935501.1 [Xenopus (Silurana) tropicalis]. 852 aa	PREDICTED: LOW QUALITY PROTEIN: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta-like
XP_685002.1 [Danio rerio]. 854 aa chromosome 21	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta isoform 1
XP_002187940.1 [Taeniopygia guttata]. 821 aa	PREDICTED: phosphodiesterase 6B, cGMP-specific, rod, beta (congenital stationary night blindness 3, autosomal dominant)
XP_424876.2 [Gallus gallus]. 822 aa	PREDICTED: rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit beta
NP_650369.3 [Drosophila melanogaster]. 1118 aa	phosphodiesterase 6
NP_491544.2 [Caenorhabditis elegans]. 710 aa	phosphodiesterase family member (pde-5)

PDE6C

Accession	Description
NP_957165.1 [Danio rerio] 852 aa chromosome 12	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
XP_543934.3 [Canis lupus familiaris]. 844 aa	PREDICTED: cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
DAA14770.1 [Bos taurus]. 855 aa chromosome 26	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
NP_776844.1 [Bos taurus]. 855 aa chromosome 26	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
XP_003133198.1 [Sus scrofa]. 859 aa chromosome 14	PREDICTED: cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'-like
NP_990317.1 [Gallus gallus]. 862 aa chromosome 6	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha

XP_002190171.1 [Taeniopygia guttata]. 854 aa chromosome 6	PREDICTED: phosphodiesterase 6C, cGMP-specific, cone, alpha prime isoform 2
XP_002190140.1 [Taeniopygia guttata]. 862 aa chromosome 6	PREDICTED: phosphodiesterase 6C, cGMP-specific, cone, alpha prime isoform 1
XP_002937230.1 [Xenopus (Silurana) tropicalis]. 853 aa chromosome unknown	PREDICTED: cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'-like
NP_006195.3 [Homo sapiens]. 858 aa chromosome 10	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'
AAA92886.1 [Homo sapiens]. 858 aa chromosome 10	cone photoreceptor cGMP-phosphodiesterase alpha' subunit
XP_001148438.1 [Pan troglodytes]. 858 aa chromosome 10	PREDICTED: cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha
XP_001375179.1 [Monodelphis domestica].	PREDICTED: cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'-like
EDL41802 [Mus musculus] 861 aa chromosome 19	phosphodiesterase 6C, cGMP specific, cone, alpha prime, isoform CRA_a
EDL41803.1 [Mus musculus] 865 aa chromosome 19	phosphodiesterase 6C, cGMP specific, cone, alpha prime, isoform CRA_b, partial
NP_001164430.1 [Mus musculus] 836 aa chromosome 19	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha' isoform 2
NP_291092.1 [Mus musculus]. 861 aa chromosome 19	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha' isoform 1
NP_001101992.1 [Rattus norvegicus]. 861 aa chromosome 1	cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'
XP_001502478.1 [Equus caballus]. 854 aa chromosome 1	PREDICTED: cone cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha'-like
NP_650369.3 [Drosophila melanogaster]. 1118 aa	phosphodiesterase 6
NP_491544.2 [Caenorhabditis elegans]. 710 aa	PhosphoDiEsterase family member (pde-5)

PDE6D

Accession	Description
NP_002592.1 [Homo sapiens]. 150 aa chromosome 2	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta

XP_001144711.1 [Pan troglodytes]. 150 aa chromosome 2B	PREDICTED: hypothetical protein LOC738319
NP_032827.1 [Mus musculus]. 150 aa chromosome 1	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta
EDL40208.1 [Mus musculus]. 133 aa chromosome 1	phosphodiesterase 6D, cGMP-specific, rod, delta, isoform CRA_c, partial
NP_001102276.1 [Rattus norvegicus]. 131 aa chromosome 9	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta
EDL75597.1 [Rattus norvegicus]. 150 aa chromosome 9	phosphodiesterase 6D, cGMP-specific, rod, delta (predicted), isoform CRA_a
NP_001003156.1 [Canis lupus familiaris]. 150 aa chromosome 9	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta
XP_003365184.1 [Equus caballus]. 131 aa chromosome 6	PREDICTED: retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta-like
NP_776845.1 [Bos taurus]. 150 aa chromosome 2	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta
XP_003483803.1 [Sus scrofa]. 150 aa chromosome 15	PREDICTED: retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta-like
XP_001373880.1 [Monodelphis domestica]. 131 aa chromosome 2	PREDICTED: retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta-like
XP_422739.2 [Gallus gallus]. 131 aa chromosome 9	PREDICTED: retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta
XP_002193783.1 [Taeniopygia guttata]. 150 aa chromosome 9	PREDICTED: similar to ADP-ribosylation factor-like 2
XP_002937389.1 [Xenopus (Silurana) tropicalis]. 149 aa chromosome unknown	PREDICTED: retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic

	phosphodiesterase subunit delta-like
NP_001002708.1 [Danio rerio]. 150 aa chromosome 6	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit delta
NP_609246.1 [Drosophila melanogaster]. 151 aa	Prenyl-binding protein
NP_495490.1 [Caenorhabditis elegans]. 159 aa	Phosphodiesterase Delta-like family member (pdl-1)

RHO

Accession	Description
DAA16827.1 [Bos taurus]. 348 aa chromosome 22	rhodopsin
NP_001014890.1 [Bos taurus]. 348 aa chromosome 22	Rhodopsin
NP_254276.1 [Rattus norvegicus]. 348 aa chromosome 22	Rhodopsin
NP_001080517.1 [Xenopus laevis]. 354 aa	Rhodopsin
NP_001090803.1 [Xenopus (Silurana) 354 aa	Rhodopsin
NP_999386.1 [Sus scrofa]. 348 aa	Rhodopsin
AAA63392.1 [Mus musculus]. 348 aa chromosome 6	Opsin
NP_663358.1 [Mus musculus]. 348 aa chromosome 6	Rhodopsin
NP_571159.1 [Danio rerio]. 354 aa protein chromosome 8	Rhodopsin
CAX13341.1 [Danio rerio]. 354 aa chromosome 8	Rhodopsin
NP_001025777.1 [Gallus gallus]. 351 aa protein chromosome 12	Rhodopsin
NP_001070163.1 [Taeniopygia guttata]. 351 aa chromosome 12	Rhodopsin
NP_001008277.1 [Canis lupus familiaris]. 358 aa	Rhodopsin
XP_001490351.1 [Equus caballus].	PREDICTED: rhodopsin-like

314 aa	
NP_000530.1 [Homo sapiens] 348 aa chromosome 3	Rhodopsin
XP_516740.2 [Pan troglodytes]. 348 aa chromosome 3	PREDICTED: rhodopsin
XP_001366225.1 [Monodelphis domestica] 348 aa	PREDICTED: rhodopsin-like
NP_502959.1 [Caenorhabditis elegans]. 192 aa	RHO (small G protein) family member (rho-1)
NP_524368.3 [Drosophila melanogaster]. 369 aa	rhodopsin 6, partial
AAN13666.2 [Drosophila melanogaster]. 369 aa	rhodopsin 6, partial

RIBEYE

Accession	Description
NP_001164215.1 [Mus musculus]. 988 aa chromosome 7	C-terminal-binding protein 2 isoform 1
NP_034110.1 [Mus musculus]. 445 aa chromosome 7	C-terminal-binding protein 2 isoform 2
NP_445787.1 [Rattus norvegicus]. 988 aa chromosome 1	C-terminal-binding protein 2
EDM11743.1 [Rattus norvegicus]. 753 aa chromosome 1	C-terminal binding protein 2, isoform CRA_a
NP_073713.2 [Homo sapiens]. 985 aa chromosome 10	C-terminal-binding protein 2 isoform 2
NP_001320.1 [Homo sapiens]. 445 aa chromosome 10	C-terminal-binding protein 2 isoform 1
CAI16101.1 [Homo sapiens]. 513 aa chromosome 10	C-terminal binding protein 2
XP_508100.3 [Pan troglodytes]. 992 aa chromosome 10	PREDICTED: LOW QUALITY PROTEIN: hypothetical protein
DAA14642.1 [Bos taurus]. 982 aa chromosome 26	C-terminal-binding protein 2
NP_783643.1 [Bos taurus]. 982 aa chromosome 26	C-terminal-binding protein 2
XP_003433684.1 [Canis lupus familiaris]. 445 aa	PREDICTED: C-terminal-binding protein 2
XP_003363448.1 [Equus caballus]. 445 aa	PREDICTED: hypothetical protein LOC100056830

XP_002193579.1 [Taeniopygia guttata]. 974 aa	PREDICTED: C-terminal binding protein 2
NP_001015064.1 [Danio rerio]. 860 aa chromosome 12	C-terminal-binding protein 2
XP_001363827.2 [Monodelphis domestica]. 552 aa	PREDICTED: c-terminal-binding protein 2-like
XP_421817.3 [Gallus gallus]. 978 aa	PREDICTED: uncharacterized protein LOC423958
NP_001016866.1 [Xenopus (Silurana) tropicalis]. 437 aa	C-terminal binding protein 2
NP_001014617.1 [Drosophila melanogaster]. 476 aa	C-terminal binding protein, isoform E
NP_731764.1 [Drosophila melanogaster]. 386 aa	C-terminal binding protein, isoform D
NP_731763.1 [Drosophila melanogaster]. 386 aa	C-terminal binding protein, isoform C
NP_731762.1 [Drosophila melanogaster]. 386 aa	C-terminal binding protein, isoform B [Drosophila melanogaster].
NP_524336.2 [Drosophila melanogaster]. 386 aa	C-terminal binding protein, isoform A
NP_508983.2 [Caenorhabditis elegans]. 727 aa	CTBP (CtBP) transcriptional co-repressor homolog family member(ctbp-1)

RPE- retinal G-protein receptor

Accession	Description
NP_002912.2 [Homo sapiens]. 295 aa chromosome 10	RPE-retinal G protein-coupled receptor isoform 1
NP_001012740.1 [Homo sapiens]. 253 aa chromosome 10	RPE-retinal G protein-coupled receptor isoform 3
EAW80357.1 [Homo sapiens]. 112 aa chromosome 10	retinal G protein coupled receptor, isoform CRA_d
NP_001012738.1 [Homo sapiens]. 291 aa chromosome 10	RPE-retinal G protein-coupled receptor isoform 2
NP_786969.1 [Bos taurus]. 291 aa	RPE-retinal G protein-coupled receptor
XP_001927876.2 [Sus scrofa]. 291 aa	PREDICTED: RPE-retinal G protein-coupled receptor-like
XP_001495838.2 [Equus caballus].	PREDICTED: RPE-retinal G protein-

291 aa	coupled receptor-like
NP_067315.1 [Mus musculus]. 291 aa chromosome 14	RPE-retinal G protein-coupled receptor
EDL24895.1 [Mus musculus]. 293 aa chromosome 14	retinal G protein coupled receptor, partial
NP_001100769.1 [Rattus norvegicus]. 291 aa	RPE-retinal G protein-coupled receptor
XP_001154823.1 [Pan troglodytes]. 253 aa chromosome 10	PREDICTED: RPE-retinal G protein-coupled receptor isoform 2
XP_001154882.1 [Pan troglodytes]. 291 aa chromosome 10	PREDICTED: RPE-retinal G protein-coupled receptor isoform 3
XP_546190.2 [Canis lupus familiaris]. 291 aa chromosome 4	PREDICTED: RPE-retinal G protein-coupled receptor isoform 1
XP_864780.1 [Canis lupus familiaris]. 299 aa chromosome 4	PREDICTED: RPE-retinal G protein-coupled receptor isoform 3
XP_864762.1 [Canis lupus familiaris]. 253 aa chromosome 4	PREDICTED: RPE-retinal G protein-coupled receptor isoform 2
XP_002193326.1 [Taeniopygia guttata]. 295 aa chromosome 6	PREDICTED: retinal G-protein coupled receptor
NP_001026387.1 [Gallus gallus]. 295 aa	RPE-retinal G protein-coupled receptor
NP_001086324.1 [Xenopus laevis]. 295 aa	retinal G protein coupled receptor
NP_001016013.1 [Xenopus (Silurana) tropicalis 295 aa	retinal G protein coupled receptor
NP_001017877.1 [Danio rerio]. 295 aa chromosome 13	RPE-retinal G protein-coupled receptor
NP_725196.1 [Drosophila melanogaster]. 353 aa	G protein alpha49B, isoform D
NP_725197.1 [Drosophila melanogaster]. 353 aa	G protein alpha49B, isoform F
NP_725195.1 [Drosophila melanogaster]. 353 aa	G protein alpha49B, isoform G
NP_741080.1 [Caenorhabditis elegans]. 375 aa	GEX Interacting protein family member (gei-16)

SIX 3

Accession	Description
NP_005404.1 [Homo sapiens].	homeobox protein SIX3

332 aa chromosome 2	
AAD11939.1 [Homo sapiens] 332 aa chromosome 2	homeobox protein Six3
EAX00268.1 [Homo sapiens]. 332 aa chromosome 2	sine oculis homeobox homolog 3 (Drosophila), isoform CRA_a
EAX00267.1 [Homo sapiens]. 332 aa chromosome 2	sine oculis homeobox homolog 3 (Drosophila), isoform CRA_a
XP_525749.3 [Pan troglodytes]. 263 aa chromosome 2A	PREDICTED: homeobox protein SIX3
XP_538477.4 [Canis lupus familiaris]. 580 aa	PREDICTED: homeobox protein SIX3
CAA62379.1 [Mus musculus]. 352 aa chromosome 17	SIX3 protein
NP_035511.2 [Mus musculus]. 333 aa chromosome 17	homeobox protein SIX3
NP_076480.1 [Rattus norvegicus]. 337 aa chromosome 6	homeobox protein SIX3
NP_001180053.1 [Bos taurus]. 328 aa chromosome 11	homeobox protein SIX3
DAA24767.1 [Bos taurus]. 328 aa chromosome 11	SIX homeobox 3-like
NP_989695.1 [Gallus gallus]. 314 aa	homeobox protein SIX3
XP_002193827.1 [Taeniopygia guttata]. 329 aa	PREDICTED: similar to sine oculis homeobox homolog 3
NP_571437.1 [Danio rerio]. 294 aa chromosome 13	homeobox protein SIX3
NP_001079171.1 [Xenopus laevis]. 291 aa	SIX homeobox 3
XP_002932435.1 [Xenopus (Silurana)tropicalis]. 301 aa	PREDICTED: homeobox protein SIX3-like
XP_001375682.1 [Monodelphis domestica]. 333 aa	PREDICTED: homeobox protein SIX3-like
NP_476733.1 [Drosophila melanogaster]. 416 aa chromosome 2R	sine oculis
NP_524695.2 [Drosophila melanogaster]. 487 aa chromosome 2R	optix, isoform A
NP_505958.1 [Caenorhabditis elegans]. 439 aa chromosome 5	Homeobox family member (ceh-32)

Accession	Description
NP_504419.1 [Caenorhabditis elegans]. 256 aa chromosome 5	Homeobox family member (ceh-34)

SIX 6

Accession	Description
NP_031400.2 [Homo sapiens]. 246 aa chromosome 14	homeobox protein SIX6
AAD49844.1 [Homo sapiens]. 298 aa chromosome 14	homeobox containing transcription factor SIX6
XP_522870.3 [Pan troglodytes]. 246 aa chromosome 14	PREDICTED: homeobox protein SIX6
XP_002200636.1 [Taeniopygia guttata]. 246 aa	PREDICTED: similar to SIX homeobox 6
NP_990325.1 [Gallus gallus] 246 aa	homeobox protein SIX6
XP_547840.3 [Canis lupus familiaris]. 246 aa	PREDICTED: homeobox protein SIX6
NP_001098463.1 [Bos taurus]. 222 aa chromosome 10	homeobox protein SIX6
DAA25109.1 [Bos taurus]. 222 aa chromosome 10	SIX homeobox 6
XP_003121893.2 [Sus scrofa]. 325 aa chromosome 1	PREDICTED: homeobox protein SIX6-like
NP_001101502.1 [Rattus norvegicus]. 246 aa	homeobox protein SIX6
NP_035514.1 [Mus musculus]. 246 aa chromosome 12	homeobox protein SIX6
EDL36517.1 [Mus musculus]. 246 aa chromosome 12	sine oculis-related homeobox 6 homolog (Drosophila)
NP_001081933.1 [Xenopus laevis]. 244 aa	SIX homeobox 6
NP_001093696.1 [Xenopus (Silurana) tropicalis]. 244 aa	SIX homeobox 6
NP_001018421.1 [Danio rerio]. 245 aa chromosome 20	sine oculis-related homeobox 6b
XP_001369291.1 [Monodelphis domestica]. 246 aa	PREDICTED: homeobox protein SIX6-like
NP_504420.1 [Caenorhabditis elegans].	Homeobox family member (ceh-33)

261 aa chromosome 5	
NP_524695.2 [Drosophila melanogaster]. 487 aa chromosome 2R	optix, isoform A
NP_476733.1 [Drosophila melanogaster]. 416 aa chromosome 2R	sine oculis

APPENDIX B:

AUDITORY PROTEIN ACCESSION NUMBERS,

IN ALPHABETICAL ORDER

CDH23

Accession	Description
NP_075859 [Mus musculus]. 3354 aa Chromosome 10	cadherin-23 isoform 1 precursor
AAG52817.1 [Mus musculus]. 3354 aa Chromosome 10	cadherin-related 23 protein
NP_001239564.1 [Mus musculus]. 3352 aa Chromosome 10	cadherin-23 isoform 2 precursor
AAK07670.1 [Mus musculus]. 3322 aa Chromosome 10	cadherin 23
NP_446096.1 [Rattus norvegicus]. 3317 aa Chromosome 20	cadherin-23 precursor
XP_001925718.2 [Sus scrofa]. 3354 aa Chromosome 14	PREDICTED: cadherin-23
NP_071407.4 [Homo sapiens]. 3354 aa Chromosome 10	cadherin-23 isoform 1 precursor
NP_001165405.1 [Homo sapiens]. 1079 aa Chromosome 10	cadherin-23 isoform 7
NP_001165402.1 [Homo sapiens]. 1061 aa Chromosome 10	cadherin-23 isoform 4 precursor
NP_001165404.1 [Homo sapiens]. 1114 aa Chromosome 10	cadherin-23 isoform 6
XP_003434519.1 [Canis lupus familiaris]. 3354 aa Chromosome 4	PREDICTED: cadherin-23 isoform 1
XP_003434520.1 [Canis lupus familiaris]. 1079 aa Chromosome 4	PREDICTED: cadherin-23 isoform 2
XP_001917733 [Equus caballus] 3354 aa Chromosome 1	PREDICTED cadherin-23
XP_421595.2 [Gallus gallus]. 3365 aa Chromosome 6	PREDICTED: cadherin-23
XP_001365044.2 [Monodelphis domestica]. 3103 aa Chromosome unknown	PREDICTED: cadherin-23, partial
XP_002939565.1 [Xenopus (Silurana) tropicalis]. 2570 aa Chromosome unknown	PREDICTED: cadherin-23-like
XP_507839.3 [Pan troglodytes]. 1422 aa Chromosome 10	PREDICTED: cadherin-23
NP_001178135.1 [Bos taurus]. 3354 aa Chromosome 28	cadherin-23 precursor

DAA14294.1 [Bos taurus]. 3354 aa Chromosome 28	cadherin-23-like
NP_999974.1 [Danio rerio]. 3366 aa Chromosome 13	cadherin-23 precursor
NP_497340.2 [Caenorhabditis elegans]. 2922 aa Chromosome 3	Cadherin family member (cdh-12)
NP_648973.1 [Drosophila melanogaster]. 1820 aa Chromosome 3L	Cad74A, isoform A

EYA1

Accession	Description
AAB48017.1 [Mus musculus]. 591 aa Chromosome 1	Eya1
NP_034294.2 [Mus musculus]. 587 aa Chromosome 1	eyes absent homolog 1 isoform 1
NP_001239121.1 [Mus musculus]. 558 aa Chromosome 1	eyes absent homolog 1 isoform 2
NP_001083888.1 [Xenopus laevis] 592 aa	eyes absent homolog 1
NP_571268 [Danio rerio]. 609 aa Chromosome 24	eyes absent homolog 1
DAA22681.1 [Bos taurus]. 589 aa Chromosome 14	eyes absent homolog 1
XP_002199188 [Taeniopygia guttata]. 593 aa Chromosome 2	PREDICTED: eyes absent homolog 1
XP_418290.3 [Gallus gallus]. 660 aa Chromosome 2	PREDICTED: eyes absent homolog 1 isoform 2
XP_859603.1 [Canis lupus familiaris]. 557 aa Chromosome 29	PREDICTED: eyes absent homolog 1 isoform 3
XP_003640064.1 [Canis lupus familiaris]. 592 aa Chromosome 29	PREDICTED: eyes absent homolog 1
XP_001492875.1 [Equus caballus]. 557 aa Chromosome 9	PREDICTED: eyes absent homolog 1 isoform 3
XP_001492823.1 [Equus caballus]. 592 aa Chromosome 9	PREDICTED: eyes absent homolog 1 isoform 1
XP_003311799.1 [Pan troglodytes]. 557 aa Chromosome 8	PREDICTED: eyes absent homolog 1
XP_001164492.1 [Pan troglodytes]. 592 aa Chromosome 8	PREDICTED: eyes absent homolog 1 isoform 4
XP_001164379.1 [Pan troglodytes].	PREDICTED: eyes absent homolog 1

592 aa Chromosome 8	isoform 1
XP_002729501.1 [Rattus norvegicus]. 587 aa Chromosome 5	PREDICTED: eyes absent 1 isoform 1
XP_578437.2 [Rattus norvegicus]. 557 aa Chromosome 5	PREDICTED: eyes absent 1 isoform 2
NP_742057.1 [Homo sapiens]. 559 aa Chromosome 8	eyes absent homolog 1 isoform a
NP_742056.1 [Homo sapiens]. 557 aa Chromosome 8	eyes absent homolog 1 isoform
NP_000494.2 [Homo sapiens]. 592 aa Chromosome 8	eyes absent homolog 1 isoform b
NP_742055.1 [Homo sapiens]. 592 aa Chromosome 8	eyes absent homolog 1 isoform b
NP_523492.1 [Drosophila melanogaster]. 760 aa Chromosome 2L	eyes absent, isoform B
NP_723188.1 [Drosophila melanogaster]. 766 aa Chromosome 2L	eyes absent, isoform A
NP_001021055.1 [Caenorhabditis elegans]. 503 aa Chromosome 1	EYA (Drosophila eyes absent) homolog family member (eya-1)
NP_001021056.1 [Caenorhabditis elegans]. 469 aa Chromosome 1	EYA (Drosophila eyes absent) homolog family member (eya-1)

FGF3

Accession	Description
NP_570830.1 [Rattus norvegicus]. 245 aa Chromosome 1	fibroblast growth factor 3 precursor
EDM12254.1 [Rattus norvegicus]. 245 aa Chromosome 1	fibroblast growth factor 3, isoform CRA_a
EDM12253.1 [Rattus norvegicus]. 245 aa Chromosome 1	fibroblast growth factor 3, isoform CRA_a
NP_032033.2 [Mus musculus]. 245 aa Chromosome 7	fibroblast growth factor 3 precursor
EDL18270.1 [Mus musculus]. 245 aa Chromosome 7	fibroblast growth factor 3, isoform CRA_a
EDL18269.1 [Mus musculus]. 245 aa Chromosome 7	fibroblast growth factor 3, isoform CRA_a
DAA13506.1 [Bos taurus]. 236 aa Chromosome 29	fibroblast growth factor 3-like
XP_002699485.1 [Bos taurus].	PREDICTED: fibroblast growth factor 3

236 aa Chromosome 29	
XP_875907.2 [Bos taurus]. 236 aa Chromosome 29	PREDICTED: fibroblast growth factor 3
NP_990658.1 [Gallus gallus]. 220 aa Chromosome 5	fibroblast growth factor 3 precursor
NP_571366 [Danio rerio]. 256 aa Chromosome 7	fibroblast growth factor 3 precursor
XP_002194557.1 [Taeniopygia guttata]. 228 aa Chromosome 5	PREDICTED: fibroblast growth factor 3 (murine mammary tumor virus integration site (v-int-2) oncogene homolog)
NP_001079132.1 [Xenopus laevis]. 237 aa	fibroblast growth factor 3 precursor
XP_001373768.1 [Monodelphis domestica]. 254 aa Chromosome 5	PREDICTED: fibroblast growth factor 3
XP_854497.2 [Canis lupus familiaris]. 170 aa Chromosome 18	PREDICTED: fibroblast growth factor 3
XP_522092.2 [Pan troglodytes]. 239 aa Chromosome 11	PREDICTED: fibroblast growth factor 3
NP_005238.1 [Homo sapiens]. 239 aa Chromosome 11	fibroblast growth factor 3 precursor
EAW74753.1 [Homo sapiens]. 239 aa Chromosome 11	fibroblast growth factor 3 (murine mammary tumor virus integration site (v-int-2) oncogene homolog), isoform CRA_a
NP_732452.1 [Drosophila melanogaster]. 770 aa Chromosome 3R	branchless, isoform A
NP_001138083.1 [Drosophila melanogaster]. 760 aa Chromosome 3R	branchless, isoform C
NP_498403.1 [Caenorhabditis elegans]. 425 aa Chromosome 3	Lethal family member (let-756)

IRF**6**

Accession	Description
DAA20980.1 [Bos taurus]. 467 aa Chromosome 16	interferon regulatory factor 6
NP_001070402.1 [Bos taurus]. 467 aa Chromosome 16	interferon regulatory factor 6
XP_001490816.1 [Equus caballus]. 467 aa Chromosome 5	PREDICTED: interferon regulatory factor 6 isoform 1
NP_999443.1 [Sus scrofa]. 467 aa Chromosome 9	interferon regulatory factor 6
CAI95692.1 [Homo sapiens]. 276 aa Chromosome 1	interferon regulatory factor 6, partial
NP_006138.1 [Homo sapiens]. 467 aa Chromosome 1	interferon regulatory factor 6 isoform 1
AEL89176.1.1 [Homo sapiens]. 467 aa Chromosome 1	interferon regulatory factor 6
NP_001193625.1 [Homo sapiens]. 372 aa Chromosome 1	interferon regulatory factor 6 isoform 2
XP_001168751.1 [Pan troglodytes]. 467 aa Chromosome 1	PREDICTED: interferon regulatory factor 6 isoform 1
XP_514168.2 [Pan troglodytes]. 467 aa Chromosome 1	PREDICTED: interferon regulatory factor 6 isoform 3
NP_001081215.1 [Xenopus laevis]. 459 aa	interferon regulatory factor 6, gene 2
NP_001025493.1 [Xenopus (Silurana) tropicalis]. 460 aa	interferon regulatory factor 6
NP_001102329.1 [Rattus norvegicus]. 467 aa Chromosome 13	interferon regulatory factor 6
NP_058547.2 [Mus musculus]. 467 aa Chromosome 13	interferon regulatory factor 6
XP_537138.3 [Canis lupus familiaris]. 462 aa Chromosome 7	PREDICTED: interferon regulatory factor 6
XP_001366308.1 [Monodelphis domestica]. 466 aa Chromosome 2	PREDICTED: interferon regulatory factor 6
XP_417990.3 [Gallus gallus]. 460 aa Chromosome 26	PREDICTED: interferon regulatory factor 6
ABB77237.1 [Gallus gallus]. 457 aa Chromosome 26	interferon regulatory factor 6, partial
XP_002196388.1 [Taeniopygia guttata].	PREDICTED: interferon regulatory factor

460 aa Chromosome 26	6
NP_956892.1 [Danio rerio]. 492 aa Chromosome 22	interferon regulatory factor 6
NP_650273.2 [Drosophila melanogaster]. 994 aa Chromosome 3R	CG8773
NP_498670.1 [Caenorhabditis elegans]. 2585 aa Chromosome 3	Protein FBN-1

MYO7A

Accession	Description
NP_694515.1 [Danio rerio]. 2179 aa Chromosome 18	myosin-VIIa
XP_002937295.1 [Xenopus (Silurana) tropicalis] 2143 aa Chromosome unknown	myosin-VIIa-like
XP_002189823.1 [Taeniopygia guttata]. 2213 aa Chromosome 1	PREDICTED: similar to Myosin VIIa
EDM18456.1 [Rattus norvegicus]. 2117 aa Chromosome 1	myosin VIIA, isoform CRA_b
EDM18455.1 [Rattus norvegicus]. 2155 aa Chromosome 1	myosin VIIA, isoform CRA_a
NP_703203.1 [Rattus norvegicus]. 2177 aa Chromosome 11	myosin-VIIa
AAB40708.1 [Mus musculus]. 2215 aa Chromosome 7	myosin VIIa
NP_001243010.1 [Mus musculus]. 2215 aa Chromosome 7	unconventional myosin-VIIa isoform 1
NP_032689.2 [Mus musculus]. 2177 aa Chromosome 7	unconventional myosin-VIIa isoform 2
EDL16329.1 [Mus musculus]. 2215 aa Chromosome 7	myosin VIIa, isoform CRA_b
EDL16328.1 [Mus musculus]. 2204 aa Chromosome 7	myosin VIIa, isoform CRA_a
XP_417277.3 [Gallus gallus]. 2206 aa Chromosome 1	PREDICTED: myosin-VIIa
XP_542292.3 [Canis lupus familiaris]. 2218 aa Chromosome 21	PREDICTED: myosin-VIIa
XP_001494652 [Equus caballus]. 2162 aa Chromosome 7	PREDICTED: myosin-VIIa
XP_002693553.2 [Bos taurus]. 2269 aa Chromosome 15	PREDICTED: myosin-VIIa

XP_870166.5 [Bos taurus]. 2251 aa Chromosome 15	PREDICTED: myosin-VIIa isoform 2
NP_001093398.1 [Sus scrofa]. 2177 aa Chromosome 9	unconventional myosin-VIIa
XP_003313297.1 [Pan troglodytes]. 2217 aa Chromosome 11	PREDICTED: LOW QUALITY PROTEIN: myosin-VIIa
EAW75023.1 [Homo sapiens]. 1958 aa Chromosome 11	myosin VIIA, isoform CRA_f
NP_001120651.2 [Homo sapiens]. 1178 aa Chromosome 11	unconventional myosin-VIIa isoform 3
NP_001120652.1 [Homo sapiens]. 2175 aa Chromosome 11	unconventional myosin-VIIa isoform 2
NP_000251.3 [Homo sapiens]. 2215 aa Chromosome 11	unconventional myosin-VIIa isoform 1
EAW75022.1 [Homo sapiens]. 2177 aa Chromosome 11	myosin VIIA, isoform CRA_e
EAW75020.1[Homo sapiens]. 2277 aa	myosin VIIA, isoform CRA_c
EAW75018.1 [Homo sapiens]. 1178 aa Chromosome 11	myosin VIIA, isoform CRA_a
EAW75019.1 [Homo sapiens]. 2215 aa Chromosome 11	myosin VIIA, isoform CRA_b
EAW75021.1 [Homo sapiens]. 791 aa Chromosome 11	myosin VIIA, isoform CRA_
NP_523571.1 [Drosophila melanogaster]. 2167 aa Chromosome 2L	crinkled, isoform B
NP_723895.1 [Drosophila melanogaster]. 2167 aa Chromosome 2L	crinkled, isoform A
NP_508420.1 [Caenorhabditis elegans]. 2098 aa Chromosome 10	Heavy chain, Unconventional Myosin family member (hum-6)

OTX1

Accession	Protein
AF424700_1 same as AAL24809.1 [Mus musculus]. 355 aa Chromosome 11	Otx1
NP_035153.1 [Mus musculus]. 355 aa Chromosome 11	homeobox protein OTX1

XP_001917473.1 [Equus caballus]. 355 aa Chromosome 15	PREDICTED: LOW QUALITY PROTEIN: homeobox protein OTX1-like
NP_037241.1 [Rattus norvegicus]. 355 aa Chromosome 14	homeobox protein OTX1
NP_001081009.1 Xenopus laevis]. 339 aa	orthodenticle homeobox 1
NP_989216.1 [Xenopus (Silurana) tropicalis]. 336 aa	orthodenticle homeobox 1
NP_571325.2 [Danio rerio]. 323 aa Chromosome 17	homeobox protein OTX1 B
XP_003640955.1 [Gallus gallus]. 314 aa Chromosome 3	PREDICTED: homeobox protein OTX1-like
NP_001192946.1 [Bos taurus]. 355 aa Chromosome 11	orthodenticle homeobox 1
DAA24647.1 [Bos taurus]. 355 aa Chromosome 11	orthodenticle homeobox 1-like
XP_852530.1 [Canis lupus familiaris] 355 aa Chromosome 10	PREDICTED: homeobox protein OTX1 isoform 1
XP_001162799.1 [Pan troglodytes]. 354 aa Chromosome 2A	PREDICTED: homeobox protein OTX1 isoform 2
NP_001186699.1 [Homo sapiens]. 354 aa Chromosome 2	homeobox protein OTX1
NP_055377.1 [Homo sapiens]. 354 aa Chromosome 2	homeobox protein OTX1
XP_003125144.1 [Sus scrofa]. 356 aa Chromosome 3	PREDICTED: homeobox protein OTX1-like
XP_001374919.2 [Monodelphis domestica]. 367 aa Chromosome 1	PREDICTED: homeobox protein OTX1-like
NP_001014727.2 [Drosophila melanogaster]. 542 aa Chromosome X	ocelliless
[Caenorhabditis elegans]. 278 aa Chromosome X	Homeobox family member (ceh-37)

PAX3

Accession	Description
NP_852124.1 [Homo sapiens]. 505 aa Chromosome 2	paired box protein Pax-3 isoform PAX3e
NP_852126.1 [Homo sapiens]. 403 aa Chromosome 2	paired box protein Pax-3 isoform PAX3g
NP_852125.1 [Homo sapiens]. 407 aa Chromosome 2	paired box protein Pax-3 isoform PAX3h
NP_852122.1 [Homo sapiens]. 479 aa Chromosome 2	paired box protein Pax-3 isoform PAX3
NP_852123.1 [Homo sapiens]. 484 aa Chromosome 2	paired box protein Pax-3 isoform PAX3d
NP_001120838.1 [Homo sapiens]. 483 aa Chromosome 2	paired box protein Pax-3 isoform PAX3i
XP_001165390.1 [Pan troglodytes]. 505 aa Chromosome 2B	PREDICTED: paired box protein Pax-3 isoform 6
XP_545664.3 [Canis lupus familiaris]. 482 aa Chromosome 37	PREDICTED: paired box protein Pax-3
NP_001193747.1 [Bos taurus]. 484 aa Chromosome 2	paired box protein Pax-3
NP_032807.3 [Mus musculus]. 479 aa Chromosome 1	paired box protein Pax-3 isoform a
NP_001152992.1 [Mus musculus]. 484 aa Chromosome 1	paired box protein Pax-3 isoform b
NP_446162.1 [Rattus norvegicus]. 484 aa Chromosome 9	paired box 3
XP_002194011.1 [Taeniopygia guttata]. 529 aa Chromosome 9	PREDICTED: similar to paired box 3
NP_989600.1 [Gallus gallus]. 484 aa Chromosome 9	paired box 3
NP_571352.1 [Danio rerio]. 509 aa Chromosome 2	paired box protein Pax-3
XP_001495210.1 [Equus caballus]. 505 aa Chromosome 6	PREDICTED: paired box protein Pax-3 isoform 2
XP_001495022.3 [Equus caballus]. 483 aa Chromosome 6	PREDICTED: paired box protein Pax-3 isoform 1
XP_003365174.1 [Equus caballus]. 479 aa Chromosome 6	PREDICTED: paired box protein Pax-3
XP_003365175.1 [Equus caballus]. 403 aa Chromosome 6	PREDICTED: paired box protein Pax-3
XP_001495229.2 [Equus caballus].	PREDICTED: paired box protein Pax-3

424 aa Chromosome 6	isoform 3
XP_001365807.1 [Monodelphis domestica]. 484 aa Chromosome 7	PREDICTED: paired box protein Pax-3
NP_001088993.1 [Xenopus laevis]. 484 aa	paired box protein Pax-3-A
NP_001006776.2 [Xenopus (Silurana) tropicalis]. 461 aa	paired box protein Pax-3
XP_003361395.2 [Sus scrofa]. 509 aa Chromosome unknown	PREDICTED: paired box protein Pax-3
NP_523863.1 [Drosophila melanogaster]. 427 aa Chromosome 2R	gooseberry
NP_523862.1 [Drosophila melanogaster]. 449 aa Chromosome 2R	gooseberry-neuro
NP_001024570.1 [Caenorhabditis elegans]. 455 aa Chromosome 10	Variable ABnormal morphology family member (vab-3)

SIX1

Accession	Description
XP_509988.2 [Pan troglodytes]. 284 aa Chromosome 14	PREDICTED: homeobox protein SIX1
NP_005973.1 [Homo sapiens]. 284 aa Chromosome 14	homeobox protein SIX1
XP_547841.3 [Canis lupus familiaris]. 284 aa Chromosome 8	PREDICTED: homeobox protein SIX1
NP_033215.2 [Mus musculus]. 284 aa Chromosome 12	homeobox protein SIX1
CAA56585.1 [Mus musculus]. 273 aa Chromosome 12	six1, partial
NP_446211.1 [Rattus norvegicus]. 284 aa Chromosome 6	sine oculis-related homeobox 1 homolog
XP_588692.2 [Bos taurus]. 484 aa Chromosome 10	PREDICTED: homeobox protein SIX1
DAA25162.1 [Bos taurus]. 567 aa Chromosome 10	SIX homeobox 1
NP_001186647.1 [Sus scrofa]. 284 aa Chromosome 1	homeobox protein SIX1
NP_001038150.1 [Gallus gallus]. 282 aa Chromosome 5	homeobox protein SIX1

XP_001377489.2 [Monodelphis domestica]. 284 aa Chromosome 1	PREDICTED: homeobox protein SIX1-like
XP_001492836.1 [Equus caballus]. 536 aa Chromosome 24	PREDICTED: homeobox protein SIX1-like
NP_001082027.1 [Xenopus laevis]. 284 aa	SIX homeobox 1
NP_001093693.1 [Xenopus (Silurana) tropicalis]. 284 aa	SIX homeobox 1
NP_996978.1 [Danio rerio]. 284 aa Chromosome 20	homeobox protein SIX1
NP_476733.1 [Drosophila melanogaster]. 416 aa Chromosome 2R	sine oculis
NP_504419.1 [Caenorhabditis elegans]. 256 aa Chromosome 5	Homeobox family member (ceh-34)

SOX10

Accessions	Description
CAG30470.1 [Homo sapiens]. 466 aa chromosome 22	SOX10
NP_008872.1[Homo sapiens]. 466 aa chromosome 22	transcription factor SOX-10
NP_990123.1 [Gallus gallus]. 461 aa chromosome 1	transcription factor SOX-10
XP_538379.3 [Canis lupus familiaris]. 468 aa chromosome 10	PREDICTED: transcription factor SOX-10 [Canis lupus familiaris].
NP_001180176.1 [Bos taurus]. 469 aa chromosome 5	transcription factor SOX-10
DAA29140.1 [Bos taurus]. 469 aa chromosome 5	SRY (sex determining region Y)-box 10
XP_003481595.1 [Sus scrofa]. 471 aa chromosome 5	PREDICTED: transcription factor SOX-10-like
NP_001093403.1 [Sus scrofa]. 469 aa chromosome 5	transcription factor SOX-10
NP_035567.1 [Mus musculus]. 466 aa chromosome 15	transcription factor SOX-10
EDM15819.1 [Rattus norvegicus]. 488 aa chromosome 7	SRY-box containing gene 10, isoform CRA_a
NP_062066.1 [Rattus norvegicus].	transcription factor SOX-10

466 aa chromosome 7	
NP_571950.1 [Danio rerio]. 485 aa chromosome 3	SRY-box containing gene 10
XP_002198867.1 [Taeniopygia guttata] 463 aa chromosome 1A	PREDICTED: SRY (sex determining region Y)-box 10
XP_001916500.2 [Equus caballus] 452 aa chromosome 28	PREDICTED: LOW QUALITY PROTEIN: transcription factor SOX-10
XP_525590.3 [Pan troglodytes]. 459 aa chromosome 22	PREDICTED: transcription factor SOX-10
NP_001082358.1 [Xenopus laevis]. 446 aa	transcription factor Sox-10
NP_001093691.1 [Xenopus (Silurana) tropicalis]. 436 aa	transcription factor Sox-10
NP_497910.1 [Caenorhabditis elegans] 257aa chromosome 3	Protein RPS-1
NP_651839.1 [Drosophila melanogaster] 529 aa chromosome 3R	Sox100B

SPT6

Accession	Description
NP_033323.2 [Mus musculus]. 1726 aa chromosome 11	transcription elongation factor SPT6
CAI24323.1 [Mus musculus]. 1726 aa chromosome 11	suppressor of Ty 6 homolog (S. cerevisiae)
AAB18950.1 [Mus musculus]. 1726 aa chromosome 11	Supt6h
NP_001178749.1 [Rattus norvegicus]. 1726 aa chromosome 10	transcription elongation factor SPT6
XP_001142885.2 [Pan troglodytes]. 1526 aa chromosome 17	PREDICTED: transcription elongation factor SPT6 isoform 3
XP_001143115.2 [Pan troglodytes]. 1726 aa chromosome 17	PREDICTED: transcription elongation factor SPT6 isoform 5
XP_003315613.1 [Pan troglodytes]. 1526 aa chromosome 17	PREDICTED: transcription elongation factor SPT6
NP_003161.2 [Homo sapiens]. 1726 aa chromosome 17	transcription elongation factor SPT6
EAW51117.1 [Homo sapiens]. 1726 aa chromosome 17	suppressor of Ty 6 homolog (S. cerevisiae), isoform CRA_b
EAW51118.1 [Homo sapiens].	suppressor of Ty 6 homolog (S.

1738 aa chromosome 17	cerevisiae), isoform CRA_c
EAW51119.1 [Homo sapiens]. 1679 aa chromosome 17	suppressor of Ty 6 homolog (S. cerevisiae), isoform CRA_d
XP_537747.2 [Canis lupus familiaris] 1726 aa chromosome 9	PREDICTED: transcription elongation factor SPT6 isoform 1
XP_001504206.1 [Equus caballus]. 1726 aa chromosome 11	PREDICTED: transcription elongation factor SPT6
XP_003484346.1 [Sus scrofa]. 748 aa chromosome unknown	PREDICTED: transcription elongation factor SPT6-like, partial
DAA19054.1 [Bos taurus]. 1726 aa chromosome 19	suppressor of Ty 6 homolog
NP_001180055.1 [Bos taurus]. 1726 aa chromosome 19	transcription elongation factor SPT6
XP_001368732.1 [Monodelphis domestica]. 1726 aa chromosome 2	PREDICTED: transcription elongation factor SPT6
XP_423183.3 [Gallus gallus]. 1726 aa chromosome 19	PREDICTED: transcription elongation factor SPT6
XP_002199580.1 [Taeniopygia guttata]. 196 aa	PREDICTED: similar to SUPT6H protein, partial
NP_660094.1 [Danio rerio]. 1726 aa chromosome 21	transcription elongation factor SPT6
NP_001072665.1 [Xenopus (Silurana)tropicalis]. 519 aa	suppressor of Ty 6 homolog isoform 1
AAF46140.1 [Drosophila melanogaster]. 1831 aa chromosome 10	Spt6
NP_651962.2 [Drosophila melanogaster]. 1831 aa chromosome 10	Spt6
NP_497969.1 [Caenorhabditis elegans]. 1521 aa 2 chromosome 3	abnormal EMBroygenesis family member (emb-5)

UGDH

Accession	Description
DAA28716.1 [Bos taurus]. 494 aa chromosome 6	UDP-glucose dehydrogenase
NP_776636.1 [Bos taurus]. 494 aa chromosome 6	UDP-glucose 6-dehydrogenase
NP_001103872.1 [Danio rerio]. 493 aa chromosome 1	UDP-glucose 6-dehydrogenase

NP_001079465.1 [Xenopus laevis]. 494 aa	UDP-glucose 6-dehydrogenase
NP_001013628.1 [Xenopus (Silurana) tropicalis]. 494 aa	UDP-glucose 6-dehydrogenase
NP_001171629.1 [Homo sapiens]. 427 aa chromosome 4	UDP-glucose 6-dehydrogenase isoform 2
NP_001171630.1 [Homo sapiens]. 397 aa	UDP-glucose 6-dehydrogenase isoform 3
NP_003350.1 [Homo sapiens]. 494 aa chromosome 4	UDP-glucose 6-dehydrogenase isoform 1
EAW92938.1 [Homo sapiens]. 494 aa chromosome 4	UDP-glucose dehydrogenase, isoform CRA_c
EAW92937.1 [Homo sapiens]. 427 aa chromosome 4	UDP-glucose dehydrogenase, isoform CRA_b
EAW92936.1 [Homo sapiens]. 494 aa chromosome 4	UDP-glucose dehydrogenase, isoform CRA_a
NP_001012599.1 [Gallus gallus]. 494 aa chromosome 4	UDP-glucose 6-dehydrogenase
EDL90061.1 [Rattus norvegicus]. 493 aa chromosome 14	UDP-glucose dehydrogenase, isoform CRA_a
NP_112615.1 [Rattus norvegicus]. 493 aa chromosome 14	UDP-glucose 6-dehydrogenase
NP_033492.1 [Mus musculus]. 493 aa chromosome 5	UDP-glucose 6-dehydrogenase
AAC36096.1 [Mus musculus]. 493 aa chromosome 5	UDP-glucose dehydrogenase
EDL37741.1 [Mus musculus]. 493 aa chromosome 5	UDP-glucose dehydrogenase
XP_003434444.1 [Canis lupus familiaris]. 427 aa chromosome 3	PREDICTED: UDP-glucose 6-dehydrogenase
XP_536254.2 [Canis lupus familiaris]. 494 aa chromosome 3	PREDICTED: UDP-glucose 6-dehydrogenase isoform 1
XP_003364775.1 [Equus caballus]. 427 aa chromosome 3	PREDICTED: UDP-glucose 6-dehydrogenase isoform 2
XP_001498065.1 [Equus caballus]. 494 aa chromosome 3	PREDICTED: UDP-glucose 6-dehydrogenase isoform 1
XP_003341477.1 [Monodelphis domestica] 494 aa chromosome 5	PREDICTED: UDP-glucose 6-dehydrogenase isoform 2
XP_001365740.2 [Monodelphis domestica]. 427 aa chromosome 5	PREDICTED: UDP-glucose 6-dehydrogenase isoform 1

XP_001142520.1 [Pan troglodytes]. 494 aa chromosome 4	PREDICTED: UDP-glucose 6-dehydrogenase isoform 8
NP_505730.1 [Caenorhabditis elegans]. 481 aa chromosome 5	SQuashed Vulva family member (sqv-4)
NP_476980.1 [Drosophila melanogaster]. 476 aa chromosome 3L	sugarless

APPENDIX C:

VISUAL PROTEIN TREES,

IN ALPHABETICAL ORDER

5e+001

AAL11911.1 Homo sapiens

ABM85298.1 Drosophila

NP_660161.1 Homo sapiens

ABM82115.1 Drosophila

XP_521492.2 PRED Pan troglodytes

AAL11912.1 Mus musculus

NP_058560.1 Mus musculus

NP_001163953.1 Rattus norvegicus

XP_546132.2 PRED Canis familiaris

XP_590240.2 PRED Bos taurus

XP_002698904.1 PRED Bos taurus

DAA14284.1 Bos taurus

NP_989999.1 Gallus gallus

NP_001079289.1 Xenopus laevis

XP_002936902.1 Xenopus tropicalis

CAD52125.1 Danio rerio

NP_571707.1 Danio rerio

NP_508410.2 Caenorhabditis elegans

NP_508725.1 Caenorhabditis elegans

ATOH7 TREE

1e+002

CAA77176.1 Homo sapiens

EAW64993.1 Homo sapiens

EAW64992.1 Homo sapiens

NP_003449.2 Homo sapiens

XP_516463.3 PRED Pan troglodytes

XP_541885.2 PRED Canis familiaris

XP_601010.4 PRED Bos taurus

XP_002697122.1 PRED Bos taurus

XP_003361436.2 PRED Sus scrofa

XP_001494301.2 PRED Equus caballus

CCD63778.1 Caenorhabditis elegans

CCD63779.1 Caenorhabditis elegans

CAA76598.1 Mus musculus

EDL21270.1 Mus musculus

NP_031593.2 Mus musculus

NP_062019.2 Rattus norvegicus

XP_001378098.2 PRED Monodelphis domestica

XP_002936530.1 PRED Xenopus tropicalis

XP_003642021.1 PRED Gallus gallus

XP_001920788.3 PRED Danio rerio

NP_001096896.1 Drosophila melanogaster

NP_001188549.1 Drosophila melanogaster

BSN TREE

OPN1SW TREE

PAX6 TREE

1e+002

NP_000530.1 Homo sapiens

XP_516740.2 PRED Pan troglodytes

XP_001490351.1 PRED Equus caballus

XP_001366225.1 PRED Monodelphis domestica

NP_524368.3 Drosophila melanogaster

AAN13666.2 Drosophila melanogaster

NP_001014890.1 Bos taurus

DAA16827.1 Bos taurus

NP_001008277.1Canis familiaris

NP_663358.1 Mus musculus

AAA63392.1 Mus musculus

NP_254276.1 Rattus norvegicus

NP_001070163.1 Taeniopygia guttata

NP_001025777.1 Gallus gallus

NP_001090803.1 Xenopus tropicalis

NP_001080517.1 Xenopus laevis

NP_571159.1 Danio rerio

CAX13341.1 Danio rerio

NP_999386.1 Sus scrofa

NP_502959.1 Caenorhabditis elegans

RHO TREE

2e+001

XP_508100.3 PRED Pan troglodytes

NP_073713.2 Homo sapiens

EDM11743.1 Rattus norvegicus

NP_445787.1 Rattus norvegicus

NP_001164215.1 Mus musculus

NP_783643.1 Bos taurus

DAA14642.1 Bos taurus

CAI16101.1 Homo sapiens

NP_001320.1 Homo sapiens

XP_003433684.1 PRED Canis familiaris

XP_003363448.1 PRED Equus caballus

NP_001016866.1 Xenopus tropicalis

NP_034110.1 Mus musculus

XP_001363827.2 PRED Monodelphis domestica

XP_421817.3 PRED Gallus gallus

XP_002193579.1 PRED Taeniopygia guttata

NP_001015064.1 Danio rerio

NP_731764.1 Drosophila melanogaster

NP_524336.2 Drosophila melanogaster

NP_731762.1 Drosophila melanogaster

NP_731763.1 Drosophila melanogaster

NP 001014617.1 Drosophila melanogaster

NP_508983.2 Caenorhabditis elegans

RIBEYE TREE

1e+002

NP_002912.2 Homo sapiens

EAW80357.1 Homo sapiens

NP_001012738.1 Homo sapiens

NP_001012740.1 Homo sapiens

XP_001154882.1 Pan troglodytes

XP_001154823.1 Pan troglodytes

NP_786969.1 Bos taurus

XP_001927876.2 Sus scrofa

XP_001495838.2 Equus caballus

XP_546190.2 Canis familiaris

XP_864762.1 Canis familiaris

XP_864780.1 Canis familiaris

EDL24895.1 Mus musculus

NP_067315.1 Mus musculus

NP_001100769.1 Rattus norvegicus

XP_002193326.1 Taeniopygia guttata

NP_001026387.1 Gallus gallus

NP_001017877.1 Danio rerio

NP_001086324.1 Xenopus laevis

NP_001016013.1 Xenopus tropicalis

NP_725197.1 Drosophila melanogaster

NP 725196.1 Drosophila melanogaster

NP_725195.1 Drosophila melanogaster

NP_741080.1 Caenorhabditis elegans

RPE TREE

2e+001

AAD11939.1 Homo sapiens

NP_005404.1 Homo sapiens

EAX00267.1Homo sapiens

EAX00268.1 Homo sapiens

XP_525749.3 Pan troglodytes

DAA24767.1 Bos taurus

NP_001180053.1 Bos taurus

XP_538477.4 Canis lupus familiaris

NP_035511.2 Mus musculus

CAA62379.1 Mus musculus

NP_076480.1 Rattus norvegicus

XP_001375682.1 Monodelphis domestica

NP_571437.1 Danio rerio

NP_989695.1 Gallus gallus

XP_002193827.1 Taeniopygia guttata

XP_002932435.1 Xenopus tropicalis

AAF63242.1 Xenopus laevis

NP_001079171.1 Xenopus laevis

NP_524695.2 Drosophila melanogaster

NP_505958.1 Caenorhabditis elegans

NP_504419.1 Caenorhabditis elegans

NP_476733.1 Drosophila melanogaster

SIX3 TREE

2e+001

AAD49844.1 Homo sapiens

NP_031400.2 Homo sapiens

XP_522870.3 Pan troglodytes

NP_035514.1 Mus musculus

EDL36517.1 Mus musculus

NP_001101502.1 Rattus norvegicus

DAA25109.1 Bos taurus

NP_001098463.1 Bos taurus

XP_547840.3 Canis familiaris

XP_003121893.2 Sus scrofa

NP_990325.1 Gallus gallus

XP_002200636.1 Taeniopygia guttata

XP_001369291.1 Monodelphis domestica

NP_001018421.1 Danio rerio

NP_001081933.1 Xenopus laevis

NP_001093696.1 Xenopus tropicalis

NP_524695.2 Drosophila melanogaster

NP 476733.1 Drosophila melanogaster

NP_504420.1 Caenorhabditis elegans

SIX6 TREE

APPENDIX D:

AUDITORY PROTEIN TREES,

IN ALPHABETICAL ORDER

5e+001

NP_001165404.1 Homo sapiens

NP_001165405.1 Homo sapiens

NP_001165402.1 Homo sapiens

NP_071407 Homo sapiens

XP_507839 Pan troglodytes

AAG52817.1 Mus musculus

NP_075859 Mus musculus

NP_001239564.1 Mus musculus

AAK07670.1 Mus musculus

NP_446096 Rattus norvegicus

DAA14294.1 Bos taurus

XP_001252945 Bos taurus

XP_001925718.2 Sus scrofa

XP_003434519.1 Canis familiaris

XP_003434520.1 Canis familiaris

XP_001917733.2 Equus caballus

XP_001365044.2 Monodelphis domestica

XP_421595 PRED Gallus gallus

XP_002939565.1 Xenopus tropicalis

NP_999974 Danio rerio

NP_648973.1 Drosophila melanogaster

NP_497340.2 Caenorhabditis elegans

CDH23 TREE

5e+001

NP_742055 Homo sapiens

NP_000494 Homo sapiens

NP_742056 Homo sapiens

NP_742057 Homo sapiens

XP_001164492.1 Pan troglodytes

XP_003311799.1 Pan troglodytes

XP_001164379.1 Pan troglodytes

XP_002729501.1 Rattus norvegicus

XP_578437 Rattus norvegicus

NP_034294 Mus musculus

AAH66860.2 Mus musculus

AAB48017.1 Mus musculus

XP_003640064.1 Canis lupus familiaris

XP_859603.1 Canis lupus familiaris

XP_001492875.1 Equus caballus

XP_001492823.1 Equus caballus

DAA22681.1 Bos taurus

NP_723188.1 Drosophila melanogaster

NP_523492.1 Drosophila melanogaster

XP_418290.3 Gallus gallus

XP_002199188.1 Taeniopygia guttata

NP_001083888 Xenopus laevis

NP_571268 Danio rerio

NP_001021056.1 Caenorhabditis elegans

NP_001021055.1 Caenorhabditis elegans

EYA1 TREE

1e+002

CAI95692.1 Homo sapiens

NP_001193625.1 Homo sapiens

AEL89176.1 Homo sapiens

XP_001168751.1 Pan troglodytes

XP_514168.2 Pan troglodytes

NP_006138 Homo sapiens

XP_001490816 Equus caballus

XP_537138.3 Canis familiaris

DAA20980.1 Bos taurus

NP_001070402 Bos taurus

NP_999443 Sus scrofa

NP_058547.2 Mus musculus

NP_001102329.1 Rattus norvegicus

XP_001366308.1 Monodelphis domestica

NP 650273.2 Drosophila melanogaster

ABB77237.1 Gallus gallus

XP_417990.3 Gallus gallus

XP_002196388 Taeniopygia guttata

NP_001025493.1 Xenopus tropicalis

NP_001081215.1 Xenopus laevis

NP_956892.1 Danio rerio

NP 498670.2 Caenorhabditis elegans

IRF6 TREE

2e+001

NP_055377 Homo sapiens

NP_001186699.1 Homo sapiens

XP_001162799 Pan troglodytes

AF424700_1 Mus musculus

NP_035153 Mus musculus

NP_037241.1 Rattus norvegicus

DAA24647.1 Bos taurus

NP_001192946.1 Bos taurus

XP_003125144.1 Sus scrofa

XP_852530 Canis familiaris

XP_001917473.1 Equus caballus

XP_001374919.2 Monodelphis domestica

XP_003640955.1 Gallus gallus

NP_571325 Danio rerio

NP_989216 Xenopus tropicalis

NP_001081009 Xenopus laevis

NP_001014727.2 Drosophila melanogaster

NP_510366.1 Caenorhabditis elegans

OTX1 TREE

2e+001

NP_852124.1 Homo sapiens

XP_001165390.1 Pan troglodytes

NP_852123.1 Homo sapiens

NP_001120838.1 Homo sapiens

NP_446162.1 Rattus norvegicus

NP_001152992.1 Mus musculus

XP_545664.3 Canis familiaris

XP_001495022.3 Equus caballus

NP_852126.1 Homo sapiens

NP_852125.1 Homo sapiens

XP_003365175.1 Equus caballus

XP_001495229.2 Equus caballus

NP_852122.1 Homo sapiens

NP_032807.3 Mus musculus

XP_003365174.1 Equus caballus

XP_001495210.1 Equus caballus

NP_001193747.1 Bos taurus

XP_001365807.1 Monodelphis domestica

XP_003361395.2 Sus scrofa

XP_002194011.1 Taeniopygia guttata

NP_989600.1 Gallus gallus

NP_001088994.1 Xenopus laevis

NP_001006776.2 Xenopus tropicalis

NP_571352.1 Danio rerio

NP_523862.1 Drosophila melanogaster

NP_523863.1 Drosophila melanogaster

NP 001024570.1 Caenorhabditis elegans

PAX3 TREE

2e+001

NP_005973.1 Homo sapiens

XP_509988.2 Pan troglodytes

XP_001492836.1 Equus caballus

NP_446211.1 Rattus norvegicus

CAA56585.1 Mus musculus

NP_033215.2 Mus musculus

XP_547841.3 Canis familiaris

NP_001186647.1 Sus scrofa

XP_588692.2 Bos taurus

DAA25162.1 Bos taurus

XP_001377489.2 Monodelphis domestica

NP_001038150.1 Gallus gallus

NP_996978.1 Danio rerio

NP_001093693.1 Xenopus tropicalis

NP_001082027.1 Xenopus laevis

NP_476733.1 Drosophila melanogaster

NP_504419.1 Caenorhabditis elegans

SIX1 TREE

2e+001

NP_003161.2 Homo sapiens

EAW51117.1 Homo sapiens

EAW51119.1 Homo sapiens

EAW51118.1 Homo sapiens

XP_001143115.2 Pan troglodytes

XP_001142885.2 Pan troglodytes

XP_003315613.1 Pan troglodytes

XP_537747.2 Canis familiaris

XP_001504206.1 Equus caballus

NP_001180055.1 Bos taurus

DAA19054.1 Bos taurus

XP 003484346.1 Sus scrofa

AAB18950.1 Mus musculus

CAI24323.1 Mus musculus

NP_033323.2 Mus musculus

NP_001178749.1 Rattus norvegicus

XP_001368732.1 Monodelphis domestica

XP_423183.3 Gallus gallus

XP_002199580.1 Taeniopygia guttata

NP_001072665.1 Xenopus tropicalis

NP_660094.1 Danio rerio

AAF46140.1 Drosophila melanogaster

NP_651962.2 Drosophila melanogaster

NP_497969.1 Caenorhabditis elegans

STP6 TREE

1e+001

NP_001171630.1 Homo sapiens

EAW92936.1 Homo sapiens

EAW92938.1 Homo sapiens

NP_003350.1 Homo sapiens

XP_001142520.1 Pan troglodytes

NP_001171629.1 Homo sapiens

EAW92937.1 Homo sapiens

DAA28716.1 Bos taurus

NP_776636.1 Bos taurus

XP_536254.2 Canis familiaris

XP_003434444.1 Canis lupus familiaris

XP_001498065.1 Equus caballus

XP_003364775.1 Equus caballus

AAC36096.1 Mus musculus

EDL37741.1 Mus musculus

NP_033492.1 Mus musculus

EDL90061.1 Rattus norvegicus

NP_112615.1 Rattus norvegicus

XP_001365740.2 Monodelphis domestica

XP_003341477.1 Monodelphis domestica

NP_001013628.1 Xenopus tropicalis

NP_001079465.1 Xenopus laevis

NP_001103872.1 Danio rerio

NP_001012599.1 Gallus gallus

NP_476980.1 Drosophila melanogaster

NP 505730.1 Caenorhabditis elegans
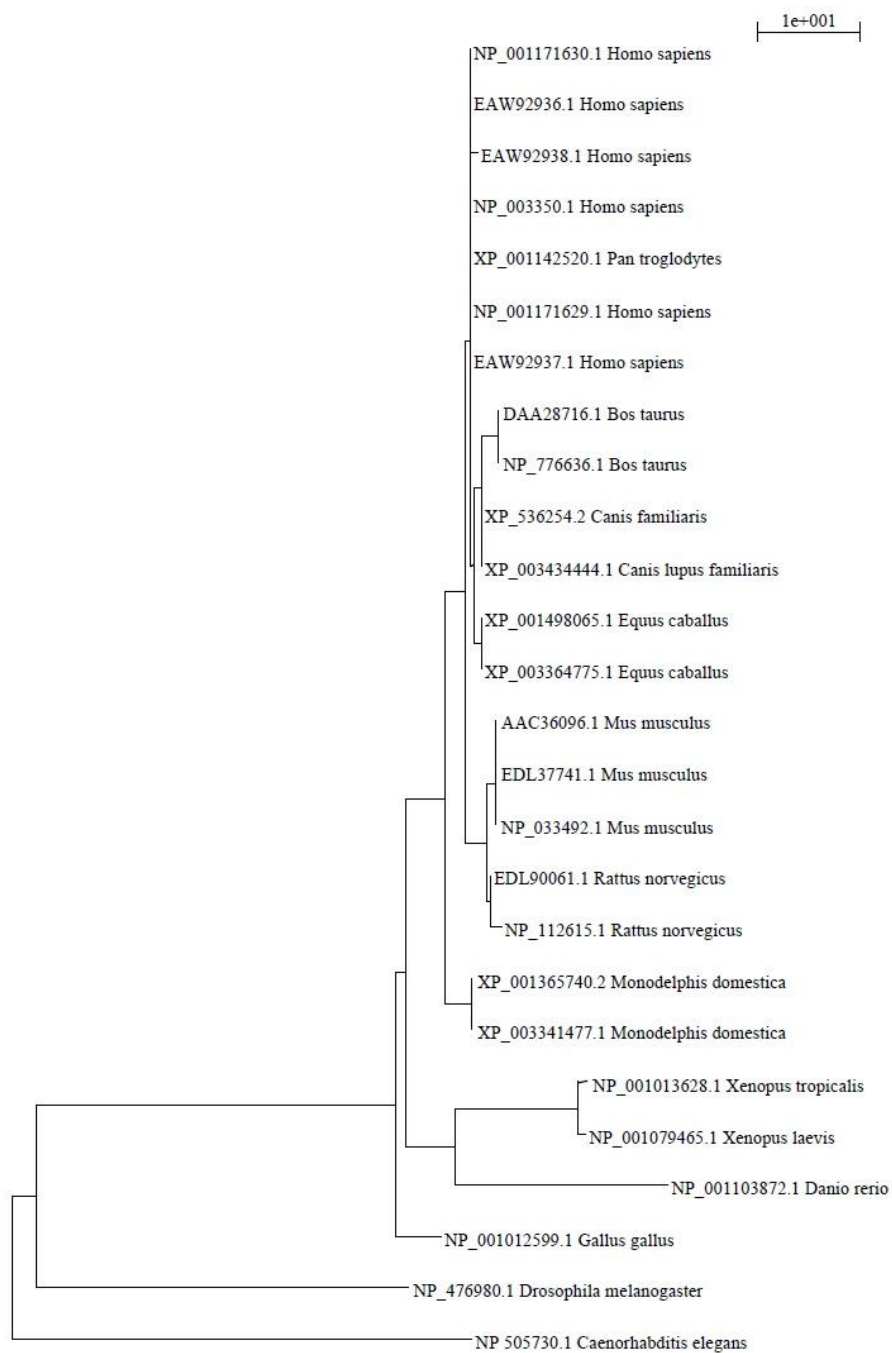
UGDH TREE

REFERENCES

Active Motif, North America. (2008). Nuclear Complex Co-IP Kit *Catalog No. 5400.* (Reprinted).

Alberts, B., Johnson, A., & Lewis, J. (2002). Molecular Biology of the Cell. (4th edition. ed., Vol.). New York: Garland Science. (Reprinted).

Altschuh, D., Lesk, A. M., Bloomer, A. C., & Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. [Journal Article; Research Support, U.S. Gov't, Non-P.H.S.]. *J Mol Biol*, 193(4), 693-707

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. [Journal Article; Research Support, U.S. Gov't, P.H.S.; Review]. *Nucleic Acids Res*, 25(17), 3389-3402

Anderson, A. M., Weasner, B. M., Weasner, B. P., & Kumar, J. P. (2012). Dual transcriptional activities of SIX proteins define their roles in normal and ectopic eye development. [Journal Article; Research Support, N.I.H., Extramural]. *Development*, 139(5), 991-1000. doi: 10.1242/dev.077255

Arnold, K., & Gosling, J. (1996). The Java Programming Language *The Java Series*. MA: Addison-Wesley. (Reprinted).

Bailey, T. L. (2008). Discovering sequence motifs. [Journal Article; Review]. *Methods Mol Biol*, 452, 231-251. doi: 10.1007/978-1-60327-159-2_12

Baldwin, T., & Lapointe, M. (2003). The Chemistry of Amino Acids. (Reprinted).

Berg, J., Tymoczko, J., & Stryer, L. (2006). Biochemistry (6thed., Vol.). New York: W H Freeman. (Reprinted).

Bergsten, J. (2005). A review of long-branch attraction (21, pp. 163-193). (Reprinted).

Bhattacharya, P., & Neamtiu, I. (2011). Assessing Programming Language Impact on Development and Maintenance: A Study on C and C++. Honolulu, Hawaii, USA. (Reprinted).

Bronner-Fraser, M. (1994). Neural crest cell formation and migration in the developing embryo. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.; Review]. FASEB J, 8(10), 699-706

Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O., Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. [Comparative Study; Journal Article]. *Genome Biol*, 5(5), R35. doi: 10.1186/gb-2004-5-5-r35

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. [Comparative Study; Journal Article; Research Support, Non-U.S. Gov't]. *Nat Genet*, 28(3), 281-285. doi: 10.1038/90129

Brown, N. L., Dagenais, S. L., Chen, C. M., & Glaser, T. (2002). Molecular characterization and mapping of ATOH7, a human atonal homolog with a predicted role in retinal ganglion cell development. [Journal Article; Research Support, U.S. Gov't, P.H.S.]. *Mamm Genome*, 13(2), 95-101. doi: 10.1007/s00335-001-2101-3

Burns, M. E., & Arshavsky, V. Y. (2005). Beyond counting photons: trials and trends in vertebrate visual transduction. [Journal Article; Research Support, N.I.H., Extramural; Review]. *Neuron*, 48(3), 387-401. doi: 10.1016/j.neuron.2005.10.014

Busch-Nentwich, E., Sollner, C., Roehl, H., & Nicolson, T. (2004). The deafness gene dfna5 is crucial for ugdh expression and HA production in the developing ear in zebrafish. [Journal Article]. *Development*, 131(4), 943-951. doi: 10.1242/dev.00961

Chatterjee, S., Kraus, P., & Lufkin, T. (2010). A symphony of inner ear developmental control genes. [Journal Article; Review]. *BMC Genet*, 11, 68. doi: 10.1186/1471-2156-11-68.

Chen, G., & Courey, A. J. (2000). Groucho/TLE family proteins and transcriptional repression. [Journal Article; Research Support, U.S. Gov't, P.H.S.; Review]. *Gene*, 249(1-2), 1-16

Chung, H., Suh, E. K., Han, I. O., & Oh, E. S. (2011). Keratinocyte-derived laminin-332 promotes adhesion and migration in melanocytes and melanoma. [Journal Article; Research Support, Non-U.S. Gov't]. *J Biol Chem*, 286(15), 13438-13447. doi: 10.1074/jbc.M110.166751

Cold, Spring, Harbor, Laboratory, & Press. (2008). (Reprinted).

Copper, Cup, & Images. (2012). Melanocytes and Keratynocytes of the inner ear. (Reprinted).

Dandekar, T., Snel, B., Huynen, M., & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. [Journal Article]. *Trends Biochem Sci*, 23(9), 324-328

Dieck tom, S., Sanmarti-Vila, L., Langnaese, K., Richter, K., Kindler, S., Soyke, A.,... Gundelfinger, E. D. (1998). Bassoon, a novel zinc-finger CAG/glutamine-repeat protein selectively localized at the active zone of presynaptic nerve terminals. [Comparative Study; Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.]. J Cell Biol, 142(2), 499-509

Dutton, K., Abbas, L., Spencer, J., Brannon, C., Mowbray, C., Nikaido, M., Whitfield, T. T. (2009). A zebrafish model for Waardenburg syndrome type IV reveals diverse roles for Sox10 in the otic vesicle. [Journal Article; Research Support, Non-U.S. Gov't]. *Dis Model Mech*, 2(1-2), 68-83. doi: 10.1242/dmm.001164

Evolve Media. (2012). Vestibular Senses. (Reprinted).

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) (3.6ed.Vol.). (Reprinted).

Fettiplace, R., & Hackney, C. M. (2006). The sensory and motor roles of auditory hair cells. [Journal Article; Research Support, N.I.H., Extramural; Review]. *Nat Rev Neurosci*, 7(1), 19-29. doi: 10.1038/nrn1828

Fitzpatrick, D. A., Logue, M. E., Stajich, J. E., & Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.]. *BMC Evol Biol*, 6, 99. doi: 10.1186/1471-2148-6-99

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. [Journal Article; Research Support, U.S. Gov't, Non-P.H.S.; Research Support, U.S. Gov't, P.H.S.]. *J Exp Zool B Mol Dev Evol*, 304(1), 64-74. doi: 10.1002/jez.b.21026

Garcia-Molina, H., Ullman, J., & Widom, J. (2008). Database Systems: The Complete Book (2nded., Vol.): Prentice Hall. (Reprinted)

Gilbert, S. (2000). Developmental Biology: Development of the Vertebrate Eye. (6th edition. ed., Vol.). Sunderland (MA): Sinauer Associates. (Reprinted).

Gillespie, P. G., & Muller, U. (2009). Mechanotransduction by hair cells: models, molecules, and mechanisms. [Journal Article; Review]. *Cell*, 139(1), 33-44. doi: 10.1016/j.cell.2009.09.010

Giorgini, F., & Muchowski, P. J. (2005). Connecting the dots in Huntington's disease with protein interaction networks. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.; Review]. *Genome Biol*, 6(3), 210. doi: 10.1186/gb-2005-6-3-210

Giribet, G., Edgecombe, G. D., & Wheeler, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. [Journal Article; Research Support, U.S. Gov't, Non-P.H.S.]. *Nature*, 413(6852), 157-161. doi: 10.1038/35093097

Halder, G., Callaerts, P., & Gehring, W. J. (1995). New perspectives on eye evolution. [Journal Article; Research Support, Non-U.S. Gov't; Review]. *Curr Opin Genet Dev*, 5(5), 602-609

Hall, B. (2011). Phylogenetic Trees Made Easy: A How-To Manual (4th ed., Vol.). (Reprinted).

Hargrave, P. A., & McDowell, J. H. (1992). Rhodopsin and phototransduction: a model system for G protein-linked receptors. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.; Review]. *FASEB J*, 6(6), 2323-2331

Harvey, B., & Wright, M. (1999). Simply Scheme: Introducing Computer Science: MIT Press. (Reprinted).

He, X., & Zhang, J. (2006). Why do hubs tend to be essential in protein networks? [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't]. *PLoS Genet*, 2(6), e88. doi: 10.1371/journal.pgen.0020088

Heller, S. (1996). Who's Afraid of C++?: Academic Press. (Reprinted).

Hirose, S. (2012). Inferring Protein-Protein Interactions (PPIs) Based on Computational Methods, Protein-Protein Interactions - Computational and Experimental Tools. In W. C. A. H. Editors (Ed.): InTech. (Reprinted).

Hsu, W. T., Pang, C. N., Sheetal, J., & Wilkins, M. R. (2007). Protein-protein interactions and disease: use of *S. cerevisiae* as a model system. [Journal Article]. *Biochim Biophys Acta*, 1774(7), 838-847. doi: 10.1016/j.bbapap.2007.04.014

Jothi, R., & Przytycka, T. (2008). Computational approaches to predict protein–protein and domain–domain interactions. In Z. A. E. Mandoiu I (Ed.) *Bioinformatics algorithms: Techniques and application.* : Wiley Book Series on Bioinformatics. (Reprinted).

Keegan, B. R., Feldman, J. L., Lee, D. H., Koos, D. S., Ho, R. K., Stainier, D. Y., Yelon, D. (2002). The elongation factors Pandora/Spt6 and Foggy/Spt5 promote transcription in the zebrafish embryo. [Journal Article; Research Support, Non-U.S. Gov't]. *Development*, 129(7), 1623-1632

Khan, K., Logan, C. V., McKibbin, M., Sheridan, E., Elcioglu, N. H., Yenice, O., Ali, M. (2012). Next generation sequencing identifies mutations in Atonal homolog 7 (ATOH7) in families with global eye developmental defects. [Journal Article; Research Support, Non-U.S. Gov't]. *Hum Mol Genet*, 21(4), 776-783. doi: 10.1093/hmg/ddr509

Kim, I., Liu, Y., & Zhao, H. (2007). Bayesian methods for predicting interacting protein pairs using domain information. [Journal Article; Research Support, U.S. Gov't, Non-P.H.S.]. *Biometrics*, 63(3), 824-833. doi: 10.1111/j.1541-0420.2007.00755.x

Lenski, R. E. (2001a). Evolutionary Rate (10, pp. 1-2): Academic Press. (Reprinted).

Li, Z., & Sheng, M. (2003). Some assembly required: the development of neuronal synapses. [Journal Article; Review]. *Nat Rev Mol Cell Biol*, 4(11), 833-841. doi: 10.1038/nrm1242

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. [Journal Article]. *Mol Biol Evol*, 19(1), 1-7

Lovell, S. C., & Robertson, D. L. (2010). An integrated view of molecular coevolution in protein-protein interactions. [Journal Article; Research Support, Non-U.S. Gov't; Review]. *Mol Biol Evol*, 27(11), 2567-2575. doi: 10.1093/molbev/msq144

Maddison, W. P. (1997). Gene Trees in Species Trees (46, pp.). (Reprinted).

Maeda, A., Maeda, T., Golczak, M., Chou, S., Desai, A., Hoppel, C. L., Palczewski, K. (2009). Involvement of all-trans-retinal in acute light-induced retinopathy of mice. [Journal Article; Research Support, N.I.H., Extramural]. *J Biol Chem*, 284(22), 15173-15183. doi: 10.1074/jbc.M900322200

Makino, T., & Gojobori, T. (2007). Evolution of protein-protein interaction network. [Journal Article; Research Support, Non-U.S. Gov't; Review]. *Genome Dyn*, 3, 13-29. doi: 10.1159/000107601

Malacinski, G. M. (2005). Essentials of Molecular Biology (4ed., Vol.). New York: Jones & Bartlett Learning. (Reprinted).

Marquardt, T., Ashery-Padan, R., Andrejewski, N., Scardigli, R., Guillemot, F., Gruss, P. (2001). Pax6 is required for the multipotent state of retinal progenitor cells. [Journal Article; Research Support, Non-U.S. Gov't]. *Cell*, 105(1), 43-55

Matthews, G., & Fuchs, P. (2010). The diverse roles of ribbon synapses in sensory neurotransmission. [Journal Article; Research Support, N.I.H., Extramural; Review]. *Nat Rev Neurosci*, 11(12), 812-822. doi: 10.1038/nrn2924

Meyer, A., Todt, C., Mikkelsen, N. T., & Lieb, B. (2010). Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity. [Journal Article; Research Support, Non-U.S. Gov't]. *BMC Evol Biol*, 10, 70. doi: 10.1186/1471-2148-10-70

Morhardt, D. R., Guido, W., & Chen, C. K. (2009). The role of Gbeta5 in vision. [Journal Article; Review]. *Prog Mol Biol Transl Sci*, 86, 229-248. doi: 10.1016/S1877-1173(09)86008-0

Muller, U. (2008). Cadherins and mechanotransduction by hair cells. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Review]. *Curr Opin Cell Biol*, 20(5), 557-566. doi: 10.1016/j.ceb.2008.06.004

Nakhleh, L. (2010). Phylogenetics: Recovering Evolutionary History. (Reprinted).

Nancy, V., Callebaut, I., El, M. A., & de Gunzburg, J. (2002). The delta subunit of retinal rod cGMP phosphodiesterase regulates the membrane association of Ras and Rap GTPases. [Journal Article; Research Support, Non-U.S. Gov't]. *J Biol Chem*, 277(17), 15076-15084. doi: 10.1074/jbc.M109983200

Naylor, G. J., & Brown, W. M. (1998). Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. [Comparative Study; Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.]. *Syst Biol*, 47(1), 61-76

Pagel, M., Meade, A., & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. [Comparative Study; Journal Article; Research Support, Non-U.S. Gov't]. *Syst Biol*, 53(5), 673-684. doi: 10.1080/10635150490522232

Pazos, F., & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. [Comparative Study; Evaluation Studies; Journal Article]. *Protein Eng*, 14(9), 609-614

Perkel, J. (2004). Validating the Interactome. (Reprinted).

Perriere, G., & Gouy, M. (1996). WWW-query: an on-line retrieval system for biological sequence banks. [Journal Article]. *Biochimie*, 78(5), 364-369

Pisani, D., & Wilkinson, M. (2002). Matrix representation with parsimony, taxonomic congruence, and total evidence. [Comparative Study; Journal Article]. *Syst Biol*, 51(1), 151-155. doi: 10.1080/106351502753475925

Radu, R. A., Hu, J., Peng, J., Bok, D., Mata, N. L., Travis, G. H. (2008). Retinal pigment epithelium-retinal G protein receptor-opsin mediates light-dependent translocation of all-trans-retinyl esters for synthesis of visual chromophore in retinal pigment epithelial cells. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't]. *J Biol Chem*, 283(28), 19730-19738. doi: 10.1074/jbc.M801288200

Restivo, G., Nguyen, B. C., Dziunycz, P., Ristorcelli, E., Ryan, R. J., Ozuysal, O. Y., Dotto, G. P. (2011). IRF6 is a mediator of Notch pro-differentiation and tumour suppressive function in keratinocytes. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't]. *EMBO J*, 30(22), 4571-4585. doi: 10.1038/emboj.2011.325

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.]. *Nature*, 425(6960), 798-804. doi: 10.1038/nature02053

Saitou, N. (1988). Property and efficiency of the maximum likelihood method for molecular phylogeny. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, Non-P.H.S.; Research Support, U.S. Gov't, P.H.S.]. *J Mol Evol*, 27(3), 261-273

Sanderson, M. J., Driskell, A. C., Ree, R. H., Eulenstein, O., & Langley, S. (2003). Obtaining maximal concatenated phylogenetic data sets from large sequence databases. [Journal Article; Research Support, U.S. Gov't, Non-P.H.S.]. *Mol Biol Evol*, 20(7), 1036-1042. doi: 10.1093/molbev/msg115

Sanderson, M. J., & Driskell, A. C. (2003). The challenge of constructing large phylogenetic trees. [Journal Article; Review]. *Trends Plant Sci*, 8(8), 374-379. doi: 10.1016/S1360-1385(03)00165-1

Schlosser, G., Awtry, T., Brugmann, S. A., Jensen, E. D., Neilson, K., Ruan, G., Moody, S. A. (2008). Eya1 and Six1 promote neurogenesis in the cranial placodes in a SoxB1-dependent fashion. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't]. *Dev Biol*, 320(1), 199-214. doi: 10.1016/j.ydbio.2008.05.523

Schmitt, S., Aftab, U., Jiang, C., Redenti, S., Klassen, H., Miljan, E.,Young, M. (2009). Molecular characterization of human retinal progenitor cells. [Journal Article; Research Support, Non-U.S. Gov't]. *Invest Ophthalmol Vis Sci*, 50(12), 5901-5908. doi: 10.1167/iovs.08-3067

Schmitz, F., Konigstorfer, A., & Sudhof, T. C. (2000). RIBEYE, a component of synaptic ribbons: a protein's journey through evolution provides insight into synaptic ribbon function. [Journal Article; Research Support, Non-U.S. Gov't]. *Neuron*, 28(3), 857-872

Schwander, M., Lopes, V., Sczaniecka, A., Gibbs, D., Lillo, C., Delano, D., Muller, U. (2009). A novel allele of myosin VIIa reveals a critical function for the C-terminal FERM domain for melanosome transport in retinal pigment epithelial cells. [Comparative Study; Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't]. *J Neurosci*, 29(50), 15810-15818. doi: 10.1523/JNEUROSCI.4876-09.2009

Shichida, Y., & Matsuyama, T. (2009). Evolution of opsins and phototransduction. [Journal Article; Research Support, Non-U.S. Gov't; Review]. *Philos Trans R Soc Lond B Biol Sci*, 364(1531), 2881-2895. doi: 10.1098/rstb.2009.0051

Simeone, A., Acampora, D., Gulisano, M., Stornaiuolo, A., & Boncinelli, E. (1992). Nested expression domains of four homeobox genes in developing rostral brain. [Comparative Study; Journal Article; Research Support, Non-U.S. Gov't]. *Nature*, 358(6388), 687-690. doi: 10.1038/358687a0

Sollner, C., Rauch, G. J., Siemens, J., Geisler, R., Schuster, S. C., Muller, U., Nicolson, T. (2004). Mutations in cadherin 23 affect tip links in zebrafish sensory hair cells. [Journal Article; Research Support, Non-U.S. Gov't]. *Nature*, 428(6986), 955-959. doi: 10.1038/nature02484

Soria-Carrasco, V., Talavera, G., Igea, J., & Castresana, J. (2007). The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. [Journal Article; Research Support, Non-U.S. Gov't]. *Bioinformatics*, 23(21), 2954-2956. doi: 10.1093/bioinformatics/btm466

Splettstoesser, T. (2012). Tobacco mosaic virus structure. (Reprinted).

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. (2006). BioGRID: a general repository for interaction datasets. [Journal Article; Research Support, Non-U.S. Gov't]. *Nucleic Acids Res*, 34(Database issue), D535-D539. doi: 10.1093/nar/gkj109

Steel, M. (2001). Some statistical aspects of the maximum parsimony method. In R. G. G. W. DeSalle (Ed.) *Molecular Systematics and Evolution: Theory and Practice* (125-140). Boston: Birkhauser. (Reprinted).

Sul, S.J., & Williams, T. L. (2011). Big cat phylogenies, consensus trees, and computational thinking. [Journal Article; Research Support, U.S. Gov't, Non-P.H.S.]. *J Comput Biol*, 18(7), 895-906. doi: 10.1089/cmb.2010.0199

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. [Journal Article; Research Support, Non-U.S. Gov't]. *Nucleic Acids Res*, 39(Database issue), D561-D568. doi: 10.1093/nar/gkq973

Thermo Scientific Pierce (2010). Protein Interaction Technical Handbook (2ed., Vol.). (Reprinted).

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. [Comparative Study; Journal Article]. *Nucleic Acids Res*, 22(22), 4673-4680

Tillier, E. R., & Charlebois, R. L. (2009). The human protein coevolution network. [Journal Article; Research Support, Non-U.S. Gov't; Validation Studies]. *Genome Res*, 19(10), 1861-1871. doi: 10.1101/gr.092452.109

Underwood, T. J., Amin, J., Lillycrop, K. A., & Blaydes, J. P. (2007). Dissection of the functional interaction between p53 and the embryonic proto-oncoprotein PAX3. [Journal Article; Research Support, Non-U.S. Gov't]. *FEBS Lett*, 581(30), 5831-5835. doi: 10.1016/j.febslet.2007.11.056

Vollrath, M. A., Kwan, K. Y., & Corey, D. P. (2007). The micromachinery of mechanotransduction in hair cells. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Review]. *Annu Rev Neurosci*, 30, 339-365. doi: 10.1146/annurev.neuro.29.051605.112917

Walker, M. G., & Blum, R. L. (1985). An introduction to Lisp. (2, pp.): M.D. Computing. (Reprinted).

Wang, G. Z., & Lercher, M. J. (2011). The effects of network neighbours on protein evolution. [Journal Article]. *PLoS One*, 6(4), e18288. doi: 10.1371/journal.pone.0018288

Weisstein, E. (2012). Wolfram Math World. (Reprinted).

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., & Koonin, E. V. (2002). Genome trees and the tree of life. [Comparative Study; Journal Article; Review]. *Trends Genet*, 18(9), 472-479

Wu, L., Pan, L., Wei, Z., & Zhang, M. (2011). Structure of MyTH4-FERM domains in myosin VIIa tail bound to cargo. [Journal Article; Research Support, Non-U.S. Gov't]. *Science*, 331(6018), 757-760. doi: 10.1126/science.1198848

Yang, J., Wang, L., Song, H., & Sokolov, M. (2012). Current understanding of usher syndrome type II. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't]. *Front Biosci*, 17, 1165-1183

Zeably Gallery Images (2012). Leucine Zipper. (Reprinted).

Zenisek, D., Horst, N. K., Merrifield, C., Sterling, P., & Matthews, G. (2004). Visualizing synaptic ribbons in the living cell. [Journal Article; Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.]. *J Neurosci*, 24(44), 9752-9759. doi: 10.1523/JNEUROSCI.2886-04.2004

Zhang, H., Chen, H., Luo, H., An, J., Sun, L., Mei, L., Feng, Y. (2012). Functional analysis of Waardenburg syndrome-associated PAX3 and SOX10 mutations: report of a dominant-negative SOX10 mutation in Waardenburg syndrome type II. [Case Reports; Journal Article; Research Support, Non-U.S. Gov't]. *Hum Genet*, 131(3), 491-503. doi: 10.1007/s00439-011-1098-2

Zheng, W., Huang, L., Wei, Z. B., Silvius, D., Tang, B., Xu, P. X. (2003). The role of Six1 in mammalian auditory system development. [Journal Article; Research Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.]. *Development*, 130(17), 3989-4000

Zhu, C. C., Dyer, M. A., Uchikawa, M., Kondoh, H., Lagutin, O. V., Oliver, G. (2002). Six3-mediated auto repression and eye development requires its interaction with members of the Groucho-related family of co-repressors. [Journal Article; Research

Support, Non-U.S. Gov't; Research Support, U.S. Gov't, P.H.S.]. *Development*, 129(12), 2835-2849

VITA

Monica Micek earned a Bachelor of Science in Biology and Psychology from Loyola University Chicago. After her undergraduate studies, Monica worked as a research assistant at Parmly Hearing Institute and performed psycho-acoustical experiments on *Carassius auratus*, *Opsanus tau* and *Dipturus innominatus*. After a year of work she decided to take a directed reading class at the graduate level with Dr. Fay to learn more about the animals she worked with. Her curiosity in emerging sciences attracted her to a class in Bioinformatics. There she met Dr. Putonti in 2008 who has since become her advisor. With a strong biological background, Monica has worked as a Physiological Biology technician for Loyola University Chicago as well as in Marine Biological Laboratory in Woods Hole, Massachusetts. Her current research areas include computational biology and microbial evolution. Presently, Monica works in the Castignetti Laboratory at Loyola as a Microbiology research assistant.

THESIS APPROVAL SHEET

The thesis submitted by Monica Micek has been read and approved by the following committee:

Catherine Putonti, Ph.D., Director
Professor of Biology and Computer Science
Loyola University Chicago

Terry Grande, Ph.D.
Professor of Biology
Loyola University Chicago

Eric Schroeter, Ph.D.
Professor of Biology
Loyola University Chicago

The final copies have been examined by the director of the thesis and the signature that appears below verifies the fact that any necessary changes have been incorporated, and that the thesis is now given final approval by the committee with reference to content and form.

The thesis is therefore accepted in partial fulfillment of the requirements for the degree of Masters of Science.

10/16/2012
Date

C. Putti
Director's Signature