Environmentally Influenced Duplication Patterns Followed by Functional Shifts Fueling the Evolution of Metazoan Sensory Systems

A thesis submitted to the National University of Ireland for the Degree of
Doctor of Philosophy

Presented by:
Sinéad C. Hamilton
Department of Biology,
NUI Maynooth,
Co. Kildare, Ireland.

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

October 2012

Supervisor: Professor James O. McInerney B.Sc., Ph.D. (Galway)
Head of Dept.: Professor Paul Moynagh, BA(mod), PhD (Dublin)

Table of Contents

Abstract

In this thesis, some of the methods by which animals use their sensory systems to interact with their environment have been extensively studied. How gene duplications have played an important role in sensory evolution by duplication followed by functional shifts resulting in neofunctionalisation has been analysed. This extensive neofunctionalisation allows for an expansion in the number of environmental signals the animal can detect. In the following chapters, some of the ways gene duplication has effected sensory perception have been shown in detail, in particular by the expansion and specialisation of sensory receptor repertoires. Chapter two describes an extensive study performed on the duplication and neofunctionalisation of opsins in animals as a result of environmental signals, leading to the evolution of colour vision. This study of vision is expanded upon in chapter three by looking at how the duplication of an entire visual pathway has led to the emergence of a new cell type and visual function in the rod and cone cells of vertebrates. Finally, in chapter four, large-scale analyses were performed of some massively expanded gene families used for olfactory and gustatory discrimination, showing the effects of extreme cases of gene duplication on animal sensory perception.

Acknowledgements

Firstly, I'd like to thank my supervisor, Prof. James McInerney for all his help and insight and for giving me the opportunity to work in his lab. I'd also like to thank the Irish Recearch Council (IRC) for my funding, without which this work would not be possible. A special thanks must go to my collaborators, Dr. Davide Pisani and Dr. Roberto Feuda. It was a pleasure to work with you! I have to mention all of my IT support over the years and everyone who took care of Darwin, especially Brian and Misha for fixing all of the mysterious IT problems.

I'd like to give my appreciation and support to the girls who started and finished with me. Without them completing this work would have been a lot less enjoyable. I would also like to thank everyone (past and present) in the Bioinformatics and Molecular Evolution Unit for making my time in the lab so memorable. I've definitely made some great friends for life! I don't think lunch times will ever again be so much fun. Good luck to you all in the future!

It would be impossible to list everyone but I have to thank all my friends and family for their support, especially my parents who always encouraged me through the good times and the not so good.

I have to thank my housemates over the years for all your support and for making the past four years some of the best times of my life. In particular I'd like to thank Leanne for always being there for tea and chats.

I'd like to thanks everyone who supported me over the years. This experience was one of the toughest of my life but also the most rewarding. I've made so many new friends for life and I'll never forget my time in the lab. Thanks!!! ☺

List of Figures

List of Tables

List of Equations

Abbreviations Used

16S RNA – Ribosomal RNA Small Subunit

AIC - Akaike Information Criterion

AU – Approximately Unbiased

BAC – Bacterial Artificial Chromosome

BI – Bayesian Inference

BIC – Bayesian Information Criterion

BLAST – Basic Local Alignment Search Tool

C-opsin – Ciliary Type Opsin

C20 – CAT Mixture Model

cAMP – Cyclic Adenosine Monophosphate

CAT – Mixture Model

CED – Certain Evolutionary Distance

cGMP – Cyclic Guanosine Monophosphate

CIR – Cox, Ingersoll, Ross

CNG-channel – Cyclic Nucleotide Gated Ion Channel

CR – Chemosensory Receptors

DAG - diacylglycerol

DNA – Deoxyribonucleic Acid

ETE – Environment for Tree Exploration

GDP – Guanosine Diphosphate

Go-opsin – Go-binding Type Opsin

GPCR – G-Protein Coupled Receptor

Gt - Transducins

GTP – Guanosine Triphosphate

GTR – General Time Reversible

HCN - Hyperpolarisation-activated cyclic nucleotide-gated Channel

HKY - Hasegawa, Kishino and Yano

I – Inflation Value

Indels – Insertion or Deletion

IP3 - Inositol trisphosphate

JC69 – Jukes and Cantor Model 2969

JTT – Jones, Taylor and Thornton

K2P – Kimura 2 Parameter

K80 – Kimura 1980

KH – Kishino and Hasegawa

LBA – Long Branch Attraction

LE – Log Expectation Score

LG – Le and Gascuel

Ln - Lognormal

LRT – Likelihood Ratio Test

LWS – Longwave Sensitive

MCL – Markov Cluster Algorithm

MCMC – Markov Chain Monte Carlo

ML – Maximum Likelihood

MLT – Melatonin Receptor

MRCA - Most Recent Common Ancestor

MSA – Multiple Sequence Alignment

MUSCLE - multiple sequence comparison by log-expectation

MWS – Mediumwave sensitive

NCBI – National Centre for Biotechnology Information

NNI – Nearest Neighbour Interchange

OR – Olfactory Receptor

PAM – Point Accepted Mutation

PAX6 - Paired Box Transcription Factor 6

PCR – Polymerase Chain Reaction

PDE - Phosphodiesterase

PhyML – Software to Estimate Large Phylogenies by Maximum Likelihood

PIP2 - Phosphatidylinositol 4,5-bisphosphate

R-opsin – Rhabdomeric Type Opsin

RAxML – Randomised Axelerated Maximum Likelihood

RELL – Resampling Estimated LogLikelihoods

RH1 - Rhodopsin

RH2 - Rhodopsin like Opsin

RH7 – Unknown Function R-Opsin

RNA – Ribonucleic Acid

RNase – RNA Digesting Enzyme

SH – Shimodaira and Hasegawa

SP – Sum of Pairs Score

SPR – Subtree pruning and Regrafting

SWS1 – Shortwave Sensitive 1

SWS2 – Shortwave Sensitive 2

T1R – Taste Receptor Type 1 (Sweet/Umami)

T2R – Taste Receptor Type 2 (Bitter)

TIGER – Tree Independent Generation of Evolutionary Rates

TRP – Transient Receptor Potential Ion Channels

TRPC – Canonical Transient Receptor Potential Channel

TRPL – TRP-like Channel

TRPV – Transient Receptor Potential Vanilloid

Ugam – Uncorrelated Gamma

UPGMA - Unweighted Pair Group Method with Arithmetic Mean

UV – Ultra Violet

V1R – Vomeronasal Receptor Type 2

V2R – Vomeronasal Receptor Type 2

WAG – Whelan and Goldman

WGS – Whole Genome Sequencing

Chapter 1 – General Introduction

In his book, On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life (Darwin 1859), Charles Darwin had but one simple diagram to explain his theory of decent with modification, a tree (see Figure 1.1). He speculated that all life on Earth evolved in a tree like manner, with one ancestral species dividing into two new, non-interbreeding populations. We now know that because of horizontal gene transfer (Doolittle 1999) this tree-like speciation process can be considered an inappropriate method of describing many of the major clades of life (Dagan and Martin 2006), such as bacteria. However, animal life still largely holds to this method of branching speciation (Mayr 1992).

In this thesis, the main focus is on the evolution and adaptation of animal sensory systems to a constantly changing environment. These studies are performed primarily using phylogenetic tree based approaches and molecular dating techniques based on the fossil record. The layout of this thesis is as follows. There will be a general introduction where some of the background information relating to the analyses will be described as well as some of the main techniques used. Then there are three results chapters, each of which has their own more specific introduction, methods and discussion sections. Finally, there will be a section overviewing the results and conclusions found in this thesis.

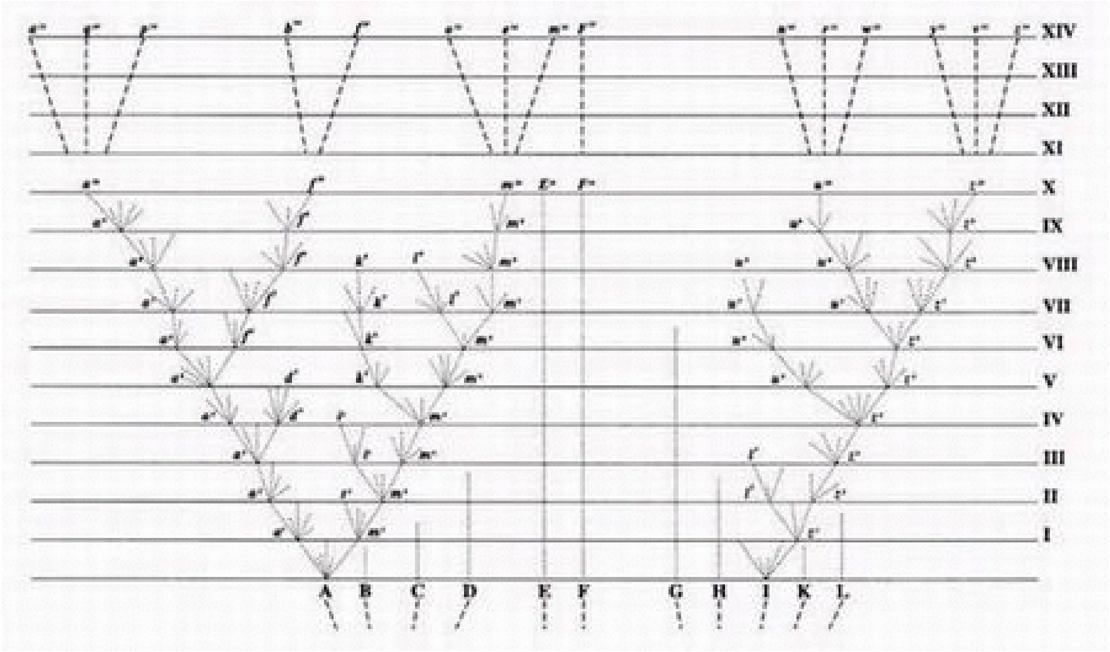


Figure 1.1: Darwin's branching pattern of decent with modification.

Figure 1.1: Darwin's branching pattern of decent with modification. The only diagram Darwin used in his book, on the Origin of Species, depicting a branching pattern of decent with modification (Darwin 1859).

1.1 Metazoan Sensory Perception

The following section will briefly discuss the evolution of the Metazoa, as well as some of the various methods used by the Metazoa to perceive their environment.

1.1.1 Metazoan Phylogenetics

The Metazoa are a diverse group of organisms commonly known as animals that include *Homo sapiens* and our closest relatives (Erwin 1991; Wray *et al*. 1996; Halanych and Passamaneck 2001; Halanych 2004). They are usually large multicellular organisms with various tissue specificities and organs for various functions. They contain multiple different cellular types due a wide variety of expression levels between different tissues. They are opisthokonts and are closely related to another common opisthokont, the Fungi (Medina *et al*. 2003). The Metazoa can be broken down into various groupings. Figure 1.2 shows a phylogenetic reconstruction of the tree topology of the Metazoa, showing each of the major animal groups, adapted from Nielsen (2011). The most basal animals are the sponges, phylum Porifera (Müller 1995). These organisms have no nervous system and very few tissue types but are still classified as animals due to their cell types. The Placozoa are also an early diverging primitive animal with few tissue types. Another group of early diverging animals are the Cnidaria (jellyfish) (Philippe and Telford 2006). These animals have multiple tissue types and tend to be motile throughout most of their life cycle. They have a basic net-like arrangement of nervous tissue. The Ctenophora are similar to the Cnidaria but are phylogenetically a separate phylum.

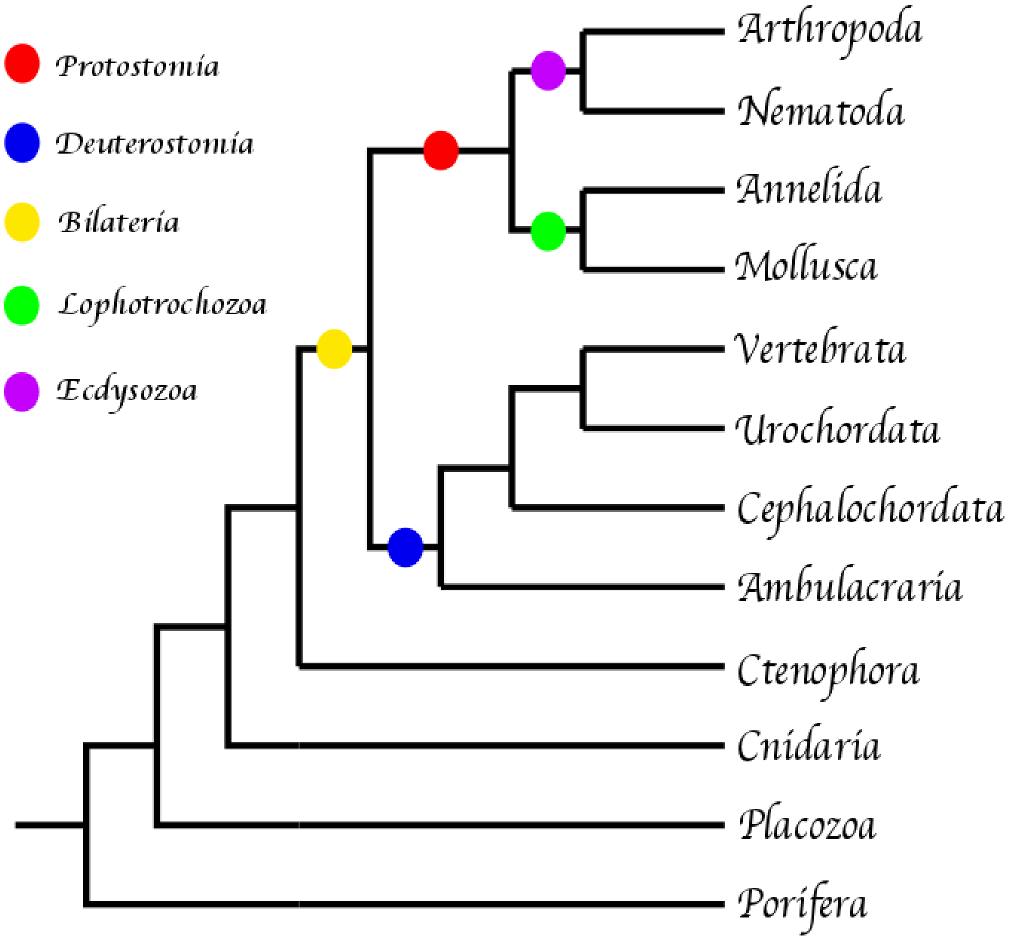


Figure 2

Figure 1.2: The Metazoan phylogeny, adapted from Nielsen (2011). The coloured circles represent the points at which particular groupings arose.

A major grouping of animals that evolved after the divergence of the Cnidaria and the Ctenophora is the Bilateria (Collins 1998). These are animals with bilateral symmetry. The Bilateria make up the majority of animal species. Within this group there are two major divisions, the protostomes and the deuterostomes. The protostomes are classified as organisms that, during gastrulation, developed their mouthparts from the initial invaginations of the blastopore (Mallatt and Winchell 2002). The deuterostomes are classified as the organisms that developed their anus from the initial invaginations of the blastopore (Blair and Hedges 2005).

The protostomes contain the vast majority of animal species within the Bilateria and include the most diverse group of all the animals, the insects. The protostomes can be further subdivided into two major groups (as well as a number of additional smaller phyla), the Lophotrochozoa and the Ecdysozoa. The Lophotrochozoa contain several phyla, the largest and most familiar of which are the Mollusca (molluscs, e.g. snails) and the Annelida (segmented worms, e.g. earth worms). The Ecdysozoa are classified as animals that construct a thick exoskeleton or cuticle that they grow and must shed, as the animal grows larger by the process of ecdysis (Philippe *et al*. 2005). The Ecdysozoa contains many phyla such as the familiar Nematoda (round worms, e.g. *Caenorhabditis elegans*) and the extremely diverse and successful Arthropoda (insects, spiders and crustaceans).

The deuterostomes can be subdivided into the Chordata and the Ambulacraria. The Ambulacraria is a group of organisms that includes the echinoderms (starfish), characterised as being non-chordate deuterostome invertebrates. The chordates are the group of animals that includes humans and other vertebrates, as well as the urochordates (e.g. tunicates) and the cephalochordates (e.g. lancelets).

In this work the focus was mainly on two groups of animals, the Arthropoda and the Vertebrata. These two groups have extremely advanced sensory systems when compared to other members of the Metazoa and often have achieved similar sensory systems in very different ways (Strausfeld and Hildebrand 1999). The early evolution of these two groups can tell us a lot about how animal sensory perception has evolved.

1.1.2 Early Cambrian Animal Evolution

The fossil record before the start of the Cambrian period is relatively scarce (Morris 2000). There are very few fossilised features that can be identified as evidence for metazoan life. Most fossils from this period are remnants of burrows in the soil from burrowing type animals that might be ancestral deuterostomes or protostomes (Knoll 2004). There are also some calcite deposits that are likely biomarkers of the first sponges (Brain *et al*. 2012). Other than these few fossils, little can be found in rocks from this time to suggest that ancestors of modern day metazoans were abundant. However, the molecular data strongly suggests that early metazoan life evolved well before the start of the Cambrian period (Wray *et al*. 1996).

After the start of the Cambrian, metazoan organisms began to flourish (Morris 2000; Jensen 2003; Marshall 2006). During this time, an evolutionary event known as the "Cambrian Explosion" occurred (Morris 1989; Abouheif *et al*. 1998; Conway Morris 2000; Butterfield 2003). This was a short period of around 10 million years where a vast amount of evolutionary divergence occurred. From this event, ancestors of the vast majority of the major animal phyla present today arose. This sudden burst in speciation and divergence may have been due to changing ecosystems around this time and the appearance of predation and competition (Bengtson 2002; Bush *et al*. 2011).

The early Cambrian period marks a very important time in the evolution of animals as they began to flourish and diversify extensively. It was during this early evolution of animals that sensory systems began to develop. The earliest known fossils of eyes date back to the early Cambrian (Lee *et al*. 2011) and many chemosensory systems would have been well established at this point.

1.1.3 Metazoan Sensory Capabilities

Sensory perception describes all the morphological and molecular characteristics that allow an organism to detect its environment. In the Metazoa these sensory mechanisms are quite sophisticated and diverse (Jacobs *et al*. 2007). For example, vision, which will be discussed in detail in section 1.1.4, is divided into the ability to detect different regions or wavelengths of light (Briscoe and Chittka 2001; Yokoyama 2002). The gustatory system is mediated by different taste receptor types that detect different tastants (Ishimaru 2009). GPCRs (G-Protein Coupled Receptors) are used to detect bitter (discussed further in section 1.1.6), sweet and umami (savoury) tastes. Certain ion channels are used to detect salts and acidic substances. The sense of smell (olfaction) is extremely complex, requiring a huge variety of receptors to detect the multitude of potential odorants (Hildebrand and Shepherd 1997). Olfaction is discussed further in section 1.1.5. Touch and hearing all rely on receptors sensitive to movement or pressure (Eberl *et al*. 2000). TRPV (transient receptor potential vanilloid) channels are also present in the skin and allow for the detection of changes in temperature, as well as certain chemicals such as capsaicin (Voets *et al*. 2004). Snakes also use TRPV channels to detect infrared light (Gracheva *et al*. 2010). It is clear that sensory perception is an extremely complex combination of systems that allow for the detection of a wide variety of physical and chemical signals.

Receptors used for arthropod sensory perception can be homologous to the vertebrate receptors, such as opsin visual receptors (Pichaud *et al*. 1999), suggesting an origin that predates the separation of the protostomes and the deuterostomes. Conversely, some receptors that appear to be similar in both arthropods and vertebrates arose independently and converged on these similar functions (Strausfeld and Hildebrand 1999). For example, vertebrate and arthropod olfactory receptors have no significant sequence similarity to suggest recent common ancestry and are therefore more likely to have similar functions as a result of convergent evolution (Strausfeld and Hildebrand 1999).

In this thesis, the focus is mainly on the chemical-binding receptor types, in particular the GPCRs. These are vision (opsin receptors use light sensitive molecules as their ligands), olfaction (airborne and water soluble small molecules are detected via the nasal cavity) and gustation (specifically bitter taste reception, as most of the other taste receptors do not use GPCRs and chemical ligands).

1.1.4 Vision in Vertebrates and Arthropods

Vertebrate and arthropod eyes are morphologically very different. Vertebrates have camera type eyes, usually with large moveable lenses whereas arthropods have compound eyes with multiple small lenses (Miller 1957; Lamb *et al*. 2007) as shown in Figure 1.3. Although, the eye morphology of these two groups is quite different they both express the same developmental protein PAX6 (Gehring 1996). Without this protein, the eye structure in both vertebrates and arthropods fails to develop properly (Mathers *et al*. 1997). The cell types of the vertebrate and arthropod photoreceptors are also very different. Vertebrates primarily use cilary cell types whereas the arthropods tend to use rhabdomeric cell types as their light receptors (Arendt 2003).

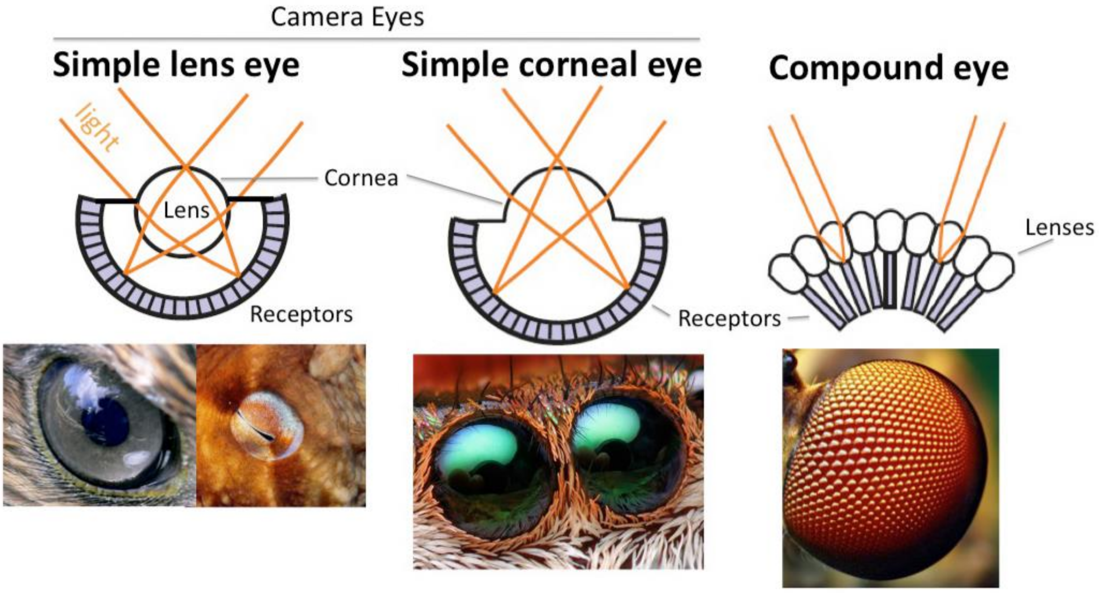


Figure 1.3: Diagram of camera and compound eyes. Camera type eyes are generally found in vertebrates and compound eyes are generally found in arthropods. Depicted showing the cell structure of how light is detected. The simple corneal eyes are a camera type eye found in some vertebrates and also arachnids. The camera eye with a simple lens that focuses the light into a cup of photoreceptors is found in cephalopods as well as vertebrates. Compound eyes, found in arthropods, are made of multiple small light detecting structures called ommatidia. Diagram adapted from Land (2005).

An analysis of the activation pathways of both arthropods and vertebrates shows that they are similar at the beginning of the pathway but the majority of the pathway is quite different (Jindrova 1998; Hardie 2001). Both arthropods and vertebrates use opsin receptors to detect photons of light, but they use different subfamilies (Terakita 2005). Once the opsin reacts to light, the activation signal is passed onto a G-protein, but different subfamilies of G-protein are used in vertebrates and arthropods. After activation of the G-protein, it goes on to activate Phosphodiesterase 6 in vertebrates, which begin hydrolysing cGMP (cyclic guanosine monophosphate). The sudden drop in cellular levels of cGMP results in the closure of cGMP-gated ion channels (CNG-channels), resulting in a hyperpolarisation of the photoreceptor membrane (Figure 1.4). In arthropods, the activated G-protein activates a Phospholipase C that hydrolyses phosphatidyl inositol 4,5-biphosphate (PIP2) to produce soluble inositol 1,4,5-triphosphate (IP3) and diacylglycerol (DAG). The drop in PIP2 levels and the rise in levels of IP3 and DAG cause the activation of TRP (Transient Receptor Potential) and TRPL (TRP-like) channels, resulting in a depolarisation of the photoreceptor membrane (Figure 1.4).

Although both arthropods and vertebrates can detect light and have colour vision, they achieve this in very different ways (Briscoe and Chittka 2001; Jacobs and Rowe 2004). In this thesis, the evolution of colour vision using both arthropod and vertebrate visual opsins is studied in chapter two and the evolution of the vertebrate phototransduction pathway is examined in detail in chapter three.

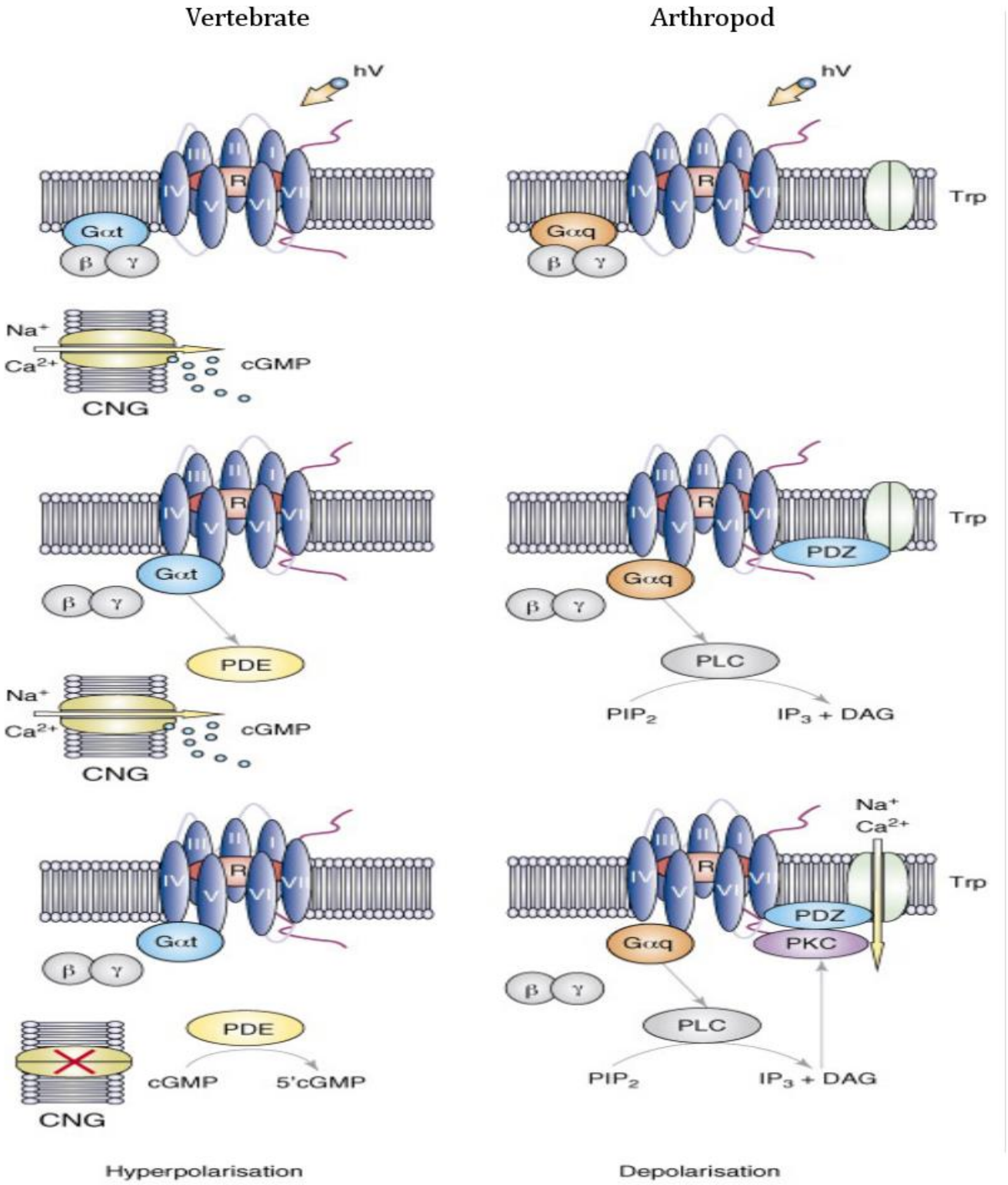


Figure 1.4: Vertebrate and arthropod phototransduction pathways. Light activation of the vertebrate opsin results in a hyperpolarisation of the cell by activating a G-protein that goes on to activate PDE6. This causes a cellular reduction in cGMP levels, closing CNG-channels. Light activation of the arthropod opsins results in a depolarisation of the cell by activating a G-protein that goes on to activate PLC. This converts PIP2 into IP3 and DAG which causes TRP channels to open. Adapted from Hankins (2008).

1.1.5 Vertebrate Olfaction

Olfaction in vertebrates allows for the detection of various airborne and water soluble chemicals (Kauer 1991; Buck 1996). Olfaction functions via the detection of these chemicals by GPCR chemical receptors within the nasal cavity. Within the nasal cavity is the olfactory epithelium. This is a region of dendrites from sensory cells with OR (olfactory receptor) proteins bound to the membrane (Morrison and Costanzo 1992). These cells are covered in a layer of mucus so that when a potential odorant is inhaled, it becomes dissolved in the nasal mucus allowing for binding to the ORs.

The olfactory epithelium can be divided up into two main regions with very different functions. The main olfactory epithelium is where the majority of odorants are detected (usually airborne odorants) and the accessory olfactory epithelium, which is primarily used for the detection of pheromones (vomeronasal receptors), or some water-soluble odorants (Restrepo *et al*. 2004). The receptor proteins primarily used to detect pheromones are not homologous to the other ORs. There are two main types of vomeronasal receptors (pheromone receptors), V1R and V2R (Boschat *et al*. 2002; Yang *et al*. 2005). In this work, the focus was on the main olfactory system that uses ORs. The study looked at receptors used to assist an animal with interacting with its environment as opposed to interactions within the same species.

The ORs are one of the largest families of proteins in vertebrates, often containing over 1000 genes in a single species (Glusman *et al*. 2001; Zhang and Firestein 2002). ORs are necessary for finding food, avoiding predators, navigation, avoiding toxins and hunting (Gittleman 1991; Barton 2006; Dixson *et al*. 2010). Numbers of receptors can vary from species to species, with some fish having tens of receptors, apes having 500-600 receptors and rodents having often over 1000 different receptors.

These ORs can be broad or narrow ranged in their detection of certain odorants. Often they are activated by a common feature of a certain molecule or by multiple small molecules. In other cases, they are quite specific to a particular odorant (Kauer 1991). Our sense of smell is based on the activation of a variety of different receptors.

The olfactory receptor activation pathway, though not as well understood as the phototransduction pathway, is similar in many ways (Lai *et al*. 2005). After the binding of an odorant to one of the ORs (GPCRs) an olfactory specific G-protein binds to the receptor and is activated. The G-protein functions as in vision, where the alpha subunit disassociates from the beta and gamma subunits after GDP is replaced with GTP. This activated alpha subunit then goes on to activate adenylyl cyclase that increases the cellular levels of cAMP (cyclic adenosine monophosphate). As the levels of cAMP increase they activate cyclic nucleotide gated ion channels (CNG), causing an influx of depolarising Na^+ and Ca^+ ions. This OR pathway differs from the visual pathway in that it increases levels of cAMP instead of decreasing them and it results in opening ion channels and a depolarisation signal as opposed to closing channels causing a hyperpolarisation signal.

In this thesis, the study performed in chapter four was on the olfactory receptor proteins. These proteins make up a huge and diverse family, whose genes have duplicated many times, making it an ideal family to study rates and patterns of duplication.

1.1.6 Vertebrate Gustation

Gustation is our sense of taste and is activated by receptor cells located in taste buds on the surface of the tongue (Ganchrow *et al*. 1993; Hara 1994; Finger 1997; Mistretta *et al*. 1999). Gustation can be divided up into five different taste types, sweet, bitter, sour, salty and umami (savoury).

Saltiness is the flavour found by the presence of sodium ions and other ions of the alkali metals. Ion channels that can be activated directly by these molecules function to detect saltiness (Heck *et al*. 1984). Sourness is the taste of acidity or levels of protons. These can also signal sourness via ion channels and possibly can directly activate cells (Huang *et al*. 2006).

The three remaining tastes, sweet, umami and bitter are all activated by the use of GPCRs. These three taste types are made up of two taste receptor families, T1R and T2R. Sweet and umami are detected by dimers of the three subtypes of T1Rs (Zhao *et al*. 2003). Sweet tastes are detected by a combination of T1R1 and T1R3. Whereas, umami tastes are detected by a combination of T1R2 and T1R3. Detection of these flavours is advantageous to determine the nutritional quality of a food source.

Bitter taste reception functions in the opposite way. It is used to determine if a potential food source is toxic. Being able to detect potentially harmful substances in a food source could have implications for the fitness of an animal. Due to the diversity of potential toxic substances, the bitter taste receptor protein family is quite large and diverse (Shi *et al*. 2003). Bitter tasting substances are detected by T2Rs (Chandrashekar *et al*. 2000). Vertebrates can have over 30 different types of T2R receptors that each detects a variety of different potentially toxic substances. Duplication patterns in T2Rs are discussed in detail in chapter four.

Bitter taste signal transduction is not as well-known as other senses but it is believed to be mediated by the Gustducin type G-protein α subunits (Yan *et al*. 2001). The α subunit goes on to activate PDE1A, which affects the cGMP/cAMP levels, as in vision. The β and γ subunits go on to mediate an increase in levels of inositol triphosphate (IP3) and diacylglycerol (DAG) through the activation of a phospholipase C protein (Yan *et al*. 2001).

1.1.7 Gene Duplications and Protein Functional Shifts

In 1936 one of the earliest observations of gene duplication was shown in *Drosophila melanogaster* (Bridges 1936). Gene duplication is prevalent in all domains of life (Zhang 2003). Gene duplication can arise from unequal crossing over or retrotransposition (Zhang 2003).

Unequal crossing over generates tandem duplications where the duplicates are connected in the chromosome. This is a result of errors in chromosomal crossing over during meiosis where the sequences are not paired precisely causing a sequence from one chromatid to be deleted and replaced with a duplicate from the other chromatid.

Retrotransposition is where transposable elements (sequences of DNA capable of moving around the genome) copy a region of DNA to RNA and then reverse transcribe it back to DNA where it is inserted back into the genome at a random location. This copying to RNA removes any introns in the sequence and the resulting duplicate will only have the exon DNA sequences.

It has been known for some time that gene duplication is a powerful force in generating functionally novel proteins (Hughes 1994). Ohno (1973) speculated that after duplication one copy of the gene is redundant and free to accumulate mutations at random. By chance, some of these mutations may alter the function of the resulting protein in some novel way (Zhang 2003). Although, there is some evidence that this is not always the case and that after duplication both copies often remain under selective constraint. Subfunctionalisation is where a bifuctional parent protein duplicates to give two child duplicates that each specialise to do one of the two possible distinct functions of the parent (Hughes 1994).

It has been shown that large amounts of gene duplication allow gene families to rapidly grow and diversify (Chang and Duda Jr 2012). Adaptive evolution after gene duplication can result in novel functions for the duplicates as well as potential new abilities for the organism, e.g. colobine monkeys adapted to a diet of leaves rather than insects after a duplication of an RNase (Zhang *et al*. 2002), known as neofunctionalisation. Another example of neofunctionalisation is the duplication of the LWS opsin in old world primates to allow for trichromatic vision (Yokoyama and Yokoyama 1989). The sensory system has adapted to detect a wide variety of signals from the environment of the organism due to extensive gene duplication. For example, the olfactory receptor gene family is exceptionally large when compared to other gene families. This is due to extensive tandem gene duplication by unequal crossing-over (Ben-Arie *et al*. 1994; Heckel 2010) as the genes are often clustered in large groups along particular chromosomes. The lack of introns in olfactory receptors (and many other GPCRs) suggests a possible early retrotransposition method of duplication.

The olfactory receptor family is a good example of how gene duplication provides raw materials in the form of duplicate genes to expand the gene family functions. As olfactory receptors function to allow the organism to better perceive its environment by the detection of various chemical odours, species-specific duplications are likely due to specialised animal environments and ecological niches. Species-specific duplications are likely to lead to species-specific gene functions and adaptations, as seen in the colobine monkeys (Zhang *et al*. 2002) previously mentioned.

1.2 Aims and Objectives

1.2.1 Overall Aims of this Thesis

The overall aim of this thesis is to gain a greater understanding of how and why sensory systems evolved and to better understand the evolutionary trends that result in duplication followed by functional shifts. In particular, to understand how natural selection as a result of environmental pressures can alter the duplication rates and the functions of certain proteins to increase the fitness of an organism.

1.2.2 Aims of Chapter 2 – Opsin Evolution

The goal of this study was to better understand how and why opsin proteins evolved. There are a number of hypotheses currently available but none give conclusive evidence with statistical significance to support their claims.

In this thesis, evidence that might give a better explanation for the evolution of colour vision in both the vertebrates and the arthropods was analyses by looking at the physical properties of light in the ocean ecosystems of early animals. These oceans would have been very different to the oceans seen today due to low oxygen levels and high toxicity levels as a result of few photosynthesising organisms and large amounts of iron and sulphur based corrosive acids from the surrounding rocks. The aim of this study was to determine if a correlation could be found between the evolution of colour vision, in these two distantly related animal groups, with any other factors influencing the global ocean environment at that time, such as atmospheric or climate changes.

An opsin dataset retrieved from previous work by Feuda *et al.* (2012) was used. In the previous work, the true opsin topology was determined by analysing a series of previously tested datasets that failed to converge on a common agreement for the pattern of opsin duplication (Plachetzki *et al*. 2007; Suga *et al*. 2008; Plachetzki *et al*. 2010; Porter *et al*. 2012) and by the addition of a newly sequenced genome from a homoscleromorph sponge, *Oscarella carmela*. Key taxa were also included from basal metazoan species; the placoazoan, *Trichoplax adherens*; the cnidarians, *Hydra magnipillata* and *Nematostella vectensis* and the demosponge, *Amphimedon queenslandica*.

By increasing the taxon sampling at uncertain regions of the tree, around more basal metazoans, the identity of the previously named group of cnidarian opsins were in fact found to be cnidarian versions of C-, R- and Go-opsins, which are not cnidarian specific. This gave a much more parsimonious explanation for the evolution of opsins. A protein sequence from the placozoan, *Trichoplax adharens* was also found that was shown phylogenetically to be an opsin, although it lacked the retinal binding site that is common to all other opsins. These results showed that opsins arose as a result of a duplication of its common ancestor with the melatonin receptors. The placozoans speciated from the other opsins before duplication, as they have a single family of opsins, described as placopsins. Then, within the common ancestors of the Neuralia (the group composed of the Cnidaria, the Ctenophora and the Bilateria) the opsins duplicated twice, to give the three main opsin sub-families, C-, R- and Go-opsins.

Finally a reduced version of the resulting tree found from extensive phylogenetic testing was used as the input tree for the work described in Chapter two of this thesis. My contribution to the previous phylogenetic analysis of the opsins was to put together the C- and Go-opsin datasets and to assist in the construction of the trees using MrBayes. Refer to the back of the thesis for the previous study (Feuda *et al*. 2012).

1.2.3 Aims of Chapter 3 – Vertebrate Phototransduction

In this chapter, the evolutionary trends that led to the emergence of two independent phototransduction pathways in vertebrates, the rod pathway and the cone pathway were analysed. The cone pathway is the ancestral type; therefore the rod pathway emerged as a result of a series of duplications at each protein along the activation pathway of the cones. It has been speculated that proteins that interact may influence each other's chances of duplicating due to the effects of co-duplication. Evidence for a co-duplication pattern in the emergence of the rod pathway was determined to test if some or all of the duplications were as a result of (1) co-duplication or (2) some other evolutionary factors, causing the proteins to be later co-opted into a new function in the phototransduction pathways.

1.2.4 Aims of Chapter 4 – Olfactory/Gustatory Evolutionary Comparisons

The goal of this section was to analyse the evolutionary trends of two large sensory protein families to see if some patterns of duplication could be detected as a result of niche occupation and other environmental changes. The protein families used were the vertebrate olfactory receptor (ORs) family and the vertebrate bitter taste receptor (T2Rs) family. These families are unique as sensory receptors as they can often have tens or hundreds of family members. This unusually large number of gene duplications followed by functional shifts in these families would have been fueled by specific evolutionary pressures and trends that must be tightly correlated with changing environments and ecological niches of the animals. In this study evidence was analysed to determine if there was a general increase in the number of sensory receptors over time, or if certain animals required specific bursts of duplications in particular types of receptors due to the natural selection of their environment.

1.3 Phylogenetics

This section describes how molecular data can be used to reconstruct the phylogenetic relationships between species, genes or proteins. Modern methods for acquiring datasets and methods for aligning sequences and phylogenetic tree reconstruction will also be described.

1.3.1 Phylogenetic Trees and Data Collection

Phylogenetics is the study of the relatedness of groups of organisms using molecular or morphological data (Nei and Kumar 2000; Zuckerkandl and Pauling 1962). Phylogenetics is used to trace the evolution of organisms, genes or proteins, generally by the construction of a phylogenetic tree (Fitch and Margoliash 1967). Phylogenetic trees are usually bifurcating trees. The leaf nodes correspond to an organism or to a sequence from a gene or protein. The internal nodes correspond to either speciation events or duplications of a gene, protein or common ancestor. The branches and nodes along the internal sections of the tree also represent ancestral sequences or species. The most recent common ancestor of two taxa (taxonomic groups) can be found by finding the node from which both taxa are decended (Figure 1.5).

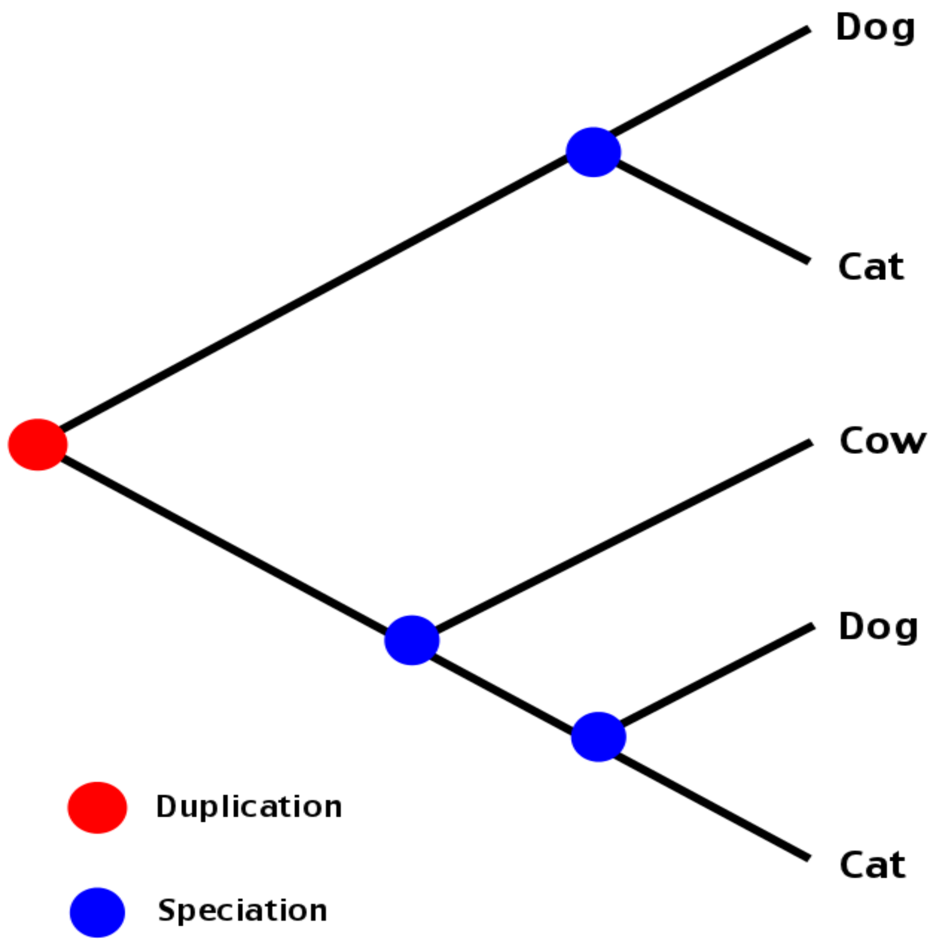


Figure 1.5: A simplified version of a gene tree. Sequences at the leaf nodes are from three species, dog, cat and cow. After the initial duplication (the red node at the base of the tree) one of the copies of the gene was lost in the cow but both the cat and the dog have two copies of the gene. The other three internal nodes in blue are speciation events.

Phylogenetic trees differ from cladograms (Hennig *et al*. 1999) in that the branch lengths can represent the amount of differences between the taxa and the ancestral node, the amount of time that passed or the rate of change in a lineage. Trees are the most commonly used way to represent phylogenetic data as speciation and duplication events are generally considered to be bifurcating processes and the branch lengths represent the amount of evolution that has occurred in the lineage (Stewart 2003).

In phylogenetics, genomic data are most often used to build the trees. These data are obtained by the sequencing of DNA from various animals using various methods. Initial sequencing of the human genome used a method called Sanger sequencing developed by Fredrick Sanger in 1977. Sanger sequencing works on the principle of identifying the bases of a DNA sequence by recording signals emitted during DNA synthesis from a template strand. It took 10 years using this method to produce the first sequence of the human genome. Shortly after this, next generation sequencing (NGS) methods were invented. They work along the same principles of Sanger sequencing but are capable of massive parallelisation of the reactions. This allows for millions of sequences to be identified at one time, rather than a small few. NGS is capable of producing five human genomes in a single week long run.

One of the more popular next generation sequencers used is the Roche/454 FLX Pyrosequencer, which was the first next generation sequencer to become available in 2004 (Buee *et al*. 2009; Hahn *et al*. 2009; Haas *et al*. 2011). A more recent sequencer that is gaining popularity is the Illumina Genome Analyzer

(Kircher *et al*. 2009; Pleasance *et al*. 2009). The Illumina sequencer outcompetes the 454 method for speed due to the lack of the PCR amplification step (Dames *et al*. 2010). The Illumina method uses single molecule amplification, which allows for extremely fast genome sequencing but it is prone to more single base errors than 454 due to mistakes in identification of the base or binding of an incorrect base.

After sequencing of the whole genome, gene and protein sequences must be identified. There are several programs available that can identify certain genes from the chromosome sequences based on certain sequence features such as start and stop codon location and base composition (Martzen *et al*. 1999; Muyzer 1999; Birol *et al*. 2009; Robertson *et al*. 2010).

1.3.2 Using BLAST to Find Homologous Sequences

BLAST (Basic Local Alignment Search Tool) (Altschul *et al*. 1990) is a program that uses the BLAST algorithm to find sequences within a database that have regions of homology with a query sequence. BLAST is a complex program that is constantly being updated (Tatusova and Madden 1999; Korf *et al*. 2003; McGinnis and Madden 2004). BLAST can be used for a wide variety of biological applications such as identification of a sequence, identification of a species, procuring datasets of gene or protein families, searching for specific domains within a sequence, identification of phylogenetic relationships, mapping a DNA sequence to a chromosome location and identification of a gene or protein function (Krauthammer *et al.* 2000; Gough *et al.* 2001; George and Heringa 2002).

There are five main versions of the BLAST program that can be used, BLASTN (nucleotide to nucleotide comparison), BLASTP (protein to protein comparison), TBLASTN (protein to a translated nucleotide comparison), BLASTX (translated nucleotide to protein comparison) and TBLASTX (protein to protein comparison both from translated nucleotides). Selecting which type of BLAST for an analysis as well as which sequence type (nucleotide or protein) is extremely important. For example, protein sequences tend to be more evolutionarily conserved than nucleotides. This is due to synonymous mutations that can occur as a result of the multiple different codons that can be used to code for a single amino acid. Often nucleotide sequences have so many synonymous mutations that the third position of the codon becomes saturated after a relatively short evolutionary distance. Therefore, it becomes very difficult to detect the evolutionary signal among the noise. Conversely, nucleotides can be extremely useful for studying differences between similar sequences. Some synonymous mutations contain detectable evolutionary signal that would otherwise be uninformative identical amino acids when looking at protein sequences.

In this thesis, BLAST is primarily used for acquiring datasets of homologous protein families for analysis. As the majority of the species used in the following analyses are relatively distantly related, only protein sequences are used.

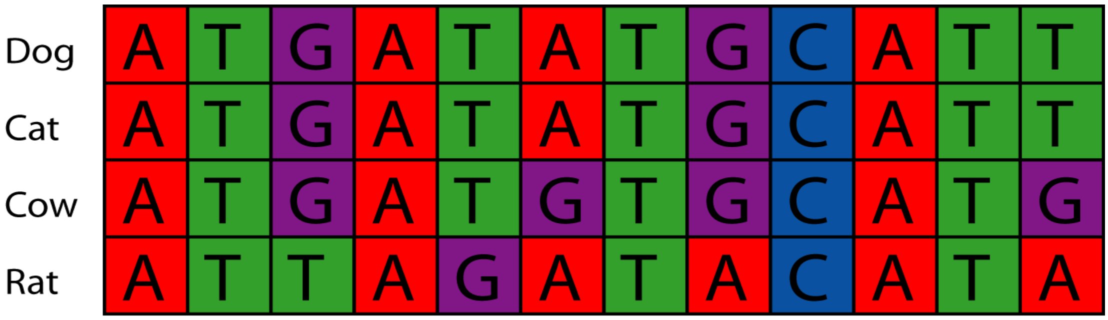
1.3.3 Building Sequence Alignments to Reflect Sequence Evolution

There are multiple types of mutations that can occur in sequence evolution. Random point mutations, such as a C -> A mutation in the codon AGC would change it to AGA, resulting in the amino acid serine being changed to arginine. This could have an effect on a binding site or a folding pattern in the protein that could be neutral, beneficial or detrimental to its function (Chang *et al*. 1990; Robbins *et al*. 1993; Turunen *et al*. 1998). Another type of random mutation is an insertion or deletion of characters in the gene/protein sequence, known as an indel. This means that if a section of the sequence was deleted or a new section added, the protein could gain or lose function (Low *et al*. 1999; De La Chaux *et al*. 2007; Ng *et al*. 2008). If the number of inserted nucleotides is not divisible by three then this could cause a frame shift mutation resulting in the order by which the nucleotide sequence is read (groups of three are a codon, each of which code for a single amino acid) being disrupted and the resulting protein being completely different to the original (Rampino *et al*. 1997; Ogura *et al*. 2001). Frame shift mutations are likely to be highly detrimental to the organism if they occur in essential proteins.

Multiple sequence alignment (MSA) (Bacon and Anderson 1986; Wallace *et al*. 2005; Edgar and Batzoglou 2006) is a tool used for the reconstruction of phylogenetic relationships. MSA is the process of finding regions of common characters between several molecular sequences to identify regions of conserved characters, as well as point mutations and indels. The resulting alignment can then be used to infer phylogenetic relationships by analysing each site in the alignment and how it has evolved or changed. It can also be used to identify critical characters within the sequences (Figure 1.6).

In this work, the alignment software MUSCLE (Edgar 2004) (multiple sequence comparison by log expectation) is primarily used as it is designed for fast analysis of large amounts of sequence data and many of the datasets used in this thesis are quite large. Other commonly used alignment programs are CLUSTAL (Chenna *et al*. 2003), T-COFFEE (Notredame *et al*. 2000) and KALIGN (Lassmann and Sonnhammer 2005). Ideally, the goal of any MSA program is to define a model of sequence evolution and give probabilities of certain sequence modifications (point mutations, indels).

The MUSCLE alignment algorithm can be split into three stages. The initial first stage focuses on speed rather than efficiency to produce a quick guide tree and alignment. First, the kmer distance is computed between each pair of sequences. The kmer distance is a score of similarity between sequences based on the fraction of small sections that the sequences have in common. This score can be computed using unaligned sequences and is, therefore, significantly faster to run than scoring methods that require comparisons of aligned sequences. This similarity information is added to a distance matrix, which is then clustered using UPGMA (Unweighted Pair Group Method with Arithmetic mean) (Sneath and Sokal 1973) to give an initial guide tree. Then, using a progressive alignment method, a MSA is constructed at each node in the tree where the two child nodes are profile aligned to produce a new profile alignment for the parent node.



Dog	A	T	G	A	T	A	T	G	C	A	T	T
Cat	A	T	G	A	T	A	T	G	C	A	T	T
Cow	A	T	G	A	T	G	T	G	C	A	T	G
Rat	A	T	T	A	G	A	T	A	C	A	T	A

Figure 1.6: A sample alignment showing a short nucleotide sequence from four species. Each row is a sequence and each column is an aligned homologous character. Aligning the sequences in this manner identifies the regions where characters are conserved and where characters differ.

In order to align profiles in a pairwise fashion, an alternative scoring function to the kmer distance is used that takes into account the alignment of the sequence profiles, called the log-expectation score (LE). Each node is aligned in this fashion, moving through the tree in a pre-order pattern, which means that each child node is visited before the parent node. As each node is an alignment of its two child nodes, this results in an MSA of the total dataset of sequences being produced at the root node of the tree.

The second stage of alignment, when using MUSCLE, is an improved progressive alignment method. Now that an initial MSA has been produced, the kimura distance (Kimura 1985) can be calculated between each pair of aligned sequences. This similarity information is added to a new distance matrix, which is again, clustered using UPGMA and a second guide tree is created. As before, the progressive alignment method moves through the tree, profile aligning the two child nodes at each node in the tree. Although, in this case, each pairwise calculation is only used on parts of the tree that differ when compared to the previous UPGMA tree that was calculated from the kmer distances. This is to improve on the speed and efficiency of the algorithm.

The third and final part of the MUSCLE algorithm is the refinement stage. An edge is chosen from the tree and deleted. A profile alignment of each of the two new subtrees are calculated and aligned together. If the sum of pairs (SP) score of the new MSA is greater than before then the new alignment is kept. Edges are chosen in order of decreasing distance from the root. This edge selection followed by profile alignment of the new subtrees is repeated to find the most efficient alignment possible, i.e. the alignment that shows the fewest number of differences between sequences. For speed, the calculations can be stopped after any stage, where an MSA is created. Although, stopping the program at an early stage can result in alignments that are not the most efficient as the true alignment has not been found yet.

Alignment errors or errors in the correct identification of homologous characters across multiple sequences and the position of gaps indicating insertions or deletions are the most common types of phylogenetic errors (Venclovas 2003). In particular, the identification of gap regions and their homologous regions is a difficult problem to solve. Often MSA software uses penalties for the opening of gaps to prevent the occurrence of a gap where there was in fact significant divergence between related sequences, although this is not always effective. Regions of an alignment that contain a lot of gaps might on the one hand represent evolutionary history characterised by a lot of length variation, but there is also the possibility that these regions are in fact poorly aligned and the alignment software, in an effort to produce a mathematically optimal alignment, has produced a region that manifests lots of indels.

Certain regions of a gene or protein can be quite variable, to the point where the phylogenetic signal is almost unrecognisable from the noise of random mutations. In this case, there tends to be a bias towards false homology between regions that are unrelated in reality, due to mutations overwriting other mutations to the point that the alignment is essentially random. False homology is where the MSA software detects similarities in sequences and aligns them as if they were homologous regions but they were in fact similar by chance as a result of the random nature of mutations. It is also often a good idea to reduce the alignment down to the most informative, more conserved regions, by use of software such as Gblocks (Talavera and Castresana 2007), or by manually looking at the alignment and deleting the highly variable regions. Often certain parts of a gene or protein are not under much selective constraint resulting in a lack of phylogenetic signal while maintaining high computational requirements. Removal of these regions can often lead to better phylogenetic trees and faster run times (Gatesy *et al*. 2006).

1.3.4 Choosing a Matrix Model

A substitution matrix or model is used in phylogenetic tree reconstruction to describe the process by which a dataset of sequences evolve (Altschul 1991). The matrix shows the likelihood of one character (nucleotide or amino acid) changing from one state to another character state, as well as the likelihood of the character remaining the same. For nucleotide sequences, a 4x4 matrix is used, describing the probability of any nucleotide changing to any other nucleotide. For protein sequences, a 20x20 matrix is used to describe the 20 possible amino acid residues and the probabilities of changing between them. A character frequency vector is also used to describe the frequency at which each character (amino acid or nucleotide) occurs in the dataset.

Some basic models, such as JC69 (Jukes and Cantor 1969) or K80 (Kimura 1980), assume character (in this case nucleotide) frequencies to be equal although this often does not fit the data very well (Keane *et al*. 2004). The Jukes and Cantor model (JC69) assumes equal base frequencies as well as equal rates of change between bases. The Kimura 2-Parameter model (K2P) (Kimura and Ohta 1972) improves on this by allowing different substitution rates for transitions and transversions. The Hasegawa-Kishino-Yano model (HKY) (Hasegawa, Kishino *et al*. 1985) improves upon the K2P model by allowing base compositions to vary. Each of these models adds another parameter that usually increases the fit of the model to the data. Using more parameter-rich models increases the complexity of the calculations. The General Time Reversible model (GTR) (Waddell and Steel 1997) allows base composition and substitution rates to vary but the rate of change from A to B must equal the rate of change from B to A. This means that this model is reversible and can be applied to unrooted trees.

Different regions of a protein are under different selective pressures (Yang and Bielawski 2000; Fares *et al*. 2002). This results in different sites of a sequence (both characters and particular regions) evolving at different rates. To account for this rate variation, a gamma distribution and rate categories are used to allow for a discrete approximation of a continuous distribution of potential rates. The gamma distribution of rates is split into regions or categories, usually four. These four categories correspond to four different regions of the rate distribution, i.e. very fast evolving sites will be given one rate category and very slowly evolving sites will be given another rate category. Therefore, the rate of change of a site is determined by the rate category that it has been assigned.

Most of the models mentioned so far in this section refer to nucleotide sequences. There are many protein sequence models available also. In 1968, Dayhoff first published a matrix called the PAM (probability of an accepted mutation) matrix (Dayhoff *et al*. 1968). Some other matrices were built using empirical data, estimated from a dataset such as the JTT matrix (Jones *et al*. 1992) (which was produced from a dataset of transmembrane proteins). In 2001, Whelan and Goldman expanded on this matrix by applying a likelihood framework, using a dataset of globular proteins, referred to as the WAG model (Whelan and Goldman 2001). These matrices are all based on empirically derived substitution rates and on the principle of General Time Reversibility (GTR).

So far, all the models mentioned assume homogeneity of rates across the sequence although the gamma distribution can be used to apply an approximation of varying rates (Yang 1996). In 2004, Lartillot and Philippe developed the site heterogeneous mixture model, CAT (Lartillot and Philippe 2004). This model splits the sequence up into columns and the substitution rates for each column are calculated separately. The CAT model requires a lot more computational power than previous models such as JTT or WAG but as the CAT model explicitly calculates the rate categories, it can often give rise to trees with higher likelihood values.

Correct model (substitution matrix) selection is hugely important for the reconstruction of a phylogeny (Keane *et al*. 2006). If a model is chosen that does not correctly fit the data then often inaccurate tree topologies or branch lengths can arise (Posada and Crandall 2001).

When selecting a model for use on a dataset, there are a number of tests that can be used to determine the best fitting model. The likelihood ratio test (LRT) calculates maximised log likelihood values for the set of possible models. The tree topology and the branch lengths are estimated from the data. This tree is assumed to be the maximum likelihood tree for every possible model. Then the maximum likelihood is calculated for the given model and the tree. The LRT then compares these maximised log likelihoods of the null and alternative models, rejecting and accepting models until a final model is found that cannot be rejected (Posada and Crandall 2001). The problem with the LRT is that the model with the additional parameters will nearly always be the better fitting model and as the models become progressively complex, the risk of overfitting the model to the data is increased. Overfitting describes when a model is overly complex and is describing random error in a dataset as opposed to the statistical relationship present.

There are a number of model selection tests that take this into account when calculating the best model to use on a particular dataset. The Akaike Information Criterion (AIC) (Posada and Buckley 2004) selects a model based on the best likelihood scores (best-fit), while penalising increased complexity. The AIC attempts to balance out overcomplexity of a model against the goodness of fit of a model. The Bayesian Information Criterion (BIC) (Posada and Buckley 2004) is related to the AIC, but uses a Bayesian formula to select the model with the maximum posterior probability.

In this thesis, the software ModelGenerator (Keane *et al*. 2004) is frequently used to select the best model for the data using the BIC and the AIC. There are drawbacks to model selection as it is possible that the selected model is the best fit of the available models, but still not accurately reflecting what the actual data is doing.

1.3.5 Building the Tree with Maximum Likelihood

There are several different types of tree-building software available that use a variety of models and tree construction methods, as well as tree alteration algorithms such as NNI (nearest neighbor interchange) (Křivánek 1986) or SPR (subtree pruning and regrafting) (Saitou and Imanishi 1989; Hordijk and Gascuel 2005). A common method of phylogenetic tree reconstruction is Maximum Likelihood (Strimmer and Von Haeseler 1996; Yang 1997; Guindon and Gascuel 2003). Maximum Likelihood (ML) was first introduced by Fisher as a mathematical concept (1912; 1921; 1922) but was first applied to the field of phylogenetics by Edwards and Cavalli-Sforza (1964). ML was first applied to molecular data by Neyman (1971) but it was popularised by Joseph Felsenstein (1981) when he showed how to make the ML calculations practical for modern large scale datasets. ML uses a basic statistical approach to determine the most likely tree hypothesis based on random point mutations to the sequence. It is superior to older methods, such as parsimony, as it allows for the possibility of hidden (superimposed) substitutions such as a lysine residue in a sequence changing to a serine residue and back to a lysine along a single branch of the tree. ML is based on calculating a probability (P) for the likelihood of observing the given data (D) (multiple sequence alignment) based on the proposed model (M) (a tree, a composition matrix and a substitution process). It can be explained by the following equation (eq.1).

$$L = P(D|M) \qquad 1$$

ML assumes a model of sequence evolution, given that we know that molecular sequence data tends to evolve in a stochastic manner (Hudson *et al*. 1987). Given that ML can account for superimposed substitutions it can calculate accurate branch lengths as all assumptions are explicitly calculated. All parts of the available data are used, as all sites are informative for ML. When a correctly fitting model is selected, ML can effectively provide the correct tree and also avoid problems with Long Branch Attraction (LBA) (Lewis 1998).

To calculate the likelihood of observing a gene or sequence of nucleotides, the probabilities of observing each character (compositional likelihood) are multiplied together. The likelihood of a tree with one branch, connecting two nucleotide sequences, is calculated as the probability of observing a certain base, multiplied by the probability of observing the transition, taken from the substitution matrix. This is then multiplied by the probability of observing the next character in the sequence, multiplied by the probability of its transition, and so on, until the entire alignment is included in the multiplication calculation.

To calculate the likelihood of a tree, the likelihood of the character composition and the likelihood of a character change for each branch are all calculated. For very short branch lengths the probability of no change in the sequence is high so the diagonal values in the substitution matrix tend to be significantly larger than the values on the off-diagonals. The previous calculations for a one-branch tree are based on one Certain Evolutionary Distance (CED). When considering a branch twice as long, i.e. 2 CED, we multiply the substitution matrix by itself. Repeated multiplication of the matrix by itself results in the values on the diagonal decreasing and the values on the off-diagonals increasing. Meaning that as the branch length increases, the likelihood of a change in a position in the sequence becomes more likely and the probability of no change to the sequence decreases. An accurate branch length can thus be calculated because as the number of CEDs is increased and the likelihood for the branch calculated, the likelihood values will peak at the most likely branch length.

ML is computationally intensive but if a model that closely fits the data is selected, ML can give an accurate tree with minimal errors. ML is vastly better than previously popular methods, such as parsimony (Stewart 2000) or neighbor joining (Saitou and Nei 1987), as it can account for superimposed substitutions, calculate accurate branch lengths and can allow for variation in evolutionary rates across the tree (Tateno *et al*. 1994). The ML theory mentioned here is discussed in the following work (Akaike 1973; Felsenstein 1981).

In this thesis, the software used that implement ML for phylogenetic tree reconstructions are PhyML (Guindon *et al*. 2010) and RAxML (Stamatakis 2006).

1.3.6 Building a Phylogenetic Tree with Bayesian Methods

Bayesian inference uses a posterior probability distribution to determine the most likely phylogenetic tree (Larget and Simon 1999; Huelsenbeck *et al*. 2001). Bayes' Theorem is central to this idea (Eq 2), which calculates the probability of a proposed new tree t_i given the prior probability distribution X, for all possible trees, t_j.

$$P(t_i \,|\, X) = \; (\frac{P(X|t_i)P(t_j)}{\sum_{j=1}^{B(s)} P(X|t_j)P(t_j)} \tag{2}$$

Bayesian inference is based on finding the probability of the hypothesis given the data. This is a reverse probability; unlike likelihood which is a forward probability calculation i.e. the probability of the data given the hypothesis. Bayesian calculations are very computationally intensive due to the denominator that requires the calculation of the probability of all possible trees, making it only practical for very small datasets.

Bayesian inference in phylogenetics uses a Markov Chain Monte Carlo (MCMC) (Metropolis *et al*. 1953; Hastings 1970) method to sample trees from the posterior probability and use these to build a majority consensus rule tree. The MCMC chain is begun with a random tree or an approximated tree (e.g. neighbor-joining tree (Saitou and Nei 1987)). The posterior probability is calculated for this starting tree. The starting tree is used as the current tree in the chain and minor changes to topology or changes to any of the model parameters, are made to produce a new possible tree. The posterior probability is calculated for this new tree and if it is a higher value than the previous tree, it is accepted as the new current tree in the chain. If it is a lower value, the chances of it being accepted depend on the ratio of the probability of the new tree to the current tree. This ratio is compared to a random number within the interval 0 and 1. If the tree ratio is higher than the random number it is accepted, if not it is rejected. Therefore small changes resulting in a tree with a lower posterior probability are somewhat likely to be accepted as the new current tree in the chain, although major changes resulting in a new tree with a large decrease in posterior probability are unlikely to be accepted. Bayesian inference uses this algorithm to decide whether to move to a new location in tree space or not (whether the new tree is accepted or rejected). Allowing minor steps down in probability allows for the crossing of "valleys" between local maxima. This algorithm is known as the Metropolis-Hastings algorithm (Chib and Greenberg 1995). The Metropolis-Hastings Algorithm is as follows:

$$R = \min[\frac{P(t'|X)}{P(t|X)} x \frac{P(t|t')}{P(t'|t)}] \quad 3$$

Here, R is the probability that the chain will move to the newly proposed state. This equation can be rewritten as the following, based on Bayes' Theorem:

$$R = \min[1, \frac{\frac{P(X|t')P(t')}{\sum_{j=1}^{B(s)} P(X|t_j)P(t_j)}}{\frac{P(X|t)P(t)}{\sum_{j=1}^{B(s)} P(X|t_j)P(t_j)}} x \frac{P(t|t')}{P(t'|t)}] \quad 4$$

The two denominators containing the sum equations are the same above and below the line so can therefore be cancelled, resulting in the following equation:

$$R = \min[1, \frac{P(X|t')}{P(X|t)} x \frac{P(t')}{P(t)} x \frac{P(t|t')}{P(t'|t)}]$$

5

It is this cancellation of the summation section of Bayes' Theorem that allows for its use on large datasets. It effectively removes the computationally intensive section of the equation. The three sections of the remaining calculation are the likelihood ratio, the prior ratio and the proposed ratio, respectively. The new tree is added to the chain as the new current tree and the process is repeated. An MCMC chain is said to have converged when the majority of the trees being sampled have similar properties, such as likelihood values, i.e. multiple chains are staying within the same region of tree space. The Bayesian Inference algorithm is effective at avoiding getting stuck in local maxima instead of the global maximum because it sometimes adds trees to the MCMC chain that are less likely than the current tree in the chain, allowing the chain to cross "valleys", shown in Figure 1.7. Multiple chains are often executed on the same dataset to ensure convergence on the global maximum.

In this thesis, Bayesian analyses were carried out using MrBayes (Ronquist and Huelsenbeck 2003) and Phylobayes (Lartillot *et al*. 2009).

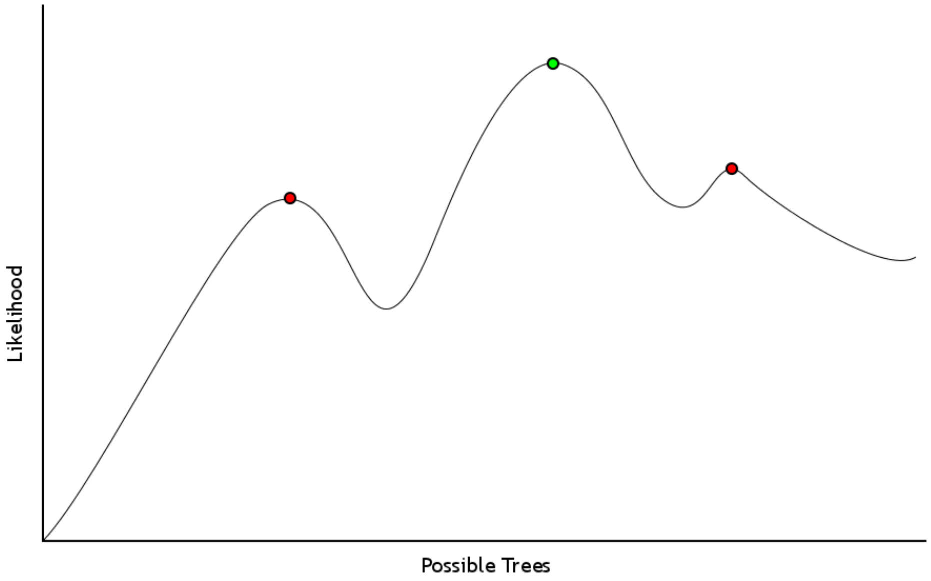


Figure 1.7: Sample graph of possible trees and their likelihood. The red circles represent local maxima whereas the green circle represents the global maximum. ML methods of tree searches might return a local maximum but as Bayesian inference allows for the crossing of the "valleys" between the maxima when searching for the best tree it is more likely to converge on the global maximum.

1.3.7 Testing Tree Topologies

To test the robustness of phylogenetic trees and to compare and contrast them, there are several methods and software that can be used that will be explained in the following section.

Bootstrap analysis of phylogenetic data was first introduced by Felsenstein (1985). It is an invaluable testing method for phylogenetic trees as it uses a non-parametric approach. Bootstrapping involves randomising the sequence data and sampling from it to produce randomly generated sequences of the same length (Figure 1.8). Multiple bootstrap replicates are usually created, and then a majority rule consensus tree is reconstructed. This is done by counting the number of times a particular clade occurs in the dataset of trees. If it occurs a majority of the time, it will be added to the final consensus tree. Each internal node on the consensus tree will be given a value showing how many times it occurred within the dataset of bootstrap replicate trees. It can be said that a node or clade has high support if it was present in 90% or more of the bootstrap replicate trees. Bootstrapping was used to verify the topologies of the majority of the trees produced in this thesis. Two programs were used. The bootstrap method implemented in PhyML (Guindon *et al*. 2009) was used in some cases. Otherwise, the software Seqboot was used to generate the bootstrap replicate alignments and after tree building, the consense module in Phylip was used to build the consensus tree (Retief 2000).

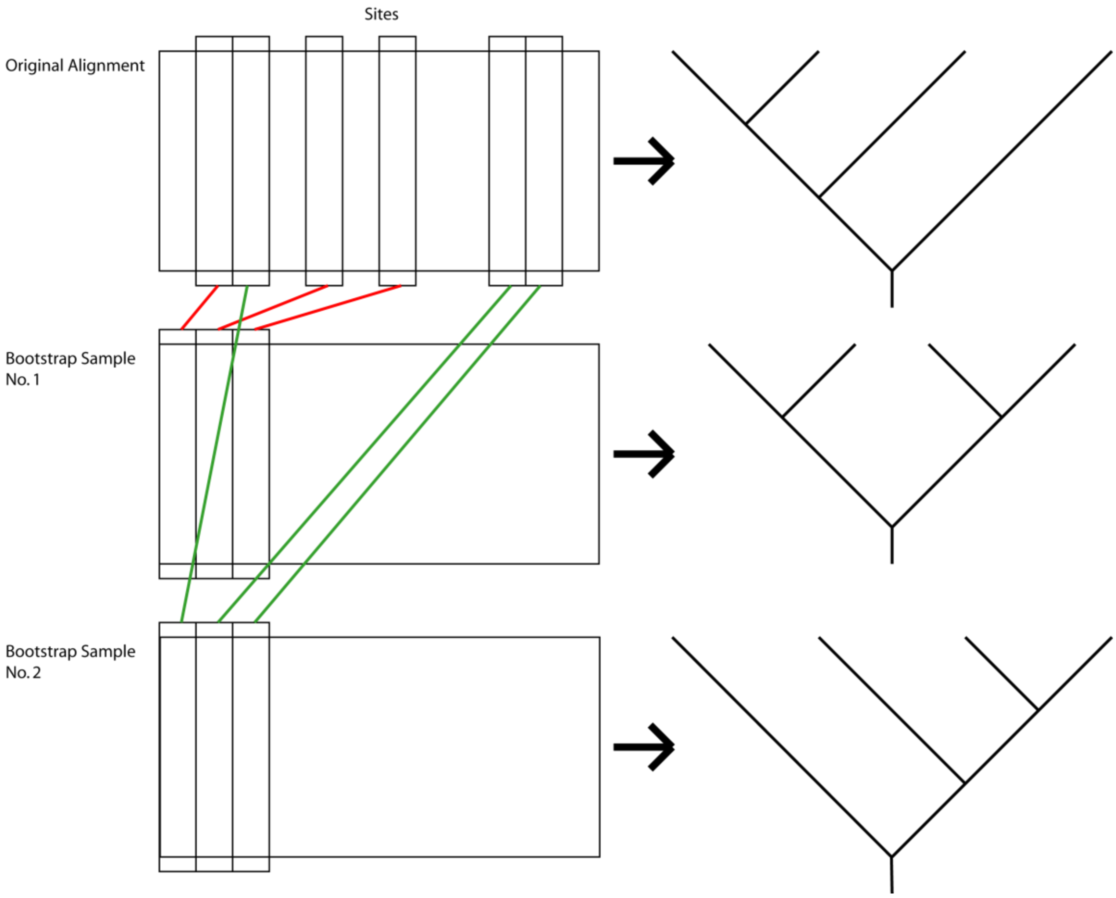


Figure 1.8: Bootstrap replicates generated by randomising the original sequence alignment. A new alignment is produced by using a random selection of sites from the original alighment. Each replicate alignment may then be used to construct a tree that may differ from the original due to different sites being in each replicate alignment. These trees are then used to construct a consensus tree where nodes that are present in a majority of the replicate trees are included. Figure adapted from Felsenstein (2004).

Paired site tests are used in phylogenetics to test if differences in topologies are significant or due to random error. Paired site tests compare parsimony or likelihood scores and calculate significance using p-values (Goldman *et al*. 2000). These tests can be applied to any data to determine significant difference but here they are only discussed in relation to determining the best tree topology using likelihood.

The software Consel (Shimodaira and Hasegawa 2001) is used in this work to calculate paired site tests for protein sequence data. It does these calculations by reading in the site-wise likelihoods that can be calculated using a separate program, e.g. PhyML (Guindon *et al*. 2009). It then generates bootstrap replicates of the log likelihoods using the RELL resampling method. The RELL method (Resampling Estimated Log Likelihoods) (Hasegawa and Kishino 1994) approximates a number of the bootstrap steps to give faster runtimes. RELL assumes the same branch length is obtained for each replicate as found in the original data. RELL keeps track of the log-likelihood values at each site in the alignment (calculated by PhyML) and adds these values together based on the sites that were resampled. The Consel output ranks the tree topologies in order of observed likelihood and show the observed difference in likelihood scores when compared to the best tree. It then gives the results of a number of paired sites tests it has calculated.

Consel uses several tests to determine if two or more tree topologies are significantly different. Examples of some of the tests used are the Kishino-Hasegawa (kh) test (Kishino and Hasegawa 1989) and the Shimodaira-Hasegawa (sh) test (Shimodaira and Hasegawa 1999) and the approximately unbiased (au) test (Shimodaira 2002). The au test is the most reliable and was developed to account for the biases of some of the other tests.

1.3.8 Tree Reconstruction Biases

There are two particularly common phylogenetic errors that can occur as a result of sequence biases, Long Branch Attraction and Compositional Biases.

Long Branch Attraction (LBA) is a phenomenon found in phylogenetic tree reconstruction whereby very long branches on a tree will tend to cluster together even if they are not phylogenetically related (Bergsten 2005). LBA is a feature of model misspecification. Similar molecular sequences will be grouped together to the exclusion of very different ones. If a sequence is quite different it will be given a long branch on the tree to show a large amount of evolutionary distance between that sequence and the rest. If there are several sequences in the dataset that have very low similarity with the majority of the dataset, the tree reconstruction program may detect some similarity between the two extremely different sequences and group them together with long branches separating them (Philippe *et al*. 2005). This is due to rapidly evolving or deeply divergent lineages being more likely to evolve the same character at a given position in an alignment due to chance rather than evolutionary history. This occurrence of homoplasy (convergent, parallel or reversal evolution) may then be mistaken for a synapomorphy (a retained trait occurring in both lineages as well as their common ancestor).

LBA can be overcome by attempting to break up the long branches in the tree by adding additional sequences that are more closely related to the Long Branch sequence (Wiens 2005) or by ensuring that the model used fits the data as closely as possible.

Bayesian Inference (BI) of phylogenetic relationships is more prone to long branch attraction errors than maximum likelihood (ML) methods as it calculates uncertainty about branch lengths by integrating over a distribution of possible values rather than estimating them from the data like ML (Kolaczkowski and Thornton 2009). This results in BI returning a somewhat LBA biased tree, particularly from datasets with more lineage specific variation.

LBA artifacts have plagued metazoan phylogenetics. The removal of LBA artifacts has been shown to confirm the monophyly of the Ecdysozoa, Lophotrochozoa and Protostomia and disprove the monophyly of the Coelomata (Philippe *et al*. 2005).

Often the base composition of a gene or genome is not evenly distributed across all four bases. Some genomes can be more GC rich than others. It has been speculated that this was due to a need for stronger bonds between the chromosome pairs in order to maintain the DNA alpha helix in organisms that live at higher temperatures. This hypothesis stems from the fact that there are three hydrogen bonds between the base pair G and C and only two between A and T. Although there has not been strong evidence to suggest that this is the true mechanism that fuels compositional biases (Hughes *et al*. 1999). Regardless, the base composition in a genome rarely shows equal frequencies of AT and GC pairs.

Compositional bias can incorrectly influence phylogenetic reconstruction by causing taxa with similar base compositions to be grouped together on a tree when they may in fact be phylogenetically quite distant. Therefore, it becomes quite difficult to determine the true genetic distance between two taxa and also the substitution rates are unclear. Some of the effects of compositional bias in phylogenetic reconstruction are discussed by Van Den Bussche (1998).

In order to overcome the problem of compositional bias, current models used in phylogenetic reconstruction also model base composition (such as the HKY model or the GTR model, mentioned in section 1.3.4) allowing for more accurate tree reconstruction.

1.3.9 Graphs, Networks and MCL Clustering

In some cases trees are not sufficient for representing certain types of phylogenetic data. In this case homology networks can be used to graph similarity between sequences (Atkinson *et al*. 2009). This is particularly useful when trying to identify protein families from a large dataset. In a phylogenetic sequence homology network each node represents a sequence and each edge represents a statement of homology between two nodes. An all vs. all BLAST search can be used to find the level of similarity of each sequence to every other sequence in the dataset. This can then be used to find clusters of closely related sequences using a clustering algorithm such as MCL (Markov Cluster Algorithm) (Enright *et al*. 2002).

MCL uses a minimum cut algorithm to identify the edges in a graph that very few random paths traverse. These edges are removed to leave the most connected clusters that have many potential random paths through them. This results in clusters of highly connected nodes of related sequences. These clusters can represent families of related protein sequences (see Figure 1.9).

Clusters in a graph are characterised by many edges between each of the nodes in that cluster, resulting in many different unique paths between two randomly selected nodes. Two random nodes, selected from two different clusters, would be expected to have significantly fewer non-overlapping paths between them compared to two nodes from the same cluster. A random walk on the graph will usually remain within clusters and rarely move between clusters due to there being more edges to choose from within a cluster than there are edges that connect different clusters. Therefore, the probability of choosing an edge that leaves a cluster is low. The probability of moving from one node to another random node is the probability of taking any one of the connecting edges, whose probabilities sum to one.

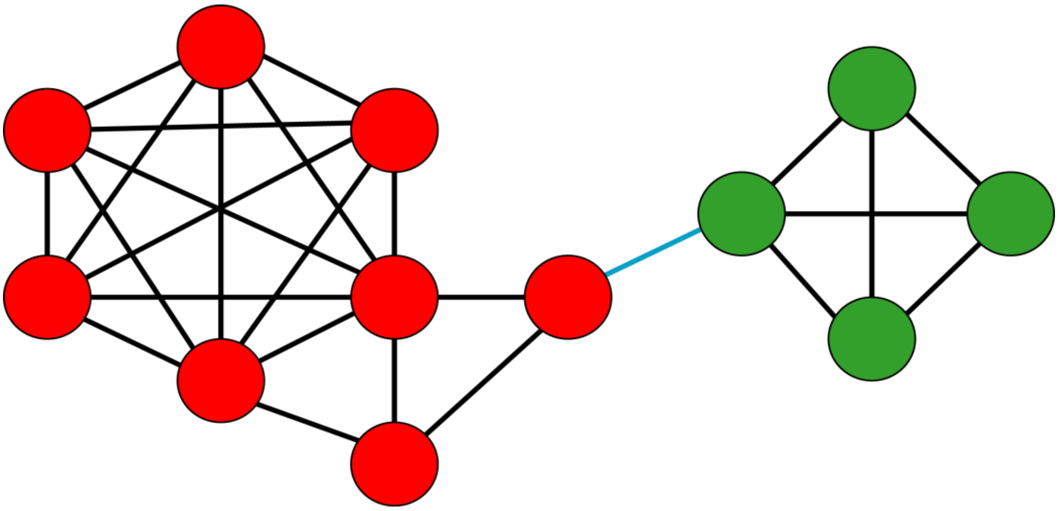


Figure 1.9: A sample graph with two maximally connected regions. The nodes represent sequences (in this thesis they are protein sequences) and the edges represent a statement of homology between them. Here, MCL would remove the blue edge, thereby splitting this graph into two protein families as there are more non-overlapping paths within the clusters than there are between them. Any path moving from a node within the red cluster to a node within the green cluster would have to traverse the blue edge.

The MCL algorithm calculates a column stochastic matrix (a square matrix where all the columns sum to one) based on the BLAST output. In this matrix a column represents all the possible edges from a single node to any other node and the probability of moving to another node. The probability of moving to another node is calculated from the ratio of the BLAST similarity hit compared to the summed BLAST similarity hits for all the edges connected to a node, which therefore sums to one.

MCL simulates random walks through the graph using two steps called expansion and inflation. The expansion step is the result of squaring the matrix. The inflation parameter uses an entrywise based power of a matrix (each entry is multiplied by itself a number of times) followed by a scaling step. In the scaling step each value is divided by the sum of all values in the column to ensure the matrix is stochastic (each column sums to one).

By using an inflation value greater than one, more probable walks will be favoured over less probable walks i.e. edges with higher BLAST similarity scores will be taken over edges with lower scores. The probability of moving from one node to another within the same cluster will generally be higher than the probability of moving between nodes of different clusters given that there are more paths that could be taken between them. Therefore, the inflation parameter has the effect of increasing the likelihood of moving within a cluster and decreasing the likelihood of moving between clusters. Repeated iteration over the expansion and inflation steps results in unlikely edges (edges with a low similarity score between the nodes i.e. low probability of being selected during a random walk) being removed and the graph becoming increasingly more granulated/clustered. Eventually no more changes can be made to the matrix so the final clusters can be considered protein families.

In this thesis, MCL is used on a network of proteins taken from genomes to identify a particular set of protein families. These families could not be identified using any tree methods and the dataset was too large to identify each protein individually.

1.4 Molecular Dating And The Fossil Record

Molecular dating is the analysis of divergence times between particular protein or nucleotide sequences (Rutschmann 2006). Using information based on rates of change across branches and known calibration points taken from the fossil record, approximations can be made on the date at which a certain node in a tree occurred. Duplication or speciation events can both be dated in this manner. Molecular dating methods (Rutschmann 2006) have been used in many different analyses such as timing the early evolution of placental mammals (Eizirik *et al*. 2001) and to accurately date events in early Cambrian animal evolution (Erwin *et al*. 2011).

1.4.1 Molecular Clock Hypothesis

The molecular clock hypothesis was first introduced by Zuckerkandl and Pauling (1962). They hypothesised that all genes are mutating at a constant rate and therefore diverging at a constant rate after duplication or speciation. The molecular clock hypothesis assumes that this global substitution rate, once calculated for a particular gene, remains constant across the entire gene tree. Therefore, the divergence times between organisms, genes or proteins can be extrapolated from this single unchanging rate by determining how many substitutions have occurred since the divergence of two sequences. This implies an ultrametric tree, i.e. a tree where the distance from the root to every leaf node is exactly the same due to the same global evolutionary rate.

Zuckerkandl and Pauling, used several proteins such as the α-globin protein, to calculate the number of substitutions found between selected species sequences when compared to a human sequence. By graphing these numbers against the estimated divergence times between each of the species from humans, based on the fossil record, they found that there was a roughly linear correlation. From this data they inferred the Molecular Clock Hypothesis.

We now know, after further testing on different proteins, that this hypothesis is not correct (Li *et al*. 1987; Howell *et al*. 2004). Although point mutations are random and therefore over large time scales may show a constant rate, their likelihood to be conserved in a sequence is dependent on natural selection and genetic drift (Lande 1976; Burger and Lynch 1995). If a mutation improves the function of the protein it will increase the fitness and chances of survival of the individual organism. This leads to more offspring with increased fitness being present in the population. Overtime and interbreeding, the mutation will likely be conserved in the species. If the mutation is detrimental to the fitness of the organism, it is likely to be lost as the individual organism is less likely to survive and produce offspring (Peck 1994). If the mutation to the protein has no effect on its function, i.e. a neutral mutation, it can be preserved in the individual and over time in the population due to population dynamics.

The effects of natural selection have a huge influence over the evolution of a protein. Often many substitutions, as well as duplications and other sequence modifications, can occur in a very small amount of time (Elena *et al*. 1996). Examples of this can be found in certain sensory proteins, such as olfactory receptors (Glusman *et al*. 2001). In other cases, some sequences can be so critical in their function that they change very little between vastly different organisms, such as the 16S ribosomal RNA small subunit (Hillis and Dixon 1991).

Although the molecular clock hypothesis has been disproved, it can still have some use for molecular dating analyses. Some sequences do evolve in a clock like manner, though not many (Palmeirim *et al*. 1997; Liu *et al*. 2004). Also, modifications have been made to the clock hypothesis to allow for local clock like behaviour on a tree. If an ancestral sequence evolved at a certain rate, it is likely that after a divergence event the rate would not be significantly different (Ihara *et al*. 1999). Although, there are some exceptions (Robinson-Rechavi and Laudet 2001).

More recently, new methods for modeling molecular evolution have been developed, such as relaxed clock models (Lepage *et al*. 2007), which are discussed further in section 1.4.3.

1.4.2 Fossils

Fossil data is invaluable to molecular dating as it can be used to calibrate certain internal nodes (speciation or duplication events) on a tree (Near *et al*. 2005; Yang and Rannala 2006). Given that molecular evolution is not a clock-like process, varying evolutionary rates can be complex, resulting in the sequence data not accurately reflecting the time scale (Howell *et al*. 2004; Hwang and Green 2004). In these cases it is essential to add in several calibration points from the fossil record to force certain nodes to remain within a known time bracket. A fossil can be dated based on stratigraphy, carbon dating (limited to recent fossils) and radiometric dating (Clifford 1968; O'Brien and Lyman 1999).

The known metazoan fossil record spans the extent of the Phanerozoic era (542 million years ago to the modern day), with very few metazoan specimens dating from before the beginning of the Cambrian Period (Morris 2000; Peterson and Butterfield 2005). There is some evidence of tunneling animals in the fossil record prior to the Cambrian period, although very little. With the use of modern computational methods, the accurate dating of events in prehistoric time can be extended into the Proterozoic era. It is likely that metazoans were present before the Cambrian but the conditions for effective fossilisation may not have been present (Lee 1999; Delgado *et al*. 2001; Smith and Peterson 2002).

When a fossil is found that is closely related to the ancestor of some extant taxa for which we have molecular data, we must first determine if the fossil is from a stem group or from the crown group. The stem group refers to a group of organisms that diverged from the lineage leading to the crown group prior to the crown groups' most recent common ancestor (MRCA). For example, the avian-like dinosaur, Archaeopteryx is a member of the stem lineage to the crown group of birds (Lee and Worthy 2011).

When examining a fossil, its characteristics must be determined to see where along the tree it can be placed. This can be difficult as some key traits do not fossilise and the skeleton may be fragmented (Kimbel 1988). Poor quality fossilisation can make it difficult to determine if a trait is absent in the organism, or if it is simply not apparent from the fossil evidence at hand.

If a fossil is found to be a common ancestor between two extant groups (within the crown group), a calibration point can be placed on that speciation node. Calibration points are given an upper and a lower bound that can be soft or hard bounds. Hard bounds are constraints that cannot be broken whereas soft bounds are constraints that can be broken to some extent, allowing the date to be slightly older or younger than the constraint. To calibrate a node (split between two clades), an ancestral fossil of a species that was present after but close to the split between the two groups must be found. The estimated date for when the organism was fossilised can then be used as a hard minimum constraint. This is because it is not possible for the split between the two clades to have occurred more recently (Benton *et al*. 2009). The maximum time that the divergence could have happened is usually a soft bound as our fossil can tell us that the divergence must have happened before a certain date but fossils cannot tell us the actual date with a greater degree of certainty.

The main problem with using fossils is that the fossil record is incomplete (Benton *et al*. 2000; Benton 2009). Charles Darwin discussed this problem in his book *On The Origin Of Species by Means of Natural Selection* (Darwin 1859) where he pointed out that many organisms have soft body parts that are unlikely to fossilise. Another problem that can eliminate fossil evidence is damage to the rock record due to erosion or plate tectonics (Ballais and Cohen 1985). Due to the many difficulties in fossilisation and obtaining large numbers of well-preserved and diverse organisms, the fossil record is both incomplete, patchy and fossils showing intermediate species are very rare.

Accurate fossil calibration points are essential for determining true molecular dates, as the molecular data is frequently misleading. In this work the majority of the calibration points used in the molecular dating analyses were taken from the work of Benton, Calibrating and Constraining Molecular Clocks (2009).

To test the effects of calibration points on a dataset a Jackknife test can be used. Jackknife tests were first introduced by Wu (1986) and Felsenstein (1985). Usually, jackknife tests sample a random 50% (delete half jackknife) or sometimes a random 75% of the data (Xia and Xie 2001; Winstanley *et al*. 2005). The data refers to the calibration points in this case but it could also be characters, similar to bootstrapping. The dating analysis can be repeated multiple times using the reduced randomly selected calibration data and the results can be compared to the original full dataset results to determine the effects of the calibration points.

1.4.3 Molecular Dating Models

As mentioned previously in section 1.4.1, from the early molecular clock hypothesis, other molecular dating models arose, that more closely reflected true evolutionary events. The relaxed clock allows the molecular clock to vary across parts of the tree (Drummond *et al*. 2006; Battistuzzi *et al*. 2010). Implementing a relaxed clock now incorporates rate heterogeneity into the molecular dating inferences (Welch and Bromham 2005; Lepage *et al*. 2007). These dating models allow for wide variation in evolutionary rates between branches of a tree. The evolutionary dating models used in these cases can be uncorrelated or autocorrelated. Autocorrelation is the process by which the evolutionary rate of a lineage after duplication or speciation is prohibited from changing drastically from its parent lineage. The daughter lineage (the new lineage that stems from the parent lineage) inherits its initial evolutionary rate from the parent lineage and can then gradually evolve a new rate. In other words, the evolutionary rates on daughter lineages are always related in some way to their parent lineages. Some molecular dating models that use autocorrelation to calculate rates are Cox-Ingersoll-Ross (*CIR*) and lognormal (*ln*). The CIR model is based on the CIR process (Chou and Lin 2006), which uses a Brownian-like motion with a spring-like component to ensure that the rate process does not drift too far away from the mean value. The lognormal model determines the rate variation from a lognormal distribution. An "autocorrelation parameter" is used that determines how likely a rate is to depart from its ancestral rate.

When using an uncorrelated model, each lineage or branch can have a different evolutionary rate, completely unrelated to any other branch in the tree. A commonly used uncorrelated model is uncorrelated gamma (*ugam*) (Drummond *et al*. 2006). The *ugam* model calculates branch length according to a gamma distribution independently of the rate process.

In this thesis, the program Phylobayes (Lartillot *et al*. 2009) is used in order to execute molecular dating analyses using the models *ln, CIR* and *ugam*.

A correct dating model must accurately reflect the evolutionary process of the gene or protein. Otherwise, the resulting chronogram (dated phylogeny) can give inaccurate dates (Near and Sanderson 2004). Bayes factors are used as a bayesian method for model selection (Kass and Raftery 1995). They can be used for any model selection questions but thus far are only implemented in the use of dating model selection (Berger and Pericchi 1996). The bayes factor test resembles a likelihood ratio test using a bayes factor integral to calculate the posterior probability that one of two given models is correct. It can be described as the ratio of the posterior odds of the model in question to its prior odds, where the odds are equal to the probability divided by one minus the probability. This Bayesian framework integrates over all parameters in each model and includes a penalty for overfitting models. It calculates whether the additional data information increases or decreases the likelihood of model 1 compared to model 2 (Goodman 1999). The bayes factor K is denoted as:

$$K = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(\theta_1|M_1)P(D|\theta_1, M_1)d\theta_1}{\int P(\theta_2|M_2)P(D|\theta_2, M_2)d\theta_2} \tag{6}$$

The probability of the data given the model is called the marginal likelihood (Chib and Jeliazkov 2001). The value θ represents the parameters. If K is greater than 1 then model 1 is more strongly supported by the data than model 2. In this thesis the Phylobayes implementation of bayes factors is used to determine the best fitting dating model.

Chapter 2 – Ocean Drive: Availability of New Ocean Ecosystems Promoted the Evolution of Colour Vision

In this chapter a new hypothesis, devised by Dr. Davide Pisani, called Ocean Drive is discussed. This hypothesis attempts to give a better understanding as to why colour vision evolved by comparing the patterns of opsin duplications to light penetration patterns in water. By comparing the timing of each of these duplications to known global environmental events that would have happened around the same time it can be hypothesised that certain environmental changes have fuelled the evolution of colour vision. The findings of the following analyses suggest that the evolution of colour vision was as a result of early animals exploring deeper ocean ecosystems.

2.1 Introduction

2.1.1 The Importance of Gene Duplication in Animal Evolution

Animals have large numbers of gene families that arose as a result of massive amounts of gene duplication and diversification (Hughes 2002; Zhang 2003). These genes can differ greatly in function and genomic location while still having relatively recent common ancestors (Zhang 2003). Duplication is an extremely important process in evolution, allowing for radically new genomic changes (Hughes 2005) and the genesis of novel proteins that can confer significant changes to the fitness of an organism (Hittinger and Carroll 2007).

An example of a gene family that has experienced large amounts of duplications is the vertebrate olfactory receptor family (Young *et al*. 2002). Duplication of this family has expanded the repertoire of odours that can be detected by those animals where the gene expansions have occurred. These genes have gone through many rounds of duplications, creating new receptors that allow animals to constantly change their olfactory perception in response to external environment changes. Another example of complex gene duplication is to be found in the chemoreceptors of nematodes that mediate chemo-detection (Robertson 1998).

In a different context, one of the major distinctions between vertebrates and invertebrates is that all vertebrates have two or more Hox gene clusters (transcription factors that have been linked to molecular complexity), while all invertebrates have only one. This has been linked to a large-scale duplication event, possibly whole genome duplication, early in vertebrate evolution (Garcia-Fernàndez and Holland 1994). Other large-scale duplication events have also been seen early in vertebrate evolution such as at the origin of the gnathostomes (Holland *et al*. 1994). Gene duplication is discussed in detail by Zhang (2003), Hurles (2004) and Donoghue and Purnell (2005).

2.1.2 The Effects of Global Environmental Changes on Animal Evolution

Global environmental changes, such as temperature fluctuation or variation in available oxygen levels, can have an effect on duplication rates and other genomic alterations (e.g. expression rates, mutation rates, epigenetic changes)

(Møller and Szép 2005). The study of epigenetics has shown how aspects of an organism's genome can be changed quickly in response to environmental changes (Jaenisch and Bird 2003). In addition, it is now possible to measure gene expression responses to climate change in plants (Garrett *et al*. 2006). It has also been shown that woodrats during the Holocene could alter their body size in response to the changing temperatures of their environment (Smith and Betancourt 2003).

It is likely that other aspects of genome evolution, such as mutation rate or duplication rate, can be altered by natural selection in response to environmental changes. However, gene copy-number variation often produces a very complex pattern of paralogy, with duplication and loss events sometimes occurring repeatedly in the same family, making it difficult to understand that family's history. In this chapter, the evolution of a duplicated family of proteins, the opsins, is studied and how their evolution has been influenced by environmental changes is analysed.

2.1.3 Vision and Light Detection

There are approximately 32 recognised animal phyla (Nielsen 2012). Most of these phyla, including representatives of the morphologically simpler ones (e.g. the sponges) have the ability to react to light (Rivera *et al*. 2012). Within the Neuralia (Cnidaria, Ctenophora and Bilateria), which represent the majority of the animal phyla, proteins belonging to the opsin family are universally used to detect light.

Visual capabilities vary greatly within the Neuralia. Some animals, such as the earthworms, have limited visual capabilities, being only able to detect the presence of light and use it for negative phototaxis. Two animal phyla, the Arthropoda and the Vertebrata, have much more well-developed visual capabilities – including image forming and polychromatic, colour vision (Land and Nilsson 2012). In these distantly related lineages, colour vision evolved convergently, through independent processes of opsin gene duplication and gene neofunctionalisation (Briscoe and Chittka 2001; Yokoyama 2002).

It is clear that the ability to detect colours can be of great benefit to an animal (e.g. in the processes of mating, food detection and escaping predators). However, the selective pressures that drove the evolution of colour vision in the Arthropoda and the Vertebrata are still unknown.

2.1.4 Function and Biochemistry of Opsins

Opsins are light detecting G-protein coupled receptors (GPCRs), primarily used for vision in many metazoans and this innovation allows for easier detection of food sources and avoidance of predators (Land and Nilsson 2012). Opsins consist of seven trans-membrane alpha-helix domains and a binding pocket for a light sensitive chromophore (Terakita 2005). The chromophore is a vitamin A derived molecule, usually 11-cis-retinal that upon reaction with light is hydrolysed to all-trans-retinal. The molecule is usually bound to a conserved lysine residue (K) at position 296 on the 7th trans-membrane helix but after hydrolysation, this Schiff base bond is broken and the chromophore reattaches further down the protein (Sugihara *et al*. 2002). This hydrolysation of the chromophore causes a conformational change in the protein, resulting in the outward movement of some of the transmembrane helices, exposing the G-protein binding site on the cellular side of the protein (Dunham and Farrens 1999). Subsequent binding of the G-protein continues the phototransduction cascade and light activation is detected. The phototransduction pathway has been extensively reviewed elsewhere (Hardie 2001; Ridge *et al*. 2003).

There are three types of opsin, R-opsins (found in Rhabdomeric cells), C-opsins (found in ciliary cells) and Go-opsins (bind to Go type G-proteins) (Terakita 2005). Although vertebrates primarily use C-opsins, they do have some R-opsins used for other non-visual functions such as entrainment of circadian rhythms, and the same can be seen in arthropods that primarily use R-opsins but do have some C-opsins for other non-visual functions. The functions of the Go-opsins are primarily unknown but some are known to act as photoisomerases at the back of the eye (McBee *et al*. 2001). They bind to all-trans-retinal, convert it back to 11-cis-retinal and release it for use by the visual opsins. Each of the subtypes of opsin (C-, R- and Go-) have duplicated many times and adapted for various functions, such as the entrainment of circadian rhythms and for image forming vision (Terakita 2005).

2.1.5 Mechanisms of Colour Vision

Colour vision is achieved by the contrasting of signals from the detection of multiple wavelengths of light. Each visual opsin subfamily maximally detects light of a different wavelength (colour). When an organism has only one visual opsin, it is monochromatic, as it cannot compare different signals to distinguish colours. It perceives either presence or absence of light. When organisms have two, three or four different visual opsin subtypes they are dichromatic, trichromatic and tetrachromatic, respectively. This means that they detect multiple light signals and can then contrast these signals to determine the colour being perceived (Nathans 1999). Humans are trichromatic, meaning that they have three colour vision receptors that maximally detect blue, green and red wavelengths of light. Each of these receptors detects a broad region of the light spectrum but they are most sensitive to a small region, their maximum absorbency. The light detecting ranges of these receptors overlap with each other. This allows for the detection of various colours by contrasting signals for each receptor (Figure 2.1).

2.1.6 Evolutionary and Environmental Changes During the Cambrian Period

The environment likely affected the evolution of primitive animals and vision. The early evolution of animals was marked by a period of unusually high rate of organismal evolution and diversification commonly known as the "Cambrian Explosion". This "explosion" of life resulted in the emergence of almost all modern animal phyla.

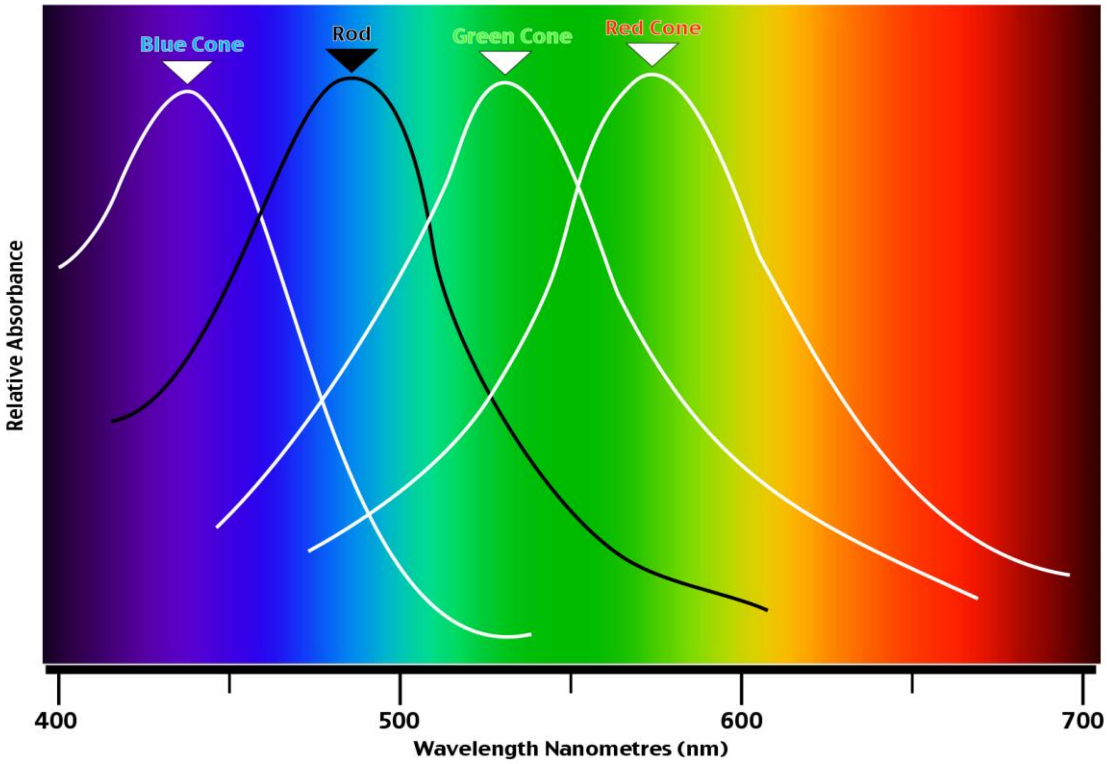


Figure 2.1: Human opsin repertoire and their wavelength absorbancies. Humans have three colour vision cone opsins that maximally detect blue, green and red light, as well as one rod opsin, that is used for dim light vision. Each opsin maximally detects a particular region of the light spectrum. Colours are detected by the contrasting of signals from each overlapping opsin.

Various causes have been proposed to explain the Cambrian Explosion, such as the emergence of vision fuelling a predator-prey arms race, or the increased levels of oxygen at this time, as well as the increase in global temperatures. Some or all of these factors may have allowed the evolution of larger, more complex animals.

By the Cambrian period, global ocean geochemistry had settled to an environment similar to that of the oceans today (Holland 2006). Oxygen levels had risen due to increased photosynthesis by cyanobacteria (Holland 2006) and acidity levels had reduced (Canfield 2005). Prior to the Cambrian, the oceans were very different environments; acidity was high and dissolved oxygen levels were low, except at shallow surface water (Li *et al*. 2010), meaning that primitive animals would have had to stay close to the surface of the oceans in order to survive in this largely acidic, anaerobic ecosystem (Narbonne 2010).

2.1.7 Eyes and Vision in the Fossil Record

Given the widespread occurrence of opsins among the Metazoa (present in both the protostomes and the deuterostomes as well as more basal animals such as the Cnidaria), it can be concluded that vision (or basic light detection) has been present for a very long time, likely predating the Cambrian Explosion and many of the major lineage separations.

Eyes tend not to fossilise well (Kear *et al*. 1995). Consequently, most fossilised eyes are derived from the mineralised eyes of trilobites (Xi-Guang and Clarkson

1990). Exceptions include the fossilised eyes found by Lee (2011) which are well developed arthropod eyes that date back to approximately 515 mya. These fossils show that high-resolution eyes were already present at this point, suggesting that the first appearance of primitive eyes was much older than 515 mya. Owing to the absence of evidence for early eyes in the fossil record, there is a general lack of precision concerning the date of the origin of eyes and vision.

2.1.8 Reasons for the Evolution of Vision

Explanations for the early evolution of colour vision are usually related to predation or foraging, suggesting that colour vision gave the possessor a selective advantage for finding food or avoiding predators (Parker 2004). The "Ocean Drive" hypothesis proposes the possibility that predation and foraging might not be directly responsible and colour vision evolved as a result of other factors.

With regard to foraging abilities, it can be seen from spider monkeys that trichromatic vision does not provide a major advantage over dichromatic vision, at least when light is only filtered by air. If trichromatic vision was selectively advantageous for spider monkeys then it might be expected to have become fixed in the population (Riba-Hernández *et al*. 2004). Both dichromats and trichromats are present in the spider monkey population, suggesting that possessing an additional photoreceptor does not confer a significant selective advantage over those without this additional photoreceptor, if indeed it confers any advantage at all.

In some cases, dichromats show the best foraging rate as they can rely better on brightness levels rather than colour, making them more efficient in certain cases, particularly in lower light conditions (Melin *et al*. 2007). This suggests that foraging abilities or predation do not automatically induce an inevitable evolutionary trend towards colour vision, although their influence cannot be ruled out entirely.

2.1.9 Correlation between Opsin Evolution and Light Penetration in Water

In this study the visual opsins used for colour vision in vertebrates and arthropods were analysed. The patterns of diversification of these receptors were studied to determine how they evolved. The Ocean Drive Hypothesis may shed light on why these duplications occurred in this manner. Certain wavelengths of light can penetrate water to a much greater degree than others (Figure 2.2). For instance, green light can penetrate coastal ocean water up to 50 meters in depth, while violet light generally only penetrates up to 10 metres. These depths are not particularly influenced by ocean acidity or oxygen content, but the maximum depth that certain wavelengths of light can penetrate does vary between coastal water and open ocean due to scattering of light by sediment and dissolved substances and also due to the presence of photosynthetic microorganisms such as phytoplankton. These organisms photosynthesise by absorbing blue and red light but they reflect green light, which has the effect of allowing more green light to penetrate deeper into the water than it would otherwise (http://oceanexplorer.noaa.gov).

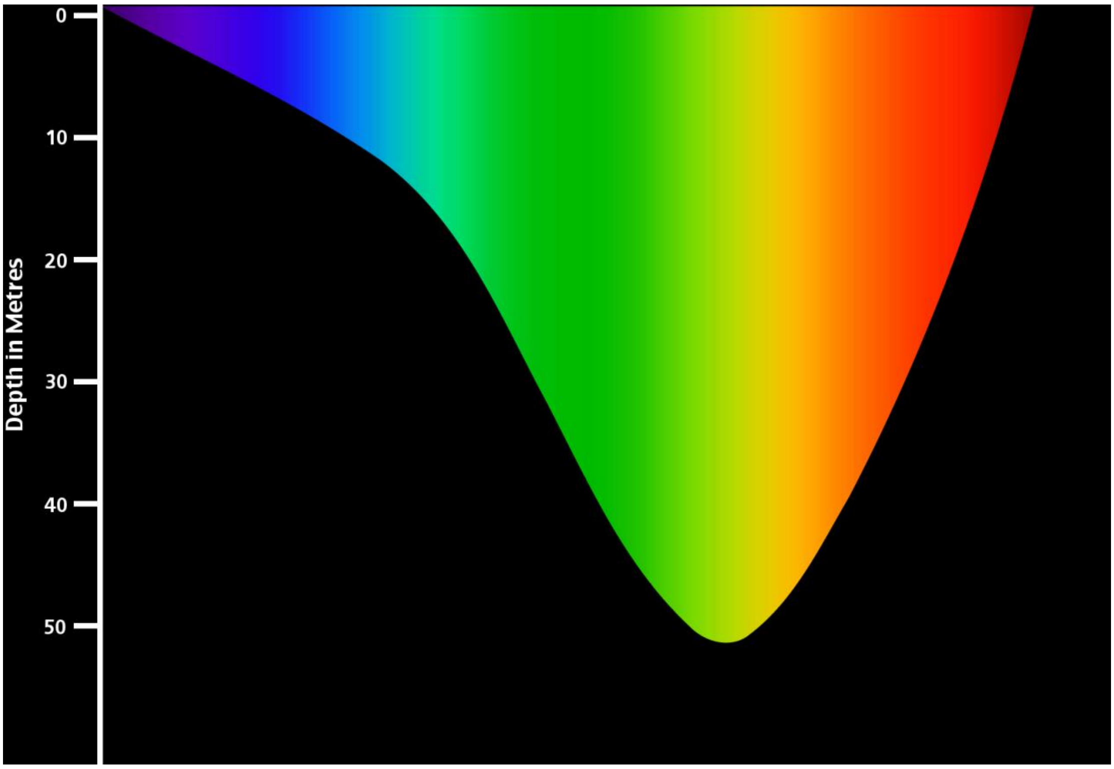


Figure 2.2: Light penetration in coastal water. Different wavelengths of light can penetrate water to different degrees. The scale on the left represents depth below the surface of the ocean in metres. The colours show that in coastal water, green light can penetrate further into the water than other colours.

In this chapter, the order in which each subfamily of opsin duplicated in animals was correlated with the water penetrating ability of the wavelength optimum of that newly arisen subfamily. This comparison made it possible to investigate whether there is a correlation between the usefulness of the opsin in water of a particular depth and the increase in the depth at which the oceans of the earth became habitable. In the absence of a correlation there would be no reason to suspect that the de-acidification and oxygenation of the oceans had an influence on metazoan vision. Conversely, in the presence of a correlation, it can be suggested that "Ocean Drive" – the change in the chemical composition and ambient conditions of the oceans - has been a primary or significant driver of metazoan vision.

2.2 Materials and Methods

2.2.1 Construction of the Opsin Tree From Previous Work

The tree was constructed from the previous work of Feuda *et al* (2012) where an opsin phylogenetic tree was reconstructed using MrBayes and the best fitting model of sequence evolution was determined to be the GTR substitution matrix using a gamma distribution approximation of rates. Some melatonin receptors were found to be the closest outgroup based on phylogenetic analyses (Feuda *et al*. 2012). The melatonin receptors were also found to be the closest outgroup by Fredriksson *et al* (2003) and were used in the work of Plachetzki (2010). Based on the results found by Feuda *et al*. (2012), the tree was modified to ensure that cnidarian opsins were included in each of the C, R and Go-opsin clades. This tree was manually pruned to only include essential taxa that describe each opsin sub-family. The main results of the previous study by Feuda are discussed briefly in section 1.2.2 and the tree used in the following analyses is shown in Figure 2.3.

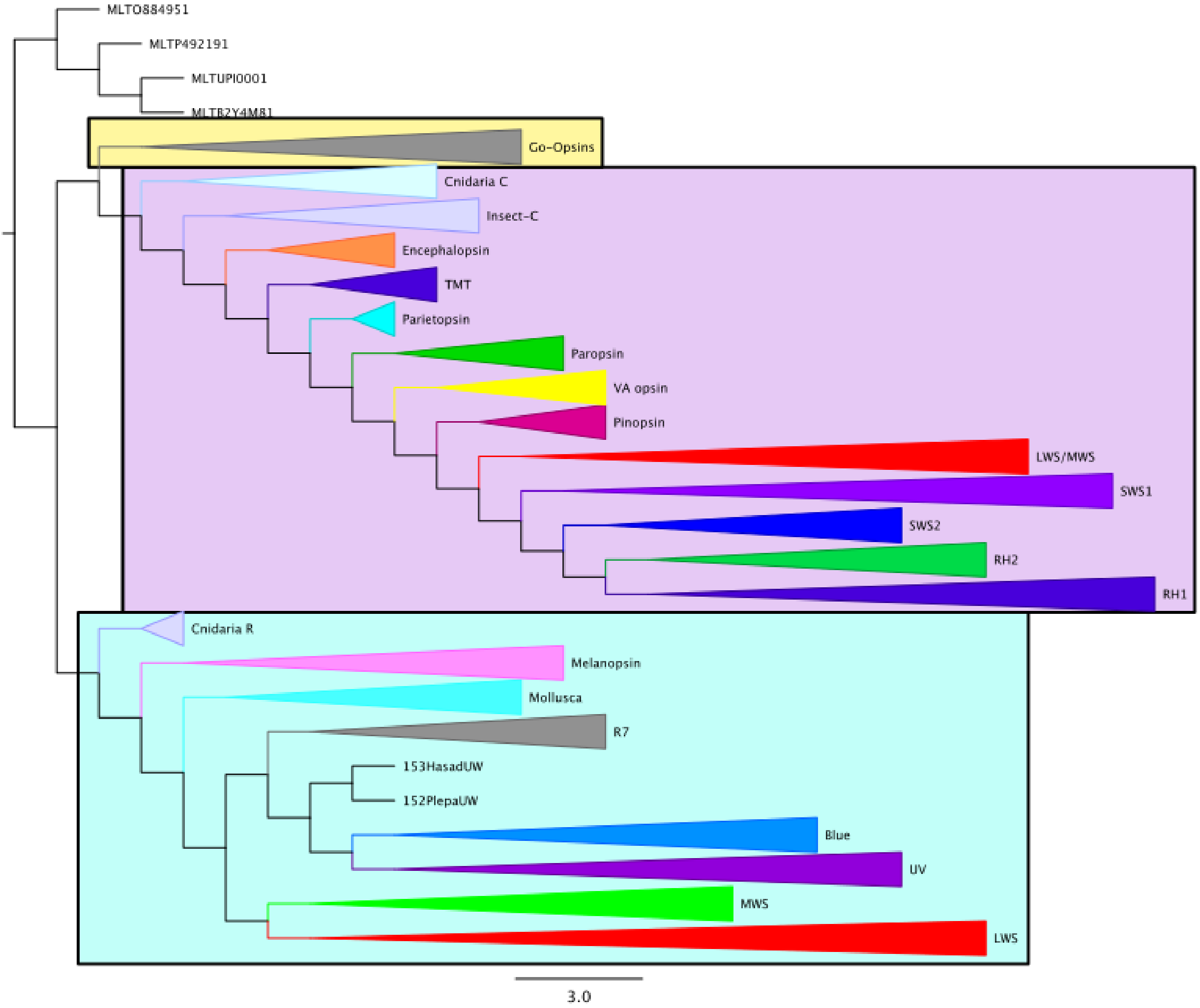


Figure 2.3: The full opsin phylogeny from Feuda *et al* (2012). The purple box marks the C-opsins, which contain the vertebrate visual opsin subfamilies (LWS, SWS1, SWS2, RH1, RH2). The blue box marks the R-opsins, which contain the arthropod visual opsins (LWS, MWS, Blue, UV, RH7, the last of which is involved in vision but its function is unknown). The yellow box marks the Go-opsins. Melatonin (MLT) receptors were used as the outgroup.

2.2.2 Sensitivity Analysis Overview

Once a phylogenetic tree was generated, a set of known dates for phyletic events were used as judged by the fossil record in order to extrapolate and interpolate the dates of other phyletic events on the phylogenetic tree. These estimations were carried out using Phylobayes v3.2c (Lartillot *et al*. 2009). The fossil calibrations were mostly taken from the work of (Benton *et al*. 2009) although the date used to describe the root node was taken from the work of (Erwin *et al*. 2011). Based on the findings of Feuda *et al* (2012), this root calibration point describes the branch separating the sponges and Trichoplax from the other animals, as dated by Erwin *et al* (2011).

Using Phylobayes, a sensitivity analysis was performed to determine the most appropriate dating model, substitution matrix and percentage soft bounds to use for the dataset. The dating models *ugam* and *ln* and the matrices GTR, CAT and LG were tested. Soft bounds were tested using the default 2.5%, as well as 5%, 10%, 20%, 30% and hard bounds. The default bounds allows the dates for a node to break the calibration points given if necessary, i.e. the dates can be up to 2.5% older or younger than the calibration point. All of phylobayes runs for the sensitivity analysis were executed for 60 hours before the parallel runs were checked for convergence using tracecomp.

Running Phylobayes, using each combination of parameters, tested whether parameter variation had a significant effect on the results. Two runs of each parameter combination (72 runs – excluding the jackknife runs, the topology tests and the final analysis) were set up and convergence was checked using the tracecomp analysis in the Phylobayes package. The resulting dates found for each of the internal nodes (taken from the .dates output file) were graphed in ascending order.

After selecting the model, matrix and bounds, a jackknife test was performed to see if any of the calibrations have had an increased effect on the dating results. A total of 100 random jackknife permutations were tested, removing a random 50% of the calibrations each time. A Phylobayes analysis was executed on each new set of calibrations. The average date found at each node was calculated from the results of each jackknife test. The averages were then compared to the original analysis, containing the full set of calibrations, to check for significant differences in results. Each jackknife analysis was executed for 24 hours.

A topology test was performed to determine if the use of certain topologies could significantly alter the dating results. Initially, the topology tested was created by moving the Go-opsins to be the sister group to the R-opsins, instead of the C-opsins (TOPA). The next topology to be tested was created by further altering TOPA by moving the Cnidarian R-opsins so that they were placed as the outgroup to all the main opsin families (C, R and Go) (TOPB). Lastly, the lamprey Rh2 was placed as the outgroup to both Rh1 and Rh2 to create the final topology (TOPC). These three topologies were tested separately to see if they had any effect on the dating results for our nodes of interest (the visual opsin duplication events). Moving branches has the effect of altering internal branches and can make calibration points inconsistent with the fossil record. Therefore any calibration points that became problematic when the topology was changed were removed to avoid the constraint of incorrect nodes. Two runs of each topology were executed and tested for convergence using tracecomp.

2.2.3 Inclusion of the Onychophora Sequences

The dataset used for the sensitivity analysis differed slightly from the final dataset from which the results are taken. The final datas*et al*so included four sequences from various species of Onychophora taken from the dataset used in Hering (2012). These were key sequences to include, given that the onychophoran lineage (velvet worms) separated from the Arthropods prior to visual opsin duplication. Therefore, the Onychophora only have one visual opsin type. This speciation node can then be calibrated, resulting in more accurate dates for the duplication nodes.

As the previous tree used in the sensitivity analysis was robustly supported, it was not altered prior to the inclusion of the Onychophora. Based on the work of Hering (2012) the position of the Onychophora within the tree was inferred, so they were added into the tree by manually altering the newick file. To add the Onychophora sequences into the alignment they were aligned separately and then profile aligned to the rest of the dataset using MUSCLE. As the dataset had previously been reduced, this was repeated with the new dataset that included the Onychophora. Any highly variable sites were removed from the alignment, including any sites where only the Onychophora had non-gap characters. This new alignment was used in the final analysis to date the opsin tree using

Phylobayes. This final analysis was executed for 72 hours with two runs in parallel. Tracecomp was used after the 72 hours to check for convergence.

2.2.4 Ancestral Trait Reconstruction

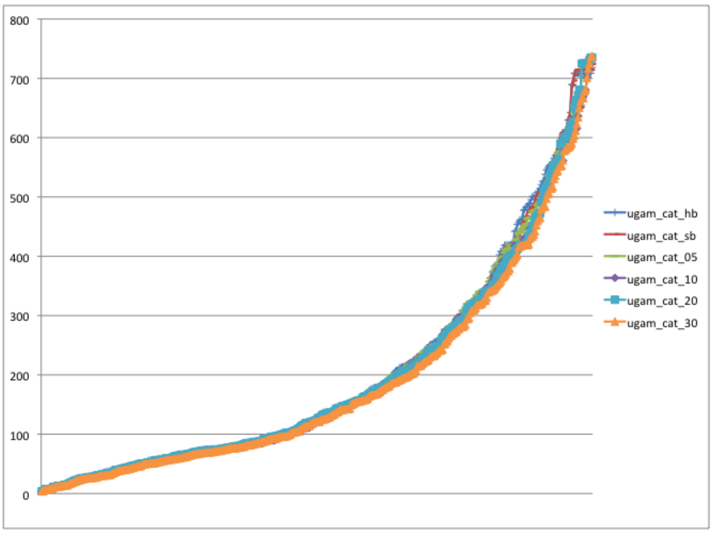
In order to determine the maximum wavelength absorbencies of the arthropod opsin repertoires prior to each of the duplications, ancestral trait reconstruction methods were used. It was not necessary to calculate the ancestral traits for the vertebrates as the tree was so ladderised, the order of the emergence of each of the subfamilies was clear. The wavelength absorbancies of some of the arthropod taxa in the tree were found from the literature. This information was used, along with the R-opsin region of the original tree to calculate the ancestral traits at each of the duplication points for the arthropods.

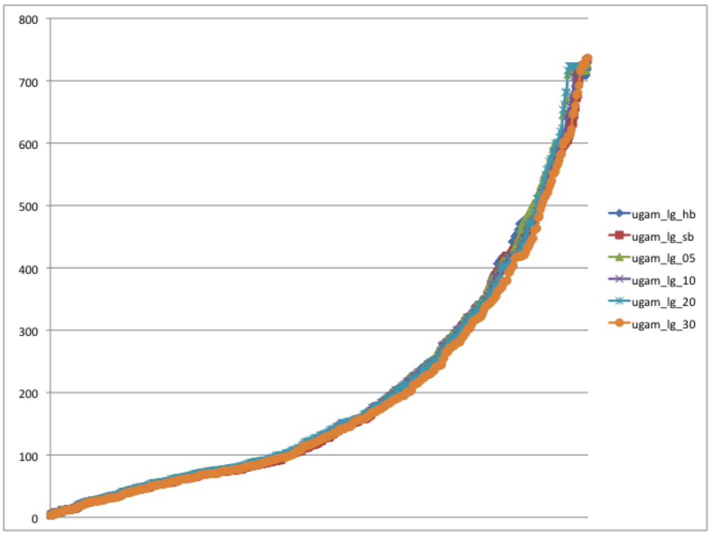
Three methods were used; parsimony calculations were performed using Mesquite, maximum likelihood calculations were performed using R v2.15.1 and Bayesian calculations were performed using BayesTraits v1.0 (although ML calculations were used with BayesTraits also). In BayesTraits, the root ancestral trait is calculated and given as the output, rather than showing the results for all the nodes on the tree. In order to determine the ancestral traits at each of the duplication points, several subtrees of the arthropod R-opsins tree were used, where the duplication points were the root nodes. BayesTraits also provided two alternative models for trait evolution, which were both tested, the directional model and the random walk model.

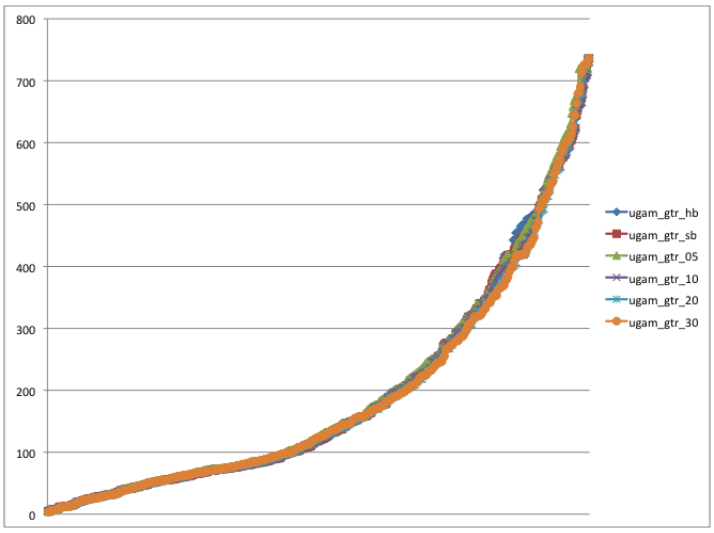
2.3 Results

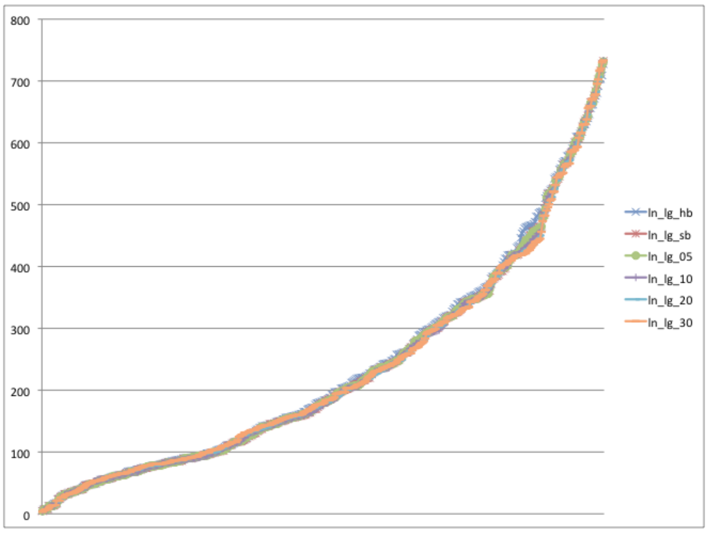
2.3.1 Percentage Soft Bounds

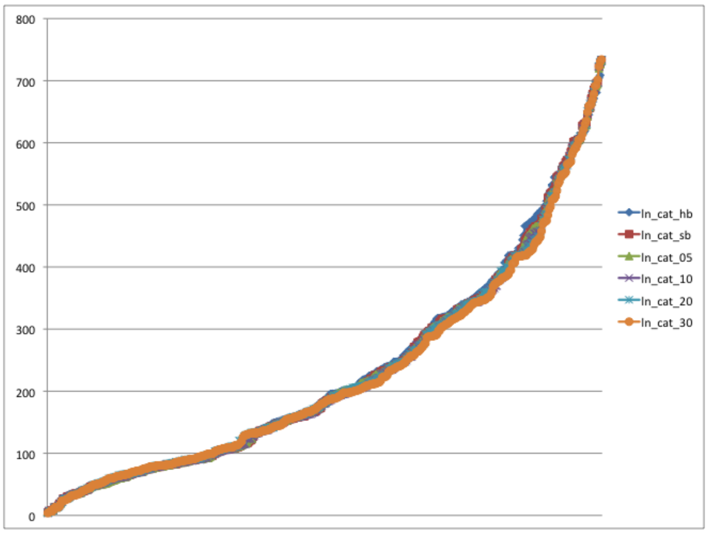
To perform the sensitivity analysis, the first parameter to be tested was the percentage bounds. Hard bounds, default soft bounds, 5%, 10% 20% and 30% soft bounds were all tested by running the analyses using all combinations of models, matrices and bounds and graphing the dating results for the internal nodes of the tree for comparison. As can be seen from the results shown in Figure 2.4, changing the percentage bounds had little effect on the dating results. Each graph shows a fixed model and matrix with the percentage bounds allowed to vary for comparison. When the resulting nodes are placed in chronological order, the curves almost exactly match up, with only minor differences.


800
700
600
500
400
300
200
100
0
ugam_cat_hb
ugam_cat_sb
ugam_cat_05
ugam_cat_10
ugam_cat_20
ugam_cat_30


800
700
600
500
400
300
200
100
0
ugam_lg_hb
ugam_lg_sb
ugam_lg_05
ugam_lg_10
ugam_lg_20
ugam_lg_30


800
700
600
500
400
300
200
100
0
ugam_gtr_hb
ugam_gtr_sb
ugam_gtr_05
ugam_gtr_10
ugam_gtr_20
ugam_gtr_30


800
700
600
500
400
300
200
100
0
ln_lg_hb
ln_lg_sb
ln_lg_05
ln_lg_10
ln_lg_20
ln_lg_30


800
700
600
500
400
300
200
100
0
ln_cat_hb
ln_cat_sb
ln_cat_05
ln_cat_10
ln_cat_20
ln_cat_30

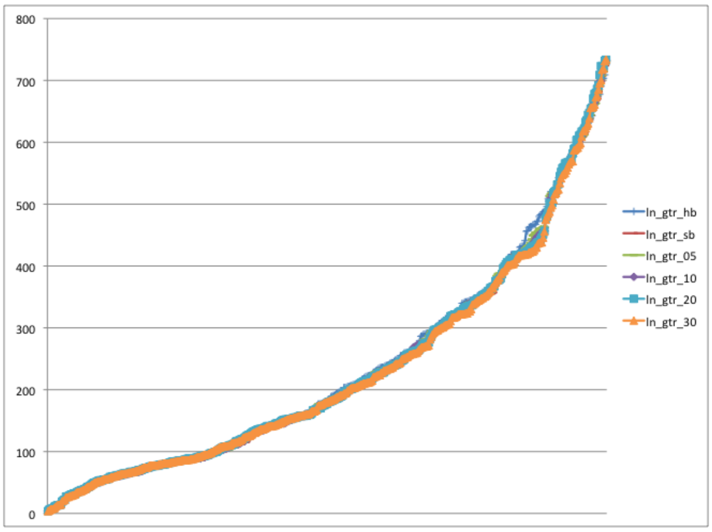

800
700
600
500
400
300
200
100
0
ln_gtr_hb
ln_gtr_sb
ln_gtr_05
ln_gtr_10
ln_gtr_20
ln_gtr_30

Figure 2.4: Sensitivity analysis results testing the percentage bounds. The six graphs represent the Phylobayes results found when each model used with each matrix and the dates found for each of the percentage bounds were used and graphed. The curves represent the dates found for each of the internal nodes. The Y-axis is the date. The nodes have been ordered in ascending order so as to allow for comparison. Changing the percentage bounds while keeping the dating model and the substitution matrix the same does not change the resulting dates found for each of the internal nodes. This is shown by the overlapping lines that represent each set of resulting dates.

2.3.2 Substitution Matrices

Having confirmed that changing the percentage bounds does not significantly affect the dating results, they were fixed at 20% soft bounds and the substitution matrices were tested. The results in Figure 2.5 show two graphs that represent both dating models, with fixed 20% soft bounds. The curves represent the dating results found for each of the three substitution matrices. When the *ln* dating model was used, changing the substitution matrix had almost no effect on the results as the curves match up almost exactly (Figure 2.5). When the *ugam* dating model was used there were minor differences to the resulting dates at some of the older nodes from about 600 mya and older. The dates of interest (the visual opsin duplications) fall within the range of approximately 500 to 600 mya. This time bracket does not differ significantly when the matrices are changed, for either model. Therefore, it can be said that for the results of this chapter, changing the substitution matrices does not have a significant effect on the resulting dates.

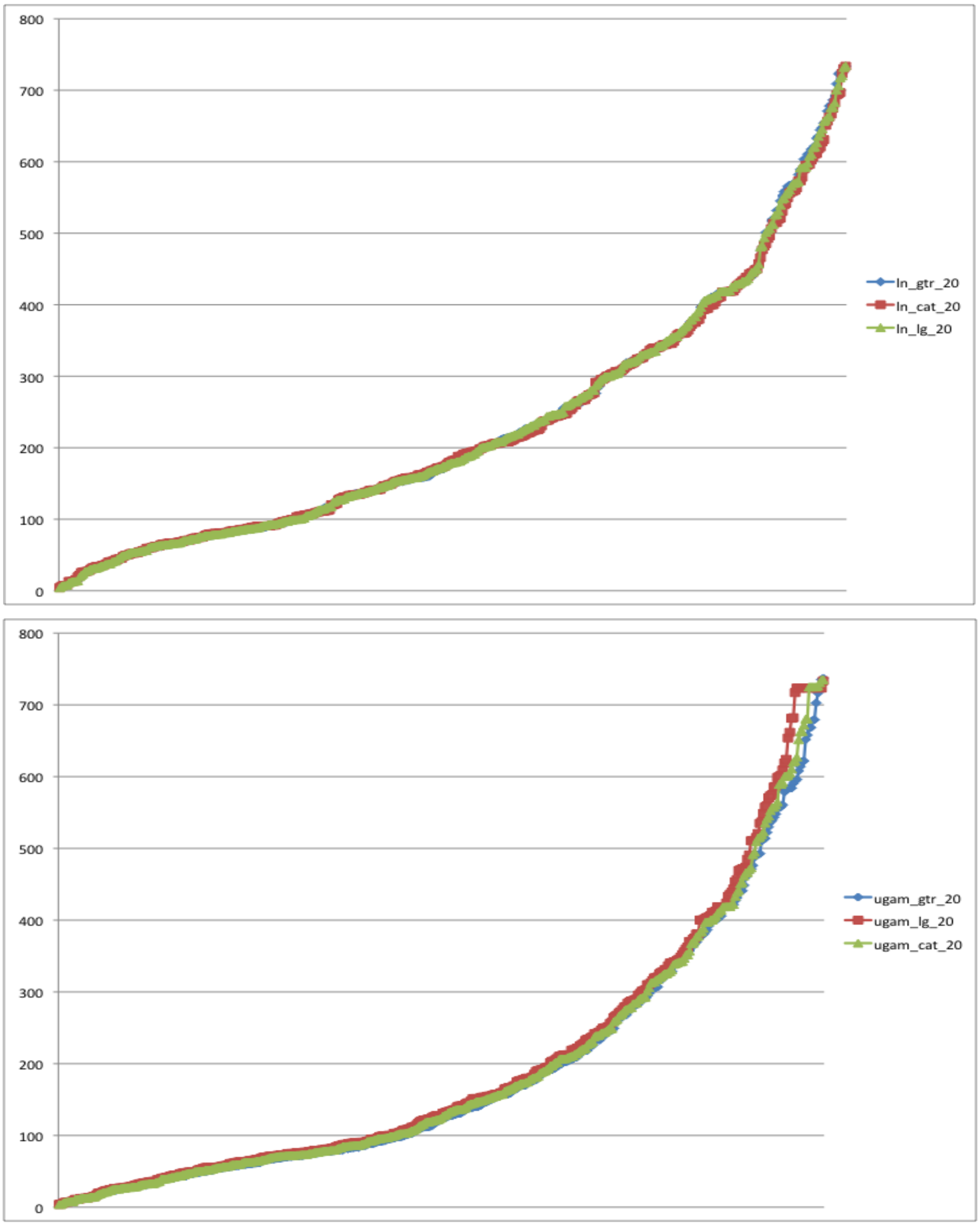


Figure 2.5: The results found after testing the effect of changing the substitution matrix on the opsin dating results. The bounds were fixed to 20% and both *ln* and *ugam* dating models were used. The dating results for each internal node was found and graphed in ascending order. The Y axis represents the date. As can be seen by the overlapping curves, changing the substitution matrix had very little effect on the results.

2.3.3 Dating Model

The dating model used was tested. Previous results have confirmed that changing the percentage bounds and changing the substitution matrix does not significantly alter the resulting dates, therefore these parameters were fixed in order to compare the differences found in the dating results when the dating model was changed. The percentage bounds were fixed to 20% and the substitution matrix was fixed to GTR. Differences were found in the resulting dates when the dating model was changed, as can been seen by the lack of overlap in large portions of the graphed dating curves (Figure 2.6). These differences only occur at the nodes that are dated approximately 450 mya or younger. As the nodes of interest for this analysis are the visual opsin duplications, and it is known that these nodes are dated at approximately 500 to 600 mya, they are not affected by changing the dating model.

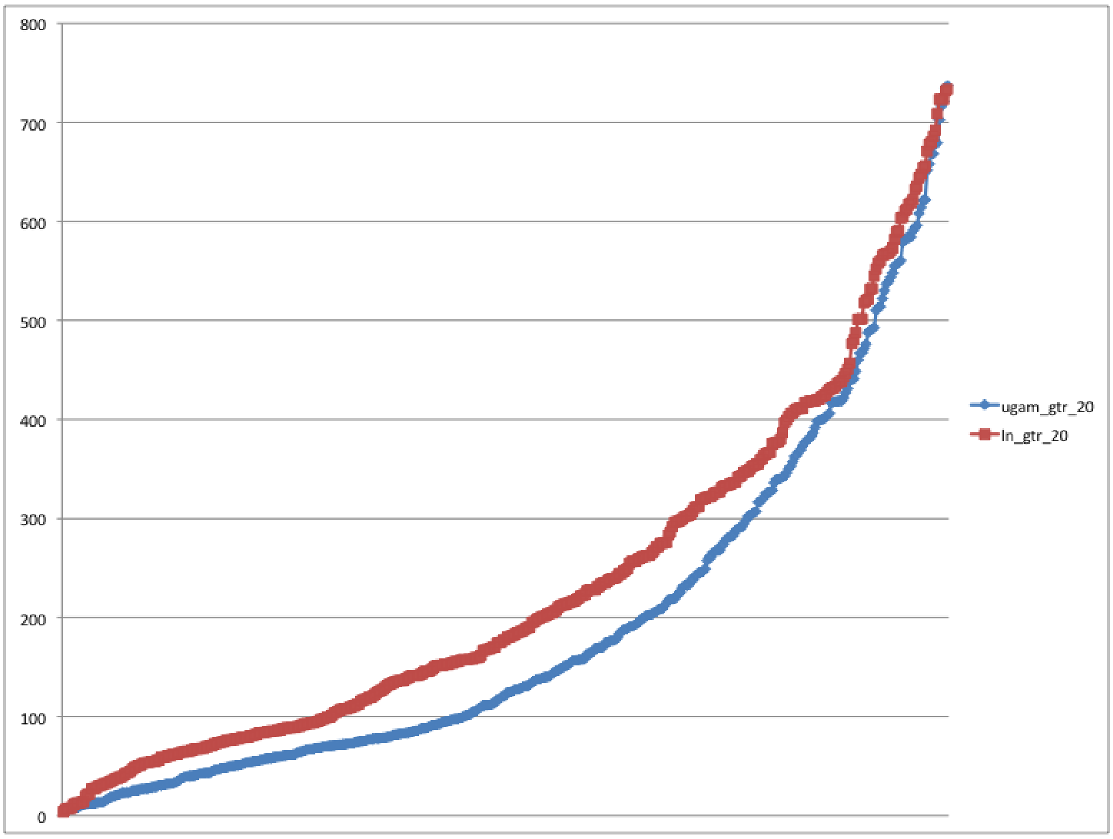


Figure 2.6: Opsin dating results found when comparing the effects of changing the dating model. The dating results found for each internal node when the matrix was fixed to GTR and the percentage bounds was fixed to 20% were graphed in ascending order for the two analyses using the dating models *ugam* and *ln*. The Y axis shows the date. The lack of overlap of the curves at the nodes dating from 400mya and younger suggests a difference in the dating results when the dating model is changed. However, in this study, the nodes of interest are the opsin duplications which are known to be between 500 and 600 mya, where there is little difference in the dating results between models.

2.3.4 Jackknife Testing On Calibration Points

The effect of changing the calibration points was analysed next by using a jackknife test. A total of one hundred 50% jackknife replicates were created from the original calibration file (see appendix) and each jackknife was used as the calibration file and analysed using Phylobayes. The resulting dates for all one hundred jackknife analyses were averaged and graphed in ascending order for comparison against the datasets where the full set of calibrations was used. All jackknife analyses were executed using 20% bounds, the GTR matrix and the *ugam* model. The jackknife test did not result in very different dates from the original dataset (Figure 2.7). The results found for the jackknife tests indicated that the dates assigned to the nodes were slightly younger, as a result of removing some calibrations, than the results of the previous *ugam* analysis. However, this only occurred on nodes that were 400 mya or younger, which does not affect the visual opsin duplications.

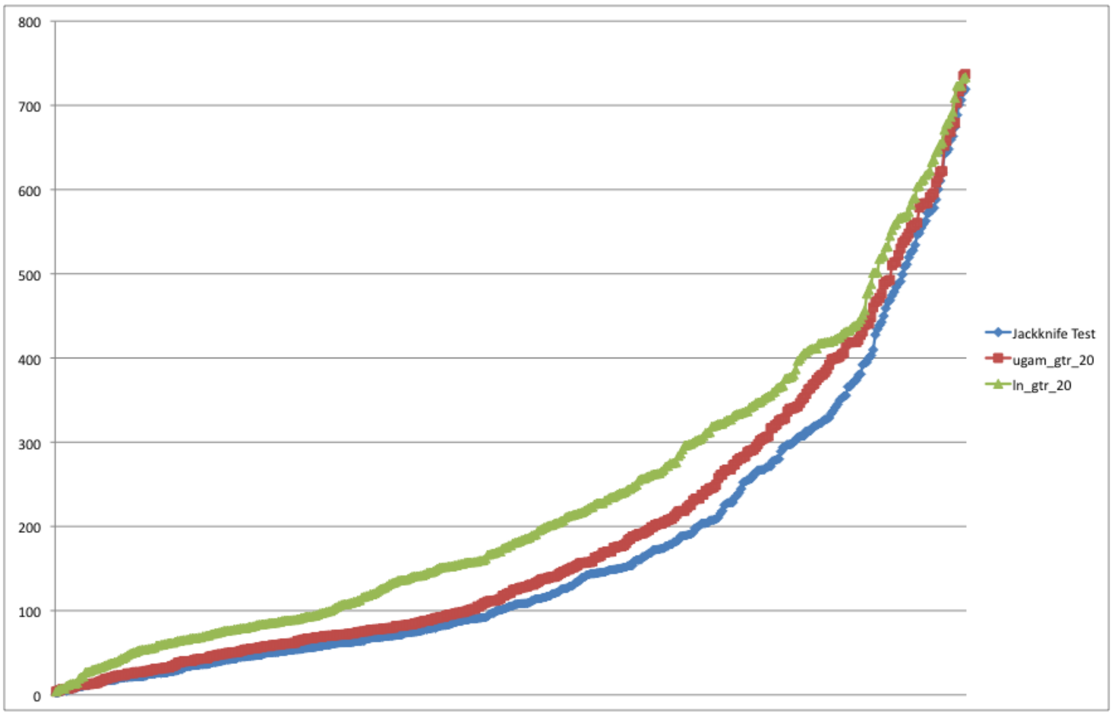


Figure 2.7: Comparison of the jackknife test results with the analyses executed with the full set of calibrations. All tests used 20% bounds and the GTR matrix. The jackknife test used the *ugam* model and the analyses with full sets of calibrations used *ugam* and *ln* models. The jackknife test dates are slightly younger than the original *ugam* result at nodes younger than 400 mya. The jackknife results are significantly different to the results found from the full set of calibrations and using the *ln* model but this can be attributed to the use of a different dating model. The dated nodes are ordered in ascending order and the Y axis denotes the date.

2.3.5 Alternative Topology Tests

Having confirmed that changing the percentage bounds, the substitution matrix, the dating model or the calibration points has no effect on the dates of the visual opsin duplications, some alternative topologies were tested to see if the resulting dates would change. Details of the alternative topologies used are discussed in section 2.2.2. All three topologies, as well as the original, were executed using 20% bounds, the GTR matrix and the *ugam* dating model. The curves produced from the dates almost exactly overlap with each other showing almost no difference (Figure 2.8).

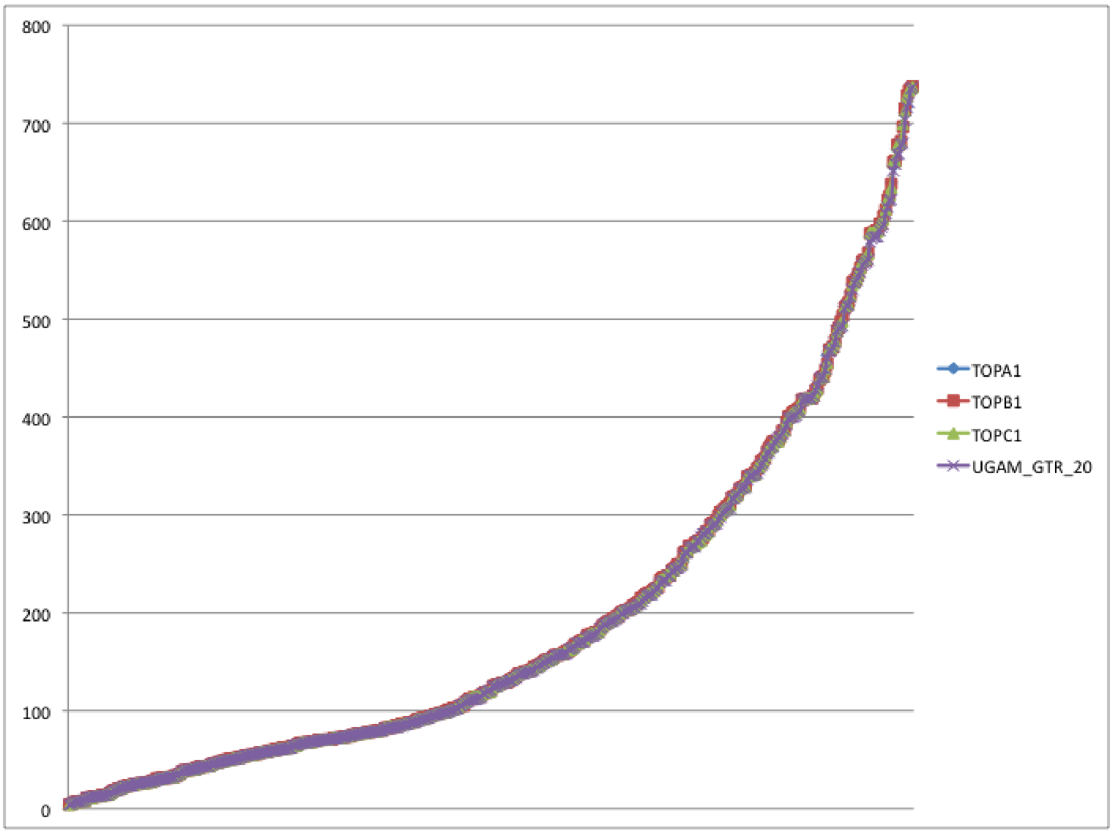


Figure 2.8: Comparing the effects of changing the topology on the opsin dating results. Three alternative topologies were used and were compared to the original analysis. All topologies were tested using the *ugam* dating model, the GTR substitution matrix and 20% soft bounds. TOPA was created by moving the Go-opsins to be the sister group to the R-opsins, instead of the C-opsins. TOPB was created by further altering TOPA by moving the Cnidarian R-opsins so that they were placed as the outgroup to all the main opsin families (C, R and Go). TOPC was created by moving the position of the lamprey Rh2 to be placed as the outgroup to both RH1 and RH2. The resulting dated internal nodes were graphed in ascending order and the Y axis denotes the date.

2.3.6 Molecular Dating Results

The molecular dating analysis was executed again using Phylobayes and 20% bounds, the GTR substitution matrix and the *ugam* dating model on the original opsin dataset combined with several Onychophora sequences. The dating results for the visual opsin duplications from this analysis are summarised in Figure 2.9.

2.3.7 Ancestral Trait Reconstruction

The ancestral wavelength absorbencies from the internal nodes of the Arthropod R-opsins region of the tree were calculated using several ancestral trait reconstruction methods. The results were extrapolated from the wavelength absorbency information of extant taxa, taken from the literature (see appendix). Multiple different software, methods and models were used (see methods for details). The duplications are summarised and labelled in Figure 2.9. The labels used in Figure 2.9 are used here to refer to the internal nodes and the results found for each of the ancestral trait reconstruction analyses are summarised in Table 2.1.

Table 2.1: Ancestral trait reconstruction at the duplication nodes. Showing the inferred wavelengths maximally absorbed by the opsin present before each duplication, i.e. at each duplication node (labels taken from Figure 2.7). The first column indicates the testing method and model, the second column lists the software used and Node A1, Node A2 and Node A4 show the results found for each of the duplication nodes, labelled A1, A2 and A4, from Figure 2.7. The wavelength results shown are in nanometres (nm).

Test/Model	**Software**	**Node A1**	**Node A2**	**Node A4**
Parsimony	Mesquite	460	473	418
ML	R	461	474	419
ML/Directional	BayesTraits	426	462	390
Bayesian/Directional	BayesTraits	426	462	390
ML/RandomWalk	BayesTraits	460	502	400
Bayesian/RandomWalk	BayesTraits	460	502	400
Average	------	449	479	403

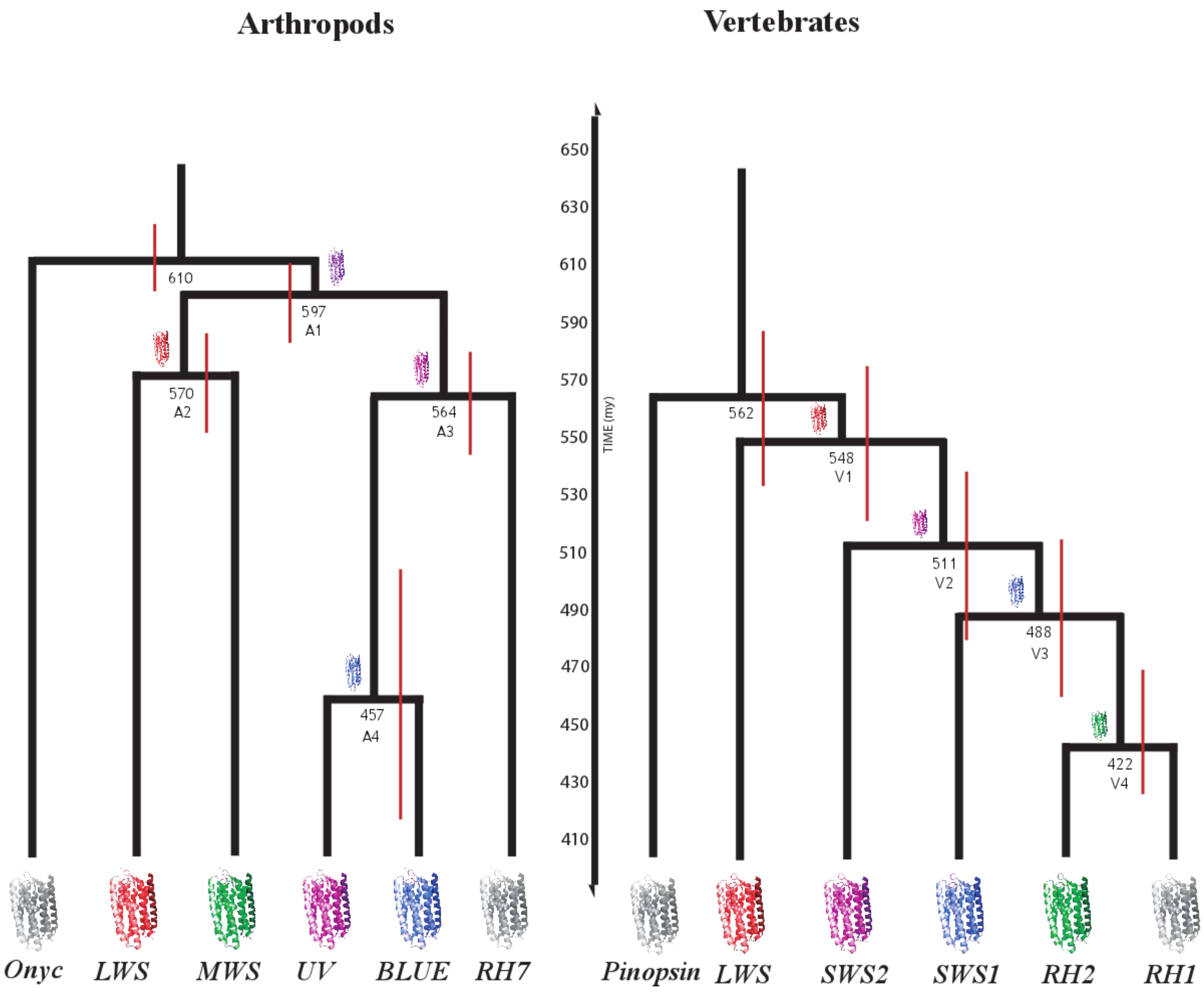


Figure 2.9: Arthropod and vertebrate dated visual opsin subtrees. Each duplication point dated (the number above each node) and confidence intervals included (the red bars). The colour of the opsins on each node correspond to the most likely ancestral state maximum wavelength absorbency before each duplication. Each node is labelled below the node with A for arthropod and V for vertebrate followed by a number, in order of when each duplication occurred.

2.4 Discussion

2.4.1 Patterns of Duplication and Light Penetration

An analysis of the order of duplications, seen in both the vertebrate and arthropod visual opsin repertoire, showed that the first opsins to evolve allowed the detection of shallow penetrating wavelengths of light. The opsins that evolved at a later stage were capable of detecting deeper penetrating wavelengths of light. These light-detecting capabilities have obviously arisen independently in both groups and use different opsin types, but yet there is a remarkable parallel in the order in which these functions arose.

The vertebrate visual opsin section of the tree showed a ladderised/imbalanced duplication pattern. The LWS clade is the first to separate from the rest of the opsin family (node V1 in Figure 2.9). This group of opsin sequences optimally detects light of 500-550 nm wavelengths, which can only penetrate coastal water to a depth of approximately 10 m. The range of light detectable by LWS opsin is capable of penetrating relatively shallow water. The next gene duplication event (node V2 in Figure 2.9) in the vertebrate opsin gene history separates SWS1 from the other opsins. SWS1 detects light at approximately 350-410 nm wavelength, which can penetrate water to a depth of approximately 10 m also. This is also relatively shallow penetrating but is at the opposite end of the light spectrum. The next duplication separates SWS2 from RH2 and Rhodopsin (node V3 in Figure 2.9). The SWS2 clade detects light of approximately 410-460 nm, which penetrates coastal water to a depth of approximately 30 m. The final duplication is that which resulted in RH2 and Rhodopsin (node V4 in Figure 2.9).

RH2 maximally detects light at 460-520 nm wavelengths, which penetrates coastal water to a depth of approximately 50 m. RH1 or rhodopsin maximally detects light at approximately 470-510 nm wavelengths, which also penetrates coastal water to a depth of approximately 50 m. At depths greater than 50 m, very little light can penetrate due to the light scattering properties of coastal sediment. In vertebrates it is quite clear that each duplication resulted in a subfamily of opsins that maximally detects wavelengths of light that are capable of penetrating deeper into ocean waters. The clades arose in the order of red, violet, blue and green, which exactly matches an increasing depth of light penetration in coastal water (wavelength detection information taken from Hisatomi and Tokunaga (2002)).

The arthropod visual opsin duplication pattern is more difficult to interpret as it shows a more balanced (palmate) tree pattern. The initial duplication (node A1 in Figure 2.9) resulted in two branches that lead to arthropod MWS and LWS in one clade and R7, UV and Blue in the other. The next duplications to occur were the split between LWS and MWS (node A2 in Figure 2.9) and also the split between R7 and the SWS (UV, Blue) clade (node A3 in Figure 2.9). It can be seen from the ancestral trait reconstructions of each of the internal nodes, the initial state was indigo, one of the shallowest penetrating wavelength of light. At the time of the first duplication (A1), the ancestral states were indigo and red, which are relatively shallow penetrating wavelengths of light. Next the red receptor duplicated (A2) to give red and green and the indigo receptor duplicated (A3) to give blue and RH7 (which has an unknown function). The last duplication (node A4 in Figure 2.9) is unrelated to water penetration as it is insect specific and there are no reported aquatic insects, so will not be discussed in this context as it is likely a relatively recent adaptation to terrestrial environments. After the initial three duplications (A1, A2, A3), arthropods would have acquired the ability to detect red, green and blue light – a situation that is similar to what is seen in the vertebrates. In the case of the arthropods, their initial state was that they could see by detecting indigo light, which is a shallow penetrating wavelength. Then they moved deeper into the ocean and acquired the ability to detect red light, which is a slightly deeper penetrating wavelength of light in coastal waters. Lastly, they acquired the ability to detect green light, the deepest penetrating wavelength in coastal waters. This shows that for both vertebrates and arthropods, the order in which each duplication occurred and the subfamilies that arose clearly follows a pattern of increasingly deeper penetrating wavelengths of light in coastal water.

2.4.2 Timing of Duplications

The internal nodes on the tree were dated using Phylobayes. It was necessary to accurately date the duplication events in both the arthropods and the vertebrates to determine if the apparent gain of opsins capable of detecting increasingly deeper penetrating wavelengths of light occurred simultaneously in both groups of animals. This would suggest that there was an environmental pressure that caused this pattern to occur in the same fashion and at the same time in such vastly different groups of organisms.

To accurately date our tree using Phylobayes, a sensitivity analysis was performed. This was done to determine the robustness of our results and to accurately select the best fitting dating model, substitution matrix and percentage soft bounds. The effects of changing the dating model, the substitution matrix and the percentage soft bounds were all separately tested, as well as the influence of the calibration points on the resulting dates and changing the topology. The sensitivity analysis showed that the molecular clock analyses are essentially robust. The only parameter to which the results are sensitive to methodological variance is the selection of the dating model that was used. However, alternative molecular clock methods did not significantly affect the inferred ages of the nodes in the opsin phylogeny that are important for this analysis.

The *ugam* GTR 20% bounds final results from the dataset including the Onychophora sequences (shown in Figure 2.9, with corresponding node labels) showed the dates for the vertebrate duplications as 442 (466-427), 488 (515-460), 511 (538-482) and 548 (573-520) (in mya) for nodes V4, V3, V2 and V1 respectively. The vertebrate visual opsins split from the non-visual opsins (pinopsin) at 562 (587-532) mya. The dating results for the arthropod duplications were 457 (504-415), 564 (581-545), 570 (586-551) and 597 (611-584) (in mya) for nodes A4, A3, A2 and A1 respectively. The arthropod visual opsins split from the Onychophoron opsins at 610 (622-598) mya. The results show that the arthropod opsins may have diversified first, although the confidence intervals do overlap significantly so they may have happened around the same time. This timescale also coincides with the oxygenation of the oceans at around 600 mya (Li *et al*. 2010).

The patterns of opsin duplication are clearly correlated with light penetration in water. The dating results show that this effect (the duplications) occurred almost simultaneously in both the arthropods and the vertebrates. This was likely due to opening up of new ecological niches in the form of deeper waters after the oxygenation of the oceans.

2.4.3 Correlation with Global Events

More than 95% of extant metazoan species possess eyes that are capable of light detection (Land and Nilsson 2012) which evolved in their ancestors around 600 mya (Figure 2.9). These early light detecting animals quite likely stayed close to the surface of the water where light and food sources would be most abundant in addition to abundant oxygen sources due to the presence of photosynthetic cyanobacteria. These cyanobacteria contributed to the early oxygenation of the oceans once they began to proliferate (Tomitani *et al*. 2006). Animals require relatively high levels of oxygen to survive, and the oxygenation of the oceans might have fuelled early animal evolution (Canfield *et al*. 2007). Eventually, oxygen became available in deeper and deeper oceans, allowing animals to explore these new ecosystems in order to find new food sources or evade predators. Light availability would vary depending on the depth, as different wavelengths of light do not penetrate water to an equal extent. The results of this study suggest that the evolution of colour vision occurred as a result of early metazoans exploring newly available deep-water niches where early animals would have evolved the ability to detect the wavelengths of light that were most abundant at particular depths.

Rhodopsin, the opsin that allows vertebrates to see in dim light, is a relatively recent adaptation (Pisani *et al*. 2006; Yokoyama *et al*. 2008) compared to bright light colour vision. It, therefore, seems likely that early vertebrates (and possibly invertebrates) remained close to the surface, where the greatest amount light was available. In aquatic animals that explore deep-sea environments, there is a relaxation of selective pressure to maintain a varied repertoire of opsins due to the limited availability of light in deep-sea environments. Therefore, deep-sea animals tend to have fewer opsins (Davies *et al*. 2012). This suggests that early animals must have remained within the photic zone during the early evolution of vision. As animals moved into deeper waters, they developed the ability to detect the available wavelengths of light for that depth. Both patterns of duplication in vertebrates and arthropods show this trend of acquiring opsin photoreceptors capable of detecting wavelengths of light that are capable of penetrating deeper and deeper water. The mirroring of the ordering of opsin acquisition is unlikely to have occurred by chance, due this pattern emerging for two major animal groups (vertebrates and arthropods) at the same time, in addition to being the pattern seen in light penetration of water.

The opsin duplications in both the vertebrates and the arthropods occurred not long after the suggested oxygenation of the oceans approximately 600 mya (Holland 2006). Arthropods were capable of the visual exploration of deeper oceans slightly earlier than vertebrates. In both cases, the duplication patterns follow a trend of acquiring opsins capable of detecting increasingly deeper penetrating wavelengths of light in coastal water, suggesting that these were acquired in order to explore deeper ocean environments where only certain wavelengths of light are abundant.

2.5 Conclusion

From the previous work by Feuda *et al.* (2012), a robust opsin phylogeny was created with a parsimonious duplication pattern. In this chapter, this phylogeny was accurately dated and the effects of each of the parameters on the results were extensively tested to give the most robust dating results. Each of the visual opsin duplications for both the vertebrates and the arthropods were dated and shown to have occurred at around the same time, suggesting a common evolutionary trend for the emergence of colour vision in these two distantly related animal groups. It was determined that the emergence of colour vision in vertebrates and arthropods coincides with the oxygenation of the oceans, which likely had a massive effect on early animal evolution, as all animals would have been aquatic at this time. In both groups of animals, the acquisition of new opsins follows a trend of acquiring opsins capable of detecting increasingly deeper water penetrating wavelengths of light. The results detailed in this chapter support the “Ocean Drive” hypothesis. The null hypothesis, that the evolution of colour vision is not affected by light penetration patterns in water, can be rejected. It seems likely that the oxygenation of the oceans played a powerful role in the evolution of colour vision. By allowing animals to travel into deeper waters, reducing the range of light visible to them, the visual sensory system evolved by duplicating the animals' opsin repertoire multiple times. Then, functional shifts to the wavelength absorbency ensured that the animal could maximally detect the wavelengths of light most prevalent at particular depths.

Chapter 3 – Timing the Emergence of the Rod Visual Pathway

The focus of the previous chapter involved a discussion on why colour vision evolved. In the following chapter, how and why dim light vision evolved is discussed. In this study, the emergence of the rod dim light visual pathway from the ancestral cone pathway is analysed by looking at the duplication patterns and timing of each of the proteins involved to investigate if there is an evolutionary trend. This trend was analysed in order to discover the evolutionary mechanisms that arose leading two these two separate but related pathways and cell types.

3.1 Introduction

3.1.1 The Importance of Vision

Vision / light detection is a highly developed and essential sensory system that is present across a wide variety of animal phyla. The ability to detect light gives animals a great selective advantage to be able to identify food, threats or mates, in cases where other senses might fail. Six of the thirty-three animal phyla have image forming eyes; Cnidaria, Mollusca, Annelida, Onychophora, Arthropoda and Chordata. These taxa comprise up to 96% of the extant animal species. The widespread occurance and success of the eye shows that once it evolved early in animal evolution, it quickly became hugly important (Fernald 2006).

3.1.2 Variation in Eye Morphology and Biochemistry Across the Metazoa

Certain proteins are expressed in the eyes of all animals with image forming eyes across all of the phyla, e.g. the PAX6 transcription factor, which is expressed throughout the animals life for various functions, but one of those functions is to mediate the development of the eyes (Gehring and Ikeo 1999). Also, the basic eye structure can be similar between very different species e.g. humans and lamprey. In contrast to these similarities, the proteins involved in the phototransduction pathways of distantly related animals can be quite different (Miller 1957; Menzel and Blakers 1976; Ashery-Padan and Gruss 2001; Kobayashi and Kohshima 2001). Vertebrate and invertebrate phototransduction pathways use different proteins to propagate visual signals which culminate in different electrical signals (Figure 1.4). Activation of the vertebrate pathway results in a hyperpolarisation (an efflux of positive ions from the cell causing a negative change in the call membrane's potential) of the cell, whereas activation of the invertebrate pathway results in a depolarisation (an influx of positive ions into a cell causing a positive change in the cell membrane's potential) of the cell (Fernald 2006). The common use of PAX6 and opsin proteins in both vertebrate and arthropod vision suggests a distant common evolutionary origin that lead to two very different mechanisms for light detection in these distantly related animal groups.

Eye morphology shows a great deal of variation (Nilsson 2004). Vertebrates have camera style eyes with moveable lenses that focus light onto the retina at the back of the eye where the photosensetive cells are present (Lamb *et al.* 2007). A similar eye structure can also be found in cephalopod molluscs, such as octopi and squids (Ogura *et al*. 2004). Upon further inspection, the similarity in structure between the eyes of cephalopods and vertebrates is clearly as a result of convergent evolution (Ogura *et al*. 2004). The morphology of the vertebrate eye is unusual in that the photoreceptor cells face away from the incoming light, towards the back of the retina, which is not the case in cephalopods (Nilsson 1996). The construction of the vertebrate eye means that light must pass through layers of blood vessels to reach the photoreceptor cells reducing the visual aquity. To overcome this problem to a degree, a small region of the retina has a reduced number of blood vessels allowing the light to gain easier access to the photoreceptors (Pumphrey 1948). This region is called the fovea and its presence explains why only the centre region of vertebrate vison is sharp, where vertebrate peripheral vision is significantly less clear. The fovea must remain as a small region because the retina requires a large blood supply to function. Cephalopods do not have this problem because their photoreceptor cells face towards the incoming light. Therefore, they have no reduced visual aquity and no need for a fovea. Cephalopod vision can be significantly better than any vertebrate due to the arrangement of their photoreceptor cells within the retina.

In arthropods, the main visual system uses a compound eye structure where multiple units, called ommatidia, with imoveable lenses, detect light separately from one region in the animals field of view. Each ommatidium sends a signal to the brain to process these separate "pixel-like" regions as an image. The number of ommatidia in a compound eye as well as the angle differences between each ommatidium determine the visual acuity of the animal (Land 1997). Dragonflies can have thousands of ommatidia giving them a significantly better ability to discriminate detail than animals with fewer ommatidia such as grasshoppers.

Basic cell structure of the photoreceptor cells can vary between vertebrates and arthropods. In arthropods rhabdomeric cells are used to detect light, whereas in vertebrates, ciliary cell types are used (Arendt 2003). Both cell types aim to achieve a larger surface area on the cell where photoreceptor proteins (opsins) can sit, increasing the opsin's chances of being activated by incoming photons of light. Rhabdomeric cells achieve greater surface area by having multiple folds of the surface membrane. Ciliary cells achieve greater surface area by having multiple expanded folds of a cilium that is extended from the cell (Figure 3.1).

In this chapter the focus is on the evolution of vertebrate visual pathways. Within the vertebrate visual system, there are two main types of photoreceptor cells, rods and cones. The rod and cone names stem from the basic shape of the cell. Cones are used for bright light, day time, colour vision due to their multiple opsin receptors and relatively reduced light sensitivity. Rods are used for dim-light, night time vision due to their single opsin type and relatively high sensitivity to small amounts of light (Bowmaker and Dartnall 1980; Pugh and Lamb 2000; Carter-Dawson and Lavail 2004).

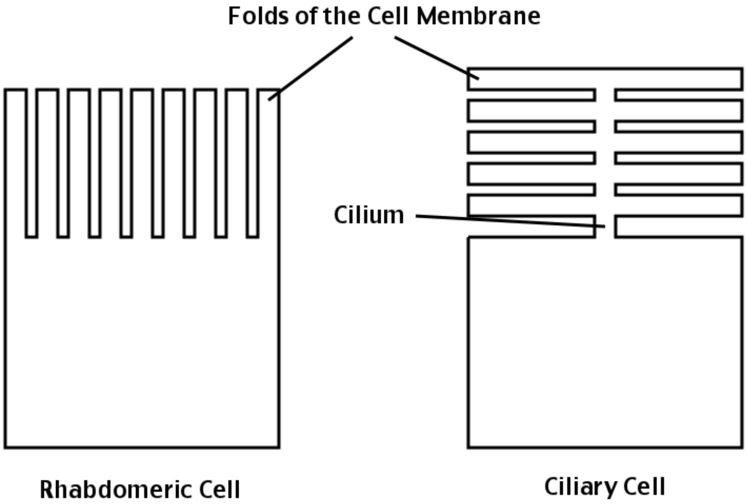


Figure 3.1: The constrasing cell types of arthropod and vertebrate photoreceptor cells. Arthropods have rhabdomeric cells and vertebrates have ciliary cells. Both cell types aim to achieve increased surface area. Rhabdomeric cells achieve this by folds to the membrane and ciliary cells achieve this by expansions to a cilium that extends from the cell. Diagram adapted from Arendt (2003).

3.1.3 Types of Rods and Cones

Rods and cones use different types of proteins to propagate their signals. An opsin protein is the photoreceptor that initially detects light in both rods and cones. Rhodopsin is the only opsin protein found in rod type cells (Khorana 1992). It is highly sensitive to small amounts of light, even picking up a single photon of light, but it becomes saturated in bright light and is useless (Terakita 2005; Yokoyama *et al*. 2008). There are several types of cone opsins, SWS1 and SWS2 (short wave sensitive), MWS (medium wave sensitive), and LWS (long wave sensitive). They are named in accordance with their maximum wavelength absorbancy. These different cone opsin types are used for bright light colour vision. Cone opsins require large amounts of light to become activated as they are much less sensitive to light than rods and the duration of their photoresponses is much shorter (Hestrin and Korenbrot 1990; Burns and Baylor 2001). As each opsin maximally detects light of a different wavelength, when a light signal is perceived the contrasting signals for the different activation levels of each opsin are used to determine the colour (Figure 2.1). If only one opsin type is present, vision is monochromatic, if two are present, vision is dichromatic (Carroll *et al*. 2001), if three are present (as in humans), vision is trichromatic (Surridge *et al*. 2003) and if four are present, vision is tetrachromatic (such as in birds) (Vorobyev *et al*. 1998; Nickle and Robinson 2007). Cone opsins, used for colour vision, require large amounts of light to function. So in low light conditions, the cone opsins are not activated, as there are insufficient amounts of light to cause a reaction (Yokoyama 2002).

The rest of the activation pathways between rods and cones are very different with each protein in the activation pathway having a specific rod and cone type. This specialisation of the rods and cones for different visual functions arose as a result of duplication of an ancestral cone pathway. This was followed by natural selection acting upon the rod duplicates to allow them to specialise their function differently from the cones in regards to their photosensitivity (Plachetzki and Oakley 2007).

3.1.4 Specialisation of Visual Systems

Mutations at certain amino acid sites can alter the wavelengths of light that activate an opsin, increasing or deceasing the wavelengths of light that the opsin maximally absorbs. In this way, mutations can give certain species a more specialised visual system for their environment. For example, the coelacanth has several mutations in its opsins causing the maximum absorbancy to be at slightly shorter (in the blue coloured range) wavelengths compared to other vertebrates. This gives the coelacanth a specialised visual system for its deep sea environment as longer wavelengths of light cannot pass through vast amounts of water as efficiently as shorter ones (Yokoyama 2000), i.e. in deep oceans blue light penetrates water to the greatest degree.

It is likely that early vertebrate life evolved in a light abundant environment, in shallow waters as deep oceans would have been highly acidic and anoxic early in animal evolution (Holland 2006). Therefore, the needs for dim light vision only occurred after animals moved into deeper ocean environments where light levels are lower. This issue has been discussed in the previous chapter where evidence was presented for a selective pressure to move into deeper oceans as they became oxygenated. Many vertebrates, such as cats, owls or deep-sea fish, have highly specialised eyes for nocturnal or dim light vision (Zhao *et al*. 2009). They have a very high percentage of rods on their retinas (80 - 100% in nocturnal animals) and large eyes that allow for the detection of much more of the available light in the environment than would be possible with smaller eyes (Yokoyama 2000). Some animals, such as the tokay gecko (*Gecko gecko*), are nocturnal and have morphologically pure rod retinas but these "rods" express cone photoreceptor proteins allowing the gecko to see some colours in dimly lit environments (Kojima *et al*. 1992).

Early in mammalian evolution, mammals were primarily nocturnal due to the dominance of the dinosaurs at the time (Ryszkiewicz and Walker 1983). This resulted in a loss of cone receptors in mammals as there was no longer any need to see in the bright light of daytime (Jacobs 2009) and explains why most mammals have only two types of cone opsins, giving them dichromatic vision (LWS/MWS and SWS1). In the primate lineage, there was a duplication of the LWS/MWS gene. This resulted in one of the copies being free to mutate, resulting in a more red shifted LWS/MWS type, which maximally absorbs red/orange wavelengths of light. This gave humans and some primates trichromatic vision, which greatly increased their visual abilities (Surridge *et al*. 2003). So, rather than having a single longwave sensitive opsin, they now have two, one that detects green and one that detects red.

3.1.5 Features and Functions of the Phototransduction Pathway Proteins

In vertebrates, the phototransduction activation pathway follows a number of steps. The light activation signal is detected by a C-opsin (ciliary opsin, present in ciliary cells) which then activates a G-protein, Transducin. This G-protein in turn, activates Phosphodiesterase 6 (PDE6), causing a reduction in the cellular levels of cGMP by hydrolysis. This drop in cGMP levels closes the cyclic nuleotide gated ion channels that are opened by the binding of cGMP, resulting in a hyperpolarisation of the cell due to the blocking of the influx of positive ions (Fain *et al*. 2010) (Figure 1.4).

Opsins are G-protein coupled receptors (GPCRs) that are bound to a light activated chromophore, usually 11-cis-retinal (Sugihara *et al*. 2002). They have 7 trans-membrane helices and bind the chromophore via a schiff base to a lysine residue on the 7th helix (Dunham and Farrens 1999). When a photon of light hits the chromophore it causes a conformational change, hydrolysing 11-cis-retinal to all-trans-retinal (Sugihara *et al*. 2002). This in turn causes a conformational change in the opsin, causing the outward movement of helices III and VI, exposing the G-protein binding site in the loop between V and VI (Bourne 1997; Tsukamoto *et al*. 2010).

G-proteins are signal transduction molecules that bind to GPCRs to transduce a signal from the receptor. G-proteins are heterotrimeric proteins that contain alpha, beta and gamma subunits (Onrust *et al*. 1997). The alpha subunit is bound to a guanosine diphosphate (GDP) molecule and it is also the subunit that interacts with the GPCR and the PDE6 (in the case of vertebrate vision) (Ridge *et* *al*. 2003). When the G-protein (Transducin in vertebrates) binds to the opsin GPCR, the GDP is replaced with guanosine triphosphate (GTP) and the G-protein becomes activated. The beta-gamma subunits disasociate as a dimer from the activated alpha-GTP subunit, which then goes on to bind to the gamma subunit of the PDE6 (Clapham 1996; Hamm 1998).

PDE6 consists of four subunits, an alpha and beta subunit in rods and two alpha subunits in cones, as well as two gamma subunits in both rods and cones. The gamma subunits bind to the binding site of cGMP on the alpha and beta subunits, preventing the binding of cGMP while the protein is inactive i.e. when the gamma subunits are not bound the G-protein. The activated G-protein alpha subunit binds to the gamma subunits, pulling them away from the cGMP binding site (Paglia *et al*. 2002). This results in the hydrolysis of cGMP, breaking it down to GMP. While the gamma subunits are bound to the activated G-protein alpha subunit, PDE6 will continue to breakdown cGMP, resulting in a decrease of cGMP concentration within the cell (Paglia *et al*. 2002).

Cyclic nucleotide gated channels (CNG-channels) are membrane bound ion channels that respond to the binding of cyclic nucleotide molecules such as cGMP or cAMP causing them to open or close. When PDE6 hydrolyses enough cGMP to reduce the cGMP concentration levels within the cell, a threshold level is crossed where there is insufficient cGMP present to maintain the CNG-channels in an open configuration. This stops the infux of Na^+ and Ca^{2+} into the cell which initiates the hyperpolarisation of the cell. When the eye detects many of these signals they are passed on to the brain to be processed as an image. CNG channels are made up of four subunits, either alpha or beta, each of which have six transmembrane spanning regions and a pore (Menini 1999; Plachetzki *et al*. 2010).

3.1.6 Co-Duplication vs. Co-Option

Two main processes drive the diversification of protein interaction pathways, causing new interactions to occur. These processes are duplication followed by mutations leading to the development of a new function and co-option of new functions by previously present proteins (Plachetzki and Oakley 2007) (Figure 3.2).

Co-duplication (Figure 3.2(a)) means that two protein networks originated by the duplication of an ancestral network. This can be seen if the resulting networks were produced by duplication events that happened roughly around the same time (Plachetzki and Oakley 2007). Co-duplication can be split into co-evolution and co-adaptation. Co-evolution is a similar duplication pattern in proteins from the same interacting network due to similar evolutionary pressures. Co-adaptation is the similarity of phylogenetic trees in an interacting network due to actual physical interactions fuelling duplication patterns (Juan *et al*. 2008). In contrast to co-duplication, proteins or genes that evolved as a result of co-option (Figure 3.2(b)) would have been due to duplication events that occur at different times. The resulting proteins will often have developed new functions as a result of being assembled to form a new interacting network (Plachetzki and Oakley 2007).

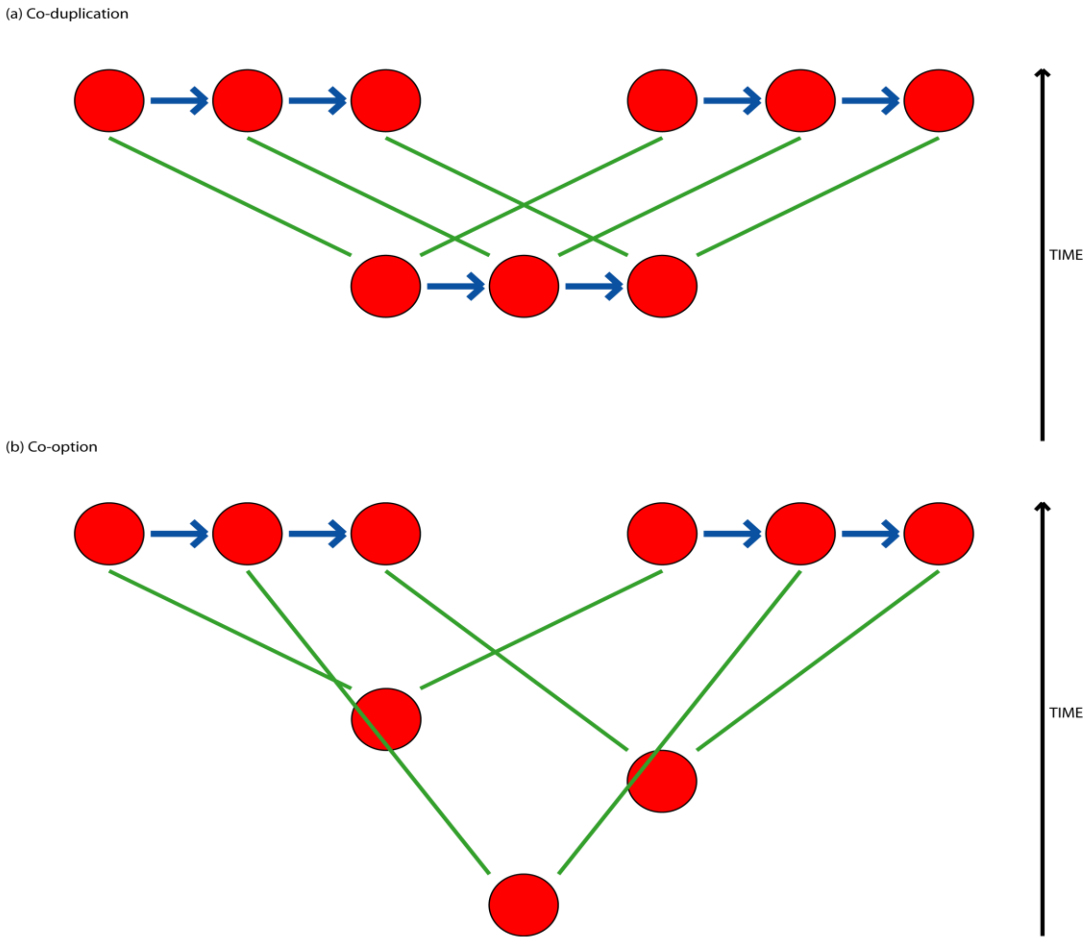


Figure 3.2: Co-duplication and co-option of pathway proteins. Adapted from Plachetzki and Oakley (2007). The red nodes represent proteins. The blue arrows represent physical interactions between two proteins. The black arrow represents the passage of time. The green lines show the ancestral protein and corresponding daughter proteins before and after duplication. Fig3.2(a) shows co-duplication, where a previously interacting pathway is duplicated due to similar environmental pressures or pressures via physical interactions. Figure 3.2(b) shows co-option where previously non-interacting proteins duplicate to give two similar pathways of proteins that now have developed a new function. In co-duplication the duplications of the proteins occur around the same time whereas in co-option, they can occur at any time, as the proteins would not be under similar selection before being co-opted into a new function.

3.1.7 How did the Rod Pathway Emerge From the Ancestral Cone Pathway?

Each protein in the phototransduction pathway has a rod and cone subtype as a result of duplication of the ancestral cone pathway proteins. In this chapter, the timing of when each of the proteins in the rod and cone pathways duplicated is analysed, to determine if the duplication of the ancestral pathway was as a result of co-duplication or co-option.

To perform this study, phylogenetic trees were constructed to trace the evolution and duplication patterns of the proteins in the activation pathways of both rods and cones. The species composition of the trees was analysed such as those done by Platchetzki and Oakley (2007) and additionally, molecular dating analyses were performed to obtain an accurate date in time for when each duplication event occurred. Then each date was compared for the emergence of the rod type across each protein in the activation pathway to determine if a pattern of co-duplication or co-option was present. The timing for the emergence of rhodopsin was taken from the dating results shown in the previous chapter.

3.2 Materials and Methods

Throughout the methods for this chapter the following software versions were used. MUSCLE v3.7 was used for constructing alignments. Seaview v4.2.8 was used to visualise the alignments. Gblocks v0.91b was used to reduce the alignment. The softwares TIGER v1.02 and PAUP v4.0b10 were used to alter the alignments based on site rates of change. ModelGenerator v0.82 was used to find the best fitting model for a dataset from a set of available models. For the reconstruction of phylogenetic trees, FastTree v2.0.1, PhyML v3.0, RAxML v7.0.4 and Leaphy v1.0 were used. Consel v1.2 was used to compare tree topologies. FigTree v1.3.1 was used to visualise the trees. Mesquite v2.75 was used to alter tree topologies. Phylobayes v3.2c was used for molecular dating analyses.

3.2.1 Transducin – Finding the Outgroup

Initially, a G-protein tree, containing all the G-protein subfamilies, was reconstructed in order to find the closest outgroup to the Transducins. A series of BLAST searches were performed using query sequences from Gt (Transducins), Go, Gq and G11 alpha subunits. An indexed NR database (downloaded and indexed in October 09) from the NCBI website (http://www.ncbi.nlm.nih.gov/) was used to allow easy identification of the sequence names from the blast output file. Then, the lengths of each sequence were calculated and any sequences that were less than 100 amino acid residues in length, or more than 1000 amino acid residues in length were removed. The identities of the sequences that were unusually long were found using the online NR database to ensure that they were not G-proteins before removing them.

Next, the four separate files containing the BLAST hit results for each of the four initial queries were added together and any duplicate sequences were removed. This reduced the number of unique sequences to 1154. The remaining sequences were aligned using MUSCLE (Edgar 2004). Then, the phylogenetic relationships between the 1154 sequences were reconstructed using FastTree (Price *et al*. 2010). The tree was reconstructed several times using multiple possible outgroups each time. The potential outgroups used were G-proteins sequences taken from several species, choanoflagellates, plants and fungi. The outgroup sequences had been aligned separately and profile aligned to the ingroup using MUSCLE. Fungi sequences were chosen as the best outgroup sequences as they were used in previous analyses and produced the shortest branches to the ingroup.

The tree was viewed using FigTree. The nematode G-protein sequences did not cluster with any of the main clades so they were removed. Several other sequences were removed from clades with large amounts of representation of closely related species, as well as taxa that were connected by unusually long branches to the tree. The remaining sequences were realigned using MUSCLE and the tree was reconstructed. From this new tree, the different clades of G-protein subfamilies were visible. The Transducin clade was located and the sequences within the Transducin clade were identified from the fasta file and separated into a new file.

In order to test for the most appropriate outgroup, a total of 57 Transducin sequences were selected from the full G-protein tree dataset. Four sequences each from several potential outgroups were selected from closely related regions of the full G-protein tree. Transducins are part of the inhibitory regulative family of G-proteins (Gi), therefore, all the potential outgroup sequences were taken from this family as well. Four sequences from the inhibitory regulative G-protein clade number 2 (Gi2), four basal Gi sequences and four sequences from a clade commonly known as Gz (another subfamily of the inhibitory regulative G-proteins, Gi) were selected as potential outgroups. The Transducins were aligned using MUSCLE and each of the three selected outgroups were aligned separately before being profile aligned to the Transducin ingroup. Trees were reconstructed using each of the alternative outgroups to find the best outgroup. PhyML (Guindon *et al*. 2009) was used to reconstruct the trees, using default settings and the WAG model.

3.2.2 Transducin – Testing the Clade Topology

The Transducins were found to have three main clades, Gt1 (expressed in rod cells), Gt2 (expressed in cone cells) and Gt3 (expressed in taste receptor cells, also known as Gustducin). Using the selected outgroup, Gi2, the robustness of the internal topology of the main three Transducin clades was tested. Using PhyML, two maximum likelihood phylogenies were reconstructed using the default settings and using the models WAG (Whelan and Goldman 2001) and JTT (Jones *et al.* 1992).

A bootstrap resampling of the data was also performed 1000 times using Phyml and the WAG model. These trees were summarised under a majority rule consensus procedure.

The robustness of the topology was also tested using the CAT model (Lartillot and Philippe 2004). The software Phylobayes was used to reconstruct the tree using the CAT model and a Bayesian framework. Modelgenerator (Keane *et al*. 2004) was used on the dataset which found that the JTT model, using a gamma approximation of rates was the best fitting model based on the AIC (Bozdogan 1987) and the BIC (Posada and Buckley 2004; Yang 2005) tests. The resulting trees were analyses to determine the most likely clade topology.

3.2.3 Transducin – Uncertainty of the Lamprey Position

The position of the lamprey sequences was uncertain due to the lamprey "Gt2" (known as the lamprey 'long' Transducin as a result of it being expressed in photoreceptor cells that detect relatively long wavelengths of light) sequences clustering with the Gt1 clade. To ensure this was not an artefact of LBA, another tree was reconstructed with the lamprey Gt1 sequence (known as lamprey 'short') removed to see if the lamprey "Gt2" sequence would then move to the Gt2 clade. PhyML was used to reconstruct the tree using the default parameters and the WAG model.

A reduced dataset was used for further analyses of the uncertainty of the lamprey sequence position. Four Gt3 sequences were used as an outgroup to a reduced set of 28 visual Transducins (Gt1 and Gt2) that included the lamprey sequences. Sequences were removed based on reducing the numbers of sequences from groups of closely related species. Trees were reconstructed from this dataset using the LG, JTT and C20 models of sequence evolution. PhyML was used to reconstruct these trees using the default settings but allowing for additional alteration of the topologies by using five random starting trees and by allowing both NNI and SPR changes to the tree.

A phylogenetic tree was also reconstructed from the reduced dataset using the phylogenetic tree reconstruction software programs RAxML (Stamatakis *et al*. 2005) and Leaphy (Whelan 2007). These programs use ML to reconstruct phylogenies but they use different tree alteration (branch swapping) techniques to find the most likely tree. These software programs were used in order to investigate whether they would produce an alternative to the PhyML phylogeny.

The software program Tree Independent Generation of Evolutionary Rates (TIGER) (Cummins and McInerney 2011) was also used to find the sites in the alignment that were the fastest evolving by splitting the sites up into 10 bins, each bin containing sites in the alignment that are judged to have evolved at similar rates. Bins 9 and 10 contained the fastest evolving sites in the alignment. These sites were then removed using PAUP and the new alignment, containing more slowly-evolving sites, was used to reconstruct a tree using the PhyML software and the LG substitution matrix, using five random starting trees. Both NNI and SPR branch swapping methods were used also to identity the most likely tree.

The position of the lamprey long sequence was unclear. In order to test the position, the trees were manually altered to force the lamprey long sequence to cluster with the Gt2 clade in order to test if this topology returned a better likelihood score. The original topology (reconstructed using PhyMl and the LG model) was used for further analysis and then it was altered manually using Mesquite (Maddison and Maddison 2001) to change the position of the lamprey long sequence to cluster with the Gt2 clade. The topology for the lamprey sequences found after the fast evolving sites were removed using TIGER was also tested by manually altering the original topology using Mesquite. All of these topologies, where the only difference was the "lamprey long" position, were executed in PhyML to determine their site likelihoods. Then, Consel (Shimodaira 2001) was used to perform paired site tests to determine if any of the topologies were significantly better at explaining the data when compared to the others.

3.2.4 Transducin – Addition of Key Sequences

A second BLAST analysis was performed using a human Gt1 sequence as the query, on a more recent NR database downloaded in April 2013. The sequences found from this analysis were aligned using MUSCLE and a tree was constructed using FastTree. This tree was used to identify any additional sequences that could be added to the previous analysis in order to increase the taxon sampling and therefore the accuracy of the results. A group of three tunicate sequences were found to be the sister taxa to the Transducins. These were selected, along with a hagfish Transducin sequence and were added to the previous dataset.

The analysis was repeated, using MUSCLE to align the sequences, Modelgenerator to identify the best fitting model and PhyML to reconstruct the trees. Two trees were reconstructed using the WAG model and the JTT model. Invariant sites were set to be estimated from the data and NNI and SPR moves were used to alter the tree. Ten random starting trees were used.

The JTT tree was altered, as before, to move the lamprey long sequence to cluster with the Gt2 sequences. Then the site likelihoods for both topologies were identified using PhyML. Then Consel was used to compare these topologies. Both topologies were selected for molecular dating.

3.2.5 Phosphodiesterase 6 – Finding the Outgroup

Initially, eleven BLAST searches were performed, using a query sequence (*Homo sapiens)* from each of the eleven subfamilies of PDE, against the indexed NR database. The blast output files were used to get the hit names and sequences. All eleven files were then combined and any duplicate sequences were removed. The remaining 2,200 sequences were aligned using MUSCLE. Then, the NCBI databases contain each genbank identification number matched to its coresponding taxonomic ID (taxid) as well as each taxonomic ID matched to each species name were downloaded. Perl scripts were written using these databases to search using the gi numbers of a selected sequence and return the corresponding taxid number and the species name. By identifying the correct species from which each sequence came, the genbank headers of the BLAST hits were altered to include the species name. This addition to the headers of the sequences permitted the automated identification of the corresponding species for each sequence, allowing for easier identification of speciation events and duplication events once the tree was reconstructed. The alignment was viewed using SeaView (Gouy *et al*. 2010). The alignment was manually curated by removing sequences that did not align well to the majority of the sequences, as well as some partial sequences. The remaining 1131 squences were realigned and the tree was reconstructed using FastTree.

3.2.6 Phosphodiesterase 6 – Confirming the Topology

The PDE6 sequences were removed from the original PDE alignment of 1131 sequences and added into another file. Four PDE5 outgroup sequences were also selected from the PDE tree to be used as the outgroup. The PDE6 sequences and the PDE5 sequences were aligned separately using MUSCLE and then aligned together using the 'profile alignment' option in MUSCLE. PhyML was used to build the tree using the default settings and the WAG model. Using this tree and the alignment as a guide, the PDE6 dataset was reduced down from 206 sequences to 144 by removing sequences from well represented closely related species or sequences that showed unusually large amounts of substitutions (suggesting possible errors) from species whose genomes have not yet been sequenced. Modelgenerator was executed using the PDE6 dataset and it confirmed that JTT was the best fitting of the available models. The robustness of the phylogenetic hypotheses was assessed using PhyML and the JTT and WAG models as well as a bootstrap resampling of the dataset followed by a summarisation by a majority rule consensus procedure.

3.2.7 CNG-Channels – Constructing the Phylogeny

Firstly, six BLAST searches were performed against the indexed NR database using one from each of the four CNG alpha subfamilies and the two CNG beta subfamilies as query sequences. Using Perl scripts, the hit names and hit sequences were taken from the result files. These four files were concatenated together and any duplicate sequences were removed, leaving 465 unique sequences. The length of each of the sequences was checked and sequences with less than 100 characters were removed. Using the previously mentioned databases, the taxanomic IDs for each sequence was found and then the full species names were extracted from the GenBank database files and added to the sequence headers. The headers were also shortened and each sequence was given a reference number. The sequences were then aligned using MUSCLE and the alignment was viewed using SeaView. Any unnecessary sequences were then removed, such as any the protostome sequences, or any sequences that did not align well or were unnecessary, such as in cases where there are many sequences from very closely related species, leaving 164 sequences.

CNG-channels are closely related to HCN voltage gated ion channels so four HCN sequences were used as an outgroup. The CNG and HCN channel sequences were aligned and profile aligned together. The alignment was then viewed using SeaView. The alignment quality was still quite poor due to the presence of long inserts in the sequences from basal deuterostome species. A tree was built from the alignment using FastTree. Another tree was reconstructed using PhyML and the WAG model, which was found to be the best fitting model by running ModelGenerator on the dataset. Using the ML software, PhyML and the WAG substitution matrix returned a tree with a better likelihood value than the FastTree reconstructed tree but some of the branch lengths were still very long and the speciation events were unclear, suggesting that large parts of the tree were incorrect.

3.2.8 CNG-Channels – Refining the Phylogeny

The dataset was manually edited by removing long indel regions using SeaView, which reduced the alignment from 5988 down to 1823 characters in length. The sequences were then realigned with MUSCLE and the tree was reconstructed using PhyML and the WAG model. Next, a bootstrap analysis was performed using Phylip by resampling the data to give 100 bootstrap replicates of the manually cleaned dataset after the protostomes and the long indels were removed. The replicates were separated into 100 different files and PhyML was used for each to build a tree using the WAG model and the default settings. Then a tree was reconstructed from the bootstrap replicates using a majority rule consensus procedure.

As manually reducing the alignment would contain biases, the software Gblocks (Castresana 2000) was also used to create an alignment with the regions containing large amounts of gaps removed. Initially, Gblocks was executed on the alignment using moderately strict parameters. The minimum number of sequences for a conserved position was set to 85. The minimum number of sequences for a flank position was set to 150. The maximum number of contiguous nonconserved positions was set to 8. The minimum length of a block was set to 10. Finally, regions of the alignment that contained some gaps were preserved as their removal would have reduced down the alignment too much. The tree was then reconstructed from this reduced alignment using PhyML under default parameters and the WAG model. Gblocks was repeated using the most flexible parameters possible in order to conserve as much of the sequence length as possible. Minimum number of sequences for a conserved position was 85, minimum number of sequences for a flank position was 85, maximum number of contiguous nonconserved positions was 32,000, minimum length of a block was 2 and gapped positions were again allowed.

To improve the alignment, profile aligning techniques to combine the sequence alignments were used. The sequences were separated into groups, the HCN outgroup, the alpha sequences, the beta sequences and the non-vertebrate deuterostomes sequences. Using MUSCLE, the alpha and beta sequences were aligned separately and then profile aligned together. Then the non-vertebrate deuterostome sequences were each profile aligned separately to the main alpha-beta alignment, one at a time. Finally the HCN outgroup was aligned and then profile aligned to the main alignment. Using this technique allows for the alignment of a sequence to a profile of the total alignment, reducing the amount of potentially incorrect indels. The tree was then reconstructed using PhyML and the LG substitution matrix. A gamma distribution was used with four rate categories. Both NNI and SPR changes were made to the tree and five random starting trees were used.

3.2.9 Molecular Dating - Transducin

The initial two Transducin trees were analysed using molecular dating techniques to identify the divergence times between the rod and cone types. The reduced dataset JTT tree with the Gt3 outgroup where both the lamprey sequences grouped together with the Gt1 clade, and the manually edited version of the same JTT tree where the lamprey long sequence was moved to group with Gt2 using Mesquite were analysed.

A Bayes Factors analysis was preformed on the dataset to find the best fitting dating model (Goodman 1999). The models *cir*, *ugam* and *ln* were each tested and the *ln* model was found to best fit the data. Both trees were dated using two parallel runs of Phylobayes. The lognormal (*ln*) dating model was used and the sequences were allowed to evolve using the birth-death process. The outgroup was specified in the out file and 8 calibration points, taken from Benton (2009), were used to calibrate certain speciation events based on the fossil record. The default softbounds were used which allowed the dating results to break the bounds of the calibrated nodes by 2.5%. The analysis was checked for convergence after 42 hours using tracecomp with a burnin of 20,000 as the trace files showed that there were over 40,000 trees recorded for each execution of Phylobayes. When convergence was reached between the parallel runs the chronogram (dated phylogeny) was reconstructed using Readdiv and a burnin of 20,000. The average date for the rod and cone divergence was recorded for both trees.

A jackknife test (Harrigan 2003) was performed on the modified JTT tree (where lamrey long was grouped with Gt2), using 50% reduced random replicates of the calibrations and ten random permutations. Originally, there were eight calibrations, so each permutation contained a random selection of four of these calibrations. These ten sets of calibrations were used for molecular dating analyses using Phylobayes, with the other parameters left unchanged. In the same manner as the previous analyses, the chronograms were reconstructed and comparisons were made between the resulting dates to determine if any calibrations were having a significantly larger impact on the results than the others.

Another two dating analyses were carried out for the Transducin sequences using an LG constructed tree and a manually altered version of this tree (altered using Mesquite) where the lamprey long sequence was moved to group with the cone type Gt2 clade. These two trees were dated as before, except this time the softbounds were set to 30% and the root was constrained using a root prior with a mean of 600 mya and a standard deviation of 100 my.

The two trees constructed with the additional taxa (three tunicates and a hagfish sequences) were also dated using Phylobayes. A new set of calibrations were taken from the fossil record to include the aditional taxa. The *ln* dating model was used and 30% softbounds were used. The analysis ran for 24 hours and convergence was tested using Tracecomp. The chronograms were reconstructed using Readdiv and a burnin of 40000.

3.2.10 Molecular Dating - Phosphodiesterase 6

Three PDE6 phylogenetic tree topologies were used in the dating analyses. These were the maximum likelihood tree that was reconstructed using WAG, the maximum likelihood tree that was reconstructed using JTT and the bootstrap consensus tree where the replicated were reconstructed using phyMl and WAG. The *cir*, *ln* and *ugam* dating models were tested for goodness of fit to the data using a Bayes Factors analysis. The Bayes Factors analysis showed that for all three phylogenetic tree topologies, the *ln* model was the best fit to the dataset. Molecular dating analyses were executed on the data using Phylobayes. The *ln* dating model and a series of calibration points from the fossil record to constrain the speciation events to certain known values were used. The birth-death process was used to describe the evolution of the sequences for each of the dating analyses. For the JTT tree analysis, softbounds were set to 20% and for the WAG tree and the bootstrap tree analyses, softbounds were set to 30%. The root prior was set to a mean of 500 my and a standard deviation of 100 my for the WAG and the bootstrap tree analyses and no root prior was set for the JTT tree analysis. Two parallel executions of Phylobayes were set up for each analysis. After 24 hours the convergence of each set of parallel runs was checked using Tracecomp. If the two executions of Phylobayes were converged the chronograms were reconstructed using Readdiv. The resulting dates found for the duplications between the rod and cone types of PDE6 from each tree were recorded and the average date was found.

3.2.11 Molecular Dating - CNG-Channels

Two CNG-channel trees were selected for molecular dating. The trees that were used were the trees constructed from the manually cleaned datast using PhyML and the WAG model and the bootstrap consensus procedure. A Bayes Factors analysis was performed on the datasets, but the CNG-channel datasets were too poorly aligned due to basal non-vertebrate deuterostomes having unusual indels in their sequences resulting in the software being unable to execute correctly. To overcome this problem, the alignment was reduced down to the most informative regions using Gblocks with the parameters set as follows. The minimum number of sequences to be included for a conserved position was 85. The minimum number of sequences for a flank position was set to 85. The maximum number of contiguous nonconserved positions was set to 32,000. The minimum length of a block was set to 2. Finally, gapped positions were allowed at all regions of the alignment. These settings were the most flexible parameters allowing for the conservation of the largest amount of the sequence possible. This reduced alignment was then analysed using a Bayes Factors analysis which showed that the best fitting dating model was *ln*. This model was presumed to be the best fit for the larger datasets also and was therefore selected as the dating model for the moleular dating analyses. Both trees were then tested with a molecular dating analysis using two parallel runs of Phylobayes. The birth-death process was used to describe the evolution of the sequences and softbounds were set to 30%. The root prior was set with a mean of 600 my and a standard deviation of 100 my. The phylobayes analyses were executed until Tracecomp showed convergence between the two runs, then the chronograms were built using Readdiv.

3.3 Results

3.3.1 Transducin – Finding the Outgroup

Initially a full G-protein tree was constructed. The Transducin clade was found in the G-protein tree (Figure 3.3) and some possible outgroups were selected from nearby sequences. The Transducin sequences were separated, along with the potential outgroup sequences and trees were constructed from the Transducins using each of the potential outgroups using PhyML and default parameters. From these trees, the sequences from the Gi2 clade were selected as the best outgroup given that they manifested the shortest branch from the outgroup to the ingroup (Figure 3.4).

3.3.2 Transducin – Testing the Clade Topology

The Transducins were found to have three main clades, Gt1, which is commonly expressed in rod cells, Gt2, which is commonly expressed in cone cells, and Gt3, which is commonly expressed in taste receptor cells. Next, the arrangement of these three clades was tested. Several models and methods were used to construct multiple trees from the data (Figure 3.5). Three of the trees showed the topology where the Gt3 taste receptor clade was the sister clade to the visual Gt1 and Gt2 clades. This topology is the most likely as any other arrangement of these clades would imply that taste receptor G-proteins arose from within a clade of visual receptor G-proteins. These trees showed the Transducins duplicating to give the taste and visual subtypes, followed by another duplication resulting in the visual rod and cone types.

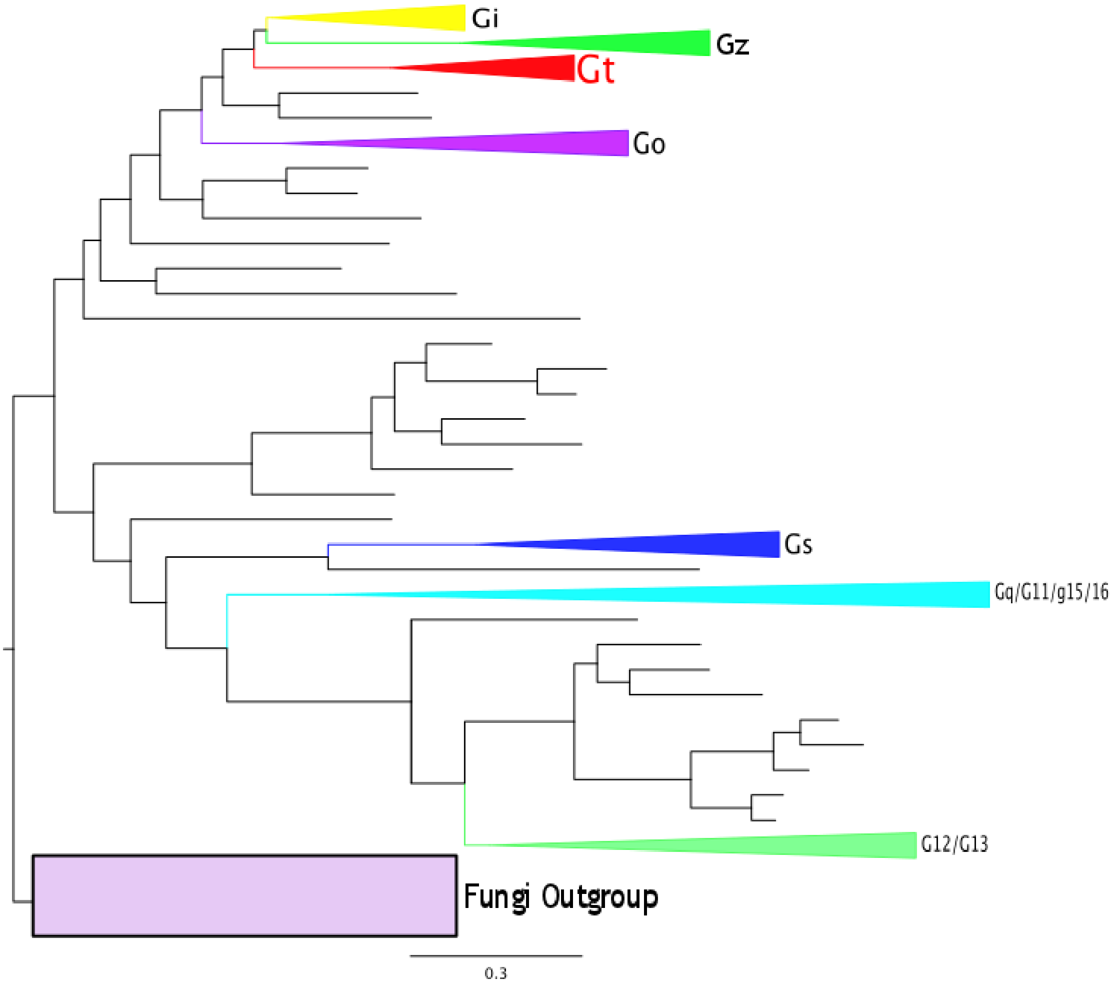


Figure 3.3: The reconstructed full G-protein tree with fungi G-protein outgroup. This tree was used to identify the Transducin (Gt clade in red) clade and identify a close outgroup. Some of the major groupings are shown as coloured labelled clades. The alignment was made using MUSCLE and the tree was constructed using FastTree. Fungi G-proteins were used as an outgroup.

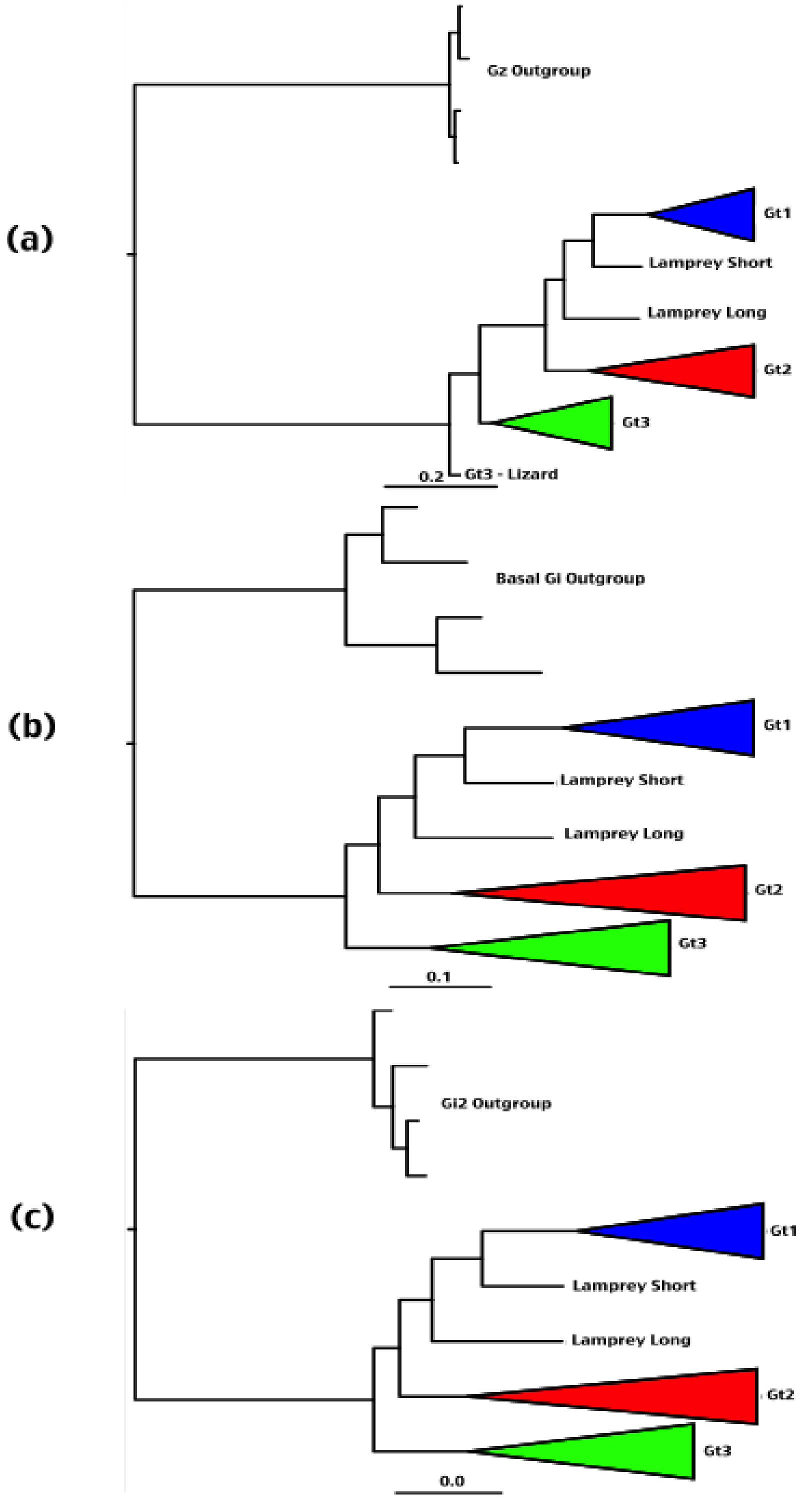
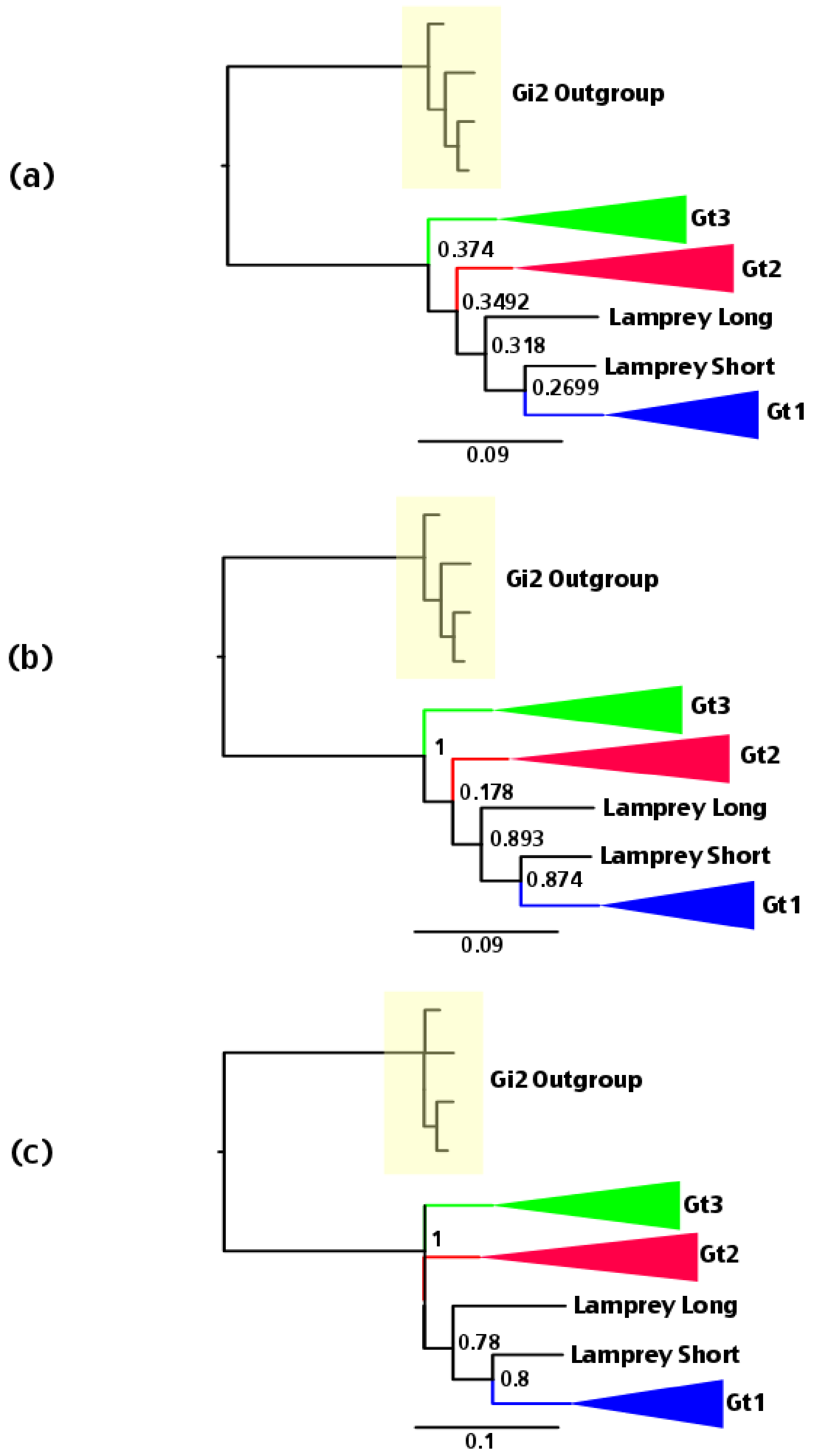


Figure 3.4: Transducin topology reconstructed using different outgroups. ((a) Gz sequences, (b) basal Gi sequences and (c) Gi2 sequences). Reconstructed using PhyML and default parameters.


(a)
Gi2 Outgroup
Gt3
0.374
Gt2
0.3492
Lamprey Long
0.318
Lamprey Short
0.2699
Gt1
0.09
(b)
Gi2 Outgroup
Gt3
1
Gt2
0.178
Lamprey Long
0.893
Lamprey Short
0.874
Gt1
0.09
(c)
Gi2 Outgroup
Gt3
1
Gt2
Lamprey Long
0.78
Lamprey Short
0.8
Gt1
0.1

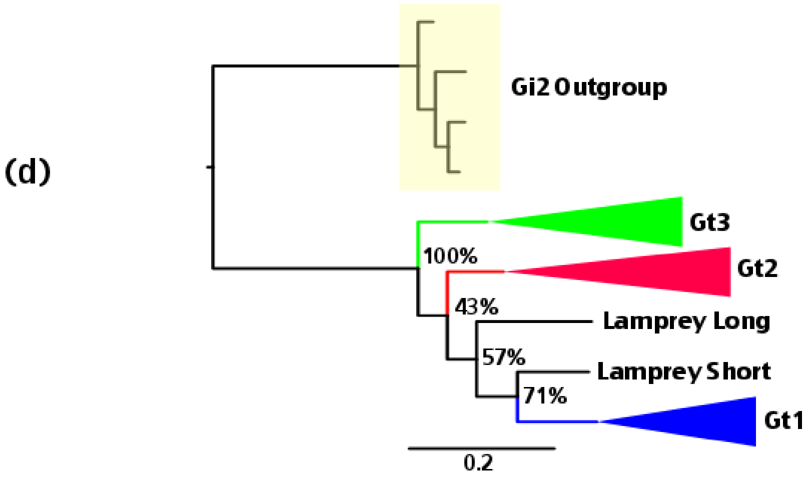


Figure 3.5: Transducin trees reconstructed using a variety of models, in order to test the robustness of the main clade topology. Tree (a) shows the Gt tree constructed using the JTT model and PhyML. Tree (b) shows the Gt tree constructed using the WAG model and PhyML. Tree (c) shows the tree constructed using the CAT model and Phylobayes. Tree (d) shows the tree constructed using a bootstrap resampling method and a majority consensus rule to produce the consensus tree. The outgroup in all four trees is Gi2 sequences. The numbers at each node represent the support for that given node. In trees a,b, and c this support value is an approximate likelihood ratio test. In tree d, the support value is a bootstrap support value. The scale bar represents the average number of substitutions per site.

The tree constructed using the CAT model and Phylobayes (Figure 3.5(c)) contains a soft polytomy at the divergence between the three clades, which suggests a lack of phylogenetically informative sites in the alignment to resolve the branching pattern. It is unlikely that it is a hard polytomy as we know from the results of the previous chapter that rhodopsin diverged from the cone opsins relatively recently, whereas taste reception is a significantly more ancient sensory method. Therefore, it is extremely unlikely that the G-proteins that are expressed primarily in these specific sensory receptors (light and taste receptors) diverged at the same time. The CAT model often requires long sequence alignments in order to resolve soft polytomies when compared to other models.

The extensive testing of the robustness of the topology confirmed that the Gt3 clade (Gustducin used in taste reception) was the sister group to Gt1 and Gt2 which are used in rod and cone phototransduction, respectively.

3.3.3 Transducin – Uncertainty of the Lamprey

Lampreys were found to have three types of Transducins. As there are three clades of Transducin (Gt1, Gt2 and Gt3), it would be expected that the lamprey genome contains one of each type of Transducin but the phylogenetic trees constructed have not always reflected this. The Transducins found in the visual receptors of lamprey are referred to as lamprey long and lamprey short rather than lamprey Gt1 and lamprey Gt2 due to their phylogenetic uncertainty. Lamprey long is expressed in the cone-like cells of the lamprey retina that maximally absorb relatively long wavelengths of light, in the red part of the light spectrum. Lamprey short is expressed in rod-like cells that maximally absorb light from relatively shorter wavelengths when compared to the lamprey long expressed photoreceptor cells. Often both lamprey visual Transducins grouped with the rod clade, Gt1, though not with significantly high support values, e.g. the bootstrap resampled majority rule consensus tree (shown in Figure 3.5(d)), constructed using the WAG model showed a bootstrap support of 57% at the node connecting the lamprey long sequence to the Gt1 clade.

To ensure this unexpected topology was not due to Long Branch Attraction (LBA), the trees were reconstructed using a dataset that did not include the lamprey short to see if the lamprey long would move to the Gt2 clade. The resulting tree (Figure 3.6) showed the lamprey long sequence as the sister taxa to a clade containing Gt1 and Gt3. As stated previously, the Gt3 clade is mostly likely the sister clade to Gt1 and Gt2, therefore, the topology of the resulting tree was likely due to a rooting error at the base of the Transducin clades. If the outgroup is ignored on this tree and the tree is considered unrooted, the lamprey long sequence is adjacent to the Gt2 sequences, as would be expected. This suggests that the lamprey long may in fact be the lamprey Gt2 sequence and that the lamprey short is the lamprey Gt1 sequence.

To further test the uncertainty of the lamprey sequence position, a reduced dataset was used to reconstruct the tree using several models, JTT, LG and C20. All three models resulted in both lamprey sequences grouping with the Gt1 clade (Figure 3.7).

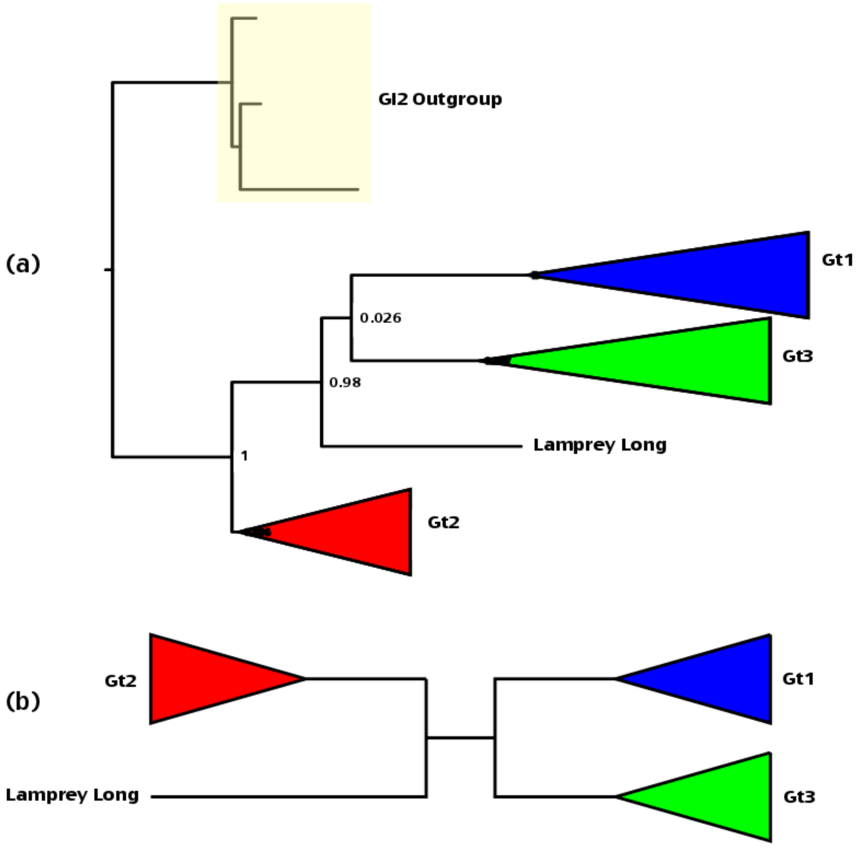


Figure 3.6: Transducin tree reconstructed with the lamprey short removed. (a) Topology found using PhyML with the lamprey short (Gt1) sequence removed and (b) the same topology unrooted. Although the support for some of the clades are high, this tree shows a very unlikely topology where Gt2 is the outgroup to the Gt1 and Gt3 clades. The support for the Gt1 and Gt3 clade is extremely low. This is most likely to be caused by a rooting problem. If the outgroup is ignored in this tree and the tree is considered unrooted, then the lamprey long sequence (Gt2) is adjacent to the Gt2 sequences, as would be expected if the lamprey long sequence was a member of cone type Transducins clade, Gt2.

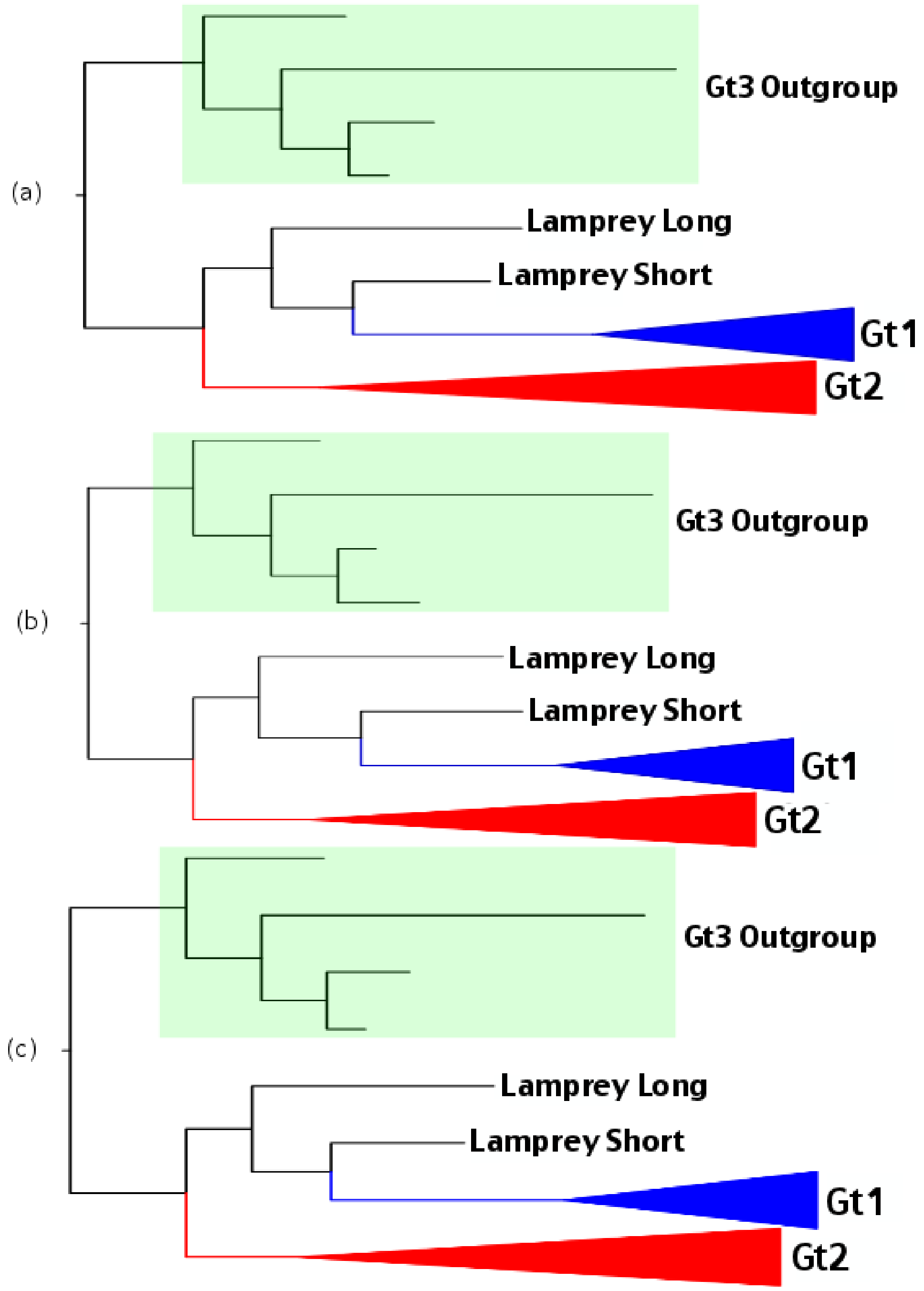


Figure 3.7: Transducin trees reconstructed using various models and a Gt3 outgroup. The trees were reconstructed using the models (a) LG, (b) JTT and (c) the fixed CAT model C20.

Several other phylogenetic software programs were used to test the lamprey position. RAxML and Leaphy were used to reconstruct the tree and the software TIGER in conjunction with PAUP was used to identify and remove the fast evolving sites. The tree was reconstructed from this reduced dataset using PhyML (Figure 3.8). The RAxML and Leaphy reconstructed trees returned the previous lamprey topology where both sequences grouped with the Gt1 clade. The tree reconstructed using PhyML after the removal of the fast evolving sites using TIGER and PAUP produced a different topology. This tree showed both lamprey sequences grouped together as a lamprey specific clade, placed as the sister group to the Gt1 clade.

Out of all of these analyses three main topologies were apparent (Figure 3.9). The first topology (Figure 3.9 (1)) shows each of the lamprey sequences group with each of the Transducin clades and this is the most parsimonious explanation, implying that there were no lamprey specific duplications or losses in addition to the two duplications that led to the three Transducin clades. The next topology (Figure 3.9 (2)) shows the lamprey long sequence as the sister taxa to the Gt1 clade (including the lamprey short), implying an additional duplication occurred where one copy was lost in all the lineages except the lamprey. The final topology (Figure 3.9 (3)) is where the two lamprey sequences cluster together as a lamprey specific suster group to the Gt1 clade, implying a lamprey specific duplication of Gt1 and a loss of the Gt2 gene.

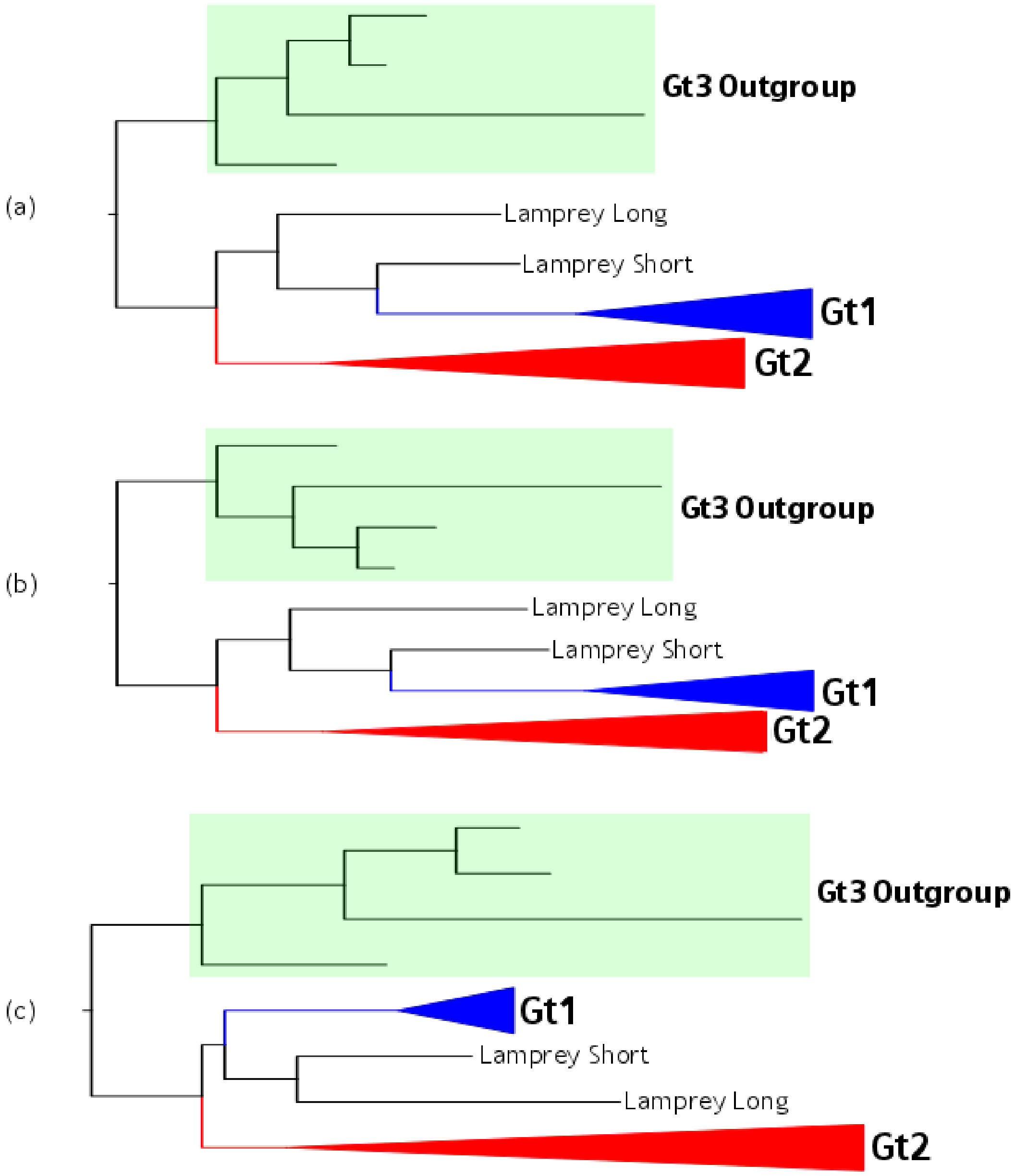


Figure 3.8: Transducin trees reconstructed using a variety of software, using (a) the ML software RAxML, (b) the ML software Leaphy, and (c) PhyML after the software TIGER was used to identify the fast evolving sites that were then removed from the alignment using PAUP. Gt3 sequences were used as an outgroup.

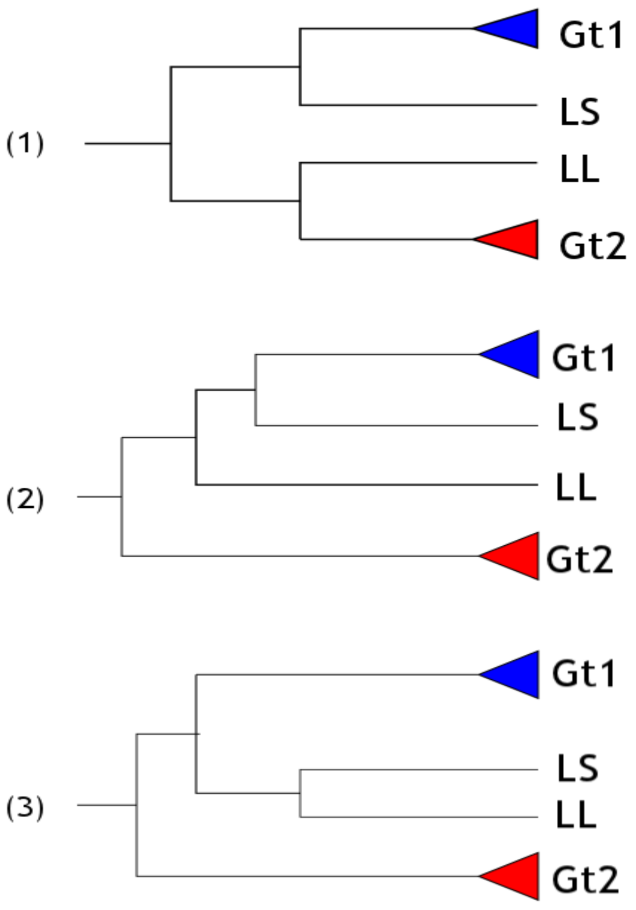


Figure 3.9: Three alternative phylogenetic tree topologies showing alternative placements of the Lamprey visual Transducin proteins. LS represents the lamprey short sequence and LL represents the lamprey long sequence. Topology 1, which places one lamprey sequence with each of the two clades, topology 2 which groups both lamprey sequences within the Gt1 clade and finally topology 3 which groups both lamprey sequences together to the exclusion of the Gt1 clade.

These topologies were all tested to see if any of them were not as good at explaining the data. The software Consel was used to perform paired site tests on the site likelihoods for each of the topologies. The results from the Consel tests are shown in Table 3.1. Although the topology where both lamprey sequences were grouped with the Gt1 clade was found the most often, the topology where each clade has one each of the lamprey sequences showed the best log likelihood value (-4043). Although Consel showed (Table 3.1) that it was not significantly better than the other topologies.

Given that Consel could not reject the sub-optimal topologies (p=0.68), two topologies were selected as possible explanations of the data for the evolution of Transducins. The selected topologies were topologies 1 and 2 from Figure 3.9, i.e. the topology where the lamprey short sequence grouped with the Gt1 clade and the lamprey long sequence grouped with the Gt2 clade. Topology 3 from Figure 3.9, where the lamprey sequences grouped together was the least likely topology and only arose from one analysis where TIGER was used to remove the fast evolving sequences; therefore, this topology was removed from further analyses.

The selected topologies were used as the input trees for molecular dating analyses to analyse the date at which the rod and cone type Transducins diverged. The JTT constructed trees and the LG constructed trees were dated (including the manually edited trees to move the lamprey long sequence to group with the Gt2 clade), i.e. four Transducin trees were dated.

Table 3.1: Consel results for three alternative Transducin topologies. Tree (item) 1 reconstructed using the JTT model in PhyML, manually edited to move the lamprey long sequence to group with Gt2 so that each of the visual Transducin clades has one lamprey sequence, tree (item) 2 where both lamprey sequences grouped with the Gt1 clade and tree (item) 3 where both lampey sequences grouped together as a sister group to Gt1. Tree 1 has the best likelihood score but the AU test could not say with a high level of significance that it is a better topology (p=0.68). Tree topologies are based on those shown in Figure 3.9.

Rank	**Item**	**Obs**	**Au-test P-values**
1	1	-4.5	0.680
2	3	4.5	0.414
3	2	6.5	0.312

3.3.4 Transducin – Additional Sequences

The addition of four new sequences to the Transducin dataset increased the taxon sampling at the basal regions of the tree where most ambiguity of the results was present. Three tunicate sequences were added which appeared to be tunicate specific Transducins that diverged before the duplications occurred. Additionally, a hagfish Transducin was added. As the hagfish, along with the lamprey, are part of the monophyletic cyclostomes (Stock and Whitt 1992; Heimberg *et al*. 2010) this sequence might add some additional information to the uncertainty of the lamprey sequence position. However, when the tree was reconstructed the lamprey long sequence still clustered with the Gt1 clade (Figure 3.10). The lamprey long sequence was then moved to the Gt2 clade by manually altering the tree using Mesquite. These trees were compared using Consel. The Consel results were inconclusive (Table 3.2) showing very little difference in the trees.

3.3.4 Phosphodiesterase 6 – Finding the Outgroup

Initially, a full Phosphodiesterase phylogenetic tree, including all eleven subfamilies, was reconstructed. This tree was used to identify the PDE6 clade and the closest outgroup, PDE5. These sequences were separated into another file and used to reconstruct PDE6 trees using a variety of methods.

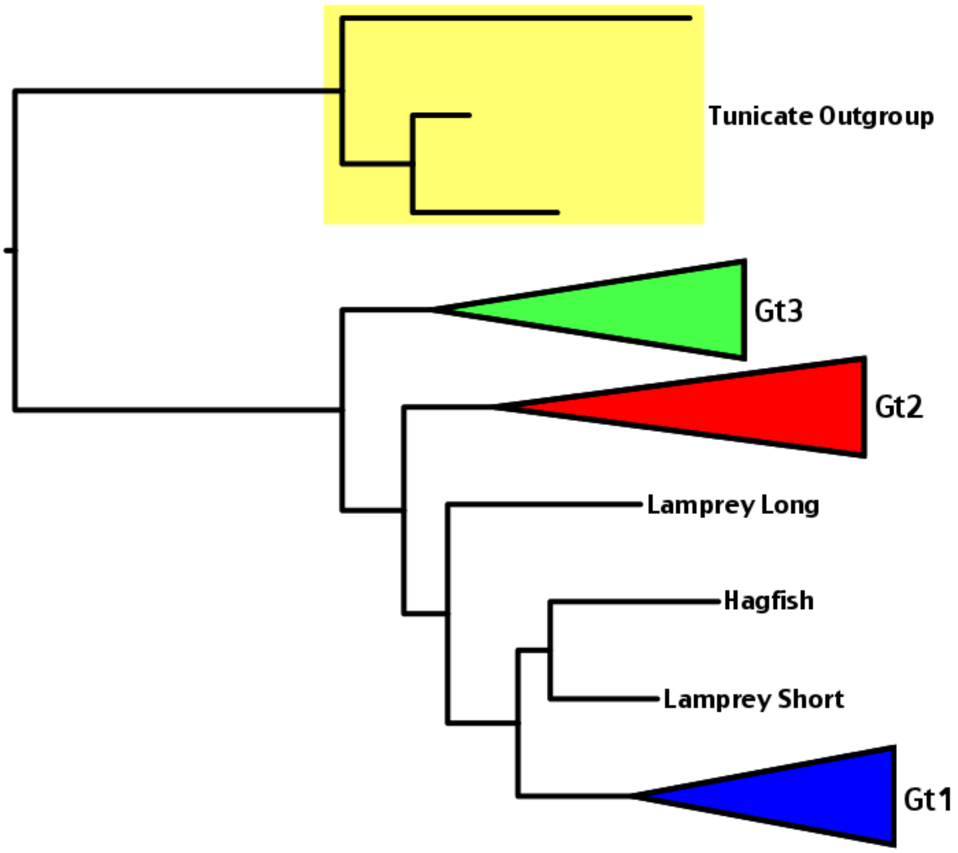


Figure 3.10: Transducin tree found when additional sequences were added. Tunicate sequences were added and used as an outgroup. A hagfish (a cyclostome closely related to the lamprey) was added also. The tree res reconstructed using PhyML and the JTT model. The topology for the lamprey sequences did not change.

Table 3.2: Consel results comparing the two topologies found when using the dataset that contained the additional taxa. The difference in the topologies is the position of the lamprey long sequence. Topology (item) 1 has the lamprey long sequence clustered with the Gt1 clade and topology (item) 2 has the lamprey long sequence clustered with the Gt2 clade. The topologies are almost identical in likelihood, showing very little difference. Therefore, one topology cannot be rejected in favour of another.

Rank	Item	Obs	Au-test P-values
1	1	-0.0	0.503
2	2	0.0	0.497

3.3.5 Phosphodiesterase 6 – Confirming the Topology

Three trees were reconstructed using the methods described in section 3.2.5. The models JTT and WAG were used along with PhyML to reconstruct the trees. A bootstrap consensus tree was also reconstructed. The PDE6 trees (Figure 3.11) showed that lamprey have only one type of PDE6 but most vertebrates, such as zebrafish, have three. The tree data did not contain any sequences from the chondrichthyes (cartilaginous fish), which would have given a more accurate representation of the evolutionary history of the PDE6 protein. From these trees we can assume that a duplication occurred after the jawed vertebrates split from the agnathans (lamprey) (as opposed to a loss in the lamprey) based on the basal position of the lamprey sequence. Although it is unclear if the duplications that led to three copies of PDE6 in bony fish and tetrapods occurred before or after the divergence of the chondrichthyes (cartilagenous fish, e.g. sharks). Molecular dating techniques allow for the identification of whether a missing species (such as the chondrichytes) diverged before or after the duplication event. This would not be possible by attempting to date duplication events with reconsiliation methods due to the lack of key species that diverged around the same time as the duplication. The same clade topology was found in each of the topologies. All of the trees found (JTT, WAG and bootstrap trees) were used in the molecular dating analyses to accurately date the divergence time between the rod and cone types of PDE6.

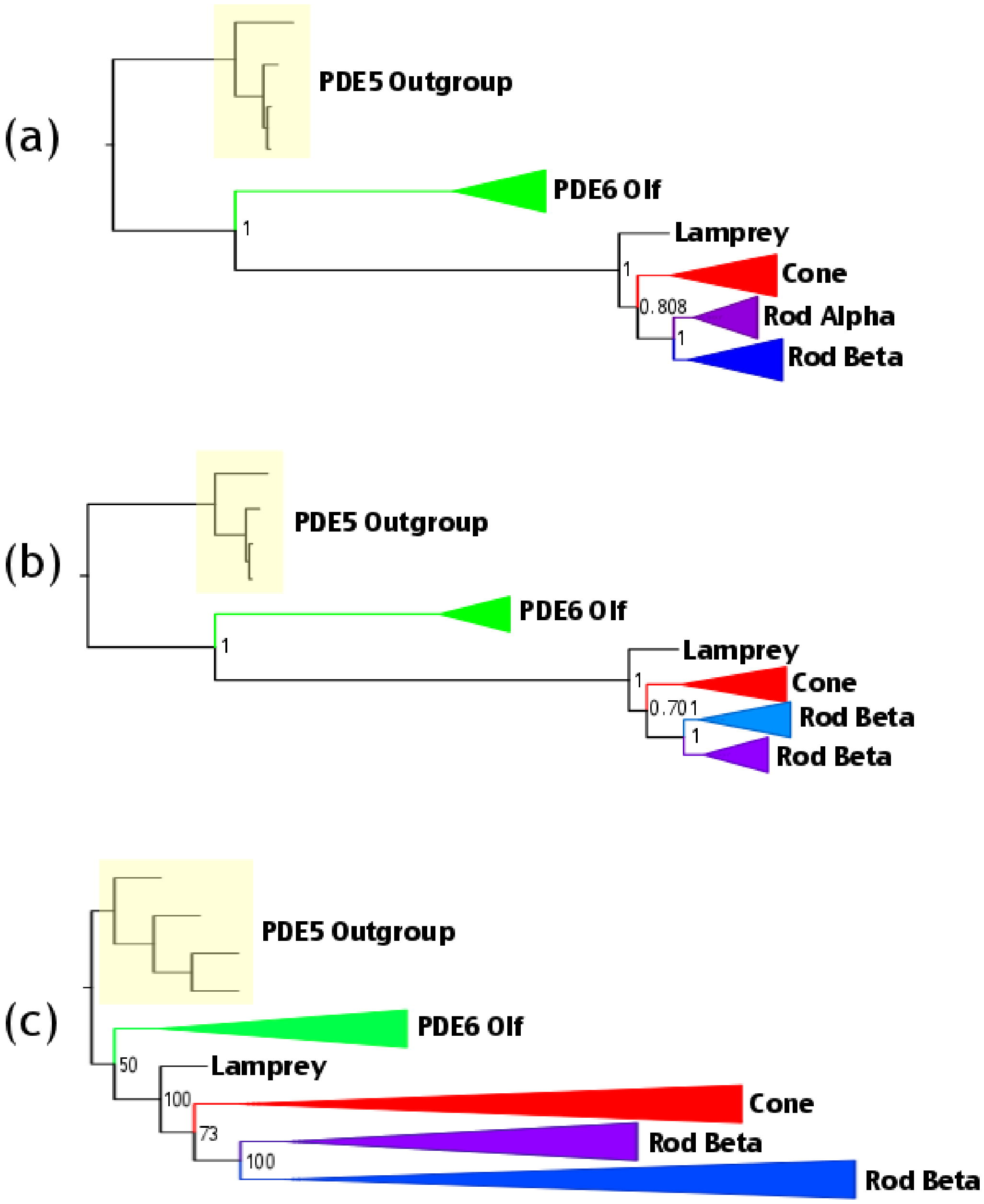


Figure 3.11: Reconstruction of the Phosphodiesterase 6 trees using (a) PhyML and the WAG model, (b) PhyML and the JTT model and (c) a bootstrap consensus tree of 100 bootstrap replicates. The outgroup used was four PDE5 sequences, which were shown to be the closest outgroup from the full PDE tree.

3.3.6 CNG-Channels

The sequence data for the CNG-channels was gathered and the trees were reconstructed using several HCN voltage gated ion channel sequences as the outgroup. The phylogenetic tree, reconstructed using PhyML and the WAG model, showed significant potential errors and unusually long branches. To fix this problem, the alignment was manually edited. The tree was again reconstructed from the reduced dataset, using PhyML and the WAG model. A bootstrap resampling analysis was also performed on the alignment and was summarised in a majority rule consensus tree (Figure 3.12). A Gblocks reduced dataset was also created from the original dataset. When flexible parameters were used, the tree reconstructed from the Gblocks reduced dataset did not show any differences in topology when compared to the tree reconstructed from the manually reduced dataset. Although, when strict Gblocks parameters were used, a different clade topology was found that split up the visual and olfactory clades. The dataset was also tested using profile alignment techniques to attempt to test the robustness of the topology. The reconstructed tree from this method also showed the same clade topology. Based on the original manually reduced topology being the most frequently found (in all trees except for one) and the other topology found by using very strict Gblocks parameters being much less parsimonious (due to the splitting up of the visual and olfactory clades), the original clade topology was selected for the molecular dating analysis. The two trees selected were the first tree that was manually cleaned of phylogenetically uninformative regions and the bootstrap consensus tree of the manually cleaned dataset.

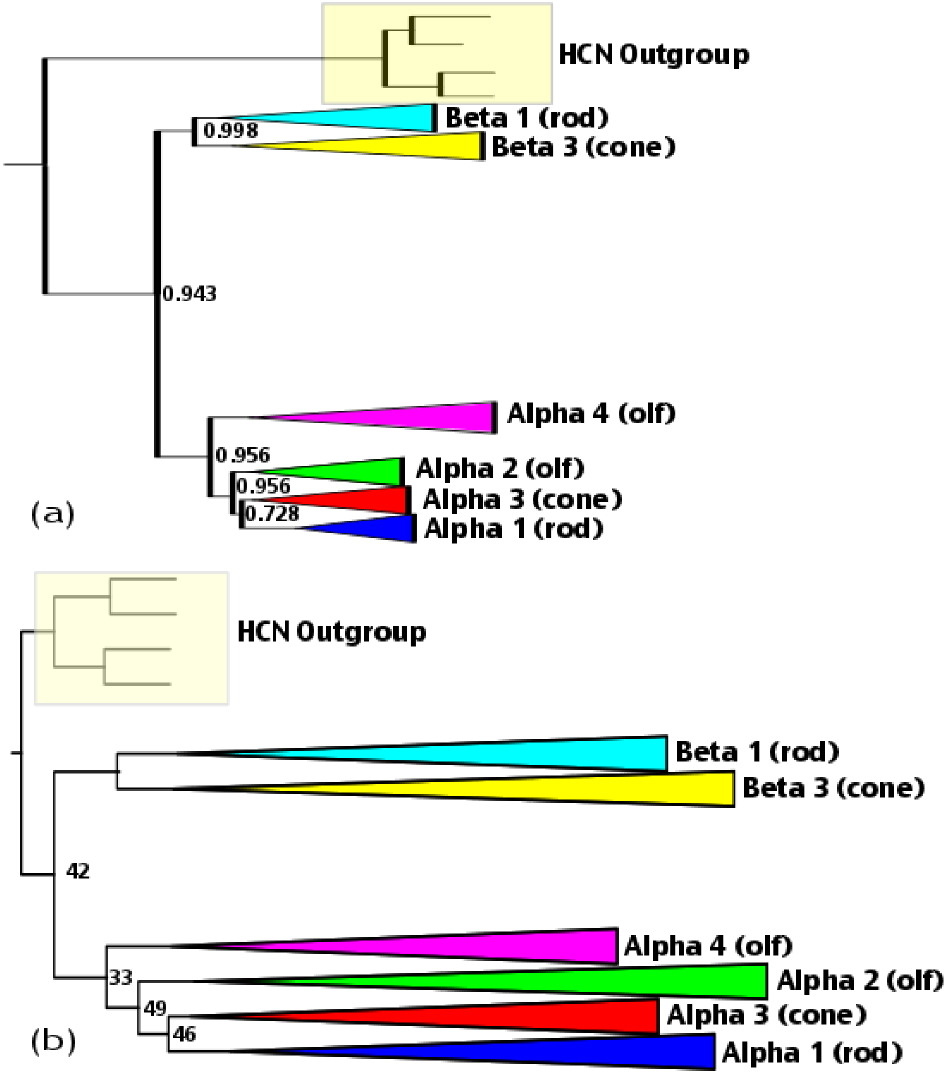


Figure 3.12: The selected CNG-channel trees reconstructed using a manually reduced dataset after the removal of regions of extreme phylogenetic uncertainty. The trees were reconstructed using (a) the WAG model in PhyML and (b) a bootstrap consensus method. The rod and cone alpha types cluster together within the olfactory clade in both trees. The beta types are a sister group to the olfactory and visual alpha clade.

3.3.7 Dating Results - Opsins

The date found for the emergence of Rhodopsin, the only opsin protein found in rod cells, was taken from the work discussed in chapter two. The split between the cone and rod types in opsin was found to be at 442 mya with a 95% confidence interval from 427 to 466 mya (Table 3.3).

3.3.8 Dating Results - Transducin

The dates found for the split between the rod and cone types of Transducin in both JTT trees reconstructed from the initial dataset was found to be 968 mya (1312-708 mya 95% confidence interval) and 1009 mya (1436 - 730 95% confidence interval). This averages to approximately 989 mya. This date predates the split between the protostomes and the deuterostomes. A BLAST search was performed to check for any protostome Transducins but there were none found. The most recent possible date based on these results, at the lower end of the confidence interval, 708 mya, predates the duplications of the other proteins in the phototransduction pathway by over 100 million years.

A jackknife test was performed to test the effects of removing calibration points on the results. The data showed that only the removal of one of the calibrations altered the resulting date for the divergence time for the rod and cone types of Tranducin. This calibration corresponded to a relatively deep node, the divergence between bony fish and tetrapods in the Gt1 clade (Table 3.4). The removal of this calibration allowed the confidence intervals for the Transducins to overlap with those of the other proteins, although this calibration is considered key and was given a relatively wide range.

Table 3.3: The dates for the divergence of the rod and cone types of each protein for each analysis performed. The opsin date (Tree1) is taken from the results of chapter two. The Transducin trees shown correspond to the original JTT constructed and manually altered trees (Tree1 and Tree2), the original LG constructed and manually altered trees (Tree3 and Tree4) and the original and manually altered tree produced from the dataset that included the additional taxa (Tree5 and Tree6). The two sets of averages for the Transducins correspond dto the average for the original dataset (Trees 1-4) and the average for the dataset that included the additional taxa (Trees 5-6). The PDE6 trees shown correspond to the WAG tree (Tree1), the JTT tree (Tree2) and the bootstrap tree (Tree3). The CNG-channel trees correspond to the manually reduced dataset tree (Tree1) and the bootstrap tree (Tree2). The values shown are in million years old, mya.

Name	**Tree1**	**Tree2**	**Tree3**	**Tree4**	**Tree5**	**Tree6**	**Average**
Opsin	442	---	---	---	---	---	442
Transducin	968	1009	703	692	559	562	843/561
PDE6	497	513	502	---	---	---	504
CNGα	518	512	---	---	---	---	515
CNGβ	455	462	---	---	---	---	459

Table 3.4: The dating results and the 95% confidence intervals for 10 sets of jackknife tests on the calibration points for the Transducin original dataset. The results showing significantly younger mean dates (2,4,6,8,10) have only one thing in common, the lack of the calibration point in the Gt1 clade that constrains the speciation event between fish and tetrapods. These results show that removal of this one calibration can force the mean date of the rod/cone split to be up to 200 million years younger. This is a key calibration point, which suggests a remarkable amount of conservation of the sequence at around this point in time, where the data would suggest that the speciation event should be much younger although the fossil record shows it to be older. Removal of this calibration point allows the confidence interval for the rod/cone type Transducin split to overlap with the dates for the emergence of the rod type in the other protein in the pathway.

Jackknife Test	Mean Date (mya)	95% Confidence Range
1	789.5	950-629
2	590	766-414
3	773	920-626
4	576.5	751-402
5	772	930-614
6	642	806-478
7	773	930-616
8	646	818-474
9	774	922-626
10	600	784-416
Average	693.6	857.7-529.5
Result Using All Calibrations	788	937-636

The dates found for the split between the rod and cone types of Transducin using the original LG reconstructed trees (including the manually altered tree that moved the lamprey long sequence to cluster with the Gt2 clade) with the constrained root prior were 703 (841-593) mya and 692 (835-584) mya (Table 3.3).

The trees constructed using the additional sequences from tunicates and lamprey were dated using Phylobayes and the resulting chronograms showed a significant difference in the rod/cone divergence dates as a result of the additional calibration point (the divergence of the tunicates and the vertebrates) applied to the root node. The divergence between the rod and cone subtypes was found to be 559 (607 - 517) mya and 562 (604 - 520) mya for the unaltered and manually altered trees, respectively, resulting in an average date of 561 mya (Table 3.3).

3.3.9 Dating Results - Phosphodiesterase 6

The date found for the split between the rod and cone types of PDE6, including the 95% confidence intervals, was 497 (540-466) mya and 513 (561-475) mya for both the WAG and JTT trees respectively. The bootstrap tree showed the rod and cone divergence time to be at 502 (546-468) mya. These dates were averaged to give 504 mya (Table 3.3).

3.3.10 Dating Results - CNG-Channels

The date found for the divergence between the rod and cone types of CNG-channels including the 95% confidence intervals were 518 mya (550-490) and 512 mya (539-482) for the alpha subunits and 455 mya (472-442) and 462 mya (487-446) for the beta subunits. This averages to 515 mya for the alpha subunits and 458.5 mya for the beta subunits (Table 3.3).

The dating results for the divergence times between the rod and cone types of each of the proteins in the phototransduction pathway are summarised and compared in Figure 3.13.

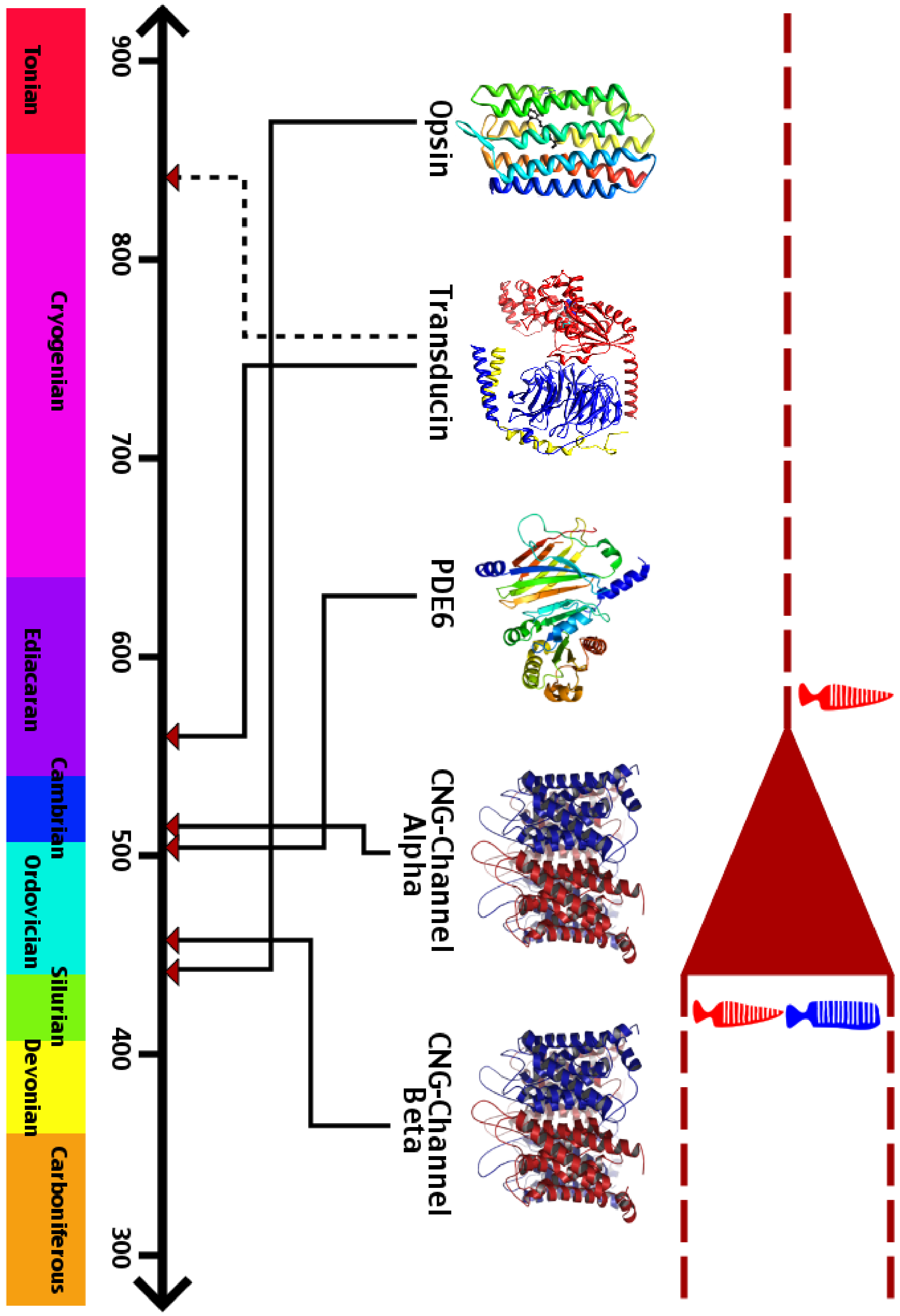

Tonian
Cryogenian
Ediacaran
Cambrian
Ordovician
Silurian
Devonian
Carboniferous
900
800
700
600
500
400
300
Opsin
Transducin
PDE6
CNG-Channel Alpha
CNG-Channel Beta

Figure 3.13: Timeline showing the point at which each protein in the phototransduction pathway duplicated to give its rod and cone subtypes. The dashed red line followed by the red triangle and two dashed red lines shows the point in time when the rod cell type arose. Each protein is shown with an arrow pointing to when it duplicated. The Transducins shows two arrows, one plain line pointing to the date found when using the dataset with the additional sequences, suggesting a co-duplication of the pathway and a second dashed line that points to the average date found for the results of the original dataset, which suggests a co-option of the Transducins from a previous function. The timeline is in millions of years (my) and the coloured bar across the bottom shows the time periods in which each duplication occurred.

3.4 Discussion

3.4.1 Co-Duplication or Co-Option?

The dates for the divergence of the rod and cone types of visual pathway proteins show that the proteins in the pathway duplicated at around the same time, 501 mya +/- 59 my (Figure 3.13).

Plachetzki and Oakley (2007) discuss the possibility that the visual pathways in rods and cones occurred as a result of co-duplication of an ancestral pathway. It is known that the cones are ancestral to the rods due to the opsins phylogenetic tree (Yokoyama 2000; Terakita 2005; Feuda *at al*. 2012), therefore the rod pathway emerged from the cone pathway by the duplication of each of the cone phototransduction proteins. To determine this result Plachetzki and Oakley (2007) used a method known as RTA (Reconciled Tree Analysis), the comparison of a gene phylogeny to a species phylogeny. The authors looked at the species present in each tree and the location of each duplication in relation to each species. Based on their trees, they came to the conclusion that the rod and cone visual pathways originated as a result of co-duplication.

Plachetzki and co-workers (2007) hypothesised that all the rod and cone proteins necessary for both cell types were present before the evolution of the first vertebrate, although their analysis was lacking in some key basal vertebrate species. The results of this chapter show that their results are not entirely correct For example, the Agnathan lamprey only possesses one type of Phosphodiesterase 6, whereas most other vertebrates have three (one cone type and two rod types). This shows that the duplications for each protein in the rod and cone phototransduction pathways had not all occurred prior to the emergence of the first vertebrate. This suggests a possible co-option solution for these phototransduction pathways rather than co-duplication and the correct evolutionary trend can only be accurately determined by the use of molecular dating techniques and calibrations taken from the fossil record.

The results of this chapter in general support the results found by Plachetzki and Oakley (2007) showing that the pathway does seem to duplicate at around the same time, with significant overlap of the confidence intervals for most of the proteins. Although, the Transducin date is somewhat older than the others (in both datasets it is the oldest date for duplication) and some of the previous analyses using the original dataset suggested it may have been co-opted from a previous function, although the evidence for either scenario is not robust enough to be conclusive. Pinpointing the exact point in time when the rod cell type arose is difficult, but it is possible that a rod-like cell may have been present relatively early, before the co-duplication of the pathway.

The mean dating results returned by the analyses are not likely to be the correct dates as the lamprey has only one type of PDE6 (compared to three in other vertebrates) but has all the vertebrate types of the other proteins (with the possible exception of the Transducins, depending on the tree topology). Therefore, this duplication likely occurred after the divergence between the Agnathans and the jawed vertebrates. The timeline should show that this duplication occurred last but it does not, although the confidence intervals do overlap significantly. This error may have been due to minor tree topology errors or rate heterogeneity in regions of the tree that could make the date seem older or younger than it is in reality. These tree topologies were relatively robust after extensive phylogenetic analysis and the confidence intervals are relatively small for each of the duplications. It seems likely that these duplications did in fact occur around the same time, although the exact order is unclear.

Certain extant species, such as the gecko, have unusual eyes. Some nocturnal geckos have a pure rod retina that expresses cone molecular pathways (Roth and Kelber 2004). As mentioned previously, cones are not normally sensitive enough to function in dimly lit conditions, but some species of gecko have modified their cones so they are expressed in a rod-like cell, allowing them to function in low light conditions. This unique ability suggests that it may have been possible that early rod cells lacked most or all of the rod type proteins but still functioned in dim light. Therefore, the rod cell may have evolved before the co-duplication event of the pathway, resulting in the cell structure causing the selective pressure for the co-duplication and specialisation of the pathway.

3.4.2 Alternative Functions for Rod Type Tranducins

The results for the original set of Transducin trees, before the addition of the tunicate and hagfish sequences gave a date for the rod/cone divergence much older than the other proteins. The lowest bounds of the confidence intervals did not overlap with those of the other proteins for any of the results found using the original dataset. The jackknife test showed that a single calibration point, if removed would result in the date being much younger. This calibration point corresponded to the divergence between the fish and the vertebrates. This was an essential calibration that was given a relatively wide range so it was assumed that when this calibration is present the date is more accurate. Even after the addition of the new sequences, that allowed for the calibration of the tunicate/vertebrate divergence, the date found for the Transducins was still older than the other proteins, although not by as much. This suggests that the co-option of the rod Transducin from a previous function cannot be rules out as a possibility.

It has previously been found that dim-light vision was present before the last common ancestor of the vertebrates and is present in lamprey (Pisani *et al.* 2006). However, there are some difficulties with the position of the lamprey Transducins, used for propagating the signal (Muradov *et al.* 2008). The date found in this study, using the original dataset, for the split between the rod and cone types of Transducin was much older than the other proteins. Also there was difficulty identifying the most likely tree, as there was uncertainty on the position of the lamprey rod and cone types. Lampreys have two types of photoreceptor cells, a pure cone type and a type that expresses both rhodopsin and cone opsins. This suggests that the rod type had not specialised enough yet to be completely functionally different from the ancestral cone type. It would be expected that the date for the rod/cone duplication in the Transducins would therefore occur not long before the last common ancestor of the lamprey and jawed vertebrates. The result found, using the older dataset, suggested that the date may have been even older than the divergence of the deuterostomes and the protostomes. Protostome genomes were analysed for the presence of Transducin-like proteins, but none were found. This date is likely to be incorrect due to rapid accelerated evolution of the Transducins around the time of the emergence of the lamprey, in addition to insufficient phylogenetic signal as a result of short sequence length. The addition of the tunicate Transducins allows for the constraining of the date to below the divergence of the vertebrates and the tunicates, assuming the node that corresponds to the split between the tunicates and the other vertebrate Transducins is a speciation event and not an older duplication event.

The Transducins also contain the basal clade Gustducin (Gt3), used for taste signal propogation. It has been shown that Gt1 (the rod type Transducin) is sometimes expressed and functional in umami taste receptors (He *et al*. 2004). It has been found that in the lizard parietal eye, Gt3 (Gustducin), rather than the usual visual G-proteins (Gt1 or Gt2) is expressed and therefore may be involved in the non-visual photostransduction cascade. This suggests a very close relationship between the visual and taste senses. It seems possible from these results that previous to its use in vision, the visual Transducins were used for taste reception and were then co-opted into the visual pathway. This would explain the duplication that led to the emergence of the rod type Transducin occurring earlier than the rest of the pathway. It is possible that the rod type Transducin had a previous function as a taste signal propogating G-protein before being co-opted into its current function as a rod visual receptor G-protein.

3.4.3 Interrelated Sensory Evolution

It is clear from the phylogenetic trees of the visual pathway proteins that many of the chemical sensory pathways are very closely related. This is the group of sensory systems that includes taste, smell and vision (vision uses a light sensitive ligand, the chromophore). The Transducins contain three subtypes commonly used for both vision and taste reception. The CNG-channel alpha clade contains groups of olfactory CNG-channels, as well as the bi-functional beta clade being used in both vision and olfaction. GPCRs, a group of transmembrane proteins that include the opsins, are receptor proteins that have a very wide range of functions, such as taste and smell. The olfactory receptor proteins, most taste receptor proteins and the visual opsins are part of a group of GPCRs known as Rhodopsin-like, or class A type GPCRs. Many of the chemical ligand binding receptor senses are very closely related and have similar proteins in their pathways, such as the Transducins and Gustducin (Fredriksson *et al*. 2003). This all suggests that the origination of the visual pathway may have been as a result of a co-option of the olfactory and taste transduction pathways already present. Other studies suggest possible older origins for the proteins involved in the visual phototransduction pathways. Plachetzki *et al*. (2010) use ancestral state reconstruction to support their hypothesis that CNG-channels were functional in the ancestral phototransduction pathway. They conclude that basal Metazoa such as the cnidarians had CNG-channel based phototransduction pathways, such as in the deuterostomes, which later swapped to TRPC channels in the protostomes. It can be easily seen from appraisal of the phylogenetic trees alone, that the "chemical" senses, olfaction, gustation and vision are extremely closely related and likely massively influenced each other's evolution.

3.5 Conclusion

There are two conclusions to be made from the analyses presented in this chapter. Firstly, it has previously been suggested that the rod visual pathway emerged as a result of co-duplication of the cone pathway (Plachetzki and Oakley 2007). Using more accurate methods of determining when and how the duplications of each protein in the phototransduction activation pathway occurred, it was determined that although there is some evidence to suggest the pathway has co-duplicated, it is not robust due to the ambiguity of the Transducins. The results found for the divergence of the rod and cone types of Transducins suggest the possibility that the rod Transducin may have had a previously unknown function in taste reception before being co-opted into the rod visual pathway, although further analyses are required to determine the certainty of either hypothesis.

Secondly, it may not be possible to consider the selective pressures on a single sensory pathway as a stand alone pathway. There are many interrelated and bi-functional proteins (such as Transducin/Gustducin and CNG-channel beta subunits) in sensory pathways that each have different selective pressures shaping their evolution. Therefore, when considering the evolution of a sensory pathway, each protein's potential multifunctionality must be considered as well.

Chapter 4 – Patterns of Duplication in Large Sensory Protein Families and the Implications for Niche Occupation by Certain Species: Analysis of Vertebrate Olfaction and Bitter Taste Reception

In this chapter the patterns of evolution of large sensory protein families are studied, in regards to numbers of duplications, likelihood of duplication, bursts of duplications and total numbers of retained receptors over time, across the vertebrates. The protein families used for the following analyses are olfactory receptors (ORs) and bitter taste receptors (T2Rs) as these are highly duplicated families.

4.1 Introduction

4.1.1 Olfaction and Gustation

Olfactory and bitter taste receptors function as chemical receptors to allow organisms to detect airborne or water soluble (olfaction) chemicals or to detect chemicals to assess the palatability of certain food sources (bitter taste reception). Both receptor types are GPCRs that react to the binding of certain chemicals or chemical features, such as a particular amino acid (Mombaerts 1999; Satoh 2005). Having a diversity of these receptors allows for the organism to obtain more information about its environment.

Olfactory receptors (ORs) are used for a wide variety of functions such as finding mates or prey, avoiding predators, finding food sources or for navigation in migrating animals (Wisby and Hasler 1954; Weissburg and Zimmer-Faust 1994; Pureswaran and Poland 2009). ORs are known to detect odors via a combination of broadly tuned receptors (Malnic *et al*. 1999). Bitter taste receptors (T2Rs) are used for the identification of bitter compounds in food. T2Rs are often an animal's only mechanism of identifying potentially poisonous or spoiled food sources.

Both T2Rs and ORs are relatively large and varied gene families due to repeated gene duplication and deletion, a process known as birth and death evolution (Nei and Rooney 2005). Copy number variations are also common within species. This process of genomic drift (random changes in genome size) has been associated with chemosensory receptor genes (CRs) by Nozawa and Nei (2008). Young *et al*. (2008) showed that although OR copy number variation between species was adaptive, copy number variation within species was a neutral process. Nei *et al*. (2008) describe several examples of adaptive copy number variation e.g. the significant expansion of OR genes in the opossum lineage.

Bitter taste receptors are the largest protein family of the gustatory receptors, with numbers of receptors varying from ~3 up to ~50 depending on the vertebrate species (Shi *et al*. 2003). Of the gustatory senses, bitter taste reception is unique in having such varied large numbers of receptors as most other gustation receptor families have only a small few receptors (Bachmanov and Beauchamp 2007). For example, sweet and umami taste reception make up only three receptor genes in mammals. This small number of receptors is likely due to there not being much of a selective advantage to being able to distinguish various sweet compounds. Many bitter compounds are toxic so it would be of great benefit to the animal to be able to identify as many as possible. It has been shown that T2Rs are under positive diversifying selection to allow the animal to recognise a wide variety of potentially poisonous substances. An animal capable of detecting a wider variety of bitter tastes is less likely to ingest harmful substances and therefore has a greater fitness. The T2R family had between 30 – 70% sequence similarity between its members (Shi *et al*. 2003).

The olfactory receptors are the largest protein family in the vertebrate genome, making up to ~3% of the total genome. The total numbers of genes encoding olfactory receptors in the vertebrate genome can vary from ~100 up to ~2000 (Niimura and Nei 2007). ORs can have >40% sequence similarity among their members and between 25 – 30% sequence similarity with their closest related GPCRs (Glusman *et al*. 2001). Olfaction is a primitive sensory method as odour detecting receptors are found in both the protostomes and the deuterostomes, although their olfactory receptors share no sequence similarity (Nei *et al*. 2008). Vertebrate and arthropod olfaction is reviewed by Kaupp (2010). Vertebrate ORs can be divided into two major classes, Class I (fish-like receptors) and Class II (tetrapod specific receptors) (Shi and Zhang 2007). Class I receptors are used in the detection of water-soluble odours and are therefore dominant in the olfactory systems of fish and aquatic mammals (Shi and Zhang 2007). Class II receptors are used for the detection of airborne odours and diversified after the emergence of the first terrestrial tetrapods (Shi and Zhang 2007).

Olfactory and bitter taste receptors are similar in that they both lack introns, which may be due to a retrotransposition method of duplication, although a lack of introns is common in GPCRs (Brosius 1999). Olfaction and bitter taste reception require large amounts of duplication and mutation to allow for a large and varied repertoire of detectable ligands (Zhang and Firestein 2002; Fischer *et al*. 2005). This pattern can be seen in their chromosomal location, as both families tend to duplicate in tandem. Both vertebrate ORs and T2Rs are clustered together on specific regions of the chromosomes. For example, the 25 human T2Rs are located on chromosomes 5, 7 and 12 and the >1000 mouse ORs are clustered in 46 genomic locations along all chromosomes except 20 and Y (Conte *et al*. 2002; Zhang and Firestein 2002). In the ORs, most subfamilies are chromosome and cluster specific (Glusman *et al*. 2001). It has been seen that clusters of ORs are interspersed with repetitive elements. These repeat regions can cause tandem duplications which might explain some of the duplications that resulted in this massively expanded gene family (Sosinsky *et al*. 2000).

4.1.2 Selective Pressures on Sensory Proteins

A massive amount of gene duplication and loss tends to occur in sensory protein families due to the fact that a constantly changing external environment generates a constantly varied selection pressure and this, in turn, drives the evolution of these genes. For example, a duplication of a red light sensitive photoreceptor allowed new world apes to see with trichromatic vision as opposed to dichromatic vision as is the case with most mammals (Surridge *et al*. 2003). This shift to trichromatic vision reduced the apes' reliance on olfaction and vomeronasal (pheromone signalling) communication. This in turn caused a reduction in natural selection on the olfactory receptors (ORs) and on the vomeronasal receptors, allowing for significant gene loss of these receptors by pseudogenisation (the accumulation of random mutations in a gene resulting in its loss of function, although it is still recognisable as previously functional gene).

Mice and humans are examples of organisms that rely on very different sensory systems for communication and foraging. These differences demonstrate the dynamic nature of the evolution of sensory perception. Mice have >1000 functional ORs while humans have approximately 500 (Young *et al*. 2002). ORs diversified by a birth-death mechanism that was fuelled by the great diversity of odorants in the environment, requiring vast amounts of gene duplication and diversification for vertebrates to detect a large proportion of them (Freitag *et al*. 1998). Humans also have a massively reduced vomeronasal receptor repertoire, having only 4 V1Rs and no functional V2Rs, whereas mice have 165 and 61 V1Rs and V2Rs, respectively (Lane *et al*. 2002). This is likely due to humans' (and other great apes') greater reliance on visual cues for finding food and for communication rather than olfactory cues (Matsui *et al*. 2010). The vast difference in the numbers of certain sensory receptors emphasises the particular senses that each organism relies on most heavily.

4.1.3 Influence of Sensory Evolution on Animals

Various methods for sensory perception in vertebrates, such as vision, olfaction or gustation, provide a unique way for the organism to perceive its environment and can increase an organism's chances of survival by better prey or mate detection, predator avoidance or avoidance of toxins (Hansen *et al*. 2003; Mueller *et al*. 2005). This means that the underlying mechanisms for environmental detection, although they are non-lethal if lost, are under constant selective pressure to better fit the organisms' specific ecological niche, which itself is always changing. These selective pressures on the animals' sensory system can even lead to speciation events by "sensory drive" (Seehausen *et al*. 2008). For example, the cichlid species of Lake Victoria in Africa have very rapid speciation rates even though the species are not geographically isolated. The various colour vision opsin pigments between the cichlid species are spectrally tuned to optimally detect varying wavelengths of light, which correlate significantly to the colouration of the males of the species. A species where the males are red in colour tend to have opsins tuned to maximally detect longer (redder) wavelengths of the visual light spectrum. Conversely, the species with blue coloured males tend to have opsins tuned to maximally detect shorter (bluer) wavelengths of the visual spectrum. This is due to alterations in the visual sensory system allowing for a species to be more visually tuned to detect individuals from their own species rather than other species with different colours. Therefore, in the case of the many closely related cichlid species of Lake Victoria, changes to their sensory system drove their speciation (Carleton *et al*. 2005).

4.1.4 Trends in Duplication Patterns

These large sensory protein families are increasing and decreasing in size as a result of duplication and loss of genes. The change in the number of receptors is due to the sensory systems constantly reacting to its environment. As changes to the environment do not occur at a constant rate, changes to these sensory families do not occur at a constant rate.

In this chapter, changes in the rate of duplication of the ORs and the T2Rs across a vertebrate phylogeny was analysed to determine if the duplication patterns showed a general trend towards gaining more receptors over time or if certain bursts of duplication were apparent in certain lineages as a result of selection to a particular ecological niche. In order to do this, the number of receptors in several vertebrate species were found and counted. Then the lineages on which each duplication occurred were found. These patterns were analysed extensively for significant changes as a result of increased duplication rates. Then the duplication patterns across all parts of the OR and T2R sensory protein trees were compared and contrasted for each species in order to detect species-specific changes in the duplication rates.

4.2 Materials and Methods

4.2.1 Obtaining and Curating the Datasets

A total of 27 deuterostome proteomes were downloaded from Ensembl (Flicek *et al*. 2011). The proteomes were concatenated together into one large file and the dataset was reduced by removing any very long or short sequences that were not likely to be GPCRs (<100 or >600 residues). Then an All vs All BLAST (Altschul *et al*. 1990) was performed on the reduced dataset of protein sequences. The BLAST output file was then filtered to only include hits where homology was found across a minimum of 70% of the query and hit sequences. Then this further reduced output file was converted to abc input format to be used in the clustering algorithm MCL (van Dongen 2007). Each line in the abc file contains the name of the query sequence, the name of the hit sequence and the e-value. This file was then analysed using MCL. Several different inflation values were used and tested for efficiency and similarity. It was found that although the highest inflation value had the greatest efficiency, all of the clusterings were very similar. Less than 1% of the edges had to be removed from one clustering to obtain the other. So in order to reduce complexity for the visualisation program Cytoscape (Shannon *et al*. 2003), the smallest inflation value was used, I 1.5.

Once the clusters were found using MCL, the sequences for each of the hits found in each cluster were placed into a file. Each file corresponded to each cluster. Next, the gene annotation information for each sequence was retrieved from the Ensembl Biomart database (Kinsella *et al*. 2011) and added to one file. The percentage of each cluster that the gene annotation information showed to be an olfactory receptor or bitter taste receptor was found and clusters with high percentages of known ORs or T2Rs were selected. The sequences for each protein in each cluster were found and put into a fastA file.

An analysis of the T2R bitter taste receptors indicated that it was possible that some fish sequences were missing. To ensure the dataset was as comprehensive as possible, a BLAST search was performed against all possible sequences from the concatenated genomes file using the sequences already found as queries. When this blast output was checked a small number of additional sequences were found and added to the dataset.

Each bitter taste and olfactory receptor sequence was used as a query sequence in a BLAST search against the NCBI NR database (http://www.ncbi.nlm.nih.gov/) that had been downloaded in January 2012 to confirm the identities of the sequences, in order to ensure that there were no additional receptors present that were not T2Rs or ORs. The top hits for each sequence was checked to ensure that each Ensembl sequence was returning a top BLAST hit from an olfactory or bitter taste receptor protein.

Certain species had unusually large numbers of receptors, therefore, further inspection of their identities was required using Biomart from the Ensembl website. For several sequences, it was found that although they had alternative protein IDs, some had the same gene and transcript IDs. Normally two sequences with the same gene IDs are alternative transcripts of the same gene but as ORs and T2Rs have no introns they have only one possible transcript.

Therefore, these alternative sequences were likely errors. They were identified from all species and removed from the dataset by identification of the "true" protein IDs using Biomart. The remaining sequences that were not found by searching using their protein IDs were checked to see if they were false alternative transcripts of previously found genes. This was tested by comparing the gene IDs, transcript IDs and protein IDs of the unknown sequences to those of the sequences that were found in the Biomart database. If they had the same gene ID as a sequence that was found then they were removed from the dataset. The remaining sequences that were not found in the Biomart database at all were ordered by gene ID. Then, arbitrarily, the first sequence with a particular gene ID was kept and the others were presumed to be the false alternative transcripts and were removed. Although this method cannot correctly identify each of the correct sequences, the sequence similarity of the false transcripts was very high so it would not likely change their position in the tree and would not affect the results.

Next, the outgroup sequences were selected in order to ensure that the OR and T2R trees were appropriately rooted. Incorrect rooting could affect the inferred number of duplications on a phylogeny. The outgroups were selected based on similarities and evolutionary distances between GPCR families from the literature (Fredriksson *et al*. 2003). For the olfactory receptors, opsins were used and for the bitter taste receptors, vomeronasal V1R receptors were used.

4.2.2 Constructing the Phylogenies

The sequences for both the ORs and the T2Rs were aligned using MUSCLE (Edgar 2004) separately to their outgroups, which were aligned and then profile aligned to the ingroup sequences. ModelGenerator (Keane *et al*. 2004) was used to determine the best-fit model for the bitter taste receptors which was the JTT substitution matrix with a gamma approximation of rates and with amino acid frequencies estimated from the data. For the olfactory receptors the numbers of sequences were too large for ModelGenerator to execute correctly, therefore a subset of receptor sequences were used and the best-fit model found was to be the JTT substitution matrix with a gamma approximation of rates and with amino acid frequencies estimated from the data.

A phylogenetic tree was reconstructed for the olfactory receptor genes using the RAxML maximum likelihood software (Stamatakis *et al*. 2005) with the fast-tree method and P-threads algorithm employed, as the extremely large number of taxa would be too computationally expensive to analyse by any other method in a reasonable amount of time. This analysis was executed using the JTT substitution model with a gamma distribution. The bitter taste receptor phylogenetic trees were reconstructed using PhyML (Guindon *et al*. 2010) with the JTT substitution model, a gamma distribution with 4 categories, empirical amino acid frequencies, using the best of NNI and SPR branch swapping, with 5 random starting trees. When the T2R tree was reconstructed, there were a number of sequences that seemed out of place on the tree, connected by a very long branch. When the identity of these sequences was checked, it was found that they were not bitter taste receptors so they were removed.

Both trees were rooted correctly using FigTree and any possible polytomies were removed using a Python ETE module method (Huerta-Cepas *et al*. 2010). The ETE module removed the polytomies by arbitrarily assigning a branching pattern at the node. Therefore, there is a significant chance that the branching pattern assigned was incorrect. However, in the absence of any alternative, this option of arbitrarily resolving polytomies was selected. The species names were added to the beginning of each sequence name to allow for easy identification. A species tree was manually constructed from each of the species in our dataset using information from the fossil record by Benton (2009). The branch lengths were calculated by taking the mean date of each of the calibration points in Benton's work. At a particular region at the base of the Eutherian mammals, the calibration points were all the same causing the mean dates to be the same. In this case the branch lengths were moved by one million years to account for the same calibration.

4.2.3 Counting the Total Numbers of Duplications per Branch

The species overlap method was then used to count the number of duplications on each branch of the species tree. This method was performed by finding nodes in the tree where a species is found either side of the bifurcating node. These nodes were recorded as duplication points and all other nodes were recorded as speciation points. This technique was used to find all the duplication points in the trees. The species present in the descendant clade from the selected node were analysed. Based on the species composition of each selected clade, the duplication was assigned to a branch in the species tree. This method was used for both trees as the species overlap method gives a more conservative number of duplications than other methods such as reconciliation. As it does not rely on the species tree, it would not be as biased by any tree reconstruction errors which are likely to occur in trees of this size and as a result of the arbitrary removal of polytomies.

Once the number of duplications per branch on the species tree for each receptor type was determined, this number was divided by the length of each branch in millions of years to calculate a lineage-specific duplication rate. Using this information for each branch, boxplots were constructed using R to detect any outliers. In other words, boxplots were used to find branches in the tree where the duplication rate was higher than expected. The default criteria for the identification of outliers in R was used such that any data point whose value was more than 1.5 times the interquartile range in distance from the lower or upper quartiles is marked as an outlier.

4.2.4 Visualisation of Regions of Expansion for Each Species

The numbers of receptors per species for both the ORs and the T2Rs were found and it was seen that certain species had massive expansions when compared to others. To determine the patterns of expansion across the tree, both trees were split up into several subtrees. A maximum number of leaves per subtree was used to account for the differences in the sizes of the trees. 25 leaves per subtree was used for the bitter taste receptors and both 200 and 650 leaves per subtree was used for the olfactory receptors. The subtrees were created by moving through the tree in a preordered fashion (each child node is visited before the parent node) in order to find the largest possible subtree where the number of taxa was fewer than the selected number of leaves (25, 200 or 650).

The numbers of receptors from each species in each subtree was recorded and visualised using stacked column charts in the Microsoft Office 2008 software Excel v12.3.6. This permitted the visualisation of the expansion or contraction of certain lineages across parts of the tree, allowing for the analysis of whether species containing large amounts of receptors expanded specific types of receptors by bursts of duplications and variation or if they expanded homogeneously across the tree with minimal losses. Stacked column charts were insufficient for visualising the olfactory receptors due to the size of the tree so the visualisation program CIRCOS (Krzywinski *et al*. 2009) was used instead.

4.3 Results

4.3.1 Dataset Statistics

27 deuterostome genomes were used in the original analysis to identify ORs and T2Rs. ORs and T2Rs were only found in the 23 vertebrate genomes used (listed in Table 4.1). No ORs or T2Rs were found in the selected four non-vertebrate deuterostomes that were included in the original analysis (*Strongylocentrotus purpuratus, Saccoglossus kowalevskii, Ciona intestinalis* and *Branchiostoma floridae*). Once the very long and short sequences were removed, only 226,733 sequences remained for analysis. After the output files from the BLAST database search were filtered to include only hits that covered at least 70% of the hit and query sequences, MCL was executed on the BLAST information. Using an inflation value of 1.5, 11,820 clusters were produced. When the clusters containing the ORs and the T2Rs were found, totals of 14,166 and 449 sequences were used for the olfactory receptors and bitter taste receptors, respectively. The average length of the ORs was 311 residues and the average length of the T2Rs was 302 residues. Both ORs and T2Rs have no introns and the T2Rs lack long C/N terminal tails, resulting in relatively short proteins.

After the outgroups were added and aligned, any additional sequences were included and the false alternative transcripts were removed, the reduced T2R alignment had 386 sequences and an alignment length of 433 amino acid residues and the olfactory receptor alignment had 12,934 sequences and an alignment length of 1,301 amino acid residues.

Table 4.1: List of species present in the dataset that were found to have ORs or T2Rs in their proteome and the corresponding numbers of each receptor type present in their genome.

Species	Name	Olfactory Receptors	Bitter Taste Receptors
Anolis carolinensis	ANC	83	2
Bos taurus	BOT	1060	20
Canis familiaris	CAF	947	15
Callithrix jacchus	CAJ	345	20
Cavia porcellus	CAP	714	16
Danio rerio	DAR	25	2
Equus caballus	EQC	885	15
Gasteroteus aculeatus	GAA	13	0
Gallus gallus	GAG	249	3
Gorilla gorilla	GOG	366	22
Homo sapiens	HOS	514	24
Loxodonta africana	LOA	1820	11
Macaca mulatta	MAM	370	27
Meleagris gallopavo	MEG	41	1
Monodelphis domestica	MOD	988	18
Mus musculus	MUM	1154	36
Pan troglodytes	PAT	349	25
Pongo abelii	POA	240	24
Rattus norvegicus	RAN	1102	36
Sus scrofa	SUS	1045	10
Taeniopygia guttata	TAG	220	3
Takifugu rubripes	TAR	10	1
Xenopus tropicalis	XET	392	52

4.3.2 Identification of Duplication Enrichment

PhyML was used to reconstruct the bitter taste receptor phylogenetic tree and RAxML was used to reconstruct the olfactory receptor phylogenetic tree. The species tree was manually constructed and the numbers of duplications for both T2Rs and ORs were counted and placed on their respective locations along the species tree branches (Figure 4.1). Rates of duplications per million years were calculated for each branch. The distance of each rate from the overall distribution of rates was determined using boxplots in R to identify outliers (Figure 4.2 and Figure 4.3). These outliers were branches of the species tree where the rate of duplication was significantly higher than the other branches. All but one of the branches that deviated significantly from the distribution of rates of duplications were located within the placental mammals (Figure 4.4). The ORs show significant increases in duplication rates in six branches, including recent species-specific enrichment in three species leaf branches, human, rat and mouse. The ORs also expanded significantly in the branches leading to the Hominoidea lineages (human-chimp divergence), the Boreoeutheria lineages (human-cow divergence) and the Zooamata lineages (dog-horse divergence). The T2Rs showed significantly high duplication rates on four branches, the branches leading to the Boreoeutheria lineages (human-cow divergence), the Zooamata lineages (dog-horse divergence), the Muridae lineages (rat-mouse divergence) and the Amniota lineages (human-lizard divergence). The first two branches have uncertain lengths due to the calibration points given by Benton (2009), therefore, they may be longer than one million year, which would eliminate the significance found.

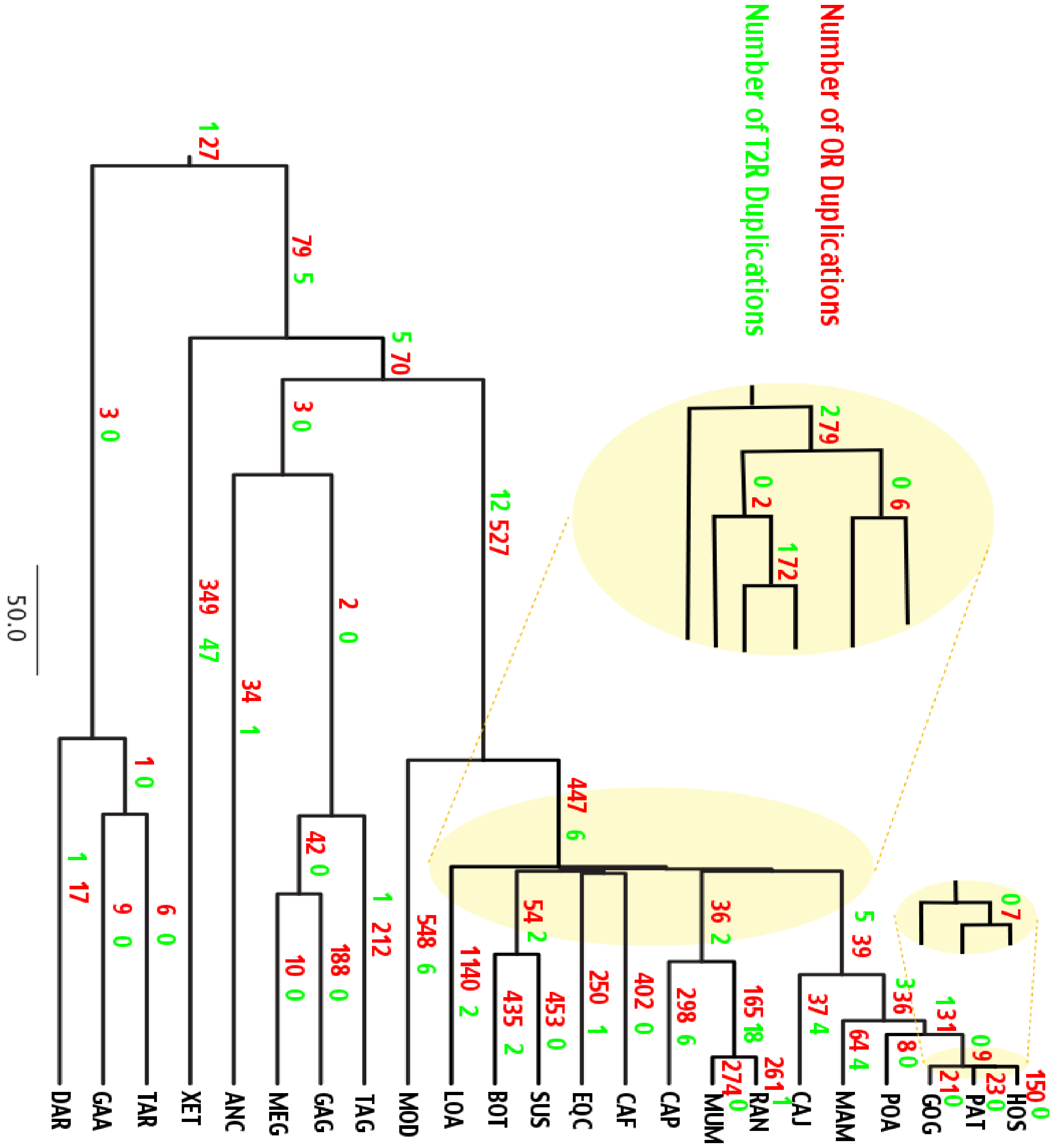


Figure 4.1: Species tree with numbers of duplications of ORs and T2Rs that occurred along each branch labelled. Duplications were inferred by the species overlap method from OR and T2R gene trees.

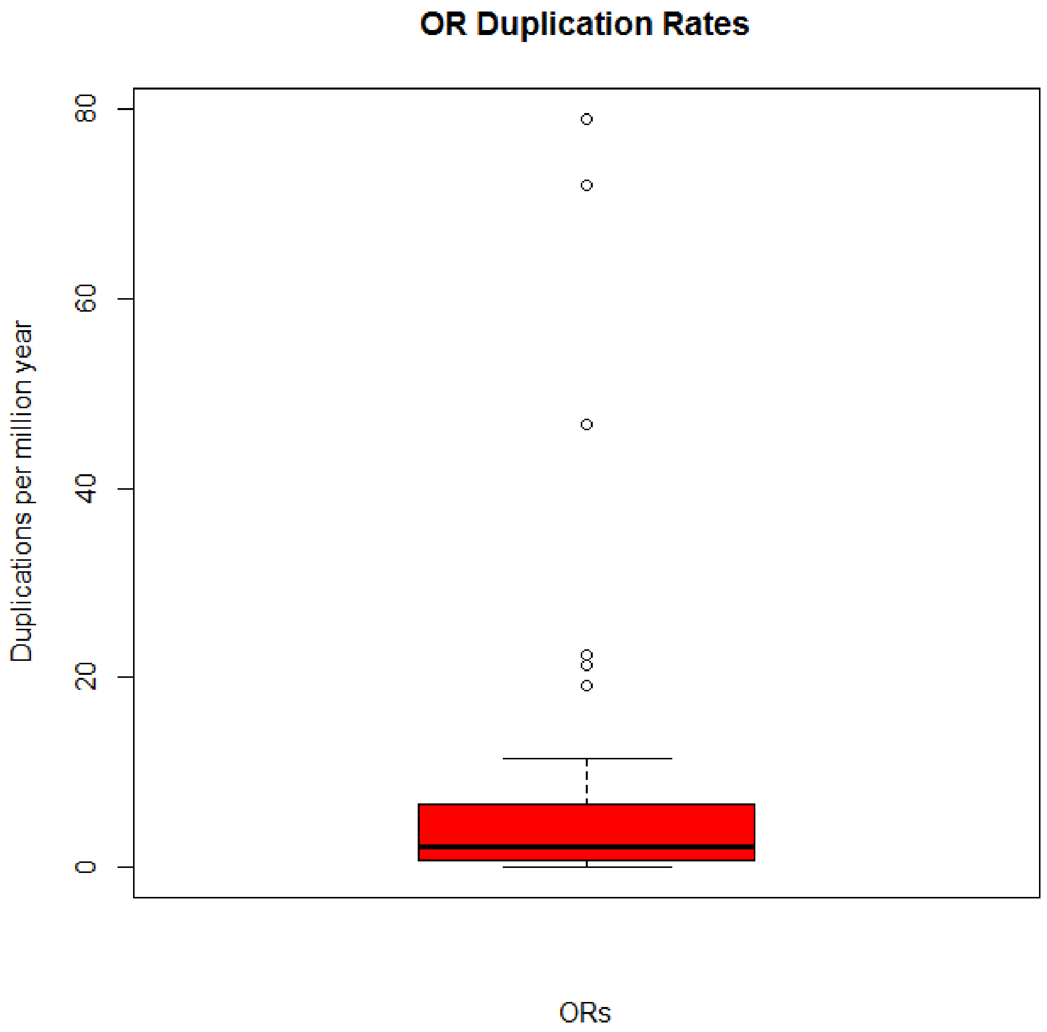


Figure 4.2: Boxplot of the distribution of the olfactory receptor (OR) duplication rates for each branch in the species tree. The red box represents the interquartile range, where 50% of the data is located around the median, represented by the thick black line. The whiskers represent the maximum and minimum data points that are located within 1.5 times the interquartile range of the lower and upper quartiles. The circles represent the data that are outside of that range, i.e. the outliers. These are the data points that represent the branches whose duplication rates are significantly higher than the others.

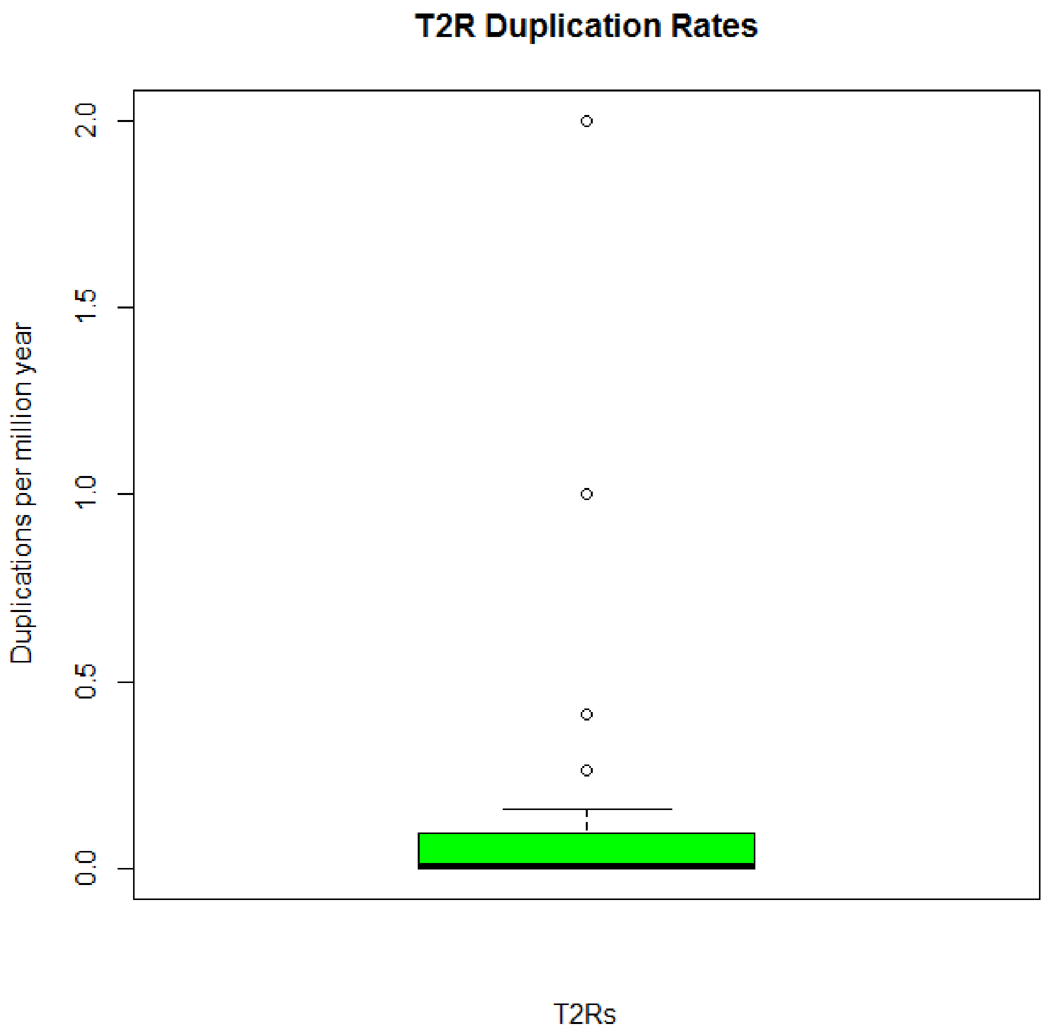


Figure 4.3: Boxplot of the distribution of the bitter taste receptor (T2R) duplication rates for each branch in the species tree. The green box represents the interquartile range, where 50% of the data are located around the median, represented by the thick black line. The whiskers represent the maximum and minimum data points that are located within 1.5 times the interquartile range of the lower and upper quartiles. The circles represent the data that are outside of that range, i.e. the outliers. These are the data points that represent the branches whose duplication rates are significantly higher than the others.

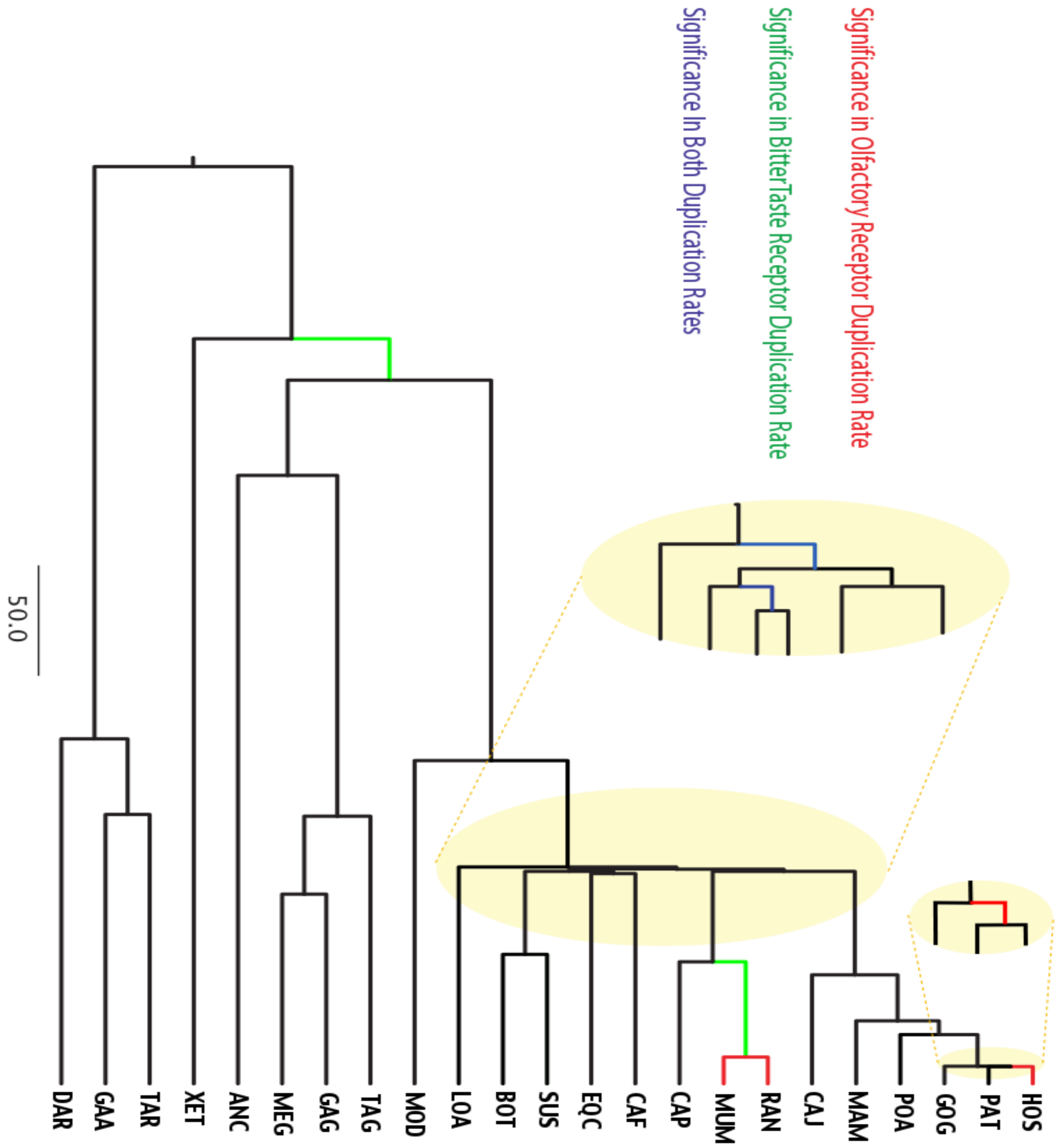

Significance in Olfactory Receptor Duplication Rate
Significance in Bitter Taste Receptor Duplication Rate
Significance In Both Duplication Rates
50.0
DAR
GAA
TAR
XET
ANC
MEG
GAG
TAG
MOD
LOA
BOT
SUS
EQC
CAF
CAP
MUM
RAN
CAJ
MAM
POA
GOG
PAT
HOS

Figure 4.4: Species tree with branches that have a significantly high rate of duplication for olfactory receptors and bitter taste receptors labelled. The branches are to scale in millions of years. The yellow circles show the branching patterns of regions of the tree where the branches are too short to be visible. The branches with colours are the ones that had significantly higher duplication rates than the others.

4.3.3 Species Receptor Enrichment Visualisation for T2Rs

The numbers of receptors in the genomes of each species was found (Table 4.1) and it was evident that certain species had much larger receptor repertoires than others and they were not necessarily species that showed significant lineage specific expansion of their receptor repertoire. To further examine the reasons behind the large amount of receptors in the genomes of certain species, the OR tree and the T2R tree were divided into subtrees and the species composition of each subtree was analysed. Large numbers of receptors in a species may have been due to massive expansion of specific subtypes by increased duplication rates or due to a relatively consistent duplication rate across all parts of the tree and minimal losses of receptor types. A maximum subtree size of 25 was allowed for the T2Rs and both 200 and 650 for the ORs, resulting in 36 T2R subtrees and 182 and 75 OR subtree, respectively.

The distribution of species within the T2R subtrees was visualised using stacked column charts in Excel. Further analysis of the species distribution across the T2R subtrees showed a number of interesting trends (Figure 4.5). The frog (XET) has 52 bitter taste receptors but they are clustered together towards the base of the tree in an amphibian specific clade. The rat (RAN) and mouse (MUM) have 36 each and Figure 4.5 shows that they are generally evenly distributed across the tree with the exception of one rodent specific clade (subtree 4 in Figure 4.5) that is massively expanded in both the rat and the mouse. The majority of species are well represented across broad regions of the tree.

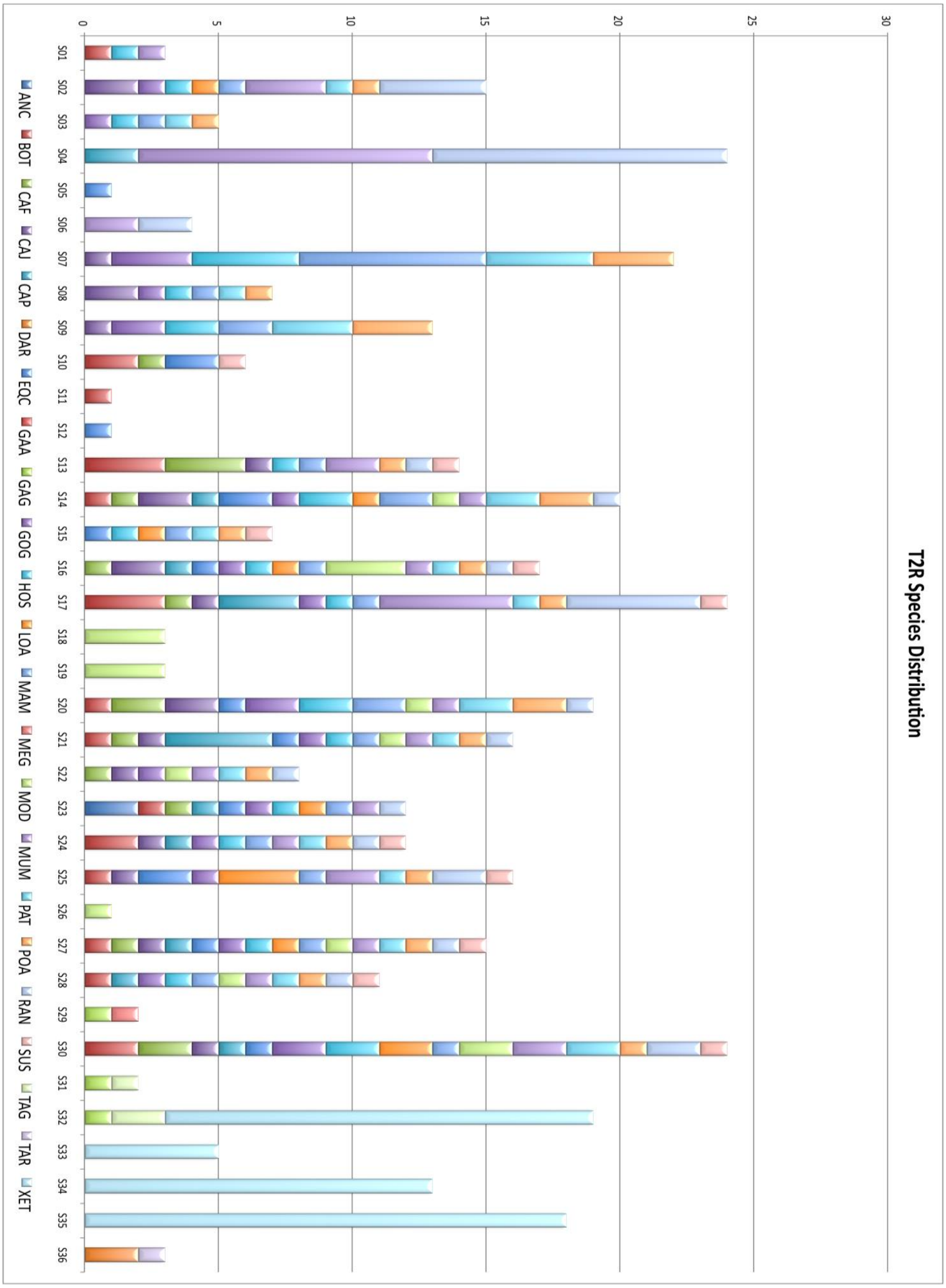


Figure 4.5: Stacked column chart showing the distribution of species among each of the subtrees of the T2R tree. Subtree number 36 is the outgroup. Each column represents a subtree and each coloured square represents the numbers of receptors found in a particular species in that subtree.

4.3.4 Visualisation Using CIRCOS for ORs

CIRCOS was used to visualise and analyse the species distribution across the many OR subtrees in both the larger (82 subtrees) and smaller (75 subtrees) datasets. Figure 4.6 shows an example of how the trees map onto the subtree sections shown in the CIRCOS diagrams, using the bitter taste receptor tree as an example. The CIRCOS images could only visualise a maximum of 100 subtrees so the 182 subtrees dataset was split up into two diagrams representing the two parts of the tree (Figure 4.7). A maximum of 650 leaves per subtree was used to reduce the number of subtrees produced to less than 100, allowing the whole tree to be represented on a single diagram of 74 subtrees (Figure 4.8). The CIRCOS images suggest that the species with unusually large numbers of ORs acquired them due to a constant duplication rate across all parts of the tree, with minimal losses. These species are all well represented in all parts of the tree with only a few having significant expansions at certain regions.

There are some regions of the tree with unusually large amounts of representation by certain species. Species such as elephant, which has the largest number of receptors among all of our selected species, clearly shows a large amount of duplications in a number of regions of the tree. Elephant specific duplications can be from the S094 and S153 in Figure 4.5. Species-specific expansions can also be seen in cows in at S035 and S105 (Figure 4.7).

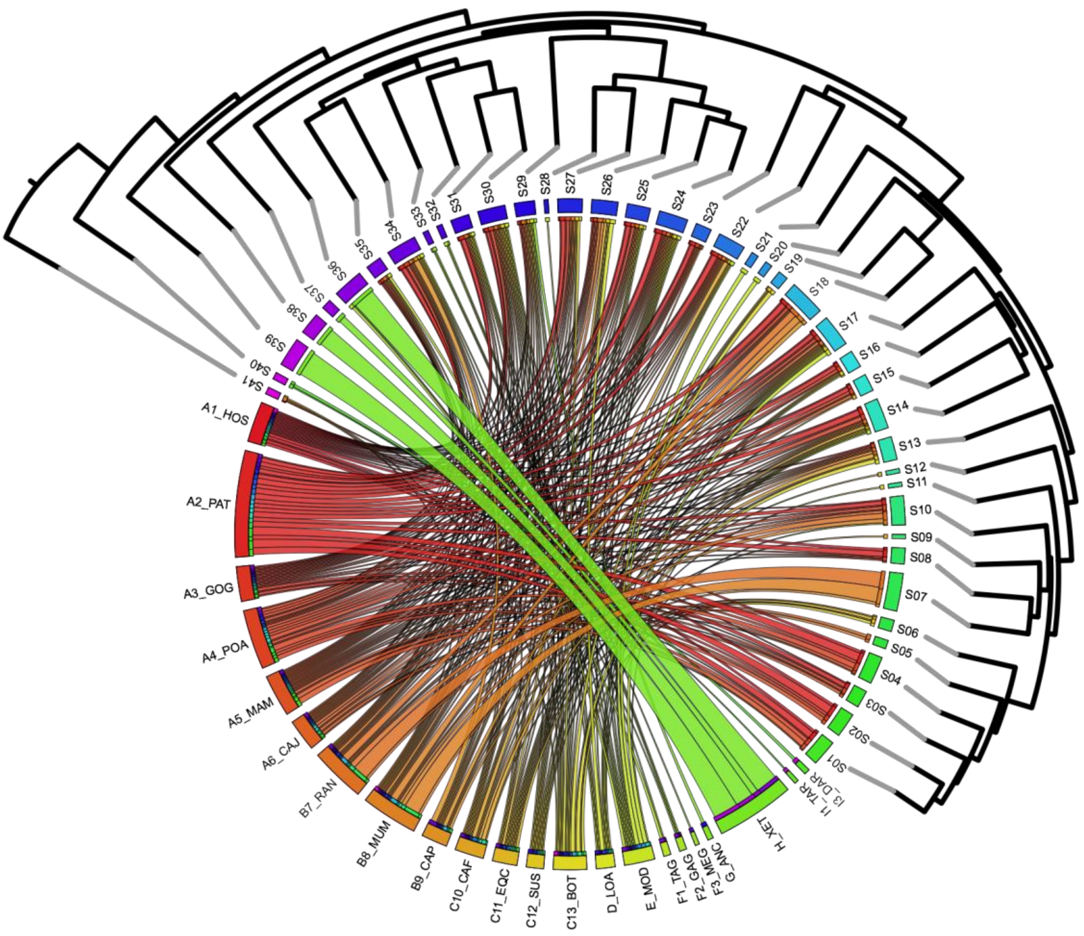
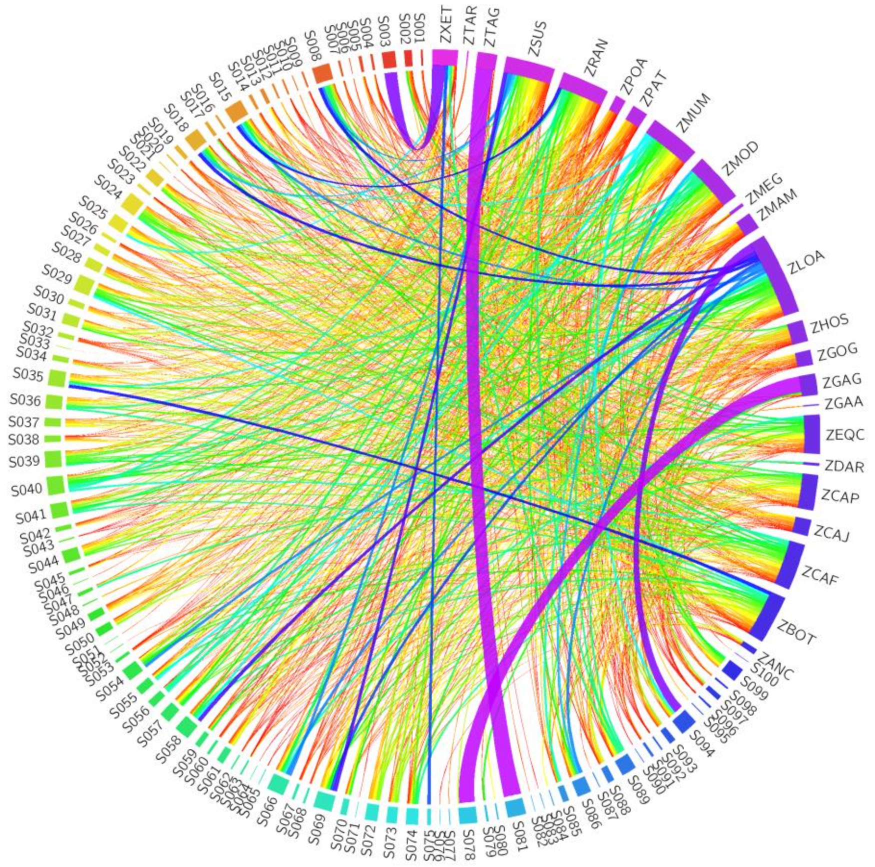


Figure 4.6: Example showing how the subtrees from the T2R tree map onto the sections of the CIRCOS diagram. The sections shown on the CIRCOS image represent clades from the connected parts of the inverted tree. The ribbons connect these clades with their corresponding species, based on representation within the clade, i.e. a clade with a large amount of cow receptors and a small amount of human receptors would have a larger ribbon connecting the clade to the cow section of the image and a smaller ribbon connecting the clade to the human section of the image. The size of each section in the image is directly proportional to the number of leaf nodes in that clade. Larger ribbons are placed on top of smaller ones in order to make regions of large species expansions clearer.


ZXET
ZTAR
ZTAG
ZSUS
ZRAN
ZPOA
ZPAT
ZMUM
ZMOD
ZMEG
ZMAM
ZLOA
ZHOS
ZGOG
ZGAG
ZGAA
ZEQC
ZDAR
ZCAP
ZCAJ
ZCAF
ZBOT
ZANC
S100
S099
S098
S097
S096
S095
S094
S093
S092
S091
S090
S089
S088
S087
S086
S085
S084
S083
S082
S081
S080
S079
S078
S077
S076
S075
S074
S073
S072
S071
S070
S069
S068
S067
S066
S065
S064
S063
S062
S061
S060
S059
S058
S057
S056
S055
S054
S053
S052
S051
S050
S049
S048
S047
S046
S045
S044
S043
S042
S041
S040
S039
S038
S037
S036
S035
S034
S033
S032
S031
S030
S029
S028
S027
S026
S025
S024
S023
S022
S021
S020
S019
S018
S017
S016
S015
S014
S013
S012
S011
S010
S009
S008
S007
S006
S005
S004
S003
S002
S001

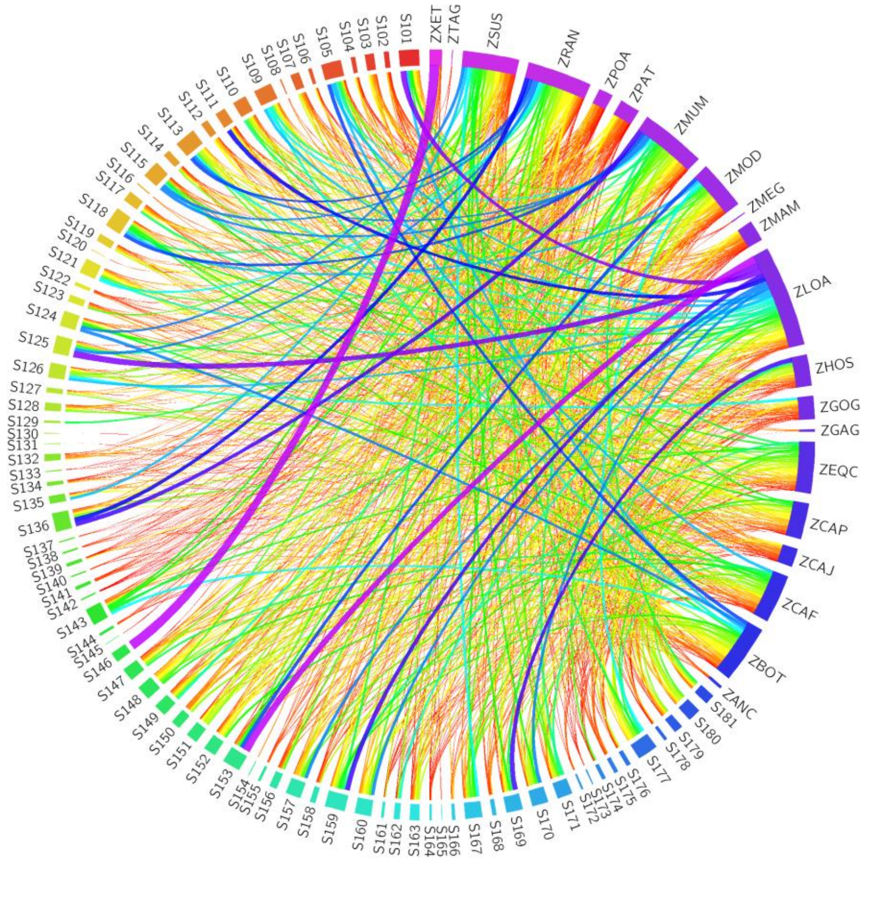

ZXET
ZTAG
ZSUS
ZRAN
ZPOA
ZPAT
ZMUM
ZMOD
ZMEG
ZMAM
ZLOA
ZHOS
ZGOG
ZGAG
ZEQC
ZCAP
ZCAJ
ZCAF
ZBOT
ZANC
S181
S180
S179
S178
S177
S176
S175
S174
S173
S172
S171
S170
S169
S168
S167
S166
S165
S164
S163
S162
S161
S160
S159
S158
S157
S156
S155
S154
S153
S152
S151
S150
S149
S148
S147
S146
S145
S144
S143
S142
S141
S140
S139
S138
S137
S136
S135
S134
S133
S132
S131
S130
S129
S128
S127
S126
S125
S124
S123
S122
S121
S120
S119
S118
S117
S116
S115
S114
S113
S112
S111
S110
S109
S108
S107
S106
S105
S104
S103
S102
S101

Figure 4.7: Olfactory receptors species composition represented by two CIRCOS diagrams of 100 and 85 subtrees each. The circle is divided into the section representing the species and the section representing the subtrees. Each ribbon connects a species to the different subtrees in which it has receptors. The larger ribbons show the subtrees where certain species are well represented. Such as the elephant (LOA) has some large expansions in S125 and S153. The birds, zebrafinch (TAG) and chicken (GAG), have large expansions on two subtrees located close together, S081 and S078, respectively. The ribbon colours blue and purple represent larger numbers of receptors, and the colours yellow green or red are smaller numbers of receptors.

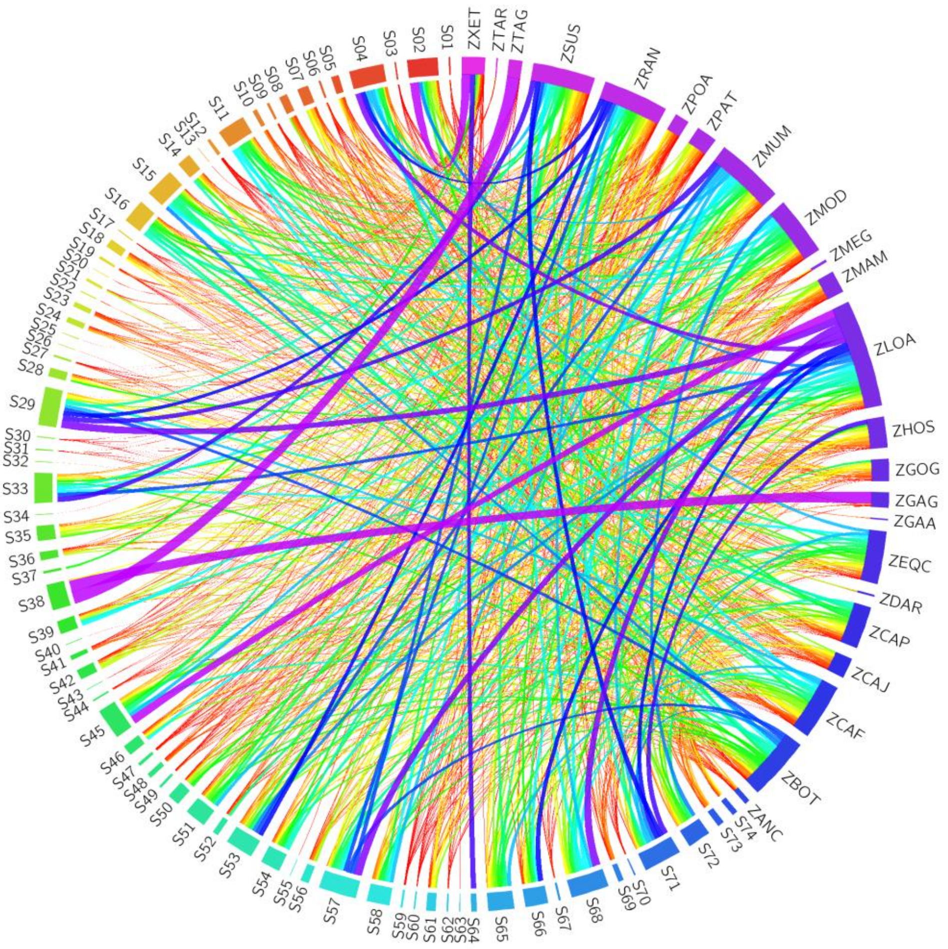


Figure 4.8: CIRCOS diagram displaying the species composition across the tree of the OR dataset, represented by 75 subtrees. The circle is divided into the section representing the species and the section representing the subtrees. Each ribbon connects a species to the different subtrees in which it has receptors. The larger ribbons show the subtrees where certain species are well represented. The birds, zebrafinch (TAG) and chicken (GAG), had large expansions on two subtrees in Figure 4.7 but the data here has been condensed and shows both those expansions on the same subtree, S38. The ribbon colours blue and purple represent larger numbers of receptors, and the colours yellow green or red are smaller numbers of receptors.

Similar to the bitter taste receptors there is a rodent specific expansion in S136 (Figure 4.7) but it is smaller in comparison to the overall tree size. In general, rodents seem to have acquired a large number of receptors by minimal losses in their evolutionary history, so their environment allowed for some selective advantage to be able to detect a large variety of odours.

Frogs have very few ORs in comparison to some of the other species, such as the rodents, but they are mostly clustered together on the tree in S003 and S146 (Figure 4.7).

This pattern of species receptor clustering is the same for the birds, chicken and zebrafinch, where their ORs are almost entirely located in one large subtree each, S081 and S078 (Figure 4.7). The clustering of the majority of the avian olfactory receptors suggests a lineage-specific expansion of a small number of ancestral ORs in birds.

Dogs and pigs also have very large OR repertoires but they do not show any unusually large expansions in specific clades in either dataset. This contrasts with what is seen in the elephant and cow genomes. The primates have smaller total numbers of ORs compared to others (cows, dogs, elephants) but they do seem to cover a wide variety of the tree. In contrast to the other primates, humans have a small species-specific expansion in S169 (Figure 4.7).

4.4 Discussion

4.4.1 Rates of Evolution and Duplication

The rate of duplication found for both the ORs and the T2Rs was very high when compared to other estimates of the likelihood of a protein to duplicate, such as that by Lynch and Conery (2000) who suggest a rate of duplication of one duplication per gene per 100 million years. From the results shown in this chapter the rate of duplication in the olfactory receptors and the bitter taste receptors is immensely larger than the expected rate of duplication found by Lynch and Conery.

The results of an analysis performed by Glusman *et al*. (2001) give a possible explanation for the unusually high duplication rate found in these chemosensory receptor families. Glusman *et al*. (2001) showed that the majority of the clusters of ORs in the human genome arose from the partial or complete duplication of two ancestral clusters of Class I and Class II ORs. Both classes of ORs were originally found on chromosome 11, prior to duplicates moving to other regions of the genome, which now contains 42% of all human ORs. This suggests that the unusually high rate of duplication found in this study may be due to large-scale segmental duplications, where multiple local genes are duplicated together. Therefore, what might appear to be multiple duplications of several genes may in fact have been due to a single large-scale duplication of a cluster of genes.

4.4.2 Changes in Receptor Repertoire as a Result of Ecology

The olfactory receptor repertoire between different organisms can be extremely varied as a result of selective pressures. Positive selection for adaptive evolution has been in shown in species such as birds (Steiger *et al*. 2010). OR repertoire sizes can have extreme variation between species (Table 4.1). It has been shown that orthologs (but not paralogs) are likely activated by the same ligand, making phylogenetic comparisons of these genes a good indicator for common identifiable ligands between species (Adipietro *et al*. 2012). This also indicates that gene duplication allows for the identification of more varied ligands. Gene gain and loss can occur extremely quickly in ORs, allowing for closely related species to differ significantly in their OR repertoire. This can be seen in humans and chimps where although the total number of ORs in their genomes are similar, approximately 20% of their functional genes are species specific (Go and Niimura 2008).

There consistently seems to be an expansion of smell and taste receptors at the base of the placental mammal clade, possible due to a change in lifestyle or diet, or due to the mammals diversifying massively at this time. The branch lengths around this point in the species tree are somewhat uncertain due to the wide calibration points used to date the species tree. Therefore, the significant increase in the duplication rates found on some of these branches might not be present if the branches in question are actually longer. However, the significant amount of speciation known to have occurred at this time suggests that there might be significance due to changes in mammalian lifestyles.

As can be seen from Figure 4.1, the rates of duplication along the branches around the base of the placental mammals are quite high. For example, the branch leading to the cow and human lineages, at the base of the Boreoeutheria, has 79 OR duplications and a rate of 79 duplications per million year due to the short branch length. The OR boxplot of duplication rate distributions (Figure 4.2) shows that a rate of approximately 20 duplications per million year or less would not be significant. The Boreoeutheria branch would have to be more than four times longer to lose its significance. As the nodes in the species tree are specifically dated, changes to the branch length of any branch would affect the branch lengths of the surrounding branches. This suggests that the presence of an unusually high duplication rate along the placental mammal radiation branches in this study may be reflecting some increase in the duplication rate but to what extent is uncertain. Uncertainty in the branch length results in uncertainty in the rate of duplication and therefore, how significant the rate is when compared to the rates found on the other branches.

When looking at the species with unusually high numbers of receptors, only the rat and the mouse have large numbers of both the ORs and the T2Rs. This might be due the scavenging nature and widespread habitats of rodents, causing there to be a selective pressure on ORs to allow for the detection of a wide variety of food sources and threats and on T2Rs to determine the potential toxicity of a possible food source.

The frog also retained high numbers of T2R receptors. This is possibly due to the frogs' production of toxic chemicals in their skin to deter predation. It seems likely that frogs have an increased number of bitter taste receptors in order to distinguish a wide variety of potentially poisonous compounds. They possibly use this heightened bitter taste response to identify members of their own species, as well as other closely related species by the chemical compounds secreted onto the skin, although this hypothesis requires further testing.

Several other species have expanded OR repertoires such as the dog, the cow, the pig and the elephant. However, they have achieved these large numbers of receptors through different means. The results of this study show that although all four of these species have receptors spanning large portions of the tree, suggesting diversity in the types of odorants these animals can detect, they are not always overlapping regions. These species have achieved niche specific receptor repertoires that differ from each other. The dog and the pig have obtained their large repertoires with little expansion of specific regions. This contrasts with the evolution of the elephant and cow repertoires which show several large expansions at specific regions, such as the cow specific expansion at S29 and the elephant expansion at S45 (Figure 4.8). These differences likely reflect alternative niche occupations such as the omnivorous diet of dogs and pigs as well as the migratory habits of elephants and cows. Both of these traits likely require a diverse olfactory receptor repertoire for identifying food sources and for navigation across large distances.

4.4.3 Evidence for Trends in Duplication Patterns

In the case of the bitter taste receptors, the distribution of most species across the tree is generally widespread (Figure 4.5), suggesting a need to preserve the ability to detect a wide variety of bitter compounds. Although, there are some exceptions to this trend such as the rodent specific clade and the primate specific clade mentioned earlier. These species-specific clades likely reflects a change in the ecological niche of the last common ancestor of these mammalian groups, requiring multiple new, specific bitter taste receptors. These results are in agreement with previous studies, such as that done by Shi *et al*. (2003) where they also found "species or lineage specific" T2Rs as well as "species general" T2Rs within the mouse and human genomes. Shi and co-workers postulated that these receptors evolved in a species-specific manner to account for niche-specific bitter tastants that a species may encounter. The authors suggest that some receptors evolve in a general species manner to account for tastants found more generally in the environment that would be encountered by a wide variety of species.

In the case of the olfactory receptors, there are clear regions of the tree that are species-specific, particularly for the more basal species (frogs and birds). These clades are due to duplications of a specific group of receptors but also losses in other groups as unlike in the bitter taste receptors, the frog and bird specific clades are not located at basal regions of the tree. If the species that are located at basal regions of the species tree have their genes located in basal regions of the gene tree, this suggests either a species specific expansion of these genes after the divergence from the other species in the tree or a loss of the MRCA of that group of genes in the ancestors of the other species in the tree. If the basal species expansions are not located at basal regions of the species tree then it suggests that there was an unusually large burst in the duplication rate of these receptors in these species and possibly that other related types of receptors were lost in the species. The presence of both species-specific and general species OR clades in the tree are supported by previous results found by Grus and Zhang (2008) who also found both multispecies clades and lineage specific clades.

For the species with unusually large numbers of ORs, their distribution across the tree shows that it is generally due to a constant duplication rate and minimal losses, although there are some regions with clade specific bursts of duplications. There is some evidence for these chemical receptors evolving as a result of bursts of duplications at certain regions of the tree. It can be seen in Figure 4.4 that there are several branches of the species tree where the duplication rates are significantly higher than expected in both the ORs and the T2Rs. The leaf branches that show significant bursts of duplications are generally the species with above average numbers of receptors, with the exception of humans. This suggests that one of the main driving forces for the evolution of ORs and T2Ts is bursts of duplications. The internal branches that tend to have increased duplication rates, in general, occur around the time of the divergence of the placental mammals in the late Cretaceous period. This was a time of bursts of general evolution, with several mammalian orders arising around this time. This increase in speciation may have affected the duplication rates seen in the ORs and the T2Rs.

However, species such as the dog and the pig do not show large lineage specific expansions although they have some of the largest receptor repertoires found in our dataset. Estimating ancestral sequence repertoires can be difficult when not using a duplication counting method that takes into account losses as well as duplications. The lack of recordable losses affects this result as some massively duplicated receptor subtypes from early in vertebrate evolution may have been lost in more recent times. It can be seen from Figure 4.1 that there may have been a general increase in the rate at which these duplications occur, as most the branches that were found to have significantly high duplication rates tend to be towards the tips of the tree, rather than the root. This suggests that there may have been a general increase in chemical sensory acuity over time. This also may have been due to segmental tandem duplications. As the cluster of duplicates increased the likelihood of more than one gene being duplicated at a time would have increased. As more duplications occurred the receptor repertoire was larger, increasing the likelihood for a duplication to occur. However, without recording losses as well as duplications and without having access to the genomic information of extinct animals, it is unclear if modern animals have larger receptor repertoires when compared to their ancestors. Although the duplication patterns found in the receptors of the pig and the dog do suggest that they evolved via a general increase in the number of receptors across large proportions of the gene tree, with minimal losses.

The results of this study show that there is evidence for both bursts of duplications and a gradual increase of duplications in the evolution of ORs and

T2Rs. However, further analyses could clarify these patterns if, in the future, a method that records losses as well as duplications could be used accurately on very large gene trees.

4.5 Conclusion

In this chapter, the patterns of duplication of the OR and T2R protein families were analysed to investigate if certain trends could be identified. These large chemosensory receptor families are ideal for this type of study because they primarily evolved via gene duplication, allowing for the detection of a wide variety of possible ligands. The results found in this study show that certain lineages along the species tree did achieve significant bursts of duplications when compared to the other branches, suggesting that these receptors evolved in part by bursts of gene duplication. Although, when the receptor repertoires of specific species were analysed, the results showed that not all the species with large amounts of receptors had acquired their repertoire by bursts of duplications. These results showed that the ecological niche of the organism can significantly shape the evolution of these large receptor families. The ORs and the T2Rs do not follow one particular trend of duplication. Rather, there is evidence for both bursts of duplications (as can be seen in the elephant) as well as a constant increase in the numbers of receptors over time (as can be seen in the dog).

Chapter 5 – Final Discussion and Future Work

The main aim of this thesis was to obtain further insight into how and why gene duplication occurs in sensory systems. Gene duplication is recognised as an important driving factor of genomic novelty and evolution (Hughes 2002; Zhang 2003). It allows for neofunctionalisation and increased complexity in animal systems (Hittinger and Carroll 2007). Gene duplication has been shown to play a pivotal role in the morphological complexity of vertebrates through the increase of *Hox* gene numbers (Garcia-Fernàndez and Holland 1994). As the senses function as an animal's toolset for detecting its environment, they are excellent model systems for the study of gene duplication in response to variable environmental factors. Sensory duplications have increased the variability and acuity of complex chemical detection systems such as the many duplications that have occurred in the evolution of ORs and T2Rs. Another example is the duplication of a longwave sensitive opsin gene in new world apes, allowing for trichromatic vision and an adaptation towards frugivory and increased intra-species visual communication. In this thesis several types of gene duplications of sensory GPCRs (as well as their activation pathways) that arose as a result of different selective pressures are discussed. Three main duplication trends are discussed in the three results chapters and include (1) Opsin duplications that allowed for the detection of a broader range of the light spectrum; (2) Duplication of the phototransduction pathway producing a second light activated pathway in vertebrates, allowing for the divergence between the rod and cone cell types; (3) Extreme large-scale duplication patterns found in ORs and T2Rs as a result of a need to detect a broad range of potential chemical ligands.

Chapter two addresses the question of why colour vision evolved while showing that the evidence in favour of the previous theories (foraging/finding prey) are not well supported (Parker 2004). The Ocean Drive hypothesis is presented in this work to possibly answer that question. This hypothesis states that colour vision evolved as a result of organisms moving into deeper regions of the ocean where different wavelengths of light are more abundant, thus, adapting their visual system to detect these wavelengths of light. Colour vision evolved by duplication of a light detecting opsin protein, followed by mutations allowing for the new proteins to maximally detect different wavelengths of light. The phylogeny of the opsins was constructed, whereby these duplications could be seen and ancestral wavelengths absorbencies could be calculated. These results showed that the pattern of the emergence of new opsin subfamilies followed a trend that matched the penetration ability of certain wavelengths of light in water. This pattern was seen in both arthropod and vertebrate visual opsins. In order to support the Ocean Drive Hypothesis, the duplications of the visual opsins of both the arthropods and the vertebrates would have to occur at around the same time, as a result of an environmental change to the oceans. Molecular dating techniques were applied to the opsin dataset to determine the dates at which the visual opsin duplications occurred. The results showed that the visual opsins in both vertebrates and arthropods duplicated at around the same time, not long after the oxygenation of the oceans. These results are robust and well supported as the effect of each of the parameters (including the calibrations taken from the fossil record) on the results were tested extensively. The results of this chapter give a parsimonious and clear explanation for the evolution of colour vision in the Ocean Drive Hypothesis. Ocean depth and light availability has a powerful effect on the evolution of the visual systems of aquatic animals, as can been seen in the work of Davies *et al.* (2012). The authors show that the opsin repertoires of aquatic animals are reduced when they are not exposed to certain wavelengths of light by moving deeper into the oceans. The Ocean Drive Hypothesis suggests exactly the opposite; opsin repertoires were expanded when animals moved deeper into the oceans due to a need to detect the wavelengths of light that were available at certain depths. The results found in this chapter show how an environmental change can result in protein neofunctionalisation in the form of the detection of different wavelengths of light.

Chapter three describes the duplication patterns found in the emergence of new protein interaction pathways. This chapter builds on the results found in chapter two by discussing the duplication patterns found after the emergence of the rod activation pathway from the ancestral cone activation pathway. This work addresses the question of whether the duplications that resulted in this newly emerging pathway arose by co-duplication or co-option (Plachetzki and Oakley 2007). In order to answer this question, phylogenetic analyses were performed on each protein in the pathway. Extensive testing was applied in order to determine the closest outgroup for each protein and which substitution matrix was the best fit to the data. The date for the emergence of the rod type opsin (Rhodopsin) was taken from the results of the chapter two. The next protein in the activation pathway, the G-protein Transducin was less well supported in its topology than the other proteins. Multiple potential topologies were found by using different tree reconstruction software and alternative analyses of the data. The position of the lamprey sequences posed the most ambiguity and three possible topologies were compared using Consel. This software compares topologies based on their site likelihoods and performs an AU test to determine if any of the possible topologies can be considered significantly less likely than the others. The most parsimonious tree, where each of the Gt clades contained one lamprey sequence, was the most likely tree although it could not be stated with very high confidence that it was significantly more likely than the other topologies. Additional basal species sequences were added to the Transducin dataset in order to clarify the position of the lamprey. Reconstruction of the phylogeny using the new data did not confirm the position of the lamprey as the Consel results once again could not confirm which topology was more likely. The analysis was repeated using the next protein in the pathway, PDE6. After the identification of the closest outgroup and the best fitting substitution matrix, the phylogeny was quite robust and required minimal additional testing. Finally, the analysis was repeated on the CNG-channels. This alignment contained some divergent non-vertebrate deuterostomes, which resulted in a relatively long alignment compared to the sequence length. To counteract this problem, the software Gblocks was used to reduce down the alignment. Manually reduced alignments were also produced. The outgroup and the best fitting substitution matrix were again identified and multiple tree reconstruction methods were used to determine the most likely tree. Once a number of robust phylogenies were identified for each protein in the pathway, these were analysed using molecular dating techniques and calibrations taken from the fossil record. In order to get the most likely date for the divergence between the rod and cone types of each protein, an average date was taken from each of the resulting dates for each phylogeny. When these results were graphed it could be seen that the vast majority of the pathway diverged at around the same time, circa 501mya, suggesting a co-duplication of the pathway. However, this trend was not observed across all Transducin dating results. The date for the split between the rod and cone type of Transducin was dated at over 800mya in the datasets that did not contain the additional sequences. The addition of more sequences included some tunicate Transducins which allowed for the calibration of the node that corresponded to the divergence of the tunicates and the vertebrates. This constrained the date for the rod and cone divergence to be much younger, at 561mya. The dataset that included the additional sequences suggests a co-duplication of the pathway as, although this is the oldest duplication, is it still relatively close to the timing of the duplications of the rest of the pathway. From these results the exact date for the divergence between the rod and cone type Transducins is not certain. The previous results suggest the possibility that the Transducins duplication significantly earlier than the other proteins. This may be due to the rod type Transducin having an alternative function in taste perception. It is possible that the rod type Transducin may have duplicated for another function in taste perception and was later co-opted into its current function in vision. The results of this chapter show that the proteins of the ancestral cone activation pathway mostly co-duplicated, resulting in the rod and cone pathways we see in vertebrates today. The ancestral visual Transducin may have co-duplicated but there is some evidence to suggest that it may have been co-opted from a previous gustation function. This result emphasises how interactions between proteins can influence the duplication patterns of their genes.

Chapter four describes the evolution of olfactory and bitter taste receptor proteins in response to niche occupation. This study once again analyses trends of duplication patterns but in two massively duplicated protein families, the ORs and the T2Rs. The evolution of these proteins is unlike what has been discussed so far in relation to vision as the range of possible ligands for smell and taste is extremely large and varied (Shi *et al*. 2003; Niimura and Nei 2007). The impact of niche occupation and environment is extremely powerful in these proteins. It has been shown that the OR repertoire of humans and chimps can be almost 25% species specific, suggesting a massive change in receptor repertoire in an extremely short space of time (Go and Niimura 2008). This study answers the question of how these genes duplicate, specifically 1) whether there were times when both receptor types expanded in response to an environmental change fueling the expansion of the genome and 2) did these genes duplicate at a constant rate, gradually expanding the organisms' gene repertoire or whether there were clear bursts of duplications at certain times. In order to answer these questions a comprehensive dataset of ORs and T2Rs were taken from the proteomes of 27 deuterostomes. These datasets were used to reconstruct the phylogenies of the ORs and the T2Rs. These families are extremely large when compared to the protein families discussed in the previous chapters, in particular the OR dataset, which contained over 14,000 sequences. Due to the size of the OR dataset, RAxML was used to reconstruct the tree using the fast tree parameters that approximate some calculations which reduced the computational time required to perform the analysis. Additionally, the size of the dataset was too large for most other tree reconstruction methods to calculate. The species overlap method for counting duplications was then used to count the number of duplications that occurred on each branch of the species tree. When the rates of duplication were calculated for each branch, a boxplot was constructed to delineate the trends found in the data. It could be seen from the boxplots that certain branches had rates which were much higher than others and appeared as outliers. These branches were identified and were compared between the two datasets. Only the branch leading to the Boreoeutheria and the Zooamata showed significant expansion in both ORs and T2Rs, suggesting that their duplication patterns occurred independently. In order to gain further insight into the trends of duplications in each species the protein trees were divided up into smaller subtrees and the species composition of each subtree was analysed. This data was then graphed using a histogram for the T2Rs and the software CIRCOS was used for the ORs due to the larger dataset. This method of representation clearly showed the regions of the tree where expansions could be seen in particular species. This was compared with the numbers of each receptor found in each of the species to determine the duplication trends that arose in the evolution of these proteins in the species with unusually large receptor repertoires. This data clearly showed evidence for both bursts of duplications in certain species as well as a gradual increase in receptor number in others. The lifestyle of the species massively influenced the number and type of receptors present in its genome. This study clearly demonstrates large scale duplication patterns across a diverse range of species and shows how environmental and niche related changes can influence the duplication rate of a gene.

This thesis demonstrates how and why gene duplications occur in several sensory systems. Firstly by showing how the duplication patterns occurred in the evolution of colour vision and by proposing the Ocean Drive Hypothesis to explain why this occurred in this manner. Secondly, the evolution of the rod and cone pathways is explained by a need for a bright light and a dim light visual system which evolved by a co-duplication of the ancestral pathway, although there is some evidence for the co-option of the rod type Transducin from taste perception. This shows the importance of the interrelatedness of sensory evolution. Thirdly, the evolution of large sensory protein families, ORs and T2Rs, is discussed in relation to the specialisation of an organism's chemical sensory system to its environment. This is shown to be as a result of both a gradual increase in the receptor repertoires by duplications as well as bursts of duplications at certain branches of the species tree. In conclusion, the work in this thesis clearly demonstrates some previously undiscovered duplication patterns found in the complex evolution of sensory systems.

There are some areas of this thesis could be expanded upon in the future. The analysis performed in chapter three compares the timing of the rod/cone duplications for each of the proteins in the phototransduction activation pathway. This pathway is more complex than this suggests, as a series of proteins are also expressed in these photoreceptor cells whose function is to deactivate the signal. An analysis to determine the timing of the duplications of the rod and cone types of each of these proteins would expand our knowledge on the trends of co-duplication in pathways such as these. Chapter four gives an overview of the phylogenies of the OR and T2R receptor repertoires of a variety of animal species. This study could be expanded upon in the future to investigate the specific types of receptors and their ligands that are common to some species or that are species specific. The regions of the tree that showed bursts of duplications in the receptor repertoire of a specific species could be analysed to determine their ligands and expression patterns and hence, their function in the sensory systems of these animals.

Bibliography

Abouheif, E., Zardoya, R., *et al.* (1998). "Limitations of metazoan 18S rRNA sequence data: implications for reconstructing a phylogeny of the animal kingdom and inferring the reality of the Cambrian explosion." Journal of Molecular Evolution 47(4): 394-405.

Adipietro, K. A., Mainland J.D., *et al.* (2012). "Functional evolution of mammalian odorant receptors." PLoS Genetics 8(7): e1002821.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle.

Altschul, S. F. (1991). "Substitution Matrices." eLS.

Altschul, S. F., W. Gish, *et al.* (1990). "Basic local alignment search tool." Journal of Molecular Biology 215(3): 403-410.

Arendt, D. (2003). "Evolution of eyes and photoreceptor cell types." International Journal of Developmental Biology 47(7/8): 563-572.

Ashery-Padan, R. and P. Gruss (2001). "Pax6 lights-up the way for eye development." Current Opinion in Cell Biology 13(6): 706-714.

Atkinson, H. J., J. H. Morris, *et al.* (2009). "Using sequence similarity networks for visualization of relationships across diverse protein superfamilies." PLoS One 4(2): e4345.

Bachmanov, A. A. and G. K. Beauchamp (2007). "Taste receptor genes." Annual Review of Nutrition 27:389.

Bacon, D. J. and W. F. Anderson (1986). "Multiple sequence alignment." Journal of Molecular Biology 191(2): 153-161.

Ballais, J. L. and J. Cohen (1985). "Problems of fossilization and interpretation of pollen from a travertine from Sidi-Masmoudi(Aures-Algeria).)." Comptes Rendus des Seances de la Societe de Biogeographie. 61(4): 118-128.

Barton, R. A. (2006). "Olfactory evolution and behavioral ecology in primates." American Journal of Primatology 68(6): 545-558.

Battistuzzi, F. U., A. Filipski, *et al.* (2010). "Performance of relaxed-clock methods in estimating evolutionary divergence times and their credibility intervals." Molecular Biology and Evolution 27(6): 1289-1300.

Ben-Arie, N., D. Lancet, *et al*. (1994). "Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire." Human Molecular Genetics 3(2): 229-235.

Bengtson, S. (2002). "Origins and early evolution of predation." Paleontological Society Papers 8: 289-318.

Benton, M., P. Donoghue, *et al*. (2009). "Calibrating and constraining molecular clocks." The Timetree of Life: 35-86.

Benton, M. J. (2009). Fossil Record: Quality, Wiley Online Library.

Benton, M. J., M. Wills, *et al*. (2000). "Quality of the fossil record through time." Nature 403(6769): 534-537.

Berger, J. O. and L. R. Pericchi (1996). "The intrinsic Bayes factor for model selection and prediction." Journal of the American Statistical Association: 109-122.

Bergsten, J. (2005). "A review of long,Äêbranch attraction." Cladistics 21(2): 163-193.

Birol, I., S. D. Jackman, *et al*. (2009). "De novo transcriptome assembly with ABySS." Bioinformatics 25(21): 2872-2877.

Blair, J. E. and S. B. Hedges (2005). "Molecular phylogeny and divergence times of deuterostome animals." Molecular Biology and Evolution 22(11): 2275-2284.

Boschat, C., C. Pélofi, *et al*. (2002). "Pheromone detection mediated by a V1r vomeronasal receptor." Nature Neuroscience 5(12): 1261-1262.

Bourne, H. R. (1997). "How receptors talk to trimeric G proteins." Current Opinion in Cell Biology 9(2): 134-142.

Bowmaker, J. and H. Dartnall (1980). "Visual pigments of rods and cones in a human retina." The Journal of Physiology 298(1): 501-511.

Bozdogan, H. (1987). "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions." Psychometrika 52(3): 345-370.

Brain, C. K. B., A. R. Prave, *et al*. (2012). "The first animals: ca. 760-million-year-old sponge-like fossils from Namibia." South African Journal of Science 108(1/2): 8 pages.

Bridges, C. B. (1936). "THE BAR" GENE" A DUPLICATION." Science (New York, NY) 83(2148): 210.

Briscoe, A. D. and L. Chittka (2001). "The evolution of color vision in insects." Annual Review of Entomology 46(1): 471-510.

Brosius, J. (1999). "Many G-protein-coupled receptors are encoded by retrogenes." Trends in Genetics 15(8): 304-305.

Buck, L. B. (1996). "Information coding in the vertebrate olfactory system." Annual Review of Neuroscience 19(1): 517-544.

Buee, M., M. Reich, *et al*. (2009). "454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity." New Phytologist 184(2): 449-456.

Burger, R. and M. Lynch (1995). "Evolution and extinction in a changing environment: a quantitative-genetic analysis." Evolution: 151-163.

Burns, M. E. and D. A. Baylor (2001). "Activation, deactivation, and adaptation in vertebrate photoreceptor cells." Annual Review of Neuroscience 24(1): 779-805.

Bush, A. M., R. K. Bambach, *et al.* (2011). "Ecospace utilization during the Ediacaran radiation and the Cambrian eco-explosion." Quantifying the Evolution of Early Life: 111-133.

Butterfield, N. J. (2003). "Exceptional fossil preservation and the Cambrian explosion." Integrative and Comparative Biology 43(1): 166-177.

Canfield, D. E. (2005). "The early history of atmospheric oxygen: homage to Robert M. Garrels." Annual Review of Earth and Planetary Sciences 33: 1-36.

Canfield, D. E., Poulton S. W. *et al.* (2007). "Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life." Science 315(5808): 92-95.

Carleton, K. L., Parry, J. W. L. *et al.* (2005). "Colour vision and speciation in Lake Victoria cichlids of the genus Pundamilia." Molecular Ecology 14(14): 4341-4353.

Carroll, J., C. J. Murphy, *et al.* (2001). "Photopigment basis for dichromatic color vision in the horse." Journal of Vision 1(2).

Carter-Dawson, L. D. and M. M. Lavail (2004). "Rods and cones in the mouse retina. I. Structural analysis using light and electron microscopy." The Journal of Comparative Neurology 188(2): 245-262.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular Biology and Evolution 17(4): 540-552.

Chandrashekar, J., K. L. Mueller, *et al.* (2000). "T2Rs function as bitter taste receptors." Cell 100(6): 703-711.

Chang, D. and T. F. Duda Jr (2012). "Extensive And Continuous Duplication Facilitates Rapid Evolution And Diversification Of Gene Families." Molecular Biology and Evolution.

Chang, L., R. Hirsch, *et al*. (1990). "Effects of insertional and point mutations on the functions of the duck hepatitis B virus polymerase." Journal of Virology 64(11): 5553-5558.

Chenna, R., H. Sugawara, *et al*. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Research 31(13): 3497-3500.

Chib, S. and E. Greenberg (1995). "Understanding the metropolis-hastings algorithm." American Statistician: 327-335.

Chib, S. and I. Jeliazkov (2001). "Marginal likelihood from the Metropolis-Hastings output." Journal of the American Statistical Association 96(453): 270-281.

Chou, C. S. and H. J. Lin (2006). "Some properties of CIR processes." Stochastic Analysis and Applications 24(4): 901-912.

Clapham, D. E. (1996). "The G-protein nanomachine." Nature 379(6563): 297-299.

Clifford, T. (1968). "Radiometric dating and the pre-Silurian geology of Africa." Radiometric Dating for Geologists: 299-416.

Collins, A. G. (1998). "Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence." Proceedings of the National Academy of Sciences 95(26): 15458.

Conte, C., M. Ebeling, *et al*. (2002). "Identification and characterization of human taste receptor genes belonging to the TAS2R family." Cytogenetic and Genome Research 98(1): 45-53.

Conway Morris, S. (2000). "The Cambrian “explosion”: Slow-fuse or megatonnage?" Proceedings of the National Academy of Sciences 97(9): 4426.

Cummins, C. A. and J. O. McInerney (2011). "A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases." Systematic Biology 60(6): 833-844.

Dagan, T. and W. Martin (2006). "The tree of one percent." Genome Biology 7(10): 118.

Dames, S., J. Durtschi, *et al*. (2010). "Comparison of the Illumina Genome Analyzer and Roche 454 GS FLX for resequencing of hypertrophic cardiomyopathy-associated genes." Journal of Biomolecular Techniques: JBT 21(2): 73.

Darwin, C. (1859). "On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life." J. Murray, London.

Davies, W. I. L., S. P. Collin, *et al*. (2012). "Molecular ecology and adaptation of visual photopigments in craniates." Molecular Ecology 21(13): 3121-3158.

Dayhoff, M. O., R. V. Eck, *et al*. (1968). Atlas of Protein Sequence and Structure, 1967-68, National Biomedical Research Foundation.

De La Chaux, N., P. Messer, *et al*. (2007). "DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage." BMC Evolutionary Biology 7(1): 191.

Delgado, S., D. Casane, *et al*. (2001). "Molecular evidence for Precambrian origin of amelogenin, the major protein of vertebrate enamel." Molecular Biology and Evolution 18(12): 2146-2153.

Dixson, D. L., P. L. Munday, *et al*. (2010). "Ocean acidification disrupts the innate ability of fish to detect predator olfactory cues." Ecology Letters 13(1): 68-75.

Donoghue, P. C. J. and M. A. Purnell (2005). "Genome duplication, extinction and vertebrate evolution." Trends in Ecology & Evolution 20(6): 312-319.

Doolittle, W. F. (1999). "Lateral genomics." Trends in Genetics 15(12): M5-M8.

Drummond, A. J., S. Y. W. Ho, *et al*. (2006). "Relaxed phylogenetics and dating with confidence." PLoS biology 4(5): e88.

Dunham, T. D. and D. L. Farrens (1999). "Conformational changes in Rhodopsin." Journal of Biological Chemistry 274(3): 1683-1690.

Eberl, D. F., R. W. Hardy, *et al*. (2000). "Genetically similar transduction mechanisms for touch and hearing in Drosophila." The Journal of Neuroscience 20(16): 5981-5988.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research 32(5): 1792-1797.

Edgar, R. C. and S. Batzoglou (2006). "Multiple sequence alignment." Current Opinion in Structural Biology 16(3): 368-373.

Eizirik, E., W. Murphy, *et al*. (2001). "Molecular dating and biogeography of the early placental mammal radiation." Journal of Heredity 92(2): 212-219.

Elena, S. F., V. S. Cooper, *et al*. (1996). "Punctuated evolution caused by selection of rare beneficial mutations." Science 272(5269): 1802-1804.

Enright, A. J., S. Van Dongen, *et al*. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Research 30(7): 1575-1584.

Erwin, D. H. (1991). "Metazoan phylogeny and the Cambrian radiation." Trends in Ecology & Evolution 6(4): 131-134.

Erwin, D. H., M. Laflamme, *et al*. (2011). "The Cambrian conundrum: early divergence and later ecological success in the early history of animals." Science 334(6059): 1091-1097.

Fain, G. L., R. Hardie, *et al*. (2010). "Phototransduction and the evolution of photoreceptors." Current Biology 20(3): R114-124.

Fares, M. A., S. F. Elena, *et al*. (2002). "A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses." Journal of Molecular Evolution 55(5): 509-521.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of Molecular Evolution 17(6): 368-376.

Felsenstein, J. (1985). "Confidence limits on phylogenies: an approach using the bootstrap." Evolution: 783-791.

Felsenstein, J. (2004). "Inferring phytogenies." Sunderland, Massachusetts: Sinauer Associates.

Fernald, R. D. (2006). "Casting a genetic light on the evolution of eyes." Science Signalling 313(5795): 1914.

Feuda, R., Hamilton, S.C., *et al*. (2012). "Metazoan opsin evolution reveals a simple route to animal vision." Proceedings of the National Academy of Sciences.

Finger, T. E. (1997). "Evolution of taste and solitary chemoreceptor cell systems." Brain, Behavior and Evolution 50(4): 234-243.

Fischer, A., Y. Gilad, *et al*. (2005). "Evolution of bitter taste receptors in humans and apes." Molecular Biology and Evolution 22(3): 432-436.

Fisher, R. A. (1912). "On an Absolute Criterion for Fitting Frequency Curves."

Fisher, R. A. (1921). "On the "probable error" of a coefficient of correlation deduced from a small sample." Metron 1(5): 3-32.

Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222(594-604): 309-368.

Fitch, W. M. and E. Margoliash (1967). "Construction of phylogenetic trees." Science 155(760): 279-284.

Flicek, P., M. R. Amode, *et al*. (2011). "Ensembl 2011." Nucleic Acids Research 39(suppl 1): D800-D806.

Fredriksson, R., M. C. Lagerstrom, *et al*. (2003). "The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints." Molecular Pharmacology 63(6): 1256-1272.

Fredriksson, R., M. C. Lagerström, *et al*. (2003). "The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints." Molecular Pharmacology 63(6): 1256-1272.

Freitag, J., G. Ludwig, *et al*. (1998). "Olfactory receptors in aquatic and terrestrial vertebrates." Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology 183(5): 635-650.

Ganchrow, J. R., D. Ganchrow, *et al*. (1993). "Aspects of vertebrate gustatory phylogeny: morphology and turnover of chick taste bud cells." Microscopy Research and Technique 26(2): 106-119.

Garcia-Fernàndez, J. and P. W. H. Holland (1994). "Archetypal organization of the amphioxus Hox gene cluster." Nature 370(6490): 563-566.

Garrett, K., S. Dendy, *et al*. (2006). "Climate change effects on plant disease: genomes to ecosystems." Annual Review of Phytopathology 44: 489-509.

Gatesy, J., R. DeSalle, *et al*. (2006). "Alignment-ambiguous nucleotide sites and the exclusion of systematic data." Molecular Phylogenetics and Evolution 2(2): 152-157.

Gehring, W. J. (1996). "The master control gene for morphogenesis and evolution of the eye." Genes to Cells 1(1): 11-15.

Gehring, W. J. and K. Ikeo (1999). "*Pax 6*: mastering eye morphogenesis and eye evolution." Trends in Genetics 15(9): 371-377.

George, R. A. and J. Heringa (2002). "Protein domain identification and improved sequence similarity searching using PSI-BLAST." Proteins: Structure, Function, and Bioinformatics 48(4): 672-681.

Gittleman, J. L. (1991). "Carnivore olfactory bulb size: allometry, phylogeny and ecology." Journal of Zoology 225(2): 253-272.

Glusman, G., I. Yanai, *et al*. (2001). "The complete human olfactory subgenome." Genome Research 11(5): 685-702.

Go, Y. and Y. Niimura (2008). "Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees." Molecular Biology and Evolution 25(9): 1897-1907.

Goldman, N., J. P. Anderson, *et al.* (2000). "Likelihood-based tests of topologies in phylogenetics." Systematic Biology 49(4): 652-670.

Goodman, S. N. (1999). "Toward evidence-based medical statistics. 2: The Bayes factor." Annals of Internal Medicine 130(12): 1005-1013.

Gough, J., K. Karplus, *et al.* (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure1." Journal of Molecular Biology 313(4): 903-919.

Gouy, M., S. Guindon, *et al.* (2010). "SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Molecular Biology and Evolution 27(2): 221-224.

Gracheva, E. O., N. T. Ingolia, *et al.* (2010). "Molecular basis of infrared detection by snakes." Nature 464(7291): 1006-1011.

Grus, W. E. and J. Zhang (2008). "Distinct evolutionary patterns between chemoreceptors of 2 vertebrate olfactory systems and the differential tuning hypothesis." Molecular Biology and Evolution 25(8): 1593-1601.

Guindon, S., J. Dufayard, *et al.* (2009). "PhyML: fast and accurate phylogeny reconstruction by maximum likelihood." Infection Genetics and Evolution 9(3): 384-385.

Guindon, S., J. F. Dufayard, *et al.* (2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." Systematic Biology 59(3): 307-321.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Systematic Biology 52(5): 696-704.

Haas, B. J., D. Gevers, *et al.* (2011). "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons." Genome Research 21(3): 494-504.

Hahn, D. A., G. J. Ragland, *et al.* (2009). "Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly Sarcophaga crassipalpis." BMC Benomics 10(1): 234.

Halanych, K. M. (2004). "The new view of animal phylogeny." Annual Review of Ecology, Evolution, and Systematics: 229-256.

Halanych, K. M. and Y. Passamaneck (2001). "A brief review of metazoan phylogeny and future prospects in Hox-research." American Zoologist 41(3): 629-639.

Hamm, H. E. (1998). "The many faces of G protein signaling." Journal of Biological Chemistry 273(2): 669-672.

Hankins, M. W., S. N. Peirson, *et al*. (2008). "Melanopsin: an exciting photopigment." Trends in Neurosciences 31(1): 27-36.

Hansen, A., S. H. Rolen, *et al*. (2003). "Correlation between olfactory receptor cell type and function in the channel catfish." The Journal of Neuroscience 23(28): 9328-9339.

Hara, T. J. (1994). "The diversity of chemical stimulation in fish olfaction and gustation." Reviews in Fish Biology and Fisheries 4(1): 1-35.

Hardie, R. C. (2001). "Phototransduction in Drosophila melanogaster." Journal of Experimental Biology 204(20): 3403-3409.

Harrigan, E. T. (2003). “Jackknife testing-an experimental approach to refine model calibration and validation.” Transportation Research Board.

Hasegawa, M. and H. Kishino (1994). "Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree." Molecular Biology and Evolution 11(1): 142.

Hasegawa, M., H. Kishino, *et al*. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." Journal of Molecular Evolution 22(2): 160-174.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." Biometrika 57(1): 97-109.

He, W., K. Yasumatsu, *et al*. (2004). "Umami taste responses are mediated by α-transducin and α-gustducin." The Journal of neuroscience 24(35): 7674-7680.

Heck, G. L., S. Mierson, *et al*. (1984). "Salt taste transduction occurs through an amiloride-sensitive sodium transport pathway." Science 223(4634): 403-405.

Heckel, D. G. (2010). "Smells like a new species: Gene duplication at the periphery." Proceedings of the National Academy of Sciences 107(21): 9481-9482.

Heimberg, A. M., R. Cowper-Sal, *et al*. (2010). "microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate." Proceedings of the National Academy of Sciences 107(45): 19379-19383.

Hennig, W., D. Davis, *et al*. (1999). Phylogenetic Systematics, Univ of Illinois Pr.

Hering, L., M. J. Henze, *et al*. (2012). "Opsins in Onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods." Molecular Biology and Evolution.

Hestrin, S. and J. I. Korenbrot (1990). "Activation kinetics of retinal cones and rods: response to intense flashes of light." The Journal of Neuroscience 10(6): 1967-1973.

Hildebrand, J. G. and G. M. Shepherd (1997). "Mechanisms of olfactory discrimination: converging evidence for common principles across phyla." Annual Review of Neuroscience 20(1): 595-631.

Hillis, D. M. and M. T. Dixon (1991). "Ribosomal DNA: molecular evolution and phylogenetic inference." Quarterly Review of Biology: 411-453.

Hisatomi, O. and F. Tokunaga (2002). "Molecular evolution of proteins involved in vertebrate phototransduction." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 133(4): 509-522.

Hittinger, C. T. and S. B. Carroll (2007). "Gene duplication and the adaptive evolution of a classic genetic switch." Nature 449(7163): 677-681.

Holland, H. D. (2006). "The oxygenation of the atmosphere and oceans." Philosophical Transactions of the Royal Society B: Biological Sciences 361(1470): 903-915.

Holland, P. W. H., J. Garcia-Fernàndez, *et al*. (1994). "Gene duplications and the origins of vertebrate development." Development 1994(Supplement): 125-133.

Hordijk, W. and O. Gascuel (2005). "Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood." Bioinformatics 21(24): 4338-4347.

Howell, N., J. L. Elson, *et al*. (2004). "African haplogroup L mtDNA sequences show violations of clock-like evolution." Molecular Biology and Evolution 21(10): 1843-1854.

Huang, A. L., X. Chen, *et al*. (2006). "The cells and logic for mammalian sour taste detection." Nature 442(7105): 934-938.

Hudson, R. R., M. Kreitman, *et al*. (1987). "A test of neutral molecular evolution based on nucleotide data." Genetics 116(1): 153-159.

Huelsenbeck, J. P., F. Ronquist, *et al*. (2001). "Bayesian inference of phylogeny and its impact on evolutionary biology." Science 294(5550): 2310-2314.

Huerta-Cepas, J., J. Dopazo, *et al*. (2010). "ETE: a python Environment for Tree Exploration." BMC bioinformatics 11(1): 24.

Hughes, A. L. (1994). "The evolution of functionally novel proteins after gene duplication." Proceedings of the Royal Society of London. Series B: Biological Sciences 256(1346): 119-124.

Hughes, A. L. (2002). "Adaptive evolution after gene duplication." Trends in Genetics 18(9): 433-434.

Hughes, A. L. (2005). "Gene duplication and the origin of novel proteins." Proceedings of the National Academy of Sciences of the United States of America 102(25): 8791.

Hughes, S., D. Zelus, *et al*. (1999). "Warm-blooded isochore structure in Nile crocodile and turtle." Molecular Biology and Evolution 16(11): 1521-1527.

Hurles, M. (2004). "Gene duplication: the genomic trade in spare parts." PLoS Biology 2(7): e206.

Hwang, D. G. and P. Green (2004). "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution." Proceedings of the National Academy of Sciences of the United States of America 101(39): 13994.

Ihara, K., T. Umemura, *et al*. (1999). "Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation1." Journal of Molecular Biology 285(1): 163-174.

Ishimaru, Y. (2009). "Molecular mechanisms of taste transduction in vertebrates." Odontology 97(1): 1-7.

Jacobs, D. K., N. Nakanishi, *et al*. (2007). "Evolution of sensory structures in basal metazoa." Integrative and Comparative Biology 47(5): 712-723.

Jacobs, G. H. (2009). "Evolution of colour vision in mammals." Philosophical Transactions of the Royal Society B: Biological Sciences 364(1531): 2957-2967.

Jacobs, G. H. and M. P. Rowe (2004). "Evolution of vertebrate colour vision." Clinical and Experimental Optometry 87(4,Äê5): 206-216.

Jaenisch, R. and A. Bird (2003). "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals." Nature Genetics 33: 245-254.

Jensen, S. (2003). "The Proterozoic and earliest Cambrian trace fossil record; patterns, problems and perspectives." Integrative and Comparative Biology 43(1): 219-228.

Jindrova, H. (1998). "Vertebrate phototransduction: activation, recovery and adaptation." Physiological Research 47: 155-168.

Jones, D. T., W. R. Taylor, *et al*. (1992). "The rapid generation of mutation data matrices from protein sequences." Computer Applications in the Biosciences: CABIOS 8(3): 275-282.

Juan, D., F. Pazos, *et al*. (2008). "Co-evolution and co-adaptation in protein networks." FEBS letters 582(8): 1225-1230.

Jukes, T. H. and C. R. Cantor (1969). "Evolution of protein molecules." Mammalian Protein Metabolism 3:21-132.

Kass, R. E. and A. E. Raftery (1995). "Bayes factors." Journal of the American Statistical Association: 773-795.

Kauer, J. S. (1991). "Contributions of topography and parallel processing to odor coding in the vertebrate olfactory pathway." Trends in Neurosciences 14(2): 79-85.

Kaupp, U. B. (2010). "Olfactory signalling in vertebrates and insects: differences and commonalities." Nature Reviews Neuroscience 11(3): 188-200.

Keane, T., C. Creevey, *et al*. (2006). "Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified." BMC Evolutionary Biology 6(1): 29.

Keane, T., T. Naughton, *et al*. (2004). "ModelGenerator: amino acid and nucleotide substitution model selection." National University of Ireland. Available at: http://bioinf.nuim.ie/software/modelgenerator.

Kear, A. J., D. E. G. Briggs, *et al*. (1995). "Decay and fossilization of non-mineralized tissue in coleoid cephalopods." Palaeontology 38(1): 105-132.

Khorana, H. (1992). "Rhodopsin, photoreceptor of the rod cell. An emerging pattern for structure and function." Journal of Biological Chemistry 267(1): 1-4.

Kimbel, W. H. (1988). "Identification of a partial cranium of *Australopithecus afarensis* from the Koobi Fora Formation, Kenya." Journal of Human Evolution 17(7): 647-656.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." Journal of Molecular Evolution 16(2): 111-120.

Kimura, M. (1985). The neutral theory of molecular evolution, Cambridge University Press.

Kimura, M. and T. Ohta (1972). "On the stochastic model for estimation of mutational distance between homologous proteins." Journal of Molecular Evolution 2(1): 87-90.

Kinsella, R. J., A. Kähäri, *et al*. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space." Database: the Journal of Biological Databases and Curation 2011.

Kircher, M., U. Stenzel, *et al*. (2009). "Improved base calling for the Illumina Genome Analyzer using machine learning strategies." Genome Biology 10(8): R83.

Kishino, H. and M. Hasegawa (1989). "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea." Journal of Molecular Evolution 29(2): 170-179.

Knoll, A. H. (2004). Life on a young planet: the first three billion years of evolution on earth, Princeton Univ Press.

Kobayashi, H. and S. Kohshima (2001). "Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye." Journal of Human Evolution 40(5): 419-435.

Kojima, D., T. Okano, *et al*. (1992). "Cone visual pigments are present in gecko rod cells." Proceedings of the National Academy of Sciences 89(15): 6841-6845.

Kolaczkowski, B. and J. W. Thornton (2009). "Long-branch attraction bias and inconsistency in Bayesian phylogenetics." PLoS One 4(12): e7891.

Korf, I., M. Yandell, *et al*. (2003). Blast, O'Reilly & Associates, Inc.

Krauthammer, M., A. Rzhetsky, *et al*. (2000). "Using BLAST for identifying gene and protein names in journal articles." Gene 259(1): 245-252.

Křivánek, M. (1986). "Computing the nearest neighbor interchange metric for unlabeled binary trees is NP-complete." Journal of Classification 3(1): 55-60.

Krzywinski, M., J. Schein, *et al*. (2009). "Circos: an information aesthetic for comparative genomics." Genome Research 19(9): 1639-1645.

Lai, P. C., M. S. Singer, *et al*. (2005). "Structural activation pathways from dynamic olfactory receptor‚Äìodorant interactions." Chemical Senses 30(9): 781-792.

Lamb, T. D., S. P. Collin, *et al*. (2007). "Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup." Nature Reviews Neuroscience 8(12): 960-976.

Land, M. F. (1997). "Visual acuity in insects." Annual Review of Entomology 42(1): 147-177.

Land, M. F. (2005). "The optical structures of animal eyes." Current Biology 15(9): R319-R323.

Land, M. F. and D. E. Nilsson (2012). Animal eyes, Oxford Univ Pr.

Lande, R. (1976). "Natural selection and random genetic drift in phenotypic evolution." Evolution: 314-334.

Lane, R. P., T. Cutforth, *et al.* (2002). "Sequence analysis of mouse vomeronasal receptor gene clusters reveals common promoter motifs and a history of recent expansion." Proceedings of the National Academy of Sciences 99(1): 291.

Larget, B. and D. L. Simon (1999). "Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees." Molecular Biology and Evolution 16: 750-759.

Lartillot, N., T. Lepage, *et al.* (2009). "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating." Bioinformatics 25(17): 2286-2288.

Lartillot, N. and H. Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process." Molecular Biology and Evolution 21(6): 1095-1109.

Lassmann, T. and E. Sonnhammer (2005). "Kalign: an accurate and fast multiple sequence alignment algorithm." BMC Bioinformatics 6(1): 298.

Lee, M. S. Y. (1999). "Molecular clock calibrations and metazoan divergence dates." Journal of Molecular Evolution 49(3): 385-391.

Lee, M. S. Y., J. B. Jago, *et al.* (2011). "Modern optics in exceptionally preserved eyes of early Cambrian arthropods from Australia." Nature 474(7353): 631-634.

Lee, M. S. Y. and T. H. Worthy (2011). "Likelihood reinstates Archaeopteryx as a primitive bird." Biology Letters.

Lepage, T., D. Bryant, *et al.* (2007). "A general comparison of relaxed molecular clock models." Molecular Biology and Evolution 24(12): 2669-2680.

Lewis, P. O. (1998). "A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data." Molecular Biology and Evolution 15(3): 277-283.

Li, C., G. D. Love, *et al.* (2010). "A stratified redox model for the Ediacaran ocean." Science 328(5974): 80-83.

Li, W. H., M. Tanimura, *et al*. (1987). "An evaluation of the molecular clock hypothesis using mammalian DNA sequences." Journal of Molecular Evolution 25(4): 330-342.

Liu, Y., D. C. Nickle, *et al*. (2004). "Molecular clock-like evolution of human immunodeficiency virus type 1." Virology 329(1): 101-108.

Low, P. J., R. Ai, *et al*. (1999). "Active sites in complement components C5 and C3 identified by proximity to indels in the C3/4/5 protein family." The Journal of Immunology 162(11): 6580-6588.

Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science 290(5494): 1151-1155.

Maddison, W. P. and D. R. Maddison (2001). "Mesquite: a modular system for evolutionary analysis."

Mallatt, J. and C. J. Winchell (2002). "Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes." Molecular Biology and Evolution 19(3): 289-301.

Malnic, B., J. Hirono, *et al*. (1999). "Combinatorial receptor codes for odors." Cell 96(5): 713-723.

Marshall, C. R. (2006). "Explaining the Cambrian “explosion” of animals." Annual Review of Earth Planetary Sciences 34: 355-384.

Martzen, M. R., S. M. McCraith, *et al*. (1999). "A biochemical genomics approach for identifying genes by the activity of their products." Science 286(5442): 1153-1155.

Mathers, P., A. Grinberg, *et al*. (1997). "The Rx homeobox gene is essential for vertebrate eye development." Nature 387(6633): 603.

Matsui, A., Y. Go, *et al*. (2010). "Degeneration of olfactory receptor gene repertories in primates: no direct link to full trichromatic vision." Molecular Biology and Evolution 27(5): 1192-1200.

Mayr, E. (1992). "Speciational evolution or punctuated equilibria." The Dynamics of Evolution: 21-48.

McBee, J. K., K. Palczewski, *et al*. (2001). "Confronting complexity: the interlink of phototransduction and retinoid metabolism in the vertebrate retina." Progress in Retinal and Eye Research 20(4): 469-529.

McGinnis, S. and T. L. Madden (2004). "BLAST: at the core of a powerful and diverse set of sequence analysis tools." Nucleic Acids Research 32(suppl 2): W20-W25.

Medina, M., A. G. Collins, *et al*. (2003). "Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa." International Journal of Astrobiology 2(3): 203-211.

Melin, A. D., L. M. Fedigan, *et al*. (2007). "Effects of colour vision phenotype on insect capture by a free-ranging population of white-faced capuchins, *Cebus capucinus."* Animal Behaviour 73(1): 205-214.

Menini, A. (1999). "Calcium signalling and regulation in olfactory neurons." Current Opinion in Neurobiology 9(4): 419-426.

Menzel, R. and M. Blakers (1976). "Colour receptors in the bee eye: morphology and spectral sensitivity." Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology 108(1): 11-13.

Metropolis, N., A. W. Rosenbluth, *et al*. (1953). "Equation of state calculations by fast computing machines." The Journal of Chemical Physics 21: 1087.

Miller, W. H. (1957). "Morphology of the ommatidia of the compound eye of Limulus." The Journal of Biophysical and Biochemical Cytology 3(3): 421-428.

Mistretta, C. M., K. A. Goosens, *et al*. (1999). "Alterations in size, number, and morphology of gustatory papillae and taste buds in BDNF null mutant mice demonstrate neural dependence of developing taste organs." The Journal of Comparative Neurology 409(1): 13.

Møller, A. P. and T. Szép (2005). "Rapid evolutionary change in a secondary sexual character linked to climatic change." Journal of Evolutionary Biology 18(2): 481-495.

Mombaerts, P. (1999). "Seven-transmembrane proteins as odorant and chemosensory receptors." Science 286(5440): 707-711.

Morris, S. C. (1989). "Burgess Shale faunas and the Cambrian explosion." Science 246(4928): 339-346.

Morris, S. C. (2000). "The fossil record and the early evolution of the Metazoa." Shaking the Tree: Readings from Nature in the History of Life: 128.

Morrison, E. E. and R. M. Costanzo (1992). "Morphology of olfactory epithelium in humans and other vertebrates." Microscopy Research and Technique 23(1): 49-61.

Mueller, K. L., M. A. Hoon, *et al*. (2005). "The receptors and coding logic for bitter taste." Nature 434(7030): 225-229.

Müller, W. E. G. (1995). "Molecular phylogeny of Metazoa (animals): monophyletic origin." Naturwissenschaften 82(7): 321-329.

Muyzer, G. (1999). "DGGE/TGGE a method for identifying genes from natural ecosystems." Current Opinion in Microbiology 2(3): 317-322.

Narbonne, G. M. (2010). "Ocean chemistry and early animals." Science 328(5974): 53-54.

Nathans, J. (1999). "The Evolution and Physiology of Human Review Color Vision: Insights from Molecular Genetic Studies of Visual Pigments." Neuron 24: 299-312.

Near, T. J., D. I. Bolnick, *et al*. (2005). "Fossil calibrations and molecular divergence time estimates in centrarchid fishes (Teleostei: Centrarchidae)." Evolution 59(8): 1768-1782.

Near, T. J. and M. J. Sanderson (2004). "Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359(1450): 1477-1483.

Nei, M. and S. Kumar (2000). Molecular Evolution and Phylogenetics, Oxford University Press.

Nei, M., Y. Niimura, *et al*. (2008). "The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity." Nature Reviews Genetics 9(12): 951-963.

Nei, M. and A. P. Rooney (2005). "Concerted and birth-and-death evolution of multigene families." Annual Review of Genetics 39: 121.

Neyman, J. (1971). "Molecular studies of evolution: a source of novel statistical problems." Statistical Decision Theory and Related Topics: 1-27.

Ng, P. C., S. Levy, *et al*. (2008). "Genetic variation in an individual human exome." PLoS Genetics 4(8): e1000160.

Nickle, B. and P. R. Robinson (2007). "The opsins of the vertebrate retina: insights from structural, biochemical, and evolutionary studies." Cellular and Molecular Life Sciences : CMLS 64(22): 2917-2932.

Nielsen, C. (2011). Animal Evolution: Interrelationships of the Living Phyla, OUP Oxford.

Nielsen, C. (2012). Animal Evolution: Interrelationships of the Living Phyla, Oxford Univ Pr.

Niimura, Y. and M. Nei (2007). "Extensive gains and losses of olfactory receptor genes in mammalian evolution." PLoS One 2(8): e708.

Nilsson, D. E. (1996). "Eye ancestry: old genes for new eyes." Current Biology 6(1): 39-42.

Nilsson, D. E. (2004). "Eye evolution: a question of genetic promiscuity." Current Opinion in Neurobiology 14(4): 407-414.

Notredame, C., D. G. Higgins, *et al*. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." Journal of Molecular Biology 302(1): 205-218.

Nozawa, M. and M. Nei (2008). "Genomic drift and copy number variation of chemosensory receptor genes in humans and mice." Cytogenetic and Genome Research 123(1-4): 263-269.

O'Brien, M. J. and R. L. Lyman (1999). Seriation, stratigraphy, and index fossils: The backbone of archaeological dating, Springer.

Ogura, A., K. Ikeo, *et al*. (2004). "Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human." Genome Research 14(8): 1555-1561.

Ogura, Y., D. K. Bonen, *et al*. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature 411(6837): 603-606.

Ohno, S. (1973). "Ancient linkage groups and frozen accidents." Nature 244: 259-262.

Onrust, R., P. Herzmark, *et al*. (1997). "Receptor and βγ binding sites in the α subunit of the retinal G protein transducin." Science 275(5298): 381-384.

Paglia, M. J., H. Mou, *et al*. (2002). "Regulation of photoreceptor phosphodiesterase (PDE6) by phosphorylation of its inhibitory γ subunit re-evaluated." Journal of Biological Chemistry 277(7): 5017-5023.

Palmeirim, I., D. Henrique, *et al*. (1997). "Avian *hairy* Gene Expression Identifies a Molecular Clock Linked to Vertebrate Segmentation and Somitogenesis." Cell 91(5): 639-648.

Parker, A. (2004). In the blink of an eye: how vision sparked the big bang of evolution, Basic Books.

Peck, J. R. (1994). "A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex." Genetics 137(2): 597-606.

Peterson, K. J. and N. J. Butterfield (2005). "Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record." Proceedings of the National Academy of Sciences of the United States of America 102(27): 9547.

Philippe, H., N. Lartillot, *et al*. (2005). "Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia." Molecular Biology and Evolution 22(5): 1246-1253.

Philippe, H. and M. J. Telford (2006). "Large-scale sequencing and the new animal phylogeny." Trends in Ecology & Evolution 21(11): 614-620.

Philippe, H., Y. Zhou, *et al*. (2005). "Heterotachy and long-branch attraction in phylogenetics." BMC Evolutionary Biology 5(1): 50.

Pichaud, F., A. Briscoe, *et al*. (1999). "Evolution of color vision." Current Opinion in Neurobiology 9(5): 622-627.

Pisani, D., S. Mohun, *et al*. (2006). "Molecular evidence for dim-light vision in the last common ancestor of the vertebrates." Current Biology 16(9): 318-319.

Plachetzki, D. C., B. M. Degnan, *et al*. (2007). "The origins of novel protein interactions during animal opsin evolution." PLoS One 2(10): e1054.

Plachetzki, D. C., C. R. Fong, *et al*. (2010). "The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway." Proceedings of the Royal Society B: Biological Sciences 277(1690): 1963-1969.

Plachetzki, D. C. and T. H. Oakley (2007). "Key transitions during the evolution of animal phototransduction: novelty, “tree-thinking”, co-option, and co-duplication." Integrative and Comparative Biology 47(5): 759-769.

Pleasance, E. D., R. K. Cheetham, *et al*. (2009). "A comprehensive catalogue of somatic mutations from a human cancer genome." Nature 463(7278): 191-196.

Porter, M. L., J. R. Blasic, *et al*. (2012). "Shedding new light on opsin evolution." Proceedings of the Royal Society B: Biological Sciences 279(1726): 3-14.

Posada, D. and T. R. Buckley (2004). "Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests." Systematic Biology 53(5): 793-808.

Posada, D. and K. A. Crandall (2001). "Selecting the best-fit model of nucleotide substitution." Systematic Biology 50(4): 580-601.

Price, M. N., P. S. Dehal, *et al*. (2010). "FastTree 2-approximately maximum-likelihood trees for large alignments." PLoS One 5(3): e9490.

Pugh, E. and T. Lamb (2000). "Phototransduction in vertebrate rods and cones: molecular mechanisms of amplification, recovery and light adaptation." Handbook of Biological Physics 3: 183-255.

Pumphrey, R. J. (1948). "The theory of the fovea." Journal of Experimental Biology 25(3): 299-312.

Pureswaran, D. S. and T. M. Poland (2009). "The role of olfactory cues in short-range mate finding by the emerald ash borer, Agrilus planipennis Fairmaire (Coleoptera: Buprestidae)." Journal of Insect Behavior 22(3): 205-216.

Rampino, N., H. Yamamoto, *et al.* (1997). "Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype." Science 275(5302): 967-969.

Restrepo, D., J. Arellano, *et al.* (2004). "Emerging views on the distinct but related roles of the main and accessory olfactory systems in responsiveness to chemosensory signals in mice." Hormones and Behavior 46(3): 247-256.

Retief, J. D. (2000). "Phylogenetic analysis using PHYLIP." Methods of Molecular Biology 132: 243-258.

Riba-Hernández, P., K. E. Stoner, *et al.* (2004). "Effect of polymorphic colour vision for fruit detection in the spider monkey *Ateles geoffroyi*, and its implications for the maintenance of polymorphic colour vision in platyrrhine monkeys." Journal of Experimental Biology 207(14): 2465-2470.

Ridge, K. D., N. G. Abdulaev, *et al.* (2003). "Phototransduction: crystal clear." Trends in Biochemical Sciences 28(9): 479-487.

Rivera, A. S., N. Ozturk, *et al.* (2012). "Blue-light-receptive cryptochrome is expressed in a sponge eye lacking neurons and opsin." The Journal of Experimental Biology 215(8): 1278-1286.

Robbins, L. S., J. H. Nadeau, *et al.* (1993). "Pigmentation phenotypes of variant extension locus alleles result from point mutations that alter MSH receptor function." Cell 72(6): 827-834.

Robertson, G., J. Schein, *et al.* (2010). "De novo assembly and analysis of RNA-seq data." Nature Methods 7(11): 909-912.

Robertson, H. M. (1998). "Two Large Families of Chemoreceptor Genes in the NematodesCaenorhabditis elegans and Caenorhabditis briggsae Reveal Extensive Gene Duplication, Diversification, Movement, and Intron Loss." Genome Research 8(5): 449-463.

Robinson-Rechavi, M. and V. Laudet (2001). "Evolutionary rates of duplicate genes in fish and mammals." Molecular Biology and Evolution 18(4): 681-683.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics 19(12): 1572-1574.

Roth, L. S. and A. Kelber (2004). "Nocturnal colour vision in geckos." Proceedings of the Royal Society of London. Series B: Biological Sciences 271(Suppl 6): S485-S487.

Rutschmann, F. (2006). "Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times." Diversity and Distributions 12(1): 35-48.

Ryszkiewicz, M. and R. S. Walker (1983). "Mammals Versus Dinosaurs: the Success of a Conspiracy." Diogenes 31(124): 78-89.

Saitou, N. and T. Imanishi (1989). "Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree." Molecular Biology and Evolution 6(5): 514-525.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Molecular Biology and Evolution 4(4): 406-425.

Satoh, G. (2005). "Characterization of novel GPCR gene coding locus in amphioxus genome: gene structure, expression, and phylogenetic analysis with implications for its involvement in chemoreception." Genesis 41(2): 47-57.

Seehausen, O., Y. Terai, *et al.* (2008). "Speciation through sensory drive in cichlid fish." Nature 455(7213): 620-626.

Sforza, C. L. L. and A. W. F. Edwards (1964). "Analysis of human evolution." Genetics Today 3: 923-933.

Shannon, P., A. Markiel, *et al.* (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Research 13(11): 2498-2504.

Shi, P. and J. Zhang (2007). "Comparative genomic analysis identifies an evolutionary shift of vomeronasal receptor gene repertoires in the vertebrate transition from water to land." Genome Research 17(2): 166-174.

Shi, P., J. Zhang, *et al.* (2003). "Adaptive diversification of bitter taste receptor genes in mammalian evolution." Molecular Biology and Evolution 20(5): 805-814.

Shimodaira, H. (2001). "Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection." Communications in Statistics-Theory and Methods 30(8-9): 1751-1772.

Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Systematic Biology 51(3): 492-508.

Shimodaira, H. and M. Hasegawa (1999). "Multiple comparisons of log-likelihoods with applications to phylogenetic inference." Molecular Biology and Evolution 16: 1114-1116.

Shimodaira, H. and M. Hasegawa (2001). "CONSEL: for assessing the confidence of phylogenetic tree selection." Bioinformatics 17(12): 1246-1247.

Smith, A. B. and K. J. Peterson (2002). "Dating the time of origin of major clades: molecular clocks and the fossil record." Annual Review of Earth and Planetary Sciences 30(1): 65-88.

Smith, F. A. and J. L. Betancourt (2003). "The effect of Holocene temperature fluctuations on the evolution and ecology of *Neotoma* (woodrats) in Idaho and northwestern Utah." Quaternary Research 59(2): 160-171.

Sneath, P. H. A. and R. R. Sokal (1973). Numerical taxonomy. The Principles and Practice of Numerical Classification.

Sosinsky, A., G. Glusman, *et al.* (2000). "The genomic structure of human olfactory receptor genes." Genomics 70(1): 49-61.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics 22(21): 2688-2690.

Stamatakis, A., T. Ludwig, *et al.* (2005). "RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees." Bioinformatics 21(4): 456-463.

Steiger, S. S., A. E. Fidler, *et al.* (2010). "Evidence for adaptive evolution of olfactory receptor genes in 9 bird species." Journal of Heredity 101(3): 325-333.

Stewart, C. B. (2000). "The powers and pitfalls of parsimony." Shaking the Tree: Readings from Nature in the History of Life: 48.

Stewart, I. (2003). "Speciation: a case study in symmetric bifurcation theory." Universitatis Iagellonicae Acta Mathematica 41: 67-88.

Stock, D. W. and G. S. Whitt (1992). "Evidence from 18S ribosomal RNA sequences that lampreys and hagfishes form a natural group." Science 257(5071): 787-789.

Strausfeld, N. J. and J. G. Hildebrand (1999). "Olfactory systems: common design, uncommon origins?" Current Opinion in Neurobiology 9(5): 634-639.

Strimmer, K. and A. Von Haeseler (1996). "Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies." Molecular Biology and Evolution 13(7): 964-969.

Suga, H., V. Schmid, *et al.* (2008). "Evolution and functional diversity of jellyfish opsins." Current Biology 18(1): 51-55.

Sugihara, M., V. Buss, *et al*. (2002). "11-cis-retinal protonated Schiff base: influence of the protein environment on the geometry of the rhodopsin chromophore." Biochemistry 41(51): 15259-15266.

Surridge, A. K., D. Osorio, *et al*. (2003). "Evolution and selection of trichromatic vision in primates." Trends in Ecology & Evolution 18(4): 198-205.

Talavera, G. and J. Castresana (2007). "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments." Systematic Biology 56(4): 564-577.

Tateno, Y., N. Takezaki, *et al*. (1994). "Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site." Molecular Biology and Evolution 11(2): 261-277.

Tatusova, T. A. and T. L. Madden (1999). "BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences." FEMS Microbiology Letters 174(2): 247-250.

Terakita, A. (2005). "The opsins." Genome Biology 6(3): 213.

Tomitani, A., A. H. Knoll, *et al*. (2006). "The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives." Proceedings of the National Academy of Sciences 103(14): 5442-5447.

Tsukamoto, H., A. Terakita, *et al*. (2010). "A pivot between helices V and VI near the retinal-binding site is necessary for activation in rhodopsins." Journal of Biological Chemistry 285(10): 7351-7357.

Turunen, O., M. Sainio, *et al*. (1998). "Structure-function relationships in the ezrin family and the effect of tumor-associated point mutations in neurofibromatosis 2 protein." Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology 1387(1-2): 1-16.

Valentine, J. W. (1997). "Cleavage patterns and the topology of the metazoan tree of life." Proceedings of the National Academy of Sciences 94(15): 8001.

Van Den Bussche, R. A., R. J. Baker, *et al*. (1998). "Base compositional bias and phylogenetic analyses: A test of the “Flying DNA” hypothesis." Molecular Phylogenetics and Evolution 10(3): 408-416.

van Dongen, S. (2007). "MCL-an algorithm for clustering graphs." URL: http://micans. org/mcl/, Access date: 1st August.

Venclovas, Č. (2003). "Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance." Proteins: Structure, Function, and Bioinformatics 53(S6): 380-388.

Voets, T., G. Droogmans, *et al.* (2004). "The principle of temperature-dependent gating in cold-and heat-sensitive TRP channels." Nature 430(7001): 748-754.

Vorobyev, M., D. Osorio, *et al.* (1998). "Tetrachromacy, oil droplets and bird plumage colours." Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology 183(5): 621-633.

Waddell, P. J. and M. A. Steel (1997). "General Time-Reversible Distances with Unequal Rates across Sites: Mixing [Gamma] and Inverse Gaussian Distributions with Invariant Sites." Molecular Phylogenetics and Evolution 8(3): 398-414.

Wallace, I. M., G. Blackshields, *et al.* (2005). "Multiple sequence alignments." Current Opinion in Structural Biology 15(3): 261-266.

Weissburg, M. J. and R. K. Zimmer-Faust (1994). "Odor plumes and how blue crabs use them in finding prey." Journal of Experimental Biology 197(1): 349-375.

Welch, J. J. and L. Bromham (2005). "Molecular dating when rates vary." Trends in Ecology & Evolution 20(6): 320-327.

Whelan, S. (2007). "New approaches to phylogenetic tree search and their application to large numbers of protein alignments." Systematic Biology 56(5): 727-740.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." Molecular Biology and Evolution 18(5): 691-699.

Wiens, J. J. (2005). "Can incomplete taxa rescue phylogenetic analyses from long-branch attraction?" Systematic Biology 54(5): 731-742.

Winstanley, H. F., S. Abeln, *et al.* (2005). "How old is your fold?" Bioinformatics 21(suppl 1): i449-i458.

Wisby, W. J. and A. D. Hasler (1954). "Effect of olfactory occlusion on migrating silver salmon (O. kisutch)." Journal of the Fisheries Board of Canada 11(4): 472-478.

Wray, G. A., J. S. Levinton, *et al.* (1996). "Molecular evidence for deep Precambrian divergences among metazoan phyla." Science 274(5287): 568-573.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis." The Annals of Statistics 14(4): 1261-1295.

Xi-Guang, Z. and E. N. K. Clarkson (1990). "The eyes of Lower Cambrian eodiscid trilobites." Palaeontology 33: 911-932.

Xia, X. and Z. Xie (2001). "DAMBE: software package for data analysis in molecular biology and evolution." Journal of Heredity 92(4): 371-373.

Yan, W., G. Sunavala, *et al.* (2001). "Bitter taste transduced by PLC-β2-dependent rise in IP3 and Œ±-gustducin-dependent fall in cyclic nucleotides." American Journal of Physiology-Cell Physiology 280(4): C742-C751.

Yang, H., P. Shi, *et al.* (2005). "Composition and evolution of the V2r vomeronasal receptor gene repertoire in mice and rats." Genomics 86(3): 306-315.

Yang, Y. (2005). "Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation." Biometrika 92(4): 937-950.

Yang, Z. (1996). "Among-site rate variation and its impact on phylogenetic analyses." Trends in Ecology & Evolution 11(9): 367-372.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Computer Applications in the Biosciences: CABIOS 13(5): 555-556.

Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." Trends in Ecology & Evolution 15(12): 496-503.

Yang, Z. and B. Rannala (2006). "Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds." Molecular Biology and Evolution 23(1): 212-226.

Yokoyama, S. (2000). "Molecular evolution of vertebrate visual pigments." Progress in Retinal and Eye Research 19(4): 385-419.

Yokoyama, S. (2002). "Molecular evolution of color vision in vertebrates." Gene 300(1): 69-78.

Yokoyama, S., T. Tada, *et al.* (2008). "Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates." Proceedings of the National Academy of Sciences 105(36): 13480-13485.

Yokoyama, S. and R. Yokoyama (1989). "Molecular evolution of human visual pigment genes." Molecular Biology and Evolution 6(2): 186-197.

Young, J. M., R. L. M. Endicott, *et al.* (2008). "Extensive copy-number variation of the human olfactory receptor gene family." American Journal of Human Genetics 83(2): 228.

Young, J. M., C. Friedman, *et al.* (2002). "Different evolutionary processes shaped the mouse and human olfactory receptor gene families." Human Molecular Genetics 11(5): 535-546.

Zhang, J. (2003). "Evolution by gene duplication: an update." Trends in Ecology & Evolution 18(6): 292-298.

Zhang, J., Y. Zhang, *et al*. (2002). "Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey." Nature Genetics 30(4): 411-415.

Zhang, X. and S. Firestein (2002). "The olfactory receptor gene superfamily of the mouse." Nature Neuroscience 5: 124-133.

Zhao, G. Q., Y. Zhang, *et al*. (2003). "The receptors for mammalian sweet and umami taste." Cell 115(3): 255-266.

Zhao, H., S. J. Rossiter, *et al*. (2009). "The evolution of color vision in nocturnal mammals." Proceedings of the National Academy of Sciences 106(22): 8980.

Zuckerkandl, E. and L. Pauling (1962). "Molecular disease, evolution and genetic heterogeneity." Horizons in Biochemistry: 189-225.

Zuckerkandl, E., L. Pauling, *et al*. (1962). "Horizons in biochemistry." Horizons in Biochemistry: 97-166.

Appendix

Table: 88 calibration points used in the opsin analysis in chapter two. Each row in the list shows the two taxa whose most recent common ancestral node is being calibrated, followed by the upper and lower bounds (mya) on each calibration point.

Species	Opsin Type	Species	Opsin Type	Upper	Lower
Pediculus humanus	R7	*Aedes aegypti*	R7	414	307.2
Papilio glaucus	UV	*Apodemia mormo*	UV	155	56
Pieris rapae	Blue	*Manduca sexta*	Blue	155	56
Nasonia vitripennis	UV	*Apis mellifera*	UV	243	152
Nasonia vitripennis	Blue	*Apis mellifera*	Blue	243	152
Branchiostoma floridae	Melanopsin	*Felis catus*	Melanopsin	-1	518.5
Branchiostoma floridae	Melanopsin	*Papilio glaucus*	Red	-1	531.5
Patella vulgata	Lopho R	*Papilio glaucus*	Red	646	531.5
Danio rerio	Melanopsin	*Felis catus*	Melanopsin	421.75	416
Oryzias latipes	Melanopsin	*Gallus gallus*	Melanopsin	421.75	416
Gallus gallus	Melanopsin	*Felis catus*	Melanopsin	330.4	312.3
Pan troglodytes	Melanopsin	*Homo sapiens*	Melanopsin	10	5.7
Patella vulgata	R	*Helicotylenchus canadensis*	R	666	605
Monodelphis domestica	Melanopsin	*Felis catus*	Melanopsin	171.2	124

Latimeria chalumnae	RH2	*Danio rerio*	RH2	421.75	416
Canis familiaris	RH1	*Felis Catus*	RH1	65.8	39.68
Taeniopygia guttata	RH1	*Gallus gallus*	RH1	86.5	66
Taeniopygia guttata	RH2	*Gallus gallus*	RH2	86.5	66
Taeniopygia guttata	SWS1	*Gallus gallus*	SWS1	86.5	66
Lethenteron japonicum	RH1	*Salmo salar*	RH1	-1	460.5
Geotria australis	SWS1	*Salmo salar*	SWS1	-1	460.5
Cavia porcellus	RH1	*Ornithorhynchus anatinus*	RH1	191.1	162.9
Mus musculus	RH1	*Rattus norvegicus*	RH1	14	10.4
Tetraodon nigroviridis	SWS1	*Takifugu rubripes*	SWS1	56	32.25
Ornithorhynchus anatinus	Neuropsin	*Danio rerio*	Neuropsin	421.75	16
Gallus gallus	RGR	*Canis familiaris*	RGR	330.4	312.6
Gallus gallus	Neuropsin	*Takifugu rubripes*	Neuropsin	86.5	66
Acyrthosiphon pisum	Insect C	*Culex quinquefasciatus*	Insect C	414	307
Anopheles gambia	Insect C	*Culex quinquefasciatus*	Insect C	294.6	238.5
Homo sapiens	RGR	*Pan troglodytes*	RGR	10	5.7
Homo sapiens	Neuropsin	*Pan troglodytes*	Neuropsin	10	5.7
Mus musculus	Encephalopsin	*Homo sapiens*	Encephalopsin	131.5	61.5
Danio rerio	VA	*Petromyzon marinus*	VA	-1	460.5
Takifugu rubripes	TMT	*Taeniopygia guttata*	TMT	150.9	96.9
Monodelphis domestica	Neuropsin	*Mus musculus*	Neuropsin	171.2	124

Ornithorhynchus anatinus	Neuropsin	*Mus musculus*	Neuropsin	191.1	162.9
Rattus norvegicus	Neuropsin	*Mus musculus*	Neuropsin	14	10.4
Xenopus tropicalis	Parapinopsin	*Uta stansburiana*	Parapinopsin	350.1	330.4
Xenopus tropicalis	RGR	*Canis familiaris*	RGR	350.1	330.4
Oncorhynchus mykiss	Parapinopsin	*Ciona intestinalis*	Parapinopsin	-1	518.5
Danio rerio	VA	*Oryzias latipes*	VA	165.2	149.85
Takifugu rubripes	TMT	*Danio rerio*	TMT	165.2	149.85
Danio rerio	Neuropsin	*Felis catus*	LWS	-1	700
Daphnia pulex	R7	*Culex pipiens*	Red	603	-1
Geotria australis	RH2	*Danio rerio*	RH2	-1	460.5
Geotria australis	SWS2	*Poecilia reticulata*	SWS2	-1	460.5
Geotria australis	LWS	*Homo sapiens*	LWS(G)	-1	460.5
Danio rerio	RGR	*Canis familiaris*	RGR	421.75	416
Nematostella vectensis	151Su	*Geotria australis*	LWS	710	-1
Papilio glaucus	Red	*Nematostella vectensis*	151Su	-1	700
Hasarius adansoni	Red	*Papilio glaucus*	Red	581	515
Hasarius adansoni	Red	*Limulus polyphemus*	Red	-1	445
Drosophila melanogaster	Red	*Papilio glaucus*	Red	414	307
Culex pipiens	Red	*Anopheles gambia*	Red	294.6	238.5
Aedes aegypti	Red	*Anopheles gambia*	Red	294.6	238.5
Papilio glaucus	Red	*Papilio glaucus*	Red	155	56
Petrolisthes cinctipes	Red	*Prorus milleri*	Red	-1	425

Apis mellifera	Red	*Camponotus abdominalis*	Red	243	152
Limulus polyphemus	M1	*Daphnia pulex*	M1	581	515
Triops granarius	M1	*Daphnia pulex*	M1	-1	410
Triops granarius	M2	*Daphnia pulex*	M2	-1	410
Eurpu	M2	*Daphnia pulex*	M1	-1	515
Eurpu	M2	*Portunus pelagicus*	M2	-1	425
Hasarius adansoni	UV	*Aedes aegypti*	UV	581	515
Ixodes scapularis	R7	*Aedes aegypti*	R7	581	515
Daphnia pulex	UV	*Branchinella kugenumaensis*	UV	-1	410
Rhopr	UV	*Aedes aegypti*	UV	307.2	238.5
Lauko	Blue	*Manduca sexta*	Blue	414	307.2
Danio rerio	RH1	*Salmo salar*	RH1	165.2	149.85
Lucania goodei	SWS1	*Cyprinus carpio*	SWS1	165.2	149.85
Gallus gallus	SWS1	*Taeniopygia guttata*	SWS1	86.5	66
Gallus gallus	LWS	*Taeniopygia guttata*	LWS	86.5	66
Bos Taurus	SWS1	*Sus scrofa*	SWS1	65.8	52.4
Loxodonta africana	SWS1	*Cavia porcellus*	SWS1	131.5	62.5
Mus musculus	SWS1	*Cavia porcellus*	SWS1	58.9	52.5
Gorilla gorilla	SWS1	*Macaca fasticularis*	SWS1	34	23.5
Gallus gallus	SWS2	*Geotria australis*	SWS2	-1	460.5
Anolis carolinensis	Pinopsin	*Gallus gallus*	Pinopsin	299.8	255.9
Setonix brachyurus	SWS2	*Cavia porcellus*	SWS2	171.2	124
Mus musculus	SWS1	*Rattus norvegicus*	SWS1	14	10.4
Anolis carolinensis	Pinopsin	*Bufo japonicus*	Pinopsin	350.1	330.4
Loxodonta Africana	LWS	*Callorhinchus milii*	LWS	462.5	421.75
Poecilia reticulata	SWS2	*Carassius auraus*	SWS2	165.2	149.85

Takifugu rubripes	TMT	*Branchiostoma belcheri*	TMT	-1	518.5
Branchiostoma belcheri	TMT	*Canis familiaris*	RGR	-1	518.5
Mus musculus	Encephalopsin	*Culex quinquefasciatus*	Insect C	-1	531.5
Daphnia pulex	Insect C	*Culex quinquefasciatus*	Insect C	581	515
Mus musculus	Encephalopsin	*Danio rerio*	Encephalopsin	421.75	416

Table: Maximum wavelength absorbencies for each opsin in chapter two. This information was used to calculate the ancestral wavelength absorbency prior the each of the opsin duplications. Each available taxon is listed with the maximum absorbency in nanometers (nm).

Taxa	Wavelength	Taxa	Wavelength	Taxa	Wavelength
Schistocerca gregaria	430	*Drosophila melanogaster*	375	*Prorus milleri*	522
Pieris rapae	425	*Drosophila melanogaster*	345	*Drosophila melanogaster*	420
Pieris rapae	453	*Hemigrapsus oregonensis*	480	*Drosophila melanogaster*	478
Heliconius erato	470	*Hemigrapsus oregonensis*	480	*Calliphora vicina*	490
Lycaena rubidus	500	*Limulus polyphemus*	530	*Apis mellifera*	499
Lycaena rubidus	437	*Limulus polyphemus*	520	*Schistocerca gregaria*	520
Danaus plexippus	435	*Archaeomysis grebnitzkii*	496	*Sphodromantis viridis*	515
Papilio glaucus	460	*Mysis diluviana*	501	*Apis mellifera*	534
Manduca sexta	450	*Euphausia superba*	487	*Cataglyphis bombycinus*	510
Papilio glaucus	360	*Neogonodactylus oerstedii*	528	*Camponotus abdominalis*	510
Manduca sexta	357	*Neogonodactylus oerstedii*	489	*Papilio glaucus*	515

Pieris rapae	360	*Neogonodactylus oerstedii*	522	*Lycaena rubidus*	568
Heliconius erato	370	*Homarus gammarus*	515	*Heliconius erato*	570
Danaus plexippus	340	*Holmesimysis costata*	512	*Danaus plexippus*	545
Lycaena rubidus	360	*Neomysis americana*	520	*Papilio glaucus*	530
Camponotus abdominalis	360	*Cambarellus shufeldtii*	526	*Papilio glaucus*	575
Cataglyphis bombycinus	360	*Cambarus ludovicianus*	529	*Pieris rapae*	563
Bombus impatiens	350	*Orconectes australis*	530		
Apis mellifera	353	*Procambarus clarkii*	533		