Structure-Aided Detection of Functional Innovation in Protein Phylogenies

by

Jeremy Bruce Adams

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2015

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Detection of positive selection in proteins is both a common and powerful approach for investigating the molecular basis of adaptation. In this thesis, I explore the use of protein three-dimensional (3D) structure to assist in prediction of historical adaptations in proteins. Building on a method first introduced by Wagner (*Genetics*, 2007, 176: 2451–2463), I present a novel framework called Adaptation3D for detecting positive selection by integrating sequence, structural, and phylogenetic information for protein families. Adaptation3D identifies possible instances of positive selection by reconstructing historical substitutions along a phylogenetic tree and detecting branch-specific cases of spatially clustered substitution. The Adaptation3D method was capable of identifying previously characterized cases of positive selection in proteins, as demonstrated through an analysis of the pathogenesis-related protein 5 (PR-5) phylogeny. It was then applied on a phylogenomic scale in an analysis of thousands of vertebrate protein phylogenetic trees from the Selectome database. Adaptation3D's reconstruction of historical mutations in vertebrate protein families revealed several evolutionary phenomena. First, clustered mutation is widespread and occurs significantly more often than that expected by chance. Second, numerous top-scoring cases of predicted positive selection are consistent with existing literature on vertebrate protein adaptation. Third, in the vertebrate lineage, clustered mutation has occurred disproportionately in proteins from certain families and functional categories such as zinc-finger transcription factors (TFs). Finally, by separating paralogous and orthologous lineages, it was found that TF paralogs display significantly elevated levels of clustered mutation in their DNA-binding sites compared to orthologs, consistent with historical DNA-binding specificity divergence in newly duplicated TFs. Ultimately, Adaptation3D is a powerful framework for reconstructing structural patterns of historical mutation, and provides important insights into the nature of protein adaptation.

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Andrew C. Doxey for his support and guidance of my MSc research. I would also like to thank him for his patience, motivation, and knowledge over the past two years. His guidance helped me during all of my research and thesis writing. I do not think I could have had a better advisor and mentor for my MSc study.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Brendan McConkey, and Dr. Barbara Moffatt, for their insightful comments, and for their questions which led me to look at my research from different perspectives.

I thank my fellow lab mates for the stimulating discussions and fun memories we shared.

I would also like to thank my family for supporting me throughout writing this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

All life on this planet owes its existence to a long and fascinating history of evolutionary adaptation. The genomes of living species can act as a historical record for these ancestral adaptations, which ultimately act on the genes and functional elements encoded at the genomic level. In this thesis, I aim to develop a method capable of reconstructing historical, molecular adaptations from existing bioinformatics datasets, and apply it *specifically* to particular gene families of interest as well as *broadly* to screen a large catalog of gene families and organisms. The development of methods to infer historical adaptations is essential in order to understand both the history of life as well as functionally interpret and annotate the vast and growing collection of incompletely characterized, genomic sequence data.

1.1 Protein adaptation and selection

An evolutionary adaptation can be described as an "adapted trait" that has evolved through natural selection, or alternatively the evolutionary process that generates such traits. According to Darwinian natural selection, traits that confer a fitness advantage in the context of environment to individuals in a population are favoured and become overrepresented compared to traits that impede an organism's ability to survive, grow, and reproduce (Demetrius & Ziehe, 2007). Whether a particular trait is beneficial to species fitness is dependent on the environment that the species finds itself in; and a trait that enhances fitness in one setting can hinder fitness in another. Because different organisms have adapted to inhabit virtually all environments on earth, each species has evolved traits that uniquely enable them to perform optimally in their respective climates and habitats.

Traits that are strongly favoured in an environment have a tendency to become fixed in a population, because individuals possessing the trait have a greater chance of passing on their characteristics to successive generations over others that do not carry the beneficial trait (Mitchell-Olds, Willis, & Goldstein, 2007; Rieseberg, Widmer, Arntz, & Burke, 2002). The process by which a trait increases to high frequency or fixation within a population is known as *positive selection*. In a genetic context, alleles that encode for the beneficial trait become more prevalent in the population over a period of generations, ensuring that progeny are more and more likely to carry the specific allele. Conversely, alleles that are detrimental to an individual's survival and reproductive ability tend to be removed from the population (Charlesworth, B., Morgan, Charlesworth, D., 1993; Hudson & Kaplan, 1995). This process is called negative selection or purifying selection and underlies observed patterns of evolutionary *conservation*.

Mutations that occur within key sites of genes and genomes can generate new traits and alleles that did not previously exist. Since biological traits can be modulated with respect to the environment, it is necessary to study the molecular basis of these traits to see how modification of molecular information can translate to whole organism physiological and morphological differences. For example, mutations in non-coding *cis*-regulatory sites can alter transcription factor binding preferences, altering gene expression patterns and changing organism development and morphology (Prud'homme *et al.*, 2006; Wray, 2007). While mutations in non-coding regions certainly affect phenotype, it can be argued that mutations in protein-coding regions can potentially have even more drastic effects on an organism's overall phenotype (Halligan *et al.*, 2013). For instance, the regulation of hundreds or thousands of genes may potentially be affected by just a few substitutions that alter the binding specificity of a single transcription factor.

Proteins are the most functionally diverse type of biological macromolecule and the basic building blocks of cellular systems, and include enzymes, structural proteins, signaling molecules, and transcription factors (Gutteridge & Thornton, 2005). Collectively, proteins, through interactions with other biological molecules (DNA, RNA, lipids, carbohydrates), perform the fundamental cellular processes of life. While many proteins are shared across the tree of life and perform consistent functions, lineage-specific adaptations occur through several mechanisms. First, orthologous proteins may diverge in function between species through modifications; this happens without gene duplication but rather through lineage-specific substitutions. Second, duplicated genes can generate new gene copies that are free to diverge in function and adopt entirely new roles, a process called *neofunctionalization* (Falciatore *et al.*, 2005; Fitch, 1970; Zhang, 2003). According to the gene duplication model of Ohno (1970), one gene duplicate becomes free to mutate, potentially takes on a new function without any detriment to the host organism because the original copy remains the same and performs the canonical function (Taylor & Raes, 2004). In this way, paralogous genes can evolve to perform similar, yet unique functions, such as catalyzing different substrates or binding to different partners (Grove, Willcox, Griffith, & Bryant, 2005; Yang *et al.*, 2013). However, it is important to note that extensive changes to a protein do not always result in a modification of the protein's canonical function, but may still play an adaptive role by maintaining optimal protein function in changing cellular conditions. Indeed, proteins tend to be optimized for narrow temperature, pH, and salinity ranges (Arfi *et al.*, 2013; Dubnovitsky, Kapetaniou, & Papageorgiou, 2005; Siddiqui & Cavicchioli, 2006; Siddiqui *et al.*, 2006). Various changes in the amino acid sequence can result in the protein being optimized for new environments that affect the organism's internal state. Poikilotherms, temperature non-regulators that can live at a wide range of temperatures have a very high degree of gene duplication and paralogy (Genge, Davidson, & Tibbits, 2013; Moon & Hochachk, 1971). A high number of gene paralogs allows for the same function to be accomplished at different temperatures by expression of a range of temperature-adapted protein variants (e.g., cold-adapted enzyme variants produced by the winter rye plant, *Secale cereale*) (Griffith & Yaish, 2004; Yaish *et al.*, 2006).

Observed rates of amino acid mutations, when evaluated in a structural context, have demonstrated that protein surfaces evolve considerably faster than the internal sites (Toth-Petroczy & Tawfik, 2011). This is thought to partially reflect ongoing re-wiring of protein-protein and protein-ligand interactions and divergence of interaction networks between species. Even slight surface mutations may affect function by modifying a variety of interactions including post-translational modifications [e.g., acetylation, phosphorylation, and methylation (Beltrao *et al.*, 2009; Ghanta, Grossman, & Brenner, 2013; Grewal & Rice, 2004; Nakayam *et al.*, 2001)], as well as protein-protein interactions and pathways (Jin *et al.*, 2013, Mintseris & Weng, 2005).

Since many selected biological traits are encoded at the molecular level, biological molecules themselves may be considered as under selection. Furthermore, because the functions of biomolecules like proteins are ultimately encoded by sequence, individual sites within these sequences are also subject to selective forces. This provides a powerful means for computationally detecting selective pressures on genes and proteins through comparative sequence analysis.

There are several classes of computational methods that have been designed to detect positive selection on genes and other regions of the genome. Detecting positive selection is of tremendous biological and evolutionary importance since it opens up the possibility of identifying where, when and possibility how proteins have likely altered function within the tree of life. We are no longer limited by the availability of genomic information for many organisms; however, our interpretation of the genomic differences between species and the adaptive forces that underlie these differences is still poor. Computational detection of selection from sequence information is therefore an important tool for making informed predictions about protein neofunctionalization and generating hypotheses for future lab experimentation.

1.2 Computational methods for detecting positive selection in proteins

There are two commonly used classes of methods for sequence-based detection of positive selection: population-genetics based methods, and K_a/K_s ratio based methods. Each of these is summarized below and thoroughly reviewed elsewhere (Vitti, Grossman, & Sabeti, 2013).

1.2.1 Population genetics based methods

A major class of computational methods for detecting positive selection involves the sequencing and analysis of a genomic locus from multiple individuals in a population. These methods operate on a *microevolutionary* scale (within population), compared to a macroevolutionary scale, which examines evolutionary phenomena between species (Vitti, Grossman, & Sabeti, 2013). Through analysis of the prevalence and genomic position of alleles, it is possible to predict whether a genomic region has likely undergone recent positive or purifying (negative) selection. Traditional analyses were focused on subsets of genomes and genomic regions (e.g., information derived from SNP arrays), whereas recent high-throughput sequencing methods have made it possible to expand these analyses to entire genomes (Begun *et al.,* 2007; Davey & Blaxter, 2010).

Positive selection acts to rapidly increase the prevalence of certain alleles within a population to high frequency or fixation (100% prevalence). It has been observed that positive selection on alleles also has an effect on loci adjacent to that allele since neighboring alleles may be co-inherited, a pattern known as *linkage disequilibium* (Harpur *et al.,* 2014; Slatkin, 2008; Akey, 2009; Bamshad & Wooding, 2003). Therefore, once a beneficial variant *sweeps* to high frequency, a set of nearby linked alleles, which collectively define a *haplotype,* also exhibit much of the same pattern (Palaisa, Morgante, Tingey, & Rafalski, 2004). There are several methods to determine the strength of a selective sweep at a genomic region within in a population by scanning genomic regions for evidence of selective sweeps.

In frequency-based methods such as Tajima's D (Tajima, 1989), selective sweeps from a positively selected allele lead to a high number of low frequency, derived alleles in the region of interest when compared to a site that is under neutral selection.

Another class of positive selection detection methods are based on analyzing patterns of linkage disequilibrium (LD) (Slatkin, 2008). Genomic regions under neutral selection are expected to have a high number of relatively equal frequency haplotypes due to historical recombination that has broken linkage patterns throughout the region. However, loci that are undergoing or have recently undergone a selective sweep have had less chance to undergo genetic recombination. This is because the beneficial allele rapidly increases in the population, leaving little time for recombination. One strategy for inferring selective sweeps in a genomic region is therefore to evaluate the sequence homogeneity and length of detected haplotypes across a region by measuring LD patterns. The extended haplotype homozygosity (EHH) test estimates the age of haplotypes by assessing the extent to which LD patterns have been broken down by recombination (Sabeti *et al.*, 2002). EHH measures this by evaluating LD patterns as a function of distance to the center of each haplotype block.

Lastly, population differentiation methods are a class of tests that infer positive selection by identifying alleles that have unusually high frequency in one portion of a population versus another (i.e., two geographically separated subpopulations) . A common measure of population differentiation for a genomic locus is the *fixation index* (F_{ST}), which compares allele frequency variance (e.g., as determined by comparing SNPs) between populations (Holsinger & Weir, 2009). If a particular genomic region has a large F_{ST} value compared to that seen elsewhere for neutral regions, then it may have been under selection. F_{ST} and related methods are useful for finding instances of geographically restricted selection: cases in which regional differences can lead to an existing allele being selected for. In all classes of tests, statistical methods are used to determine if the polymorphism observed significantly deviates from polymorphic models that are expected in regions that undergo neutral selection. Population genomics based methods have yielded important insights into the selective forces acting on the genomes from a variety of species. Positive selection studies in humans collectively give us the largest map of selected sites in any species. A previous study identified 722 human genomic regions containing 2465 genes under positive selection (Akey, 2009; Vallender & Lahn, 2004). Many of these genes have been found to be involved in olfaction, cell cycle regulation, and synaptic transmission, among many other functions according to assessment of gene ontology (GO) terms. One well known example of positive selection detected in the human population by genome-wide screening is the detected selection on variants within the *LCT* gene (Bersaglieri *et al.*, 2004) which encodes the lactase enzyme. Positive selection on the *LCT* gene is thought to coincide with the spread of dairy farming after the colonization of Europe. Another interesting recent example includes detection of strong selection on East Asian variants of the *EDAR* gene, which play a role in ectodermal tissue development (Grossman *et al.*, 2010; Kamberov *et al.*, 2013). These methods are not restricted to human population studies. For instance, a recent population genomics study on honeybees uncovered positive selection on genes related to worker traits and sociality (Harpur *et al.*, 2014).

Population genomic methods are very useful in making predictions about candidate regions that have had an advantageous and significant effect on a species' physiology. Such analyses can help bridge the gap between genotype and phenotype by finding likely genomic loci that are responsible for functional differences between species. Population genomics methods are exceptionally useful because they can be used to find selection across all regions of a genome including non-protein coding, *cis*-regulatory sites. However, they operate over relatively short (microevolutionary) time scales, require information on population genome variation and often can identify general regions but cannot pinpoint precisely which sites within an apparently selected region represent causal, selected variants.

1.2.2 Detection of selection using the K_a/K_s ratio

The K_a/K_s ratio is a method template for determining whether purifying or positive selection has taken place on a protein-coding gene, and operates more on the macroevolutionary (between species) than microevolutionary (within population) scale. In its most basic form, the method compares the aligned DNA sequences from two homologous genes (Yang & Bielawski, 2000). K_a refers to the number of nonsynonymous substitutions per nonsynonymous site that have occurred between the sequence pair. K_s refers to the number of synonymous mutations per synonymous site that have occurred between the sequences. Using the genetic code (codon translation table), and a pairwise sequence alignment, it is possible to calculate these two ratios quite simply. The final ratio of K_a/K_s, also called ω, is the measure of how much selection has occurred between the sequences. A sequence pair that is under neutral selection is expected to have a relatively equal rate of nonsynonymous and synonymous substitutions, which is therefore equivalent to a K_a/K_s ratio close to 1 (Hurst, 2002). A sequence pair under purifying selection is expected to have a much lower rate of nonsynonymous substitution compared to synonymous substitution. This is because non-synonymous substitutions likely lead to changes that cause protein dysfunction, and are thereby purged from the gene pool. Therefore, a K_a/K_s score of less than 1 signifies purifying or negative selection. Lastly, a sequence pair that is under positive selection is expected to have a much higher rate of nonsynonymous mutation compared to synonymous mutation (Hurst, 2002; Nei & Gojobori, 1986). This is because the non-synonymous substitutions cause adaptive changes in the protein's biochemical function, and so an overrepresentation of amino acid changes means that there has been positive selection to change amino acid properties and ultimately, protein function. Therefore, a K_a/K_s score of greater than 1 signifies positive selection.

Although this is the most fundamental use of the K_a/K_s ratio, there are some inherent drawbacks with this method. First, genes do not necessarily need to have many amino acid changes throughout its sequence to cause a significant change to the encoded protein's function. Functional changes can occur in proteins by very few amino acid substitutions in key binding or catalytic sites (Hughes, 2008; Yokoyama, Tada, Zhang, & Britt, 2008; Doxey et al., 2006; 2010). For example, single amino acid mutations in opsin genes are sufficient to cause changes in light spectra absorbance (Hughes, 2008; Yokoyama, Tada, Zhang, & Britt, 2008). Since general K_a/K_s

statistically requires many non-synonymous mutations throughout the sequence to infer positive selection, they are inherently tailored to finding positive selection that has occurred broadly and repeatedly throughout a protein (Hughes, 2007).

In response to some of these drawbacks, there have been several enhancements to the K_a/K_s ratio model, including site-specific K_a/K_s models (Nielsen & Yang, 1998), branch-specific K_a/K_s models (Yang, 1998), and branch-site K_a/K_s models (Yang & Bielawski, 2000).While the simple K_a/K_s model determines a general level of selection between two sequences, site-specific K_a/K_s models (Nielsen & Yang, 1998) finds specific sites along gene segment where selection has occurred (Creevey & McInerney, 2002; Yang & Bielawski, 2000). Using the multiple sequence alignment, a K_a/K_s score is determined for each individual codon in the alignment, with the expectation that each codon will have its own unique rate of nonsynonymous substitution depending on how critical that amino acid is to protein function. If the K_a/K_s score of a single codon significantly deviates from the average observed K_a/K_s ratio, positive selection is said to have occurred at that point in the sequence (Yang & Bielawski, 2000). Statistical methods are employed to make claims about positive or purifying selection based on the divergence between the sequences involved (Fu, 1996; Tajima, 1989).

Although the site-specific K_a/K_s model is used to assert where in the multiple sequence alignment selection has occurred, there is no directionality to the selection. That is, the site-specific model determines whether the average K_a/K_s for a site over all species/lineages is greater than that observed for other sites. Based on this averaging, it is impossible to identify which lineages or combinations of lineages selection has occur (McClellan, 2013; Wong, Yang, Goldman, & Nielsen, 2004). Also, it cannot be inferred which states or characteristics are more derived, and which are ancestral.

Branch-specific models are a third class of K_a/K_s methods that take as input a phylogenetic tree and a multiple sequence alignment. An evolutionary model, as represented by the phylogeny, takes into account K_a/K_s ratios and branch lengths, and determines in which branches positive or purifying selection has occurred (Nickel, Tefft, Goglin, & Adams, 2008). Here, the average K_a/K_s ratio for all sites is computed and compared between branches. The branch K_a/K_s model therefore overcomes the lack of directionality of site-specific models, but loses information on which specific sites are under selection. Also, these algorithms are optimized to deal with closely related sequences and recent evolutionary divergence, which is not the case with site-specific models.

Site-specific and branch-specific K_a/K_s methods are used to make more specific claims about where and when positive selection has taken place in the evolution of a gene or protein family. The natural progression of these models is thus to combine the two into a model of tests that both determine where in a phylogenetic tree and where specifically in a gene sequence that selection has occurred (Yang & Bielawski, 2000; Zhang, Nielsen, & Yang, 2005). This method is called the branch-site model and uses a likelihood ratio test (LRT) for identifying both lineage- and site-specific positive selection given a tree and multiple alignment, allowing K_a/K_s to vary both across sites and across lineages.

All types of K_a/K_s methods have been used extensively to make claims about positive selection. The traditional K_a/K_s ratio was first used to find positive selection in human MHC1 compared to chimpanzee (Hughes, & Nei, 1988; Hughes, & Nei, 1989; Hughes, Ota, & Nei, 1990). An entire database of positive selection called *Selectome* (Moretti *et al.,* 2014; Proux *et al.*, 2009) has been compiled to detect positive selection across a wide array of protein families in many different vertebrate species. Many studies are published each year claiming detection of positively selected genes from diverse taxonomic groups including mammals, fish, and insects (Areal, Abrantes, & Esteves, 2011; Dunning *et al*, 2013; Tong *et al.*, 2015).

Although the above methods have been used to detect positive selection in many different protein families and in many different clades, they have been critiqued for a variety of reasons. Some have stated that the K_a/K_s tests have little grounding in real biology, and as such, a significant K_a/K_s ratio may predict instances where there is a large degree of mutation, but not necessarily mutation that is likely to significantly affect the protein's function or biochemical properties (Hughes, 2007; Hughes, 2008). Another critique is that some of the early K_a/K_s models do not distinguish between positive selection and the relaxation or absence of purifying selection (Arbiza, Dopazo, J., & Dopazo, H., 2006). More complicated algorithms have been developed to correct for this (Zhang, Nielsen, & Yang, 2005).

Critiques of the branch-site models are based on evidence of high rates of false positive predictions produced when there are low numbers of nucleotide substitutions in foreground branches (Nozawa, Suzuki, & Nei, 2009). Schmid and Yang (2008) found that site-specific models of positive selection are likely to produce false results because they fail to correct for multiple hypothesis testing. A study found that several factors led to an inflation of false-positive positive selection results, including small numbers of nonsynonymous mutations in foreground branches, and when incorrect assumptions were made regarding certain parameters (Suzuki, 2008). With these critiques, it is evident that there is a lot of room for improvement in the field of selection detection algorithms.

1.2.3 Wagner's method for detection of clustered mutation

A third and less commonly used approach for detecting positive selection is based on the idea of identifying linearly or spatially *clustered* substitutions in a protein (Wagner, 2007), which is in part motivated by the notion that selection often acts only on small regions of proteins.

The basis of Wagner's method is that positive selection may drive the accumulation of beneficial variants within particular regions of proteins at a faster rate than other regions, and thus lead to "clustering" of observed substitutions. Wagner's method therefore detects what he calls "variation clusters", which are groups of aggregated substitutions that are too close to one another to have arisen by chance alone. Conceptually, this idea is not too different from genome-wide methods that seek to identify particular regions or genes (which can be considered genomic "clusters") containing an excess of derived mutations.

Wagner (2007) developed two statistical approaches for determining variation clusters. First, Wagner examined the positions of all observed substitutions in a linear sequence alignment, and compare the distances between substitutions to that expect by a null (Poisson) distribution. Second, 3D clustering of mutations was evaluated by measuring the Euclidean distances between all k substitutions when mapped onto a protein 3D structure, computing the average pairwise distance d_{k}, comparing d_{k} to a distribution obtained by a random sampling, and computing the statistical significance as a permutation-based P-value (P_{3D}). In other words, this test evaluates whether the observed substitutions are closer to one another than that expected by chance assuming a null model in which mutations occur uniformly throughout a sequence.

To account for several alternative explanations for clustered mutation, Wagner (2007) demonstrated that observed variation clusters were not caused by high mutability of CpG-rich regions, low complexity regions, or simply relaxed selection (high mutation rates) as determined by four-fold degenerate sites. Finally, Wagner showed that among the highest-scoring predictions of clustered mutation were previously known examples such as BRCA1 and VN1R1.

Wagner's P_{3D} method has since been expanded, refined, and applied in several other studies (Zhou, Enyeart, & Wilke, 2008) including analysis of cancer mutations (Ye *et al.*, 2010).

Wagner's P_{3D} method, though potentially powerful, has not yet been integrated with phylogenetic analysis in order to identify historical and lineage-specific clustered substitution. If clustered mutation could be evaluated at all stages of protein evolution given a phylogenetic tree, this would represent a potentially new approach to identify lineage-specific positive selection. One possible approach for integrating phylogenetics with spatial clustering, would be to infer all ancestral sequences within a phylogenetic tree and compute Wagner's P_{3D} statistic across all branches. This will be the core approach used in my proposed Adaptation3D pipeline. First, I will review methods for ancestral sequence reconstruction below.

1.3 Ancestral Sequence Reconstruction

Closely linked to the idea of detecting positive selection in gene and protein family evolution is the idea of ancestral sequence reconstruction. That is, if it is possible to accurately infer ancestral sequences, then the sequences at ancestral versus derived nodes in a tree may be directly compared using methods described above.

Most evolutionary analyses compare present-day (extant) sequences in a horizontal manner. Only through the incorporation of phylogenetic trees does the analysis take on a vertical, or time-dependent perspective. Ancestral reconstruction involves using extant sequences to infer ancestral sequences that existed before speciation or gene duplication events (Cai, Pei, & Grishin, 2004; Williams, Pollock, Blackburne, & Goldstein, 2006). Thus, ancestral reconstruction methods will use current sequences and their evolutionary history to "reconstruct" or hypothesize the most likely sequence that existed at each ancestor node in a phylogenetic tree. Ancestral reconstruction is important in understanding protein function, and how evolutionary events produce protein functional shifts (Chang, Ugalde, & Matz, 2005).

Similar to phylogenetic tree construction algorithms, there are three main groups of ancestral reconstruction algorithms. There are maximum parsimony based methods, maximum likelihood based methods, and Bayesian inference methods.

Maximum parsimony ancestral reconstruction methods are the oldest and simplest model for recreating ancient sequences (Fitch, 1971). The principle of parsimony is predicated on the idea that an evolutionary history hypothesis with the least number of transition or mutation events must be the correct hypothesis (Minaka *et al.,* 2008; Omland, 1999; Williams, Pollock, Blackburne, & Goldstein, 2006). In the Fitch (1971) algorithm for parsimony reconstruction, each type of nucleotide or amino acid mutation is weighted equally, and ancestral states are inferred that minimize the total number of mutations in the tree. This hypothesis of sequence reconstruction can be faulty in some of its assumptions. For example, it is not true that evolution works quickly towards some goal in the future, so designing a method in which the fewest changes is the best answer will not necessarily be true (Mooers, & Schluter, 1999). Also, the parsimony method assumes that the rate of evolution is constant in all branches of the tree, which is not accurate either. To overcome many of the drawbacks of maximum parsimony based sequence reconstruction, the maximum likelihood methods were developed.

Maximum likelihood methods allow for more complex scenarios than parsimony reconstruction because they are likelihood frameworks that incorporated sophisticated models of character evolution and even variable rates of change at different sites and branches within the tree. Transition and transversion substitutions are not considered equally likely to happen, and so the weighting inherent to maximum likelihood reflects different probabilities of events occurring (Pagel, 1999). Some of the drawbacks of maximum likelihood reconstruction include an overestimation of protein stability because it assumes that the proteins provided are always very optimized and stable (Williams, Pollock, Blackburne, & Goldstein, 2006). Some have critiqued maximum likelihood for being too computationally intensive for not enough of a benefit over maximum parsimony methods (Doornik, & Ooms, 2003). In general, maximum likelihood is considered to provide more accurate reconstructions compared to maximum parsimony methods overall (Guindon & Gascuel, 2003).

Bayesian inference methods are considered the most versatile and robust of the three paradigms of ancestral sequence reconstruction. Hypotheses about ancestral amino acid states are tested by combining the likelihood of observed data with the likelihood that a certain order of events have occurred (Cunningham, Omland, & Oakley, 1998). Bayesian methods are considered advantageous over maximum parsimony and maximum likelihood methods primarily because Bayesian methods provide a distribution of likely possible trees, as opposed to a single estimate (Huelsenbeck, & Ronquist, 2001).

To date, the Phylogenetic Analysis by Maximum Likelihood (PAML) software suite is the most used tool for analyses for ancestral sequence reconstruction (Yang, 1997; Yang, 2007). Many studies on ancestral protein function are published each year that use PAML to hypothesize the amino acid sequences of ancient proteins (Assis & Bachtrong, 2015; Wallis, 2015). One critique of PAML is its inability to handle gaps in a multiple sequence alignment (McGuire, Denham, & Balding, 2001). Codeml, the amino acid reconstruction program of PAML, has a tendency to overestimate amino acids as ancestral states for gap-containing positions, because it treats ambiguous residues and likely gaps the same: that is, as a site to be filled in with some character. This results in a tendency for ancestral sequences to be longer than the extant sequences, with gaps not appearing in the phylogeny until more recent, derived lineages (Yang, 2007). A second issue related to this is that divergent alignments lead towards very large ancestral amino acid sequences. This is also unrealistic biologically, and hence codeml is generally applicable to cases where there is strong conservation in an alignment. For these reasons, the codeml maximum-likelihood framework has not been used to infer ancestral protein sequences in the pipeline developed in this thesis. Instead, a program called FastML has been used because of its proper treatment of gap characters (Ashkenazy *et al.,* 2012; Pupko, Pe’er, Shamir, & Graur, 2000).

1.4 Overview of Adaptation3D: A novel method to predict protein adaptation

As described above, it is evident that there is great room for improvement of methods for positive selection and/or protein adaptation detection. The basic K_a/K_s method is underpowered to find important adaptation events because it requires a high rate of nonsynonymous mutation over the whole length of the protein (Hughes, 2007; Hughes, & Nei, 1988; Tajima, 1989). On the other hand, the site-specific and branch-specific K_a/K_s models are often too sensitive and incorrectly detect natural variations and misalignments in MSAs as sites of selection (Nozawa, Suzuki, & Nei, 2009; Schmid, & Yang, 2008; Suzuki, 2008). Wagner’s method for detecting spatial clustering of mutation is potentially powerful and overcomes some of these limitations, however it has only been developed and applied to pairwise sequence comparison.

This thesis therefore entails the design and development of a novel software pipeline called Adaptation3D for detecting positive selection and functional adaptation in proteins. Adaptation3D integrates several of the methods described earlier, specifically combining phylogenetic ancestral sequence reconstruction and Wagner's method for detecting spatially clustered mutation. Adaptation3D therefore detects positive selection in specific proteins and evolutionary lineages by identifying **lineage-specific clustered mutation.** Furthermore, Adaptation3D has been implemented as a **high-throughput screening tool** and thus is capable of searching large databases of protein phylogenies for adaptations by automatically retrieving and analyzing structural data from the Protein Data Bank (PDB).

In the following thesis, I will explain the methodology behind this novel pipeline, as well as how this tool has been applied in various ways to address different question in protein adaptation. In chapter 2, I will explain, in detail, how the pipeline works, and the reasoning behind some of the models and assumptions inherent in the method. In chapter 3, I will outline and discuss the results of running the pipeline on four different biological datasets. First, I will discuss how the novel pipeline predicted the correct sites and lineage of adaptation in the pathogenesis related protein PR-5D. Second, I will discuss the predictions about functional adaptation made by the novel pipeline when analyzing the Selectome database of positive selection (Moretti *et al.,* 2014; Proux *et al.*, 2009). Third, I will apply Adaptation3D to analyze the structural evolution of vertebrate transcription factors, and uncover significant differences in the evolutionary dynamics of paralogs versus orthologs in the vertebrate lineage.

Chapter 2

The Adaptation3D framework for protein adaptation detection: method design

2.1 Methodology overview and input

Adaptation3D detects positive selection in proteins by identifying lineage-specific, spatially clustered mutation. Given a phylogeny, multiple sequence alignment, and protein structure, Adaptation3D infers ancestral sequences, and for each branch it computes the set of substitutions that have occurred and their degree of spatial clustering. Adaptation3D therefore addresses the question of protein adaptation in two dimensions: in which lineage/taxonomic group has the change occurred, and where on the protein tertiary structure has the change occurred.

To start off, Adaptation3D requires as input an amino acid multiple sequence alignment file in FASTA format, and a phylogenetic tree file in Newick format (Cardona, Rossello, & Valiente, 2008; Lipman, & Pearson, 1985). In order for the program to execute, there must be perfect correspondence between the FASTA sequence headers and the names of the terminal nodes in the tree file. For example, if there is an unequal number of sequences and node termini, an error will be raised. Similarly, an error will be raised if one or more sequence header(s) in the FASTA file do not have a terminus with the exact same name in the Newick file. Adaptation3D will use the sequences for ancestral state reconstruction which is why there must be a real sequence for each end position in the tree.

2.2 Precomputation of PDB structural features

Adaptation3D uses protein structures in the PDB as template 3D models for query proteins of interest (Berman *et al.*, 2000; Drew *et al.*, 2011). It is these template structures that are used to assign key structural information (pairwise residue distances and solvent accessibility) to the query protein family being analyzed, which is in turn used to detect spatial clustering.

An inefficient strategy is to BLAST a query family against the PDB and compute structural information on-the-fly, whereas an efficient strategy is to pre-compute structural features for all entries in the PDB database and store these features as a database for later fast retrieval. This avoids unnecessary re-computation, and only the PDB-BLAST step is required to determine a query/template residue mapping and transfer of pre-computed structural information.

Pre-computation of structural features was performed as follows. Solvent accessible surface area (ASA) was calculated for all PDB entries using the POPS algorithm (Cavallo, Kleinjung, & Fraternali, 2003) and stored in a database. Positions of residue α-carbons were also retrieved to calculate inter-residue distances, which were also stored as a database for later use. Finally, a BLAST database representing a snapshot of the PDB (May 30, 2015) was constructed out of the sequences to facilitate later BLAST searches (Altschul *et al.*, 1990; Camacho *et al.*, 2008).

2.3 Protein family ancestral sequence reconstruction and comparison

Adaptation3D takes as input a multiple sequence alignment (MSA) file and a phylogenetic tree file. The PDB-BLAST step is performed on-the-fly by the software. Every terminal node in the tree must have a corresponding sequence in the MSA file. Often a tree and

MSA will depict a gene tree of multiple orthologs/paralogs across several species. The method infers and reconstructs ancestral sequences at all nodes in the phylogenetic tree unaware of the orthologous versus paralogous relationships between proteins. Ancestral sequence reconstruction is done using the FastML program (Ashkenazy *et al.*, 2012; Pupko *et al.*, 2000). The user supplies a protein phylogenetic tree and amino acid sequence alignment. Joint reconstruction is used to reconstruct ancestral sequences (Pupko, *et al.*, 2000). Aligned extant sequences are not directly compared to each other, but instead ancestral sequences are compared and checked for substitutions to their immediate descendants. This structure of the analysis makes it possible to determine what mutations have occurred in what specific lineage. Ultimately, the program determines not only whether clustered mutation (and thus positive selection) has occurred or not, but also determines at what evolutionary time and what clade(s) have been affected by this ancestral event. An example of how the program checks for substitutions between ancestral and derived sequences is displayed in Figure 1.

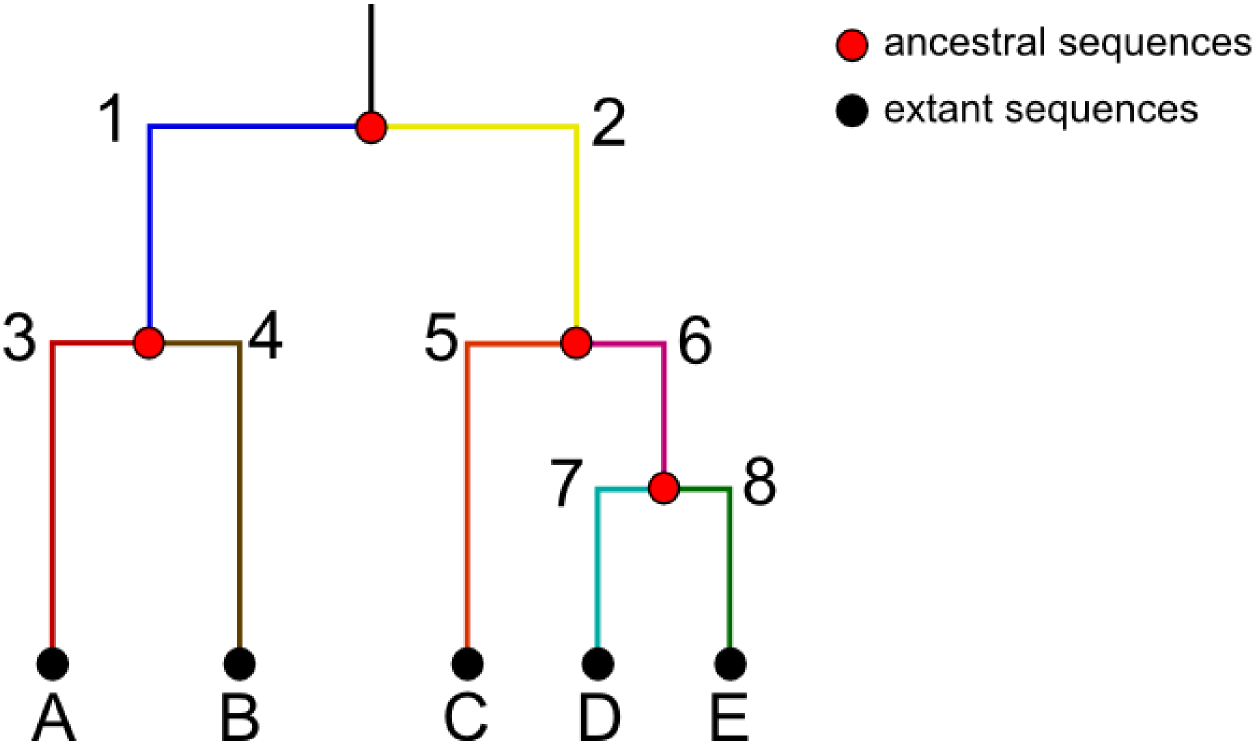


Figure 1: A hypothetical input phylogenetic tree for Adaptation3D, subdivided into ancestral and extant sequences. A multiple sequence alignment of 5 extant sequences (black circles) and a tree is supplied to the Adaptation3D program. An ancestral sequence is ascribed to each internal node of the tree (red circles). Substitutions are mapped along branch segments from ancestral sequences to immediately derived child nodes. Unique branch segments are numbered and coloured individually.

2.4 Alignment of sequences to PDB structural representative(s)

Protein structural information is required for Adaptation3D's assessment of lineage-specific spatial clustering and other structural features to be analyzed later. A single extant sequence that is the most similar to the remaining extant sequences (i.e., the most representative sequence of the family) is determined and used as the query for a protein BLAST (BLASTP) search against the pre-computed snapshot of the PDB (Camacho *et al.*, 2008; Berman *et al.*, 2000). The structure of the query protein can be assumed to be closely related to that of its PDB template over the region for which significant homology is detected by BLAST, a principle that is also the basis of all homology-based modeling methods. By the alignment of the query protein sequence to the PDB structure, it is possible to assign spatial positions to substituted sites, using the PDB target as the template structure. Substitutions are mapped from ancestral to derived sequences at every stage in the phylogenetic tree, and these mutated sites are given 3D positions and geometrical characteristics. Figure 2 depicts how substitutions between ancestral and derived sequences are mapped to structure.

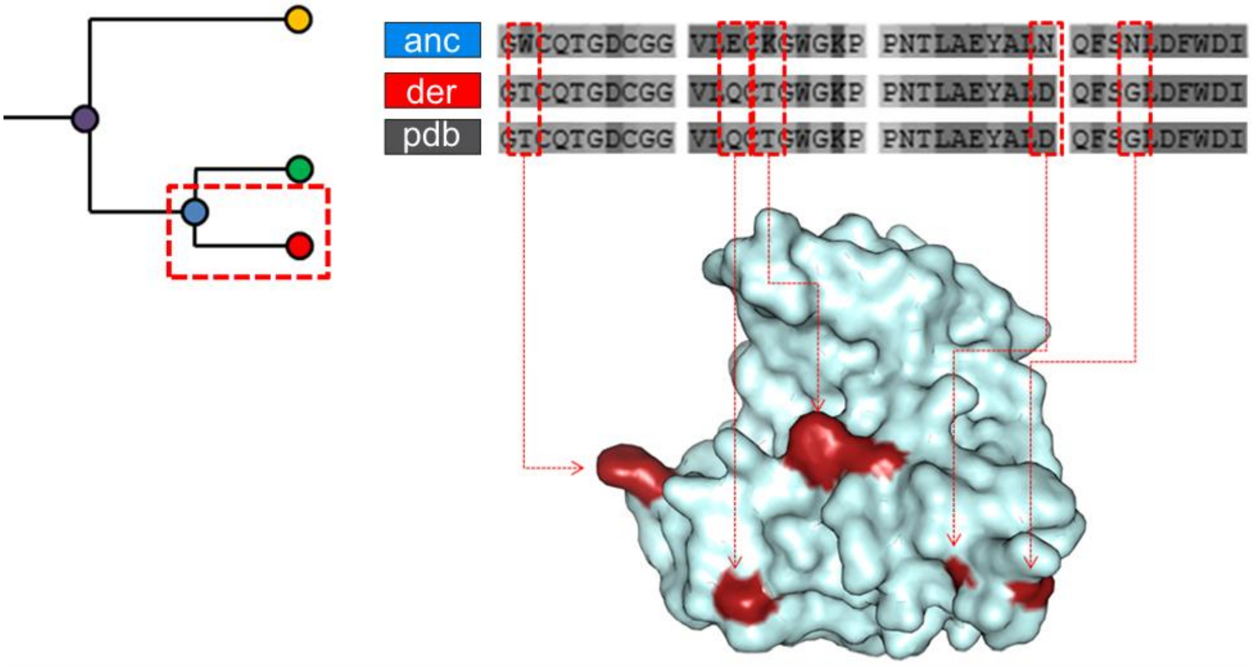

anc
GWCQTGDCGG VLECKGWGKP PNTLAEYALN QFSNLDFWDI
der
GTCQTGDCGG VLQCTGWGKP PNTLAEYALD QFSGLDFWDI
pdb
GTCQTGDCGG VLQCTGWGKP PNTLAEYALD QFSGLDFWDI

Figure 2: Substitutions between ancestral and derived sequences mapped to protein structure. The ancestral-derived sequence pair represented is highlighted in the red box. Sequence mismatches between the ancestral sequence (anc) and derived sequence (der) are given a sequence position to the PDB sequence through the alignment. Substitutions are placed on the PDB structure through the correspondence between PDB sequence residue positions and 3D coordinates.

2.5 Structural features used in detection of protein adaptation

The structural features evaluated by Adaptation3D are derived from the template PDB structures detected via BLAST. The two basic properties evaluated are listed below and one value for each of these criteria is computed for each group of branch-specific substitutions:

1. Mean Euclidean inter-α-carbon distance between substituted sites
2. Mean relative residue side chain solvent-accessible surface area (RASA)

RASA is a measure of the degree of residue side chain burial or solvent exposure. Residues in the core of the protein have a low RASA, while residues on the surface with side chains oriented facing the external face of the protein have a high RASA. A residue's RASA is a ratio of the observed accessible surface area (ASA) of the side chain of a specific residue in the PDB relative to the "ideal" ASA of that side chain, that is, the ASA given if the amino acid existed independently of a protein. Each amino acid type has its own ideal ASA. Ideal ASA calculations for each amino acid type were determined by running the POPS algorithm on a PDB file containing a single amino acid (1 file per amino acid type) (Cavallo, Kleinjung, & Fraternali, 2003). This algorithm was also used to pre-compute ASA information for all residues in all structures for a snapshot of the PDB. Ideal side chain ASA values for each amino acid type are listed in Table 1.

Although RASA was not specifically used as a criterion or structural feature by Wagner (2007), it was included here for later analyses of clustered mutation occurring in surface versus interior regions. Exposed substituted sites are likely to be under different selective pressures and functional constraints compared to buried residues, and though potentially over-simplistic, it makes sense to attempt to classify residues into these two categories.

Table 1: Ideal side chain ASA values for isolated amino acids computed using the POPS algorithm.

Residue	ASA (Å^2)	Residue	ASA (Å^2)
Alanine	102.77	Leucine	205.91
Arginine	279.20	Lysine	246.36
Asparagine	174.75	Methionine	213.64
Aspartic Acid	172.69	Phenylalanine	224.78
Cysteine	143.36	Proline	170.15
Glutamic Acid	214.78	Serine	126.89
Glutamine	220.26	Threonine	182.33
Glycine	47.92 (α-carbon)	Tryptophan	243.47
Histidine	221.38	Tyrosine	261.73
Isoleucine	241.44	Valine	204.59

The decision to compute RASA versus actual ASA values for each sidechain becomes important when considering how to infer the solvent-accessibility of new substitutions. For instance, if an exposed alanine (small sidechain) in the template substitutes for an arginine (large sidechain) in the query, it is not appropriate simply to transfer alanine's actual computed ASA from the template to the query. Thus, Adaptation3D instead transfers the RASA from the ancestral to derived amino acid. For example, if a mutation at position 25 in the MSA aligns to a proline residue with an ASA of 85.075 in a PDB structure, then the RASA at position 25 is 0.5 (85.075 / 170.15). The RASA value for a substituted site group is the mean RASA for each site in the group.

An analysis was performed to determine the accuracy of comparing relative ratios of side chain solvent accessibility between substituted sites on a PDB structure. The Mutaprot database of homologous PDB structures was used to identify amino acid substitutions on protein structures (Eyal, Najmanovich, Sobolev, & Edelman, 2001). Two correlation coefficients were computed from this dataset: the correlation of raw side chain solvent accessibilities between substituted sites, and the correlation of side chain solvent accessibilities relative to their ideal accessibility between substituted sites. Plots of these distributions are displayed in Figure 3.

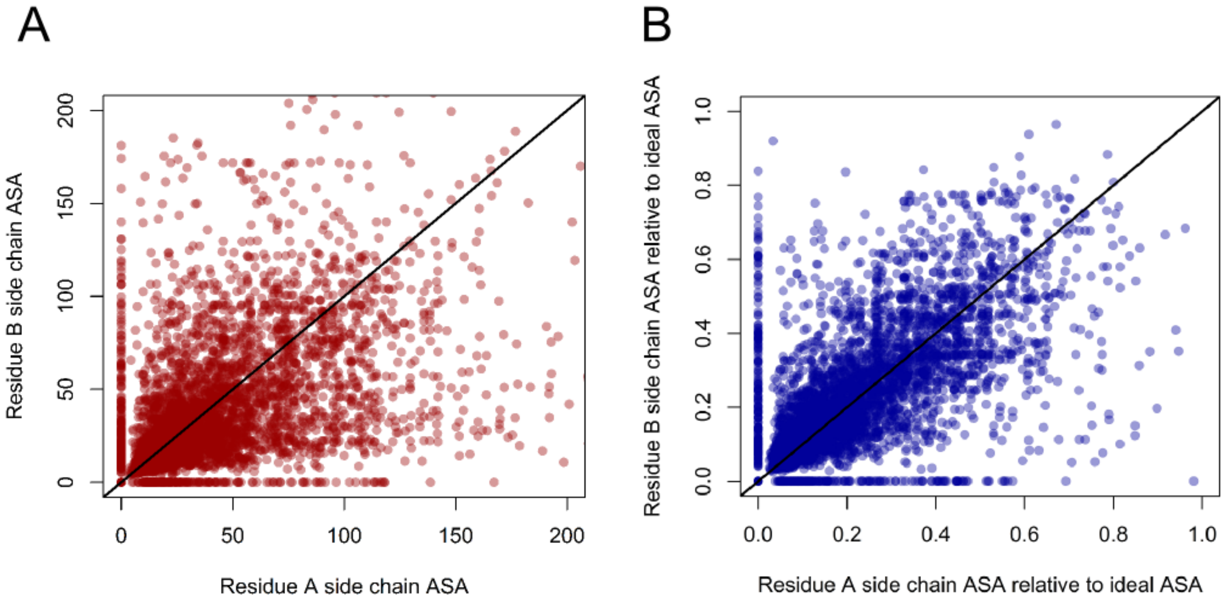


Figure 3: Correlation between sidechain solvent-accessible surface areas for observed pairs of amino acid substitutions. A: Raw solvent accessible surface area; B: Solvent accessible surface area normalized to ideal solvent accessibility for each unique residue.

Pearson correlation coefficients were calculated for both distributions. The correlation coefficient (*r*) for raw solvent accessibility values was 0.520, whereas the coefficient for solvent accessibility values normalized to ideal solvent accessibility was 0.650. This indicates that normalizing solvent accessibilities to a unique maximum value for each amino acid improves the accuracy of assessing solvent accessibility between substituted sites on protein structures.

Euclidean inter-α carbon distance is a measure of the degree to which a set of mutations are spatially clustered (localized together) in 3D space. A group of substituted sites that are aggregated more closely together than that expected by chance is indicative of adaptive or function-altering change compared to substituted sites that occur throughout the protein in an uncoordinated manner (Wagner, 2007; Zhou *et al.*, 2008). Residue α-carbon positions are pre-computed for all PDB entries, and the α-carbon position of a site that has undergone a mutation from the ancestral to derived sequence is directly transferred from that of the of the PDB residue that it aligns to.

In the event that a query protein BLASTs to multiple PDB structures, the information from each structural alignment may be integrated. That is, if structure A provides solvent accessibility or distance information for an N-terminal portion of the protein and structure B provides this information for the C-terminal portion, both data can be combined. This is depicted in Figure 4, which illustrates how a residue-to-residue distance matrix may combine information across multiple PDB templates. In these cases, if there is more than one structure contributing information on a particular site, values are simply taken from the best aligned structure. This is a beneficial methodological feature of Adaptation3D as it captures as much information as possible from multiple PDB templates.

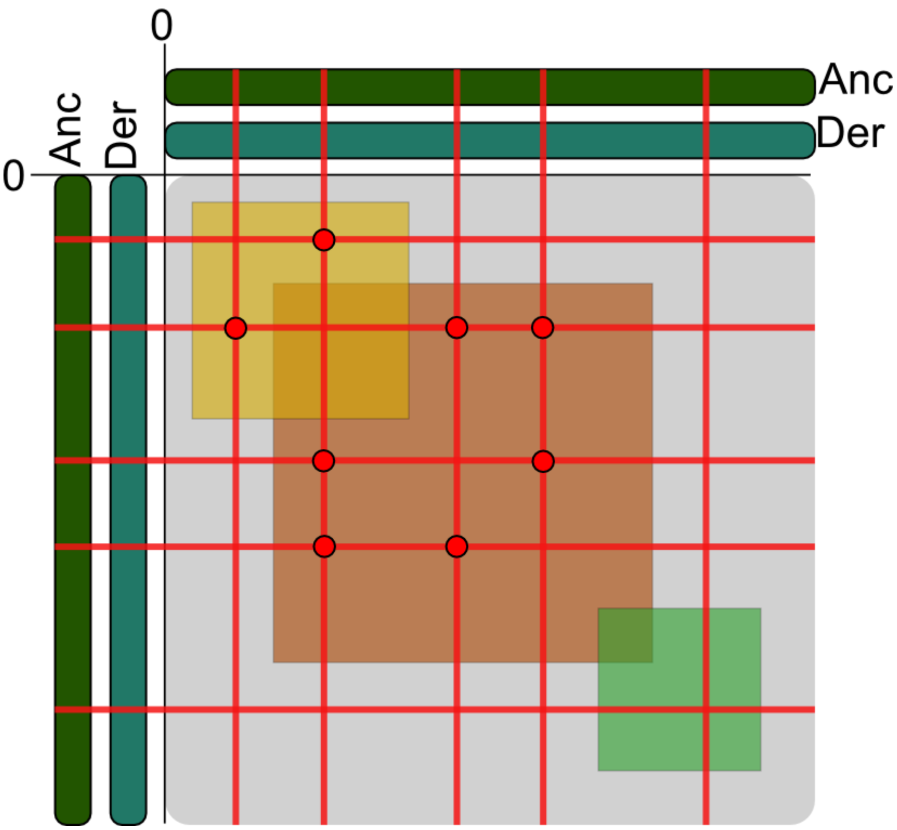


Figure 4: A graphical illustration of Adaptation3D's use of distance information from multiple PDB templates. In this example, substitutions between ancestral and derived sequences (red lines) are mapped to three different structures (represented by the yellow, orange, and green squares). Distances between substituted sites can only be compared when they fall on the same PDB structure.

Scoring values (i.e., average pairwise distance or solvent accessibility) for a group of substitutions is determined by the methods described above. However, measures must be taken to assess the statistical significance of these observed values (Wagner, 2007). For reference, a collection of an aligned ancestral, derived, and template (PDB) sequences that are analyzed as a unit will be referred to as an "ancestral, derived, PDB sequence alignment triad" (ADPST). For a given ADPST, two resampling distributions are generated: one for each criterion. This is done by randomly sampling amino acid groups and calculating their properties for many (10^4) iterations.

The RASA random distribution for an ADPST is built by randomly selecting 10,000 sets of non-overlapping residues on the part of PDB structure aligned to the ancestral and derived sequences, and getting the mean RASA for each set. The distance random distribution for an ADPST is built by randomly selecting 10,000 sets of non-overlapping residues on the part of the PDB structure aligned to the ADSP, and calculating the mean distance between all pairs of residue combinations for each set. For all cases, the number of values selected before averaging is equivalent to the number of observed mutations for that ADPST. While this ensures statistical consistency, it is also important to note that having a greater number of substitutions does increase statistical power. The null distributions are built using randomized data pulled from sampled residues from the PDB structural alignment.

Determination of P-values

With the observed scoring values for the ADPST, and the random distribution background, it is possible to determine the significance of the group of substitutions, reflecting the degree to which the average distance or accessibility deviates from the expected, null distribution. With the observed scoring value, the cumulative density function (CDF) is used on the distribution to see what percentile of records are less than the observed value. The result of the CDF shows the extremity of the observed value on the background distribution, and hence is a *P*-value. If the observed value falls at one tail end of the distribution ($P < 0.05$), then the observed statistic deviates from the random beyond what can be expected due to chance and is a candidate for positive selection (Wagner, 2007). The process of randomly selecting groups of sites to build a distribution of mean distances, and subsequent determination of distance *P*-values (P_{3D}) is represented in Figure 5. This process is also used to determine if the observed group of substitutions is statistically significant in terms of how buried or exposed the residues are (P_{asa}). The P_{asa} statistic can be seen as a two-tailed distribution to find groups of mutations that are either significantly buried, or significantly exposed. An overall flowchart of the entire method for a single MSA file and Newick file is displayed in Figure 6.

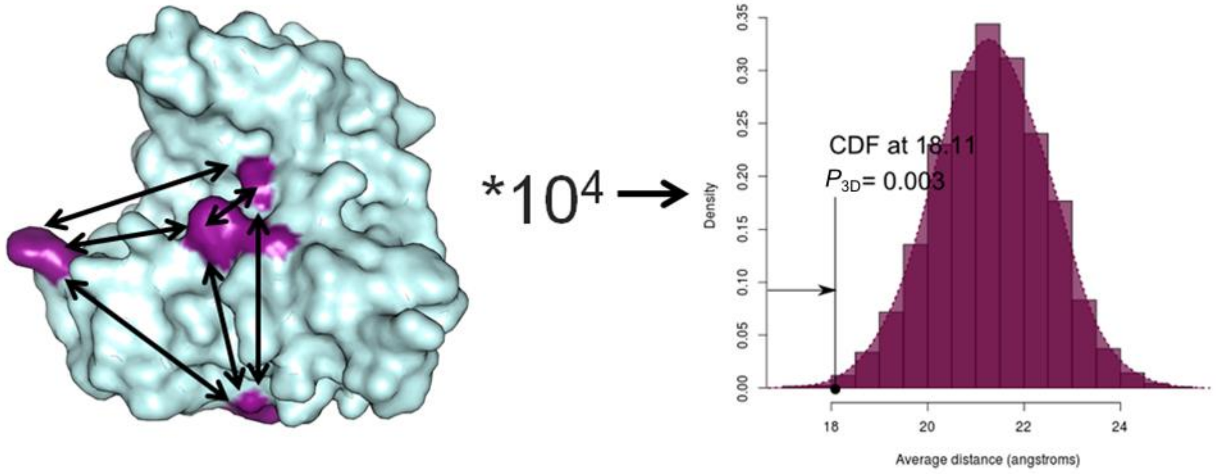


Figure 5: Calculation of P_{3D} values based on resampling. Amino acid sites that fall within the alignment bounds of the ADPST are randomly selected, and the mean Euclidean distance for all pairwise distances is calculated (left). This process is repeated 10,000 times to build a random distribution for the ADPST (right). The CDF function on the random distribution at the observed mean Euclidean distance determines the distance *P*-value (P_{3D}).

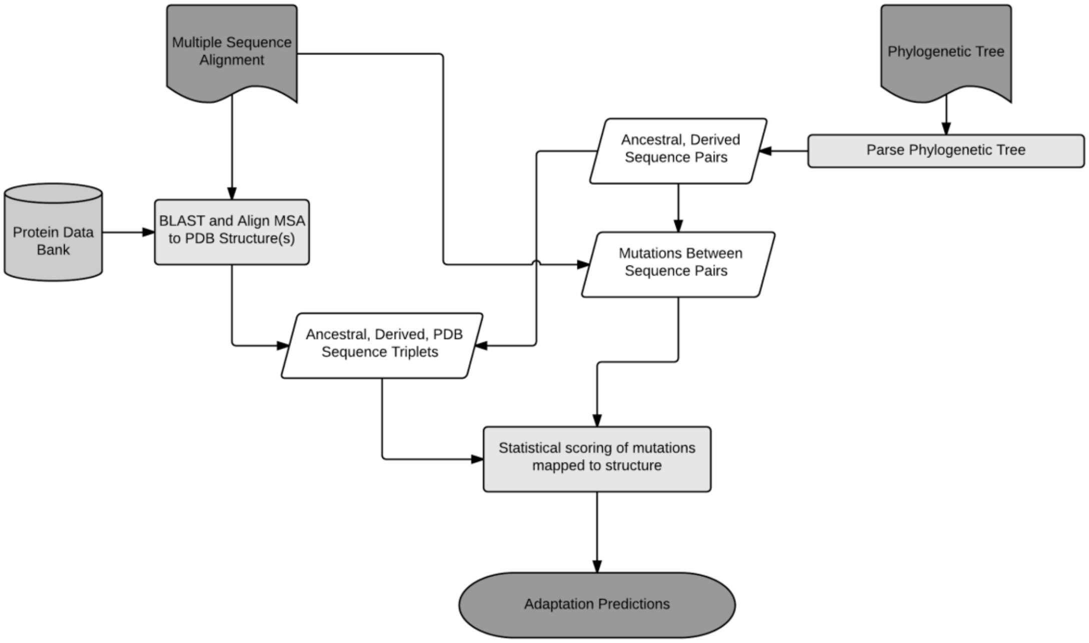

Multiple Sequence Alignment
Phylogenetic Tree
Ancestral, Derived Sequence Pairs
Parse Phylogenetic Tree
Protein Data Bank
BLAST and Align MSA to PDB Structure(s)
Mutations Between Sequence Pairs
Ancestral, Derived, PDB Sequence Triplets
Statistical scoring of mutations mapped to structure
Adaptation Predictions

Figure 6: Flowchart of the Adaptation3D method for a single multiple sequence alignment file and phylogenetic tree file.

2.6 Novelty and distinguishing features of the Adaptation3D method

There are several important considerations and features of Adaptation3D that distinguish it from previous methods that warrant further explanation.

1) Focus on proteins: Adaptation3D does not rely on any DNA sequence information whatsoever, and instead uses amino acid sequence information to infer function modification. By examining and interpreting directly the changes in amino acid sequence, it is possible to obtain direct insight into the molecular determinants of functional modification; whereas analyzing protein function modification through DNA substitution rates can be seen as more indirect or roundabout.

2) Focus on structure: Adaptation3D uses protein tertiary (3D) structural information to make claims about functional modification and adaptation. Since proteins perform their function as folded structures, then a structural perspective on adaptation should theoretically have more power to accurately detect functionally relevant changes. For example, suppose there is a small group of residues within an enzyme that are directly involved in ligand binding or catalysis. Even if that enzyme is shown to possess a high number of substitutions, these substitutions may have occurred in a functionally irrelevant area of the protein. Structure information can therefore be used to inform algorithms about regions that are likely to result in functional change when mutated. Such knowledge may include physicochemical characteristics, pockets and clefts, surface versus interior regions, secondary structural information, and other spatial features.

3) Phylogenetic perspective: Adaptation3D uses both a phylogenetic perspective and ancestral sequence reconstruction to infer specifically the specific evolutionary lineage in which protein adaptation has occurred. In the case of pairwise or site-specific K_a/K_s, it is impossible to determine the direction of adaptation (Hughes, 2007; Hughes & Nei, 1988; Tajima, 1989), and hence which characteristics are derived and which are ancestral. This phylogenetic information is critical for understanding where and when selection has occurred because this may provide clues about *why* selection has occurred biologically.

With ancestral sequence reconstruction, it is theoretically possible to go back up the evolutionary tree and retrace the mutation events that have happened in a phylogeny to lead to the modification of function (Cai, Pei, & Grishin, 2004; Chang, Ugalde, & Matz, 2005). With a given phylogenetic tree and a set of extant sequences, we can hypothesize protein sequences at each point in the tree and find significant patterns of mutation going from more ancestral sequences to more derived sequences. This leads to the pinpointing of lineage-specific functionally significant mutation. This process also makes it more natural to bridge the gap between protein evolution and phenotypic shifts, because specific groups of mutations can be attributed to taxonomic clades and/or gene duplication and divergence events.

Chapter 3

Applications of Adaptation3D to specific protein families and entire proteomes

Adaptation3D detects branch-specific clustered mutation by incorporating sequence, structural, and phylogenetic data. It uses statistical random sampling of mutations mapped to protein tertiary structures in order to detect significant mutational clustering and does so for all branches within a phylogenetic tree. This metric is used as a means to infer functional modification in a protein lineage, because spatially concentrated changes within a particular region may be indicative of positive selection on specific functional sites compared to situations involving spatially scattered patterns of mutation (Wagner, 2007; Zhou, *et al.*, 2008). As a proof of concept, Adaptation3D was applied to a previously studied model of protein family adaptation – that being the PR-5 pathogenesis-related protein family in plants (Doxey *et al.*, 2010), in which key carbohydrate-binding motif adaptations in the Solanaceae-specific (PR-5d) subfamily have been detected and experimentally validated.

3.1 Detecting lineage-specific clustered mutation: Application to the PR-5 protein family

3.1.1 Background on the PR-5 protein family

The *Solanaceae* are a commercially important family of flowering plants (Olmstead & Bohs, 2007). Members of this family include potato, tomato, peppers, eggplant, and tobacco. Many species within this family are targets of a variety of pathogens, particularly oomycete species (Latijnhouwers, de Wit, & Govers, 2003; Woloshuk *et al.*, 1991; Zevenhuizen, & Bartnicki-Garcia, 1969; Zhang, McCarthy, & Smart, 2008). It has been observed that there has been a large radiation of novel genes in these plant species as a result of duplication of osmoregulatory genes including osmotin. These genes are hypothesized to play a large role in pathogen defense processes (Campos *et al.*, 2002; Kuboyama, 1998; Ruiz, Herrera, Ghislain, & Gebhardt, 2005).

One such osmotin-like protein found within the Solanaceae is called pathogenesis-related protein, PR-5d (Koiwa *et al.,* 1997). Expressed in root cells and vascular tissues, PR-5D has been proposed to bind to oomycete cell walls, thereby preventing them from spreading through the plant (Kitajima, Koyama, Yamada, & Sato, 1998; Koyama, Kitajima, & Sato, 2001). However, how PR-5d interacts with the pathogen cell wall was initially unclear.

A previous study demonstrated that tobacco PR-5D contains a unique pattern of three coplanar surface tryptophan residues that are present in other Solanaceae PR-5d proteins but are lacking in other members of the PR-5 superfamily (Doxey et al, 2010). In *Solanum lycopersicum*, these surface tryptophans are found at residue positions 34, 36, and 196. The triple tryptophan motif was computationally predicted as a cellulose-binding motif based on its geometric similarity to binding sites in other cellulose-binding proteins, and this prediction was validated experimentally through a tobacco cellulose pulldown assay which identified PR-5d as a major cellulose-binding protein of tobacco (Doxey et al., 2010).

Interestingly, oomycete cell walls are rich in many different carbohydrate molecules such as cellulose (Latijnhouwers, de Wit & Govers, 2003; Zevenhuizen, & Batnicki-Garcia, 1969), and cellulose is lacking from the cell walls of fungi. Therefore, PR-5d’s unique acquisition of a cellulose-binding site was described as a possible evolutionary adaptation for targeting oomycete pathogens such as *Phytopthora infestans* (Doxey et al., 2010).

Here, I have re-examined the PR-5 pathogenesis-related protein family using Adaptation3D in order to determine whether significant, lineage-specific examples of clustered mutation can be detected within the PR-5 phylogeny. I hypothesize that the triple tryptophan mutation that occurs within the PR-5d subfamily is one such example of positive selection, that it may score significantly using the P_{3D} statistic, and may have an elevated score compared to other PR-5 lineages.

3.1.2 Structural phylogenetic prediction of protein adaptation in the PR-5 family

PR-5 related sequences from *Solanaceae* and non-*Solanaceae* plants were retrieved from the NCBI non-redundant database using blastp with *Solanum lycopersicum* PR-5D as the query sequence (RefSeq Accession NP_001234351.1). Twenty-one sequences were retrieved from various source organisms and databases (Fernandez-Pozo *et al.*, 2015; Goodstein *et al.*, 2012; Kersey *et al.*, 2014; Lamesch *et al.*, 2012; Pruitt *et al.*, 2014; Szklarczyk *et al.*, 2011; Yu *et al.*, 2015). The sequences used to construct the multiple sequence alignment are listed in Table 2. A multiple sequence alignment was produced using MUSCLE in Seaview (Edgar, 2004; Gouy, Guindon, & Gascuel, 2010) (Figure 7). To reconstruct the phylogeny tree for this protein family, the PhyML maximum likelihood algorithm was used as implemented in Seaview (Guindon *et al.*, 2010). This process produced two files: a multiple sequence alignment file and a corresponding phylogenetic tree file in Newick format, which served as input for the Adaptation3D algorithm.

Table 2: Sequences used in phylogenetic reconstruction of PR-5 protein family.

Number	Accession	Species	Source	Description
1	PGSC0003DMP400005490	*Solanum tuberosum*	PTGBase	Thaumatin-like
2	Solyc08g080660	*Solanum lycopersicum*	Ensembl Plants	pathogenesis related protein 5
3	PGSC0003DMP400005491	*Solanum tuberosum*	PTGBase	Thaumatin-like
4	Solyc08g080670	*Solanum lycopersicum*	Ensembl Plants	PR protein 5-like
5	Solyc08g080620	*Solanum lycopersicum*	Sol Genomics Network	osmotin-like, pathogenesis related
6	PGSC0003DMP400005465	*Solanum tuberosum*	PTGBase	Thaumatin-like
7	PGSC0003DMP400005466	*Solanum tuberosum*	PTGBase	Thaumatin-like
8	Solyc08g080640	*Solanum lycopersicum*	Sol Genomics Network	osmotin-like, pathogenesis related
9	PGSC0003DMP400005467	*Solanum tuberosum*	PTGBase	Thaumatin-like
10	Solyc08g080650	*Solanum lycopersicum*	Sol Genomics Network	pathogenesis related protein 23
11	PGSC0003DMP400005463	*Solanum tuberosum*	PTGBase	Thaumatin-like
12	Solyc08g080610	*Solanum lycopersicum*	Sol Genomics Network	osmotin-like, pathogenesis related
13	GSVIVT01019849001	*Vitis vinifera*	Phytozome	thaumatin-like
14	Thhalv10028920m	*Eutrema salsugineum*	Phytozome	thaumatin-like
15	Bra033138	*Brassica rapa*	String DB	osmotin 34
16	Thhalv10028948m	*Eutrema salsugineum*	Phytozome	thaumatin-like
17	scaffold_603571	*Arabidopsis lyrata*	Ensembl Plants	thaumatin-like
18	scaffold_603568	*Arabidopsis lyrata*	Ensembl Plants	thaumatin-like
19	Carubv_10001823m	*Capsella rubella*	Phytozome	thaumatin-like
20	fgenesh2_kg	*Arabidopsis lyrata*	Ensembl Plants	thaumatin-like
21	AT4G11650	*Arabidopsis thaliana*	TAIR	osmotin 34

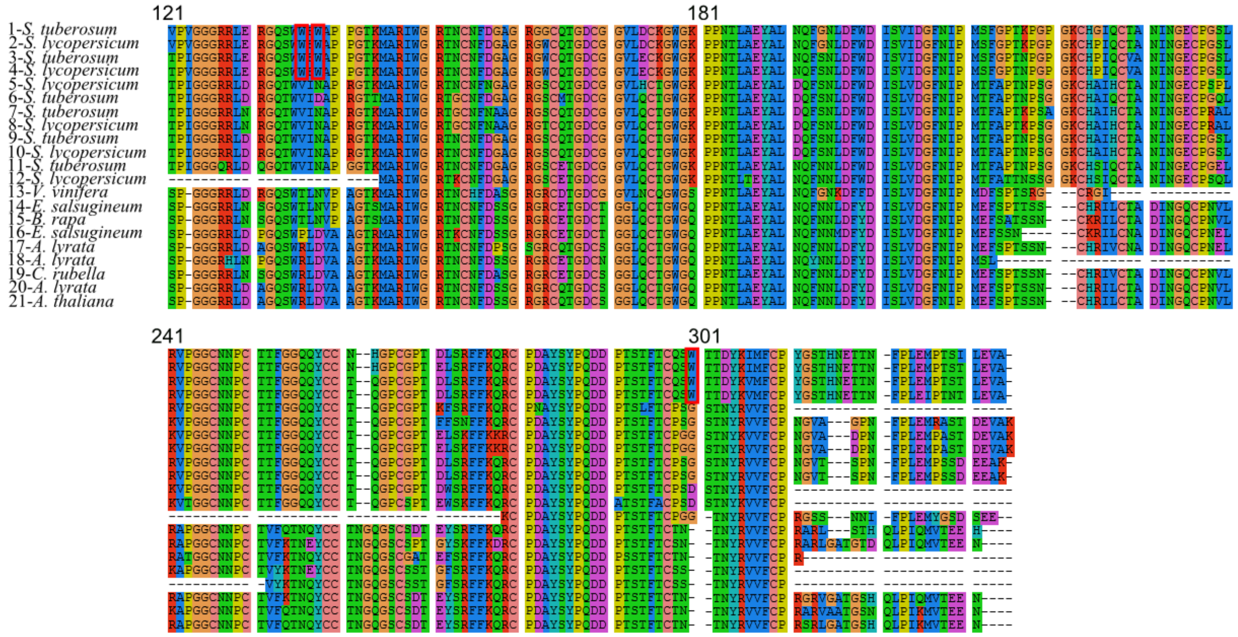


Figure 7: Multiple sequence alignment of thaumatin-like proteins from several species. The putative triple tryptophan (WWW) carbohydrate-binding motif in PR-5d are highlighted in red boxes. The displayed alignment begins at position 121 due to a long gap region present at the beginning of most sequences in the alignment.

With these two files as input, Adaptation3D identified two PDB structures as templates for the PR-5 family: 1AUN chain A, and 2I0W chain A (Ghosh & Chakrabarti, 2008; Koiwa *et al.*, 1999). By mapping the sequence mismatches between ancestral hypothetical reconstructed sequences and their derived neighbours to the PDB structures, Adaptation3D produced a set of distance *P*-values (P_{3D}) for each branch segment in the protein phylogeny as displayed in Figure 8. The PR-5 representative BLASTed to two template structures: 1AUN chain A and 2I0W chain A. The subsequent *P*-values were retrieved from Adaptation3D when 1AUN chain A was used as the template structure.

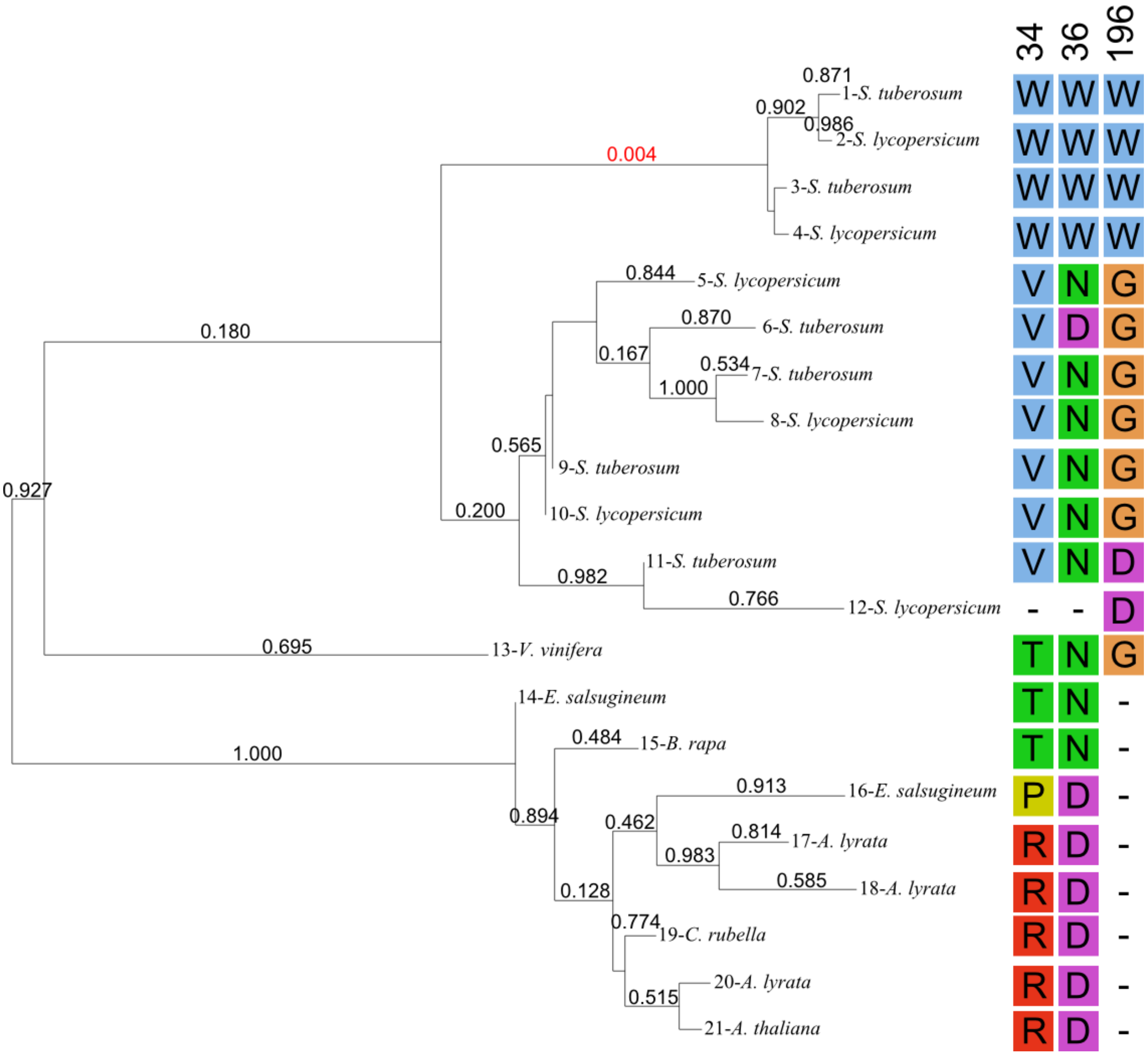


Figure 8: Phylogenetic tree with associated distance *P*-values. Residue states corresponding to the putative carbohydrate-binding motif in pr-5d are displayed to the right of its respective sequence header. *P*-values for each branch segment are based on the sequence alignment to PDB structure 1AUN chain A. Significant *P*-values (< 0.05) are displayed in red.

Not only is there only one P_{3D} value from the Adaptation3D analysis found to be significant (P_{3D} = 0.004, highlighted red) (Figure 8), but this branch segment corresponds exactly to the ancestral lineage containing the *Solanaceae* PR-5D subfamily. However, this P_{3D} value after false discovery rate (FDR) correction is only 0.108 (Benjamini & Hochberg, 1995). The sequence alignment between the ancestral, derived, and PDB 1AUN chain A sequences for this lineage is also displayed in Figure 9B.

As Figure 8 contains the residue states at positions 34, 36, and 196, it is possible to see how the P_{3D} of 0.004 is biologically relevant with respect to the putative origin of the cellulose-binding motif. The residues at positions 34, 36, and 196 are not conserved when viewing the sequences outside of this clade, and in the ancestral reconstructed sequence were V, N, and G, respectively (Figure 9B). However, these residues all mutated to conserved tryptophans for the sequences beneath the derived node. As such, the derived reconstructed sequence had residue states of W, W, and W at positions 34, 36, and 196 respectively (Figure 9B). Thus, this branch segment represents the possible gain of a specialized PR-5 function, which can be detected through branch-specific non-random spatial clustering involving substitutions concentrated to a surface patch on the tertiary structure. These substitutions occurred so close in 3D space that Adaptation3D found them to be significantly clustered together. Therefore, we can reject the null hypothesis that this spatial cluster of substitutions happened by random chance.

The positions of substituted sites in the lineage with the significant distance *P*-value are highlighted in Figure 9A. Figure 9A displays where these mutations mapped to the PDB structure 1AUN. Figures 9Ai and 9Aii display the osmotin binding cleft. Since PR-5D is a paralog of osmotin, PR-5D retains the osmotin binding cleft. The residues of the binding cleft are largely conserved. Figures 9Aiii and 9Aiv display the putative carbohydrate-binding surface, which was achieved by a 90 degree rotation along the X-axis compared to images i and ii.

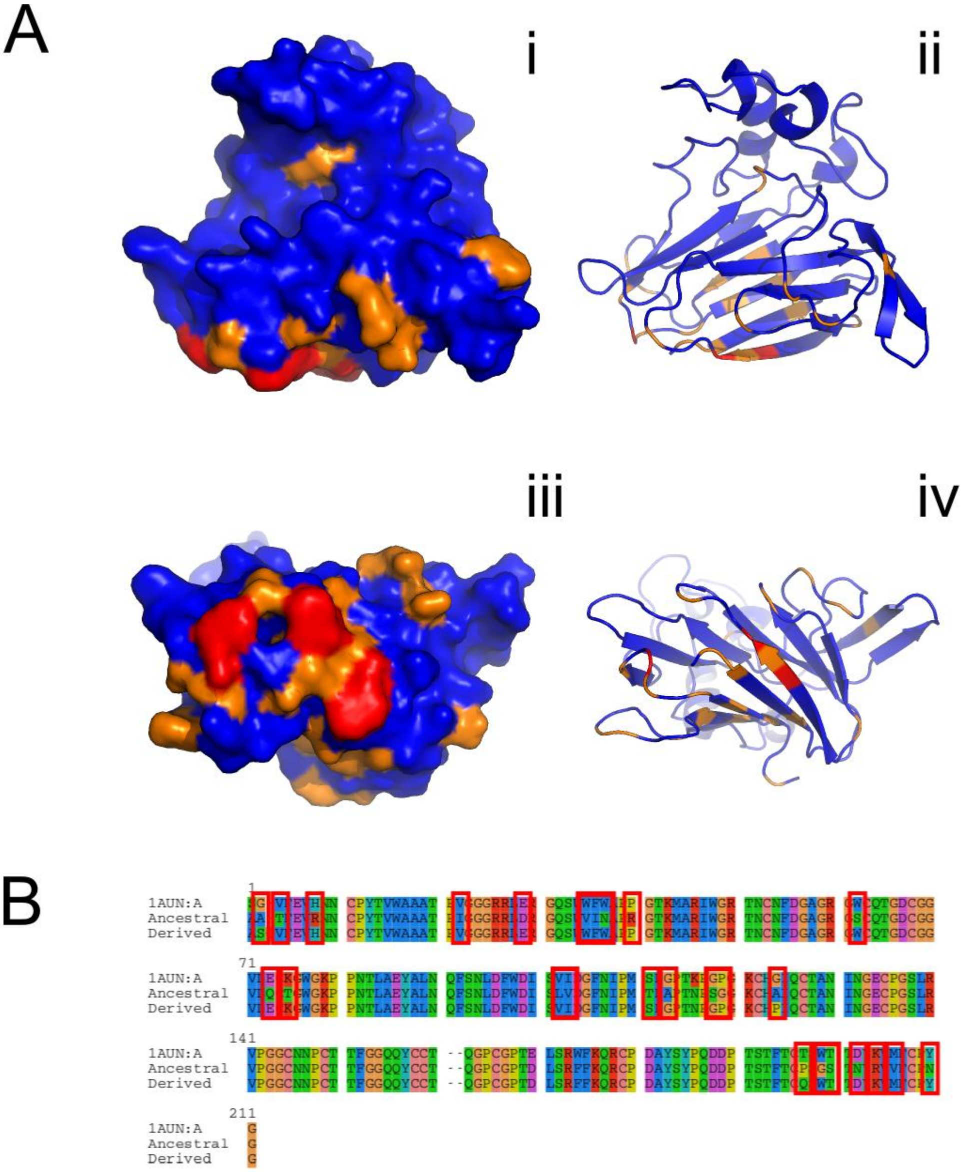


Figure 9: Structure and alignment highlighting substitutions at the branch segment with significant P_{3D}. A: Cartoon and surface structures of pr-5d highlighting mutations from the phylogenetic branch with a significant P_{3D} value. i: space-fill structure of pr-5d; ii: cartoon structure of pr-5d; iii: space-fill structure of pr-5d, carbohydrate-binding patch view; iv: cartoon structure of pr-5d, carbohydrate-binding patch view. Non-mutated residues are labelled blue, mutated residues are labelled orange; mutated residues corresponding to residues 34, 36, and 196 are labelled red. B: Ancestral, derived, PDB sequence alignment between sequences where distance P_{3D} = 0.004. Mismatches between the ancestral and derived sequences are in red boxes.

From Figure 9A, it is evident that very few substitutions mapped close to the osmotin binding cleft of PDB structure 1AUN in the PR-5d branch. On the other hand, most of the occurred in a very small space relative to the overall structure. In addition to the mutation of positions 34, 36, and 196 to tryptophan, many other amino acids underwent mutation in this lineage. However, these mutations were still clustered around the three residues of interest. It is possible that these residue changes also play a role in mediating carbohydrate binding. Thus, by visualizing where the amino acid mutations fall on the tertiary structure, it is evident that the low distance p-value of 0.004 does correspond to highly spatially clustered groups of mutations.

To further inspect if the $\boldsymbol{P}_{\mathbf{3D}}$ value is accurately representing the degree of clustering of a group of mutations, we can compare the mutation cluster from the lineage with a significant $\boldsymbol{P}_{\mathbf{3D}}$ value to the mutation cluster in a lineage with a non-significant $\boldsymbol{P}_{\mathbf{3D}}$ value. We can visually inspect the alignment for the ADPST where the derived sequence is a thaumatin-like paralog from *Vitis vinifera* (accession GSVIVT01019849001), and the lineage-specific $\boldsymbol{P}_{\mathbf{3D}}$ value is 0.695. The alignment between ancestral, derived, and PDB sequences for this branch segment is displayed in Figure 10B. The positions of substituted sites in the lineage with the $\boldsymbol{P}_{\mathbf{3D}}$ value of 0.695 are highlighted in Figure 10A. Figure 10A displays where these mutations mapped to the PDB structure 1AUN. Figures 10Ai and 10Aii display the osmotin binding cleft. Few mutations occurred in the osmotin binding cleft in this lineage as well. The residues of the binding cleft are largely conserved. Figures 10Aiii and 10Aiv display a 90 degree rotation along the Y-axis compared to images 10Ai and 10Aii. No mutations occurred on the carbohydrate-binding motif surface of PR-5D. Instead, they appear randomly distributed throughout the protein structure as predicted by Adaptation3D.

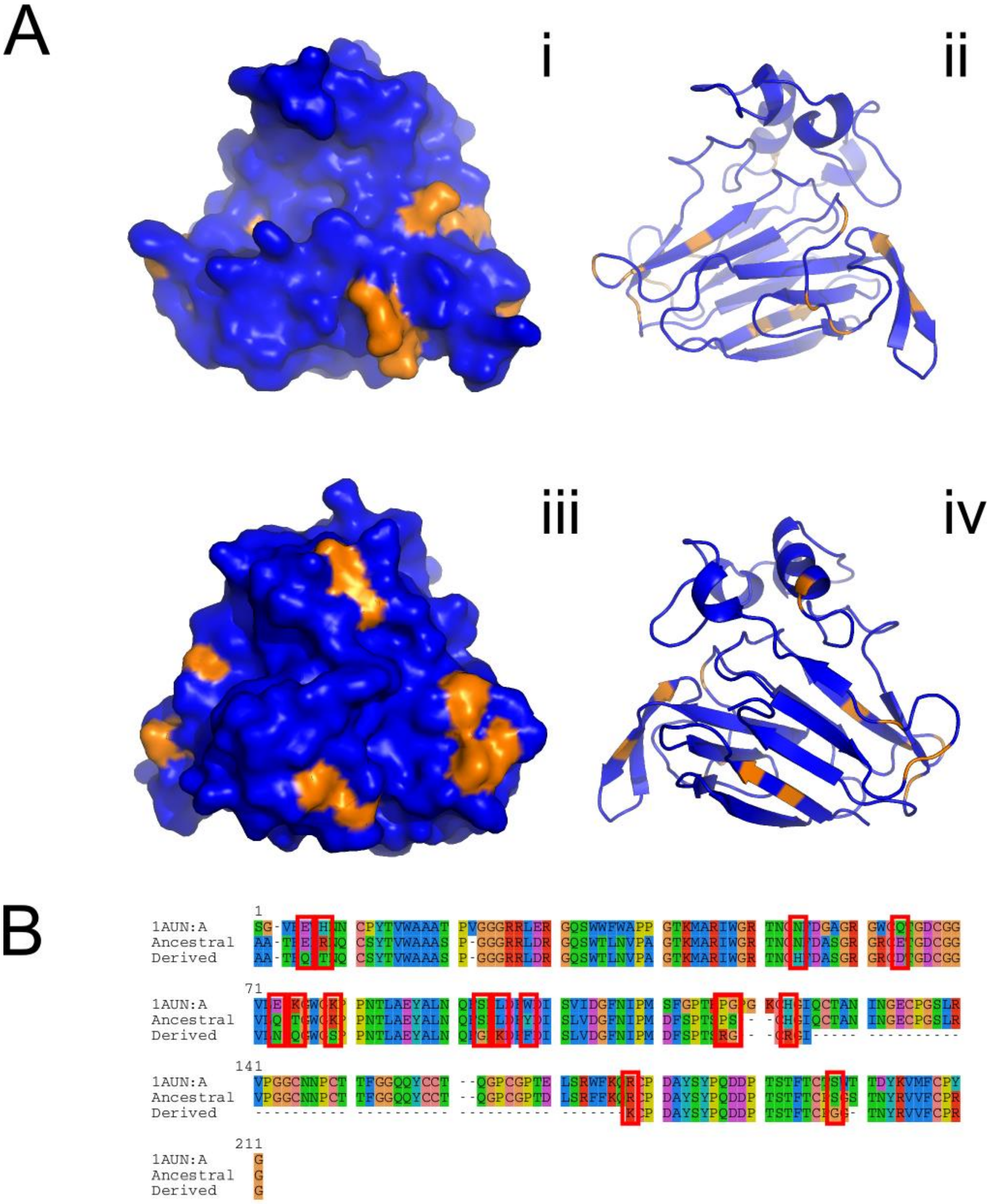

A
i
ii
iii
iv
B
1
1AUN:A SG-VFEVHNN CPYTVWAAAT PVGGGRRLER GQSWWFWAPP GTKMARIWGR TNCNFDGAGR GWCQTGDCGG
Ancestral AA-TFEVRNQ CSYTVWAAAS P-GGGRRLDR GQSWTLNVPA GTKMARIWGR TNCNFDASGR GRCETGDCGG
Derived AA-TFQVTNQ CSYTVWAAAS P-GGGRRLDR GQSWTLNVPA GTKMARIWGR TNCHFDASGR GRCDTGDCGG
71
1AUN:A VLECKGWGKP PNTLAEYALN QFSNLDFWDI SVIDGFNIPM SFGPTKPGPG KCHGIQCTAN INGECPGSLR
Ancestral VLQCTGWGKP PNTLAEYALN QFSNLDFYDI SLVDGFNIPM DFSPTSPS- -CHGIQCTAN INGECPGSLR
Derived VLNCQGWGSP PNTLAEYALN QFGNKDFFDI SLVDGFNIPM DFSPTSRG- -CRGI----- ----------
141
1AUN:A VPGGCNNPCT TFGGQQYCCT --QGPCGPTE LSRWFKQRCP DAYSYPQDDP TSTFTCTSWT TDYKVMFCPY
Ancestral VPGGCNNPCT TFGGQQYCCT --QGPCGPTD LSRFFKQRCP DAYSYPQDDP TSTFTCPSGS TNYRVVFCPR
Derived ---------- ---------- ---------- -------KCP DAYSYPQDDP TSTFTCPGG- TNYRVVFCPR
211
1AUN:A G
Ancestral G
Derived G

Figure 10: Structure and alignment highlighting substitutions at the branch segment with non-significant P_{3D}. A: Cartoon and surface structures of pr-5d highlighting mutations from the phylogenetic branch with a P_{3D} value of 0.695. i: space-fill structure of pr-5d; ii: cartoon structure of pr-5d; iii: space-fill structure of pr-5d, carbohydrate-binding patch view; iv: cartoon structure of pr-5d, carbohydrate-binding patch view. Non-mutated residues are labelled blue, mutated residues are labelled orange. B: Ancestral, derived, PDB sequence alignment between sequences where the derived sequence is thaumatin-like from *V. vinifera*, and the P_{3D} value = 0.695. Mismatches between the ancestral and derived sequences are in red boxes.

The results from the initial Adaptation3D screen on this case of PR-5D motif evolution provide an initial affirmation of the effectiveness of using a general distance *P*-value (**P_{3D})** for the detection of functional innovation in the context of protein structure. The previous study identified the evolution of the coplanar surface tryptophans through a specific screen for that structural motif (Doxey *et al.*, 2010). However, Adaptation3D discovered a likely adaptation event through a much more general structural algorithm, and independently arrives at the same site. This provides even stronger evidence that the triple tryptophan mutation in the PR-5d branch represents a historical adaptation. Furthermore, if the Adaptation3D algorithm is generalized to work with any protein family and alignment, it may be potentially useful as a screening tool to identify lineage-specific positive selection and structural adaptation on a large scale.

Through visual inspection and comparison of Figures 9B and 10B, we can see that mutations in both lineages occur all along the alignment length, and not necessarily in any particular pattern. When we compare the structural mapping of mutations between lineages in Figure 9A and 10A, we can see significant differences in the way the mutations fall onto tertiary structure. The significant lineage (P_{3D}=0.004) contains many substitutions that occur in one region of the protein, and very few mutations occur outside of this region. On the other hand, the non-significant lineage (P_{3D}=0.695) has substitutions that are quite separated from each other spatially. Thus, the P_{3D} distance statistic appears to capture the intended structural phenomenon of clustered mutation. These results also illustrate how mutation and adaptation detection through sequence analysis alone does not provide the full picture. While the substitutions in both lineages occur over the whole length of the primary sequence, this does not accurately reflect what is happening on the tertiary structure. This is further indication for the usefulness of visualizing mutations on folded proteins.

Although these are interesting computational results, these predictions should be followed up with *in vitro* assays to experimentally validate *in silico* results. A potential wet-lab analysis could test for differential carbohydrate molecule binding preferences between paralogs containing the surface tryptophan motif (PR-5D), and paralogs lacking this motif (osmotin). It is possible that the surface tryptophans coordinate with the canonical osmotin binding cleft to bind different carbohydrates.

3.2 Extending Adaptation 3D to phylogenomic scale analysis

3.2.1 High-throughput adaptation screening of the Selectome Database

The results from the previous section demonstrate that Adaptation3D may be used to reconstruct branch-specific adaptations in selected protein families of interest. The next goal is generalizing and extending Adaptation3D to screen databases of protein phylogenies and corresponding multiple alignments. In this way, Adaptation3D could be applied as a high-throughput screening tool to identify specific proteins families, evolutionarily lineages, and sites under positive selection. In addition, this could uncover macro-evolutionary patterns involving recurring proteome-wide adaptation in certain biological processes or functions (i.e., enriched functions). In the following study, I have automated Adaptation3D as a screening tool, have applied it to analyze a large database of vertebrate protein families from the Selectome database (Proux *et al.*, 2009), and have analyzed the results in the context of broader protein structural and functional trends.

The Adaptation3D algorithm was used to detect clustered adaptation on a large scale using pre-computed phylogenetic trees and multiple sequence alignments from the Selectome database (Proux *et al.*, 2009.). Specifically, all branch-specific substitutions from 13,709 protein phylogenies from the Eutelestomi taxonomic cluster of Selectome were mapped to 3D structures, and their significance of spatial clustering (P_{3D}) was measured. All branches (in all trees) were then ranked according to their P_{3D}, followed by solvent accessibility *P*-value (P_{asa}), and lastly by the number of mutations that occurred in the lineage that could be mapped to a specific PDB structure.

In total, 13709 phylogenies from the Selectome database were analyzed. This included a total of 423691 individual lineages (branch segments) that were analyzed. Of the total branch segments that the Adaptation3D method was performed on, 14677 branch segments had a significant $P_{asa} < 0.05$, 46923 branch segments had a significant $P_{asa} > 0.95$, and 50985 branch segments had a significant $P_{3D} < 0.05$.

The solvent accessibility *P*-value (P_{asa}) facilitates identification of branch-specific substitutions that are significantly exposed or significantly buried compared to that expected from random sampling. It is important to define these two sets of behavior, since surface-exposed sites are known to inherently undergo a greater rate of mutation and so sets of buried substitutions may be interpreted as more statistically valid and less likely to occur by chance. Therefore, an additional ranking was performed specifically for cases of lineage-specific clustered adaptation occurring in core regions of the protein (lineages with a $P_{3D} < 0.05$ and $P_{asa} < 0.05$). Multiple hypothesis testing correction using false discovery rate (FDR) was also performed for all *P*-value results (Benjamini, & Hochberg, 1995).

3.2.2 Overview of *P*-value distributions following Adaptation3D analysis of Selectome

The distribution of *P*-values (P_{3D} and P_{asa}) from Adaptation3D's analysis of all protein families and branches in Selectome are displayed in Figure 11. In an analysis of purely random data, these distributions are expected to be flat with an even frequency spread across all *P*-value bins.

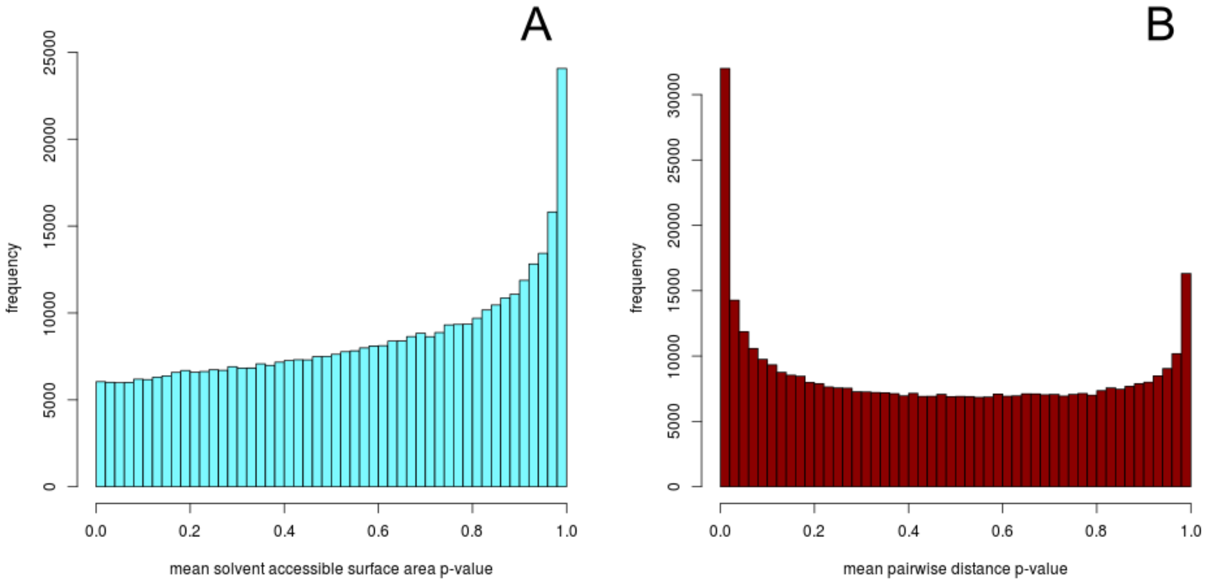


Figure 11: Histograms of *P*-values for large-scale analysis of branches from phylogenies in the Selectome database. A: P_{asa}; B: P_{3D}.

However, this is not the case since both *P*-value frequency distributions are clearly non-random and deviate from a flat distribution (Figure 11). First, there is a clear trend of increasing frequency of high P_{asa} values, which corresponds to highly exposed sets of branch-specific mutations (Figure 11A). On the other hand, Figure 11B, which displays the P_{3D} distribution, has a largely uniform spread of *P*-values with a strong spike of P_{3D} values that approach 0. This signifies an overrepresentation of tightly clustered mutation groups in the Selectome dataset, on top of a background rate of false discovery that can be visualized as the even spread across other *P*-value bins.

This analysis is important since it demonstrates that the test statistics and overall phenomenon of clustered mutation display a non-random distribution. It is therefore reasonable to focus on individual examples, functional summaries and other trends for these predictions in following sections.

3.2.3 Significant cases of detected adaptation in Selectome families

Ranked first by P_{3D} value, Table 3 lists the top 20 most significant cases of detected clustered mutation in Adaptation3D's screen of the Selectome database. The table describes the protein families, specific branches defined by the ancestral and derived lineage taxonomic name, number of mutations detected along that branch, and the raw and FDR-corrected *P*-values for non-random spatial clustering and non-random solvent-accessibility (P_{asa}). The solvent-accessibility related P_{asa} value is a two-tailed *P*-value and thus captures either significantly internal or significantly buried sets of residue substitutions. Regarding branch definitions, it is also important to note that in some cases where there has been a gene duplication event in the tree, the ancestral and derived branch may be equivalent. In the following tables, instances where the ancestral clade name and derived clade name are the same represent gene duplication events, as opposed to speciation events represented by different clade names.

The three protein families containing the most significant degree of spatial mutation anywhere in their tree include: Zinc finger protein 526 (a putative DNA-binding transcription factor), Interleukin 20 receptor beta (a cytokine-mediated signaling receptor), and thioredoxin domain containing 11 (a possible redox regulator involved in thyroid H_20_2 generation). Functional descriptions are based on their UniProt annotations (Bateman *et al.*, 2015). The role of any of these specific proteins in vertebrate evolution is unknown, however the prediction of numerous zinc-finger adaptations (including the top ranked prediction) is intriguing given the considerable literature on zinc-finger expansions and adaptive evolution in vertebrates (Emerson & Thomas, 2009; Schmidt, & Durrett, 2004; Siggers, Reddy, Barron, & Bulyk, 2014). Divergence of zinc-finger binding specificity for instance may play a role in vertebrate developmental evolution.

Predictions further down the list are also of potential evolutionary and functional interest, such as a predicted adaptation within *SLC7A2*, a mammalian-specific paralog that functions as a pregnancy-associated amino acid transporter (Gao *et al*., 2009) and thus may have played a role in the numerous evolutionary innovations associated with mammalian female reproduction.

Table 3: Top 20 Selectome hits for proteins with mutations that clustered together (P_{3D} < 0.05).

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3D}	FDR P_{3D}
Znf526	Euteleostomi	Clupeocephala	2JP9	56	4.24E-2	1.0	1.24E-280	3.75E-276
Il20rb	Euteleostomi	Clupeocephala	4DOH	50	7.71E-2	1.0	3.49E-246	8.21E-242
Txndc11	Euteleostomi	Percomorpha	2B5E	31	6.03E-1	1.0	7.55E-255	1.45E-220
Dhx29	Eukaryota	Eukaryota	3KX2	42	6.38E-1	1.0	4.28E-189	6.97E-185
Abcd1	Clupeocephala	Percomorpha	4F4C	19	7.93E-1	1.0	1.08E-182	1.58E-178
Ttc6	Euteleostomi	Sarcopterygii	1W3B	34	2.83E-1	1.0	1.78E-163	2.22E-159
Kirrel2	Euteleostomi	Tetrapoda	3DMK	39	8.49E-1	1.0	9.8E-161	1.16E-156
Asph	Euteleostomi	Euteleostomi	2NR7	43	2.2E-5	6.8E-1	3.70E-158	4.24E-154
Scrib	Eukaryota	Eukaryota	4MN8	23	6.73E-1	1.0	2.85E-146	2.68E-142
Prdm14	Percomorpha	Tetraodontidae	1MEY	15	2.67E-1	1.0	7.36E-130	5.47E-126
Anapc11	Boreoeutheria	Catarrhini	2MT5	40	3.38E-2	1.0	7.69E-129	5.52E-125
Unknown	Amniota	Testudines	1MU2	40	9.41E-1	1.0	5.36E-121	3.39E-117
Ttc6	Sauria	Phasianidae	1W3B	30	5.71E-2	1.0	3.49E-120	2.17E-116
Lmln	Eukaryota	Eukaryota	1LML	52	2.35E-2	1.0	3.21E-91	1.44E-87
Abcd1	Euteleostomi	Clupeocephala	4F4C	18	9.49E-1	1.0	1.66E-89	7.40E-86
Slc7a2	Mammalia	Mammalia	3GI9	19	1.37E-2	1.0	1.68E-88	7.34E-85
Znf585a	Euteleostomi	Danio	2EE8	61	1.53E-1	1.0	1.45E-86	6.08E-83
Znf526	Clupeocephala	Percomorpha	2JP9	38	6.97E-2	1.0	1.85E-85	7.54E-82
Znf576	Clupeocephala	Holacanthopterygii	2I13	22	9.66E-1	1.0	8.60E-85	3.47E-81
Abcb1	Murinae	Murinae	4KSB	114	1.67E-1	1.0	2.32E-81	8.78E-78

Surface and cartoon structure diagrams highlighting lineage-specific substitutions mapped to structure for statistically significant clustered mutation are displayed in Figure 12. The highlighting of substitutions on these structures show that the sites of substitution do cluster relatively close together given the overall size of the protein structure. Therefore, through visual inspection the P_{3D} statistic can be said to be an accurate metric of capturing clustered mutation on protein tertiary structure.

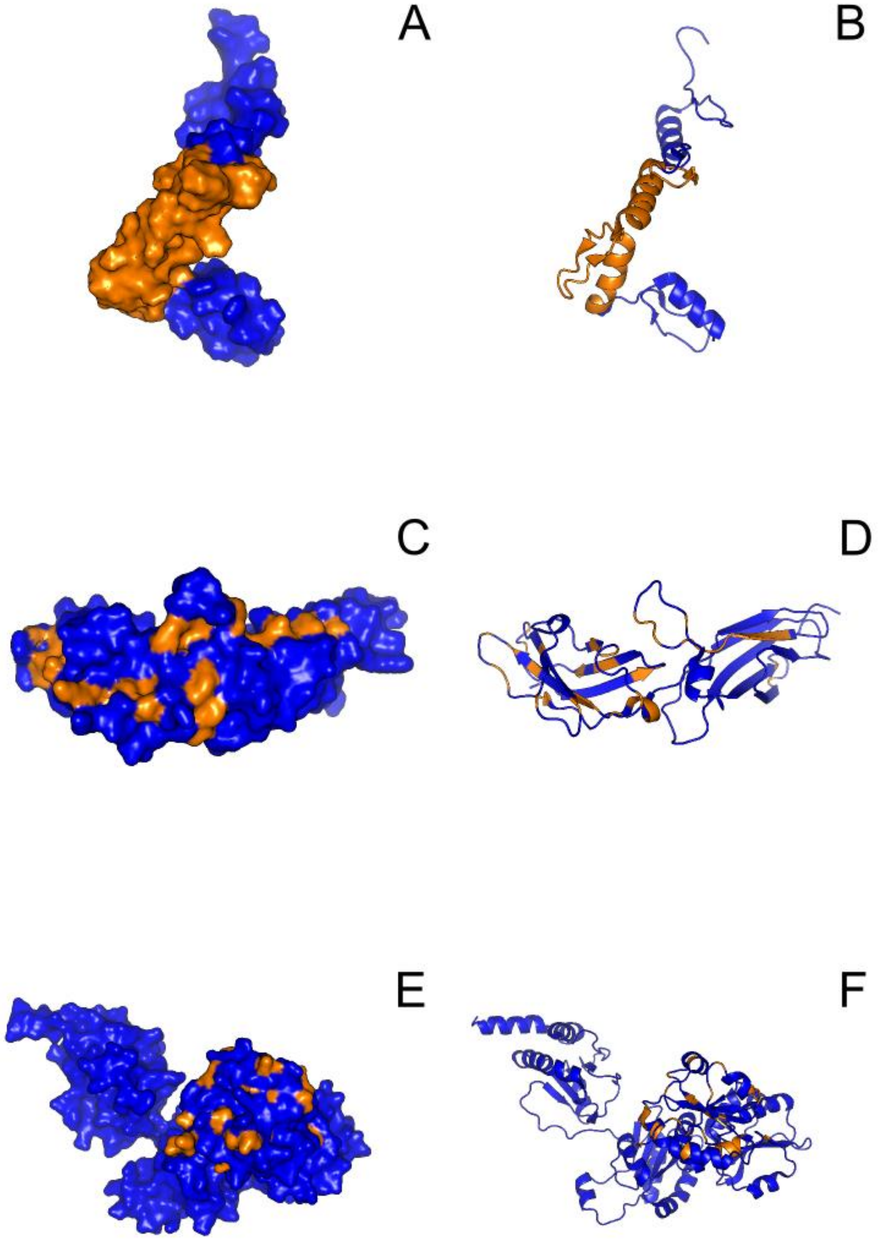


Figure 12: Structural illustrations of proteins displaying clustered mutation ($P_{3D} < 0.05$) in lineages from the Selectome dataset. A: Surface representation of mutations on the ZNF526 protein in the Clupeocephala lineage, (PDB key 2JP9); B: Cartoon representation of the structure from A; C: Surface representation of mutations on the Il20rb protein in the Clupeocephala lineage, (PDB key 4DOH); D: Cartoon representation of the structure from C; E: Surface representation of mutations on the Txndc11 protein in the Percomorpha lineage, (PDB key 2B5E); F: Cartoon representation of the structure from E. Substituted sites in the lineage of interest are highlighted in orange.

3.2.4 Clustered mutation in exposed versus internal regions of protein structures

Visual structural examination of the predictions revealed cases of clustered mutation occurring predominantly in exposed, surface regions, as well as a second class of predictions involving internal, buried mutation clusters. It can be hypothesized that these may represent different evolutionary and functional phenomena.

Lineage-specific clustered adaptation that occurred in exposed regions of the protein (lineages with a $P_{3D} < 0.05$ and $P_{asa} < 0.05$) were tabulated. The top 20 results of clustered mutation in exposed residues are listed in Table 4. Top scoring phylogenies include: anaphase promoting complex subunit 11, RAS oncogene family member 20, scribbled planar cell polarity protein, and intraflagellar transport protein 80.

Table 4: Top 20 Selectome hits for proteins with mutations that clustered together (P_{3D} < 0.05) and were significantly buried (P_{asa} < 0.05).

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3D}	FDR P_{3D}
Znf526	Euteleostomi	Clupeocephala	2JP9	56	4.24E-2	1.0	1.24E-280	3.75E-276
Asph	Euteleostomi	Euteleostomi	2NR7	43	2.2E-5	6.8E-1	3.70E-158	4.24E-154
Anapc11	Boreoeutheria	Catarrhini	2MT5	40	3.38E-2	1.0	7.69E-129	5.52E-125
Ttc6	Sauria	Phasianidae	1W3B	30	5.71E-2	1.0	3.49E-120	2.17E-116
Lmln	Eukaryota	Eukaryota	1LML	52	2.35E-2	1.0	3.21E-91	1.44E-87
Slc7a2	Mammalia	Mammalia	3GI9	19	1.37E-2	1.0	1.68E-88	7.34E-85
Mdga2	Euteleostomi	Tetrapoda	3JXA	26	3.63E-2	1.0	7.12E-63	1.95E-59
Rab20	Euteleostomi	Sarcopterygii	2FG5	27	3.65E-2	1.0	1.67E-59	4.26E-56
Asph	Eukaryota	Euteleostomi	2NR7	47	6.78E-5	1.0	9.90E-55	2.26E-51
Ankrd28	Euteleostomi	Tetrapoda	4OAU	14	1.31E-4	1.0	9.71E-52	2.08E-48
Scrib	Eukaryota	Eukaryota	4LI2	24	1.42E-3	1.0	3.89E-41	6.02E-38
Ift80	Percomorpha	Smegmamorpha	2YMU	23	3.67E-2	1.0	1.07E-40	1.64E-37
Gpsm2	Clupeocephala	Percomorpha	4JHR	78	4.54E-2	1.0	4.06E-29	4.29E-26
Pus3	Amniota	Theria	4NZ6	16	3.76E-3	1.0	1.43E-15	8.24E-13
Psmd4	Euteleostomi	Clupeocephala	1YX4	19	4.53E-3	1.0	1.09E-15	6.33E-13
F1nm06	Testudines	Neognathae	3SQW	41	1.78E-3	1.0	3.45E-16	2.06E-13
Ddx60	Boreoeutheria	Hominoidea	2XGJ	13	1.86E-3	1.0	4.13E-16	2.47E-13
Herc4	Theria	Eutheria	1A12	9	3.35E-2	1.0	2.21E-16	1.33E-13
Amz2	Eukaryota	Eukaryota	2X7M	8	1.48E-3	1.0	7.44E-14	3.92E-11
F6rwe9	Silurana	Silurana	4DJH	7	4.78E-2	1.0	1.44E-13	7.38E-11

Lineage-specific clustered adaptation that occurred in exposed regions of the protein (lineages with a P_{3D} < 0.05 and P_{asa} > 0.95) were tabulated. The top 20 results of clustered mutation in exposed residues are listed in Table 5. Top scoring phylogenies include: ubiquitinconjugating enzyme E2Z, retinoblastoma binding protein 7, enoyl-CoA Delta isomerase 1, and Rho Guanine Nucleotide Exchange Factor 1.

Table 5: Top 20 Selectome hits for proteins with mutations that clustered together ($P_{3D} < 0.05$) and were significantly exposed ($P_{asa} > 0.95$).

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3D}	FDR P_{3D}
Abcd1	Euteleostomi	Clupeocephala	4F4C	18	9.50E-1	1.0	1.66E-89	7.40E-86
Znf576	Clupeocephala	Holacanthopterygii	2I13	22	9.66E-1	1.0	8.60E-85	3.47E-81
Ube2z	Percomorpha	Tetraodontidae	2GRN	22	9.99E-1	1.0	4.70E-76	1.61E-72
Opn4xb	Percomorpha	Tetraodontidae	2KS9	31	9.85E-1	1.0	5.38E-38	7.57E-35
Rbbp7	Haplorrhini	Simiiformes	3CFS	8	9.99E-1	1.0	1.00E-33	1.26E-30
Eci1	Laurasiatheria	Laurasiatheria	1SG4	6	9.67E-1	1.0	2.96E-20	2.15E-17
Ptar1	Clupeocephala	Percomorpha	4EHM	17	9.57E-1	1.0	3.07E-20	2.23E-17
Klf14	Euarchontoglires	Murinae	2JP9	15	9.99E-1	1.0	5.96E-19	4.11E-16
Ptar1	Theria	Metatheria	4EHM	8	9.88E-1	1.0	1.03E-18	6.98E-16
Arhgef1	Euteleostomi	Clupeocephala	2OMJ	8	9.99E-1	1.0	5.98E-14	3.16E-11
Ube2z	Euteleostomi	Sarcopterygii	2GRN	20	9.82E-1	1.0	8.32E-12	3.84E-9
Zmat3	Euteleostomi	Clupeocephala	1ZU1	16	9.94E-1	1.0	3.39E-10	1.37E-7
Atp11b	Sciurognathi	Murinae	3TLM	28	9.53E-1	1.0	3.64E-9	1.33E-6
Phlpp2	Euteleostomi	Tetrapoda	4MN8	60	9.65E-1	1.0	1.07E-8	3.75E-6
Nu4m	Rodentia	Murinae	3RKO	57	9.66E-1	1.0	1.26E-8	4.38E-6
Flot2	Sciurognathi	Murinae	1WIN	12	1.0	1.0	3.01E-8	1.00E-5
Gm16603	Eukaryota	Eukaryota	1YPZ	14	9.80E-1	1.0	4.59E-8	1.50E-5
Kcnq3	Amniota	Amniota	3LUT	4	9.96E-1	1.0	4.87E-7	1.36E-4
Gstp1	Eukaryota	Eukaryota	1GLP	9	9.66E-1	1.0	6.53E-7	1.77E-4
Sall2	Theria	Metatheria	3W5K	5	9.72E-1	1.0	8.33E-7	2.23E-4

Sample structures with highlighted mutations in lineages with significant *P*-values are displayed in Figures 13 and 14. Figure 13 shows surface and cartoon renderings of clustered adaptation in buried protein regions (P_{3D} <0.05 and P_{asa} < 0.05). Figure 14 shows surface and cartoon renderings of clustered adaptation in exposed protein regions (P_{3D} < 0.05 and P_{asa} > 0.95).

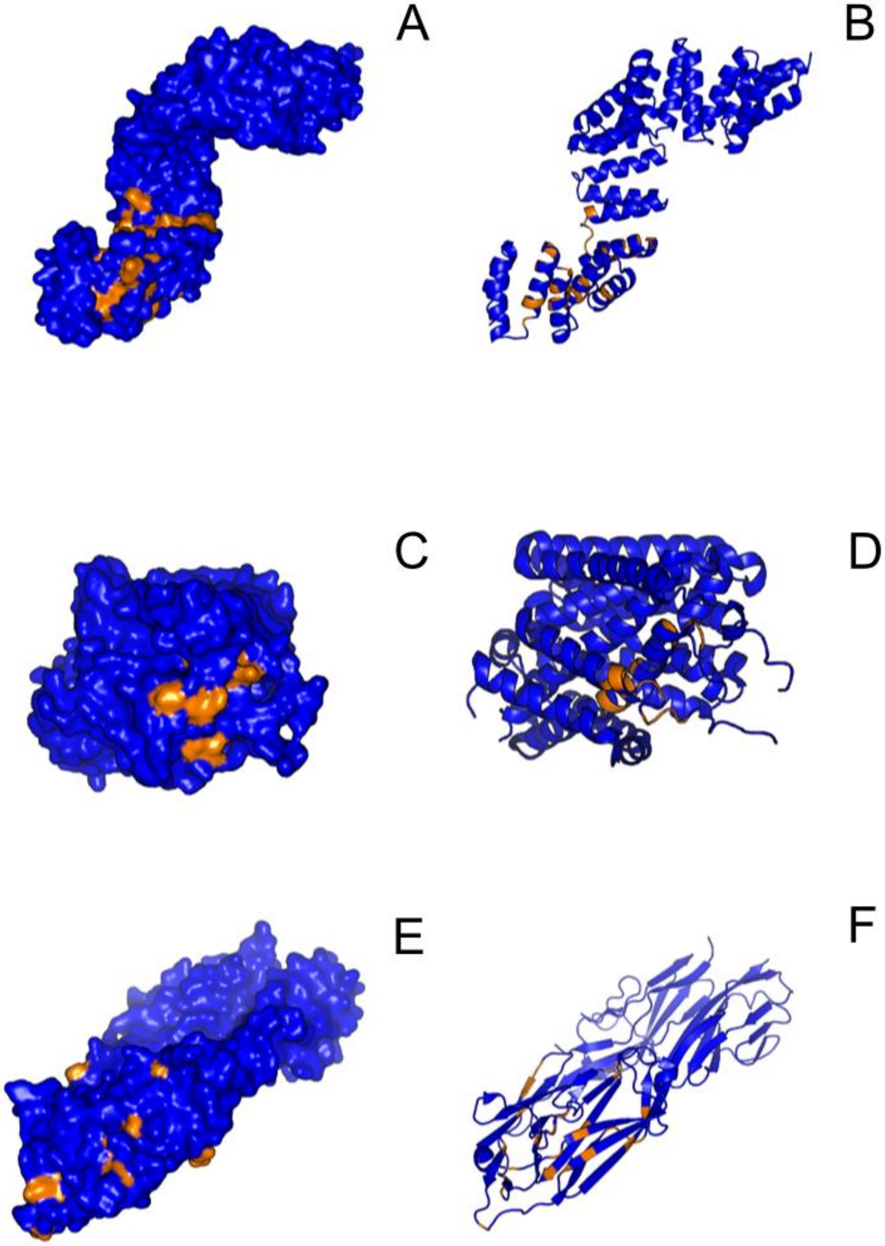


Figure 13: Structural illustrations of proteins displaying clustered mutation in buried regions (P_{3D} < 0.05 and P_{asa} < 0.05) in lineages from the Selectome dataset. A: Surface representation of mutations on the TTC6 protein in the Phasianidae lineage, (PDB key 1W3B); B: Cartoon representation of the structure from A; C: Surface representation of mutations on the SLC7A2 protein in the Mammlia lineage, (PDB key 3GI9); D: Cartoon representation of the structure from C; E: Surface representation of mutations on the MDGA2 protein in the Tetrapoda lineage, (PDB key 3JXA); F: Cartoon representation of the structure from E. Substituted sites in the lineage of interest are highlighted in orange.

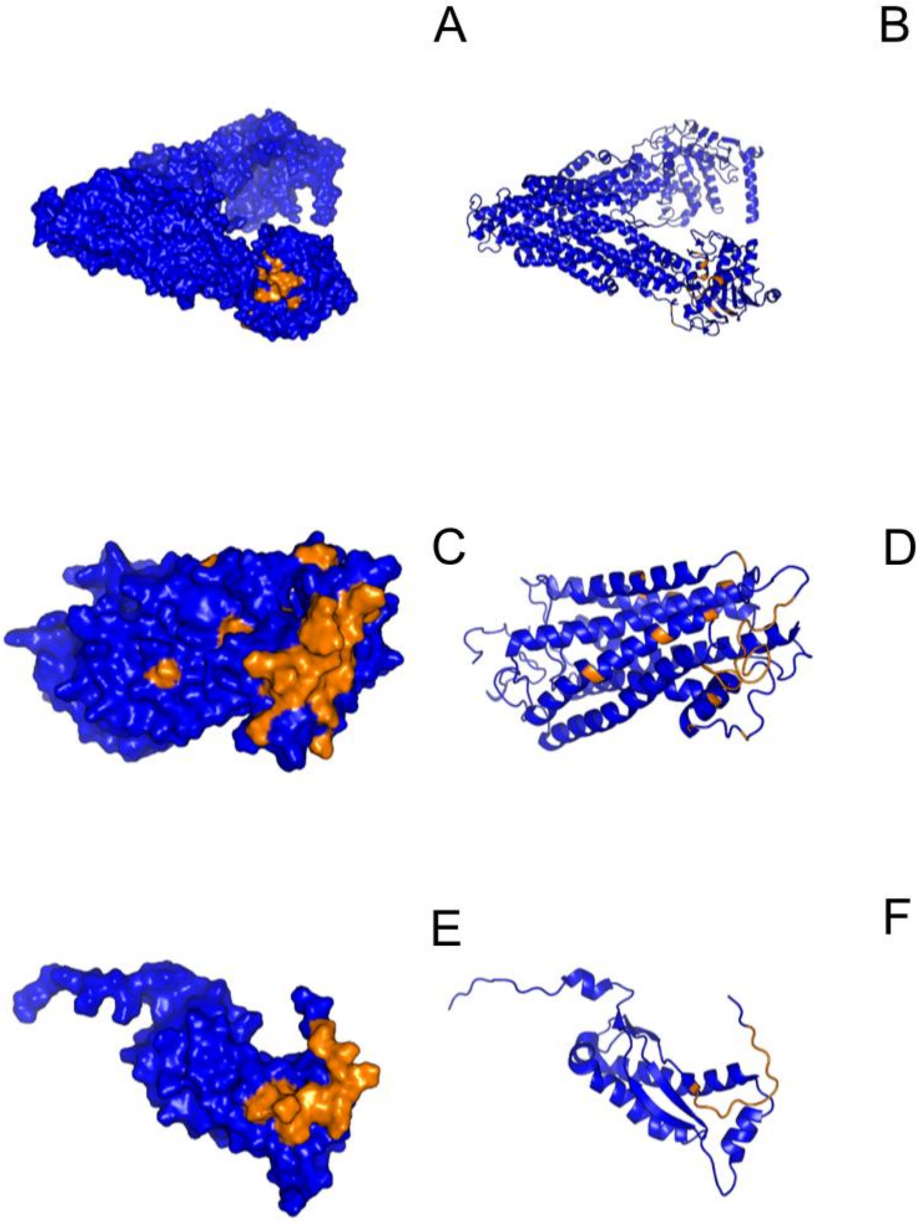


Figure 14: Structural illustrations of proteins displaying clustered mutation in exposed regions ($P_{3D} < 0.05$ and $P_{asa} > 0.95$) in lineages from the Selectome dataset. A: Surface representation of mutations on the ABCD1 protein in the Clupeocephala lineage, (PDB key 4F4C); B: Cartoon representation of the structure from A; C: Surface representation of mutations on the OPN4XB protein in the Tetraodontidae lineage, (PDB key 2KS9); D: Cartoon representation of the structure from C; E: Surface representation of mutations on the FLOT2 protein in the Murinae lineage, (PDB key 1WIN); F: Cartoon representation of the structure from E. Substituted sites in the lineage of interest are highlighted in orange.

Figure 13 and 14 display for sample high scoring (i.e., low P_{3D}) hits from the Selectome dataset. Figure 13 shows Selectome hits with significantly clustered adaptation ($P_{3D} < 0.05$) that occurred in buried protein regions ($P_{asa} < 0.05$). In all sample structures, the mutations are relatively close together in comparison with the overall size of the protein chain. Also, the substituted sites are relatively buried and occur at non-exposed sites. Figure 14 shows Selectome hits with significantly clustered adaptation ($P_{3D} < 0.05$) that occurred in exposed protein regions ($P_{asa} > 0.95$). In these structures, the substituted sites are also clustered closely together relative to the overall protein chain. The mutated sites are at highly surface accessible and exposed sites. Thus, both the P_{3D} and P_{asa} metrics can be said to be an accurate representation of the mutational patterns they are attempting to capture.

3.2.5 Function-enrichment analysis of top-scoring Selectome candidates

Lists containing Ensembl gene ids were used to perform functional enrichment analysis using the David functional annotation web tool (Cunningham *et al.*, 2015; Huang, Sherman, & Lempecki, 2009). Different gene id lists were prepared according to significant hits according the various criteria the Adaptation3D assesses (i.e., clustered mutation, clustered mutation in buried residues, clustered mutation in exposed residues, and exposed/surface mutations). The background list used with all four of these foreground lists contained the gene ids from all protein phylogenies in the Selectome Euteleostomi taxonomic cluster. Table 6 displays significant (Benjamini *P*-value < 0.05) enriched functions for lineages with clustered adaptation (FDR corrected $P_{3D} < 0.05$). Enriched functions include various nucleotide binding related keywords, ATP-binding, and Zinc finger type DNA-binding.

Table 6: Enriched functions for proteins with mutations that clustered together (FDR corrected P_{3D} < 0.05).

Category	Term	Count	*P*-Value	Benjamini *P*-Value
Swiss-Prot	Nucleotide-binding	287	7.9E-27	4.8E-24
Swiss-Prot	atp-binding	235	1.1E-21	3.3E-19
GO Terms	purine ribonucleotide binding	312	3.1E-19	3.6E-16
GO Terms	purine nucleotide binding	321	2.9E-18	1.6E-15
INTERPRO	Serine/threonine protein kinase-related	92	6.9E-16	1.2E-12
GO Terms	adenyl ribonucleotide binding	260	2.5E-15	7.3E-13
GO Terms	ATP binding	258	5.2E-15	1.2E-12
GO Terms	adenyl nucleotide binding	269	1.8E-14	3.0E-12
INTERPRO	Protein kinase, core	104	8.0E-14	5.0E-11
GO Terms	Protein kinase, ATP binding site	99	9.7E-14	4.5E-11
INTERPRO	Zinc finger, C2H2-type/integrase, DNA binding	63	2.6E-10	9.5E-8

Table 7 displays significant enriched functions for lineages with clustered adaptation (FDR corrected P_{3D} < 0.05) in buried regions of the protein (P_{asa} < 0.05). Enriched functions include protein kinase related keywords and nucleotide binding keywords.

Table 7: Enriched functions for proteins with mutations that clustered together (FDR corrected $P_{3D} < 0.05$) and were significantly buried ($P_{asa} < 0.05$).

Category	Term	Count	*P*-Value	Benjamini *P*-Value
Swiss-Prot	nucleotide-binding	59	9.2E-8	2.7E-5
Swiss-Prot	atp-binding	48	2.7E-6	4.0E-4
GO Terms	purine nucleotide binding	63	1.8E-5	3.9E-3
GO Terms	purine ribonucleotide binding	61	1.6E-5	7.0E-3
GO Terms	ribonucleotide binding	61	1.6E-5	7.0E-3
GO Terms	adenyl nucleotide binding	52	2.4E-4	1.3E-2
Swiss-Prot	Serine/threonine-protein kinase	20	1.4E-4	1.4E-2
GO Terms	ATP binding	50	2.0E-4	1.4E-2
GO Terms	adenyl ribonucleotide binding	50	2.3E-4	1.4E-2
GO Terms	nucleoside binding	53	1.9E-4	1.6E-2
GO Terms	purine nucleoside binding	53	1.6E-4	1.7E-2

Table 8 displays significant enriched functions for lineages with clustered adaptation in exposed regions of the protein (FDR corrected $P_{3D} < 0.05$ and $P_{asa} > 0.95$). Enriched functions include nucleotide-binding, and metal-binding.

Table 8: Enriched functions for proteins with mutations that clustered together (FDR corrected P_{3D} < 0.05) and were significantly exposed (P_{asa} > 0.95).

Category	Term	Count	*P*-Value	Benjamini *P*-Value
Swiss-Prot	Nucleotide-binding	52	3.7E-8	1.1E-5
Swiss-Prot	atp-binding	40	1.1E-5	1.6E-3
Swiss-Prot	metal-binding	57	9.4E-5	9.0E-3
GO Terms	ribonucleotide binding	56	2.4E-5	9.8E-3
GO Terms	Purine nucleotide binding	57	4.7E-5	9.8E-3
GO Terms	purine ribonucleotide binding	56	2.4E-5	9.8E-3
GO Terms	nucleotide binding	62	1.4E-4	2.0E-2
GO Terms	purine nucleoside binding	48	3.2E-4	2.2E-2
GO Terms	adenyl ribonucleotide binding	46	2.8E-4	2.3E-2
GO Terms	ATP binding	46	2.5E-4	2.6E-2
GO Terms	adenyl nucleotide binding	47	5.1E-4	2.6E-2
Uniprot	zinc finger region:C2H2-type 1	15	1.3E-4	4.9E-2

Tables 6, 7, and 8 display statistically enriched biological processes and functions for significant mutation clusters compared to a background of all the Selectome hits. These function enrichment tables were produced using the DAVID analysis tool (Huang, Sherman, & Lempecki, 2009). Table 7 records enriched functions for mutation clusters that occurred in buried regions of the protein, while Table 8 records enriched functions for mutation clusters that occurred in exposed regions of the protein. Many of the individual keywords are similar to one another, but keywords pertaining to nucleotide binding and serine/threonine protein kinases come up as being statistically enriched in the foreground dataset. It is possible that nucleotide binding proteins have had to radiate and diversify in function more than most other classes of proteins. Studies have shown that the landscape of transcription factor binding specificity is extremely complex, with each factor typically displaying their own unique binding preferences (Badis *et al.*, 2009; Luscombe, & Thornton, 2002). Rapid diversification has led to a high degree of regulatory complexity through high specificity in DNA subsequence binding preference.

The serine/threonine kinases are another class of proteins that have likely undergone statistically significant clustered adaptation over evolutionary time compared to other protein classes. For example, the Mitogen-activated protein (MAP) kinases JNK and p38 have been shown to have duplicated from an ancient hyperosmolarity pathway protein and developed their own substrate specificity (Caffrey, O'Neill, & Shields, 1999). In another comparative genetics study, kinases of similar function between human and fly, which are lacking in worms, suggests that these kinase families duplicated and acquired their own specific functions following the divergence between nematodes and other metazoans (Manning, Plowman, Hunter, & Sudarsanam, 2002). Thus, kinases have duplicated and diversified to allow for more varied and more complex types and cell signaling. Therefore, the enrichment of DNA-binding related and kinase related keywords in the Selectome clustered mutation dataset as determined by DAVID does have precedent in the data and metadata of the literature. Diversification of function of nucleotide binding and kinase protein classes through historical duplication and neofunctionalization have led to more complex phenotypes through highly specific gene regulation and cell signaling.

Table 9 displays significant enriched functions for lineages with mutations in buried regions of the protein ($P_{\text{asa}} < 0.05$). Benjamini p-values are also included for keywords pertaining to highly polymorphic proteins (i.e., polymorphism, sequence variant). Enriched functions include nucleotide binding, ATP binding, and immunoglobulin function.

Table 9: Enriched functions for proteins with mutations that were significantly buried (P_{asa} < 0.05).

Category	Term	Count	*P*-Value	Benjamini *P*-Value
GO Terms	purine nucleotide binding	519	2.9E-9	5.0E-6
GO Terms	ribonucleotide binding	495	6.9E-6	6.1E-6
GO Terms	Purine nucleoside binding	444	1.6E-8	7.1E-6
GO Terms	ATP binding	416	2.5E-8	8.6E-6
SMART	IGc2	54	5.5E-7	2.3E-4
Swiss-Prot	disulfide bond	448	1.4E-6	2.7E-4
Swiss-Prot	nucleotide-binding	447	3.7E-7	2.8E-4
Swiss-Prot	polymorphism	2425	1.2E-6	3.0E-4
INTERPRO	Leucine-rich repeat	58	2.1E-7	6.4E-4
Uniprot	sequence variant	2506	2.8E-6	3.9E-3
Swiss-Prot	signal	616	2.3E-4	1.7E-2
Swiss-Prot	mitochondrion	233	3.3E-4	2.3E-2

Table 10 displays significant enriched functions for lineages with mutations in exposed regions of the protein (P_{asa} > 0.95). Benjamini p-values are also included for keywords pertaining to highly polymorphic proteins (i.e., polymorphism, sequence variant). Enriched functions include nucleotide binding and immunoglobulin function.

Table 10: Enriched functions for proteins with mutations that were significantly exposed ($\boldsymbol{P_{asa}}$ > 0.95).

Category	Term	Count	*P*-Value	Benjamini *P*-Value
Swiss-Prot	nucleotide-binding	596	1.7E-13	1.5E-10
Uniprot	sequence variant	3400	3.3E-13	2.9E-9
Swiss-Prot	polymorphism	3273	2.3E-11	6.6E-9
GO Terms	ribonucleotide binding	656	5.1E-9	2.1E-6
GO Terms	ATP binding	548	3.2E-8	1.1E-5
Swiss-Prot	disease mutation	486	5.5E-7	6.7E-5
Swiss-Prot	receptor	307	3.6E-6	3.9E-4
Swiss-Prot	oxidoreductase	191	1.2E-5	1.0E-3
Swiss-Prot	ank repeat	96	2.3E-5	1.7E-3
Swiss-Prot	cell membrane	459	4.7E-5	3.1E-3
Swiss-Prot	Immunoglobulin domain	110	9.2E-5	5.6E-3
GO Terms	plasma membrane	883	1.1E-5	8.9E-3

Tables 9 and 10 display statistically enriched biological keywords for mutation groups that were either significantly buried or significantly exposed, respectively. These tables show keywords for mutation results that did not necessarily cluster tightly together (i.e., significant or non-significant P_{3D}). Overall, the biological process keywords are similar to those found in Tables 6, 7, and 8. However, two keywords that appear in Tables 9 and 10 that are absent in Tables 6, 7, and 8 are “polymorphism” (Swiss-Prot keyword), and “sequence variant” (Uniprot keyword). These terms refer to proteins that are hypervariable and tend to mutate rapidly with little to no deleterious effects on function or fitness, and tend to play a role in antibody evasion in antigenic proteins (Johnsson *et al.,* 1998; Lannergard *et al.*, 2011).

3.2.6 Analysis of selected lineages: inferred historical protein adaptations in Eutheria and Amniota

The Adaptation3D algorithm captures mutation events that occur across specific historical lineages. This made it possible to determine the specific taxonomic cluster that the ancestral mutations occurred in according to the Selectome trees and the NCBI taxonomy database (Sayers *et al.,* 2009). This facet of the method opens up another type of analysis that we can perform on the data. It is possible to look at a specific historical lineage and assess the significant protein modifications that have occurred in that lineage. Taxonomic clusters can be categorized by notable phenotypic qualities of the species belonging to that clade. This analysis can be performed to see if there are any significant protein modifications that could correlate with a phenotypic shift we see at that point in the evolutionary tree.

Two sample lineages were selected to see if we can find correlation between protein modification and phenotype in the data. These two lineages selected are represented by the taxonomic keywords Amniota and Eutheria.

Species belonging to the Amniota are characterized by their behavior and ability to lay eggs on land or retain the fertilized egg inside the mother (Benton, 1997; Benton, & Donoghue, 2006). Adaptations that allow for eggs to be laid on land include several additional membranes surrounding the egg. Specifically, reptiles, birds, and mammals comprise the Amniota clade, whereas fish and amphibians are non-amniotes (anamniotes) who lay their eggs in water (Colbert, & Morales, 2001). Another large phenotypic shift in amniotes compared to anamniotes is the change in environment from aquatic to terrestrial.

The Selectome data was also used to find clustered adaptation that occurred in specific taxonomic lineages. Table 11 lists the top 20 ranked clustered mutation hits that occurred in the Amniota lineage.

Table 11: Top 20 Selectome hits for proteins with mutations that clustered together (FDR corrected P_{3D} < 0.05) and occurred in the Amniota lineage.

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3D}	FDR P_{3D}
GALNT13	Amniota	Amniota	2D7I	8	7.57E-1	1.0	2.57E-51	5.34E-48
C1orf216	Tetrapoda	Amniota	4QPL	9	8.98E-1	1.0	8.07E-30	8.81E-27
NWD1	Sarcopterygii	Amniota	2YMU	85	6.27E-1	1.0	8.63E-19	5.89E-16
PDIA5	Euteleostomi	Amniota	3APO	42	2.94E-1	1.0	4.58E-19	3.19E-16
UNC5C	Euteleostomi	Amniota	4V2A	21	6.47E-1	1.0	3.07E-17	1.94E-14
BRIP1	Tetrapoda	Amniota	2VSF	18	7.27E-1	1.0	6.08E-10	2.39E-7
WFIKKN1	Sarcopterygii	Amniota	1BIK	18	6.37E-2	1.0	2.12E-9	7.89E-7
ZNF507	Sarcopterygii	Amniota	2I13	10	4.05E-1	1.0	5.10E-8	1.65E-5
NAP1L4	Sarcopterygii	Amniota	3HFD	21	9.29E-1	1.0	3.33E-7	9.51E-5
KCNQ3	Amniota	Amniota	3LUT	4	9.96E-1	1.0	4.87E-7	1.34E-4
CACNA1C	Tetrapoda	Amniota	4DXW	8	8.52E-1	1.0	8.38E-7	2.24E-4
PPP1R37	Euteleostomi	Amniota	2BNH	17	8.40E-1	1.0	1.22E-6	3.15E-4
MRPS10	Tetrapoda	Amniota	3J6V	20	9.47E-1	1.0	2.44E-6	5.88E-4
RHOU	Euteleostomi	Amniota	2J0V	22	6.31E-1	1.0	3.30E-6	7.74E-4
NFKB1	Sarcopterygii	Amniota	1NFI	24	8.70E-1	1.0	4.13E-6	9.43E-4
UBE2Z	Tetrapoda	Amniota	2GRN	4	3.62E-1	1.0	9.98E-6	2.10E-3
LRRC30	Euteleostomi	Amniota	4MN8	19	5.58E-1	1.0	2.18E-5	4.18E-3
GALNT13	Amniota	Amniota	2FFU	8	5.60E-1	1.0	2.80E-5	5.23E-3
FAM110A	Sarcopterygii	Amniota	4UQW	5	7.78E-1	1.0	3.79E-5	6.74E-3
ADAMTSL1	Tetrapoda	Amniota	3B43	44	4.36E-1	1.0	4.28E-5	7.54E-3

The Adaptation3D results of the Amniota subset of the Selectome data can be parsed through to see if any of these adaptation predictions line up with the changes in phenotype between amniotes and anamniotes. In Table 11, mutation in the UNC5C protein phylogeny has come up as significantly clustered in the Amniota lineage. UNC5C is a member of a family of secreted proteins (netrins) that guide axon extension and migration during development (Leonardo *et al.,* 1997). Studies have found that the development of the diaphragm during embryonic development differentiates between the amniote (specifically mammals and reptiles) and anamniote split (Hirasawa & Kuratani, 2013). The phrenic nerve, which innervates the diaphragm, is guided in part by the UNC5C gene (Burgess, Jucius, & Ackerman, 2006). This phenotypic transition could have been brought about in part by the historical modification of a protein in the UNC5C family.

Species belonging to the Eutheria include placental mammals with several skeletal morphological features (including the absence of epipubic bones) that differentiate them from noneutherians (Reilly & White, 2003; Rook & Hunter, 2014). Eutheria has also often been used to refer to all placental mammals, differentiating them from the metatheria (marsupials) and prototheria (monotremes) (Luo, Yuan, Meng, & Ji, 2011).

Table 12 lists the top 20 ranked clustered mutation hits that occurred in the Eutheria lineage.

Table 12: Top 20 Selectome hits for proteins with mutations that clustered together (FDR corrected $P_{3D} < 0.05$) and occurred in the Eutheria lineage.

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3d}	FDR P_{3d}
SLITRK1	Theria	Eutheria	4LXR	28	8.32E-2	1.0	5.07E-48	9.76E-45
BLVRB	Eutheria	Eutheria	1HDO	14	3.62E-1	1.0	4.31E-45	7.64E-42
MZF1	Theria	Eutheria	2I13	35	5.75E-1	1.0	2.97E-33	3.71E-30
NAV3	Theria	Eutheria	2YRN	7	2.09E-1	1.0	3.54E-33	4.39E-30
UBE2Z	Theria	Eutheria	2GRN	17	8.76E-1	1.0	7.34E-31	8.34E-28
RHOU	Theria	Eutheria	1KZ7	15	2.47E-1	1.0	2.29E-29	2.46E-26
ZNF507	Theria	Eutheria	2I13	9	7.82E-1	1.0	1.66E-28	1.70E-25
CSNK1G3	Amniota	Eutheria	2IZR	6	4.55E-1	1.0	6.11E-17	3.81E-14
AP4S1	Eutheria	Eutheria	2VGL	8	1.29E-1	1.0	8.30E-17	5.13E-14
HERC4	Theria	Eutheria	1A12	9	3.35E-2	1.0	2.21E-16	1.33E-13
ZNF770	Theria	Eutheria	2I13	26	8.74E-1	1.0	2.53E-16	1.52E-13
MCHR1	Theria	Eutheria	4EIY	20	2.35E-1	1.0	3.26E-16	1.96E-13
UBE2Z	Eutheria	Eutheria	2GRN	5	3.39E-1	1.0	2.53E-15	1.44E-12
ZNF689	Eutheria	Eutheria	2I13	23	4.55E-1	1.0	2.02E-14	1.10E-11
RHOU	Theria	Eutheria	2J0V	15	1.77E-1	1.0	2.37E-13	1.20E-10
ABCC8	Theria	Eutheria	3QF4	33	9.31E-1	1.0	6.08E-11	2.63E-8
EFTUD1	Theria	Eutheria	1U2R	17	8.18E-1	1.0	2.27E-9	8.41E-7
CSNK1G3	Amniota	Eutheria	4HGL	5	3.28E-1	1.0	6.51E-9	2.34E-6
ASMTL	Theria	Eutheria	2P5X	16	4.05E-1	1.0	8.89E-9	3.15E-6
BRSK1	Amniota	Eutheria	1ZMU	19	9.11E-1	1.0	2.82E-8	9.42E-6

Functional annotation enrichment analysis was also performed to determine if there are characteristic phenotype or functional shifts that have occurred at specific points in evolutionary history. Table 13 lists enriched functions from clustered adaptation that occurred in the Amniota lineage. Enriched functions include immunoglobulin related keywords. Table 14 lists biological phenotypes that are represented by clustered adaptation that occurred in the Amniota lineage, even though they may not be statistically enriched in the results. Biological processes that are represented by clustered adaptation in proteins in the Amniota lineage include terms pertaining to keratinocyte differentiation, brain development, and body axis determination.

Table 13: Enriched functions for proteins with mutations that were significantly clustered together ($P_{3D} < 0.05$) and occurred in the Amniota lineage.

Category	Term	Count	*P*-Value	Benjamini *P*-Value
INTERPRO	Immunoglobulin I-set	17	9.9E-6	8.1E-3
KEGG	ErbB signaling pathway	9	1.1E-4	1.6E-2
INTERPRO	Immunoglobulin subtype 2	19	6.1E-5	1.7E-2
KEGG	Pathways in cancer	19	4.2E-4	3.1E-2
INTERPRO	Immunoglobulin-like	23	2.7E-4	4.3E-2
Swiss-Prot	leucine-rich repeat	24	1.2E-4	4.8E-2

Table 14: Biological processes/phenotypes that are represented by clustered mutation (P_{3D} < 0.05) and occurred specifically in the Amniota lineage. Biological processes are not necessarily statistically enriched.

Category	Term	Count	*P*-Value	Benjamini *P*-Value
BIOCARTA	Keratinocyte differentiation	6	6.3E-3	5.6E-1
GO Terms	cell cortex part	5	7.3E-2	9.2E-1
BIOCARTA	TNF/Stress Related Signaling	4	8.3E-2	9.8E-1
GO Terms	regulation of protein polymerization	4	5.1E-2	1.0E0
GO Terms	forebrain development	8	6.4E-2	1.0E0
GO Terms	anterior/posterior pattern formation	6	3.8E-2	1.0E0
GO Terms	positive regulation of cell differentiation	10	6.7E-2	1.0E0
GO Terms	ectoderm development	9	3.7E-2	1.0E0
GO Terms	cell morphgenesis involved in neuron differentiation	12	3.5E-2	1.0E0
GO Terms	axonogenesis	12	1.6E-2	1.0E0

If we look at the results in Table 14, we can see that there are several terms that have come up pertaining to ectoderm differentiation, neuron formation, and body patterning and morphogenesis. Specific genes that contribute to these keywords include secreted frizzled related protein 1 (SFRP1). SFRP1 has been found to be present in mesenchymal stem cells from human amniotic fluid and is part of the pathway to contribute to neurogenic cell lineages (Savickiene *et al.,* 2015). Again, this function may have been acquired in part through the clustered modification that occurred in the historical Amniota lineage.

Another interesting term that appears in Table 14 is keratinocyte differentiation. One of the proteins that was found to have clustered adaptation is the epidermal growth factor (beta-urogastrone). Epidermal growth factor has been found to be necessary for the promotion of the extraembryonic membranes in amniote embryonic development (Albergotti, Hamlin, McCoy, & Guillette, 2009; Cross *et al.,* 2003; Jojovic, Wolf, & Mangold, 1998). It is possible that the role of epidermal growth factor for extraembryonic tissue development came about through a historical adaptation in the protein phylogeny.

Table 15 lists enriched functions from clustered adaptation that occurred in the Eutheria lineage. Enriched functions include immunoglobulin related keywords. Table 16 lists biological phenotypes that are represented by clustered adaptation that occurred in the Eutheria lineage, even though they may not be statistically enriched in the results. Biological processes that are represented by clustered adaptation in proteins in the Eutheria lineage include terms largely pertaining to organ development.

Table 15: Enriched functions for proteins with mutations that were significantly clustered together (P_{3D} < 0.05) and occurred in the Eutheria lineage.

Category	Term	Count	*P*-Value	Benjamini *P*-Value
INTERPRO	Immunoglobulin subtype 2	26	1.1E-6	1.1E-3
SMART	IGc2	26	1.4E-5	3.1E-3
INTERPRO	Immunoglobulin	19	3.1E-5	1.6E-2
GO Terms	plasma membrane	169	8.4E-5	3.3E-2
INTERPRO	Leucine-rich repeat	24	1.6E-4	3.4E-2
INTERPRO	Immunoglobulin I-set	18	1.3E-4	3.4E-2
INTERPRO	Leucine-rich repeat, typical subtype	15	1.3E-4	4.4E-2

Table 16: Biological processes/phenotypes that are represented by clustered mutation (P_{3D} < 0.05) and occurred specifically in the Eutheria lineage. Biological processes are not necessarily statistically enriched.

Category	Term	Count	*P*-Value	Benjamini *P*-Value
GO Terms	urogenital system development	10	1.1E-2	6.6E-1
GO Terms	activation of immune response	9	3.4E-2	8.6E-1
GO Terms	kidney development	8	4.0E-2	8.8E-1
GO Terms	positive regulation of muscle cell differentiation	4	6.0E-2	9.3E-1
GO Terms	hindbrain development	5	5.8E-2	9.4E-1
GO Terms	tube development	13	6.6E-2	9.4E-1
GO Terms	feeding behaviour	7	6.4E-2	9.4E-1
GO Terms	heart development	12	7.9E-2	9.5E-1

Table 16 displays several keywords that pertain to biological processes or phenotypic characteristics. Although none of these terms are statistically overrepresented, some of the individual genes that contribute to these keywords may be of relevance to the development of key phenotypes in the Eutheria lineage. Some of the proteins that contribute to “urogenital system development” and “kidney development” include integrin-linked kinase, inversin, and the potassium inwardly-rectifying channel protein. Mutations in these genes leads to nephronophthisis, left-right axis determination abnormalities of renal and urinary system development, renal agenesis, and uterine dysfunction (Lange, *et al.*, 2009; McCloskey *et al.*, 2014; Otto *et al.*, 2003). This demonstrates their importance in the development of the mammalian urogenital system. It is possible that gene duplication and diversification through clustered adaptation of these developmentally important genes has led to differential development of the urogenital system between placentals (which possess different sinuses for urinary and reproductive functions), and monotremes (which possess only one sinus for both functions) (Kobayashi & Behringer, 2003).

3.2.7 Analysis of clustered mutation in the Selectome database: A Summary

Applying the Adaptation3D algorithm to a large dataset like Selectome can provide insights into the nature of how clustered mutation can lead to changes in organismal phenotype and biological processes. The Adaptation3D algorithm can tell us not only where in the protein clustered mutation has occurred, but, with information from the NCBI taxonomy database, it is also possible to dissect what taxonomic lineage and time period the adaptation event has occurred (Sayers *et al.*, 2009). Since the algorithm has been used to make many predictions about adaptation, it is interesting to see if any of these predictions can be found to correlate with the scientific literature on protein functional modification.

It is important for the Adaptation3D algorithm to not highlight hypervariable regions of proteins, because that would indicate that the algorithm highlights mutations arising from relaxed purifying selection as opposed to adaptive mutation. Predictions from Adaptation3D on hypervariable proteins could thus be considered false positive predictions. The use of spatially clustered mutation to determine adaptation on its own is enough to weed out hypervariable proteins (as indicated by the lack of “polymorphism” and “sequence variant” as enriched identifiers in Table 6). However, combining multiple statistics (such as distance and solvent accessibility) may prove even more useful to eliminate false positives arising from protein region hyper variability.

From the analysis of the Selectome dataset, we can see that the Adaptation3D algorithm makes interesting predictions about protein adaptation that have a precedent in the scientific literature. As such, the algorithm may eventually be a useful tool in data-driven exploration of historical protein evolutionary events. The overall results, and the lack of significant biological process keywords from the DAVID enrichment analysis, seem to indicate that only few genes involved in a biological process need to change to cause great effects in phenotype. Thus, it does not appear that statistical overrepresentation of mutations pertaining to a certain biological process is a necessity for phenotypic change, but rather that marked changes in one or a few key proteins can be enough to cause large phenotypic shifts.

3.3 Targeted identification of clustered mutation of transcription factors

3.3.1 General structural and function-enrichment analysis

Both the individual families showing the strongest extent of mutational clustering and the functions most enriched among all predicted adaptations, pointed to adaptation of DNA-binding and transcription factor families (Tables 3, 6, 7, and 8). This is interesting from the perspective of detecting phenotypic adaptation, because even slight changes in the DNA-binding specificity of a transcription factor could have drastic phenotypic effects by altering its downstream regulatory landscape (Lynch, & Wagner, 2008; Mukherjee & Burglin, 2007; Wang, & Zhang, 2007). In this section, Adaptation3D has therefore been focused toward protein families classified as DNA-binding transcription factors from the Selectome database. Transcription factor phylogenies were selected out from Selectome by finding all the phylogenies that BLASTed to a PDB structure entries within the "Biological Interaction database for Protein-nucleic Acid" (BIPA) database (Lee & Blundell, 2009; Worth *et al.,* 2007).

Predicted cases of branch-specific clustered adaptation (lineages with a FDR corrected $P_{3D} < 0.05$) for DNA-binding transcription factor families were tabulated. The top 20 results of clustered mutation in transcription factor phylogenies are listed in Table 17. Also reported is the degree of mutation enrichment within the DNA-binding site (*R*, see Equation 1*)*, which is described in detail later. Interestingly, the overwhelming majority of sequences, including the pluripotency regulator Prdm14 and the Hic1 (Hypermethylated in Cancer 1), are zinc-finger DNA-binding transcription factors.

Table 17: Top 20 hits of clustered adaptation (FDR corrected $P_{3D} < 0.05$) in DNA-binding domains of transcription factors.

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3d}	FDR P_{3d}	R
Prdm14	Percomorpha	Tetraodontidae	1MEY	15	2.67E-1	1.0	7.36E-130	5.48E-126	7.04E-1
Znf576	Clupeocephala	Holacanthoptherygii	2I13	22	9.66E-1	1.0	8.60E-85	3.47E-81	2.64E-1
CU 655961.6	Danio	Danio	2I13	38	3.92E-1	1.0	1.16E-56	2.81E-53	2.49
Znf507	Theria	Metatheria	2I13	16	7.37E-1	1.0	6.84E-45	1.21E-41	2.34E-1
Hic1	Euteleostomi	Clupeocephala	2I13	13	9.37E-1	1.0	7.19E-42	1.14E-38	2.49E-1
Znf770	Amniota	Neognathae	2I13	13	4.50E-1	1.0	2.24E-40	3.38E-37	8.23E-1
Znf507	Amniota	Mammalia	2I13	8	9.10E-1	1.0	4.41E-38	6.25E-35	2.25E-1
Znf775	Boreoeutheria	Laurasiatheria	2I13	8	8.07E-1	1.0	1.49E-34	1.94E-31	7.77E-1
Znf507	Clupeocephala	Tetraodontidae	2I13	19	7.34E-1	1.0	6.60E-34	8.45E-31	3.25E-1
Mzf1	Theria	Eutheria	2I13	35	5.75E-1	1.0	2.97E-33	3.72E-30	5.65
Znf576	Holacanthopterygii	Holacanthopterygii	2I13	14	9.12E-1	1.0	4.42E-29	4.66E-26	7.60E-2
Znf507	Theria	Eutheria	2I13	9	7.82E-1	1.0	1.66E-28	1.70E-25	1.95E-1
Znf507	Euteleostomi	Clupeocephala	2I13	26	4.84E-1	1.0	3.10E-26	2.94E-23	2.62E-1
Klf14	Euarchontoglires	Murinae	2I13	16	8.03E-1	1.0	5.37E-25	4.92E-22	2.92E-1
ENSMODP00000029716	Didelphimorphia	Didelphimorphia	2I13	14	6.45E-1	1.0	5.87E-24	5.13E-21	6.72
Znf576	Holacanthopterygii	Percomorpha	2I13	88	6.94E-1	1.0	9.21E-24	7.96E-21	3.28E-1
Znf251	Danio	Danio	2I13	15	6.86E-1	1.0	1.06E-23	9.13E-21	6.07
Znf775	Euarchontoglires	Murinae	2I13	31	1.18E-1	1.0	1.77E-19	1.27E-16	1.36
Klf14	Euarchontoglires	Murinae	1TF6	12	8.61E-1	1.0	4.52E-18	2.96E-15	2.24E-1
Znf770	Theria	Eutheria	2I13	26	8.74E-1	1.0	2.53E-16	1.52E-13	1.48

The overrepresentation of zinc fingers and even particular zinc finger proteins in Table 17 is interesting. The protein encoded by *ZNF507* for instance, a gene implicated in neurodevelopmental disorders (Talkowski *et al.*, 2012), is listed five times in this table alone, which indicates significant spatial clustering of branch-specific mutations in five different lineages. Interestingly, this detected phenomenon is highly consistent with previous literature, which has identified the *ZNF507* protein (Zfp507) as having undergone a complex history of divergence in some evolutionary lineages and strong conservation in others due to frequent deletions and missense mutations involving a selective set of positions (Liu *et al.*, 2014). Liu et al. (2014) also noted that most of the differences in ZNF507 between species involve in-phase insertion or deletions of ZNF motifs, which may in this case be the underlying mechanism for generating the apparent clustered mutation. Figure 15 includes a visual depiction of clustered mutation detected in a ZNF507 lineage as well as two other zinc-finger proteins from Table 17. Each example illustrates highly localized patches of mutation, indicating positive selection on distinct structural regions.

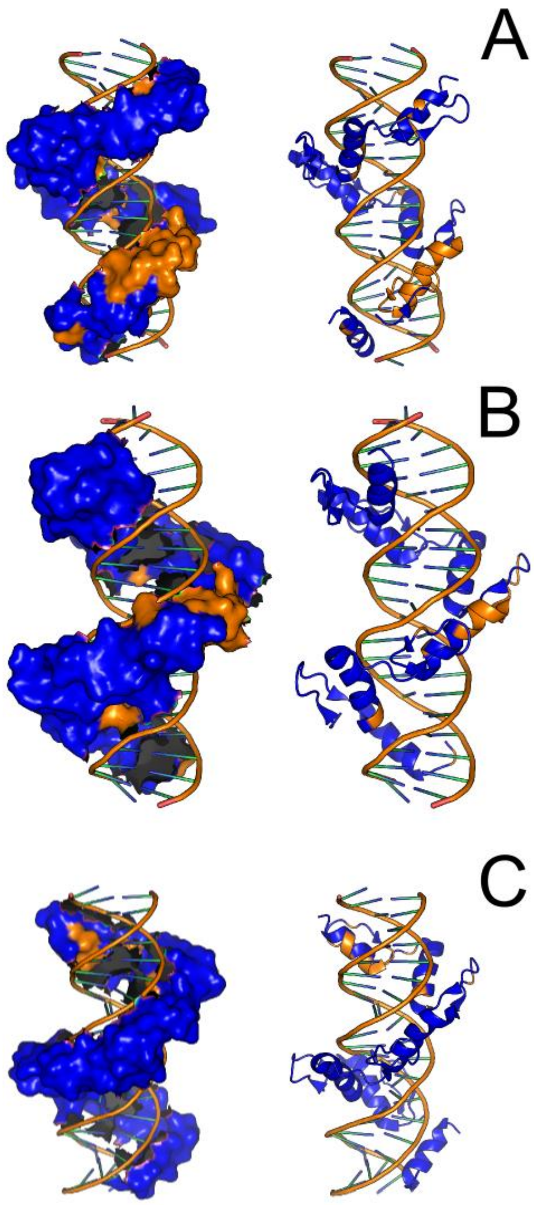


Figure 15: Structural visualization of identified clustered mutation (P_{3D} < 0.05) in DNA-binding proteins. A: Surface and cartoon representation of mutations on a protein of unknown function in the Danio lineage, (PDB key 2I13); B: Surface and cartoon representation of mutations on the ZNF507 protein in the Tetraodontidae lineage, (PDB key 2I13); C: Surface and cartoon representation of mutations on the ZNF251 protein in the Danio lineage, (PDB key 2I13). Substituted sites in the lineage of interest are highlighted in orange.

Table 18 lists the top 20 results of clustered adaptation occurring in buried regions of the protein (lineages with a P_{3D} < 0.05 and P_{asa} < 0.05) for DNA-binding phylogenies. These predictions represent possible positive selection on core regions of protein structures that may affect protein motion, stability, and folding patterns.

Table 18: Top 20 hits of clustered, structurally internal adaptations (P_{3D} < 0.05 and P_{asa} < 0.05) in DNA-binding transcription factors.

Protein	Ancestral Clade	Derived Clade	PDB	# of	P_{asa}	FDR P_{asa}	P_{3D}	FDR	R
			ID	mutations				P_{3d}	
GM10323	Mus	Mus	2I13	16	2.05E-2	1.0	4.10E-6	9.37E-4	1.03
NFKB2	Amniota	Theria	1SVC	33	4.37E-2	1.0	3.25E-5	5.96E-3	0.00
ENSACAP 00000010745	Danio	Danio	1TF6	5	3.38E-2	1.0	7.04E-4	5.61E-2	2.35E-1
zgc@173816	Danio	Danio	2I13	36	3.69E-2	1.0	7.41E-4	5.80E-2	2.82
ENSMODP 00000028376	Didelphimorphia	Didelphimorphia	2I13	4	2.50E-2	1.0	1.00E-3	5.98E-2	4.13E2
ZFP64	Euteleostomi	Eutheria	1TF6	66	3.42E-2	1.0	1.32E-3	7.30E-2	4.83E-1
ENSACAP 00000019475	Polychrontinae	Polychrotinae	2I13	12	4.32E-2	1.0	1.73E-3	8.88E-2	1.25E1
Q5ZHS5	Theria	Metatheria	2C6Y	2	1.3E-2	1.0	2.00E-3	8.94E-2	0.00
NFKB2	Eutheria	Afrotheria	1SVC	5	1.51E-3	1.0	4.83E-3	1.50E-1	0.00
ZFP760	Mus	Mus	2I13	18	8.07E-3	1.0	9.36E-3	2.06E-1	5.70
ENSTNIP 00000005721	Taeniopygia	Taeniopygia	2I13	5	1.50E-2	1.0	1.30E-2	2.37E-1	3.64
FOXP1	Euteleostomi	Sarcopterygii	2C6Y	3	2.90E-2	1.0	1.60E-2	2.60E-1	7.64E-1
FOXN1	Euteleostomi	Amniota	2C6Y	12	1.61E-2	1.0	2.29E-2	3.03E-1	1.53E-1
ZFP275	Eutheria	Euarchontoglires	1TF6	3	6.23E-3	1.0	2.30E-2	3.03E-1	0.00
FOXJ2	Euteleostomi	Amniota	2C6Y	17	4.04E-2	1.0	2.38E-2	3.09E-1	4.28E-1
HMGB4	Eutheria	Eutheria	2GZK	7	3.06E-2	1.0	2.59E-2	3.19E-1	9.94E-1
ENSGACP 00000022515	Clupeocephala	Holacanthopterygii	2C6Y	10	2.80E-2	1.0	2.80E-2	3.29E-1	6.58E-1
NFKB2	Clupeocephala	Percomorpha	1SVC	25	2.58E-3	1.0	3.55E-2	3.65E-1	0.00
GM3604	Murinae	Rattus	2I13	14	4.88E-2	1.0	3.98E-2	3.84E-1	6.43
ZNF789	Theria	Theria	2I13	13	4.38E-2	1.0	4.76E-2	4.14E-1	4.35

Interesting predictions in Table 18 include ancestral clustered mutation in the Foxp1 transcription factor, which along with Foxp2 are widely implicated in vertebrate- or mammalian-specific neuronal development, cognitive and language development. Another forkhead transcription factor, FoxN1 (also known as *WHN*) was predicted to have accumulated a set of 12 highly clustered mutations along the branch from Euteleostomi to Amniota. FoxN1 is implicated in the development of the thymus and differentiation of keratinocytes and hair follicles (RefSeq Annotation, Apr 2013) (Mecklenburg, Tychsen, & Paus, 2005; Nakamura *et al.*, 2008).

Table 19 lists the top 20 results of clustered adaptation occurring in exposed regions of the protein (lineages with a $P_{3D} < 0.05$ and $P_{asa} > 0.95$) for DNA-binding phylogenies. These predictions include potential positive selection on particular surface regions in transcription factors, which may affect transcription factor interactions with other protein cofactors or DNA targets. Again ZNF576 is detected as a top-scoring candidate. In addition, top scoring hits include ancestral lineages of homeobox A7 and homeobox containing 1 transcription factors. Interestingly, a clustered surface patch of mutations was detected for *HOXA7*, a transcription factor involved in keratinocyte differentiation.

Table 19: Top 20 hits of clustered adaptation the associated to exposed residues (P_{3D} < 0.05 and P_{asa} > 0.95) in sequences that BLASTed to DNA-binding domain PDB structures.

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3D}	FDR P_{3D}	R
ZNF576	Clupeocephala	Holacanthopterygii	2I13	22	9.66E-1	1.0	8.60E-85	3.47E-81	2.64E-1
HOXA7	Amniota	Testudines	1AHD	6	9.61E-1	1.0	1.73E-5	3.42E-3	0.00
Q5ZJP8	Euteleostomi	Sarcopterygii	2BGW	17	9.58E-1	1.0	8.34E-5	1.29E-2	2.46E-1
HIC1	Tetrapoda	Amniota	2I13	11	9.74E-1	1.0	3.14E-4	3.24E-2	3.01E-1
ZNF576	Euarchontoglires	Euarchontoglires	2I13	10	9.92E-1	1.0	4.33E-4	4.06E-2	5.00E-2
ANKRD60	Amniota	Sauria	1AWC	3	9.92E-1	1.0	8.65E-4	5.98E-2	0.00
ZBTB2	Sarcopterygii	Tetrapoda	1MEY	4	9.67E-1	1.0	8.83E-4	5.98E-2	0.00
HOXA7	Euteleostomi	Euteleostomi	1AHD	4	9.96E-1	1.0	1.54E-3	8.16E-2	0.00
KLF14	Boreoeutheria	Euarchontoglires	2I13	4	9.60E-1	1.0	1.98E-3	8.94E-2	2.60E-1
PAX6	Amniota	Testudines	6PAX	2	9.78E-1	1.0	2.00E-2	8.94E-2	1.22
HOXA7	Glires	Sciurognathi	1AHD	3	9.89E-1	1.0	1.87E-3	8.94E-2	0.00
HMBOX1	Percomorpha	Percomorpha	1IC8	3	9.75E-1	1.0	6.59E-3	1.77E-1	0.00
OVOL3	Tetrapoda	Amniota	2I13	15	9.83E-1	1.0	1.10E-2	2.20E-1	4.43E-1
ANKRD60	Theria	Eutheria	1AWC	3	9.95E-1	1.0	1.60E-2	2.60E-1	0.00
MGMT	Eukaryota	Eukaryota	1YFH	5	9.92E-1	1.0	1.70E-2	2.67E-1	0.00
MEF2B	Theria	Eutheria	1N6J	2	9.98E-1	1.0	3.00E-2	3.38E-1	0.00
A4PET4	Simiiformes	Catarrhini	2I13	3	9.97E-1	1.0	3.20E-2	3.49E-1	0.00
URB1	Laurasiatheria	Caniforma	2G8F	3	9.78E-1	1.0	3.33E-2	3.53E-1	0.00
ARID4B	Euteleostomi	Percomorpha	1KQQ	4	9.66E-1	1.0	5.00E-2	4.21E-1	0.00
ZNF687	Tetrapoda	Amniota	2I13	3	9.84E-1	1.0	4.9E-2	4.18E-1	4.68E-1

Function enrichment analysis

Lists containing Ensembl gene ids for the DNA-binding proteins were again used to perform functional enrichment analysis using the David functional annotation web tool (Cunningham *et al.*, 2015). Different gene id lists were prepared according to significant hits according to the various Adaptation3D metrics (i.e., clustered mutation, clustered mutation in buried residues, clustered mutation in exposed residues, and exposed/surface mutations). The background list used with all four of these foreground lists contained the gene ids from all protein phylogenies in the Selectome Euteleostomi taxonomic cluster that BLASTed to DNA-binding transcription factor PDB structures from the BIPA database (Camacho *et al.*, 2008; Lee, & Blundell, 2009). Table 20 displays significant (Benjamini *P*-value < 0.05) enriched functions for DNA-binding protein phylogeny lineages with clustered adaptation (FDR corrected P_{3D} < 0.05). Enriched functions include keywords pertaining to zinc finger transcription factors. This suggests a non-random, recurring pattern of clustered mutation in the evolutionary history of zinc-fingers.

Table 20: Enriched functions for proteins with mutations that clustered together (FDR corrected $P_{3D} < 0.05$) from DNA-binding phylogenies.

Category	Term	Count	*P*-Value	Benjamini *P*-Value
INTERPRO	Zinc finger, C2H2-type/integrase, DNA-binding	50	2.8E-7	1.8E-5
INTERPRO	Zinc finger, C2H2-type	54	7.2E-7	2.3E-5
INTERPRO	Zinc finger, C2H2-like	54	7.2E-7	2.3E-5
SMART	ZnF C2H2	54	8.5E-6	1.7E-4
GO Terms	zinc ion binding	55	1.5E-5	4.9E-4
GO Terms	transition metal ion binding	55	2.2E-5	4.6E-4
GO Terms	cation binding	56	8.0E-6	5.1E-4
GO Terms	ion binding	56	8.0E-6	5.1E-4
GO Terms	metal ion binding	56	8.0E-6	5.1E-4

No statistically significant functional enrichments were found for lineages with clustered mutations in buried regions of DNA-binding proteins ($P_{3D} < 0.05$ and $P_{asa} < 0.05$) for DNA-binding protein phylogenies.

Table 21 displays enriched functions for lineages with clustered mutations in exposed regions of the protein ($P_{3D} < 0.05$ and $P_{asa} > 0.95$) for DNA-binding protein phylogenies. Statistically enriched functions include keywords pertaining to zinc finger transcription factors. This indicates that zinc-fingers show a significant tendency for clustered mutation on exposed, surface regions.

Table 21: Enriched functions for proteins with mutations that clustered together (P_{3D} < 0.05) and were significantly exposed (P_{asa} > 0.95) from DNA-binding phylogenies.

Category	Term	Count	*P*-Value	Benjamini *P*-Value
Uniprot	zinc finger region: C2H2-type 2	32	5.5 E-5	8.0 E-3
GO Terms	transition metal ion binding	35	1.4 E-3	2.9 E-2
GO Terms	cation binding	36	5.1 E-4	3.1 E-2
GO Terms	ion binding	36	5.1 E-4	3.1 E-2
GO Terms	zinc ion binding	35	1.2 E-3	3.5 E-2
Swiss-Prot	metal-binding	35	2.3 E-3	5.5 E-2
SMART	ZnF C2H2	32	4.0 E-3	5.9 E-2
Swiss-Prot	zinc	35	1.9 E-3	8.9 E-2
Swiss-Prot	zinc-finger	33	8.1 E-3	1.3 E-1
INTERPRO	Zinc finger, C2H2-like	32	3.7 E-3	1.5 E-1

3.3.2 Identification of clustered mutation in protein-DNA interfaces: candidates for positive selection on divergence of DNA-binding specificity

In Section 3.3.1, the Adaptation3D algorithm was used to make predictions about spatially clustered adaptation on the trees and alignments pertaining to transcription factor phylogenies in the Selectome database. A subsequent analysis was performed to determine the specificity of substitution in known DNA-binding sites over the course of transcription factor molecular evolution.

The BIPA database contains not only PDB structures that are involved in DNA-binding, but also tabulates the amino acid sites that participate within the protein-DNA binding interface (Lee, & Blundell, 2009; Worth *et al.*, 2007). For a given lineage with mutations, a simple metric was therefore used to calculate the relative enrichment of mutations within DNA-binding site compared to those outside the DNA-binding site as defined using the pre-computed information from the BIPA database. This was done ultimately to determine if it was possible to observe instances of highly specific mutation in DNA-binding interfaces. If so, this would indicate a tendency for diversification of protein-DNA specificity in transcription factors and thus a set of predicted historical adaptations with functional consequences.

Equation 1 demonstrates the calculation for mutation enrichment in DNA-binding interfaces. Again, this calculation is performed for the set of mutations calculated along all branches (lineages). Suppose a protein is of length *N* residues and has a binding site composed of N_{B} residues, and has *M* total branch-specific mutations and M_{BS} mutations that occur in binding site residues. We can calculate the binding site mutation enrichment *R* as: the proportion of binding site mutations divided by the proportion of non-binding site mutations (Equation 1):

$$R \ = \frac{\frac{M_{\mathrm{BS}}}{N_{\mathrm{Bs}}}}{\frac{(M - M_{\mathrm{BS}}) + c}{(N - N_{\mathrm{Bs}})}}$$

Equation 1: Equation to calculate a ratio representing the degree of mutation specificity in the transcription factor binding site. $\mathbf{M_{Bs}}$ = number of substitutions occurring in the binding site, $\mathbf{N_{Bs}}$ = total number of residues in the binding site, M = total number of substitutions, N = total number of residues, C = pseudo-count of 0.5 (to avoid division by 0 errors).

Here, high values of *R* correspond to a high enrichment of mutations in the binding site. Figure 16 displays an X-Y plot of substitution specificity outside the DNA-binding site versus substitution specificity within the DNA-binding site for all transcription factor lineages. The results from Figure 16 appears to show a greater proportion of lineages have a higher degree of non-binding site substitution than binding site substitution. However, a subset of results exhibit binding-site substitution specificity.

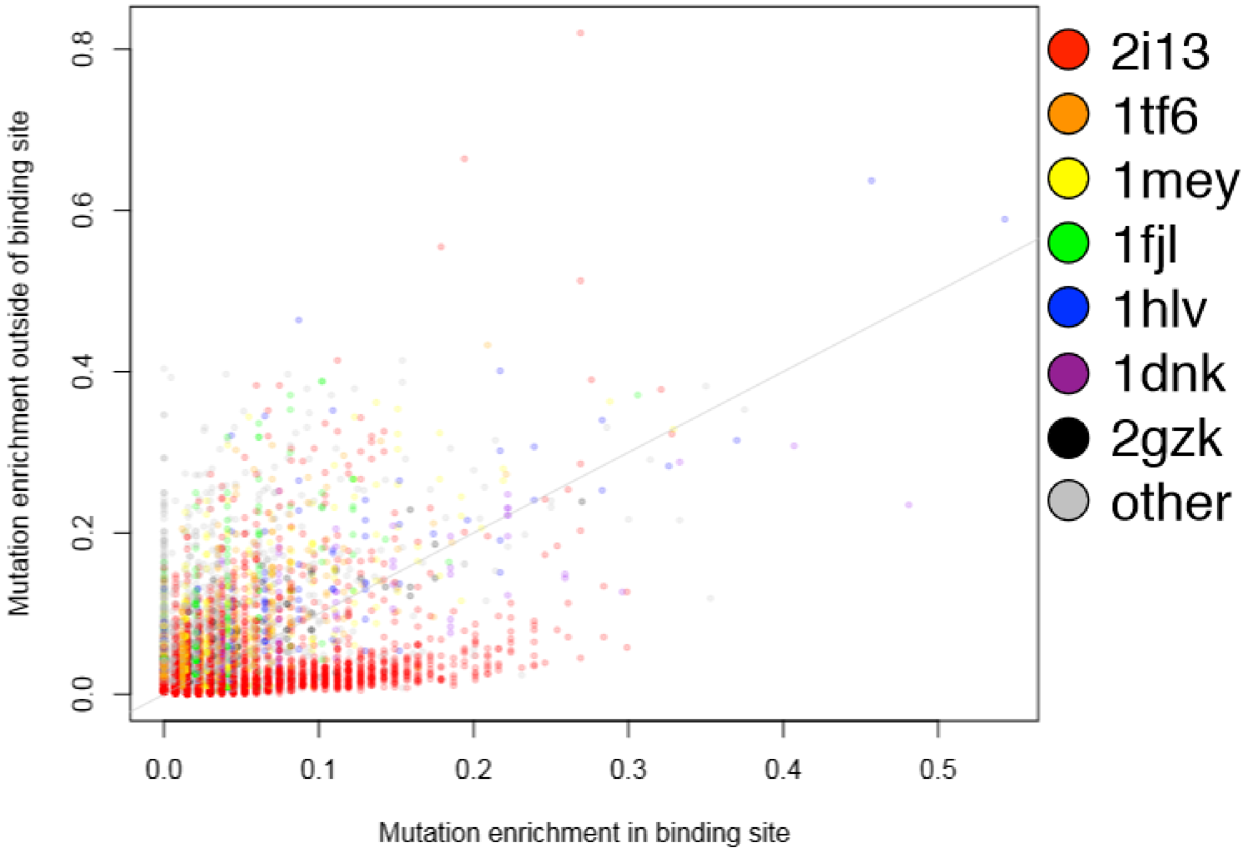


Figure 16: Observed mutation enrichment within versus outside the DNA-binding site for all transcription factor lineages. The fraction of mutations outside of the DNA-binding site (y-axis) versus the fraction of mutations inside the DNA-binding site (x-axis) is depicted. Points are coloured according to representative PDB ID. Colors are shown for the top seven most abundant PDB families, while all others are grey.

Figure 17 displays a histogram of the $\log_{10}$ of the overall binding site specificity ratio *R* for all transcription factor lineages analyzed. Interestingly, the distribution of binding site specificity ratios appears to be somewhat bimodal, indicating a tendency toward biased substitution either within ($\log_{10}(R) > 0$) or outside of DNA binding sites ($\log_{10}(R) < 0$).

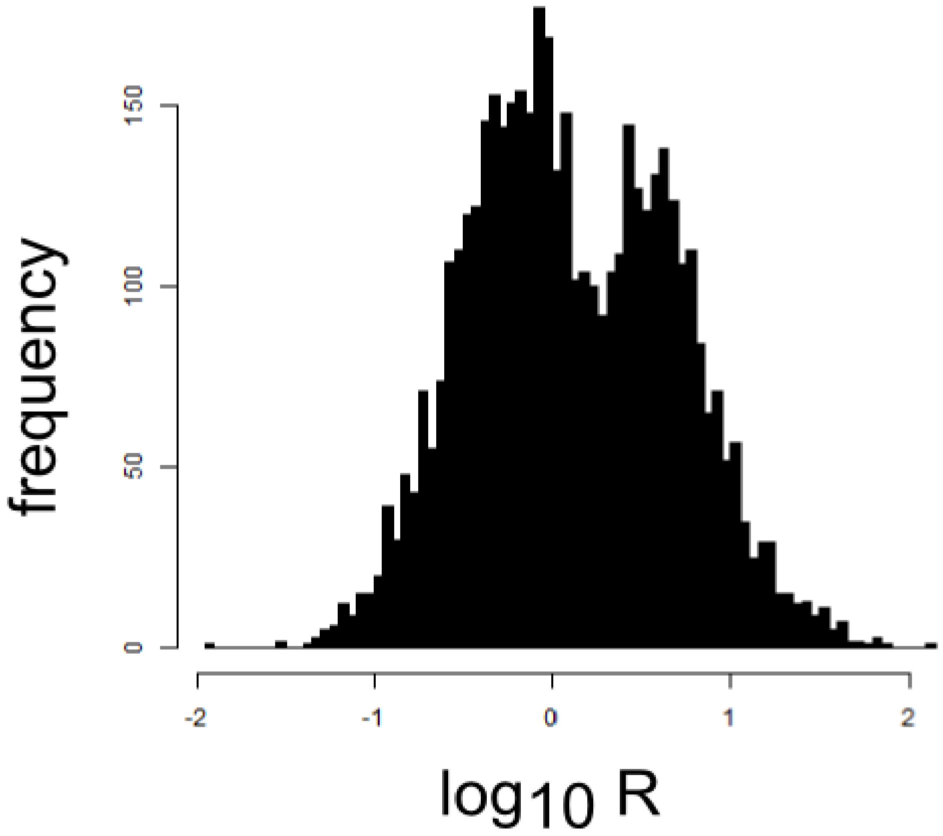


Figure 17: Distribution of observed mutation enrichment within DNA-binding sites. The plot is a histogram of the $\log_{10}$ of the overall binding site specificity ratio *R* for all transcription factor lineages from Selectome.

Table 22 lists the top Adaptation3D predictions for transcription factors ranked by the ratio *R* value defined in Equation 1. Interestingly, the list is dominated by cases in which a transcription factor has accumulated only a few mutations within its DNA-binding site along a relatively short evolutionary branch. Many of these predictions also correspond to branches in which the ancestral and derived clades are identical, implying examples of lineage-specific gene duplication. This is a very interesting result since it suggests that the most extreme examples of positive selection for changes in DNA-binding sites occur in newly duplicated transcription factor paralogs. Top scoring predictions include include a range of DNA-binding domain families including zinc-fingers, SMAD family, and forkhead box transcription factors.

Table 22: Top 20 hits of mutations that specifically occurred in known binding site residues in sequences that BLASTed to DNA-binding domain PDB structures.

Protein	Ancestral Clade	Derived Clade	PDB ID	# of mutations	P_{asa}	FDR P_{asa}	P_{3D}	FDR P_{3D}	R
ZNF121	Euarchontoglires	Euarchontoglires	2I13	10	1.19E-1	1.0	2.52E-1	7.61E-1	128.74
ZFP60	Mus	Mus	2I13	5	3.00E-1	1.0	7.13E-1	9.93E-1	74.82
ENSECAP00000005030	Murinae	Murinae	2I13	5	8.14E-1	1.0	1.18E-1	5.89E-1	68.42
ZNF596	Xenartha	Xenartha	2I13	5	3.46E-1	1.0	3.53E-1	8.43E-1	67.81
ENSACAP00000019712	Metatheria	Didelphimorphia	2I13	3	5.41E-1	1.0	7.65E-1	1.00	57.84
ENSACAP00000021182	Polychrotinae	Polychrotinae	2I13	3	6.83E-1	1.0	8.46E-1	1.00	52.77
ZNF121	Boreoeutheria	Boreoeutheria	2I13	4	1.02E-1	1.0	1.34E-1	6.16E-1	51.61
ENSSHAP00000004029	Sarcophilus	Sarcophilus	2I13	3	4.23E-1	1.0	3.52E-1	8.43E-1	48.97
ZNF236	Laurasiatheria	Cetartiodactyla	2I13	2	1.82E-1	1.0	5.61E-1	9.49E-1	46.79
ZNF234	Xenopodinae	Xenopodinae	2I13	5	2.53E-1	1.0	1.71E-2	2.68E-1	41.92
ENSMODP00000028376	Didelphimorphia	Didelphimorphia	2I13	4	2.50E-2	1.0	1.00E-3	5.98E-2	41.26
ZNF596	Cingulata	Cingulata	2I13	3	2.20E-1	1.0	1.24E-1	5.99E-1	40.72
ENSVPAP00000006282	Boreoeutheria	Hominidae	2I13	3	2.24E-1	1.0	2.65E-1	7.74E-1	40.50
ZNF208	Pelodiscus	Pelodiscus	2I13	8	5.03E-1	1.0	3.13E-1	8.14E-1	38.49
ENSSHAP00000004029	Sarcophilus	Sarcophilus	2I13	8	4.66E-1	1.0	6.00E-2	4.54E-1	38.04
ENSSHAP00000021844	Metatheria	Metatheria	2I13	8	9.03E-1	1.0	4.87E-1	9.18E-1	37.27
ENSXETP00000044932	Tetrapoda	Tetrapoda	2I13	2	2.90E-1	1.0	2.50E-2	3.14E-1	36.33
ENSMODP00000013553	Didelphimorphia	Didelphimorphia	2I13	2	3.02E-1	1.0	2.35E-1	7.45E-1	35.34
ENSACAP00000019712	Polychrotinae	Polychrotinae	2I13	11	1.35E-1	1.0	2.16E-1	7.25E-1	34.70
ENSPSIP00000009740	Pelodiscus	Pelodiscus	2I13	3	5.15E-1	1.0	5.58E-1	9.48E-1	33.78

An analysis was performed to determine if binding-site specific change occurs more often in gene duplication lineages compared to speciation lineages. This was done by comparing the number of lineages that exhibited binding-site specific modification between two sets of lineages: paralogs and orthologs. Overall, 2494 lineages were used to make this assessment, 883 of which occurred in gene duplication lineages and 1611 of which occurred in speciation lineages. 566 gene duplication lineages had more specific substitution inside the binding site versus outside (i.e., a binding site ratio > 1). 317 gene duplication lineages had more specific substitution outside the binding site than inside (i.e. binding site ratio < 1).

463 speciation lineages exhibited binding site specific substitution, and 1148 speciation lineages exhibited non-binding site specific substitution.

A Fisher's exact test was performed to determine if the difference between these ratios was statistically significant (Fisher, 1922). A matrix representing the inputs to the Fisher's exact test is represented in Table 23.

Table 23. Matrix representation of values used for Fisher's exact test to assess statistical significance of differences in prevalence of binding-site specific modification between paralog and ortholog lineages.

	Modification in binding site ($R > 1.0$)	**Modification outside of binding site ($R < 1.0$)**
Paralogs	566	317
Orthologs	463	1148

The *P*-value for this test was 6.75E-66, rejecting the null hypothesis that the odds ratio between the sets of counts is equal to 1. This suggests that most instances of functionally-relevant transcription factor binding-specific changes have occurred during instances of gene duplication (i.e., between paralogs), rather than during instances of speciation (i.e., between orthologs).

3.3.3 Summary and Discussion

Applying the Adaptation3D algorithm to DNA-binding transcription factors from the Selectome dataset can provide insights into the nature of how clustered mutation can lead to changes in organism phenotype. Clustered mutation in DNA-binding sites can potentially lead to alterations in transcription factor DNA subsequence binding preferences, thereby altering sequence-specific transcription factor binding preferences (Jarvela *et al.*, 2014; Nitta *et al.*, 2015). This can lead large-scale modifications of downstream gene regulatory networks (Erwin & Davidson, 2009; Peter & Davidson, 2011). It is interesting to see if there are any classifications of transcription factors that have undergone extensive clustered modification in the evolutionary timeline compared to other types.

Table 21 displays statistically enriched biological processes and functions for significant mutation clusters in DNA-binding protein phylogenies compared to a background of all the DNA-binding phylogenies from the Selectome dataset. These function enrichment tables were produced using the DAVID analysis tool (Huang, Sherman, & Lempecki, 2009). Table 21 indicates that there is a statistically significant enrichment of clustered adaptation in Zinc-finger transcription factors compared to other classifications of transcription factors. One study found that there has been diversification of transcription factor paralog DNA-binding specificity through modulation of residues that outside of the common-core binding sites (Siggers, Reddy, Barron, & Bulyk, 2014). The poly-zinc-finger gene family of transcriptional repressors have displayed a great degree of duplication and divergence in primate lineages, including human (Emerson & Thomas, 2009). d_N/d_S analyses of these genes have revealed that many of these lineages have undergone positive selection, most likely to affect DNA-binding specificity. The enrichment of zinc-finger related keywords in the DNA-binding Selectome data subset correlates with some of the observations about historical adaptation that have been made in the literature.

Figure 16 displays the ratio of substitution inside the DNA-binding site compared to the ratio of substitution specificity outside of the DNA-binding site for transcription factor lineages. By looking at the number of points above and below the y=x line, it is evident that most lineages exhibit a higher degree of substitution outside the binding site. However, there is still a large number of lineages with substitution specificity in the binding site. The points on the plot are coloured according to the PDB ID of the structure that the lineage BLASTed to. The plot shows that the majority of transcription factor lineages with binding site specific substitution (i.e., below the y=x line) are coloured red. These lineages aligned to PDB structure 2I13, which is a six-finger zinc finger transcription factor (Segal, *et al.*, 2006). Therefore, this may be an initial indication that zinc-finger transcription factors have undergone more DNA-binding site specific mutation over the course of evolutionary history compared to other TF classes.

Figure 17 displays a histogram of the *R* ratio value for all transcription factor lineages. From the data, it can be seen that the distribution of binding site specificity ratios is bimodal, displaying one mode of lineages with a greater degree of non-binding site specific substitution ($\log_{10}(R) < 0$), and one mode of lineages with a greater degree of binding site specific substitution ($\log_{10}(R) > 0$). This suggests that there are possible multiple ways in which DNA-binding proteins can undergo adaptation with relation to structure.

The results from the Fisher Exact Test displayed in Table 23 indicates that there is a statistically significant increase in the prevalence of binding-site specific substitution in gene duplication (paralog) lineages versus speciation (ortholog) lineages. This suggests that gene duplication events enable modification to DNA-binding preference in the newly duplicated gene, and that these events lead to greater regulatory complexity compared to speciation events (Pougach *et al.*, 2014).

Chapter 4

Discussion and future directions

In the previous chapters, I have demonstrated that the Adaptation3D algorithm can be useful for finding statistically significant, spatially clustered sets of substitutions that have occurred in the history of a protein family, thereby inferring functional shifts at certain points in protein phylogenies. This structure-aware approach for detecting positive selection can be used in conjunction with more traditional methods to make more educated predictions about the evolution of novel protein functions.

4.1 Summary of Main Findings

Overall, applying Adaptation3D to a wide range of phylogenies and lineages has demonstrated that clustered mutation is a widespread evolutionary phenomenon. This can be seen in Figure 11, where there is a spike in *P*-values that are indicative of spatially clustered groups of substitutions. Thus, although identifying mutation clusters has traditionally only been applied to select protein classes or phylogenies (Wagner, 2007; Zhou *et al.,* 2008), there is ultimately value in performing large-scale analyses to detect clustered adaptation in many different protein phylogenies throughout the tree of life.

Adaptation3D has been demonstrated to predict functionally relevant mutation clusters that correlate with the scientific literature in a number of cases. The foremost of this is in the case of PR-5d. Adaptation3D detected one significant lineage with clustered adaptation within the osmotin/pathogensis-related protein phylogeny. This lineage corresponds to the emergence of a derived WWW structural motif, which is potentially indicative of a novel carbohydrate-binding surface patch (Doxey *et al.,* 2010; Koyama *et al.,* 2001). The epidermal growth factor beta-urogastrone was found to have undergone clustered adaptation in the Amniota lineage. This change could correspond with the development of extraembryonic membranes that is characteristic of Amniote development (Albergotti *et al.,* 2009; Cross *et al.,* 2003; Jojovic *et al.,* 1998).

Adaptation3D has also been used to detect certain functional classes of proteins that are statistically overrepresented as having undergone clustered adaptation. Some of these functional classifications include DNA-binding protiens, protein kinases and immunoglobulins. In the case of transcription factors, Zinc-finger DNA-binding proteins have a greater incidence of clustered adaptation compared to other TF classifications.

Lastly, an analysis of transcription factor phylogenies and lineages shows that there is a significant difference in the likelihood of DNA-binding site specific substitution between paralog and ortholog lineages. Proteins resulting from gene duplication were found to be more likely to have binding site specific adaptation compared to proteins resulting from speciation. These results support the Ohno (1970) model that functional divergence is likely to occur in one copy of a duplicate gene following a gene duplication event. In effect, this result also supports the ortholog conjecture, that is, gene orthologs are more likely to retain the same function following speciation than paralogs are following duplication (Altenhoff, Studer, Robinson-Rechavi, & Dessimoz, 2012).

4.2 Potential future improvements to the Adaptation3D method

This section will discuss some potential advancements to the Adaptation3D algorithm that were explored in a preliminary fashion but were not pursued as core components. These features could be expanded upon and implemented in future versions of the Adaptation3D method to provide more detailed predictions with a greater focus on *types* of protein functional modification.

Adaptation3D uses statistical random sampling and *P*-values based on observed distributions of pairwise Euclidean distances and solvent accessible surface area values. It was earlier hypothesized that this process and resampling statistic could be applied to other properties as well. If a group of amino acid substitutions led to a statistically significant change in quantitative amino acid property values, for instance, this could signify some kind of protein adaptation. Thus, one of the additional properties analyzed in this way was amino acid hydropathy index. Hydropathy index is a measure of the hydrophobicity versus hydrophilicity of an amino acid (Biro, 2006; Kyte & Doolittle, 1983). Strongly hydrophobic residues score high on the hydropathy scale, while strongly hydrophilic residues score low. One may hypothesize that mutations that significantly alter amino acid hydropathy index would be much more likely to alter aspects of the protein’s function than mutations the only slightly affect hydropathy index. In some cases, large hydropathy index changes could affect conformational and stability properties, and alteration of interactions with small ligands and other proteins.

A second amino acid property change that can be explored in greater detailed is side chain mass change. Groups of amino acid substitutions that led to large changes in the side chain mass may be more likely to lead to functional changes compared to groups of amino acid substitutions that only slightly changed reside mass.

To obtain *P*-values for both hydropathy index change and mass change, substitutions from the entire phylogeny were randomly sampled, and RMSD values were calculated (repeated 10,000 times to build random distribution). The CDF of the RMSD of hydropathy index change or mass change for the observed group of substitutions on its respective random distribution represented the *P*-value for that property. Appendix A Supplementary Figure S1 displays matrices that can be used to calculate root mean squared deviation (RMSD) for hydropathy index changes (S1A) and side chain mass change (S1B) for a given set of amino acid substitutions.

Some amino acid residues are more frequently involved in functional sites than others. An algorithm that is aware of this phenomenon would potentially have more power to detect function-altering substitutions. To determine which residues are most likely to be involved in protein functional sites, frequencies of amino acids can be measured in key functional databases such as the Catalytic Site Atlas (CSA), the extended Catalytic Site Identification database (CSI), and the Protein Family Interaction database (iPFam) (Finn, Miller, Clements, & Bateman, 2014; Kirshner, Nilmeier, & Lightstone, 2013; Porter, Bartlett, & Thornton, 2004). To compute background frequencies for comparison, a random sampling of the PDB database can be performed. A bar chart displaying the proportions of different amino acid types in different types of functional sites is displayed in Figure 18.

From Figure 18, we can see that certain residues are more prevalent in functional databases compared to a random sampling of residue frequencies from the PDB. Alanine, phenylalanine, isoleucine, leucine, methionine, proline, glutamine, threonine, and valine are generally overrepresented in the PDB background and underrepresented in functional sites. On the other hand, cysteine, aspartic acid, glutamine, histidine, lysine, asparagine, arginine, tryptophan, and tyrosine are overrepresented in one or more of the functional site databases. Therefore, mutations toward these residues could be considered more likely as candidates for causing functional modifications in proteins, with specific attention made to the type of mutation in question. For example, mutations to asparagine would be most likely to contribute to catalytic site changes (highly prevalent in the CSA), whereas mutations to arginine would be more likely to lead to ligand binding site changes (highly prevalent in iPFam).

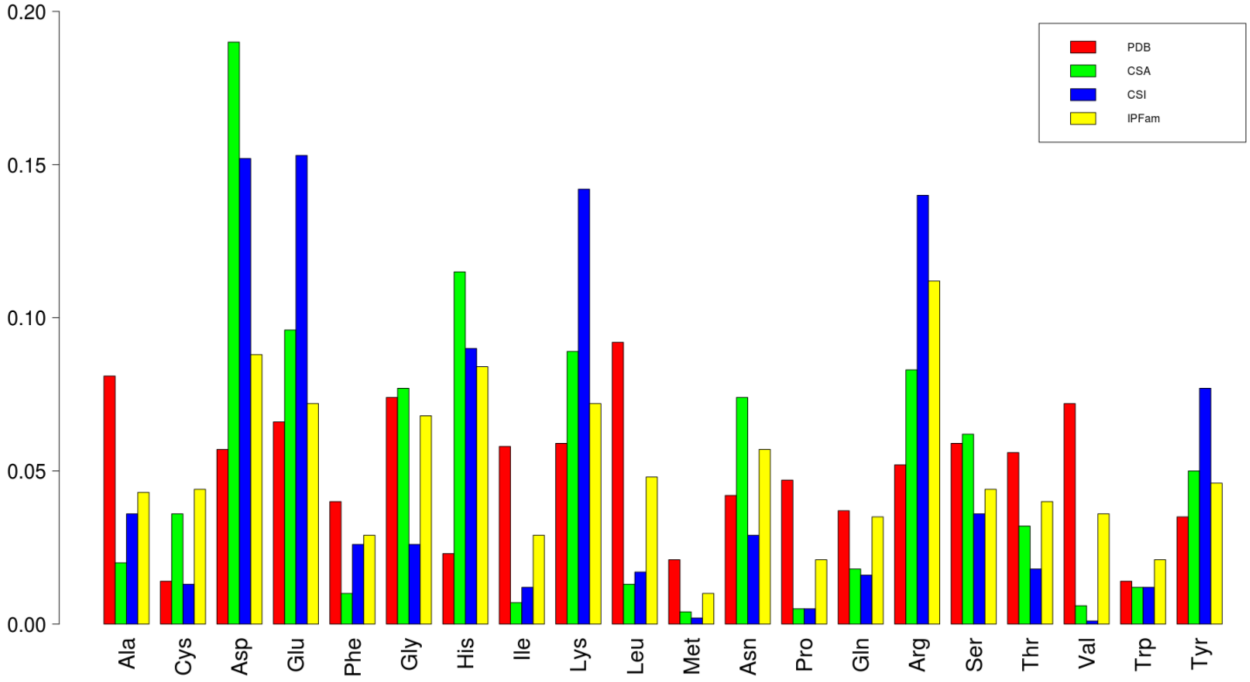


Figure 18: Bar chart displaying the prevalence of each amino acid type in various databases. red: PDB, green: CSA, blue: CSI, yellow: iPFam.

Another way of advancing the Adaptation3D algorithm is to incorporate solvent accessibility and distance information extracted from functional databases. Supplementary Figure S2 (Appendix A) illustrates how different residues are overrepresented in certain functional categories. It is reasonable to assume that certain residue types may have preferred ranges of solvent accessibility at which they perform binding or catalytic functions. These solvent accessibility ranges may deviate from that seen in randomly sampled positions. Solvent accessibilities were measured for all residue types in the CSA, CSI, and iPFam databases, and these distributions were compared to those from randomly sampled residues from the PDB.

Kernel density functions were used to generate density curves for all lists used to generate Supplementary Figure S2. It is evident that certain residue types do indeed exhibit very different solvent accessibility profiles when they are involved in functional sites versus when they are randomly sampled. Alanine, histidine, isoleucine, leucine, methionine, glutamine, and tyrosine have similar solvent accessibility density curves between functional and randomly sampled categories. On the other hand, aspartic acid, glutamic acid, phenylalanine, lysine, serine, tryptophan show distinct solvent accessibility distributions for functional sites. Therefore, mutations to any of these residues that exhibit the statistically preferred range of solvent accessibility for functional sites could be used as evidence supporting functional adaptation.

4.3 Conclusion

In conclusion, Adaptation3D shows promise as a new, integrative approach for detecting positive selection in protein phylogenies. It is the first method to my knowledge that integrates sequence, structural and phylogenetic information into a single framework for detecting positive selection. The Adaptation3D method has been applied both on the protein family scale as demonstrated using the PR-5 protein family, and as a screening tool for phylogenomic-scale protein adaptation detection. When applied on a large scale to the Selectome Database, the Adaptation3D approach has revealed a widespread evolutionary phenomenon of clustered substitution. This clustered substitution has occurred disproportionately in key lineages and protein families and may signify historical episodes of positive selection. Clustered mutation was widespread in some DNA-binding domain transcription factor families, notably including vertebrate zinc-fingers. Further analysis of transcription factors in general show that clustered mutation in many cases coincides with the DNA-binding site which likely represents functional alterations of DNA-binding specificity. Furthermore, a significantly greater degree of DNA-binding site divergence has occurred in new TF duplicates (paralogs) versus TFs evolving by speciation (orthologs). This not only provides some validation for the Adaptation3D method to predict functional divergence, but it is consistent with other studies on the functional divergence of paralogs versus orthologs (Altenhoff *et al.*, 2012; Conant & Wolfe, 2008), and confirms Ohno's (1970) classic model of functional divergence following gene duplication.

References

Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research, 19*(5), 711-722.

Albergotti, L.C., Hamlin, H.J., McCoy, M.W., & Guillette, L.J. (2009). Endocrine activity of extraembryonic membranes extends beyond placental amniotes. *Plos ONE, 4*(5), e5452.

Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., & Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *Plos Computational Biology, 8*(5), e1002514.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215,* 403-410.

Arbiza, L., Dopazo, J., & Dopazo, H. (2006). Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *Plos Computational Biology, 2*(4), 288-300.

Areal, H., Abrantes, J., & Esteves, P.J. (2011). Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. *BMC Evolutionary Biology, 11,* 368.

Arfi, Y., Chevret, D., Henrissat, B., Berrin, J.G., Levasseur, A., & Record, E. (2013). Characterization of salt-adapted secreted lignocellulolytic enzymes from the mangrove fungus *Pestalotiopsis sp. Nature Communications, 4,* 1810.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozi, G., Zomer, O., & Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research, 40,* W580-W584.

Assis, R., & Bachtrong, D. (2015). Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evolutionary Biology, 15,* 138.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., . . . Bulyk, M.L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science, 324*(5935), 1720-1723.

Bamshad, M., Wooding, S.P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics, 4*(2), 99-111.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., . . . Zhang, J. (2015). UniProt: a hub for protein information. *Nucleic Acids Research, 43,* D204-D212.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W., . . . Langley, C.H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans. Plos Biology, 5*(11), 2534-2559.

Beltrao, P., Trinidad, J.C., Fiedler, D., Roguev, A., Lim, W.A., Shokat, K.M., … Krogan, N.J. (2009). Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *Plos Biology, 7*(6), e1000134.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289-300.

Benton, M.J. (1997). *Vertebrate Paleontology.* London: Chapman & Hall.

Benton, M.J., & Donoghue, P.C.J. (2006). Palaeontological evidence to date the tree of life. *Molecular Biology and Evolution, 24*(1), 26-53.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., . . . Bourne, P.E. (2000). The protein data bank. *Nucleic Acids Research, 28*(1), 235-242.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., . . . Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics, 74*, 1111-1120.

Biro, J.C. (2006). Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theoretical Biology and Medical Modelling, 3*, 15.

Burgess, R.W., Jucius, T.J., & Ackerman, S.L. (2006). Motor axon guidance of the mammalian trochlear and phrenic nerves: dependence on the netrin receptor Unc5c and modifier loci. *Journal of Neuroscience, 26,* 5756-5766.

Caffrey, D.R., O'Neill, L.A.J., Shields, D.C. (1999). The evolution of the MAP kinase pathways: coduplication of interaction proteins leads to new signaling cascades. *Journal of Molecular Evolution, 49*(5), 567-582.

Cai, W., Pei, J., & Grishin, N.V. (2004). Reconstruction of ancestral protein sequences and its applications. *BMC Evolutionary Biology, 4,* 33.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T.L. (2008). BLAST+: architecture and applications. *BMC Bioinformatics, 10,* 421.

Campos, M.A., Ribeiro, S.G., Rigden, D.J., Monte, D.C., & De Sa, M.F.G. (2002). Putative pathogenesis-related genes within *Solanum nigrum L. var. americanum* genome: isolation of two genes coding for PR5-like proteins, phylogenetic and sequence analysis. *Physiological and Molecular Plant Pathology, 61*(4), 205-216.

Cardona, G., Rossello, F., & Valiente, G. (2008). Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics, 9,* 532.

Cavallo, L., Kleinjung, J., Fraternali, F. (2003). POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Research, 31*, 3364-3366.

Chang, B.S.W., Ugalde, J.A., & Matz, M.V. (2005). Applications of ancestral protein reconstruction in understanding protein function: GFP-like proteins. *Methods in Enzymology, 395*, 652-670.

Charlesworth, B., Morgan, M.T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics, 134*(4), 1289-1303.

Colbert, E.H., & Morales, M. (2001). *Colbert's Evolution of the Vertebrates: A History of the Backboned Animals Through Time.* New York: John Wiley & Sons.

Conant, G.C., & Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics, 9*(12), 938-950.

Creevey, C.J., & McInerney, J.O. (2002). An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. *Gene, 300*, 43-51.

Cross, J.C., Baczyk, D., Dobric, N., Hemberger, M., Hughes, M., Simmons, D.G., . . . Kingdom, J.C. (2003). Genes, development and evolution of the placenta. *Placenta, 24*, 123-30.

Cunningham, C.W., Omland, K.E., & Oakley, T.H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution, 13*(9), 361-366.

Cunningham, F., Amone, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research, 43*, D662-D669.

Davey, J.L., & Blaxter, M.W. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, *9*(5-6), 416-423.

Demetrius, L., & Ziehe, M. (2007). Darwinian fitness. *Theoretical Population Biology, 72*, 323-345.

Doornik, J., & Ooms, M. (2003). Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models. *Computational Statistics & Data Analysis, 42*(3), 1-14.

Doxey, A.C., Yaish, M.W., Griffith, M., & McConkey, B.J. (2006). Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nature Biotechnology, 24*(7), 852-855.

Doxey, A.C., Cheng, Z.Y., Moffatt, B.A., & McConkey, B.J. (2010). Structural motif screening reveals a novel, conserved carbohydrate-binding surface in the pathogenesis-related protein PR-5d. *BMC Structural Biology, 10*, 23.

Drew, K., Winters, P., Butterfoss, G.L., Berstis, V., Uplinger, K., Armstrong, J., . . . Bonneau, R. (2011). The proteome folding project: proteome-scale prediction of structure and function. *Genome Research, 21,* 1981-1994.

Dubnovitsky, A.P., Kapetaniou, E.G., & Papageorgiou, A.C. (2005). Enzyme adaptation to alkaline pH: Atomic resolution (1.08 A) structure of phosphoserine aminotransferase from *Bacillus alcalophilus. Protein Science, 14*(1), 97-110.

Dunning, L.T., Dennis, A.B., Thomson, G., Sinclair, B.J., Newcomb, R.D., & Buckley, T.R. (2013). Positive selection in glycolysis among Australasian stick insects. *BMC Evolutionary Biology, 13*, 215.

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

Emerson, R.O., & Thomas, J.H. (2009). Adaptive evolution in zinc finger transcription factors. *Plos Genetics, 5,* e1000325.

Erwin, D.H., & Davidson, E.H. (2009). The evolution of gene regulatory networks. *Nature Reviews Genetics, 10*(2), 141-148.

Eyal, E., Najmanovich, R., Sobolev, V., & Edelman, M. (2001). MutaProt: a web interface for structural analysis of point mutations. *Bioinformatics, 17*, 381-382.

Falciatore, A., Merendino, L., Barneche, F., Ceol, M., Meskauskiene, R., Apel, K., & Rochaix, J.D. (2005). The FLP proteins act as regulators of chlorophyll synthesis in response to light and plastid signals in *Chlamydomonas*. *Genes & Development, 19*(1), 176-187.

Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., Strickler, S.R., . . . Mueller, L.A. (2015). The Sol Genomics Network (SGN)-from genotype to phenotype to breeding. *Nucleic Acids Research, 43*(D1), D1036-D1041.

Finn, R.D., Miller, B.L., Clements, J., & Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research, 42*(D1), D364-D373.

Fisher, R.A. (1922). On the interpretation of X^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85,* 87-94.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology, 19*(2), 99-113.

Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology, 20*(4), 406-416.

Fu, Y. (1996). New statistical tests of neutrality for DNA samples for a population. *Genetics, 143,* 557-570.

Gao, H., Wu, G., Spencer, T.E., Johnson, G.A., & Bazer, F.W. (2009). Select nutrients in the ovine uterine lumen. III. Cationic amino acid transporters in the ovine uterus and peri-implantation conceptuses. *Biology of reproduction, 80*(3), 602-609.

Genge, C.E., Davidson, W.S., & Tibbits, G.F. (2013). Adult teleost heart expresses two distinct troponin C paralogs: cardiac TnC and a novel and teleost-specific ssTnC in a chamber- and temperature-dependent manner. *Physiological Genomics, 45*(18), 866-875.

Ghanta, S., Grossman, R.E., & Brenner, C. (2013). Mitochondrial protein acetylation as a cell-intrinsic, evolutionary driver of fat storage: chemical and metabolic logic of acetyl-lysine modifications. *Critical Reviews in Biochemistry and Molecular Biology, 48*(6), 561-574.

Ghosh, R., & Chakrabarti, C. (2008). Crystal structure analysis of NP24-I: a thaumatin-like protein. *Planta, 228*(5), 883-890.

Goodstein, D.M., Shu, S.Q., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., . . . Rokshar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research, 40*(D1), D1178-D1186.

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

Grewal, S.I.S., & Rice, J.C. (2004). Regulation of heterochromatin by histone methylation and small RNAs. *Current Opinion in Cell Biology, 16*(3), 230-238.

Griffith, M., & Yaish, M.W. (2004). Antifreeze proteins in overwintering plants: a tale of two activities. *Trends in Plant Science, 9*(8), 399-405.

Grossman, S.R., Shylakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., . . . Sabeti, P.C. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science, 327*(5967), 883-886.

Grove, D.E., Willcox, S., Griffith, J.D., & Bryant, F.R. (2005). Differential single-stranded DNA binding properties of the paralogous SsbA and SsbB proteins from *Streptococcus pneumoniae. Journal of Biological Chemistry, 280*(12), 11067-11073.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology, 59*(3), 307-321.

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

Gutteridge, A., & Thornton, J.M. (2005). Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences, 30*(11), 622-629.

Halligan, D.L., Kousathanas, A., Ness, R.W., Harr, B., Eory, L., Keane, T.M., . . . Keightley, P.D. (2013). Contributions of protein-coding and regulatory changes to adaptive molecular evolution in murid rodents. *Plos Genetics, 9*(12), e1003995.

Harpur, B.A., Kent, C.F., Molodtsova, D., Lebon, J.M.D., Alqarni, A.S., Owayss, A.A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *PNAS USA, 111*(7), 2614-2619.

Hirasawa, T., & Kuratani, S. (2013). A new scenario of the evolutionary derivation of the mammalian diaphragm from shoulder muscles. *Journal of Anatomy, 222*(5), 504-517.

Holsinger, K.E., & Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating, and interpreting F(ST). *Nature Reviews Genetics, 10*(9), 639-650.

Huang, D.W., Sherman, B.T., & Lempecki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57.

Hudson, R.R., & Kaplan, N.L. (1995). Deleterious background selection with recombination. *Genetics, 141*(4), 1605-1617.

Huelsenbeck, J.P., Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8), 754-755.

Hughes, A.L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity, 99,* 364-373.

Hughes, A.L. (2008). The origin of adaptive phenotypes. *PNAS USA, 105*(36), 13193-13194.

Hughes, A.L., & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature, 335*(6186), 167-170.

Hughes, A.L., & Nei, M. (1989). Evolution of the major histocompatibility complex-independent origin of nonclassical class-I genes in different groups of mammals. *Molecular Biology and Evolution, 6*(6), 559-579.

Hughes, A.L., Ota, T., & Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class-I major-histocompatibility-complex molecules. *Molecular Biology and Evolution, 7*(6), 515-524.

Hurst, L.D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics, 18*(9), 486-487.

Jarvela, A.M.C., Brubaker, L., Vedenko, A., Gupta, A., Armitage, B.A., Bulyk, M.L., & Hinman, V.F. (2014). Modular evolution of DNA-binding preference of a Tbrain transcription factor provides a mechanism for modifying gene regulatory networks. *Molecular Biology and Evolution, 31*(10), 2672-2688.

Jojovic, M., Wolf, F., & Mangold, U. (1998). Epidermal growth factor, vascular endothelial growth factor and progesterone promote placental development in rat whole-embryo culture. *Anatomy and Embryology, 198*(2), 133-139.

Jin, Y.L., Turaev, D., Weinmaier, T., Rattei, T., & Makse, H.A. (2013). The evolutionary dynamics of protein-protein interaction networks inferred from the reconstruction of ancient networks. *Plos One, 8*(3), e58134.

Johnsson, E., Berggard, K., Kotarsky, H., Hellwage, J., Zipfel, P.F., Sjobring, U., & Landahl, G. (1998). Role of the hypervariable region in streptococcal M proteins: binding of a human complement inhibitor. *Journal of Immunology, 161*(9), 4894-4901.

Kamberov, Y.G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., . . . Sabeti, P.C. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell, 152*(4), 691-702.

Kersey, P.J., Allen, J.E., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., …, Staines, D.M. (2014). Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research, 42*(D1), D546-D552.

Kirshner, D.A., Nilmeier, J.P., Lightstone, F.C. (2013). Catalytic site identification-a web server to identify catalytic site structural matches throughout PDB. *Nucleic Acids Research, 41*(W1), W256-W265.

Kitajima, S., Koyama, T., Yamada, Y., & Sato, F. (1998). Constitutive expression of the neutral PR-5 (OLP, PR-5d) gee in roots and cultured cells of tobacco is mediated by ethylene-responsive cis-element AGCCGCC sequences. *Plant Cell Reports, 18,* 173-179.

Kobayashi, A., & Behringer, R.R. (2003). Developmental genetics of the female reproductive tract in mammals. *Nature Reviews Genetics, 4*, 969-980.

Koiwa, H., Kato, H., Nakatsu, T., Oda, J., Yamada, Y., & Sato, F. (1997). Purification and characterization of tobacco pathogenesis-related protein PR-5d, an antifungal thaumatin-like protein. *Plant Cell Physiology, 38*(7), 783-791.

Koiwa, H., Kato, H., Nakatsu, T., Oda, J., Yamada, Y., & Sato, F. (1999). Crystal structure of tobacco PR-5d protein at 1.8 angstroms resolution reveals a conserved acidic cleft structure in antifungal thaumatin-like proteins. *Journal of Molecular Biology, 286*(4), 1137-1145.

Koyama, T., Kitajima, S., & Sato, F. (2001). Expression of PR-5d and ERF genes in cultured tobacco cells and their NaCl stress-response. *Bioscience Biotechnology and Biochemistry*, *65*(5), 1270-1273.

Kuboyama, T. (1998). A novel thaumatin-like protein gene of tobacco is specifically expressed in the transmitting tissue of stigma and style. *Sexual Plant Reproduction, 11*(5), 251-256.

Kyte, J., & Doolittle, R.F. (1983). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*(1), 105-132.

Lamesch, P., Berardini, T.Z., Li, D.H., Swarbreck, D., Wilks, C., Sasidharan, R., . . . Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research, 40*(D1), D1202-D1210.

Lange, A., Wickstrom, S.A., Jakobson, M., Zent, R., Sainio, K., & Fassler, R. (2009). Integrin-linked kinase is an adaptor with essential functions during mouse development. *Nature, 461*, 1002-1006.

Lannergard, J., Gustafsson, M.C.U., Waldemarsson, J., Norrby-Teglund, A., Stalhammar-Carlemalm, S., & Lindahl, G. (2011). The hypervariable region of *Steptococcus pyogenes* M protein escapes antibody attack by antigenic variation and weak immunogenicity. *Cell Host & Microbe, 10*(2), 147-157.

Latijnhouwers, M., de Wit, P.J., & Govers, F. (2003). Oomycetes and fungi: similar weaponry to attack plants. *Trends in Microbiology, 11*, 462-469.

Lee, S., & Blundell, T.L. (2009). BIPA: a database for protein-nucleic acid interaction in 3D structures. *Bioinformatics, 25*(12), 1559-1560.

Leonardo, E.D., Hinck, L., Masu, M., Keino-Masu, K., Ackerman, S.L., & Tessier-Lavigne, M. (1997). Vertebrate homologues of *C. elegans* UNC-5 are candidate netrin receptors. *Nature*, *386*(6627), 833-838.

Lipman, D.J., & Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science, 227*(4693), 1435-1441.

Liu, H., Chang, L.H., Sun, Y., Lu, X., & Stubbs, L. (2014). Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome Biology and Evolution, 6*(3), 510-525.

Luo, Z.X., Yuan, C.X., Meng, Q.J., & Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature*, *476*(7361), 442-445.

Luscombe, N.M., & Thornton, J.M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology, 320*(5), 991-1009.

Lynch, V.J., & Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. *Evolution, 62*(9), 2131-2154.

Manning, G., Plowman, G.D., Hunter, T., & Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences, 27*(10), 514-520.

McClellan, D.A. (2013). Directional Darwinian selection in proteins. *BMC Bioinformatics, 14,* S6.

McCloskey, C., Rada, C., Bailey, E., McCavera, S., van den Berg, H.A., Atia, J., . . . Blanks, A.M. (2014). The inwardly rectifying K^+ channel KIR7.1 controls uterine excitability throughout pregnancy. *EMBO Molecular Medicine, 6*(9), 1161-1174.

McGuire, G., Denham, M.C., & Balding, D.J. (2001). Models of sequence evolution for DNA sequences containing gaps. *Molecular Biology and Evolution, 18*(4), 481-490.

Mecklenburg, L., Tychsen, B., & Paus, R. (2005). Learning from nudity: lessons from the nude phenotype. *Experimental Dermatology, 14*(11), 797-810.

Minaka, N., Suemara, T., Okano, K., Sugiura, N., Yamamoto, H., & Machii, K. (2008). Ancestral character-state reconstruction and its applications using BALANCE, an integrated software for calculating large phylogenies under the maximum parsimony criterion. *Cladistics, 24*(1), 98-99.

Mintseris, J., & Weng, Z.P. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *PNAS USA, 102*(31), 10930-10935.

Mitchell-Olds, T., Willis, J.H., & Goldstein, D.B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics, 8*(11), 845-856.

Mooers, A.O., & Schluter, D. (1999). Reconstructing ancestor states with maximum likelihood: support for one- and two-rate models. *Systematic Biology, 48*(3), 623-633.

Moon, T.W., & Hochachk, P.W. (1971). Temperature and enzyme activity in poikilotherms – isocitrate dehydrogenase in rainbow-trout liver. *Biochemical Journal, 123*(5), 695-705.

Moretti, S., Laurenczy, B., Gharib, W.H., Castella, B., Kuzniar, A., Schabauer, H., . . . Robinson-Rechavi, M. (2014). Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Research, 42,* D917-D921.

Mukherjee, K., & Burglin, T. (2007). Comprehensive analysis of animal TALE homeobox genes: new conserved motifs and cases of accelerated evolution. *Journal of Molecular Evolution, 65,* 137-153.

Nakamura, Y., Ichinohe, M., Hirata, M., Matsuura, H., Fujiwara, T., Igarashi, T., . . . Fukami, K. (2008). Phospholipase C-delta1 is an essential molecule downstream of Foxn1, the gene responsible for the nude mutation, in normal hair development. *FASEB Journal, 22*(3), 841-849.

Nakayam, J., Rice, J.C., Strahl, B.D., Allis, C.D., & Grewal, S.I.S. (2001). Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science, 292*(5514), 110-113.

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution, 3*(5), 418-426.

Nickel, G.C., Tefft, D.L., Goglin, K., & Adams, M.D. (2008). An empirical test for branch-specific positive selection. *Genetics, 179,* 2183-2193.

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics, 148*(3), 929-936.

Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., . . . Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife, 4*, e04837.

Nozawa, M., Suzuki, Y., & Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *PNAS USA, 106*(16), 6700-6705.

Ohno, S. (1970). *Evolution by gene duplication*. Berlin: Springer-Verlag.

Olmstead, R.G., & Bohs, L. (2007). A summary of molecular systematic research in *Solanaceae*. *Acta Horticulturae, 745*, 255-268.

Omland, K.E. (1999). The assumptions and challenges of ancestral state reconstructions. *Systematic Biology, 48*(3), 604-611.

Otto, E.A., Schermer, B., Obara, T., O'Toole, J.F., Hiller, K.S., Mueller, A.M., . . . Hildebrandt, F. (2013). Mutations in INVS encoding inversin cause nephronophthisis type 2, linking renal cystic disease to the function of primary cilia and left-right axis determination. *Nature Genetics, 34*(4), 413-420.

Pagel, M. (1999). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology, 48*(3), 612-622.

Palaisa, K., Morgante, M., Tingey, S., & Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *PNAS USA, 101*(26), 9885-9890.

Peter, I.S., & Davidson, E.H. (2011). Evolution of gene regulatory networks controlling body plan development. *Cell, 144*(6), 970-985.

Porter, C.T., Bartlett, G.J., & Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified enzymes using structural data. *Nucleic Acids Research, 32,* D129-D133.

Pougach, K., Voet, A., Fyodor, A.K., Voordeckers, K., Christiaens, J.F., Baying, B., . . . Verstrepen, K.J. (2014). Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Nature Communications, 5,* 4868.

Proux, E., Studer, R.A., Moretti, S., & Robinson-Rechavi. (2009). Selectome: a database of positive selection. *Nucleic Acids Research, 37,* D404-D407.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M., Yeh, S.D., . . . Carroll, S.B. (2006). Repeated morphological evolution through cis-regulatory changes in pleiotropic gene. *Nature, 440*(7087), 1050-1053.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen F., Astashyn, A., Ermolaeva, O., . . . Ostell, J.M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research, 42*(D1), D756-D763.

Pupko, T., Pe'er, I., Shamir, R., Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino-acid sequences. *Molecular Biology and Evolution, 17*(6), 890-896.

Reilly, S.M., & White, T.D. (2003). Hypaxial motor patterns and the function of epipubic bones in primitive mammals. *Science, 299*(5605), 400-402.

Rieseberg, L.H., Widmer, A., Arntz, A.M., & Burke, D.B. (2002). Directional selection is the primary cause of phenotypic diversification. *PNAS, 99,* 12242-12245.

Rook, D.L., & Hunter, J.P. (2014). Rooting around the eutherian family tree: the origin and relations of the Taeniodonta. *Journal of Mammalian Evolution, 21*(1), 75-91.

Ruiz, R.A.C., Herrera, C., Ghislain, M., & Gebhardt, C. (2005). Organization of phenylalanine ammonia lyase (PAL), acidic PR-5 and osmotin-like (OSM) defence-response gene families in the potato genome. *Molecular Genetics and Genomics, 274*(2), 168-179.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z. Richter, D.J., Schaffner, S.F., . . . Lander, E.S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature, 419*(6909), 832-837.

Savickiene, J., Treigyte, G., Baronaite, S., Valiuliene, G., Kaupinis, A. Valius, M., . . . Navakauskiene, R. (2015). Human amniotic fluid mesenchymal stem cells from second- and third-trimester amniocentesis: differentiation potential, molecular signature, and proteome analysis. *Stem Cells International, 2015,* 319238.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., . . . Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 37,* D5-D15.

Schmid, K., & Yang, Z. (2008). The trouble with sliding windows and the selective pressure in BRCA1. *Plos One, 3*(11), e3746.

Schmidt, D., & Durrett, R. (2004). Adaptive evolution drives the diversification of zinc-finger binding domains. *Molecular Biology and Evolution, 21*(12), 2326-2339.

Segal, D.J., Crotty, J.W., Bhakta, M.S., Barbas, C.F., & Horton, N.C. (2006). Structure of Aart, a designed six-finger zinc finger peptide, bound to DNA. *Journal of Molecular Biology, 363, 405-421.*

Siddiqui, K.S., & Cavicchioli, R. (2006). Cold-adapted enzymes. *Annual Review of Biochemistry, 75,* 403-433.

Siddiqui, K.S., Poljak, A., Guilhaus, M., De Francisci, D., Curmi, P.M.G., Feller, G., . . . Cavicchioli, R. (2006). Role of lysine versus arginine in enzyme cold-adaptation: modifying lysine to homo-arginine stabilizes the cold-adapted alpha-amylase from *Pseudoalteromonas haloplanktis. Proteins – Structure Function and Bioinformatics, 64*(2), 486-501.

Siggers, T., Reddy, J., Barron, B., & Bulyk, M.L. (2014). Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding. *Molecular Cell, 55*, 640-648.

Slatkin, M. (2008). Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics, 9*(6), 477-485.

Suzuki, Y. (2008). False-positive results obtained from the branch-site test of positive selection. *Genes and Genetic Systems, 83*(4), 331-338.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., . . . von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research, 39,* D561-D568.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123,* 585-595.

Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., . . . Gusella, J.F. (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell, 149*(3), 525-537.

Taylor, J.S., & Raes, J. (2004). Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics, 38,* 615-643.

Tong, C., Zhang, C.F., Shi, J.Q., Qi, H.F., Zhang, R.Y., Tang, Y.T., . . . Zhao, K. (2015). Characterization of two paralogous myostatin genes and evidence for positive selection in Tibet fish: *Gymnocypris przewalskii*. *Gene*, *565*(2), 201-210.

Toth-Petroczy, A., & Tawfik, D.S. (2011). Slow protein evolutionary rates are dictated by surface-core association. *PNAS USA*, *108*(27), 11151-11156.

Vallender, E.J., & Lahn, B.T. (2004). Positive selection on the human genome. *Human Molecular Genetics, 13*, R245-R254.

Vitti, J.J., Grossman, S.R., & Sabeti, P.C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics, 47,* 97-120.

Wagner, A. (2007). Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics, 176*(4), 2451-2463.

Wang, X., & Zhang, J. (2007). Rapid evolution of primate ESX1, an X-linked placenta- and testis-expressed homeobox gene. *Human Molecular Genetics*, *16*, 2053-2060.

Wallis, M. (2015). Coevolution of insulin-like growth factors, insulin and their receptors and binding proteins in New World Monkeys. *Growth Hormone & IGF Research, 25*(4), 158-167.

Williams, P.D., Pollock, D.D., Blackburne, B.P., & Goldstein, R.A. (2006). Assessing the accuracy of ancestral protein reconstruction methods. *Plos Computational Biology, 2*(6), 598-605.

Woloshuk, C.P., Meulenhoff, J.S., Sela-Buurlage, M., van den Elzen, P.J., & Cornelissen, B.J. (1991). Pathogen-induced proteins with inhibitory activity toward *Phytophthora infestans. Plant Cell, 3,* 619-628.

Wong, W.S.W., Yang, Z.H., Goldman, N., & Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics, 168, 1041-1051.*

Worth, C.L., Bickerton, G.R., Schreyer, A., Forman, J.R., Cheng, T.M., Lee, S., . . . Blundell, T.F. (2007). A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *Journal of Bioinformatics and Computational Biology, 5*(6), 1297-1318.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics, 8*(3), 206-216.

Yaish, M.W., Doxey, A.C., McConkey, B.J., Moffatt, B.A., & Griffith, M. (2006). Cold-active winter rye glucanases with ice-binding capacity. *Plant Physiology, 141*(4), 1459-1472.

Yang, J.N., An, J.F., Li, M., Hou, X., & Qiu, X.H. (2013). Characterization of chicken cytochrome P450 1A4 and 1A5: inter-paralog comparisons of substrate preference and inhibitor selectivity. *Comparative Biochemistry and Physiology: C-Toxicology & Pharmacology, 157*(4), 337-343.

Yang, Z. (1998). On the best evolutionary rate for phylogenetic analysis. *Systematic Biology, 47*(1), 125-133.

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

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

Yang, Z.H. (2007). PAML4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution, 24*(8), 1586-1591.

Ye, J., Pavlicek, A., Lunney, E.A., Rejto, P.A., & Teng, C.H. (2010). Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics, 11*, 11.

Yokoyama, S., Tada, T., Zhang, H., & Britt, L. (2008). Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *PNAS USA, 105*(36), 13480-13485.

Yu, J., Ke, T., Tehrim, S., Sun, F., Liao, B., & Hua, W. (2015). PTGBase: an integrated database to study tandem duplicated genes in plants. *The Journal of Biological Databases and Curation, 2015*, 1-10.

Zevenhuizen, L.P.T.M., Bartnicki-Garcia, S. (1969). Structure of the insoluble hyphal wall glucan of *Phytophthora cinnamoni*. *Biochemistry, 8,* 1496-1502.

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

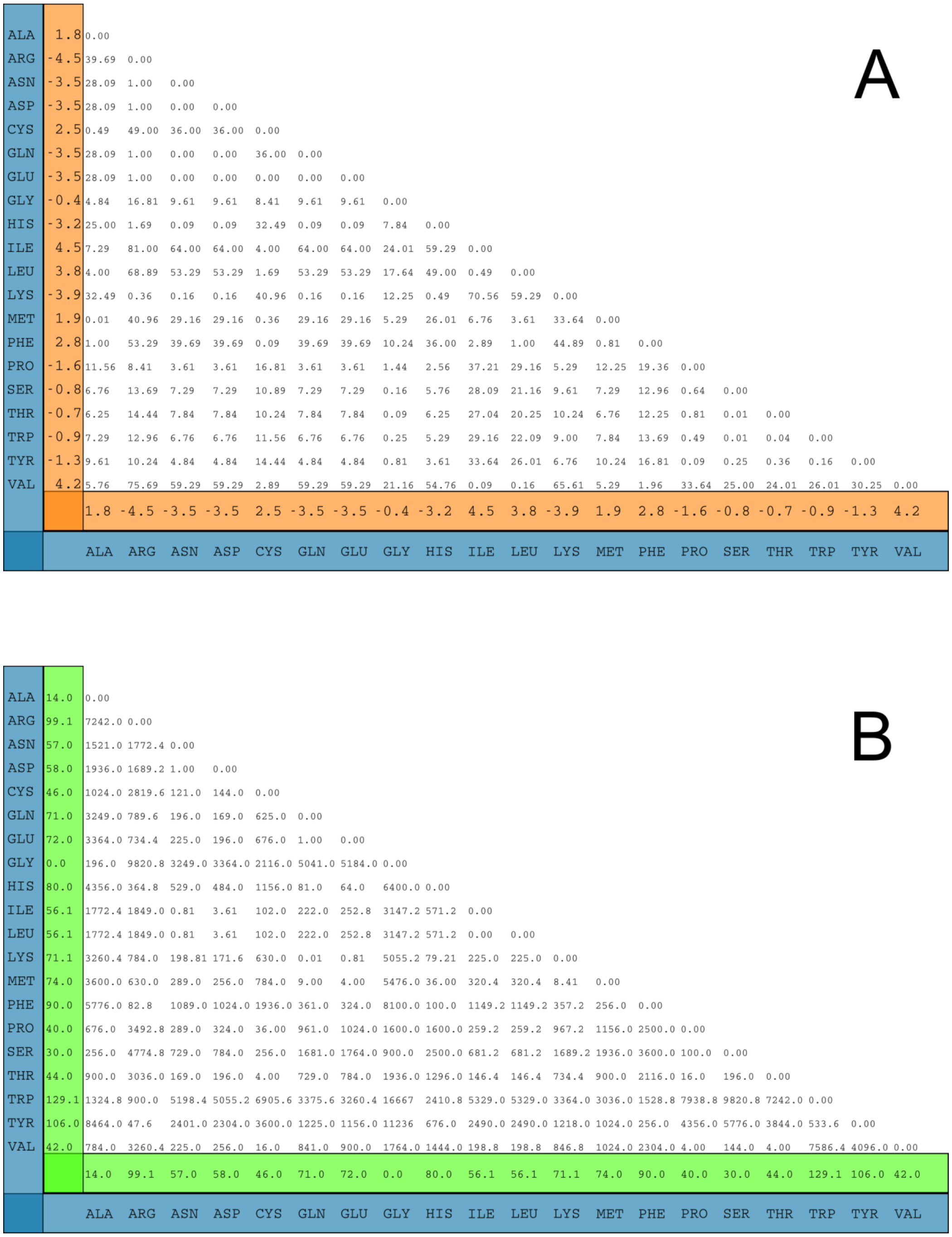
Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology & Evolution, 22*(12), 2472-2479.

Zhang, N., McCarthy, M.L., & Smart, C.A. (2008). A macroarray system for the detection of fungal and oomycete pathogens of solanaceous crops. *Plant Disease, 92*(6), 953-960.

Zhou, T., Enyeart, P.J., & Wilke, C.O. (2008). Detecting clusters of mutations. *Plos One, 3*(11), e3765.

Appendix A

Supplementary Figures



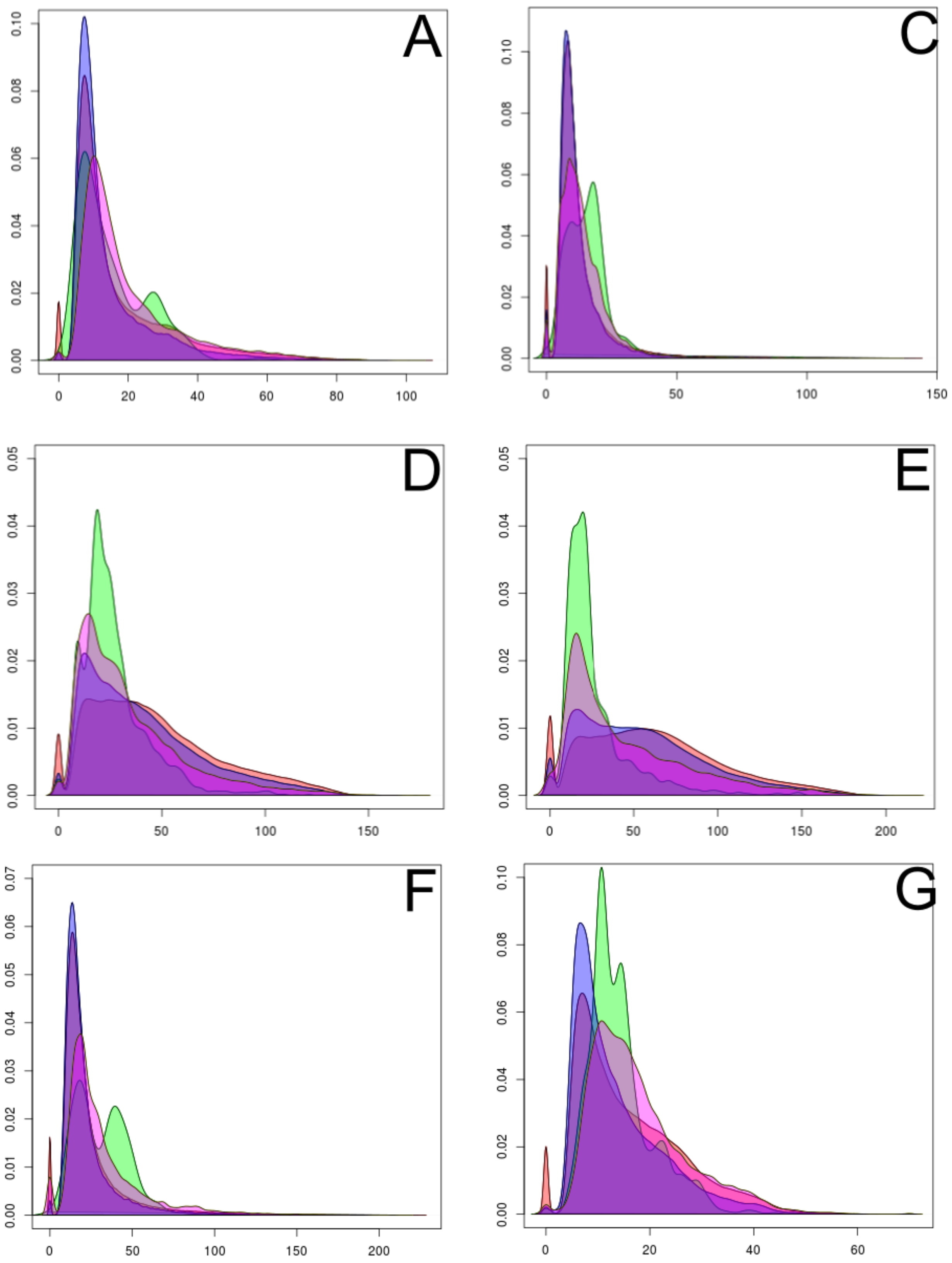
A

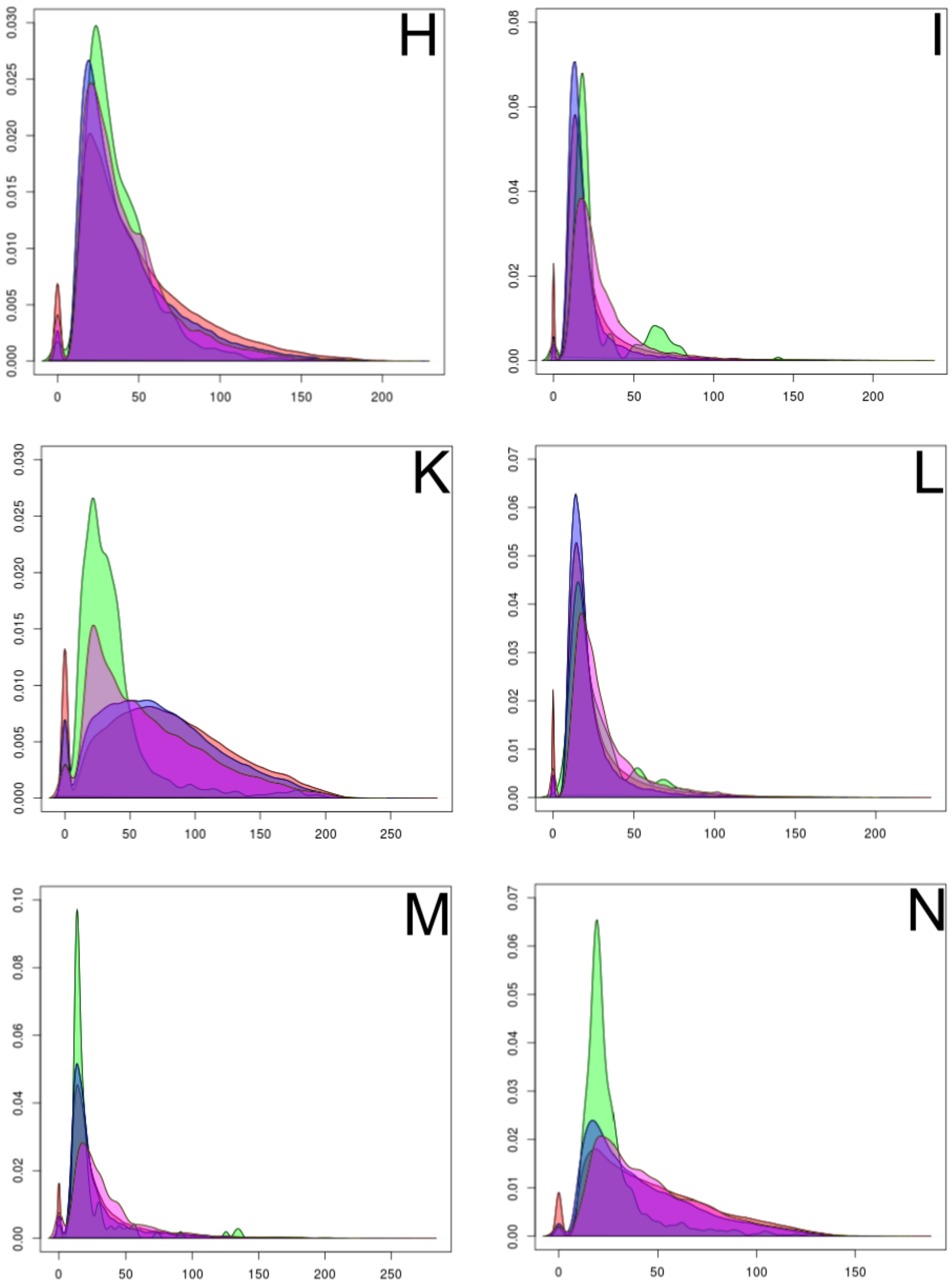
ALA	1.8	0.00																			
ARG	-4.5	39.69	0.00																		
ASN	-3.5	28.09	1.00	0.00																	
ASP	-3.5	28.09	1.00	0.00	0.00																
CYS	2.5	0.49	49.00	36.00	36.00	0.00															
GLN	-3.5	28.09	1.00	0.00	0.00	36.00	0.00														
GLU	-3.5	28.09	1.00	0.00	0.00	0.00	0.00	0.00													
GLY	-0.4	4.84	16.81	9.61	9.61	8.41	9.61	9.61	0.00												
HIS	-3.2	25.00	1.69	0.09	0.09	32.49	0.09	0.09	7.84	0.00											
ILE	4.5	7.29	81.00	64.00	64.00	4.00	64.00	64.00	24.01	59.29	0.00										
LEU	3.8	4.00	68.89	53.29	53.29	1.69	53.29	53.29	17.64	49.00	0.49	0.00									
LYS	-3.9	32.49	0.36	0.16	0.16	40.96	0.16	0.16	12.25	0.49	70.56	59.29	0.00								
MET	1.9	0.01	40.96	29.16	29.16	0.36	29.16	29.16	5.29	26.01	6.76	3.61	33.64	0.00							
PHE	2.8	1.00	53.29	39.69	39.69	0.09	39.69	39.69	10.24	36.00	2.89	1.00	44.89	0.81	0.00						
PRO	-1.6	11.56	8.41	3.61	3.61	16.81	3.61	3.61	1.44	2.56	37.21	29.16	5.29	12.25	19.36	0.00					
SER	-0.8	6.76	13.69	7.29	7.29	10.89	7.29	7.29	0.16	5.76	28.09	21.16	9.61	7.29	12.96	0.64	0.00				
THR	-0.7	6.25	14.44	7.84	7.84	10.24	7.84	7.84	0.09	6.25	27.04	20.25	10.24	6.76	12.25	0.81	0.01	0.00			
TRP	-0.9	7.29	12.96	6.76	6.76	11.56	6.76	6.76	0.25	5.29	29.16	22.09	9.00	7.84	13.69	0.49	0.01	0.04	0.00		
TYR	-1.3	9.61	10.24	4.84	4.84	14.44	4.84	4.84	0.81	3.61	33.64	26.01	6.76	10.24	16.81	0.09	0.25	0.36	0.16	0.00	
VAL	4.2	5.76	75.69	59.29	59.29	2.89	59.29	59.29	21.16	54.76	0.09	0.16	65.61	5.29	1.96	33.64	25.00	24.01	26.01	30.25	0.00
		1.8	-4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2
		ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL

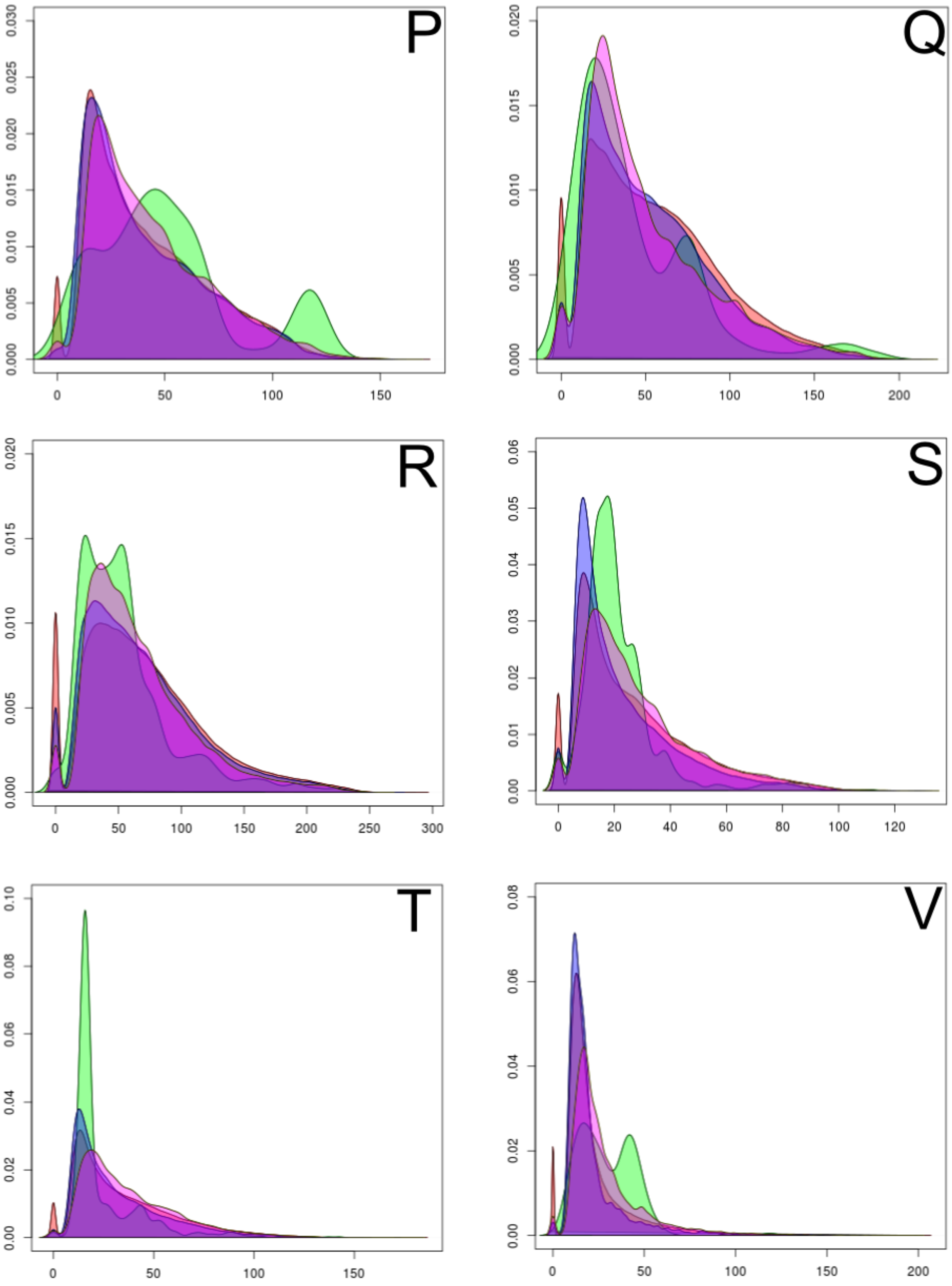
B

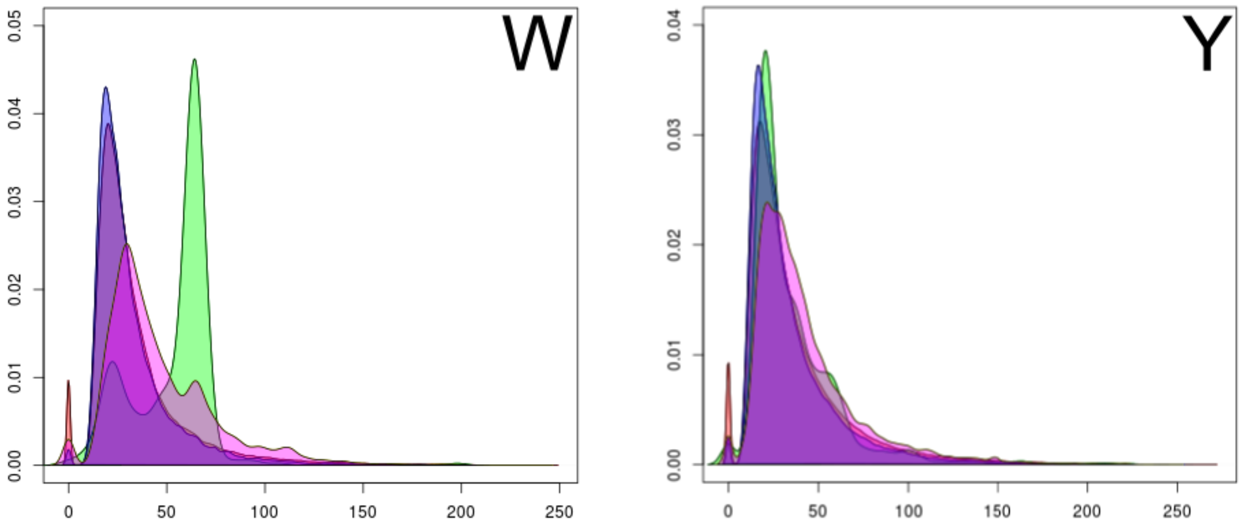
ALA	14.0	0.00																			
ARG	99.1	7242.0	0.00																		
ASN	57.0	1521.0	1772.4	0.00																	
ASP	58.0	1936.0	1689.2	1.00	0.00																
CYS	46.0	1024.0	2819.6	121.0	144.0	0.00															
GLN	71.0	3249.0	789.6	196.0	169.0	625.0	0.00														
GLU	72.0	3364.0	734.4	225.0	196.0	676.0	1.00	0.00													
GLY	0.0	196.0	9820.8	3249.0	3364.0	2116.0	5041.0	5184.0	0.00												
HIS	80.0	4356.0	364.8	529.0	484.0	1156.0	81.0	64.0	6400.0	0.00											
ILE	56.1	1772.4	1849.0	0.81	3.61	102.0	222.0	252.8	3147.2	571.2	0.00										
LEU	56.1	1772.4	1849.0	0.81	3.61	102.0	222.0	252.8	3147.2	571.2	0.00	0.00									
LYS	71.1	3260.4	784.0	198.81	171.6	630.0	0.01	0.81	5055.2	79.21	225.0	225.0	0.00								
MET	74.0	3600.0	630.0	289.0	256.0	784.0	9.00	4.00	5476.0	36.00	320.4	320.4	8.41	0.00							
PHE	90.0	5776.0	82.8	1089.0	1024.0	1936.0	361.0	324.0	8100.0	100.0	1149.2	1149.2	357.2	256.0	0.00						
PRO	40.0	676.0	3492.8	289.0	324.0	36.00	961.0	1024.0	1600.0	1600.0	259.2	259.2	967.2	1156.0	2500.0	0.00					
SER	30.0	256.0	4774.8	729.0	784.0	256.0	1681.0	1764.0	900.0	2500.0	681.2	681.2	1689.2	1936.0	3600.0	100.0	0.00				
THR	44.0	900.0	3036.0	169.0	196.0	4.00	729.0	784.0	1936.0	1296.0	146.4	146.4	734.4	900.0	2116.0	16.0	196.0	0.00			
TRP	129.1	1324.8	900.0	5198.4	5055.2	6905.6	3375.6	3260.4	16667	2410.8	5329.0	5329.0	3364.0	3036.0	1528.8	7938.8	9820.8	7242.0	0.00		
TYR	106.0	8464.0	47.6	2401.0	2304.0	3600.0	1225.0	1156.0	11236	676.0	2490.0	2490.0	1218.0	1024.0	256.0	4356.0	5776.0	3844.0	533.6	0.00	
VAL	42.0	784.0	3260.4	225.0	256.0	16.0	841.0	900.0	1764.0	1444.0	198.8	198.8	846.8	1024.0	2304.0	4.00	144.0	4.00	7586.4	4096.0	0.00
		14.0	99.1	57.0	58.0	46.0	71.0	72.0	0.0	80.0	56.1	56.1	71.1	74.0	90.0	40.0	30.0	44.0	129.1	106.0	42.0
		ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL

Supplementary Figure S1: Matrices displaying squared quantitative property changes for amino acid substitutions. A: squared hydropathy index changes; B: square residue side chain mass changes.


A
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
C
0.00
0.02
0.04
0.06
0.08
0.10
0
50
100
150
D
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
E
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
200
F
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50
100
150
200
G
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60


H
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
50
100
150
200
I
0.00
0.02
0.04
0.06
0.08
0
50
100
150
200
K
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
50
100
150
200
250
L
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50
100
150
200
M
0.00
0.02
0.04
0.06
0.08
0.10
0
50
100
150
200
250
N
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
50
100
150


P
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
50
100
150
Q
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
R
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
250
300
S
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
20
40
60
80
100
120
T
0.00
0.02
0.04
0.06
0.08
0.10
0
50
100
150
V
0.00
0.02
0.04
0.06
0.08
0
50
100
150
200



Supplementary Figure S2: Density curves for side chain solvent accessible surface area for all residue types from different databases. Observed solvent accessible surface values in squared angstroms are on the x-axis. Frequency is displayed on the y-axis. Databases. red: PDB; green; CSA; blue: CSI; yellow: iPFam. Plots are labelled according to amino acid one-letter codes.