Correlated Evolution of Teleost Fish Rhodopsin with Habitat:
Linking Molecular Changes to Ecology

By

Amir-Arsalan Sabouhanian

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Department of Cell and Systems Biology
University of Toronto

© Copyright by Amir-Arsalan Sabouhanian, 2015
Correlated Evolution of Teleost Fish Rhodopsin with Habitat:
Linking Molecular Changes to Ecology

Amir-Arsalan Sabouhanian
Master of Science
Department of Cell and Systems Biology
University of Toronto
2015

Abstract
A critical challenge in molecular evolutionary biology is to determine how changes in protein coding sequence correlate with ecology, physiology, or morphology. While statistical approaches aiming to investigate signatures of selection at the molecular level have been developed, methods designed to correlate these findings with macroscopic traits remain understudied. Here, I investigated amino acid replacements in the visual pigment rhodopsin (RH1) of teleost fishes that show significant association with shifts in habitat, using both existing phylogenetic comparative methods, as well as novel implementation of sequence simulations and ancestral reconstructions. Through an unbiased examination of amino acid sites in RH1, I found several substitutions that are known to affect spectral sensitivity or other properties of rhodopsin. Thirteen additional control genes showed markedly weaker associations between amino acid substitutions and habitat compared with teleost RH1. The framework established in this thesis can similarly be used to study other genes and macroscopic traits.
Acknowledgements

I would like to thank my supervisor Belinda Chang, for her guidance and trust on me since I started as an undergraduate in her lab. This project would not have been possible without Belinda’s incredible support, patience and supervision. I am also very grateful to Alan Moses and Jason Weir, who graciously agreed to be on my supervisory committee; their invaluable comments and feedback greatly improved this thesis.

I would like to thank David Yu for initiating this project and his insight on this work, Alex Van Nynatten for his invaluable help and advice in the development of the illustrations and figures used here, and Bhawandeep Panesar and Samantha Beddington for their scripts to automate several steps of this study.

All other past and present Chang lab members, including Nihar Bhattacharya, Gianni Castiglione, Jing Du, Sarah Dungan, Eduardo Gutierrez, Frances Hauser, James Morrow, Benjamin Scott, Ilke van Hazel, Cameron Weadick, and especially Ryan Schott also helped me a lot in various aspects of my work over years; I thank them for all their feedback on my oral presentations, posters, and written drafts. Additional thanks go to Nihar Bhattacharya, Gianni Castiglione, Frances Hauser, Ryan Schott, and Alex Van Nynatten for proofreading my thesis.

I am very thankful to my family and friends for their overwhelming encouragement and support throughout my studies.

Lastly, I would also like to thank for the funding provided by NSERC for this project.
Table of Contents

Abstract .................................................................................................................................................. ii
Acknowledgements .............................................................................................................................. iii
Table of Contents ................................................................................................................................ iv
List of Tables ......................................................................................................................................... vi
List of Figures ....................................................................................................................................... vii

Chapter 1: General Introduction ....................................................................................................... 1

1.1 Phylogenetic Comparative Methods ....................................................................................... 1
1.2 Codon Based Models for Evolution of Molecular Sequences ............................................. 3
1.3 Ancestral Reconstruction of Sequences, Discrete and Continuous Characters ............... 5
1.4 Vision, visual pigments, and phototransduction cascade .................................................... 9
1.5 Teleost Fishes .......................................................................................................................... 14
1.6 Vision in the deep-sea .............................................................................................................. 15
1.7 Objectives ................................................................................................................................... 17

Chapter 2: Correlated Evolution of Euteleost Fish Rhodopsin with Habitat Depth: Linking
Molecular Changes to Ecology of Organisms ................................................................................. 24

2.1 Abstract ....................................................................................................................................... 24
2.2 Introduction ................................................................................................................................... 25
2.3 Materials and Methods ............................................................................................................... 29

2.3.1 Sequence Collection, Alignment and Phylogeny Construction .................................. 29
2.3.2 Compiling Ecological Data and Testing Models of Evolution .................................. 30
2.3.3 Statistical Analyses and Correcting for Phylogenetic Signal ......................................... 31
2.3.4 Correlations in Other Genes Used as Control ............................................................... 33
2.3.5 Homology Modeling of Amino Acid Substitutions on RH1 Crystal Structure ........... 33

2.4 Results ......................................................................................................................................... 34

2.4.1 Non-phylogenetic Kruskal-Wallis Tests ......................................................................... 34
2.4.2 Phylogenetic Generalized Least Squares ...................................................................... 35
List of Tables

Chapter 2: Correlated Evolution of Euteleost Fish Rhodopsin with Habitat Depth: Linking Molecular Changes to Ecology of Organisms

Table 1. Summary of sites in rhodopsin which correlate with habitat depth after phylogenetic correction, using the time-calibrated phylogeny ................................................................. 46
Table 2. Comparison of correlations with depth in RH1 to control genes used in this study ... 47
Supplementary Table S1. Summary of sites in rhodopsin which correlate with habitat depth after phylogenetic correction, using codon or unity branch lengths ........................................SM1
Supplementary Table S2. Number of sites in correlation with habitat depth when RH1 data set is pruned to only include sequences from species with available control gene sequence .....SM2
Supplementary Table S3. Likelihood values, corrected Akaike Information Criterion, and parameter estimates of the models tested for habitat depth .......................................................SM2
Supplementary Table S4. Comparison of false positive rates of non-phylogenetic and phylogenetic tests for different branch lengths, estimated from sequence simulations ........SM3
Supplementary Table S5. Comparison of number of RH1 sites in correlation with habitat depth across different methods and branch length specification .......................................................SM3
Supplementary Table S6. Accession numbers of sequences downloaded from NCBI……SM4

Chapter 3: Investigating Correlations between habitat water temperature of teleost fishes and amino acid substitutions in rhodopsins

Table 1. Summary of sites in rhodopsin which correlate with habitat water temperature after phylogenetic correction ........................................................................................................ 64
Table 2. Likelihood values, corrected Akaike Information Criterion, and parameter estimates of the models tested for habitat water temperature .......................................................... 66
List of Figures

Chapter 1: General Introduction

Figure 1. Effect of phylogenetic relatedness on statistical associations .......................... 20
Figure 2. Crystal structure of rhodopsin in the dark and light-activated states .................. 21
Figure 3. Chronogram of ray-finned fishes ................................................................. 22
Figure 4. Attenuation of sunlight with increased depth in marine habitats ...................... 23
Figure 5. Vertebrate phototransduction cascade .......................................................... 23

Chapter 2: Correlated Evolution of Euteleost Fish Rhodopsin with Habitat Depth: Linking Molecular Changes to Ecology of Organisms

Figure 1. Phylogeny of euteleost fishes in this study and their habitat depths ..................... 48
Figure 2. Distribution of p-values for the sites in rhodopsin compared to simulated sequences, across the four approaches taken in this study ...................................................................... 49
Figure 3. Box-plots showing the difference in habitat depth across different residues for three RH1 sites correlating with habitat depth .................................................................................. 51
Figure 4. (a) Sequence logo showing the change in prevalence of residues at some of the significant sites in accord with photic zones (b) Substitutions at site 292 in parallel with inferred changes in habitat depth .................................................................................. 52
Figure 5. RH1 crystal structure showing sites which correlate with habitat depth .............. 53
Figure 6. RH1 crystal structure showing four substitutions near the chromophore ligand binding pocket which correlate with habitat depth ........................................................................... 54
Figure S1. Chronogram showing the relationship between euteleost fishes used here.......SM13
Figure S2. Rhodopsin snake structure showing significant sites .....................................SM18
Figure S3. Box-plots showing the difference in habitat depth across different residues for some of the RH1 sites correlating with habitat depth .....................................................................SM19
Figure S4. Distribution of p-values for the sites in rhodopsin compared to the control genes, across the four approaches taken in this study .........................................................SM20
Chapter 3: Investigating Correlations between habitat water temperature of teleost fishes and amino acid substitutions in rhodopsins

Figure 1. Box-plots showing the difference in water temperature across different residues for some of the sites with significant correlation................................................................. 67

Figure 2. RH1 crystal structure showing six substitutions which appear to have occurred in parallel and correlate with habitat water temperature in the ostariophysan data set................. 69

Figure S1. Phylogeny of ostariophysan fishes used in this study ..............................................SM47
Chapter 1: General Introduction

Phylogenetic Comparative Methods

Comparative approaches to the study of organismal form and function have been among the most commonly used and productive methods to study evolutionary processes (Garland, et al. 2005). However, performing statistical analyses on trait values of extant taxa violates the underlying assumption of independence of data points (Felsenstein 1985). This is because species are differentially related due to their shared ancestry, making some species more closely related than others. Phylogenetic comparative methods (PCMs) address this issue of “phylogenetic signal” by incorporating a phylogeny and an explicit model of character evolution in their estimations (Garland, et al. 2005). Felsenstein’s Independent Contrast (IC) is perhaps the best known and most widely used PCM; it calculates the difference for values of a trait between two nodes, standardizes the difference by branch lengths, and calculates the correlation between the standardized contrasts of the two traits instead of the raw data at the tips (Felsenstein 1985). Originally, Independent Contrast assumed a Brownian Motion (BM) model of evolution and forced the regression line through the origin; however, these assumptions were later relaxed in the Phylogenetic Generalized Least Squares (PGLS) method (Grafen 1989; Martins and Hansen 1997). In PGLS, the phylogenetic tree is converted into a variance-covariance matrix with the main-diagonal indicating the expected variances, taken as the total branch distance from the root to each tip, and the off-diagonal elements are the distance from the root to the last common ancestor of each tip (Garland, et al. 2005). IC can be considered a special case of PGLS and given the same tip data, phylogeny and model of evolution, they yield similar results (Blomberg,
et al. 2012; Garland, et al. 2005). For an ultrametric tree, i.e., a phylogeny with contemporaneous tips, these methods would generate statistics identical to conventional, non-phylogenetic methods if the hierarchical structure of the tree is ignored and a star phylogeny is assumed (Figure 1). In order to evaluate sensitivity of conclusions to errors in the phylogeny or the model of evolution, some studies conduct their analyses using different tree topologies, branch lengths, and evolutionary models (Laurin 2004; Symonds and Elgar 2002). The evolution of traits can also be conducted under a Bayesian framework to account for uncertainties in the phylogeny and parameter estimations (Botero, et al. 2014; Hadfield 2010; Huelsenbeck and Rannala 2003).

To investigate correlated evolution of two binary characters, the most widely used approach is to fit two nested models and compare their likelihood scores given a bifurcating phylogeny and extant character states (Pagel 1994). The null model (Independent evolution model) has four parameters which are transition rates (q) between two states of each character. If $A_0$ and $A_1$ represent the two states of character A while $B_0$ and $B_1$ denote the two possible states of character B, four transition rates sufficiently describe the independent evolution of A and B: $q_1$ for $A_0$ to $A_1$, $q_2$ for $A_1$ to $A_0$, $q_3$ for $B_0$ to $B_1$, and $q_4$ for $B_1$ to $B_0$. However, if A and B have evolved in a correlated fashion, a model that estimates transition rates between two states of one character depending on the state of the second character will provide a better fit to the data. The alternate (Dependent) model, has eight parameters which correspond to $q_1$ for $A_0$ to $A_1$ when state at B is 0 ($B_0$), $q_2$ for $A_0$ to $A_1$ when state at B is 1 ($B_1$), and so on. The maximum likelihood values of Independent and Dependent models can be compared via a Likelihood Ratio Test (LRT) with four degrees of freedom (the difference in the number of parameters between the two models) to determine whether the more complex, alternate model fits the data significantly better than the null. Sometimes, the transition rates may not all differ significantly from each other; for
instance, forward and reverse rates might be equal for one or both characters, in which case the
degrees of freedom (d.o.f.) may need to be reduced to three or two. One can constrain some of
the transition rates to be equal, compare the likelihood value to a model without any constraint to
evaluate whether they differ significantly, and then adjust the number of parameters (transition
rates) in the null and alternate models accordingly. Comparison to a simulated distribution
instead of a standard chi-square table is also suggested. Nevertheless, four d.o.f. should provide
the most stringent alternate compared to the null model and significant results indicate high
confidence in correlated evolution of two traits. As an alternative to the ML approach, Bayesian
estimation of parameters can be achieved through MCMC simulations. Uncertainties in tree
topology and branch lengths can be accounted for by using more than one phylogeny and
significance established by calculating Bayes factors (Huelsenbeck and Rannala 2003; Pagel and
Meade 2006; Pagel, et al. 2004). An important consideration when utilizing the phylogenetic
comparative methods is that they discern between traits that show convergent evolution from
traits which show similarity as a result of shared ancestry. However, from an adaptive
standpoint, maintenance of a trait in species after divergence from their ancestor could still
indicate the adaptive significance of that trait (Thornhill and Fincher 2013).

**Codon-Based Models for Evolution of Molecular Sequences**

Comparing rates of non-synonymous ($d_N$) to synonymous ($d_S$) substitutions in protein-
coding DNA sequences has provided a framework to identify instances of adaptive molecular
evolution. $d_N$ and $d_S$ are defined as the number of non-synonymous substitutions per non-
synonymous sites and the number of synonymous substitutions per synonymous sites,
respectively, over an evolutionary timescale. Their ratio, $\omega = d_N/d_S$, can be used to estimate the
type of selective pressure acting on a gene; \( \omega = 1 \) denotes neutral selection where rate of fixation of amino acid substitutions is the same as synonymous changes. Deleterious amino acid substitutions would be fixed at a lower rate than synonymous substitutions, so the gene would be under negative purifying selection with \( \omega < 1 \). Diversifying positive selection, \( \omega > 1 \), is the dominant force when amino acid replacements are fixed at a higher rate than synonymous mutations. The most widely used methods to estimate \( d_N \) and \( d_S \) currently use explicit models of codon evolution under a maximum likelihood (ML) or Bayesian framework. These models commonly include parameters for transition to transversion rates (\( \kappa \)), estimates of equilibrium codon frequencies, and use inferences of phylogenetic relationships between sequences to estimate \( \omega \).

Purifying selection tends to dominate in evolution of functional genes, since mutations at most sites can disrupt gene function. Because of this, the rate of non-synonymous substitutions in the remaining sites needs to be extremely high for the gene to have an overall \( d_N / d_S > 1 \) (Nielsen 2005). This assumption, except for few genes such as those involved in pathogen and host immune response mechanisms, is generally unrealistic and so codon models have been developed to allow for rate heterogeneity among sites. These models either estimate \( \omega \) at each individual site or utilize a statistical distribution, typically gamma or beta distribution, to describe variation of \( \omega \) across sites by assigning them to different site classes drawn from that distribution. One can then test if the model allowing for rate heterogeneity has significantly higher likelihood score than the constant rate model. Additionally, a likelihood ratio test can be conducted to determine whether an alternate model allowing for a class of sites with \( \omega > 1 \) fits the data significantly better than the null model which does not have a positively selected site class (Yang 2002). Bayes theorem can then be used to identify individual sites with significantly
elevated $\omega > 1$ based on their posterior probabilities. These site models assume equal substitution rates across all branches of the phylogeny; models have been developed that allow $\omega$ for the entire gene or a subset of sites to vary across all or specific branches or between clades. Similar to rate heterogeneity across sites, assumption of rate homogeneity across the phylogeny can then be tested using an LRT.

Codon-based ML and Bayesian models have their own limitations: they cannot detect selection that does not cause excessive non-synonymous substitutions such as balancing selection, very divergent or similar sequences can make parameter estimations unreliable, and mechanistic codon-based models also generally assume all possible amino acid changes have the same $d_S/d_S$ ratio. This latter assumption can be especially problematic because amino acid substitution rates tend to correlate with their electrochemical properties. Unlike the mechanistic codon models, amino acid substitution rate matrices are typically estimated from one or more protein families. These empirical matrices reflect substitution tendencies between amino acids which can differ from one protein family to another. However, some more general matrices, such as WAG, represent average substitution tendencies over many protein families by sacrificing resolution at the level of genes (Miyazawa 2013). Despite the limitations of codon-based models, they were still found to fit the data significantly better than the existing empirical amino acid models, reflecting their importance for molecular evolutionary studies (Seo and Kishino 2008).

Ancestral Reconstruction of Sequences, Discrete and Continuous Characters

Ancestral Sequence Reconstruction (ASR) is generally accomplished via maximum likelihood or Bayesian methods. Given a model of nucleotide, codon or amino acid evolution, an alignment of gene sequences, and a phylogeny for those sequences, ancestral estimates of states
at each site in the alignment and at each internal node can be calculated (Hanson-Smith, et al. 2010). For discrete characters, including molecular sequences, maximum likelihood estimation can be achieved through either joint or marginal reconstruction. The former considers all ancestral nodes at the same time and estimates the set of characters at a site that maximizes the likelihood of reconstructed ancestral states. In marginal reconstruction, the character state at one ancestral node that maximizes likelihood for that particular node, integrating over all other nodes and states, is chosen as the best reconstruction. In the program Phylogenetic Analysis by Maximum Likelihood (PAML), marginal reconstruction is implemented when heterogeneity of substitution rates among sites is assumed, while both joint and marginal reconstruction methods work for the constant rate across sites model. However, marginal and joint reconstructions usually give consistent results except for when the posterior probabilities of two or more reconstructed states at a site are similar. PAML can fit a number of codon-based models and calculate their likelihood values. Ancestral sequences can be reconstructed under these models along with estimations of branch lengths, $\kappa$, and $\omega$ parameters. Some of the most commonly used random-site models include: M0, which assumes equal non-synonymous to synonymous substitution rates across all sites; M1a and M2a allow a proportion of sites to have $\omega < 1$, and a second site class to have $\omega = 1$ while M2a has an additional site class which can have $\omega > 1$; M3 estimates $\omega$ for a predetermined number of site classes without restricting $\omega$ values; M7 and M8 use a discrete beta distribution to fit variability of $\omega$ across sites for a specified number of site classes where $0 \leq \omega \leq 1$, while M8 has an additional site class which can have $\omega > 1$ (Yang, et al. 2000). As mentioned previously, while these random-site models allow for rate heterogeneity across sites, they assume these rates are equal over the entire tree. However, substitution rates may also vary across different clades or branches of the phylogeny, in which case use of branch-
site or clade models may provide a better fit to the data. The substitution process is assumed to be homogeneous over time due to computational and mathematical considerations, rather than biological reality (Anisimova and Kosiol 2009). If a priori information exists about branches or clades of the tree along which the gene might have evolved at a different rate, using models that allow rate heterogeneity across the phylogeny may result in better estimation of ancestral sequences. In the absence of a priori information, however, ancestral sequence reconstructions from existing models estimating different $\omega$ values for each branch of the tree are not computationally feasible, especially for large datasets, and can also lead to over-fitting the data. Uncertainties in inferences of ancestral reconstructions can introduce systematic and random biases in subsequent estimations (Anisimova and Kosiol 2009; Nielsen 2002; Yang 2002). Bayesian methods incorporate uncertainty in phylogeny or parameter estimates by integrating likelihoods over a distribution of trees or parameter values, so they might provide more reliable inferences of ancestral sequences. However, it has been shown that incorporating phylogenetic uncertainty very rarely changes the inferred ancestral states and does not improve the accuracy of reconstructions in most cases (Hanson-Smith, et al. 2010). Ancestral sequence reconstructions using codon models that incorporate electrochemical properties of amino acids may enhance the accuracy of inference as well. Combined empirical and mechanistic codon models are being developed which appear to provide better fit to the data (Anisimova and Kosiol 2009; Doron-Faigenboim and Pupko 2007; Kosiol, et al. 2007), but whether they actually provide more accurate parameter estimates and reconstructions requires further investigation.

Continuous characters can also be reconstructed given a phylogeny and a model of evolution. The simplest model, BM, is a neutral model that assumes the continuous character of interest has evolved completely due to random drift (i.e., changes are proportional to branch
lengths with longer branches showing greater changes). The Ornstein-Uhlenbeck (OU) process has an additional selection-strength parameter, alpha, and fits a random walk with a tendency towards an optimum or multiple optima. Another routinely used model is the Early-burst, EB, or accelerating-decelerating, ACDC, model which estimates if the rate of evolution increases or decreases exponentially over time (Blomberg, et al. 2003; Harmon, et al. 2010). Three models proposed by Pagel, lambda, delta, and kappa, have parameters which can increase the fit of BM model over the tree by transforming branch lengths (Pagel 1999). Lambda can be interpreted as the level of phylogenetic association and multiplies internal branches with its estimated value ($\lambda = 0$ produces a star phylogeny which would result in statistics equivalent to non-phylogenetic tests in the case of ultrametric trees because species would be equally distant from each other, while $\lambda=1$ implies Brownian Motion and keeps branches unchanged). Kappa raises branch lengths to an estimated power, implying more changes than estimated by BM for longer branches if $\kappa > 1$ and less changes than expected by BM for longer branches if $\kappa < 1$ ($\kappa = 0$ sets all branch lengths equal to unity and is the punctuational mode of evolution where trait divergence is proportional to number of speciation events). Delta estimates changes in the rate of trait evolution from the root to the tips and adjusts branch lengths accordingly; $\delta > 1$ suggests recent rate of evolution has been relatively fast, while $\delta < 1$ suggests recent evolution has been relatively slow (Harmon, et al. 2008; Pagel 1999). These branch length transformations mostly have a statistical perspective rather than mimicking a biologically meaningful evolutionary model.

For discrete characters, maximum likelihood reconstruction of ancestral states uses Markov models to estimate transition rates between different states. The simplest, one-parameter model, Mk1, assigns equal rates for forward and backward transitions between all states (Lewis
In the Asymmetric model, different forward and backward rates are estimated which are equal for transitions between all states. The most general model has a unique parameter for each transition, which can be constrained to get less complex models including the Mk1 or Asymmetric models. A likelihood ratio test can be performed for nested models, or Akaike Information Criterion can be implemented if models are not nested, to determine if any of the transition rates do not differ significantly from each other and can be constrained to be equal. Regardless of whether a continuous or discrete character is being reconstructed, confidence in reconstructions would be higher if inferences are consistent across different evolutionary models and phylogenies.

Vision, Visual Pigments, and the Phototransduction Cascade

The visual system is a primary means of interpreting the external world for most vertebrates, and has largely evolved in conjunction with the ecology of organisms (Yokoyama 2008). The vertebrate retina contains several layers of neural cells, the deepest of which are photoreceptors that absorb light and relay the signal to horizontal, bipolar, amacrine, and ganglion cells and finally to the brain through the optic nerve (Lamb, et al. 2007). There are two types of photoreceptor cells, rods and cones, which function under dim and daylight conditions, respectively. Rods are more sensitive to light than cones, but cones have faster photoresponse kinetics and are used for color discrimination (Shichida, et al. 2013). Within the photoreceptor cells are the visual pigments: light-sensitive molecules consisting of an integral protein moiety covalently bound to a chromophore via a Schiff base linkage to a conserved lysine residue (Bowmaker 2008; Yokoyama 2008). In the ground state of opsins, except for UV-sensitive cone pigments, the Schiff base is protonated. The terminal amino group of the lysine (K296) has a net
positive charge, which helps to stabilize the dark-state configuration via formation of a salt bridge with the negative charge on an adjacent glutamate residue (E113), called the “counterion” (Tsutsui and Shichida 2010). The chromophore is an aldehyde of vitamin A, either 11-cis-retinal or 11-cis-3,4-dehydroretinal, and the integral proteins, opsins, are members of a class of seven transmembrane domain G protein-coupled receptors approximately 350 amino acids long which enclose a ligand-binding pocket (Okada, et al. 2001). Vertebrate visual pigments are divided into five major groups according to their amino acid sequences. Rhodopsin (RH1) is the pigment typically expressed in rods and is maximally sensitive to wavelengths of light between 480–510 nm (Yokoyama 2002). The rod photoreceptor outer segment is packed with rhodopsin molecules, which increases the probability of absorption of light (Lamb and Pugh 2004). The other four classes of opsins are generally expressed in cones and include: short wavelength-sensitive type 1 (SWS1) which absorbs in ultraviolet to violet region (355–440 nm), short wavelength-sensitive type 2 (SWS2) maximally sensitive to violet to blue region of the visible spectrum (410–490 nm), RH1-like (RH2) absorbing in green region between 480–535 nm, and the long to middle wavelength-sensitive opsin (LWS/MWS) with maximal sensitivity between 490–570 nm (Bowmaker 2008). Interaction of amino acid side chains in the binding pocket with the retinal chromophore is the primary factor influencing the spectral sensitivity of these visual pigments (Bowmaker 2008). A series of gene duplication events has given rise to different classes of opsins and phylogenetic analysis shows that RH1 and RH2 are more closely related to each other, then to SWS2, SWS1, and LWS/MWS in that order (Yokoyama 2008).

Vision is initiated upon absorption of a photon of light by the retinal chromophore, which is isomerized from an 11-cis to all-trans conformation (Shichida, et al. 2013). Isomerization of the chromophore puts the protein in a strained configuration, so it undergoes a series of
conformational changes back to a lower energy state. The first several intermediates are very short-lived and within a few microseconds of photoisomerization, metarhodopsin I (MI) intermediate is formed (Lamb and Pugh 2004). Within a millisecond at mammalian body temperature, MI is converted to the active intermediate metarhodopsin II (MII), which is accompanied by deprotonation of the Schiff base (Lamb and Pugh 2004). Conversion of rhodopsin to MII involves two major conformational changes: first is motion of transmembrane 5 (TM5) towards TM6 linked to a rearrangement of hydrogen-bonding network between residues in TM3-TM5, and the second change involves rotational tilt of TM6 resulting in its cytoplasmic end to move away from TM3 ((Choe, et al. 2011); Figure 2). Deprotonation of the Schiff base linkage during MII formation shifts the wavelength of maximal absorbance ($\lambda_{\text{max}}$) of visual pigments from the visible to near UV region of the spectrum (Lamb and Pugh 2004). MII can be further converted to metarhodopsin III in which the Schiff base becomes protonated again (Lamb and Pugh 2004), however, MII is the species that maximally activates the heterotrimeric G protein transducin (Hofmann 1986). Binding of the light-activated MII to the G protein accelerates its GTPase activity which leads to dissociation of $\alpha$- and $\beta\gamma$-subunits for hundreds of transducin molecules per second (Shichida, et al. 2013). The G protein $\alpha$ subunit then activates cGMP-phosphodiesterase which decreases cGMP concentration in the outer segment of retina resulting in the closure of cGMP-gated cation channels in the plasma membrane, ultimately causing hyperpolarization of the photoreceptor cells (Shichida, et al. 2013). The hyperpolarization leads to decreased release of neurotransmitter glutamate at the photoreceptor terminal (Arshavsky, et al. 2002), which results in opening of cation channels and elevated levels of cGMP in ON-bipolar cells causing their depolarization (Shiells and Falk 1990). The signal is eventually transmitted to ganglion cells, and from there to the brain via the optic nerve (Shichida,
et al. 2013). Visual pigments can also undergo spontaneous thermal activation in the absence of light; since this event, called “dark noise”, produces a cellular response identical to photoactivation, it can interfere with the function of photoreceptors (Kefalov 2012). Because cone opsins have lower activation energy than rhodopsin, they have substantially higher rates of thermal activation which contributes to the lower photosensitivity of cones compared to rod cells (Kefalov 2012). It is also postulated that the level of dark noise is related to spectral properties of the pigments (Gozem, et al. 2012; Liu, et al. 2011).

Termination of the phototransduction cascade is initiated when rhodopsin kinase phosphorylates the active MII and the capping protein arrestin binds to the phosphorylated metarhodopsin. The covalent bond between all-trans retinal and the protein moiety is then hydrolyzed and the chromophore is released (Lamb and Pugh 2004). Although the active MII is rapidly inactivated, it is unable to signal the arrival of another photon because its all-trans retinoid needs to be replaced by a new molecule of 11-cis retinal (Lamb and Pugh 2004). In vertebrates, the conversion back to 11-cis conformation is achieved through a long series of reactions known as the retinoid cycle (Lamb and Pugh 2004). The reconverted 11-cis retinal then binds to an opsins, forms a covalent bond with it, and the visual pigment is regenerated (Kefalov 2012). In rods, 11-cis retinal is supplied from the adjacent retinal pigment epithelium while there appears to be a second pathway in cones, involving Muller cells, which contributes to the faster regeneration of cone opsins (Blomhoff and Blomhoff 2006; Lamb and Pugh 2004).

The best studied aspect of visual pigment function is the effect of amino acid replacements on the wavelength of maximal absorbance, $\lambda_{\text{max}}$ (Fasick and Robinson 2000; Hunt, et al. 2001; Sugawara, et al. 2005). Visual pigments of fishes inhabiting greater depths are generally blue-shifted compared to shallow-dwelling species and terrestrial organisms; for
instance, microspectrophotometric data obtained from rods and cones of cottoid fish from lake Baikal demonstrate a general blue-shifting trend with increase in depth (Bowmaker, et al. 1994). Most mutations causing a shift in $\lambda_{\text{max}}$ are near the chromophore binding pocket and show a change in polarity. D83N, E122Q and A292S are natural variants which are among the best-characterized substitutions that blue-shift rhodopsin absorption spectra. Effects of mutations on other properties of opsins are not as extensively investigated, but are receiving increasing attention. For instance, mutations in proximity to the two retinal channels in rhodopsin, one between TM1/TM7 and the other between TM5/TM6, are shown to alter the rate of retinal entry and/or release (Piechnick, et al. 2012), which may be important for length and intensity of signalling and regeneration time of rhodopsin, although other members of the phototransduction cascade may set these rates in-\textit{vivo} as well (Kefalov 2012). Similarly, effects of mutations on the level of transducin activation (Marin, et al. 2000), thermal stability of rhodopsin (Liu, et al. 2013), and hydroxylamine reactivity, an indication of how susceptible the Schiff base is to attack by the bulk solvent (Janz, et al. 2003; Noorwez, et al. 2004), are studied. The rate of active MII formation can also have adaptive significance. Faster MII formation caused by D83N substitution in RH1 of some cichlids and bats is hypothesized to be a dim-light adaptation to increase photosensitivity (Sugawara, et al. 2010). In native membranes, rhodopsin molecules, and GPCRs in general, can exist as dimers or higher order oligomers which may serve as platforms for activation of G protein (Fotiadis, et al. 2006; Guo, et al. 2005; Jastrzebska, et al. 2006). The dimerization interface appears to be between TM4-TM5 (Fotiadis, et al. 2006), and mutations near this interface may affect stability of rhodopsin in the lipid membrane as well as its interaction with transducin.
Rhodopsin is the most widely studied visual pigment and its role in visual pathway is relatively well-established. It is also used as a phylogenetic marker to determine evolutionary relationships; hence, the rhodopsin gene from numerous species has been sequenced and deposited in NCBI. Its 3D structure has been resolved in both the inactive dark state and active MII conformation, providing an opportunity to investigate how amino acid substitutions may alter rhodopsin structure \textit{in silico}. All these features make rhodopsin an ideal candidate gene on which to conduct comparative analyses.

Teleost Fishes

Fishes are a very diverse group of aquatic vertebrates which include jawless fishes (Agnatha), cartilaginous fishes (Chondrichthyes), and bony fishes (Osteichthyes); the latter group is more closely related to tetrapods than to jawless and cartilaginous fishes (Volff 2005). Actinopterygians, ray-finned fishes, are a class of Osteichthyes which constitute more than 30000 species and represent over 95% of all living fish and almost half of all extant vertebrates ((Near, et al. 2012); Figure 3). Teleosts comprise more than 99.8% of ray-finned fishes and exhibit an intriguing level of diversity which is evident in their ecology, morphology, and behavior (Volff 2005). Teleost species live in almost all marine and freshwater habitats from deepest regions of oceans, coral reefs, to freezing Antarctic waters (Near, et al. 2013). The diversity in habitats can exert different selective pressures on evolution of sensory systems in these fishes, including the visual system. This habitat diversity along with recent advances in resolving taxonomic relationships and divergence time estimations among ray-finned fishes (Betancur, et al. 2013; Near, et al. 2012) provide a suitable framework for conducting comparative analyses on these species.
Vision in teleost fishes ranges from rod monochromacy in some deep-sea species to tetrachromacy in a number of shallow water-dwelling fish. Adult teleosts inhabiting epipelagic zone, the illuminated zone at the surface of sea and ocean, tend to have lost or no longer express their SWS1 or LWS or both, leaving them with dichromatic or trichromatic vision, while fishes living in highly turbid or stained waters have generally kept their longer-wave LWS and RH2 opsins (Bowmaker 2008). These variations generally reflect differences in wavelengths of light most abundant in their environments. It has been proposed that a whole-genome duplication occurred early in the evolution of ray-finned fishes and consequently they generally have more copies of opsin genes than other vertebrates. Throughout evolution, some species have experienced expansion and contraction of the opsin gene family through tandem duplication(s) of individual genes, pseudogenization (mutations rendering certain opsin genes non-functional), or neo-functionalization (gene duplicates with disparate function) (Bowmaker 2008; Ohno 1970).

**Vision in the Deep Sea**

The deep sea can be defined as “that part of the ocean beyond the edge of the continental shelf, giving it an upper limit of around 200 m” (Douglas, et al. 1998). The upper 200 meters is called the epipelagic zone and is also the region where sufficient light is available for photosynthesis. Down-welling sunlight becomes progressively dimmer and bluer with increase in depth (Figure 4); it is the extended light source up to nearly 1000 meters, the mesopelagic zone, but in the bathypelagic zone, below 1000 m, the only available source of light becomes the point-like bioluminescence produced by marine species (Warrant and Locket 2004). With the exception of several genera of dragonfish which emit red light, most bioluminescence produced by marine organisms is blue (Herring and Cope 2005). Eye design shows the greatest variation in
species inhabiting the mesopelagic zone where point-like (bioluminescent flashes) and extended (dim, down-welling sunlight) sources of light may be visible simultaneously (Warrant and Locket 2004). Fishes in the bathypelagic zone generally have smaller eyes and are therefore less sensitive to light sources, but have higher spatial resolution ability, which is well-adapted to localisation of bioluminescent flashes at ecologically meaningful distance (Warrant and Locket 2004). The cost of good spatial resolution is reduced sensitivity to a dim, extended scene because each ganglion cell (which performs the final level of visual processing in the retina and their axons carry the signal to brain via the optic nerve) receives inputs from a pool of photoreceptors, typically rods in deep-sea fishes, via bipolar cells (Hughes 1977). As a result, smaller pools provide higher resolution, but are also less sensitive (Warrant and Locket 2004). Larger eyes with larger pupils can capture more light and with increase in depth in the mesopelagic zone, eye size to body size ratio tends to get larger, although with many exceptions. The trend reverses upon reaching the bathypelagic zone such that eyes tend to get smaller again below 1000m while relative size of pupil to eye generally becomes greater, which increases fishes ability to localise bioluminescent flashes (Warrant and Locket 2004). Large eyes with high photosensitivity might actually be disadvantageous for bathypelagic fishes, since smaller and less sensitive eyes are both energetically less expensive and also avoid detection of distant flashes that might be useless to pursue because the flash source is transient (Warrant and Locket 2004). However, some bathypelagic fishes, such as alepocephalids, still have large and highly developed eyes and presumably rely on vision more than on other senses (Warrant and Locket 2004).

Nearly 90% of deep-sea fishes investigated to date have a single visual pigment, as expected for species inhabiting a near-monochromatic world (Warrant and Locket 2004). Another adaptation for vision of deep-sea fishes proposed to increase sensitivity is that instead of
having a single layer of rods in their retina, their rods are often stacked several layers deep (Warrant and Locket 2004). Most deep-sea fishes also have visual pigments with a blue-shifted \( \lambda_{\text{max}} \) between 468-494 nm, which would better match the attenuated sunlight wavelengths available in their habitat, but also appear to be better adapted for maximizing sensitivity to bioluminescent sources than to down-welling daylight (Warrant and Locket 2004). Another consideration regarding photosensitivity, as discussed earlier, is that the phototransduction cascade can also be activated by thermal energy and produce electrical signals indistinguishable from those produced by light sources (Barlow 1957). This “dark noise” sets the limit of photosensitivity and can contaminate visual signals, especially at very low light levels (Aho, et al. 1988). Although levels of dark noise diminish at low water temperatures in the deep sea, they are still sufficient to limit visibility of bioluminescent flashes (Warrant and Locket 2004). Altogether, marine eyes must adapt to not only the light intensity, but also to its spatial organization, level of visual pigment thermal isomerization, and other factors which might impact visual performance upon increases in depth (Warrant and Locket 2004).

**Objectives**

In order to link changes in macroecological patterns to substitutions in molecular sequences, this study aims to investigate how habitat shifts of teleost fishes correlate with amino acid substitutions in the visual pigment rhodopsin. The main focus of the thesis is to correlate habitat depths of marine euteleosts with substitutions in RH1 under a phylogenetic framework. Additionally, similar associations with habitat temperature are also investigated in both marine euteleosts and freshwater ostariophysans. Ostariophysi is one of the largest superorders within
teleosts with approximately 6500 species, and apart from the order Gonorynchiformes, almost all other ostariophysans are exclusively freshwater inhabitants (Saitoh, et al. 2003).

**Objective 1: Establishing a phylogenetic framework to investigate correlations between habitat depth of euteleost fishes and amino acid substitutions in their visual pigment rhodopsin**

Photic properties of water differ drastically with increase in depth in marine environments. Down-welling sunlight is greatly attenuated in general, and the degree of attenuation varies for different wavelengths. Wavelengths of maximum sensitivity ($\lambda_{max}$) of visual pigments in fishes inhabiting greater depths are commonly blue-shifted to better match the ambient light and amino acid substitutions at sites near visual pigments chromophore binding pocket are known to tune $\lambda_{max}$. This project aims to: 1) Develop a methodology to establish a phylogenetic framework for testing correlations between amino acid replacements and macroecological patterns, 2) Conduct comparative analyses of a dataset of rhodopsin sequences and habitat depth in euteleost fishes, 3) Evaluate robustness of the results using simulations, different phylogenies and phylogenetic models, and through comparison with other nuclear and mitochondrial genes used as controls, and 4) Use information from homology modeling on resolved crystal structures of dark-state and light-activated bovine rhodopsin and published literature to propose how substitutions found to correlate with shifts in depth can alter rhodopsin function. Since spectral properties of water can differ drastically between marine and freshwater habitats, such that the clearest waters are found in the open oceans while higher concentrations of dissolved organic compounds in rivers, lakes and coastal waters make them appear more yellow-
orange in colour (Warrant and Locket 2004; Jerlov 1976), we only included marine species within the euteleost lineage for this project.

**Objective 2: Investigating correlations between habitat temperature of teleost fishes and amino acid substitutions in their visual pigment rhodopsin**

Thermal energy can activate visual pigments and this rate of thermal activation can affect visual sensitivity and the degree of dark noise. In poikilotherms, variation in ambient environmental temperature influences their body temperature; consequently, fishes living in colder waters are less likely to experience thermal activation of their rhodopsins compared to fishes living in warmer waters. The main aims of this project include: 1) Use the established framework for objective one to examine if significant differences exist between habitat temperature of euteleosts with different residues in rhodopsin, 2) Investigate correlation with temperature in a separate group of teleosts, freshwater ostariophysan fishes, and 3) Assess if there are any overlapping RH1 sites between euteleosts and ostariophysans which correlate with temperature, as well as sites that are significant in both the depth and temperature analyses.
Figure 1. Effect of phylogenetic relationships on correlations of tip data, modified from Garland, et al. 2005. (a) Significance cut-off of simulations on a contemporaneous, star phylogeny is the same as conventional, non-phylogenetic statistics. (b) Significance cut-off becomes more stringent than conventional statistics for a time-calibrated phylogeny which takes into consideration the hierarchical relationships of species.
Figure 2. Crystal structures of bovine rhodopsin in the dark (left) and light activated (right) states. Retinal chromophore is shown in red; extracellular (Ext.) and cytoplasmic (Cyt.) domains as well transmembrane helices 1, 5, 6 and 7 (H1, H5, H6, H7) are labeled for clarity.
Figure 3. Time-calibrated phylogeny of bony fishes, adapted from Betancur, et al. 2013 with permission. Red circles show the clades used in this study.
Figure 4. Attenuation of sunlight with increasing depth in marine habitats. Blue wavelengths are generally less attenuated and can reach depths down to nearly 1000 meters.

Figure 5. Vertebrate phototransduction cascade, adapted from Lamb 2013. Proteins in rod cells are in blue and those found in cones in red.
Chapter 2: Correlated Evolution of Euteleost Fish Rhodopsin with Habitat Depth: Linking Molecular Changes to Ecology

Abstract

Statistical methods have been developed to investigate correlated evolution of phenotypic traits or to detect signatures of selection at the molecular level under a phylogenetic framework. However, there have been minimal attempts to discern changes in protein coding sequences that correlate with ecology, physiology, or morphology. Here we investigated amino acid replacements in euteleost fish rhodopsins (RH1) that show significant association with shifts in habitat depth using existing phylogenetic comparative methods, as well as novel implementation of sequence simulation and ancestral reconstructions. Through an unbiased examination of amino acid sites in RH1, we found several substitutions known to affect spectral, thermal, and kinetic properties of rhodopsin which occurred in parallel with changes in depth. Robustness of the findings was evaluated using various models of evolution and by testing different branch lengths. Among all the sites tested in fourteen nuclear and mitochondrial genes, A292S replacement in RH1, a blue-shifting substitution reported previously to have occurred repeatedly in several deep sea fish, showed the most significant correlation with depth. Two other known spectral tuning sites that additionally alter the rate of active intermediate metarhodopsin II formation or decay, sites 83 and 122, were also correlated with habitat depth. Interestingly, substitutions at sites 97, 185, and 271 which appear to affect rhodopsin thermal stability
correlated with depth as well. Several other non-conservative substitutions near the chromophore binding pocket, putative retinal entry and exit channels, transducin binding sites, and dimerization interface of rhodopsin were also identified that require further investigation. Other genes and macroscopic traits can be studied similarly using a combination of codon-based models and phylogenetic comparative methods discussed here. Developing likelihood or Bayesian models that can incorporate molecular sequences and continuous traits under the same framework eliminates some limitations of the current study. Estimating correlations between phenotype or ecology with genetic changes would illuminate aspects of selection previously overlooked.

**Introduction**

Cross-species comparison of physiological, morphological and ecological traits as a means of elucidating evolutionary adaptations has intrigued biologists for centuries (Garland, et al. 2005), and includes studies ranging from correlations between brain and body size (Lande 1979), biological attributes affecting extinction risk (Purvis, et al. 2000), to dispersal capacity and species richness (Riginos, et al. 2014). However, the assumption of independent observations in conventional statistics is violated when comparing species’ traits because of shared ancestry and variable degrees of taxonomic relatedness. Phylogenetic comparative methods (PCMs) address this issue by incorporating phylogeny to reconstruct ancestral values (Felsenstein 1985) or through variance-covariance matrices (Grafen 1989; Martins and Hansen 1997), thus discerning true correlations driven by parallel evolution of characters from those caused by phylogenetic inertia. These methods generally incorporate either a purely neutral model of evolution, Brownian Motion (BM), or an Ornstein-Uhlenbeck (OU) process with one
or more optima (Butler and King 2004). Robustness of findings can be evaluated by the consistency of results obtained under different models or topologies, or by applying various branch length transformations (Garland, et al. 1993; Grafen 1989; Pagel 1999).

The increased availability of molecular sequences over the past few decades has risen the number of phylogenies for use in comparative studies (Weber and Agrawal 2012); it has also contributed to the growing interest in detecting selection pressures at the molecular level through codon-based likelihood/Bayesian models in order to analyze rates of evolution across codon sites and/or genes (Pond and Frost 2005; Yang, et al. 2000). Functional genes are expected to be under strong purifying selection to avoid mutations which can compromise protein function, with the rate of non-synonymous substitutions ($d_N$) being significantly lower than synonymous substitutions ($d_S$). Similarly, conserved sites are likely under tight selective constraints, while a $d_N/d_S$ ratio elevated above one is taken as evidence of positive selection (Yang 2007). These models have proven useful for the detection of differences in selective constraint in a variety of systems and are being continuously improved (Khan, et al. 2011; Metzger and Thomas 2010; Moury and Simon 2011; Swanson, et al. 2001). Given a phylogeny and substitution rate matrix, codon-based models can also be used to reconstruct ancestral sequences and infer changes that have occurred over an evolutionary time scale (Chang, et al. 2002; Kratzer, et al. 2014; Thornton 2004).

While both macroscopic comparative studies and molecular analyses have generated a wealth of knowledge regarding mechanisms of selection and adaptation at different scales, the two fields have remained largely disparate. The difficulties in integrating genomic and phenotypic information stem from differences in both scale and methodology, as well as the lack of appropriate systems in which to address the problem. Recently, a phylogenetic model was
developed for investigating correlated evolution of continuous traits and substitution rates (Lartillot and Poujol 2011); however, molecular rates need not necessarily change in parallel with changes in ecology or macroscopic traits of organisms. One or a few amino acid (AA) replacements in critical sites of a protein can alter its properties; consequently, convergent substitutions are expected to occur in species exposed to similar selective pressures. Here, we seek to bridge this gap by investigating the evolutionary relationship between habitat depth of euteleost fishes, and the dim-light visual pigment rhodopsin (RH1). In the marine habitats of most euteleosts studied here, daylight is drastically attenuated with increased depth, such that only a narrow, blue-shifted band of the visible spectrum is available below 200 meters (Douglas, et al. 1998) and none reaches depths below 1000 meters (Kenaley, et al. 2014).

Absorption of a single photon by the visual pigment rhodopsin can initiate the phototransduction cascade, leading to a series of downstream signaling events in the retina, ultimately causing the perception of light (Lamb and Pugh 2006). Visual pigments are a class of seven transmembrane domain G protein-coupled receptors that undergo conformational change when their covalently-linked retinal chromophore is isomerized upon photon absorption (Okada, et al. 2001). Studies have shown amino acid substitutions near the chromophore binding pocket can tune the wavelength of maximum sensitivity ($\lambda_{\text{max}}$) of visual pigments closer to the red or blue end of the visual spectrum, depending on the type and position of the substitution (Bowmaker 2008). In an evolutionary context, spectral tuning of visual pigments to characteristics of the ambient light environment is thought to be a common sensory adaptation, particularly in aquatic systems (Hunt, et al. 2001; Larmuseau, et al. 2009; Sugawara, et al. 2005; Weadick, et al. 2012). However, other aspects of visual pigment function, such as those
mediating photosensitivity, have not been similarly investigated, although there is some evidence that these aspects may be at least as important (Sugawara, et al. 2010).

The diverse ecology of euteleosts provides a suitable framework to conduct a comparative analysis on how amino acid substitutions in rhodopsin might have occurred in correlation with changes in habitat (Fig. 1). Wavelengths of light in the blue region of the visible spectrum become proportionally more abundant with increased depth in marine habitats because of the greater attenuation of other wavelengths, and at depths greater than 1000 m, the bathypelagic zone, the predominant source of light becomes the blue bioluminescence produced by some deep-sea species. Blue-shifting substitutions in euteleost RH1 are therefore expected to have occurred in parallel with inhabiting deeper oceanic zones. Additionally, amino acid replacements that increase the rate of active metarhodopsin II (MII) intermediate formation, or its stability, may increase photosensitivity, which can be highly beneficial in the dim environment of the deep-sea. Visual pigment dark-state thermal stability also influences photosensitivity, since thermal isomerization of the retinal chromophore, called “dark noise”, generates signals indistinguishable from photoisomerization; hence, substitutions that increase thermal stability may have also occurred in the deep-sea euteleosts RH1. However, because of the very low water temperature in the deeper zones, poikilotherms rhodopsin would likely experience less dark noise events and so may have fixed substitutions which decrease RH1 thermal stability. Assuming that colder temperatures decrease kinetics of cellular processes, substitutions which accelerate MII decay and retinal release may also be beneficial for faster regeneration of rhodopsin.

By taking advantage of the existing phylogenetic comparative methods and codon-based models, we tried to identify RH1 sites with significantly different residues between shallow and
deep sea fishes. Sequence simulations were used in order to evaluate and adjust the significance cut-off for our novel implementation of PCMs and ancestral reconstructions. We then checked the consistency of findings across different methods and branch length specifications, and compared results for RH1 with thirteen other nuclear and mitochondrial genes used as controls to verify if the discovered associations were unique to the vision gene rhodopsin.

Methods

Sequence Collection, Alignment and Phylogeny Construction

Rhodopsin sequences of 194 euteleosts, covering the majority of the orders sampled in the most recent phylogenetic studies of euteleosts (Betancur, et al. 2013; Near, et al. 2012), as well as sequences for Gonorynchus greyi and Alepocephalus bairdii as outgroups, were downloaded from NCBI along with sequences for seven other nuclear (ENC1, MLL, MYH6, RAG1, RAG2, SH3PX3, ZIC1) and six mitochondrial genes (COI, CYTB, ND1, ND2, ND4, ND5), as available. Number of available sequences ranged from 53 for ND5 to 173 for COI. Sequences were aligned by MUSCLE (Edgar 2004) as implemented in Mega5.0 (Tamura, et al. 2011). We used the time-calibrated phylogeny generated by Betancur et al. 2013 and pruned it to include the 196 taxa for which RH1 sequences were obtained. If we had different species of the same genus in our data set from the chronogram generated by Betancur et al., we assumed their relationship and divergence times to be the same.
Compiling Ecological Data and Evolutionary Selection Regime

Light spectral content and other aspects relevant to opsin function, such as temperature, are different at various depths; thus euteleosts depth ranges were collected from online databases (Fishbase, Aquamaps, Encyclopedia of Life). Midpoint values of Depth Ranges, or preferred depth ranges if available, (MDR), were used for all statistical analyses. Using depth range, maximum, or minimum boundaries instead of MDR produced similar statistics as midpoint depth and were not investigated further since they were all highly correlated with MDR. We compared several evolutionary models and branch length transformations implemented with the fitcontinuous function in Geiger (Harmon, et al. 2008) to determine the model that best explains the extant depth ranges (Table S3). Ancestral depths were then reconstructed under Brownian Motion, Ornstein–Uhlenbeck with a single optimum, and the best estimated model in Geiger determined via AICc score, Pagel’s λ model in this case (see below). In short, OU has an additional selection-strength parameter, α, compared to BM and fits a random walk with tendency towards an optimum (the multiple optima model was not used here). Pagel’s λ, δ, and κ parameters increase the fit of BM over the tree by transforming branch lengths. λ can be interpreted as the level of phylogenetic association and multiplies internal branches with the estimated value (λ=0 produces a star phylogeny and λ=1 keeps branches unchanged) while κ raises branch lengths to an estimated power (κ=0 is the punctuational mode of evolution where trait divergence is proportional to number of speciation events). δ estimates changes in the rate of trait evolution from the root to the tips and adjusts branch lengths accordingly (Harmon, et al. 2008; Pagel 1999).
**Statistical Analyses and Correcting for Phylogenetic Signal**

Our sequence alignment included 272 sites of rhodopsin (sites 32-303 using bovine numbering), which corresponds to the region of the gene sequenced for at least 50% of the species. We tested 85 sites, for which a second residue identity could be found at a site in more than five sequences, and investigated if significant differences exist in amino acid residues at individual sites between deep- and shallow-water dwelling fish. Beginning with a non-phylogenetic approach, a Kruskal–Wallis (KW) test (non-parametric equivalent of ANOVA) (Kruskal and Wallis 1952) was used to assess differences in rank of MDRs among residues at each RH1 site since midpoint depths were not normally distributed as determined by Shapiro-Wilk test of normality (p-value<0.001).

Multiple approaches were then taken to correct for phylogenetic inertia, in order to determine if similar sites were detected with different methods. First, we adapted existing PCMs to study correlations between amino acid substitutions and species habitat depths. One of the most widely known and used PCMs is independent contrast (Felsenstein 1985) and its later extension, phylogenetic generalized least squares (PGLS) (Grafen 1989; Martins and Hansen 1997). We used midpoint depths as the dependent variable and the two most prevalent residues at individual RH1 sites as dummy explanatory variables. Correlations were tested using the package nlme in R 3.1.1. We also used Pagel’s method for correlated evolution of discrete traits (Pagel 1994). This approach compares rates of evolution of two binary characters under two models of independent and dependent evolution and a likelihood ratio test determines whether the more complex model, with eight transition rate parameters, fits the data significantly better than the less complex, independent model with four parameters. To fit the binary requirements of the model, we binned species into euphotic (minimal occlusion of sunlight, <200 meters), or...
aphotic (very narrow bandwidth to no sunlight), based on the midpoint value of their depth range, and tested for correlation between these depth zones and amino acid residues at RH1 sites. We only tested the two most abundant amino acids at sites where more than two residues were present. All analyses were conducted using the package CorHMM in R 3.1.1.

As another approach to correct for phylogenetic signal, we reconstructed ancestral RH1 sequences in PAML4.7 under the M8 model, which was the best codon model as determined by the AIC. Ancestral depths were reconstructed in Geiger under BM, OU, and after applying branch length transformation from the best model discussed above (λ, in this case), in order to evaluate consistency of results. To test for correlations, amino acid replacements at individual RH1 sites along branches of the phylogeny were traced with shifts in depth from ancestral to daughter nodes and statistical significance was evaluated using KW test. In this Ancestral Reconstruction approach, AR, the difference in inferred depths at the two nodes of a branch was taken as response and substitutions at a site as factor levels for KW.

Finally, we used Evolver in PAML 4.7 (Yang 2007) to simulate sequences based on M8 estimates of ω (ratio of non-synonymous to synonymous substitutions), κ (transition to transversion ratio), branch lengths and codon frequencies for rhodopsin. 10000 sites were simulated and KW tests were conducted on the sites for which a second residue identity could be found in more than 5 sequences in order to construct a simulated distribution of p–values for comparison with KW probabilities from RH1 sites. Since sequences were simulated on the phylogeny, more closely related species would generally have higher sequence similarity; therefore, comparison of probabilities obtained from the Kruskal-Wallis test on simulated sites to non-phylogenetically corrected p-values can determine the adjusted level of significance beyond what is driven by phylogenetic inertia. For instance, at a p = 0.05 significance level, sites with
uncorrected p-values smaller than 5% of simulations would be considered significant.
Additionally, we used these simulated sites to assess the type I error rate (false positives) of the phylogenetic approaches mentioned above and to determine their adjusted significance levels.

**Correlations in Other Genes**

Similar procedures as above were conducted for the other genes used in this study to test if there are sites in those genes with significantly different residues in shallow- and deep-water fish, with the expectation that genes with functions unrelated to changes in water depth should not show any trends. We compared results for these genes to those obtained from RH1 in terms of number of significant sites and strength of correlations. Since less sequences were available for control genes, we also pruned our RH1 data set to only include sequences found for control genes, in order to minimize the effect of sample size differences in our interpretations.

**Homology Modeling of RH1**

RH1 substitutions which correlated with habitat depth were modeled on bovine rhodopsin (PDB code:1U19, (Okada, 2004)) using Modeller (Sali and Blundell 1993) in order to examine how they might affect RH1 structure. One hundred models were generated by optimizing the Modeller objective function and the model with the lowest DOPE score (Shen and Sali 2006) selected for further assessment and visualization. Model quality was checked using ProSA-web (Wiederstein and Sippl 2007) to ensure the model and template structures’ energies are comparable (as determined via z-scores), and by ProCheck (Laskowski, et al. 1993), to ensure bond lengths and angles do not have unusual stereochemical conformations.
Results

We first used a non-phylogenetic approach, the Kruskal-Wallis test, to examine each RH1 site individually and test for correlations between species’ habitat depths and particular amino acid residues. Using this approach, 55 sites (out of the 85 variable sites tested) were significant at $\alpha = 0.05$ and 22 remained significant after Bonferroni correction. Interestingly, site 292 showed the most significant correlation between depth and amino acid identity among all the sites tested in rhodopsin and thirteen other nuclear and mitochondrial genes (Table 1, Fig. 2). This is a spectral tuning site previously known to have experienced repeated Ala to Ser substitutions in deep-dwelling fishes (Hunt, et al. 1997; Sugawara, et al. 2005).

Differences in species relatedness might underlie these correlations, however, so they were then tested under a phylogenetic framework. We simulated sequences on the phylogeny based on M8 estimates of RH1 $\omega$, $\kappa$, and codon frequencies in order to generate a new null distribution for comparison with the statistics from sites in RH1. Sites 292, 213, and 168 had smaller p-values than 5% of the simulations using KW test (Fig. 2). Sequence simulation was also used to estimate the type I error rate of the phylogenetic approaches and to adjust their significance cut-offs by comparing the distribution of p-values from the simulated sites to the distribution of RH1 sites for each method. As expected, the ancestral reconstructions approach was very conservative and only 0.6% of the simulations were significant at nominal $\alpha = 0.05$. Phylogenetic generalized least squares and Discrete had higher false positives with over 8% and 22%, respectively (Table S4). Significance cut-off for each method was therefore determined through simulations such that only the RH1 sites with a p-value smaller than 5% of simulations would be considered significant.
We then used PGLS and Discrete, two of the most widely implemented phylogenetic comparative methods, for correlations between habitat depth and substitutions in RH1 sites. Using PGLS, site 292 still showed the most significant correlation. The median habitat depth of species with Ser at this site is over 500 meters deeper than those with Ala (Fig. 3), which is expected given A292S substitution is known to cause ~10 nm blue-shift. Two other known blue-shifting substitutions, D83N and E122Q, were also identified and showed significant correlation with inhabiting deeper zones (Table 1, Figs. 2, 4). These sites are known to additionally affect rates of metarhodopsin II formation and/or decay (Imai, et al. 2007; Sugawara, et al. 2010).

Other substitutions near the chromophore binding pocket which were found to be correlated with habitat depth using PGLS include T97S, G114A, A124S, C185V, and V271T (Figs. 5, 6). Site 97 is known from experimental studies to spectrally tune other opsins when a polar residue is substituted with a non-polar amino acid (Takenaka and Yokoyama 2007; Yokoyama, et al. 2007). Euteleost RH1 sequences in this study all had only polar residues at this site, Ser or Thr, but median habitat depth of species with Ser was 260 m compared to 30 m for Thr (Fig. 3). T97S mutation in bovine rhodopsin did not change $\lambda_{\text{max}}$, although $A_{500}/A_{280}$ height was substantially lower in the mutant (Yokoyama and Takenaka 2004). G114A substitution was only found in four sequences but since they occurred along four different lineages, this site was included in the analysis and showed significant correlation with shifts to deeper habitats. Substitutions at site 124 are known to tune $\lambda_{\text{max}}$ and affect retinal release rate, while site 185 has been shown to alter rhodopsin thermal stability. V271 is within 10 Å of the chromophore (Fig. 5, 6) and substitution to Thr may change the electrochemical properties of the binding pocket. Site 271 is also adjacent to two residues (A272 and F273) that form retinal channel B and have been shown to slow chromophore uptake when mutated in bovine RH1 (Piechnick, et al. 2012).
Since Discrete is limited to binary characters, species were binned in two groups: species living in the epipelagic zone, which is the illuminated zone of the sea (≤200 m), and species inhabiting greater depths where ambient light is highly reduced. For each amino acid site in the protein, two models were implemented: a dependent model where the rate of evolution between two states of the first character depends on the state of the second character, and an independent model where the rates are independent. A Likelihood Ratio Test (LRT) determines whether the dependent model fits the data significantly better than the independent model, which would imply that the characters had evolved in a correlated fashion. Using this approach, site 292 showed the most significant correlation with depth. T251S was the second most significant substitution identified by Discrete. Thr251 forms hydrogen-bonded networks with other polar and charged residues in the second and third cytoplasmic loops of rhodopsin and mutations at this site are shown to alter the level of transducin activation. Substitutions at sites 112 (adjacent to the E113 counterion), 166 (dimerization interface), and 189 (near the retinal binding pocket) were other interesting sites that correlated with habitat depth using Discrete (Fig. 2).

While PGLS and Discrete are widely used by comparative biologists, they may not be ideal for estimating evolution of molecular sequences. In order to address this limitation, we used codon-based models to reconstruct ancestral sequences and combined them with habitat depth reconstructions. Ancestral RH1 sequences were reconstructed under M8 in PAML, which was the best-fitting codon model tested here, and ancestral depths were estimated separately using Geiger. The most significant association was once again found for site 292 where A292S and S292A paralleled shifts to deeper and shallower habitats respectively (Fig. 2, 4). Similarly, site 83, another spectral tuning site also identified by PGLS, passed the significance cut-off for AR. Another non-conservative substitution, A168S, in close proximity to the retinal chromophore and
E122-H211 hydrogen bond network, was significantly associated with an increase in habitat depth (Fig. 5, 6); this change in polarity may have an impact on the peak wavelength of sensitivity and other aspects of rhodopsin function. Sites 46, 97, 166, 259, and 266 were among the significant sites using the AR method, which were also identified by PGLS or Discrete (Fig. 2). These results were similar whether using Brownian Motion, Ornstein–Uhlenbeck, which was a better fit than BM, or Pagel’s λ model, which was the model with the lowest AICc, for reconstruction of habitat depths.

Finally, we wanted to evaluate how sensitive our findings are to changes in branch lengths (BLs), since PCMs can sometimes yield different results when very inaccurate BLs are used. Consistency of results under different BLs (and tree topologies) would provide further the confidence in conclusions. The effect of equal branch lengths (unity) was compared with branch lengths calculated from divergence times and from nucleotide substitutions per codon estimated from M8 in PAML. Simulations showed type I error was similar for the AR approach regardless of the BL specification, it was highest when divergence times were used for Discrete, and highest when BLs were arbitrarily set to unity for PGLS. Nevertheless, most sites in rhodopsin that correlated with habitat depth in the time-calibrated tree were still recovered and passed the simulation adjusted significance cut-off when using alternative models for estimating branch lengths (Tables S1, S5). Site 292, remained the most significant site in correlation with depth irrespective of BLs.

With the aim of assessing whether the associations found with habitat depth are due to changes which impose a selective pressure on vision and on the evolution of rhodopsin, we performed similar analyses using thirteen other nuclear and mitochondrial genes (Table 2). Unlike RH1, these genes are not generally expected to show correlated evolution of amino acid
substitutions with depth. However, adaptation to deep water environments (such as increased pressure, decreased temperature, and decreased oxygen content) might influence the evolution of the other genes as well, and any interpretation of the results should be made with caution. For instance, mitochondrial genes involved in the electron transport chain might be under selective pressures at depths where available oxygen is minimal. Nevertheless, when compared to rhodopsin, the number and proportion of sites with significant correlations with habitat depth, after phylogenetic correction, were substantially less for control genes and reflected the expected false positive rates, on average. Without phylogenetic correction, 33.1% of the sites were significant overall in the control genes using KW test, while the proportions dropped to 3.7%, 0.3%, and 5.0% after simulation adjusted cut-offs for PGLS, Discrete, and AR, respectively (Table 2). For rhodopsin, these proportions were 17.6%, 8.2%, and 18.8% of the sites, respectively (Table 2).

Given that less sequences were available for the additional genes than for RH1, we pruned our RH1 data set to include only species for which we could find a corresponding sequence in the other nuclear or mitochondrial genes. This was done to ensure larger RH1 sample size was not biasing our interpretations, and to check if correlations still hold in smaller data sets. The results were generally congruent with the larger data set originally tested, and a substantially higher number of sites in rhodopsin correlated with habitat depth compared to the control genes with the same number of sequences from the same species (Table S2). Additionally, the strength of correlations was higher for rhodopsin than the control genes.
**Discussion**

Here we identified sites in rhodopsin that show significant correlations with shifts in habitat depth after correcting for phylogenetic inertia. We used a combination of existing phylogenetic comparative methods, as well as new methods that make use of both ancestral reconstructions and sequence simulations, to investigate these associations. The consistency of results across different models of evolution and branch length assignments was compared, and the methods were found to be relatively robust to these differences. Site 292, a previously known spectral tuning site, was found to show the most significant correlation using an unbiased statistical approach. Substitutions at sites 83 and 122 also correlated with depth; experimental studies have shown effects on spectral and kinetic properties of rhodopsin when these sites are mutated. Amino acid replacements at three generally conserved sites across vertebrate rhodopsins, T97S, C185V, and T251S, correlated with shifts to deeper habitats. Mutations at these sites have been experimentally characterized, and seem to affect at least one aspect of rhodopsin function including thermal stability and level of G protein activation. Another highly conserved site, G114, showed multiple substitutions to Ala in accord with habitation of deeper zones, and this mutation may alter the orientation of the adjacent counterion’s (E113) side chain. Non-conservative substitutions at several sites in proximity to the chromophore, particularly A168S and V271T, were identified as well, and warrant further investigation. Other RH1 sites identified in this study to correlate with depth may also alter rates of retinal entry and exit, metarhodopsin formation and stability, transducin activation, thermal stability, dimerization ability, and spectral properties of rhodopsin. Genes used in this study as controls had substantially weaker associations, and less sites correlating with habitat depth than rhodopsin, near the expected level of false positives on average. These results further validated the
associations found for RH1 and established a framework to investigate correlated changes in other ecological or phenotypic characters and molecular sequences.

While phylogenetic comparative methods have been used to study correlations between a wide range of morphological and ecological traits (Barker and Pagel 2005; Friedman, et al. 2009; Varela, et al. 2007), the current study is one of the first to examine correlations between amino acid substitutions at individual sites and shifts in habitat under a statistical framework. The abundance of possible confounding factors is a likely reason for the lack of such studies. For instance, the same mutation may have different effects on rhodopsin depending on the surrounding residues with which it interacts. The photic properties of seawaters can also differ at similar depths depending on the amount of dissolved minerals (Jerlov 1976). Additionally, very different adaptive mechanisms can be utilized in response to similar ecological changes making statistical evaluations particularly challenging. Nevertheless, we were able to find functionally important sites in rhodopsin that correlated with shifts in habitat depth.

Amino acid substitutions at specific sites near the chromophore binding pocket change the wavelength of maximum absorbance of rhodopsin (and other opsins) and can serve as a mechanism for optimizing species’ visual sensitivity to ambient light. Other AA changes are also known to affect the retinal release rate of visual pigments (Bickelmann, et al. 2012; Piechnick, et al. 2012), or their thermal stability, since thermal energy can also initiate signaling [dark noise; (Aho, et al. 1988; Barlow 1957; Janz, et al. 2003)]. These AA substitutions are more likely to be observed when comparing species inhabiting diverse environments. Since only a very narrow spectrum of light near the blue region can reach depths more than 200 m in oceans (Douglas, et al. 1998), and sunlight is almost entirely occluded by 1000 m leaving short-wave bioluminescence produced by some deep-sea organisms the predominant source of light
(Johnsen, et al. 2012; Kenaley, et al. 2014), rhodopsins of deep water-dwelling fish are expected to have adapted by blue-shifting their $\lambda_{\text{max}}$. Other proposed dim-light adaptations in visual pigments include faster formation or slower decay of Metarhodopsin II (Sugawara, et al. 2010), and higher thermal stability (Liu, et al. 2011). An increased rate of rhodopsin activation may be an adaptive mechanism to ensure a physiological response is elicited in the presence of scarce photons of light. Alternatively, faster all-trans retinal release can lead to faster regeneration of rhodopsin, which might be adaptive under certain light conditions, especially if kinetic processes involved in pigment regeneration are slowed in fishes of colder, deeper waters. Also, a more thermally stable visual pigment prevents false signals in the absence of light, which is of great importance in dim-light habitats to ensure high photosensitivity. On the other hand, due to decreased temperature with increased depth, the chances of thermal activation of RH1 would be substantially lower; hence, amino acid replacements which lower rhodopsin thermal stability may be more tolerated in deep sea fishes compared to shallow dwelling organisms due to a relaxation of selective constraint. Similarly, bathypelagic fishes may actually benefit from less photosensitive eyes to avoid detection of distant flashes which might be pointless to pursue (Warrant and Locket 2004).

Some of the sites identified in this study have been extensively investigated through site directed mutagenesis or computational modeling. The rhodopsin substitution found to most significantly correlate with increased depth, A292S, has been previously found to occur in RH1 of several deep sea fishes and mammals (Fasick and Robinson 2000; Hope, et al. 1997); it has been shown to cause a nearly 10 nm blue-shift (Hunt, et al. 2009) and accelerate the decay of the light activated intermediate in vitro (Janz and Farrens 2001). Here we established a statistical association between this substitution and habitat shifts in euteleosts after testing almost all
variable sites in the seven transmembrane region of rhodopsin. Two other known spectral tuning sites, 83 and 122, were identified by both the PGLS and AR approach, and PGLS only, respectively. D83N and E122Q were found to occur in parallel with increased habitat depth and are both blue-shifting substitutions, as expected based on the greater proportion of blue wavelengths of light in deeper regions. Both of these substitutions are also known to affect the equilibrium between rhodopsin dark-state and light activated intermediates. D83N was shown to increase the rate of metarhodopsin II formation, and was proposed to be a mechanism for dim-light adaptation by enhancing rhodopsin photosensitivity in bats and in some African lake cichlids which inhabit deeper water (Sugawara, et al. 2010). Similar adaptation can be expected here for marine euteleosts of mesopelagic and bathypelagic zones. On the other hand, E122Q appears to slow MII formation and decrease rhodopsin photosensitivity, but accelerates MII decay and formation of opsin-retinal bond (Imai, et al. 2007; Kefalov 2012). Although pigment regeneration was not affected by the E122Q substitution in vivo (Kefalov 2012), the conditions may be different in the deep-sea where scarce photons of light are available and the rate of retinal release and covalent linkage formation may set the speed of rhodopsin regeneration rather than the rate of retinal delivery to photoreceptors. Substitutions at site 124 are also shown to affect $\lambda_{\text{max}}$ and retinal release rate. Mutations at site 185, away from Cysteine, were shown to decrease the dark state thermal stability of rhodopsin and although it was not directly measured for T97S mutation, its $A_{500}/A_{280}$ absorbance ratio hints to a change in thermal stability as well. These two sites are highly conserved in most vertebrate rhodopsins. Reduction in RH1 thermal stability of some deep-sea fishes may reflect the decrease in probability of dark noise events because of the lower temperature while substitutions which increase thermal stability would enhance photosensitivity. Some of the sites identified here are close to the putative retinal release
and uptake channels, so they can modulate how fast the chromophore is ejected from, or taken up by, the protein moiety. Sites 144, 137, and 251 in the second and third cytoplasmic loops are in proximity to transducin binding sites. In fact, mutations at the latter two sites are shown experimentally to affect the level of transducin activation (Acharya, et al. 1997; Andrés, et al. 2003; Ridge, et al. 1995; Shi, et al. 1995). A number of sites are in the dimerization interface and may have an effect on how rhodopsin forms dimers in the lipid bilayer (Fotiadis, et al. 2006; Jastrzebska, et al. 2006). Two sites adjacent to the counterion E113, also correlated with habitat depth, sites 112 and 114. They may interfere with the counterion interactions by altering the orientation of its side chain, especially given that G114 is conserved in most vertebrate rhodopsins. Among the remaining sites identified, A168S and V271T are particularly interesting since they are relatively close to the chromophore and substitution to a polar residue is observed in both in parallel with shifts to deeper habitats. Homology modeling of these substitutions did cause a drastic change in the pigment structure, but even small changes in the orientation of side chains or their polarity can affect electrochemical properties of visual pigments.

We applied several methods to test for correlations between amino acid replacements and changes in habitat depth. Among the existing PCMs, the phylogenetic generalized least squares method has been widely applied to study correlations between various continuous traits, including when the explanatory variable is discrete. However, a major limitation of this method, and other PCMs in general including Discrete, is that the evolution of amino acid residues is generally inadequately described by an evolutionary model designed for phenotypic and ecological continuous or discrete traits. Using amino acid indices, such as polarity and volume, could be more appropriate for these methods especially if one expects a corresponding effect on function based on the residue position in the 3D structure. However, even in this case,
substitutions to amino acids with similar properties may not necessarily alter protein function in the same direction across different sites in the protein. Codon based models can be used to infer how genes have evolved and to reconstruct ancestral sequences. We combined these models with those designed for evolution of continuous characters and investigated parallel changes. This approach was very conservative, however, since we only used information along branches where a substitution had occurred. Sequence simulations were utilized to assess the type I error of each approach, which allowed the significance cut-off to be adjusted accordingly.

An interesting future avenue of research lies in the incorporation of ecological parameters into molecular evolutionary models in a likelihood/Bayesian framework. Detecting molecular selection at the level of genes through codon-based models has been widely used in the last two decades (Goldman and Yang 1994; Pond and Frost 2005; Yang, et al. 2000). Functional significance of the positively selected sites inferred from these models can be experimentally tested through site directed mutagenesis and in vitro assays. There is growing interest in incorporating known information about physiology, behaviour, and ecology of organisms to generate more guided hypothesis about the role and significance of sites under positive or diversifying selection. Recent advances in clade models allow for the partitioning of two or more groups to investigate divergences in dN/dS rates (Bielawski and Yang 2004; Weadick and Chang 2012). For instance, divergent selective pressures on cichlid rhodopsins were found to be predominantly driven by differences in lake vs. riverine environments (Schott, et al. 2014). Correlations between life-history traits and substitution rates have also been investigated in a Bayesian framework using a variance-covariance matrix (Lartillot and Poujol 2011). These models are useful in identifying the relationship between macroscopic traits or ecological attributes and substitution rates or selection pressures at the molecular level. However, changes
in amino acid identity may not always be reflected by changes in substitution rates. The approach taken here can be used for any gene or trait under study to address this limitation. One might start with using conventional statistics to search for correlations and then proceed to correct for phylogenetic signal to examine if the correlation still holds. It would be recommended to evaluate consistency of results across a number of topologies, branch lengths, and evolutionary models. There are several factors that might prevent identifying true correlations between amino acid replacements and macroscopic traits, such as different adaptive mechanisms being used in response to similar stimuli or differences in some uncharacterized/unmeasured aspects of ecology in the otherwise similar habitats.

Despite these limitations, we found significant correlations between depth and amino acid identity in RH1 sites known to modulate its function, as well as other potentially interesting substitutions that have not been well characterized yet. Development of models that can incorporate molecular sequences and continuous traits and estimate correlated changes in traits with amino acids substitutions under the same framework can further enhance the current study and illuminate aspects of selection generally overlooked.
Table 1. List of sites in rhodopsin that correlated with habitat depth, using the time-calibrated phylogeny, in at least one method. Only residues found in more than five sequences at each site are shown (except for site 114 where Gly to Ala had occurred along four different lineages) and are listed in descending order of prevalence. Significance was determined by comparing p-values for each site with the distribution obtained from simulations: *: p < 0.05, **: p < 0.01, ***: p < 0.001, PGLS: Phylogenetic Generalized Least Squares, DISC: Discrete, AR: Ancestral Reconstructions, KW: Kruskal-Wallis

<table>
<thead>
<tr>
<th>Site</th>
<th>Residue</th>
<th>PGLS</th>
<th>DISC</th>
<th>AR</th>
<th>KW</th>
<th>Functional Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>A, F, S</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td>Near retinal channel A</td>
</tr>
<tr>
<td>46</td>
<td>L, F</td>
<td></td>
<td>*</td>
<td>**</td>
<td></td>
<td>Near retinal channel A</td>
</tr>
<tr>
<td>83</td>
<td>N, D</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td>Spectral tuning, MII kinetics</td>
</tr>
<tr>
<td>97</td>
<td>T, S</td>
<td>**</td>
<td>*</td>
<td></td>
<td></td>
<td>Thermal stability</td>
</tr>
<tr>
<td>112</td>
<td>L, I, V</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near counterion E113</td>
</tr>
<tr>
<td>114</td>
<td>G, A</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>Near counterion E113</td>
</tr>
<tr>
<td>122</td>
<td>E, Q, M</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Spectral tuning, MII kinetics</td>
</tr>
<tr>
<td>124</td>
<td>A, S, G</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Spectral tuning</td>
</tr>
<tr>
<td>137</td>
<td>V, M, L, I</td>
<td>***</td>
<td></td>
<td></td>
<td></td>
<td>Transducin activation</td>
</tr>
<tr>
<td>144</td>
<td>S, A</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>Transducin activation</td>
</tr>
<tr>
<td>159</td>
<td>F, L, C</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Dimerization interface</td>
</tr>
<tr>
<td>160</td>
<td>T, S, A</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Dimerization interface</td>
</tr>
<tr>
<td>166</td>
<td>A, S, T</td>
<td></td>
<td></td>
<td>**</td>
<td>*</td>
<td>Dimerization interface</td>
</tr>
<tr>
<td>168</td>
<td>A, S, T</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>Near E122-H211 H-bond, Dimerization interface</td>
</tr>
<tr>
<td>185</td>
<td>C, V</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Thermal stability</td>
</tr>
<tr>
<td>189</td>
<td>V, I</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>MII kinetics for I189P</td>
</tr>
<tr>
<td>213</td>
<td>L, C, S, M, T, A, I</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Dimerization interface</td>
</tr>
<tr>
<td>214</td>
<td>I, T, V</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>Dimerization interface</td>
</tr>
<tr>
<td>217</td>
<td>T, V, I, F, A</td>
<td></td>
<td></td>
<td>**</td>
<td></td>
<td>Dimerization interface</td>
</tr>
<tr>
<td>251</td>
<td>T, S</td>
<td></td>
<td></td>
<td>***</td>
<td></td>
<td>Transducin activation</td>
</tr>
<tr>
<td>259</td>
<td>I, V</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>266</td>
<td>L, V, C</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>270</td>
<td>S, G</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near retinal channel B</td>
</tr>
<tr>
<td>271</td>
<td>V, T</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near retinal channel B</td>
</tr>
<tr>
<td>273</td>
<td>W, V</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near retinal channel B</td>
</tr>
<tr>
<td>274</td>
<td>W, Y, F</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near retinal channel B</td>
</tr>
<tr>
<td>277</td>
<td>T, C, S, L</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near retinal channel B</td>
</tr>
<tr>
<td>278</td>
<td>H, N</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>Near retinal channel B</td>
</tr>
<tr>
<td>281</td>
<td>S, A, T</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>282</td>
<td>E, D</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>292</td>
<td>A, S</td>
<td></td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>300</td>
<td>I, V</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Number and proportion (in brackets) of sites with significant \( p < 0.05 \) correlation with habitat depth in rhodopsin and seven control genes. Significance was determined by comparison with simulated distributions. PGLS: Phylogenetic Generalized Least Squares, DISC: Discrete, AR: Ancestral Reconstructions, KW: Kruskal-Wallis

<table>
<thead>
<tr>
<th>Gene</th>
<th>PGLS</th>
<th>DISC</th>
<th>AR</th>
<th>KW</th>
</tr>
</thead>
<tbody>
<tr>
<td>RH1</td>
<td>15 (17.6%)</td>
<td>7 (8.2%)</td>
<td>16 (18.8%)</td>
<td>3 (3.5%)</td>
</tr>
<tr>
<td>COI</td>
<td>1 (1.5%)</td>
<td>1 (1.5%)</td>
<td>7 (10.8%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>CYTB</td>
<td>2 (1.8%)</td>
<td>1 (0.9%)</td>
<td>6 (5.5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>RAG1</td>
<td>8 (6.7%)</td>
<td>0 (0%)</td>
<td>8 (6.7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>MYH6</td>
<td>3 (8.8%)</td>
<td>1 (2.9%)</td>
<td>2 (5.9%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ZIC1</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>MLL</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>4 (6.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ND2</td>
<td>9 (6.0%)</td>
<td>0 (0%)</td>
<td>6 (4.0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ND1</td>
<td>2 (3.4%)</td>
<td>0 (0%)</td>
<td>3 (5.2%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>SH3PX3</td>
<td>3 (10.0%)</td>
<td>0 (0%)</td>
<td>3 (10.0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ENC1</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ND4</td>
<td>4 (3.8%)</td>
<td>0 (0%)</td>
<td>3 (2.9%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>RAG2</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (1.2%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>ND5</td>
<td>5 (3.2%)</td>
<td>0 (0%)</td>
<td>7 (4.5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>control genes avg.</td>
<td>37 (3.7%)</td>
<td>3 (0.3%)</td>
<td>50 (5.0%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
Figure 1. Euteleost phylogeny used in this study and the distribution of their habitat depths. [Alex Van Nynatten made significant contribution to the development of this figure]
(a) **KW**

(b) **PGLS**
Figure 2. Distribution of p-values for the sites in rhodopsin compared to simulated sequences. The broken line shows the 5% significance cut-off, sites passing this cut-off for each method are labeled (using bovine RH1 numbering). KW: Kruskal-Wallis, PGLS: Phylogenetic Generalized Least Squares, DISC: Discrete, AR: Ancestral Reconstructions.
Figure 3. Box-plots showing the difference in median habitat depths of species with different residues at RH1 sites 97, 271, and 292. Species with S97, T271, and S292 tend to live in the deep sea while species with T97, V271, and A292 are generally shallow-dwelling.
Figure 4. Comparison of RH1 residues found in euteleosts inhabiting euphotic and aphotic zones. (a) Sequence logo showing relative prevalence of residues at some sites near the chromophore binding pocket. (b) Change in habitat depth along branches of the phylogeny where a substitution at site 292 had occurred. Bold lines show the average change in depth. [Alex Van Nynatten made significant contribution to the development of these figures]
Figure 5. Crystal structure of the dark-state bovine rhodopsin with 11-cis retinal shown in green. (a) All the sites correlating with habitat depth in at least one of the methods are shown in red. (b) Sites correlating with habitat depth which are near the chromophore binding pocket are shown; residues found in bovine and shallow-dwelling fish RH1 are in red and the residues in deep-sea fish are modeled in blue.
Figure 6. Higher resolution RH1 structure for four of the sites correlating with habitat depth: 97, 168, 271, and 292. (a), (c), (e), and (g) are residues found in bovine and shallow-dwelling fish rhodopsins; (b), (d), (f), and (h) are homology modeled residues found in the deep sea fish.
Chapter 3: Investigating Correlations between Habitat Water Temperature of Teleost Fishes and Amino Acid Substitutions in Rhodopsin

Introduction

Thermal energy is also known to activate visual pigments and can therefore affect photosensitivity. Using the framework established for investigating correlations between amino acid substitutions in rhodopsin and habitat depth of euteleost fishes in the previous chapter, correlations with water temperature can be similarly studied. Increasing depth in marine habitats is associated not only with the sharp attenuation of sunlight, but also with decreasing temperature, increases in pressure and other changes to which deep-sea species need to adapt. Lower temperatures in the deep-sea may have interesting implications for the vision of poikilotherms, since dark noise events would presumably be less frequent in the colder temperatures of the mesopelagic and bathypelagic zones. The rate of thermal activation of visual pigments is likely still high enough to impose a limit on opsin light sensitivity, in which case, those fishes relying on vision to navigate their environment would require thermally stable pigments for high photosensitivity; however, the constraints on thermal stability may not be as high as they are for surface-dwelling fish. Additionally, if colder temperatures slow kinetic processes of visual pigment regeneration, mutations which accelerate those processes - such as by increasing the rate of retinal release from light activated metarhodopsin II- may be beneficial for vision in deep-sea/polar water environments. Given this, investigating putative correlations
between RH1 amino acid substitutions with water temperature may reveal sites of importance for the thermal stability and kinetics of rhodopsin. This analysis may highlight new sites which were not found to be in significant correlation with depth: for instance, if residues at a site differ between temperate and tropical fishes which inhabit similar depths. Conversely, sites that overlap with those found in the depth analysis may indicate their effect on multiple aspects of rhodopsin function, or that the significant correlations with depth were primarily a result of association with decreases in temperature, rather than changes in the photic environment.

In order to conduct temperature analyses two separate data sets were used. First, the identical data set of euteleosts used in the previous chapter was analyzed for correlation with temperature. Second, a data set of ostariophysan teleosts was also compiled and examined for correlated RH1 substitutions with water temperature. Since ostariophysan fishes are predominantly freshwater species, any information regarding their depth ranges is generally sparse while more information is available about their habitats water temperature. The ostariophysan data set provides us with a parallel system to compare evolution of rhodopsin in these species with euteleosts and to investigate if a similar or different set of sites show correlation with temperature in these two groups.

Methods

Euteleost data set

For the euteleost data set, the identical RH1 sequence alignment and time-calibrated phylogeny of 196 species was used as in the previous chapter. Information about ranges of habitat water temperature were collected from online databases (Fishbase, Aquamaps, Encyclopedia of Life) and midpoint values of preferred temperature ranges were used for all
statistical analyses. RH1 sites for which the second most prevalent residue was found in more than five sequences were first subjected to non-phylogenetic testing; midpoint temperatures were not normally distributed as determined by Shapiro-Wilk test of normality (p-value<0.001), so Kruskal–Wallis test was used similar to the previous analysis for depth ranges. In order to correct for phylogenetic signal, PGLS, Discrete, and AR methods were utilized. All phylogenetic analyses were conducted on the phylogeny with branch lengths in units of chronological time. Implementation of PGLS and Discrete were similar to the habitat depth analysis: the two most abundant residues at each site were used as explanatory variable for PGLS and as one of the two binary characters for Discrete; midpoint water temperatures were taken as the response variable for PGLS, and divided to tropical (> 25 °C) or temperate (< 25 °C) groups for Discrete as binned in Chen et al. 2014. Similarly, for ancestral reconstructions method, the fit of Brownian Motion and Ornstein-Uhlenbeck models were compared in Geiger for extant temperatures; OU model had lower AICc and was chosen for reconstructing ancestral temperatures. Reconstructed ancestral rhodopsin sequences from the M8 model (best fit for codon models) in PAML were also used as in the previous chapter. Statistical significance was then assessed using the KW test for correlating amino acid substitutions with changes in temperature along branches of the phylogeny.

**Ostatiophysan data set**

For the ostariophysan data set, nucleotide RH1 sequences of 105 species for which information about habitat water temperature could be obtained from FishBase were downloaded from NCBI, along with rhodopsin sequences for outgroups *Lepisosteus oculatus* and *Osteoglossum bicirrhosum*. Sequences were aligned by MUSCLE (Edgar 2004) as implemented in Mega5.0 (Tamura, et al. 2011) and trimmed to only include sites available in at least 50% of
sequences (sites 32 to 304, using bovine RH1 numbering). Phylogenetic relationships were constructed based on published literature (Fang, et al. 2009; Imoto, et al. 2013; Mayden and Chen 2010; Mayden, et al. 2009; Schoenhuth, et al. 2008; Tang, et al. 2011; Tang, et al. 2010; Wang, et al. 2013; Yang and Mayden 2010; Yang, et al. 2010). The topology was then constrained and nucleotide branch lengths were estimated in RaxML (Stamatakis 2006) using option –g and GTRGAMMA model of nucleotide substitution. Habitat water temperature ranges were obtained from Fishbase, and their midpoint values were considered for statistical analyses. All non-phylogenetic and phylogenetic methods utilized were in a similar manner to the analyses for the euteleost data set, except that instead of a time-calibrated phylogeny, codon branch lengths from M8 model in PAML (Yang 2007) were used.

Results

Correlation between RH1 Amino Acid Substitutions and Habitat Water Temperature in Euteleosts

Using the non-phylogenetically corrected Kruskal-Wallis test, 48 out of 85 sites (56.5%) tested in rhodopsin (sites with the second most prevalent residue present in more than five sequences) were significant at nominal \( \alpha = 0.05 \). Nineteen sites (22.4%) still remain significant after applying the conservative Bonferroni correction for multiple testing. Interestingly, in contrast to correlations with depth which showed site 292 to have the smallest \( p \)-value of all the sites tested by several orders of magnitude, it was not among the top 20 sites in correlation with temperature, and did not pass Bonferroni- adjusted cut-off. Here, site 251 had the smallest \( p \)-value, and was also one of the most significant sites in both the non-phylogenetically corrected depth analysis, and when using Discrete. In order to correct for phylogenetic signal, PGLS,
Discrete and AR were used. Yet, since simulations were not conducted to estimate the 5% significance cut-off (as was done in the depth analyses), nominal p-values were considered for each approach instead. It should be noted that AR would likely be highly conservative whereas Discrete might have higher false positives than it should according to the results from the previous chapter. Table-1 lists some of the interesting sites which show correlated substitutions with water temperature. Sites 97, 137, 144, 160, 217, 251, 270, 271, 274, and 278 are among the sites which were found to also correlate with habitat depth in at least one of the phylogenetic methods. These overlapping sites, especially sites 144, 217, 251, 270, 274, and 278, tend to display a more significant correlation with temperature than with depth. This may indicate the functional importance of these sites on the thermal stability or kinetics of rhodopsin. Moreover, the apparent correlation of these sites with depth may in turn be a result of decreases in relative temperature within the mesopelagic and bathypelagic zones. T97S and V271T show correlated substitutions with shifts to colder waters; given their position on rhodopsin 3D structure and relative proximity to the chromophore binding pocket, these substitutions may alter several aspects of rhodopsin function. Several other sites found in the depth analysis including 46, 112, 166, 168, 259, 281, 292, are only significant using Discrete in the temperature analysis, although adjusted significance levels remain to be determined by simulations. An intriguing observation is that when using PGLS, F261Y, a substitution known to cause a red-shift in rhodopsin and cone opsins, is significant in correlation with colder habitat water temperatures in euteleosts. If red- shifting substitutions make rhodopsin less thermally stable (Gozem, et al. 2012), this association may reflect constraints for highly stable RH1 are relaxed in fishes of colder waters.
Correlation between RH1 Amino Acid Substitutions and Habitat Water Temperature in Ostariophysans

Investigating correlations with water temperature in a separate group of fishes provides a second system to compare evolution of rhodopsin and examine if the same or distinct set of sites are identified in the two data sets. In general, sequence variation was less for the ostariophysan rhodopsins and 68 sites were tested (i.e. sites which the second most prevalent residue was found in more than five sequences). Using non-phylogenetically corrected KW test, 40 sites (58.8%) were significant at nominal $\alpha = 0.05$ and 23 sites (33.8%) after Bonferroni correction. The two most significant sites, L128 and F212, were generally conserved across euteleosts except in a number of ostariophysan lineages which possessed substitutions L128I and F212A in RH1. In contrast, some of the significant sites in the euteleost analyses such as T97, T251, and A292 were entirely conserved in ostariophysans, or variable in only a few sequences in the case of F46, D83, E122, and T160. Some of the RH1 sites found to correlate with water temperature in ostariophysan fishes using phylogenetic methods are listed in Table-1. Sites 137, 144, 166, 259, 261, and 270 overlap with the sites identified to correlate with temperature in euteleosts. Their correlation with temperature in a separate data set further implicates their possible effect on rhodopsin function. Substitutions at sites 128, 212, 256, 258, 259, and 261 appear to have occurred in parallel along a single ancestral branch. Their proximity to each other on the crystal structure suggests that they may function synergistically to alter rhodopsin function, but further experimental characterization would be required to test this hypothesis (Figure 2).
Discussion

While spectral tuning mechanisms in rhodopsin are relatively well-characterized, thermal and kinetic studies are limited in comparison and only recently have begun to gain more widespread attention. It has been known for decades however that thermal energy can activate visual pigments and this “dark noise” sets the limit of photosensitivity. In the previous chapter, we studied correlations between amino acid replacements in rhodopsin and species habitat depths. Increase in depth is associated with attenuation of downwelling sunlight and a shift in the spectral properties of water. However, temperature also decreases sharply in deeper regions and has implications on the evolution of visual pigments. This chapter shows that studying correlations with habitat water temperature can reveal sites in rhodopsin which may have an effect on its thermal and/or kinetic properties, including sites that overlap with those identified in the depth analysis. Using two separate data sets of teleost fishes gives us a comparative framework to examine if similar or distinct set of sites are found to correlate with temperature.

Among the sites identified in the euteleosts, some have been functionally characterized in previous studies using in-vitro or in-silico methods. T251 is in the third cytoplasmic loop of rhodopsin and is known to reduce transducin activation when mutated to Alanine or Glycine (Acharya, et al. 1997; Shi, et al. 1995). It is hypothesized to donate an H-bond to the C-terminal carboxylate of the transducin alpha subunit (Fanelli and Dell'orco 2008) and is a member of H-bond network (ionic lock) between TM3/6 in dark state RH1, and between TM5/6 in active MII state (Choe, et al. 2011). Species with Serine at this site tend to inhabit colder waters; although Threonine to Serine substitution does not change the polarity of the residue at this site, removal of the methyl group may change the distance of side chain hydroxyl from transducin and alter its activation. A functional effect is especially likely given how conserved T251 is in the RH1 of
other species, including the ostariophysans studied here and other vertebrates in general. S144 is located in the second cytoplasmic loop, and is also shown to interact with transducin in a computational model (Nikiforovich, et al. 2007). It is also relatively conserved in teleosts as well as in most vertebrates. S144A correlates with shifts to colder water, so removal of the OH group may interfere with the interaction of rhodopsin with transducin. Site 137 is more variable and has hydrophobic residues Val, Met, Leu, Ile; mutations at this site are also shown to alter the level of transducin activation (Acharya, et al. 1997; Andrés, et al. 2003; Ridge, et al. 1995). Site 97 spectrally tunes $\lambda_{\text{max}}$ in cone opsins when a polar amino acid is substituted to a non-polar residue (Takenaka and Yokoyama 2007; Yokoyama, et al. 2007). The substitution T97S was shown not to cause a spectral shift, but the A500/A280 absorbance ratio was substantially smaller suggesting it may reduce rhodopsin stability (Yokoyama and Takenaka 2004). Site 97 has also been implicated in the formation and stability of folded rhodopsin by computational modeling (Rader, et al. 2004). Together these studies highlight the importance of this site for rhodopsin function. The F261Y substitution shows significant correlation with habitat water temperature using PGLS; this replacement red-shifts RH1 $\lambda_{\text{max}}$ by 8 nm (Yokoyama, et al. 1995) but its effect on the thermal and kinetic properties of rhodopsin remains to be studied. However, presence of Tyr at this site in species inhabiting colder waters may be interpreted as a substitution allowed to occur where a high thermally stable pigment is not as important. If, as proposed by Gozem et al. 2012, red-shifting substitutions decrease thermal stability of rhodopsin, the less stable Y261 pigment would not be as prone to dark noise events in colder habitats. Other sites are not as well characterized, but sites 270, 271, and 274 are near retinal channel B (Hildebrand, et al. 2009) and sites 157, 160, and 217 are near dimerization interface (Fotiadis, et al. 2006; Guo, et al. 2005). Site 160 is also a member of an H-bonding network with N78 and W161, two highly conserved
residues in GPCRs, which form a polar clamp cluster that may provide favorable stabilizing interactions for rhodopsin (Adamian and Liang 2002). Significant correlation of these sites with water temperature warrants further experimental investigation of their effects on rhodopsin function.

Testing correlations in a separate group of teleost fishes allowed us to compare results across the two data sets. First, the distribution of habitat water temperatures in marine euteleosts and freshwater ostariophysans studied here are quite different, as expected. While the OU model was a significantly better fit (lower AICc) than BM in both, strength of selection or how quickly return to optimum is achieved is substantially different between them as indicated by the selection strength parameter \( \alpha \) (Table 2). Although branch length specifications differed for the euteleost and ostariophysan data sets (chronological time versus nucleotide substitutions per codon, respectively) selection strength was still smaller for the former when same units of branch lengths were used (Table 2). This is not surprising given that the habitats of marine euteleosts examined ranged from freezing Antarctic waters to tropical oceans whereas tolerance to thermal variation in habitat water may be less in freshwater ostariophysans. For euteleosts, selection strength is higher for temperature than depth, 1.855 compared to 0.279, respectively, when using codon branch lengths and 0.009 compared to 0.005 when using time-calibrated tree, indicating that variation in habitat depth is not as tightly constrained.

Sites 137, 144, 166, 261, and 270 identified in euteleost temperature correlations were also significant in ostariophysans, further supporting their potential effect on modulating rhodopsin function in a manner related to its thermal or kinetic properties. While some of the sites such as T97, T160, and T251 are largely conserved in ostariophysans, three of the most highly conserved residues in euteleosts and other vertebrates, L128, F212, and V258, are
substituted in ostariophysans to Ile, Ala, and Gly, respectively. In fact, substitutions L128I, F212A, I256L, V258G, I259F, and F261Y appear to have occurred along an ancestral branch in a coordinated fashion in accord with a shift to colder water temperatures. Aside from site 261 mentioned above, limited experimental characterization or computational modeling has been performed on the other sites, except that mutation to Ala at site 258 has shown to increase transducin activation (Han, et al. 1996). Also, under molecular dynamic studies, F212 is shown to undergo a conformational change and fill in the cavity created by the movement of β-ionone ring during transition from MI to MII intermediates (Tikhonova, et al. 2008). Homology modeling of these substitutions did not reveal a significant change in overall structure, but showed how changes in the bulkiness of residues, especially for F212A, may affect the ease of retinal release. Simulations were not conducted to determine the adjusted level of significance, so these sites may not remain statistically significant after adjusting the cut-off; however, clustering of the sites near each other in rhodopsin crystal structure and their proximity to the retinal chromophore, especially side chains of F212 and F261, indicate their potentially important and synergistic effect on modulating rhodopsin function such as on the rate of retinal release. Correlations between amino acid replacements in RH1 and habitat water temperature in teleosts have identified several interesting sites, based on previous studies or their position in the 3D structure, making them suitable candidates for in-vitro assays.
Table 1. Summary of sites in rhodopsin which correlate with habitat water temperature. Residues found in more than five sequences at each site are listed in descending order of prevalence; if prevalence differs between euteleost and ostariophysan data sets, the latter is shown in brackets. *: p-value < 0.1, **: p-value < 0.05, ***: p-value < 0.01, NA: the second most prevalent residue was not found in more than five sequences in that data set.

<table>
<thead>
<tr>
<th>Site</th>
<th>Residue</th>
<th>Euteleost Prevalence</th>
<th>Ostariophysan Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>Residue 1</td>
<td>0.93</td>
<td>0.85</td>
</tr>
<tr>
<td>Site 2</td>
<td>Residue 2</td>
<td>0.67</td>
<td>0.75</td>
</tr>
<tr>
<td>Site 3</td>
<td>Residue 3</td>
<td>0.80</td>
<td>0.70</td>
</tr>
<tr>
<td>Site 4</td>
<td>Residue 4</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>Site 5</td>
<td>Residue 5</td>
<td>0.72</td>
<td>0.80</td>
</tr>
</tbody>
</table>

*: p-value < 0.1, **: p-value < 0.05, ***: p-value < 0.01, NA: the second most prevalent residue was not found in more than five sequences in that data set.
<table>
<thead>
<tr>
<th>Site</th>
<th>Residue</th>
<th>Euteleosts</th>
<th></th>
<th>Ostariophysans</th>
<th></th>
<th>Functional significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PGLS</td>
<td>DISC</td>
<td>AR</td>
<td>PGLS</td>
<td>DISC</td>
</tr>
<tr>
<td>97</td>
<td>T, S</td>
<td>*</td>
<td>**</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>119</td>
<td>L, H</td>
<td>***</td>
<td>*</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>128</td>
<td>L, I</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>137</td>
<td>V, M, L, I</td>
<td>*</td>
<td>**</td>
<td>**</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>(M, L, V, I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>S, A</td>
<td>***</td>
<td>***</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(S, T)(^a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>157</td>
<td>L, V</td>
<td>***</td>
<td>*</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>(V)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>T, S, A</td>
<td>**</td>
<td>*</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>166</td>
<td>A, S, T</td>
<td>***</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(S, A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>212</td>
<td>F, A</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>217</td>
<td>T, V, I, F, A</td>
<td>**</td>
<td>***</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(I, F, V, C, T)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>251</td>
<td>T, S</td>
<td>*</td>
<td>***</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>I, L, M</td>
<td>**</td>
<td></td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(I, L)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>258</td>
<td>V, G</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td>259</td>
<td>I, V</td>
<td>***</td>
<td></td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(I, F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>261</td>
<td>F, Y</td>
<td>**</td>
<td>*</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>270</td>
<td>S, G</td>
<td>*</td>
<td></td>
<td>***</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>271</td>
<td>V, T</td>
<td>**</td>
<td></td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>274</td>
<td>W, Y, F</td>
<td>***</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Y, W)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>278</td>
<td>H, N</td>
<td>***</td>
<td>**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\): Threonine at site 144 is found in five sequences in ostariophysans.
Table 2. AICc of Brownian Motion and Ornstein-Uhlenbeck models for habitat water temperature. The maximum likelihood value of the selection strength parameter, $\alpha$, is shown for the OU model.

<table>
<thead>
<tr>
<th></th>
<th>Euteleosts chronogram</th>
<th>Euteleosts codon BL</th>
<th>Ostariophysan codon BL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>1283.6</td>
<td>1261.8</td>
<td>570.8</td>
</tr>
<tr>
<td>OU</td>
<td>1268.4</td>
<td>1243.8</td>
<td>549.0</td>
</tr>
<tr>
<td></td>
<td>($\alpha = 0.009$)</td>
<td>($\alpha = 1.855$)</td>
<td>($\alpha = 4.143$)</td>
</tr>
</tbody>
</table>
Figure 1. Box-plots of habitat water temperature for some of the RH1 sites with significant correlation. (a) and (b) are sites significant in the euteleost data set but generally conserved in the ostariophysans; (c) and (d) are sites significant in the ostariophysan data set but generally conserved in the euteleosts; (e) and (f) are sites significant in both data sets where (e I) and (f I) are box-plots for the euteleost and (e II) and (f II) are for the ostariophysan data sets.
Figure 2. Crystal structure of dark-state rhodopsin showing sites which correlate with water temperature. (a) Sites with significant correlation in the euteleost and/or the ostariophysan data sets are shown in blue, retinal chromophore is in green. (b) Clustering of six sites in ostariophysan data set which correlate with water temperature. Residues found in most fishes (identical to bovine identity) are shown in blue and the homology modeled substitutions in red.
Chapter 4: Conclusions and Future Directions

Conclusions

Here we have established a framework for investigating correlated evolution of amino acid substitutions in molecular sequences and aspects of ecology or phenotypic traits of organisms. We used rhodopsin as a candidate gene and euteleost fish habitat depth as a model ecological attribute in order to examine how our novel use of existing phylogenetic comparative methods and ancestral reconstruction approaches perform. The well-established role of rhodopsin in the visual pathway, an abundance of experimental and computational studies on this gene, a large number of available sequences, and the resolved crystal structure of RH1 made it very suitable for our purposes. The diversity in habitats of euteleosts and the drastic differences in their photic environment, provided an opportunity to test how this variation might have shaped the evolution of rhodopsin, in these fishes. We identified several sites in rhodopsin in which substitutions are significantly correlated with shifts in habitat depth, even after accounting for phylogenetic inertia. The most significant association was found for a site known to tune $\lambda_{\text{max}}$ such that the blue-shifting substitution, A292S, had occurred repeatedly in parallel with inhabiting greater depths. This site showed the highest level of correlation in all the methods used here and among all the sites tested for rhodopsins and thirteen other nuclear and mitochondrial genes. Other spectral tuning sites and substitutions that are known to affect thermal and kinetic properties of rhodopsin were also identified, including D83N and E122Q. Some interesting amino acid replacements near the chromophore binding pocket, dimerization interface, retinal entry and exit channels, and transducin binding sites were also identified which would be suitable candidates for site-directed mutagenesis and experimental characterization.
We used the codon based models to simulate sequences based on codon frequency and substitution rate parameters estimated for RH1, then performed KW, PGLS, DISC, and AR tests on the simulated sites while keeping species habitat depths unchanged. Since sequences were simulated on the phylogeny, more closely related species would have higher sequence similarity, so we could compare the distribution of p-values for simulations to rhodopsin and determine which (if any) sites in RH1 were more significant than 95% of the simulated sites. We did not simulate habitat depths in order to keep the distribution of extant depths identical for comparison between RH1 and sequence simulations. The type I error (false positive rate) of each approach could also be determined by simulations; we expected the non-phylogenetic KW test to have high false positives because of the habitat depth and sequence similarity between closely related species; PGLS and DISC should have had close to 5% error if they were properly correcting for phylogenetic relatedness and if our novel way of implementing these methods did not violate their assumptions; AR was expected to be highly conservative since only branches along which an amino acid substitution had occurred were used in the analysis for each site. Our findings were largely congruent with our expectations such that type I error was nearly 64% for KW and only 0.6% for AR at nominal $\alpha = 0.05$, PGLS had slightly higher false positives than expected with 8.7%, but DISC error was inflated to over 22%. For the phylogenetic methods, we tested the effect of changing branch lengths from units of divergence time to nucleotide substitutions per codon or setting them all equal to one. The error rate for AR remained nearly unchanged (0.6%-0.7%), it went down for DISC to between 12%-13%, and was elevated for PGLS to 16% when unity branch lengths were used but decreased to under 4% with codon BLs. Altogether, these results suggest that simulations are needed to adjust the significance level for all methods. The significance cut-off needs to be relaxed for AR and be more stringent for DISC; PGLS has
relatively acceptable error rate when informative branch lengths are utilized but use of arbitrary (unity in this case) branch lengths substantially increased its false positives. Also, results of AR and PGLS appear to be in more agreement with each other in terms of the significant sites identified, which could be because of the requirement for binning depth values for DISC.

While we expected to observe correlated amino acid substitutions in RH1 with changes in species habitat depth because of the variation in the photic environment, we did not expect to observe a similar pattern in the control genes used in this study. When we tested thirteen control nuclear and mitochondrial genes, we only found significant correlations close to the expected number of false positives on average. Fewer sequences were available for these control genes compared to rhodopsin, however, so the difference in the size of data sets could underlie the observed differences in results for RH1 and the controls. We examined this possibility by pruning our RH1 data set to only include the sequences for which a corresponding control gene sequence could be found. The results did not change in general: there were both substantially higher number of significant sites and stronger correlations with habitat depth in RH1 than in the control genes. Therefore, a larger sample size for rhodopsin, compared to the control genes, was not inflating the associations we found and rhodopsin is unique among these genes for evolving in accord with species habitat depths. Nevertheless, some of the control genes, such as cytochrome oxidase (COI) and other mitochondrial genes involved in electron transport chain, may in fact be under selection with increasing depth due to the changes in oxygen content and pressure; further investigation can reveal if such associations exist.

We also tested correlations between amino acid substitutions in rhodopsin and water temperature of fishes. We used two data sets for this analysis: first we examined the marine euteleosts used for the depth analysis; we also examined correlations in freshwater ostariophysan
fishes. For the former, we found several significant sites which were identified in the depth analysis as well, such as sites 137, 144, and 251. These sites are in close proximity to where transducin binds to rhodopsin at the second and third cytoplasmic loops; so amino acid substitutions might alter the dynamic of its binding, for instance, by increasing their affinity to each other, potentially accelerating the signalling kinetics, which would be beneficial at colder temperatures for poikilotherms where the rate of cellular processes might be slower. Several other sites overlapped with those identified in the depth analysis, which indicates their correlation with habitat depth was probably due to the decrease in water temperature at greater depths. Substitutions at these sites are then expected to primarily affect thermal and/or kinetic properties of rhodopsin rather than its spectral sensitivity. In fact, one of these amino acid replacements, T97S, has been shown not to tune $\lambda_{\text{max}}$, but appears to have an effect on the dark-state thermal stability of rhodopsin as well as on the formation and stability of its folded structure. A few RH1 sites showed significant correlation with habitat water temperature but did not correlate with depth. L119H was one of these substitutions which given its proximity to the functionally important residue E122, and the nature of the replacement from a hydrophobic to a positively charged amino acid, warrant experimental characterization. One of the most intriguing correlations with temperature was between a known red-shifting substitution, F261Y, and colder waters. Thermal assays for this mutation have not been done but the Barlow correlation suggests reduced thermal stability for red-shifting substitutions in rhodopsin. If true, then further investigation is needed to examine how a less thermally stable pigment may be advantageous in colder habitats where cellular processes may be slower in poikilotherms. For instance, if the equilibrium between the dark and light-activated states gets shifted towards MII, hence accelerating MII formation and causing greater signal amplification. A less thermally stable
pigment could also be simply more tolerated because of the lower probability of dark noise events at colder temperatures. In ostariophysans, F261Y and five other substitutions in close proximity to each other in the rhodopsin 3D structure have occurred in parallel along a single ancestral branch in accord with shifts to colder waters. How these residues might interact and their effect on RH1 function requires experimental characterization.

This thesis showed several amino acid substitutions in rhodopsin correlating with the ecology of teleost fishes. Some of these sites are functionally characterized and shown to alter spectral, thermal, and kinetic properties of rhodopsin. Other sites have non-conservative substitutions near the chromophore binding pocket, retinal entry or exit channels, dimerization interface, or transducin binding sites which would be ideal candidates for site directed mutagenesis and experimental assay to examine their effect on rhodopsin function. These sites were identified using a combination of existing phylogenetic comparative methods, ancestral reconstructions, and sequence simulations. We utilized two of the most widely used PCMs in a novel manner, and also combined ancestral sequence reconstruction form codon-based models with reconstruction of continuous characters to find statistical associations in a fashion not tested previously. The framework established in this work linked changes in molecular sequences to differences in habitats of species and can be used to investigate correlations between any macroscopic trait and gene of interest.

**Future Directions**

The current project can be expanded in many directions ranging from experimental characterization of the RH1 sites identified, testing other genes in the visual pathway that are expected to have evolved in accord with habitats of species, and enhancing the methodology to
overcome some of the limitations of the current approach. As described earlier, there are RH1 substitutions which based on their location in crystal structure can alter one or several properties of rhodopsin, but have not yet been experimentally assayed. For instance, V271T and A168S both have substitutions to more polar residues, are within 10 Å of the retinal chromophore, and correlated with shifts to deeper habitats, but as far as we know they have not been tested in vitro or in vivo. Other substitutions, such as T97S, have been studied for spectral tuning but their effect on RH1 thermal stability or metarhodopsin II intermediate formation and decay need to be investigated. Results from functional characterization of these substitutions in conjunction with their correlation with habitat depth or temperature can provide a better understanding of their adaptive significance for vision.

Other genes involved in the phototransduction cascade may also be under selection with increase in habitat depth. For instance, substitutions at sites in the G protein transducin which interact with rhodopsin may change the affinity of binding and alter the level of transducin activation. Similarly, duration of signalling may be altered if substitutions in rhodopsin kinase and arrestin affect how quickly rhodopsin is phosphorylated and inactivated. A preliminary analysis of one of the genes involved in the release of retinol from photoreceptors after light exposure, interphotoreceptor retinoid-binding protein (IRBP), has revealed some interesting results. IRBP is a glycoprotein in interphotoreceptor matrix shown to promote release of 11-cis retinal from retinal pigment epithelium cells and release of all-trans retinol from photoreceptor cells following exposure to light (Jin, et al. 2009; Wu, et al. 2007). Tetrapod IRBP contains four repeats (called “modules”) that are possibly the result of several gene duplication events, while teleosts IRBP has only two repeats, although some have a second IRBP gene. A mutation from Asp to Asn in the fourth repeat of human IRBP has been linked to a recessive form of retinitis
pigmentosa and it is proposed that this mutation alters a salt bridge with a conserved Arg two residues away. These two residues, Asp and Arg, were conserved in the available IRBP sequences, belonging to the first module, of the euteleosts used in this study; however, we identified a substitution from Tyr to His at the adjacent site to the conserved Arginine which significantly correlated with habitat depth using PGLS. Median habitat depth of species with His at this site is over 150 m deeper than species with Tyr. This substitution can potentially alter the electrochemical environment in the vicinity of the proposed salt bridge. It is also found in the first repeat of some tetrapod IRBP, including humans, but further investigation is needed to assess its role. Additionally, substitution from Leu to Met at the site between the conserved Arg and Asp has also occurred multiple times in parallel with shifts to deeper regions, although it did not remain significant after phylogenetic correction.

Investigating correlations between species’ characters can be hindered by factors which can be divided into three broad categories: measurement and statistical errors, differences in adaptive response to similar stimuli, and presence of confounding factors. First, inaccuracies in measurements of ecological traits, intraspecific variations in macroscopic characters and molecular sequences, uncertainties in inferences about species relationships and in branch length estimations, as well as errors in statistical modeling and testing can all complicate detection of correlation between amino acid substitutions and macroscopic traits. For instance, here we used midpoint values of species’ depth ranges as a proxy for their habitat depth since most phylogenetic comparative methods can only take a single value in their analysis. Besides the concerns with reliability of depth range measurements, two major problems can potentially arise: the first issue is that the midpoint value might not be the depth at which a species is most frequently found, for instance if members of a species predominantly occupy the two extremes of
the range or if majority of the members occupy a range closer to either end. In the case of latter, we tried to mitigate the problem by using midpoint value of “usual” depth ranges, corresponding to the 10th to the 90th percentile of range, where available. As for the former, occupying two extremes of the range might indicate presence of two different populations of the same species which in turn might experience different environmental pressures and demonstrate different adaptive mechanisms. Including population-level information for those species in the analyses, if available, can enhance the statistical inference. The second issue with using midpoints is that species with very different distribution of ranges can be assigned similar values. One might expect the selection pressure on generalists, which occupy broader ranges of habitats, to differ compared to that on the specialists inhabiting narrower ranges. For instance, midpoint of habitat depth for a species inhabiting 180 to 200 meters is the same as the species inhabiting 0 to 380 meters, but the selective constraints based on availability of light may vary substantially. Developing models that can incorporate ranges of values in their calculations can address this issue in future. Here, we obtained similar results when using the difference between maximum and minimum value of ranges instead of midpoint so the aforementioned limitation may not be a serious concern in this case.

Besides measurement and statistical errors, another major impediment in conducting organismal correlational studies is that adaptation to similar stimuli can take various forms and different species may use different strategies to achieve similar response. For instance, spectral tuning of visual pigments through amino acid replacements near the chromophore binding pocket is a mechanism by which many species better match the available wavelengths of light in their habitat. But chromophore switching, changing opsin expression levels, and expressing different visual pigment genes are other adaptive mechanisms that species are known to employ in
response to changes in ambient light. Also, while spectral properties of water vary according to depth, other ecological factors can modulate availability and quality of light as well. For instance, marine and freshwater habitats have drastically different spectral properties both in terms of which wavelengths penetrate the most and how much of sunlight is absorbed at various depths; sunlight is largely available for species with a habitat depth of 50 meters in an open ocean while it is almost entirely occluded around 50 meters in some rivers. Also, reef-associated fishes may experience different degrees of sunlight occlusion than other marine fish. Additionally, most substitutions alter more than one properties of visual pigment function. A substitution that tunes $\lambda_{\text{max}}$ can change kinetics of active intermediate formation or decay, thermal stability, or hydroxylamine reactivity of opsins. Hence, while the evolutionary adaptive force for fixation of a mutation at a site in a visual pigment may be to maximize its sensitivity to surrounding wavelengths of light in one species, the same mutation may have been fixed due to its effect in changing the rate of visual pigment activation in another species. There are also considerations that need to be made regarding effect of amino acid substitutions on protein function. For visual pigments, forward and reverse substitutions at the same site may not necessarily have opposite effects on $\lambda_{\text{max}}$, a substitution can tune $\lambda_{\text{max}}$ while its reverse does not change it, or the magnitudes of tuning can be different for the two (Yokoyama 2008). The background protein in which a mutation is introduced is also important, so the same substitution can change spectral properties of a visual pigment differently from one species to another, possibly due to other amino acid differences between the orthologs. Since substitutions at several sites can alter opsin function similarly, a substitution can be followed by a compensatory substitution at a different site to restore initial function if environmental pressures change throughout evolutionary timescale. Incorporating as much information as known about ecology
and effect of substitutions in the analyses can minimize the effect of confounding factors and improve conclusions derived from the observed associations.

In the future, the methodology employed in the current study can also be improved to address some of the limitations associated with ancestral inferences and correlations. For instance, ancestral sequences were reconstructed in this study under M8 in PAML, which was the model with the best fit across the random-site models tested. It should be noted that while these random-site models allow for rate heterogeneity across sites, they assume these rates are equal over the entire tree. However, substitution rates may also vary across different clades or branches of the phylogeny, in which case use of branch-site or clade models may provide a better fit to the data. We did not test these models because we did not have a specific, *a priori* hypothesis regarding which branches or clades may have different substitution rates compared to the rest (which is a requirement for all branch, branch-site, and clade models in PAML although there are other models that do not require *a priori* specification of branches or clades, such as branch-site REL model in Hyphy). Nevertheless, even if they provide a better fit compared to the random-site models, it is not clear if ancestral reconstructions from models with a better fit would be more accurate than reconstructions under random-site models. More parameter-rich models can also increase the risk of overfitting the data. Yet, testing more models and examining how consistent results are across them can establish a higher degree of confidence about our conclusions. Finally, developing models which incorporate molecular sequences and ecological or phenotypic traits and test for correlations between amino acid substitutions and changes in the traits under a Likelihood or Bayesian framework would further enhance the current study.
References


Larmuseau MHD, Raeymaekers JAM, Ruddick KG, Van Houdt JKJ, Volckaert FAM. 2009. To see in different seas: spatial variation in the rhodopsin gene of the sand goby (*Pomatoschistus minutus*). *Molecular Ecology*. **18**: 4227-4239.


Moury B, Simon V 2011. dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus y. *Molecular Biology and Evolution* **28**: 2707-2717.


