Spatial clustering of differences in measured homoplasy with respect to protein structure

by

Kent Allan Vander Velden

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Gavin Naylor, Co-major Professor
Vasant Honavar, Co-major Professor
Irvin Hentzel
Drena Dobbs

Iowa State University

Ames, Iowa

2002

Graduate College
Iowa State University

This is to certify that the master's thesis of

Kent Allan Vander Velden

has met the thesis requirements of Iowa State University

_____
Committee Member

_____
Committee Member

_____
Co-major Professor

_____
Co-major Professor

_____
For the Major Program

Table of Contents

List of Equations

List of Figures

List of Tables

Abstract

The identification of sites in amino acid sequence alignments that hold misleading phylogenetic signals and the identification of amino acid residues that are of functional significance are intertwined. Advances in one area can support the other because misleading phylogenetic signals come from the comparison of residues from sites in alignments that are not evolving as an unconstrained random process. This is a study of the distribution of misleading phylogenetic signals contained within five proteins and identified through comparing a widely accepted phylogenetic tree to those inferred from sequence data. Through the analysis of these distributions one goal is the discovery of properties that can be used to improve the inference of phylogenetic trees, but another goal is the identification of functionally important residues. A new metric, RI Difference and based on Retention Index, is suggested measuring the relative support that individual sites provide for two trees. By identifying sites that harbor misleading phylogenetic signal, we attempt to identify residues that are cooperating to define the function of the protein. This information is presented in the context of the structure of the protein where spatial clustering patterns (or lack of) are observed for the implicated residues. A new bioinformatic software tool, RI Compare, is presented implementing the metric and blending heterogeneous information from protein alignments and structures and phylogenetic trees. Results are presented followed by speculations as to what might be causing erroneous trees to be inferred. The relationship of the implicated residues to those of known importance is also discussed. While results do not suggest that the RI Difference measure can be used to identify functionally important residues in all proteins, there is evidence to suggest it may be applicable to transmembrane proteins. Assessment of the correctness of the results has been based solely on the proximity of the implicated residues to ligands, other chains, and residues of known importance. However, even if the RI Difference measure is identifying residues other than the functional significant ones, the fact that the cluster patterns are unlikely to occur at random is intriguing and warrants further investigation.

Chapter 1. Introduction

Project Goal and Document Overview

The goal of this study was initially to improve phylogenetic inference procedures through the identification of collections of residues in an alignment that did not conform to the model of evolution being used. This goal was later expanded to searching for functionally significant residues after considering reasons that alignment sites were misleading. What follows is an introduction to the concepts required for understanding and interpreting the results of this study including reviews of recent relevant material. The introduction is followed by descriptions of the methods used and an exhaustive presentation of the actual results. The interpretation of these results is followed by concluding remarks and suggestions for future researchers in this area.

Genomic Data Background

During the past decade the world has witnessed an explosion in the development of methods and hardware for the collection and analysis of genomic sequence and related data. These developments have come from both public and private labs, often working in cooperation as much as in competition, while captivating the imagination of the public.

The first genome to be completely sequenced was of the prokaryote bacterium *Hemeophilus influenzae* in 1995, published in Science with a list of 40 authors. This accomplishment was soon followed during the same year with another genome from the prokaryote *Mycoplasma genitalium*. During the following year, the first sequence from a member of the archaeae family, *Methanococcus jannaschii*, and the first genome of a eukaryote, *Saccharomyces cerevisiae* (also known as baker's or budding yeast) were completed. While the sequencing race was only getting started an example was now available from a representative of each of the three major lineages of life. Several additional model organisms were sequenced in the following years, complementing years of previous knowledge, including *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*. Then in June 2000, a working draft of the largest genome to date, and one with a special significance, the human genome, was completed.

Proponents of whole genome sequencing projects have touted their potential for curing diseases, increasing food production, and genetic engineering. The implied goal is better understanding of life and the ability to influence its destiny. While the goals of the coming genetic revolution have been popularized by both the mainstream media and the scientific community, both have done a disservice to people outside of the discipline. Even those within the discipline sometime lose connection with the limits of our current knowledge and abilities and also the direction the field is moving. While it is true that a great deal of information has been amassed during the past few years and contributed to the realm of biology, perhaps more raw information than has been gathered for several decades prior, what have been the benefits?

Sequence data by itself is as useful as a book in a language that the reader does not understand, and at the moment we are at our infancy in out understanding of an organism's genetic language. It is impossible to gauge the thoroughness of our understanding of biology since we do not know its depths, but we are likely far from the understanding necessary to engineer a biological entity with a specified function. Furthermore, unifying principles have been very sparse which only adds to the apparent complexity.

There have been comparisons drawn between different disciplines of science. The comparisons have been built on how successfully the area can be analyzed. There are sciences in which systems can be broken down into finer and finer partitions, where each system assumes the collective function of its components in an additive fashion. These sciences allow a person to examine the finest scale building block where understanding may be easiest and reassemble the system to yield a complete understanding of the whole. This is a very accessible method for a person to tease apart the governing laws of a system. After all, surely the smaller item, being a building block of the larger, will be easier to understand. However, there are sciences where this method of dissection fails where analysis of the complete system at progressively finer scales complicates the problem to the point of becoming intractable. While this progressive dissection approach may be appealing, one must consider if an appropriate scale is being used. It can be more efficient to conduct such an analysis at the level of whole objects or at some other scale that is reasonable for the problem. Biological problems have been approached recently by continuing to break them down into finer and finer scales until today we have the governing sequence itself. However, this may not be the best level to address all problems. We must be careful to choose the correct scale. As we shall see, properties of proteins which are not apparent at the sequence level can be revealed when complementary information, such as the structure, is considered. Further information can be extracted when the evolutionary relationships are

considered. In this case, understanding, which was not forthcoming at the sequence scale alone, comes from an analysis at multiple scales.

What we have done for the past decade is listen to our own desires to have better lives through the understanding of what makes life function so it can be modified to better accommodate us. Perhaps there was also a need to find proof of our own uniqueness and separation from other species. This has led to the sequencing of many organisms including ourselves, but during the rush it seems we have overlooked the question: what do we do with it all?

The question that should have been asked at the beginning of the sequencing process has waited until the crown jewel was claimed in the form of the completion of the Human Genome Project. Sequencing continues with the belief that the sequence level is the correct scale to be analyzing biological systems and with the hope that future scientists will be able to decipher the data. The collection of data for future generations of scientists is an accomplishment with merit as long as the information is not being collecting blindly.

Scientists are aware of the need to carefully select what information to collect. While there are a few groups that appear to simply collect data without an apparent guiding goal, most choose experiments that compliment others or existing knowledge. Along with sequence data, also collected have been expression, kinetic and structural data. The complimentary nature of expression, kinetic, and structural data often provide a richer means to gain additional understanding. Sequence data provides the raw genomic data and some notion of the variation present in alleles. Expression data provides an indication of the degree to which genes are expressed under varied conditions. Protein kinetic and structural data provide information at the level of molecular interaction within a cell. All these data complement each other and lead to a more thorough understanding and new discoveries.

Sequence data has been collected to aid in the understanding of life processes. The fact that biology is one of the oldest sciences but seems to have an ever growing list of questions to answer suggests that life is highly complex. The paramount charge of biology is the understanding of the organization and function of organisms. We also suspect that genetic material, which is unique to each individual and popularized as “the blueprint of life,” controls the organization and function of that individual. By sequencing the genome we expose the mechanics of the organism, to varied extents, in all three major time frames – past, present, and future. Remnants of past infections, genes that have lost their functions, and other wide scale genomic alterations are still present in genomes today partially hidden by mutations which have

built up over millions of years and provide historical information. Functional genes along with knowledge of what causes the machinery of the cell to activate them give information of the state of the system as it is today. Finally, future insight may be gained through experiments that "improve" genes using directed and random mutations. It may even be possible to predict the effects of various environmental agents since detailed knowledge of the operations of the genome brings understanding of how the organism will respond to the stresses of any given environment.

The traditional method that has been used to further the understanding of an organism, at least at the scale that molecular biology explores, has been to analyze a single protein in depth for an extended period and often in isolation from other proteins that may exist in the biological system. It is not uncommon for a molecular biologist to spend their entire lifetime studying a single protein perhaps even at the end not understanding this particular protein completely. And this is just one of possibly thousands of proteins in a single cell! This situation is further complicated by the fact that proteins rarely, if ever, are autonomous, but instead work in tandem with other proteins. This apparent slow progress should not be taken to reflect a lack of effort on the part of the researcher, but instead should be testament to the complexity of even a tiny part of the cellular system.

Today even more data is being amassed. Sequence data has not been enough to crack the complete machinery of life, so now researchers are turning to expression data for both mRNA and proteins, structural and kinetic data of proteins and interactions, tissue specific, and organismal specific data. In an effort to better understand life processes. Thus far however, these additional sources of information have not clarified the picture. Indeed, in most cases they have made it hazier. While these types of data have been available in the past, the rate at which they are currently being gathered has quickly outpaced our capacity to interpret them.

This synopsis may seem to present a fairly disillusioned view of the field. However, the intent has been to summarize the forces that have brought us up to this point and to remind us of certain limitations. The genomic era has only begun, and while there have already been successes, the most ambitious predictions remain a long way off. But perhaps, even if a complete understanding of life is not forthcoming, the rewards will be large enough.

Evolution Background

Disagreement has existed since Darwin proposed his theory of evolution between those that sought to disprove it completely, those that favored not applying the theory to humans, and those who have viewed it as a universal law. The debate has been based in scientific fact as well as in faith, and the genome wide organism comparison projects will likely provide further fuel for this debate. The sequence data provides a quantitative indication of our relationship to other organisms. Not only are we very closely related to the apes, but also to a decreasing degree, we are related to rodents, to plants, and even to microbes. While not proof of evolution, these relationships certainly lend credence to the theory.

Whether one accepts the concept of evolution or not, mutations continue to take place in the genomes of organisms as the result of failure of copying and proofreading mechanisms and environmental factors. These mutations compound and provide variation in the genes. By examining variation of a single gene within a species it is possible to rank the variability of different parts of the gene's sequence. It can be hypothesized that regions that are invariable are of critical importance while variable regions are "placeholders" between the critical regions. This idea is supported by the observation that if all regions had equal importance then substitutions would be randomly spread out over the gene. However, if a segment is so important that if it was altered the protein's function would be compromised, this area would show small amounts of variation. The failure of a critical protein could be lethal or result in an organism being severely impaired. Because of the importance of these areas, they are seen as constants in the genes of the living organisms. It is not so much that those sites never mutate, it is only that those variations are never seen because the resulting organism is not viable.

Comparing alleles of a gene from different individuals in a single species can provide important information, especially for a gene that mutates quickly. However, for slowly evolving genes this method may never provide an adequate observation of variation. Even with a fast changing gene, by restricting our view to a single species a huge number of individuals may need to be examined before the variation becomes evident. Thankfully, the diversity of life can help with this problem.

While different species may appear outwardly unique, many of their internal processes are highly similar. The human genome has been suggested to be 98% identical to chimpanzee, 85% similar to mouse, and even 75% similar to *C. elegans*. However, such measures of similarity can be misleading. For example, the human-chimpanzee comparison is a projection based upon about

40 genes sequenced from both species and the relations extrapolated to the entire genome (the genome of the chimpanzee has not yet been sequenced). Even when the genomes of both organisms are available they can not be directly compared because of rearrangement, unique genes, different numbers of chromosomes, *etc.* And while there are still problems with this comparison method, such as questions regarding correct correspondence of genes, various studies have yielded similar values.

The reuse of the same components of life among different species allows researchers to explore a much richer source of variation. The diversity of life removes the need to gather sequence information from a large number of individuals of a species in the hope of finding unique alleles. The amount of evolutionary time that separates different species increases the chances of finding variations in gene sequences and perhaps functions. Sampling additional species, especially those with more distant relationships, should help to show more variation because of independently accumulated mutations.

Protein Background

Many of the components being reused between biological systems occur at the level of proteins. Proteins are remarkable biological entities. They are composed of only twenty different amino acid residues that connect to form strings that fold into beautiful three-dimensional structures. Occasionally, additional molecules are associated with structures such as ligands or metal ions. People, familiar only with mechanical devices, may be surprised to learn that proteins perform similar tasks of movement and alteration of other components. Proteins can also communicate information throughout a cell and between cells within an organism and even between organisms. Often tasks are performed by numerous individual proteins acting in concert. There are even examples of proteins that have the ability to perform several different functions.

The amazing diversity and methods used by proteins to run life's machinery has captivated the attention of many scientists. While the scientists' primary drive is for understanding of the biological system, there are often practical spin offs that can emerge from unlocking the secrets of proteins. Examples include increased food production and elimination of genetic diseases. To tease apart the protein's secrets, scientists have sought to identify specific amino acid residues critical for its function. This endeavor relies on the assumption that most of a protein's residues

only provide a structural scaffold for the critical functional residues. While the shape of the protein is important for function, it has already been constructed. If the initial goal is not to design a protein with unique function *de novo*, but to improve upon an already present function, it is likely sufficient to focus on only certain residues.

Current approaches for elucidating the importance of specific residues in proteins typically exploit information gleaned from analysis of at least two of the following: protein sequence, protein structure, and evolutionary information. These methods exist on a gradient between those based solely on experimental evidence and those based solely on computational information. While once the norm was to use solely experimental evidence, a shift has occurred recently to more computational methods. A summary of much of the work that has been done in this area is available in Todd *et al.* (2001). The automated search for functionally significant residues is expanding as researchers seek to discover new functions of proteins on an organismal level. These methods move us further along the continuum towards pure computational methods (Teichmann *et al.* 2001, Aloy *et al.* 2001, Elcock 2001).

The development of techniques to identify functionally significant residues *in silico* has only begun. While techniques are still fairly analysis intensive, they are less time consuming than unassisted lab work. In later sections we present and test a tool developed as part of this study to aid in this research. While the tool is far from perfect, hopefully the unique method can aid the advancement of the field by providing an alternative view.

Alignments and Phylogenetic Trees

Corresponding genes from different species often have corresponding sites that are a consequence of the genes being inherited from a common ancestor. This correspondence allows the genes to be aligned. The traditional view of an alignment is to have the sequence for a particular organism listed left to right in rows where corresponding sites are numbered columns called sites. Not all genes have correspondence at all sites as a result of a loss or introduction of a site over the course of evolution. In these cases, gaps must be inserted to act as placeholders. In extreme cases, sequence data may not show any correspondence and one must align the sequences with the help of the structure since structure is assumed to be more conserved than sequence.

Examining sequence data in the form of an alignment can be misleading. E.g. we may find there are only two residues represented at a particular site. Initially, this would appear to be a site of

low variation and possibly represent a functionally significant site. Alternatively, this could be the result of a slow mutation rate or representative of low sequence diversity. However, another explanation may come from examining the proposed evolutionary history in the form of a tree. Two hypothesized extreme cases are discussed below.

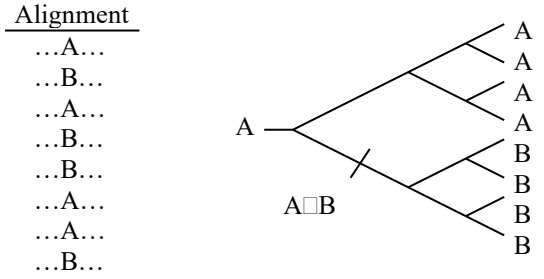


Figure 1. Alignment and Tree of Low Variation

One possible explanation is that all this variation resulted from a single change. If we assume that the ancestor of all eight organisms had a character ‘A’ at this site, we can see that all the variation can be explained with a single change along one of the main lineages. This particular tree is called the most parsimonious one for this dataset since it explains the data with the least number of changes. Notice, at least with this dataset, the original assumption is not rigid. The ancestral state could have been a ‘B’ but the main part of the argument would not change. Since only a single change is present, we can say that this site exhibits low variation.

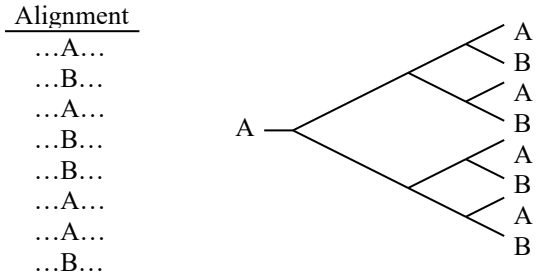


Figure 2. Alignment and Tree of High Variation

An alternative possibility is high variation, which is shown in figure 2. Again, we can assume without loss of generality that the ancestor had a character ‘A’ at this position. However, this is an example of the greatest amount of variation possible with all the observed change taking place

close to the tips of the tree. Additional changes could have occurred earlier in the tree and likely did if this site is as free to vary as it appears to be. However, without some knowledge of the ancestral states it would be impossible to place a likelihood measure on these intermediate state assignments.

Alternatively, we could be searching for residues that are covarying with each other. The idea here is that the change of one residue may disrupt the function of the protein, but that change can be compensated for by a simultaneous change at the covarying site (Kim *et al.* 1994). While a very useful and potentially informative discovery, again, when we restrict our view to only the alignment we see how we may be misled.

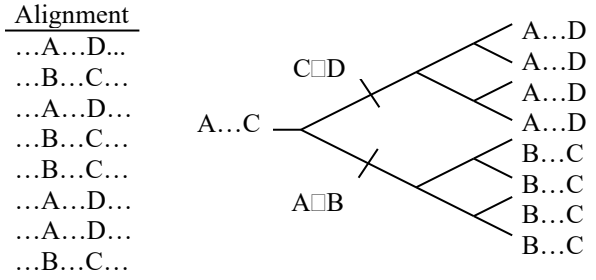


Figure 3. Alignment and Tree of Covariance

In the above alignment example whenever an ‘A’ occurs at the first site a ‘D’ occurs at the second site. Further, whenever a ‘B’ occurs at the first site a ‘C’ always occurs at the second site. If we submit this observation to a statistical test we would find this observation to be very unlikely to occur at random and so must be very significant. However, by examining the evolutionary relationships of the organisms we see that the apparent covariation is not that spectacular. There were two changes from the ancestral state along different lineages of the tree and these sites have not changed since then. Of course, with all these examples the tree needs to be correct, since our explanation rests on the accuracy of the tree.

We have examined where trees may be used to provide alternative explanations to patterns in alignments that initially appear to be highly significant. But where do these trees come from? All trees are inferred, but the source and quality of the data from which trees are being inferred varies. Some choices of characters that have been used include morphological characters such as skeletal and anatomical structures and even the calls of songbirds. Others have included more modern molecular sources such as DNA. Another source is the comparative anatomy and stratigraphic record provided by fossils. With the help of geologists, paleontologists can deduce

the age of a fossil by its depth in the sedimentary rock layers and radiometric dating. By finding several examples of a fossil, boundaries can be placed on the species' existence in history. By comparison of these relative boundaries in conjunction with an analysis of morphological characters a tree of life can be built that is fairly non-controversial. The tree constructed from traditional comparative anatomical data in conjunction with the stratigraphic range will be considered to be the correct or true tree throughout this paper, and will be the tree to which all others are compared. There can be no certainty that these trees are correct. Nevertheless, there is broad acceptance of these trees and which are based on a variety of disciplines.

The inference of trees from sequence data makes three major assumptions about the characteristics of the data, which are required properties of data to yield the correct tree. The first assumption is that all sites are independent, meaning that a change in any site has no affect on any other site. The second assumption is that the frequency of each residue type is equally represented across species. The third assumption is that the probability for mutation is equal regardless of the residue type. All these assumptions have a further restriction that they be present in the extant sequences (those at the tips of the tree and those for which sequence data is available) as well as throughout the lineages of the tree. Unfortunately, the data often violate these assumptions of the inference methods and suggest an incorrect tree as a result.

Studies addressing covariation of residues in proteins have provided examples of where the mutation of a single residue disrupts the function but a mutation in another residue restores the function (Korber *et al.* 1993, Rongey *et al.* 1993, Kim *et al.* 1994). This relationship strongly suggests that these residues are working in tandem and thus are not independent. The affect that such covariation has on the inference of phylogenic trees is unequal weighting of the contribution of sites. Inference methods take the collective suggestions made by each site and construct a tree that summarizes these suggestions. If sites are covarying they are dependent to the point that knowledge of either site can be used to determine the state of the covarying site. The level to which the determination can be made is a measure of the possible covariance (Shannon 1948, Clarke 1995, Lapedes *et al.* 1997). Maximum covariance would allow a perfect prediction of either site given the other and so this particular suggestion is being made multiple times and being overemphasized.

If sequences have uneven base compositions (unequal occurrence frequencies) there is a tendency to favor some residue types over others. The affect that this has on inference procedures is that sequences with similar base compositions have a greater probability of appearing similar and thus will be clustered nearer to each other and may not be representative of

the correct phylogeny. It is possible that a change has occurred causing an entire group of organisms to have similar base compositions. This will not harm the overall inference, since the group should be clustered together anyway, but may affect the resolution or the accuracy within the group. However, if distant organisms have similar base compositions they may be clustered erroneously.

Characteristics of different species may give them different mutation rates. Consider the birthrates of organisms with the understanding that a lot of genetic variation is introduced by the young of the species through recombination events at conception. Two distant species that have high mutation rates may be clustered together by mistake after all of the differences that separated these sequences have been obliterated by a fast mutation rate. Unfortunately, inference procedures are often susceptible to failures caused by assuming the substitution rate is constant.

An implied assumption suggested by the need for the mutation rate to be similar is that changes should be irreversible. While irreversibility of state would be ideal for inference procedures, it certainly does not reflect nature. When examining amino acid sequences there are only 20 characters to choose from and nucleic sequences only have four. Even if each change at a site yielded a different character there is an inherent limit on the number of changes a site could accommodate before information is lost by the repeat of a previous character.

Another problem referred to as “among-site-rate-variation” addresses the observation that the mutation rate is not even constant across the single sequence of an organism. There are hotspots along a sequence where mutations occur rapidly while other areas have little to no visible evidence of mutation. Observation of these constant sites can be very useful in determining the critical residues responsible for the function of the protein, but are uninformative to phylogenetic inference methods.

The among-site-rate-variation problem was the first one that we have examined that suggests that the assumptions can be broken on a site basis (other than perhaps the site independence assumption). The assumptions have been described at the level of entire sequences, but more generally violations occur to varied degrees on a site-by-site basis. Mutation rates, base compositions, *etc.* can all be varied at different sites.

Not all inference methods make all these assumptions. Nor are the effects that the violation of these assumptions may have on a particular method equal. There have been methods designed to explicitly address certain methodological weaknesses, but these often require detailed knowledge

of the sequences themselves. There is currently no method that should be blindly used for phylogenetic inference. There are methods that can be used to judge the quality of an inferred tree, and of course one can compare the results of different methods. This should always be done carefully before drawing a conclusion.

Motivation

People collect sequence and expression data and to some extent kinetic and structural data because it is easy, at least when compared to the higher level goal of understanding function of the genes and the interacting networks in which they exist. And while the collection of these data provide an important starting point for future work, independently their full potential is unlikely to be realized. Improved understanding of the biological system will likely come from combining different sources of information.

One method that has traditionally been used to gain understanding of the function of a gene is site directed mutagenesis. This is a procedure where individual or groups of residues in a protein are altered or completely removed. The altered protein is then observed, *in vivo* (in a living system) or *in vitro* (in a test tube), for characteristics different from the native version. A major drawback to this procedure is cost. This method can be very time consuming at best or prohibitive considering the size of some proteins. Also, the mutation of a single residue at a time may not be enough to tease apart the function. Even the complete disruption of the particular protein may not be enough if the organism has another protein that is able to replace it functionally. Mutation of several key residues or mutation of different subunits may be needed yielding an explosion of combinations.

Today it is possible to collect several types of data including sequence and protein structural data for a large number of genes across a large number of organisms relatively efficiently. Phylogenetic trees for a fairly broad range of the animal kingdom are known and are fairly uncontroversial. While there may be debate about the relationships at the species level, the relationships among genera and families are less controversial.

We have discussed some of the motivations for collecting genome sequence data. We have also discussed the fact that many organisms share homologous genes with corresponding sites that can be aligned in a tabular arrangement, referred to as a multiple alignment. Insights can be gained from the multiple alignment of the areas of a gene that are constricted in some way preventing change. We also examined cases where a phylogenetic tree can present a more likely explanation for apparently unlikely events in sequence data when only the alignment is considered. We note is that for several proteins there is structural data to complement the sequence data. It is the general structure of proteins that is assumed to give each class of proteins its unique function. We ask the question: Is there some way that the alignment,

phylogenetic tree and structural information can be harnessed to ease the search for the most elusive property of the protein, namely its function?

In a previous section the inference of phylogenetic trees was discussed in conjunction with the assumptions required by the inference procedures. Most methods require that the sequences behave as strings of independent residues without any reversion of character states for the inference procedures to infer the correct tree. However, structural requirements, folding pathways, and other constraints place restrictions on the selection of the characters in the gene that code for the protein.

The restrictions placed on the possible mutations that can occur in a gene and remain functional are tightly correlated with the function of the protein. Site directed mutagenesis studies can help deduce the function of the protein and which residues are critical for that function, but in the absence of a criterion to choose which sites should be mutated the procedure is prohibitive. In this study we show how the combination of phylogenetic trees with sequence and structural data can be used to identify candidate sites for mutagenesis experiments. Residues that violate the assumptions of the evolutionary model are identified through a comparison of the inferred phylogenetic tree to a known tree topology. Some force is causing these residues to behave non-randomly and those are the ones that should be examined.

Further information can come from combining the analysis of the phylogenetic support of individual residues with the positions of the residues on the protein structure. Residues that are changing randomly and have no constraints should be scattered all about on the structure. By contrast, residues that cooperate in carrying out a localized function should be spatially clustered on the protein.

This paper is a case study of the comparison of the widely accepted tree to a tree that has been inferred from sequence data of five proteins. A new metric is introduced which is derived from a commonly used metric of measuring the quality of the support that a particular site gives for a tree under parsimony. This information is presented in the context of the structure of the protein and the clustering (or lack of) patterns observed. A software tool implementing the metric and blending the information from alignments, phylogenetic trees, and structure developed for this research is presented. The results are presented followed by some speculation to what might be causing erroneous trees to be inferred. The relationship of the implicated residues to those of known importance is discussed. Unfortunately, no experimental tests of the implicated residues have been performed at this time.

Chapter 2. Materials and Methods

Random Sampling

In the Results section the spatial clustering of amino acid residues seen in each dataset is subjected to a statistical test by comparison to a distribution built from random sampling. By comparing the value obtained from the cluster to the distribution we can determine the chance that this particular cluster would appear at random. The values are computed by simply summing the squared differences in distance between all pairs in a cluster, i.e.

$$S = \sum_i \sum_j \left[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \right]$$

Equation 1. Sum of Squares

Here x, y, and z are the corresponding Cartesian coordinates of the Cα positions of an amino acid residue along the three major axes of the protein. Values are squared to ensure there are no negative values. Only the Cα positions are considered in the sampling. Alternatives can be imagined such as sampling the extents or centers of the side chain residues. While this would be perfectly acceptable, using the Cα positions only is a commonly used approximation. This also protects a person to some extent from errors that may be in the structure that will be more exaggerated in the side chains.

The following figures, from left to right, are examples of results of tests where the clustering is clustered more than expected by random, clustered as one would expect by random, and more dispersed than one would expect by random.

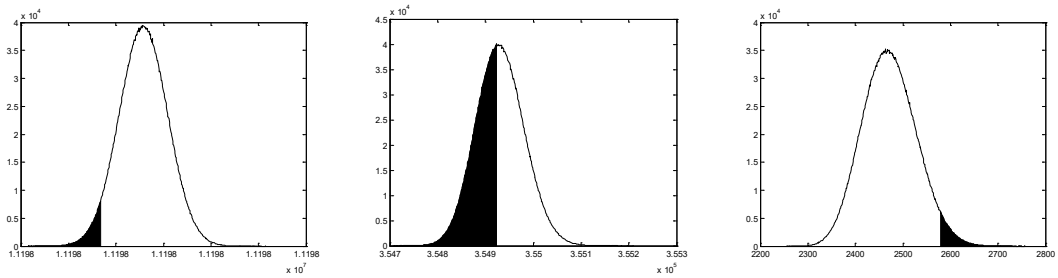


Figure 4. Examples of Randomly Sampled Distributions

Retention Index

The Retention Index (RI) was suggested by Farris (1989) as a quantitative measure to assess the amount of homoplasy that individual residues have in an alignment with respect to a phylogenetic tree. It was offered as an alternative to the previously used consistency index (Kluge and Farris 1969, Archie 1989, Klassen *et al.* 1991) since RI has the advantage of being normalized in a range [0,1]. Farris defined RI for site *i* as

$$RI_i = \frac{M_i - t_i}{M_i - m_i}, \text{ for } M_i \neq m_i$$

Equation 2. Retention Index

where given the residues at site *i*, M_i is the maximum number of changes possible, m_i is the minimum number of changes possible, and t_i is the number of changes implied by the maximum parsimony criteria and this tree. Unfortunately, there are some critical areas where *RI* is undefined. Namely, $RI = \infty$ for any site where the $M_i = m_i$ which will occur at constant sites (only one residue type present in all taxa) and at sites where it just happens that $M_i = m_i$, but $M_i, m_i > 0$.

Retention Index Difference

The Retention Index Difference measure is a method introduced by this paper. This is simply the difference between the corresponding RI values for the same alignment between two trees. Expressed in vector form this would be:

$$RI_{diff} = RI_{tree1} - RI_{tree2}$$

Equation 3. RI Difference

This measure gives a sense of the relative degree of homoplasy (identical states not the result of a shared ancestor) of a dataset with respect to the trees being compared. Since RI values have the range [0,1] corresponding to a gradation of high homoplasy to no homoplasy respectively, we immediately see that the RI_{diff} values have the extended range of [-1,1]. For RI_{diff}=-1, we must have the situation where RI_{tree1}=0 and RI_{tree2}=1. This situation would happen when the dataset suggests that tree1 has maximum homoplasy and that the dataset perfectly supports tree2. When RI_{diff}=1, the opposite must be true, namely RI_{tree1}=1 and RI_{tree2}=0. This happens when the dataset suggests that tree1 has perfect support and tree2 has maximum homoplasy.

RI Compare

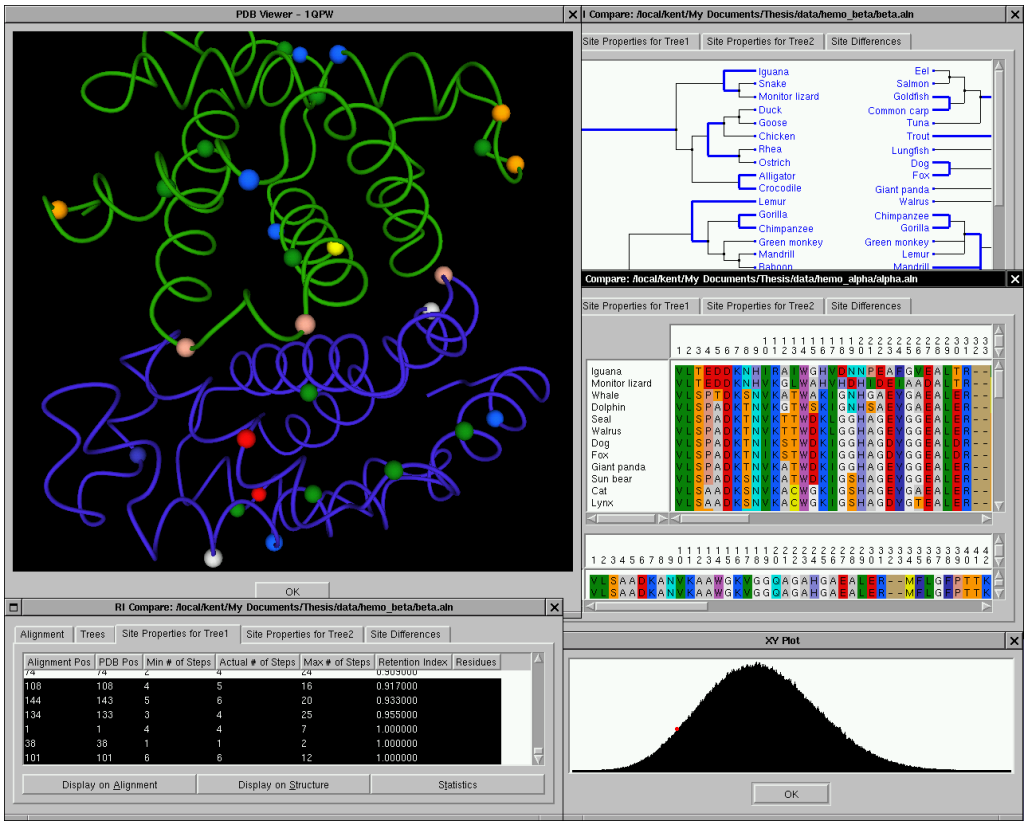


Figure 5. RI Compare Interface

A major portion of the time invested in this research project was the development of the tools used to collect and analyze the data. The analysis tool that was developed focused on the presentation of the aforementioned RI Difference, but also allows the user to explore other properties and measures of the data related to the RI Difference measure. These alternate measures include the raw RI values, measurement of the variability, residue types, *etc.* For a complete explanation of how the tool is used the reader is referred to the RI Compare User Manual, a brief description of the main components will be discussed here.

This section will provide an overview of how the user interacts with the tool and how the tables of values are computed. When the program starts the user is presented with a blank invocation of the method and as the information is provided the interface will expand to resemble that shown in the above figure. Multiple invocations are possible allowing cumulative information to be displayed. This is useful in the situation where there are multiple chains in a single protein

allowing an analysis to be performed on each chain in its natural combined context giving hints to possible interactions between chains.

The input to the tool is a set of aligned sequences in FASTA format, two trees relative to the aligned sequences in nexus format, and a PDB file containing the structure on which residues are to be highlighted. The tool performs pairwise alignments to find which chain of the structure best aligns with which sequence from the alignment forming a map between the alignment and the structure. If multiple such pairs exist, the user is given the choice of which pair to use. While a structure is not required for computation of the RI and related values, it is required for cluster analysis and discovery of spatial patterns.

As the user loads the tree files they appear side-by-side. Clades that are common to both trees are highlighted in boldfaced blue. Otherwise, the clades are drawn in black. This helps draw attention to the regions of the trees that differ between the two topologies.

Analysis of the information starts when the user selects the site properties tab for either of the trees. When these tabs are selected, several properties are computed for each site including the minimum, maximum, and actual number of steps implied by the tree, the retention index, and the represented residues. This information is computed for every site in the aligned sequences for both trees.

Steps are transitions between residues implied by a particular tree. To count the number of steps a matrix is first created where each element is the number of times that a particular residue is implied to change to a different residue (the total number of steps is then a sum of the all the elements). This information is found by making a pass through a tree after the residues of the internal nodes have been estimated using parsimony (the program does this). For each node on the tree the residues of the children are examined and changes are counted in a recursive fashion.

The minimum and maximum possible numbers of steps are independent of the tree and computed in similar ways. Conceptually, the goal is to find a tree topology that implies the minimum number of steps and another for the maximum number of steps for each site separately. It is not necessary to try all possible trees to find the minimum and maximum trees, and in fact no trees need to be tested at all. The minimal possible tree length for a site is simply the number of types of a residue minus one. The maximum tree is found by greedily clustering dissimilar residues together and then finishing by clustering the greedily assembled clusters in any order. Instead of using an iterative procedure suggested by their descriptions, a closed form equation

exists for both of these values. The minimal number of steps is the number of unique residues at the leaves minus one, and the maximum number of steps is the number of sequences minus the number of times the most frequent residue occurs at this site.

While the minimum and maximum numbers of steps are topologically independent, the actual number of steps can only be inferred using a tree. The tree's leaves are first populated with the taxa's residues for a given site. A pass is then made through the tree to populate the internal nodes using an unordered soft polytomous algorithm. This algorithm is well suited for machines because it is simply a number of set operations performed at each node recursively. After the internal nodes are populated the implied changes over the tree are scored and summed.

For each site in the aligned sequences the retention index is computed using

$$RI = \frac{S_{max} - S_{tree}}{S_{min} - S_{tree}}$$

where S_{max} is the maximum number of steps that could represent this data, S_{min} is the minimal number of steps, and S_{tree} is the number of steps implied by a tree using parsimony. Remember, while S_{tree} is restricted to the supplied tree, S_{max} and S_{min} are not restricted to that particular topology. The retention index is used as a normalized consistency index to show relative support for a particular tree among sites.

Having computed the retention index for both trees, we proceed to the site differences tab. This tab displays the site position and the difference in retention index values between the two topologies as well as the individual RI values. The more positive the RI_{diff} value, the greater the support this site provides for tree1. The more negative, the greater the support is at this site for tree2. Sites with a value of zero are sites that support both trees equally.

In both the site properties and site differences tabs the user can select and unselect multiple sites. If a PDB file has been loaded then the selected site will be highlighted on the structure. The user may interact with the structure through rotating, zooming, and translating to explore the relationship of the highlighted sites.

Selected sites can also be subjected to a basic statistical test to help the user evaluate the significance of an apparent spatial cluster. The test uses a collection of residues selected by the user and computes a UPGMA tree based on the relative spatial distances of the selected residues. The cluster structure of the UPGMA tree helps the user assess the presence of multiple clusters in the data. Testing continues after the user selects the portion of the UPGMA tree containing

the subset of residues to examine, which may be the entire tree. After the subgroup is selected, random sampling of the same number of residues as that selected is performed with each random sample being fed to the sum of squares formula. The random sampling is done several million times (the exact number is controlled by the user) and a distribution of the results is computed. The sum of squares result for the selected residues is also calculated and compared to the randomized distribution. P-values are computed and displayed along with the distribution and relative position in that distribution of the sum of squares value of the selected residues. This method is used since the data may not match any theoretical distribution and there is no need to estimate any distribution parameters. The p-values can be examined and appropriate cutoff values such as 5% or 95% applied to determine if the particular clustering is more tightly clustered or more dispersed than one would expect from a randomly selected collection of residues from the protein. However, one needs to be careful to consider the possibility that multiple clusters may exist in the selected residues which may have significant p-values independently, but when considered as a single group may have low significance. This problem is partially addressed by presenting the UPGMA tree prior to sampling giving the user a chance to examine relative distances between potentially separate clusters.

Also present are options to allow the user to quickly highlight residues or sites on both the structure and alignment using cutoff values. While the main focus of the project was examining the presence of clusters using a new measure and showing one way that protein structure can give insights into the relationships of implicated residues that alignments can not, the need to view these sites in context of the alignment still exists. Because of this need the alignment options remain available.

The primary use of the RI Compare tool, in addition to being a demonstration of the RI Difference measure, is to be a hypothesis generator. Researchers are given the ability to analyze and explore data in a new context and generate questions that are addressed by new research. It is important to understand that no particular hypothesis is being addressed by these procedures.

Data Collection

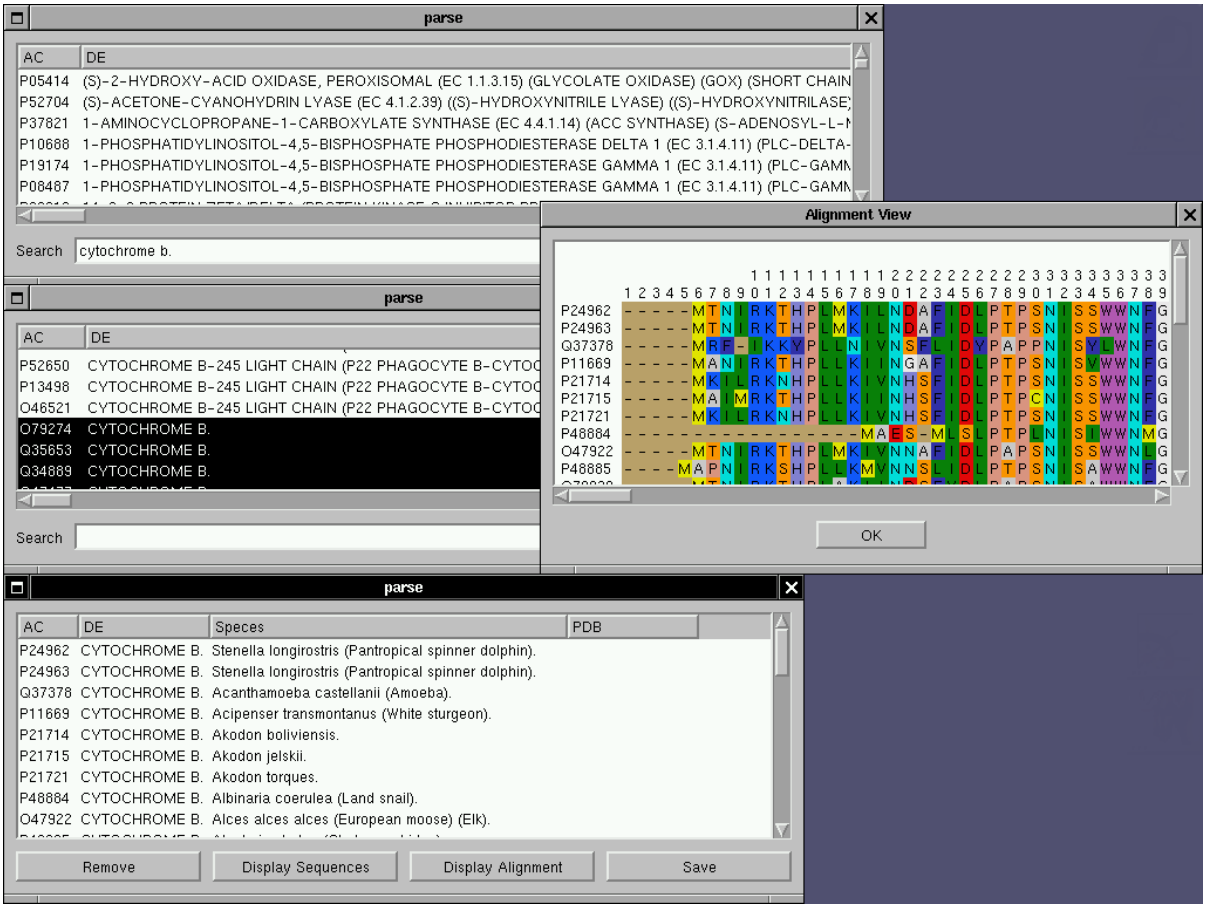


Figure 6. Data Collection Tool

The data collection tool developed and used in this project, SP-Parse (SWISS-PROT Parse), warrants explanation. For the bulk of this project six datasets were used: cytochrome b, rhodopsin, myoglobin, and hemoglobin α and β. However, for this tool to be verified and the power fully extracted a separate tool was created to aid in building datasets. While the development of this tool was not yet finished at the time of writing, it was complete enough to be significant help. The missing components were not critical to the completion of this project and mainly involve making the tool easier to use as future data becomes available, namely incremental updates and some basic machine learning techniques to automate some of the work.

In the above figure a snapshot of the dataset builder is shown. The idea of this tool is straightforward, but implementation was more complicated. The input to the software is the entire SWISS-PROT and TrEMBL (Bairoch *et al.* 2000) datasets and any associated sequence

addition or revision or annotation update files. These databases are flat files of nearly half a million proteins with detailed annotations totaling nearly 1 gigabyte in size.

The program first reads the supplied databases collecting information about each protein including the description, accession number (a unique identifier for the protein), source species, any associated PDB filenames, and file position information for quick random access. Since our interest is in comparing true trees to those generated by inference, the true trees must involve fairly uncontested areas of the tree of life. As a result, as the data is loaded, proteins from viruses, bacteria, and archea families of life are excluded. This leaves only proteins from the organisms of the eukarya kingdom though this pruning may be altered by the user.

Once records have been loaded for each of the proteins to be used, the user is presented with a list of proteins that have known structures. Few of the proteins that are in SWISS-PROT or TrEMBL have associated PDB files. However, since we are interested in displaying information about a set of sequences for a protein in context of a protein structure it is important to first start with proteins that have associated structures. The user begins to build a dataset by entering a search string or by selecting one of the listed proteins, which builds a search string based on the description of the selected protein. It is the user's responsibility to find the appropriate search string as it would be too unreliable to expect the machine to select the most appropriate keywords from the description. The search string is used to search the descriptions of all of the proteins tying together related proteins. These selected proteins are shown to the user in context of all of the available proteins in a new panel.

After the initial search has been performed, the selected proteins can be used to construct an initial working set. Further searches can be done to add to the collection of proteins. The working set is displayed in a third panel to allow the user to manipulate searches independent of each other. The SWISS-PROT or TrEMBL entry for each of the proteins presented in any of the three panels can be viewed at any time, but the working set proteins get special attention. The user can display the sequences of the selected proteins in a quick unaligned form or in an aligned fashion (the multiple alignment of the sequences performed by an external call to ClustalW).

The user must evaluate the quality of the data based on the multiple alignment the descriptions and related information available in the associated SWISS-PROT and TrEMBL entries. For closely related proteins this is generally fairly easy to do. However, distantly related proteins can occasionally appear to share no relationship at all and may require additional work after the data has been saved. It is easy to see how unrelated sequences could be added since the main

mechanism used to construct the datasets is keyword searches. For instance, one may search for the name of a protein but also get proteins that are described as proteins that bind to the sought after protein. If the user is working with closely related proteins, such extraneous proteins generally are readily apparent in the multiple alignment.

Once the user is satisfied with the quality of the data set, the final step is to save the dataset to an empty directory. The raw SWISS-PROT entries, unaligned and aligned FASTA files of the sequences from the selected protein, copies of the related PDB files, two initial phylogenetic trees, and a log of choices and searches that the user performed while constructing the dataset are placed into this empty directory. The log file can be used to reconstruct or update the dataset and also as a training set for the software to associate important keywords building an ontology for the researcher's domain. The two phylogenetic trees that are generated are based on different data and often give surprisingly different results. One tree is generated from applying the neighbor-joining algorithm to the protein sequences while the construction of the second tree is guided by the lineages of the species from which the proteins originated. These trees can later be more finely resolved by the user or provide a starting point from which the true tree can be created. A future addition would be to consult an online database of phylogenetic relations such as The Tree of Life (Maddison *et al.* 1994) or prune a massive supplied tree to construct the true tree for the user.

Chapter 3. Results

Overview

The goal of this study was to improve phylogenetic inference procedures through the identification of collections of residues in an alignment that did not conform to the model of evolution being used. This goal was later expanded to searching for functionally significant residues after considering reasons that alignment sites were misleading. Five separate datasets constructed from the proteins cytochrome b, rhodopsin, myoglobin, and hemoglobin α and hemoglobin β were analyzed. For each protein a short functional and structural description is given followed by a brief comparison of the correct and inferred trees and finally the results of the analysis with discussion. While both trees (the tree built from external information and the tree inferred from sequence data) are included in each section, the alignments are found in the appendices. Also found in the appendices are complete lists of RI and RI_{diff} values.

The following shows the layout that is used for the results along with descriptions of what is found in each pane of the layout:

<table>
<tr><td rowspan="2">Structure of the dataset's protein with the residues matching the particular criteria highlighted. The color of the residues is dependent on the amino acid type at that position in the PDB from which the structure was derived. A table of the residue color, type, properties, etc. is available in the appendices.</td><td>UPGMA tree with branch lengths of spatial positions of the residues in this category. Note: Branch lengths are consistent in each tree but are not normalized across each tree.</td></tr>
<tr><td>Graph of the distribution built from random sampling of the same number of residues as in the category from all residues in the protein. Highlighted portion shows relative position of observed sum of squares value in the distribution. Also shown are values related to the statistical test.</td></tr>
</table>

Cytochrome b Dataset

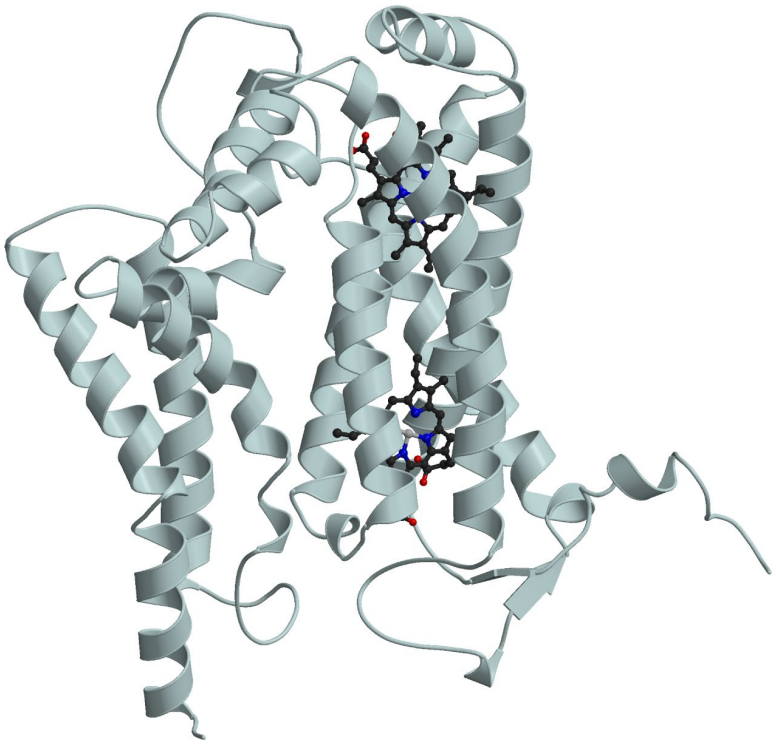


Figure 7. Cytochrome b

Functional and Structural Description

Cytochrome b is a protein involved in electron transfer across the cell membrane. This transmembrane protein, part of the larger cytochrome bc_1 complex (a dimer consisting of 11 monomers), has either one or two noncovalently bonded heme groups where part of the heme group is always associated with a highly conserved histidine residue. Cytochrome bc_1 works cooperatively with NADH dehydrogenase and cytochrome oxidase to provide aerobic respiration in the mitochondrion.

While the structure of cytochrome b has only recently become available (Xia *et al.* 1997), thousands of sequences from diverse organisms are available in repositories. This availability of data has been a factor in the use of cytochrome b for phylogenetic studies. Also, since cytochrome b is a mitochondrial gene, it's inherited only maternally in plants and animals and does not undergo recombination events. Also, because of the shear number of mitochondria in a single cell, the number of copies of any mitochondrial gene is much greater than any nuclear gene facilitating data collection. Also, the rate of evolution is approximately an order of

magnitude faster than nuclear genes (Pesole *et al.* 1999). The increased rate of mutation is important for finer resolution of closely related organisms, but does have a greater potential for multiple mutations to cause reversions to previous states.

A reminder about mutation rate is called for since this is the first dataset to be examined – mutation rate is not constant between species. We have made the assumption that the mutation rate is fairly constant, but because our datasets are restricted to vertebrates with fairly uncontroversial relationships and an occasional invertebrate to form an outgroup we are fairly safe.

Phylogenetic Analysis

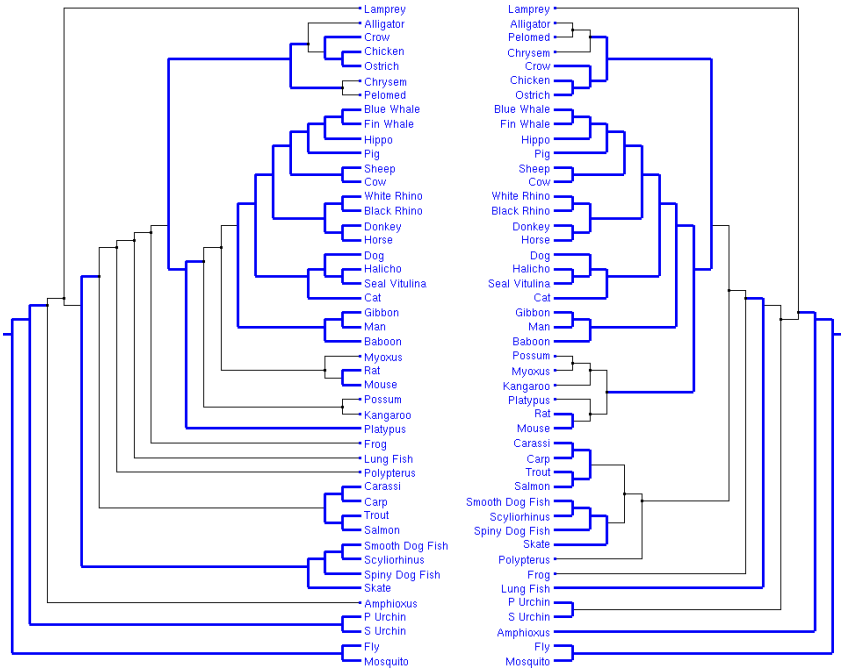


Figure 8. Cytochrome b Phylogenetic Tree Comparison

Shown above is a highlighted illustration of the similarities between the assumed true tree on the left and the maximum parsimony consensus (MPC) tree on the right. The differences between these two trees are examined below. *Amphixous*, a member of the phylum chordata, is separated from the other members of chordata by the incorrect insertion of the urchin clade in the MPC tree. Another problem with the MPC tree is the mix up of the marsupials and rodents. *Myoxus*, the dormouse and clearly a rodent, has been inserted between the marsupials, possum and kangaroo. Similarly, the platypus is clustered with the rodents instead of the marsupials. The MPC tree also has a problem with the relationship of the birds, alligator, and turtles (*pelomedusa* and *chrysemys*). One might suspect that alligator would be related to turtles. However, extensive comparative anatomy indicated that alligators are actually more related to birds. The last main problem with the MPC tree is the mixing of the boney and the cartilaginous fishes. There are other problems which can be easily identified by examination of both trees, but the major ones have been stated.

Results

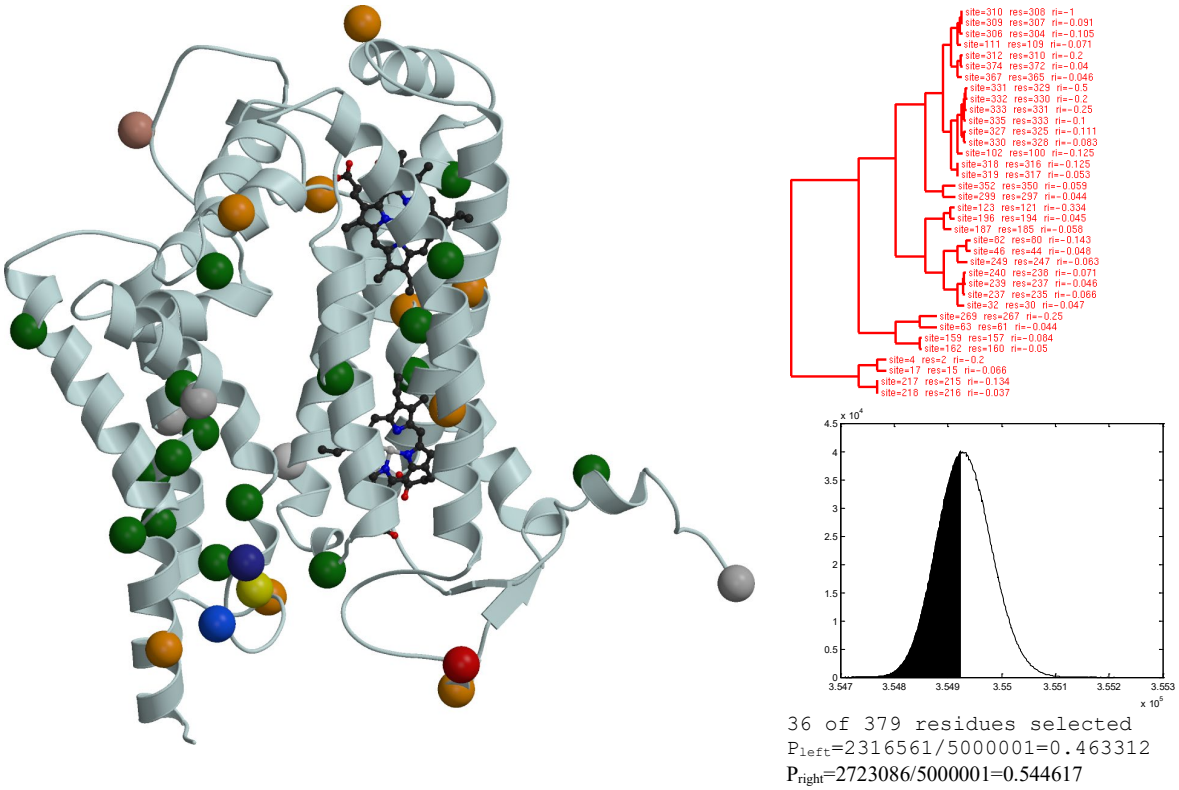


Figure 9. Cytochrome b – RI Difference < 0.0 Highlighted

With the residues in the $RI_{diff} < 0.0$ category highlighted we see that most lie in the transmembrane region. While the distribution does not suggest significant clustering, there do appear to be two or three main clusters in the UPMGA tree. There is also an abundance of green residues representing the hydrophobic residues isoleucine, leucine, and valine. The significance of this observation is questionable since the majority of residues in the transmembrane region are of these types.

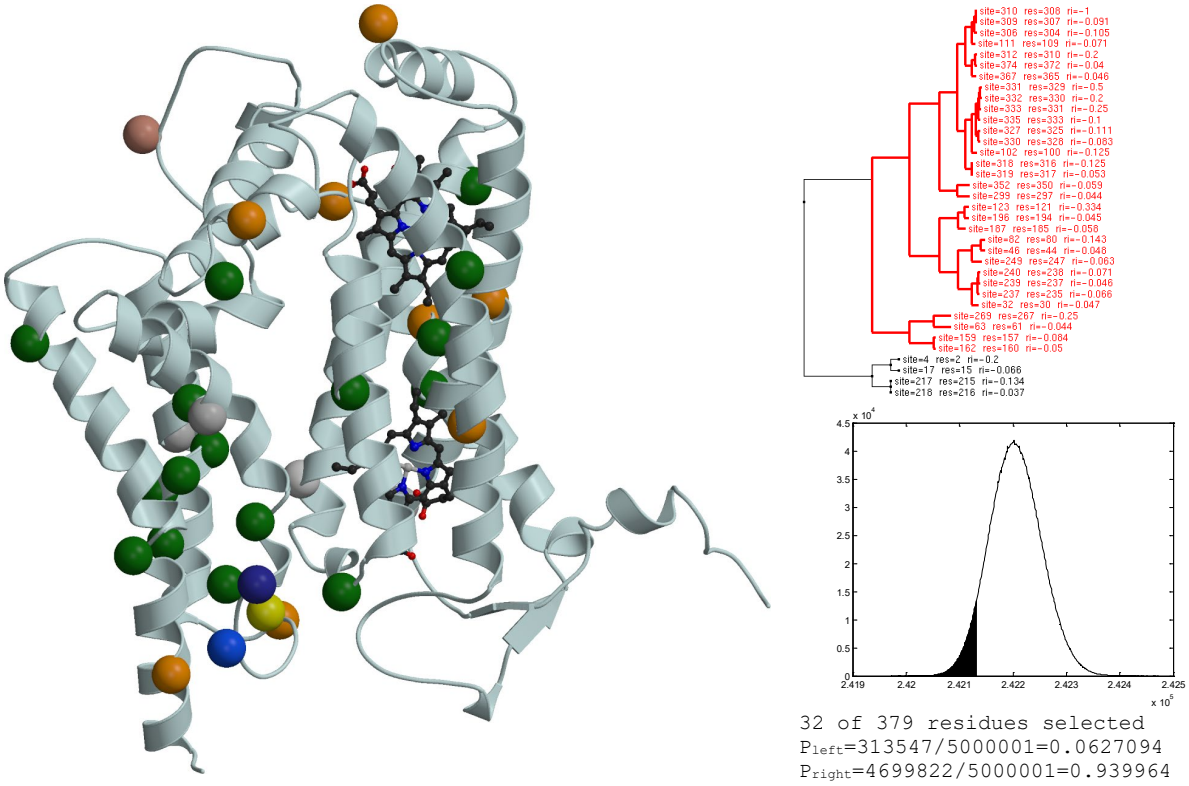


Figure 10. Cytochrome b – RI Difference < 0.0 Highlighted (Alternate)

Having suspected clustering in the previous experiment we test this by removing the outermost group of residues in the UPGMA tree from consideration. This will increase the significance, but we must be careful not to be misled by this. Four residues were removed from outside of the transmembrane region. These residues were on the extreme edges of the protein and had no apparent relationship to each other.

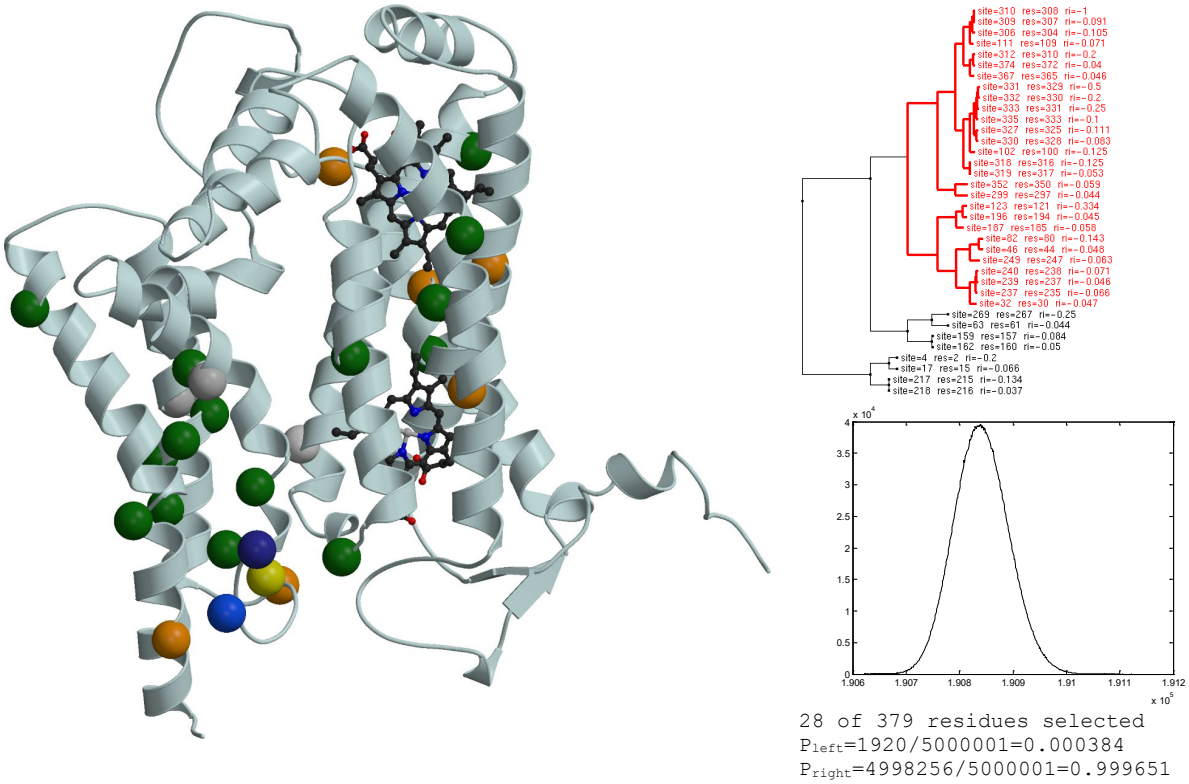


Figure 11. Cytochrome b – RI Difference < 0.0 Highlighted (Alternate 2)

Continuing to peel off the outer groups of residues from the UPGMA tree, further residues outside of the membrane are removed from consideration. The significance increases further, but this will happen whenever outer groups of the UPGMA tree are removed. The clustering within the transmembrane region is becoming clearer now though. While no rigid criteria were applied for removing particular residues in the $RI_{diff} < 0.0$ category from consideration, doing so has provided a means to more clearly see the clustering that was in the original plot but was obscured by the residues outside of the transmembrane region. Considering this plot, there appears to be significant clustering on the set of helices not associated with the two hemes, and perhaps to a lesser extent some clustering on the helices with the hemes.

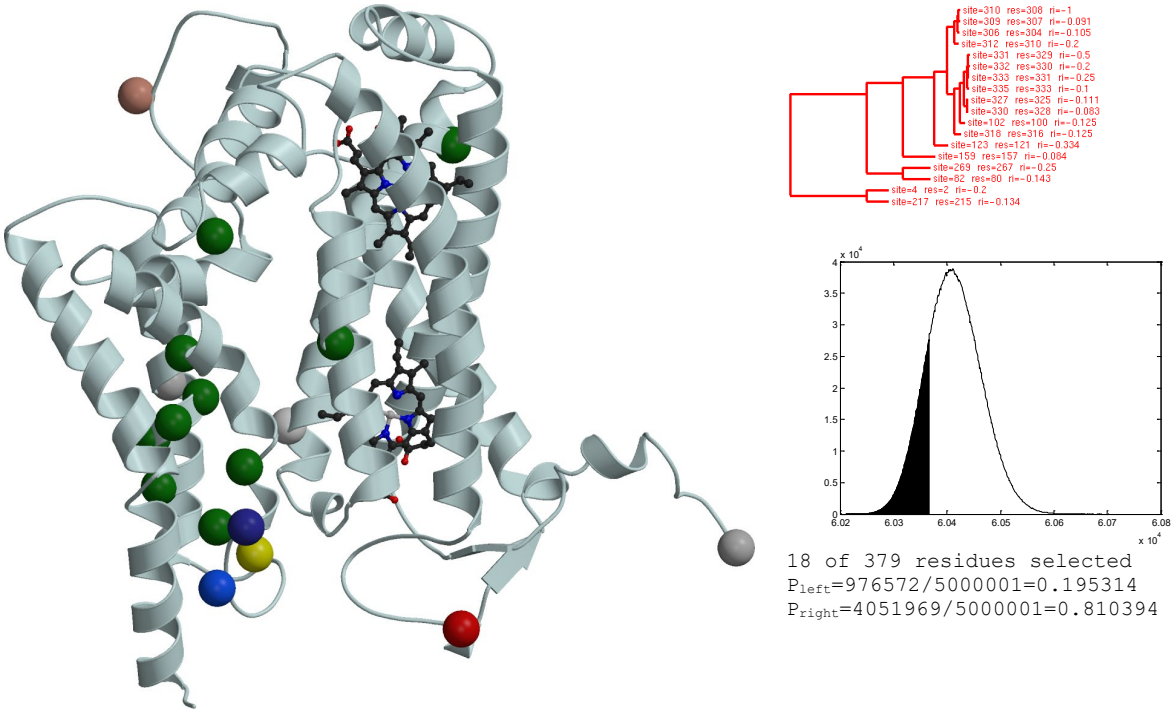


Figure 12. Cytochrome b – RI Difference <= -0.0830 Highlighted

The application of a cutoff value to the $RI_{diff} < 0.0$ category helps to isolate the tightest cluster. There are still residues that are outside of the membrane, but the clustering is fairly clear without removing any portion of the UPGMA tree.

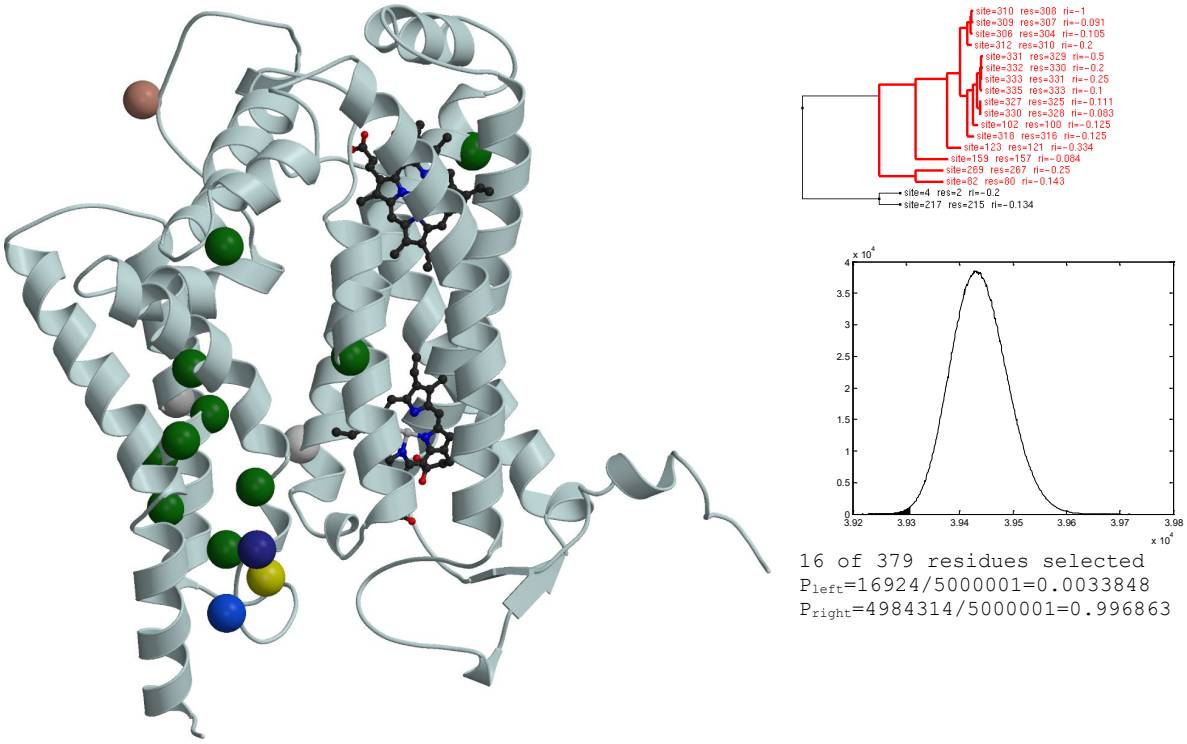
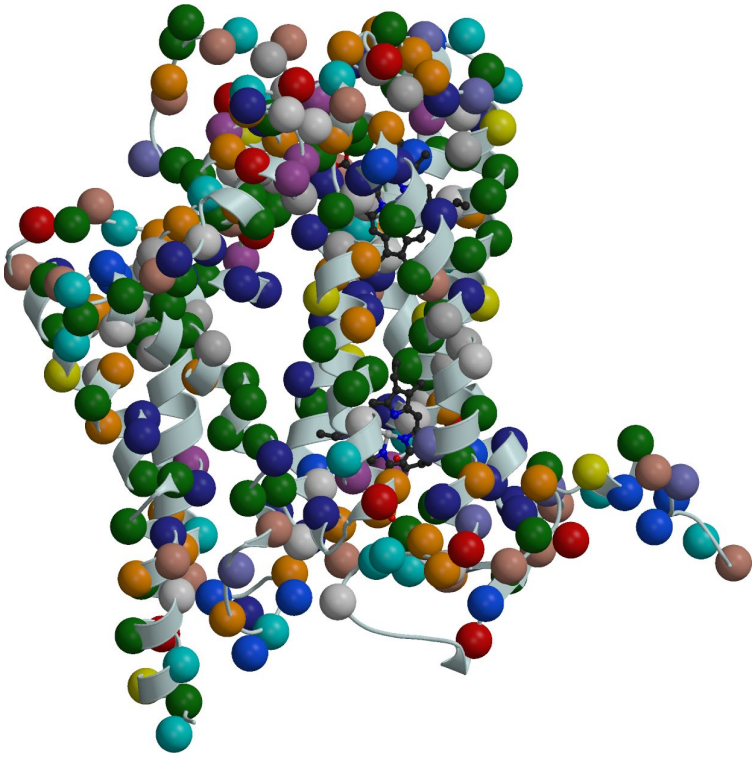
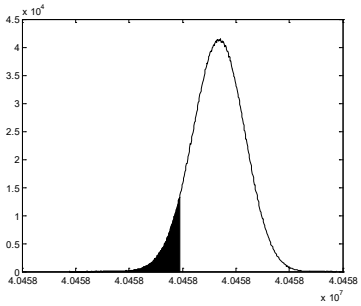


Figure 13. Cytochrome b – RI Difference <= -0.0830 Highlighted (Alternate)

By removing only two residues from consideration, the significance increases considerably. The tight cluster on the helices not associated with the two heme groups is clear. A few extraneous residues still exist, which may be a result of the crudeness of the method or data.



UPGMA tree not shown due to its extreme size.



297 of 379 residues selected
P_{left}=366726/5000001=0.0733452
P_{right}=4646958/5000001=0.929391

Figure 14. Cytochrome b – RI Difference = 0.0 Highlighted

While the statistical test shows a weakly significant clustering, it is hard to see this visually with residues highlighted all over the protein without any clear pattern. Perhaps this significance result is because the residues are not evenly distributed in the volume of the protein. Basically, this graph has residues dispersed throughout the entire structure, but does provide a nice example showing the disproportionate distribution of the residues types in the three main regions (the two non transmembrane regions and the transmembrane region).

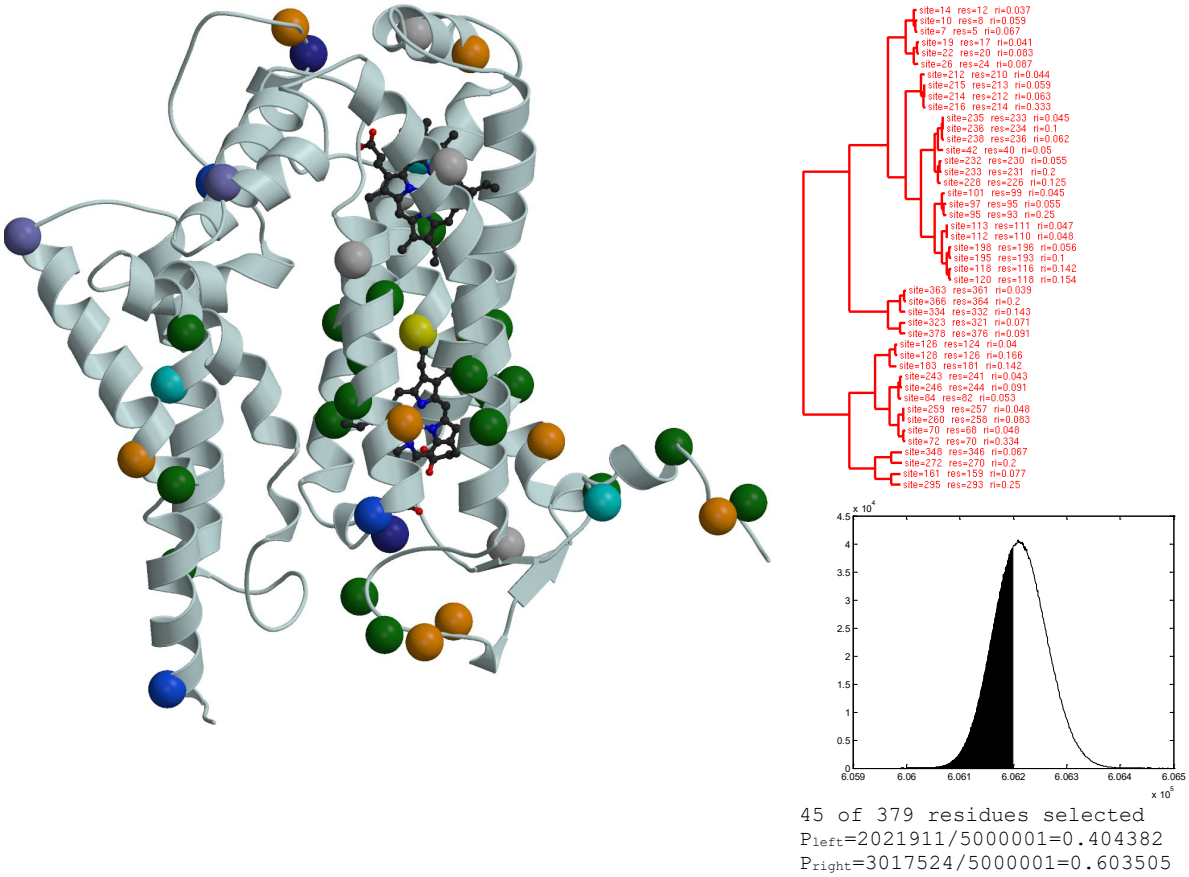


Figure 15. Cytochrome b – RI Difference > 0.0 Highlighted

No apparent patterns are visible when considering the residues with positive RI_{diff} values. There are several in the transmembrane regions and most of those are around one of the heme groups favoring the portion of the transmembrane region complementary to those in the $RI_{diff} < 0.0$ category. The residues in the transmembrane region in this category typically had smaller RI_{diff} values; however, the largest value was 0.33.

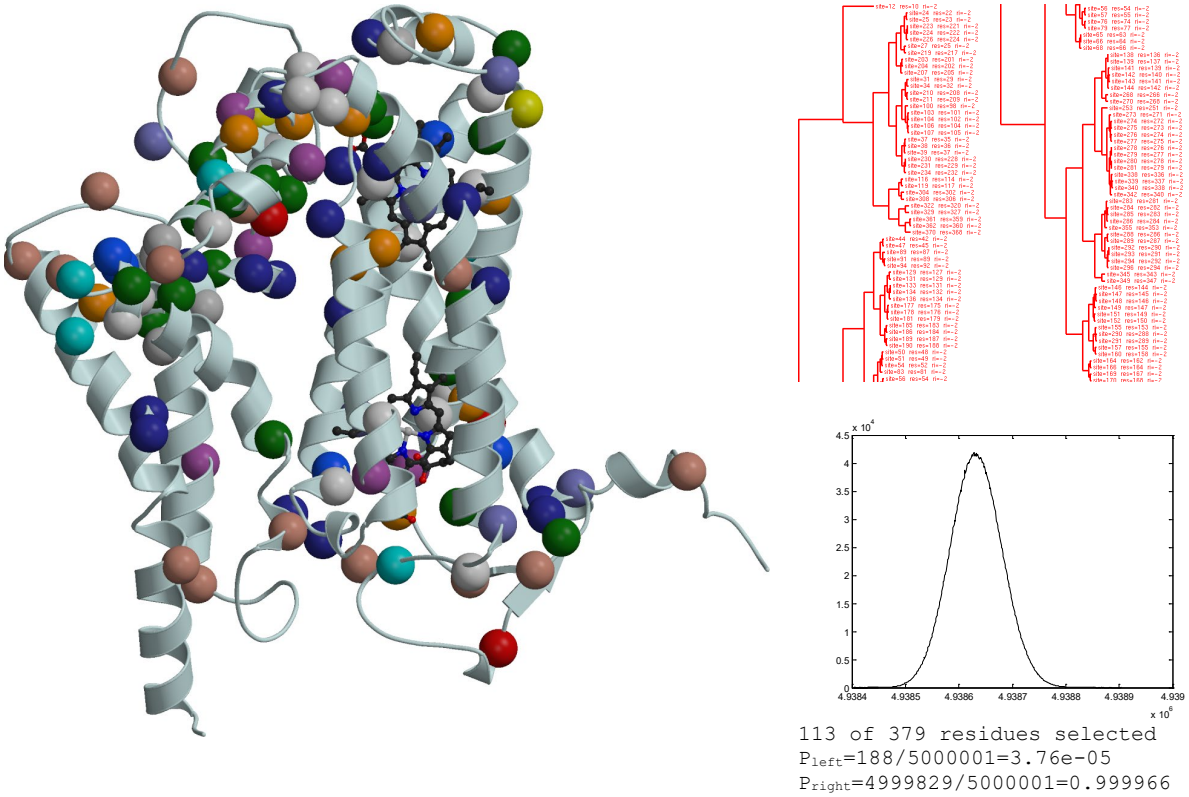


Figure 16. Cytochrome b – Sites with No Change Highlighted

While the clustering that was observed in previous plots examined residues based on RI_{diff} values in the transmembrane regions, invariant residues behave in a complementary fashion by favoring the non-membrane regions. After highlighting the invariant residues, there appears to be tight clustering on both sides of the transmembrane region. While there are a few scattered throughout the membrane, it is interesting to note that these are not green residues (isoleucine, leucine, and valine residues) which are by far the most common in that region. Also, while there appears to be a favoring for the residues to be clustered on a particular side of the protein, this is only a consequence of the distribution of the residues. The side with the higher number has several helices which help to pack the area with more residues than the opposite side that mainly has strands as the primary secondary structure. Strands are basically linear arrangements of residues and therefore can not pack the residues nearly as tightly. One should note that most of these residues have Max Change = Min Change = 1. Examination of residues in the subset of this group with Max Change > 1 did not appear to have the same clustering pattern.

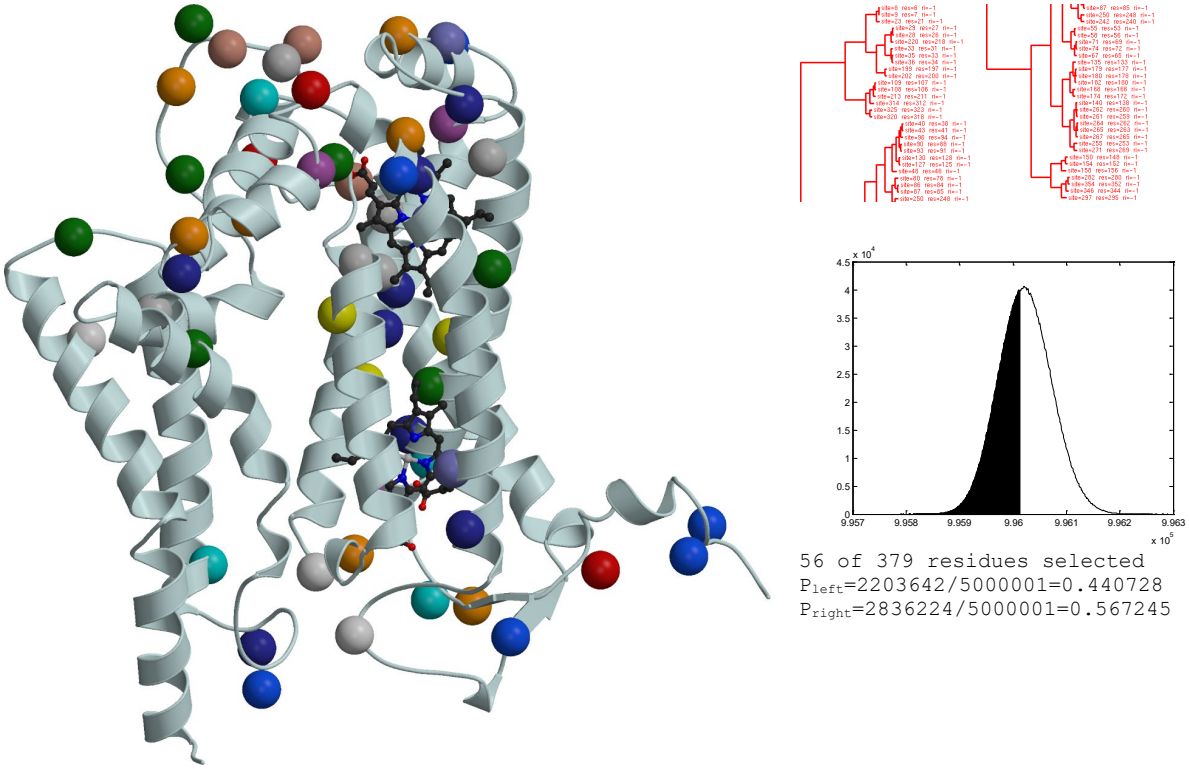


Figure 17. Cytochrome b – Sites with RI = ∞ Highlighted

Highlighting the residues that have undefined RI values gives an interesting graph. Again, as with the invariant residues, there appears to be clustering on both sides of the transmembrane region, or at the least if there is no clustering there is a tendency to avoid the transmembrane region. The residues that are in the transmembrane region are along the helices associated with the two heme groups. Also, these residues are generally not the hydrophobic residues isoleucine, leucine, and valine (shown in green) that predominate this region.

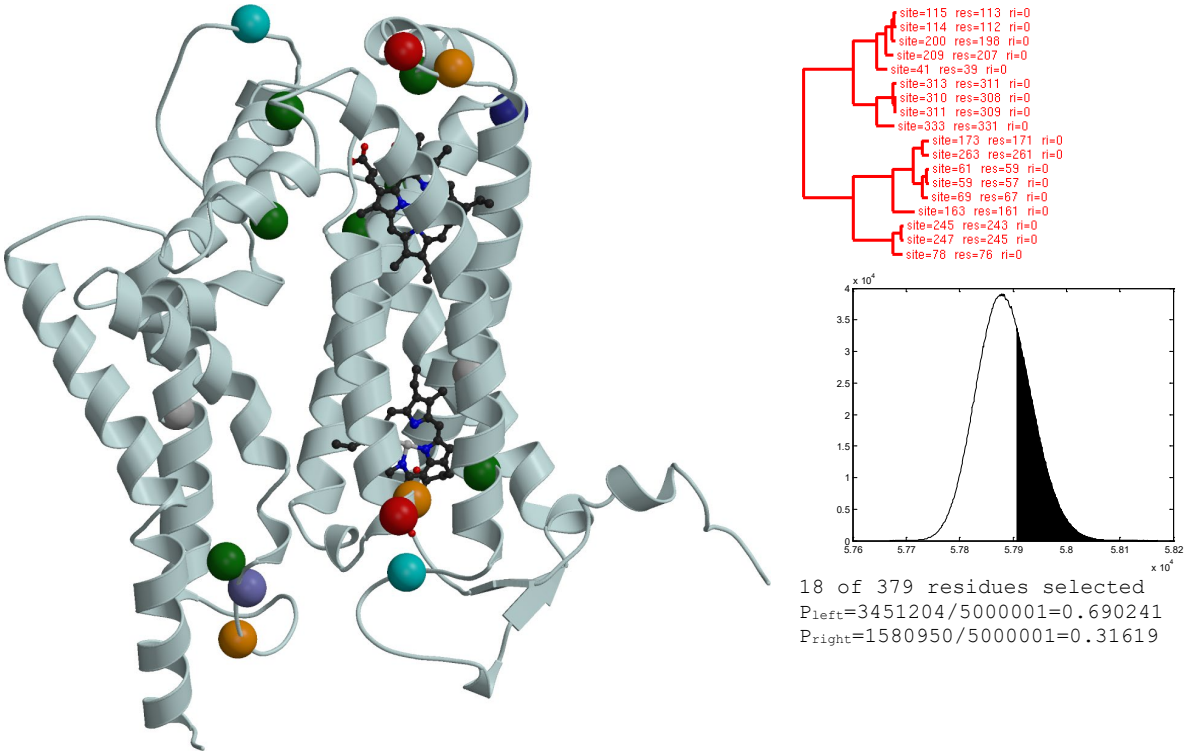


Figure 18. Cytochrome b – Sites with True Tree RI = 0.0 Highlighted

Examining residues that have the lowest allowable RI values with respect to the true tree, we see there is a definite avoidance of the transmembrane region. The distribution shows an arrangement that is more dispersed than the mean is, but it is not significant at the 0.05 level. It should also be clear that from the arrangement present in the graph, we should expect the test to indicate a more dispersed pattern. This is because there are two groups of residues straddling an area without any. When randomly sampled, the region without any highlighted residues is also sampled. This example also helps to show that the statistical test alone is insufficient to determine if there is clustering. While there are clearly two clusters in this graph, because they have such a great distance between them, the test shows no evidence of clustering. However, the UPGMA tree hints at it by showing the large branch length between the two clusters. Visual inspection affirms the UPGMA observation.

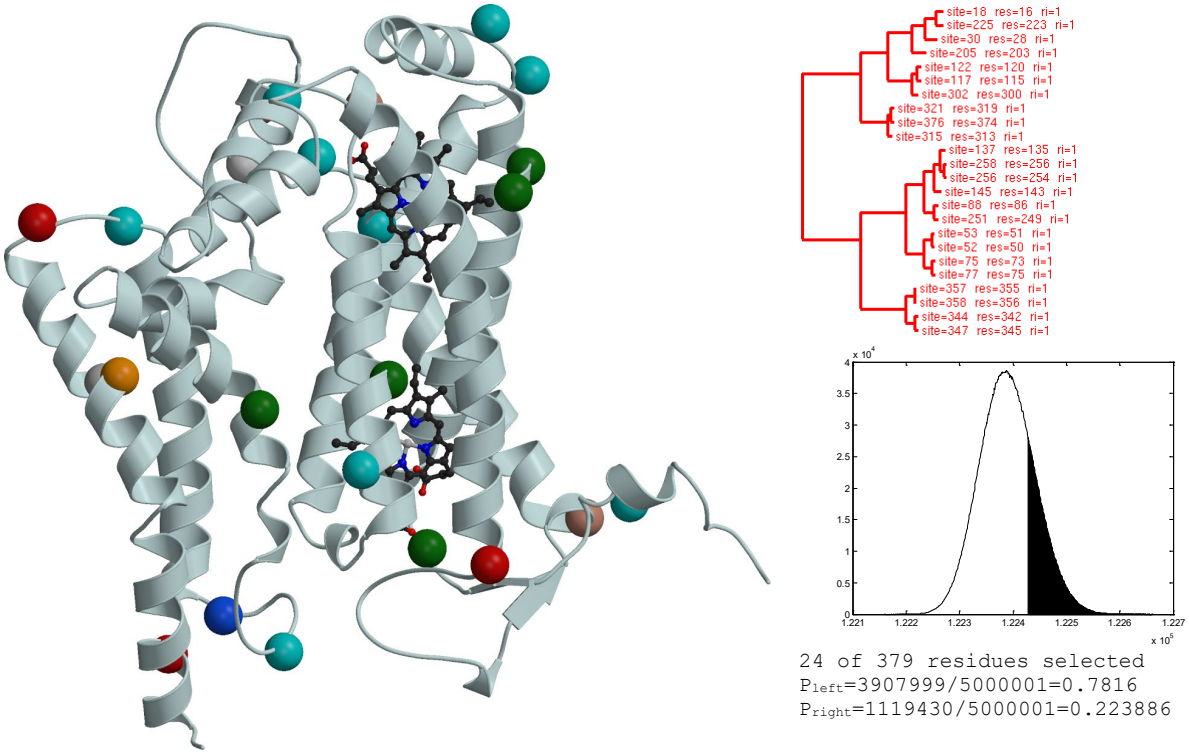


Figure 19. Cytochrome b – Sites with True Tree RI = 1.0 Highlighted

Highlighting the residues that have the greatest allowable RI values with respect to the true tree shows no apparent patterns. There appears to be a greater portion of the residues in the transmembrane region than there was with $RI_{true} = 0.0$.

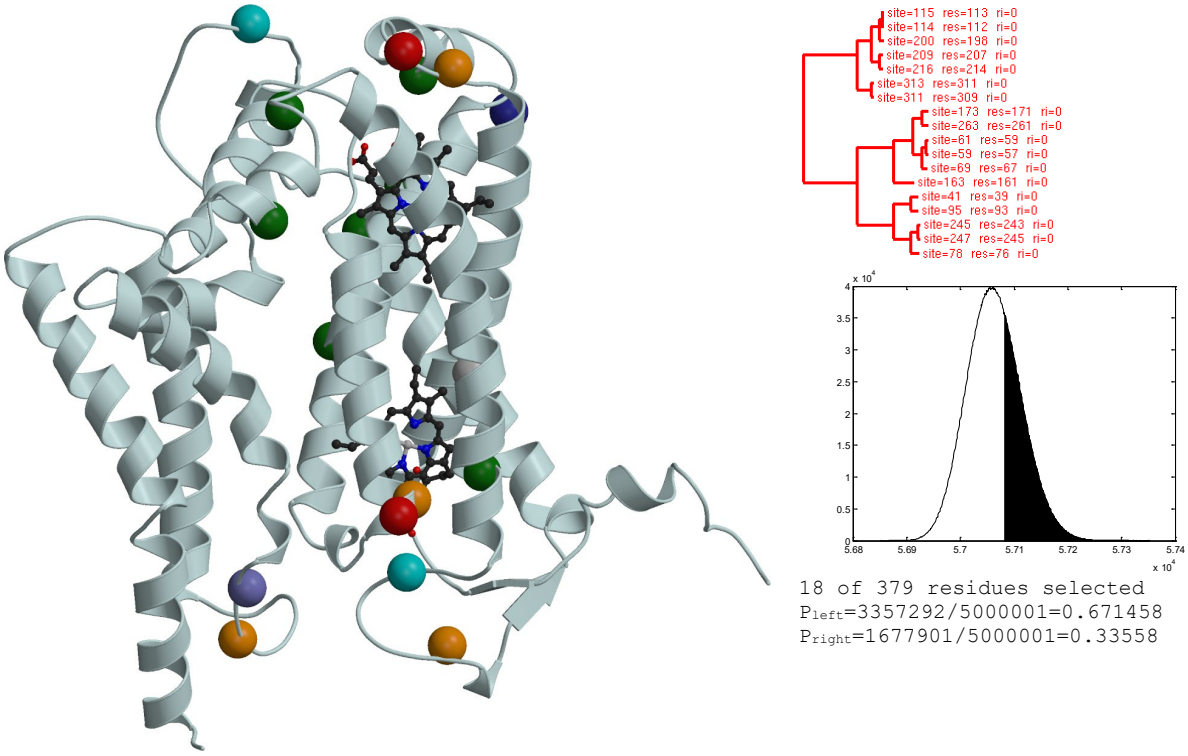


Figure 20. Cytochrome b – Sites with MPC Tree RI = 0.0 Highlighted

Only a few residues are different than those in the RI_{true} = 0.0 category. Refer to the notes for that section.

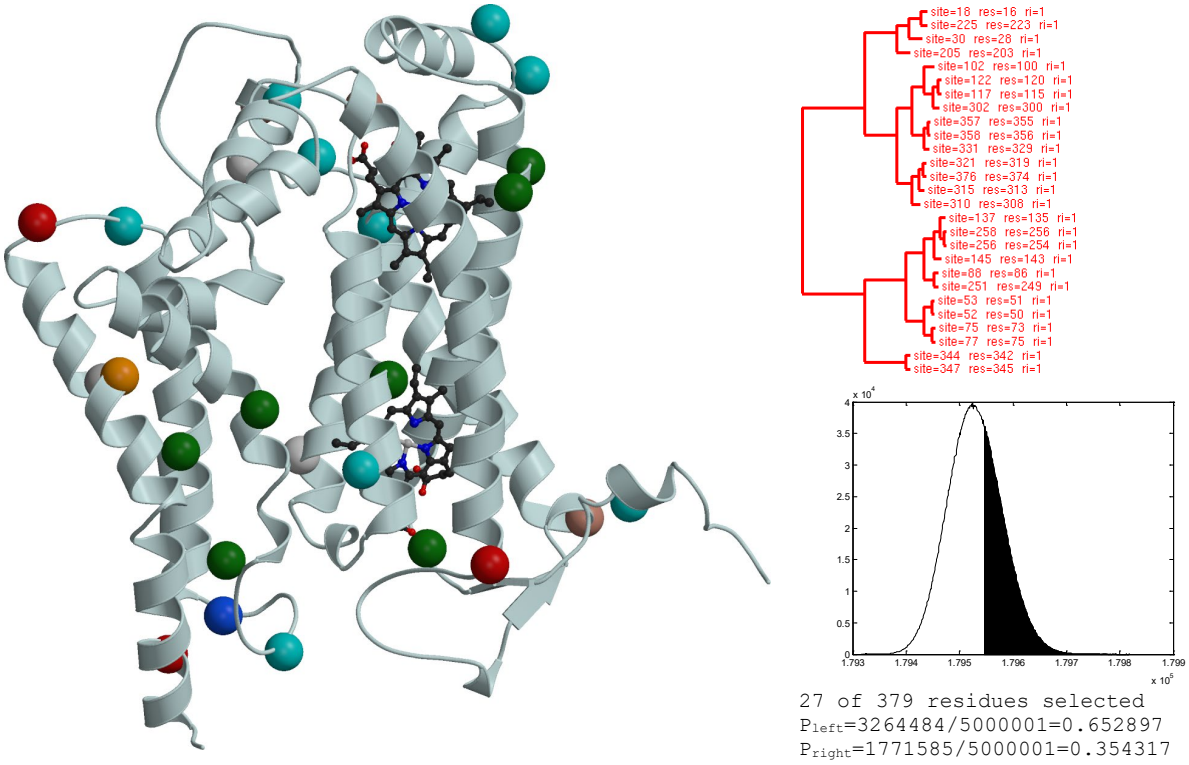


Figure 21. Cytochrome b – Sites with MPC Tree RI = 1.0 Highlighted

Only a few residues are different than those in the RI_{true} = 1.0 category. Refer to the notes for that section.

Rhodopsin Dataset

Figure 22. Rhodopsin

Functional and Structural Description

Human eyes have two main photoreceptors, rods and cones, which cooperate to allow us to see in color and also have some sensitivity in the dark and to motion. The cones, concentrated in the center portion of the eye, provide the eye's sensitivity to various colors as well as the highest visual acuity. There are at least three different cones to provide three different response curves to different colors of light. Interestingly, there are also instances of rare mutations, only present in human females, of a fourth cone that is sensitive to a blend of red and green type, bestowing tetrachromatic vision. But, there are also birds that commonly have more advanced color vision systems such as tetra- and pentachromatic vision (Pichaud *et al.* 1999).

The rods predominately occupy areas where cones are not present. While the majority of the cones are present at the center of the eye, rods are concentrated in the surrounding areas. Rods can not discern between different colors, and in fact are completely insensitive to reds, but provide some degree of night vision (rods are nearly 1000 times more light sensitive than the cones) and peripheral vision (explained by the distribution of the rods favoring areas of the eye other than the center), and are much more sensitive to motion than the cones (due to a quicker response times). What gives the rods their light sensitivity is a protein called rhodopsin which has a photosensitive chromophore called retinal.

Rhodopsin is similar to cytochrome b in that it is a member of a large family of proteins called G protein coupled receptors. Rhodopsin is also a seven helix transmembrane protein. Rhodopsin is a fairly conserved protein demonstrated by an 87% conservation of the 348 or so residues between human and cow. Also some information is known concerning functional constraints of rhodopsin. These constraints include the existence of a disulfide bond, but also folding requirements to hold and interact with the retinal chromophore and rhodopsin kinase (Hwa *et al.* 1999, 2001).

Rhodopsin may also have a second function suggested by recent work (Crandall *et al.* 1997). Organisms that are never exposed to light, such as cave dwellers, have rhodopsin with a similar rate of evolution and structural arrangement as organisms that live in the sunlight. This suggests that the functional constraint has not been lost even though the light sensitivity function is not needed. It was hypothesized that rhodopsin may also play a critical role in circadian rhythms.

Phylogenetic Analysis

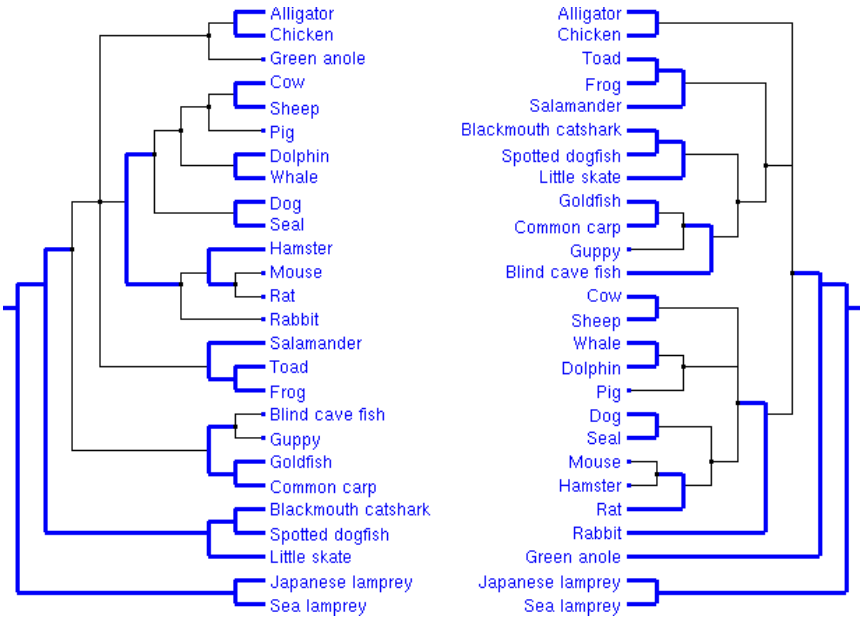


Figure 23. Rhodopsin Phylogenetic Tree Comparison

The inferred tree from the rhodopsin data was fairly close to the assumed true tree. While a superficial inspection suggests several differences, many of these are minor. For instance, several of the individual clades, such as the rodents and fish, have only a single branch that is out of place. The major conflicts between the two trees arise in the arrangement of the smaller clades. The mammal clade, comprised of several smaller clades, is where most of the disagreement exists. Also, the green anole is positioned incorrectly to a significant extent.

It should be noted that it's not uncommon to see some small disagreement at the level of a clade, such as the rodents shown in these trees, even if the tree is comprised of distinct species. Rhodopsin is a fairly conserved protein and so has a fairly slow mutation rate. If the organisms within the clade have not had sufficient time to diverge, the inference procedures may make erroneous associations as a result of noise. In other cases, lack of resolution can occur if no changes are present, and a polytomy will be formed, which is seen in this dataset.

Results

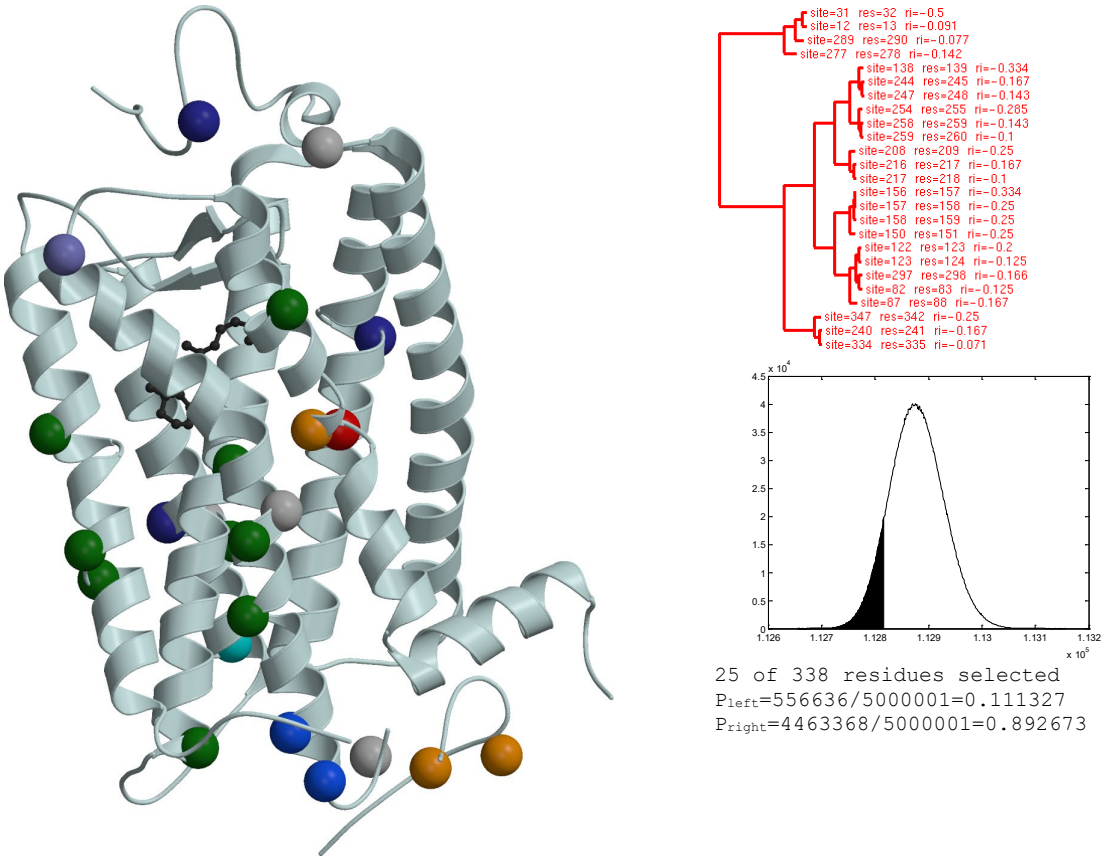


Figure 24. Rhodopsin – RI Difference < 0.0 Highlighted

There appears to be a clustering of the $RI_{diff} < 0.0$ residues associated with the transmembrane region spilling into the adjacent membrane region to one side of the protein. The statistical test indicates the clustering is not significant, but this might be the result of two competing clusters as discussed earlier. Alternatively, one could view these residues as being related by association with the helices connected to the retinal group.

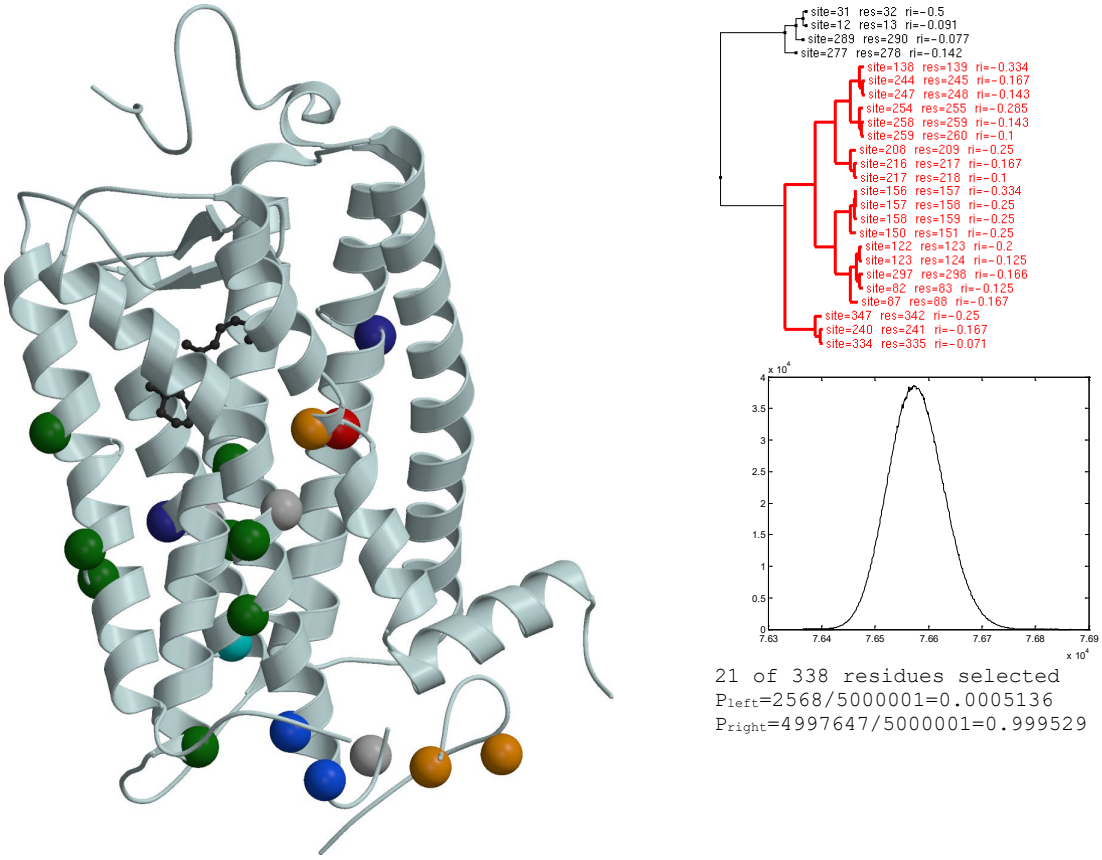


Figure 25. Rhodopsin – RI Difference < 0.0 Highlighted (Alternate)

By removing the four residues on the side of the protein where there was not apparent clustering, the significance value increases and the clustering becomes more apparent. There is an apparent clustering on the helices involved with the retinal ligand continuing into one of the non-transmembrane regions.

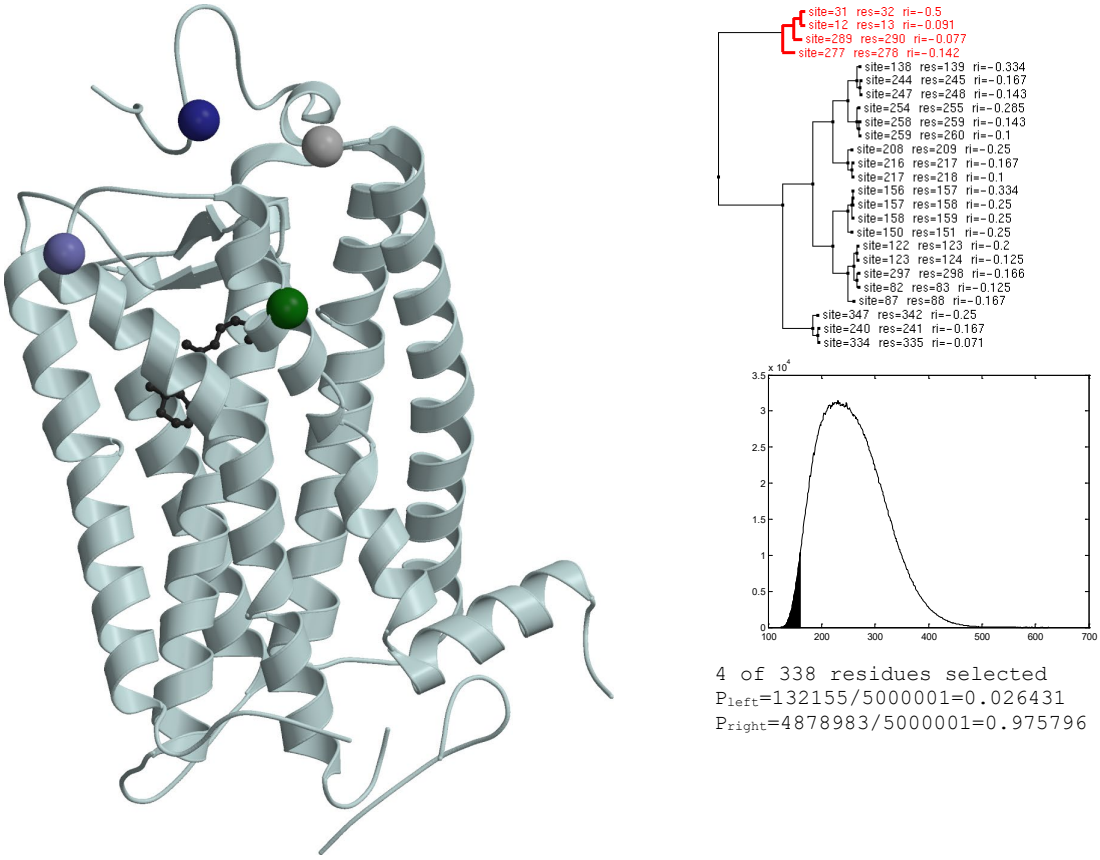
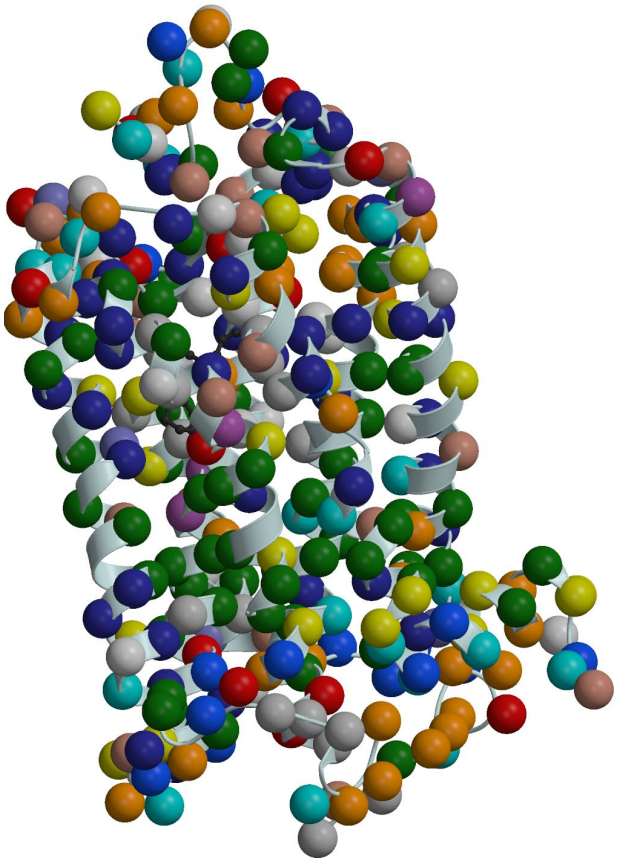
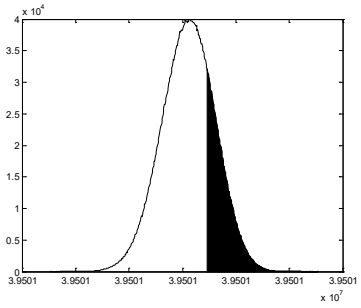


Figure 26. Rhodopsin – RI Difference < 0.0 Highlighted (Alternate 2)

This graph shows the clustering of the four residues that were removed to generate the previous graph. While the cluster has a high significance value, it is predisposed to be so given that it is a small group selected from the UPGMA tree. An interesting related artifact arises when the number of residues randomly sampled is reduced. The null distribution shifts to the left and deforms slightly. As the number of residues randomly sampled is increased, the distribution shifts to the center and is more of typical bell shape.



UPGMA tree not shown due to extreme size.



300 of 338 residues selected
P_{left}=3842062/5000001=0.768412
P_{right}=1189350/5000001=0.23787

Figure 27. Rhodopsin – RI Difference = 0.0 Highlighted

Rhodopsin is a fairly conserved protein so it is not too surprising to see that the majority of the residues fall into the $RI_{diff} = 0.0$ category. While having the majority of the residues in this category does not necessarily suggest a conserved protein. The degree of conservation is more easily addressed by examining the alignment and residues with no change. Because of the shear number of residues highlighted, it is hard to extract anything meaningful from visual inspection. Instead of focusing on what is highlighted, we could examine what is not highlighted. The residues that are missing from this graph are simply those that were shown previously ($RI_{diff} < 0.0$) and those that we are about to examine ($RI_{diff} > 0.0$). We can perhaps notice the pattern here already, the residues that are missing are mainly in the transmembrane regions.

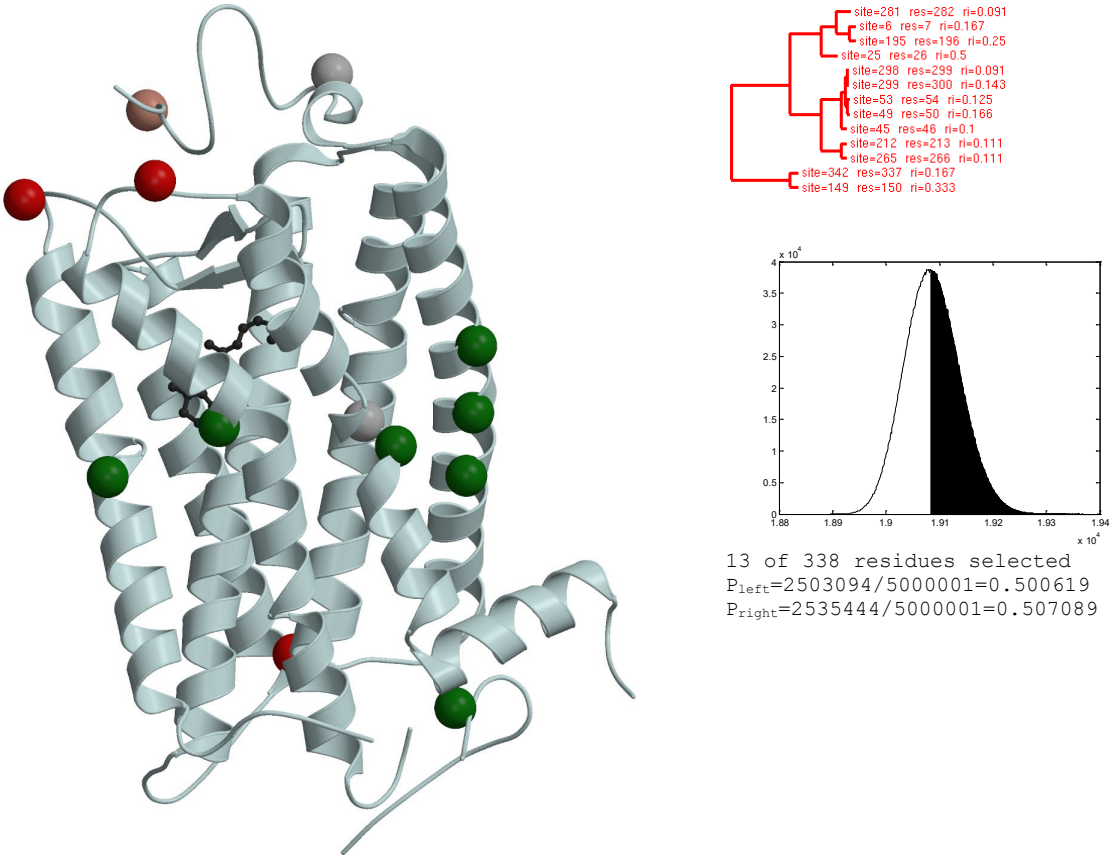


Figure 28. Rhodopsin – RI Difference > 0.0 Highlighted

This category includes residues in all three major regions of rhodopsin. While the $RI_{diff} < 0.0$ residues clustered around the side closest to the heme, here the $RI_{diff} > 0.0$ residues seem to be favoring the helices not associated with the heme. Of course since there are residues in all three regions and the number in each region is fairly small the significance and existence of clustering has to be questioned. While the other residues may be questionable, there is a tight clustering of residues in the transmembrane region. All the residues in the transmembrane region are of the hydrophobic varieties isoleucine, leucine, and valine residues (represented in green). This is a similar pattern as that observed in the cytochrome b dataset. While the residue type may initially appear to be of importance in graphs such as these, one must be careful to consider from where residues have come. In this case, many of the residues in the transmembrane region are of this type and so the chance of randomly selecting a number of residues is higher than if the residue types were all equally represented. This does not negate the fact they are all hydrophobic, but is rather a reminder that we must always consider the background context.

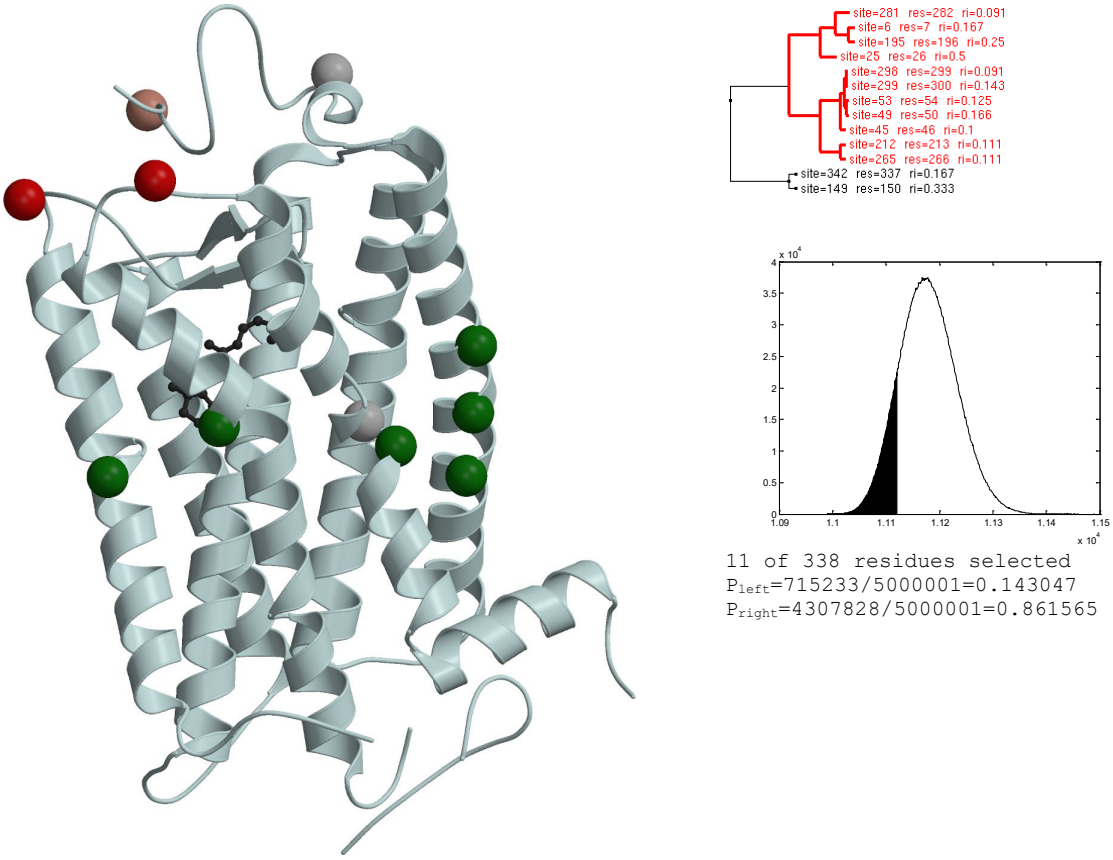


Figure 29. Rhodopsin – RI Difference > 0.0 Highlighted (Alternate)

This graph is similar to the previous graphs except two of the extreme position residues have been removed from consideration for the statistical test. As might be expected, the significance increased considerably as these two residues were removed. There is additional evidence that multiple clusters exist besides the clustering shown on the structure. This is reflected in the relatively longer branch lengths in the UPGMA tree separating the three main clusters.

The underlying justification for this apparently haphazard removal of residues from consideration is the belief that multiple clusters may exist in addition to erroneously highlighted residues. Multiple clusters could occur if there are multiple regions of the protein that are under constraint. Erroneous residues could be picked up either as the result of noise in the data or due to the crudeness of the method. These issues will be addressed further in the Discussion section.

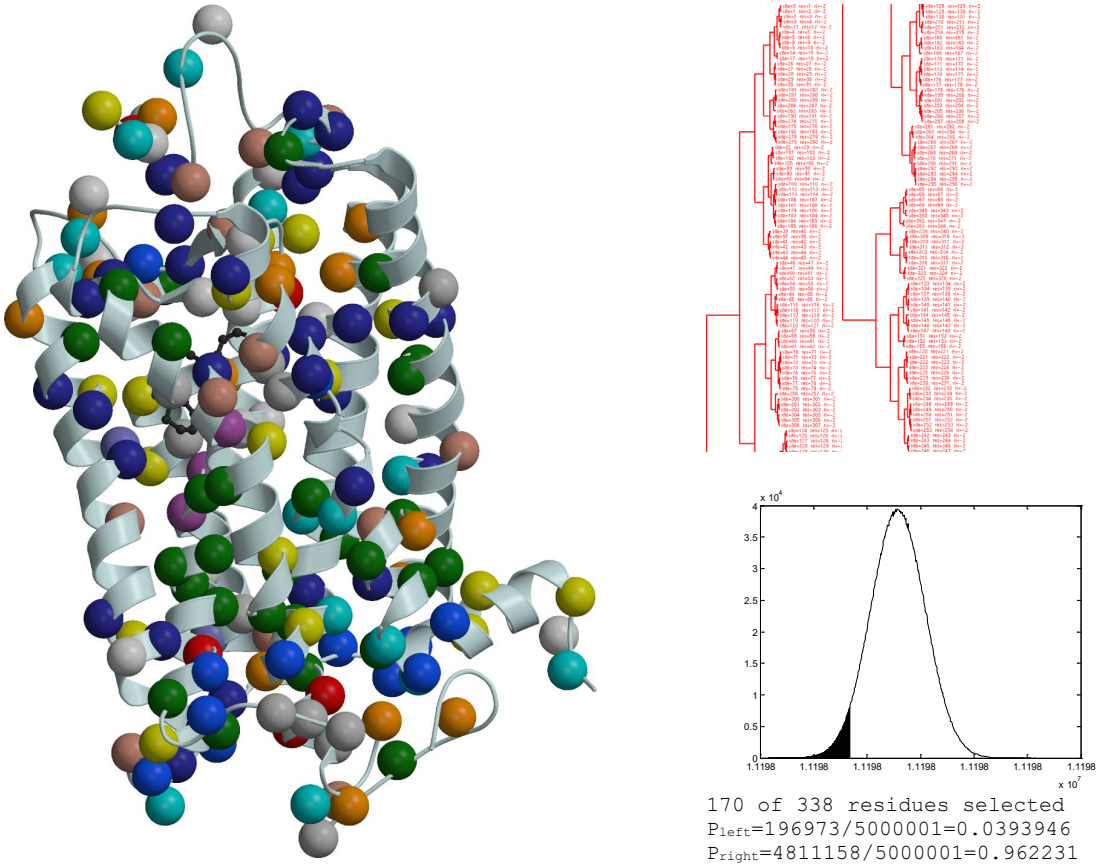


Figure 30. Rhodopsin – Sites with No Change Highlighted

As mentioned earlier, rhodopsin is a fairly conservative protein. Over half of all the residues in this dataset were invariant. Interestingly, even with such a high percentage of the residues falling into this category, the random sampling test still indicants a high significance in the clustering. One would not expect this considering the density of the highlighting on the protein's structure. The only other apparent observation that seems possible is that there are several residues that are not in this category in the transmembrane region, suggesting that the transmembrane region may be freer to change.

A traditional method for teasing out the functionally critical residues would be to assume that the invariant residues are under a constraint that prevents them from changing. By disrupting putatively conserved residues and examining changes in functional behavior of the protein, one could deduce which residues are critical. As can be seen by this example, such an approach would be time consuming.

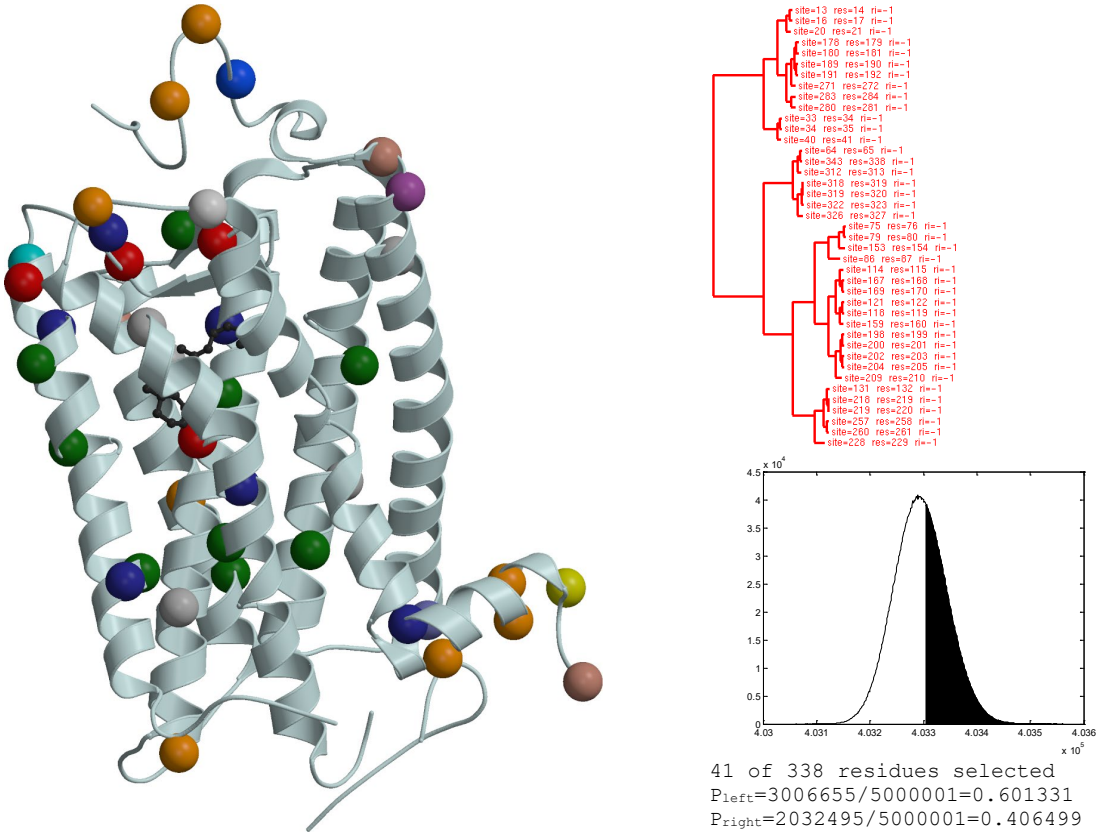


Figure 31. Rhodopsin – Sites with RI = ∞ Highlighted

When examining the residues in rhodopsin with undefined RI values, we see that the residues fall into all three major regions. There are two helices in the transmembrane region that are free of any residues in this category, however. In the transmembrane region and extending outside the membrane on one side, these residues seem to favor the helices nearest the retinal group. Of the residues whose maximum number of changes possible is greater than one, all but one are in the membrane region.

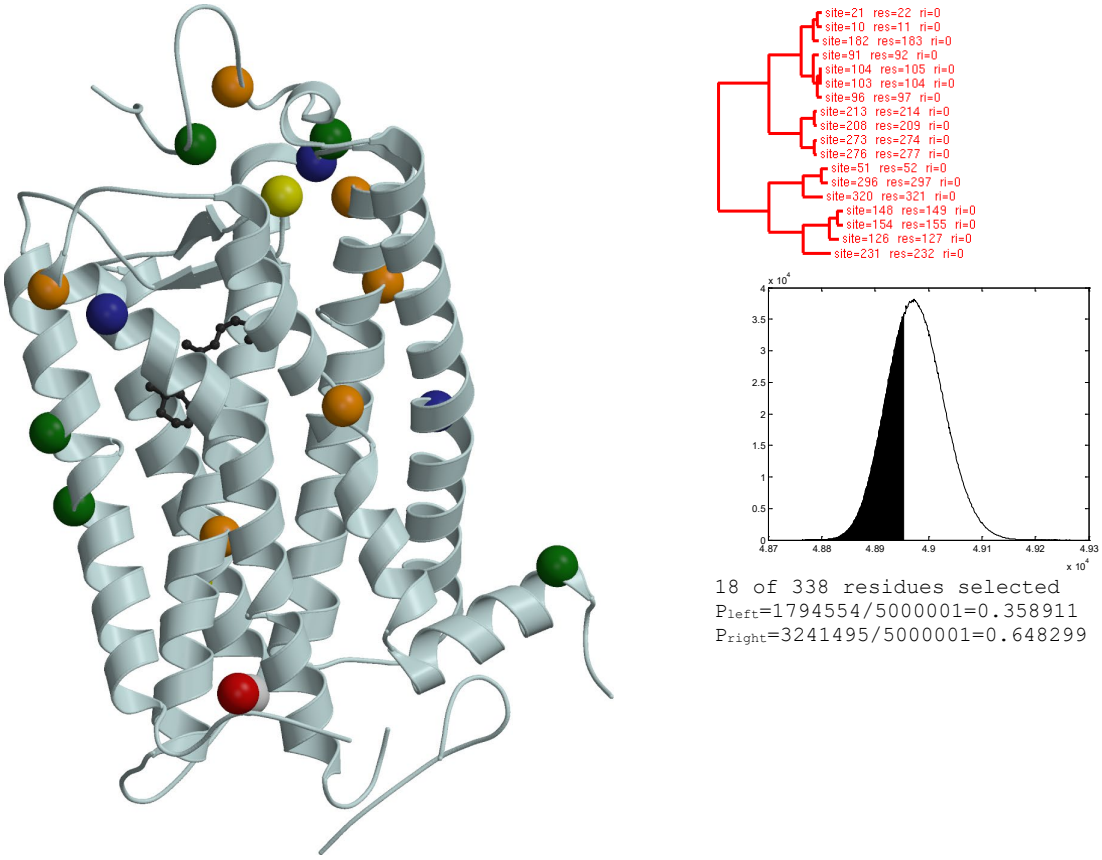


Figure 32. Rhodopsin – Sites with True Tree RI = 0.0 Highlighted

Most residues are in the transmembrane region, but with no apparent clustering or relationship.

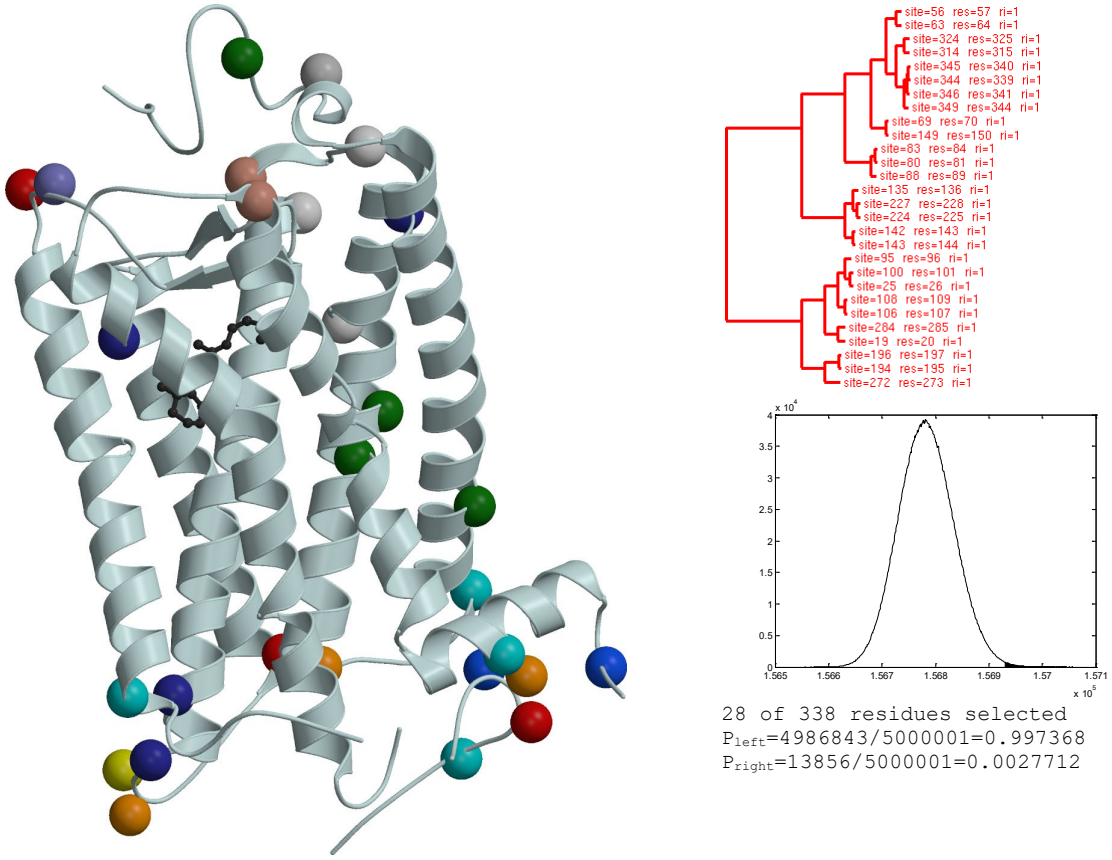
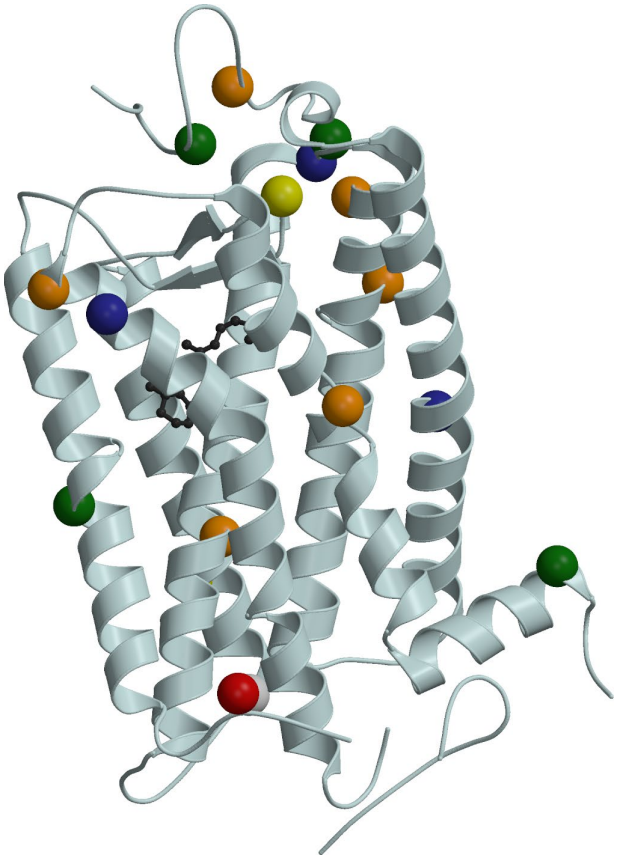
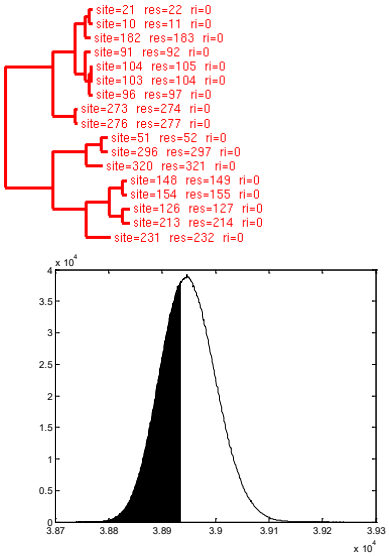


Figure 33. Rhodopsin – Sites with True Tree RI = 1.0 Highlighted

Residues that support the true tree to the greatest extent possible definitely tend to fall outside of the membrane region, as was the case for cytochrome b. There are four residues that are in the transmembrane region that have no apparent connection to the clusters on the outside. The statistical test shows a positioning that is more dispersed than random and results from there being two fairly equal density and size of clusters on the extreme ends of the protein.





17 of 338 residues selected
P_{left}=2015786/5000001=0.403157
P_{right}=3022349/5000001=0.60447

Figure 34. Rhodopsin – Sites with MPC Tree RI = 0.0 Highlighted

These results are mostly identical to the corresponding results discussed for the true tree.

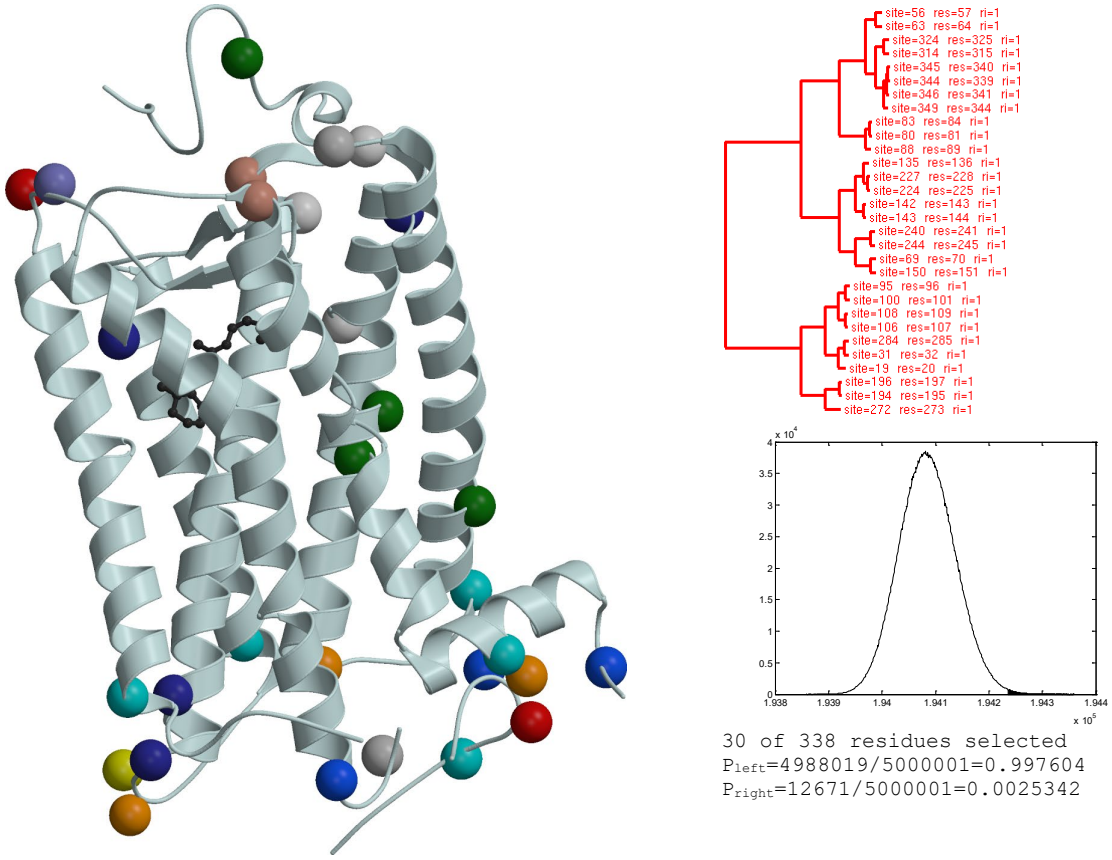


Figure 35. Rhodopsin – Sites with MPC Tree RI = 1.0 Highlighted

These results are mostly identical to the corresponding results discussed for the true tree.

Myoglobin Dataset

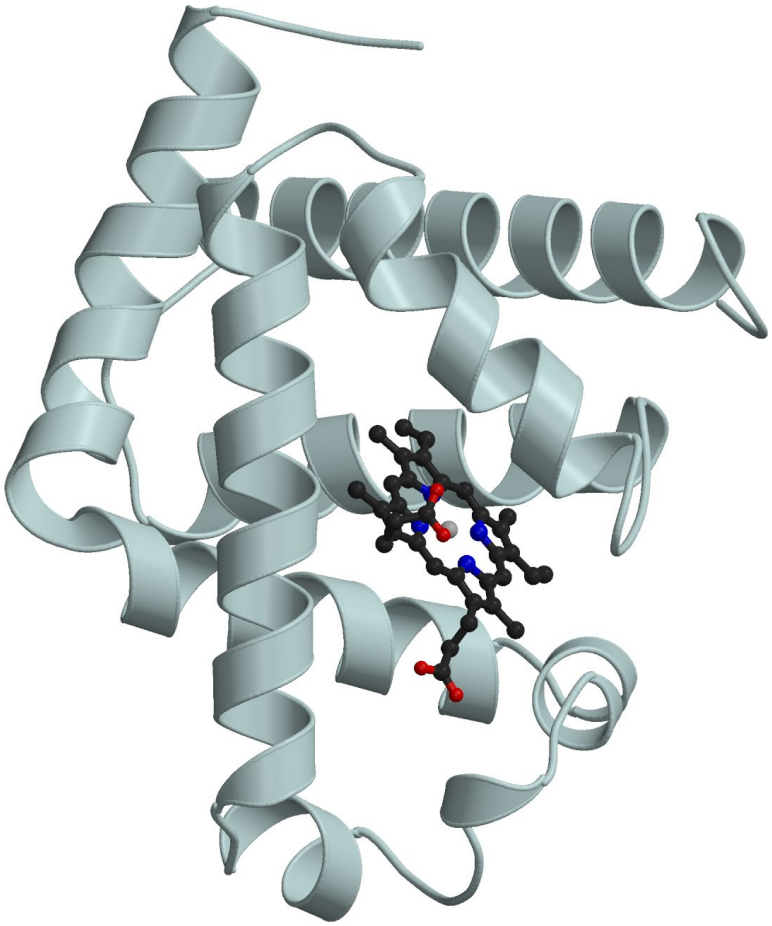


Figure 36. Myoglobin

Functional and Structural Description

The function of myoglobin is oxygen (O_2) storage in the muscle tissues of animals. This is done in cooperation with hemoglobin, which transports oxygen and will be described in the next dataset. Myoglobin has a much higher affinity for oxygen than does hemoglobin and thus will uptake it easily from hemoglobin. The higher affinity, especially at lower concentrations of oxygen, means the stored oxygen is only released during strenuous activity where hemoglobin would not be able to deliver fresh oxygen quickly enough.

The structure of myoglobin is a single monomeric protein of roughly 153 amino acids forming eight helices that surround the oxygen storing heme component. At the core of the heme is an

iron ion where oxygen binds. This part of the heme also bonds to the distal histidine 93 residue, which is conserved across species.

Because of the similarities, both functionally and structurally, between hemoglobin and myoglobin (each of the subunits of hemoglobin resembles myoglobin) much of the information discussed in the hemoglobin section is also applicable to myoglobin.

Phylogenetic Analysis

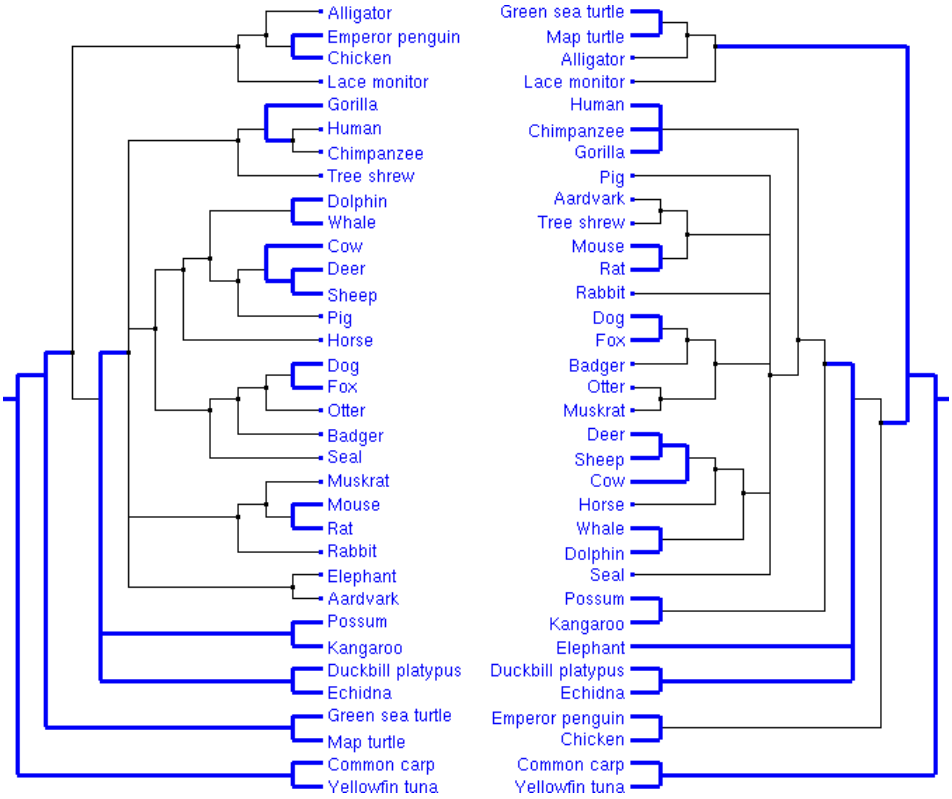


Figure 37. Myoglobin Phylogenetic Tree Comparison

In the tree inferred from myoglobin we see some of the familiar errors evidenced in the other datasets. Once again, alligator (and lace monitor) are clustered with the turtles instead of with the birds. There are also several errors in the major mammal clade. These errors include basic lack of resolution, but also several instances of animals being inserted in this wrong clade. Since there does not seem to be any pattern to the mistakes, instead of simply listing the errors the reader is referred to trees themselves.

Results

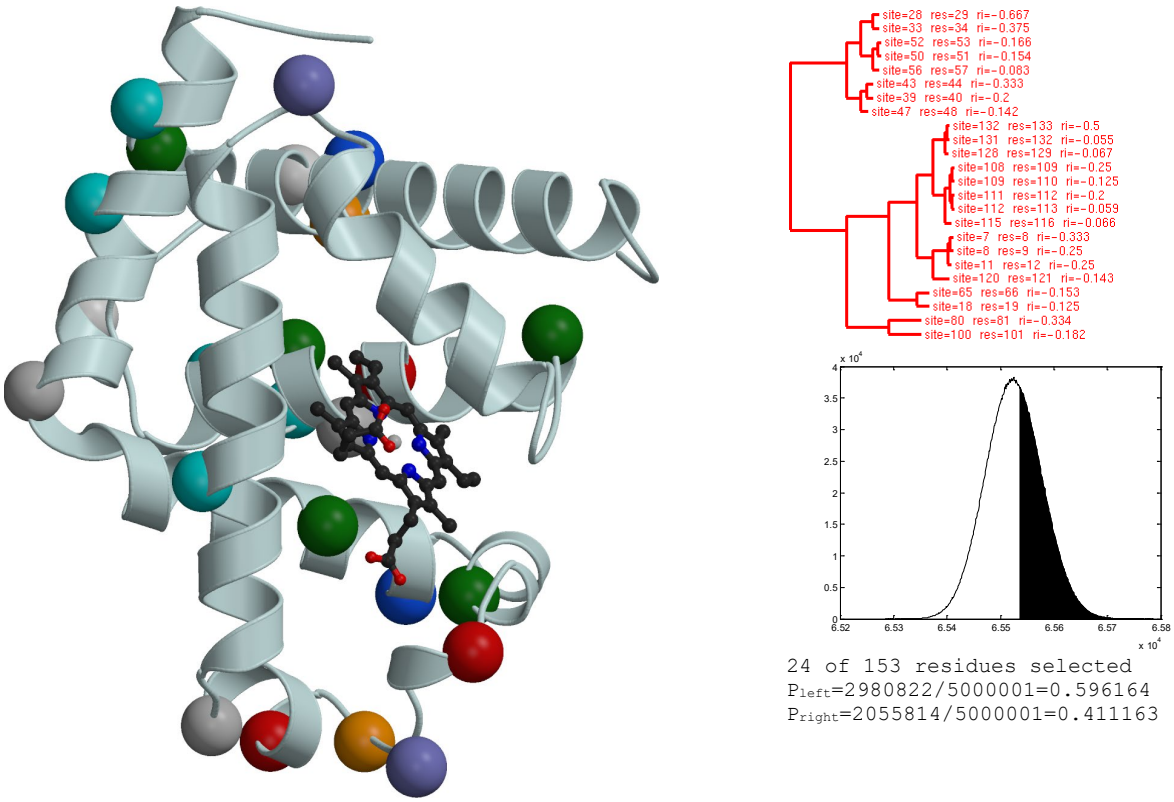


Figure 38. Myoglobin – RI Difference < 0.0 Highlighted

No apparent clustering or patterns are observed.

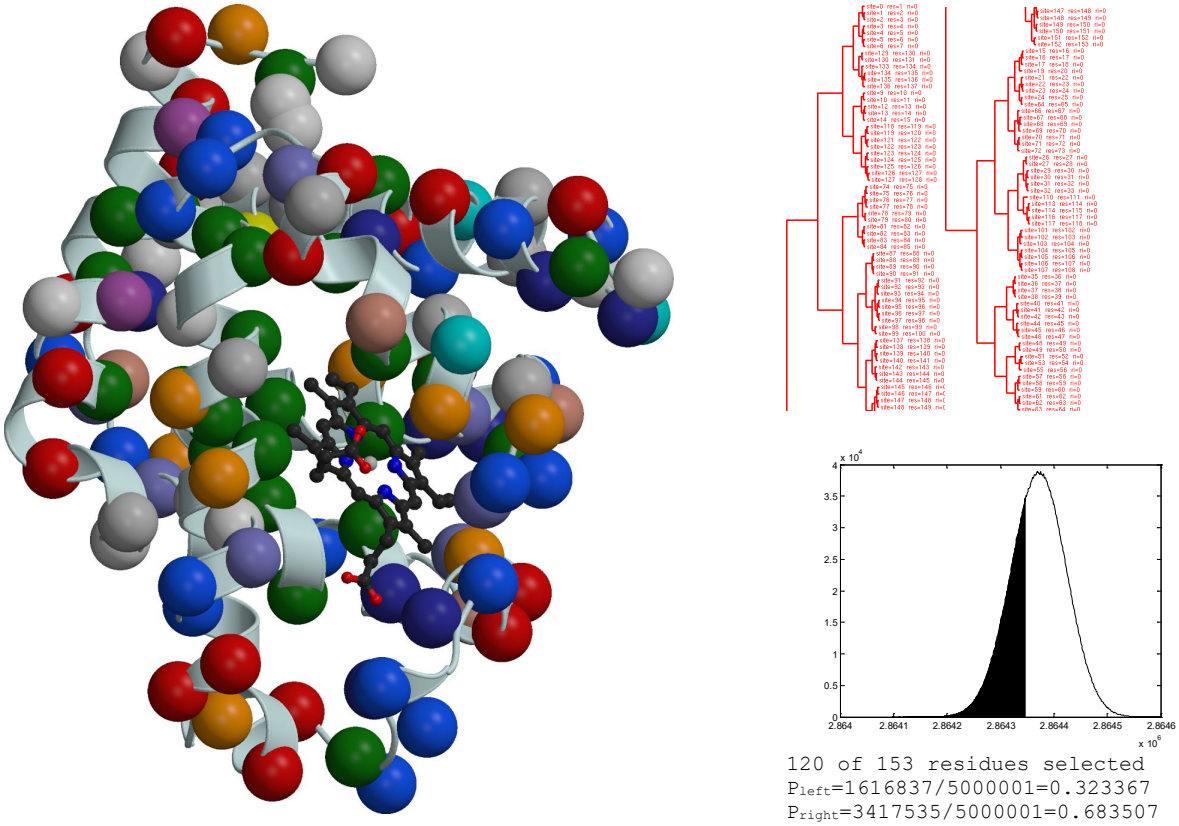


Figure 39. Myoglobin – RI Difference = 0.0 Highlighted

No apparent clustering or patterns are observed.

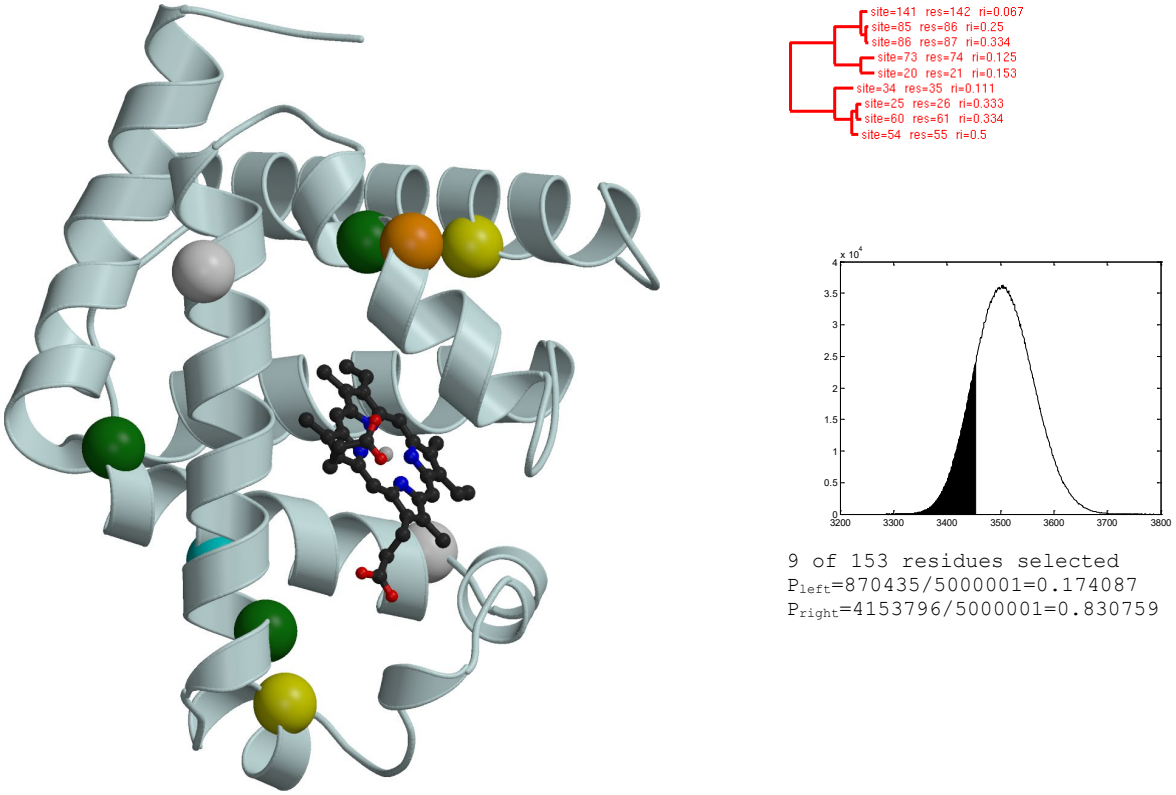


Figure 40. Myoglobin – RI Difference > 0.0 Highlighted

No apparent clustering or patterns are observed.

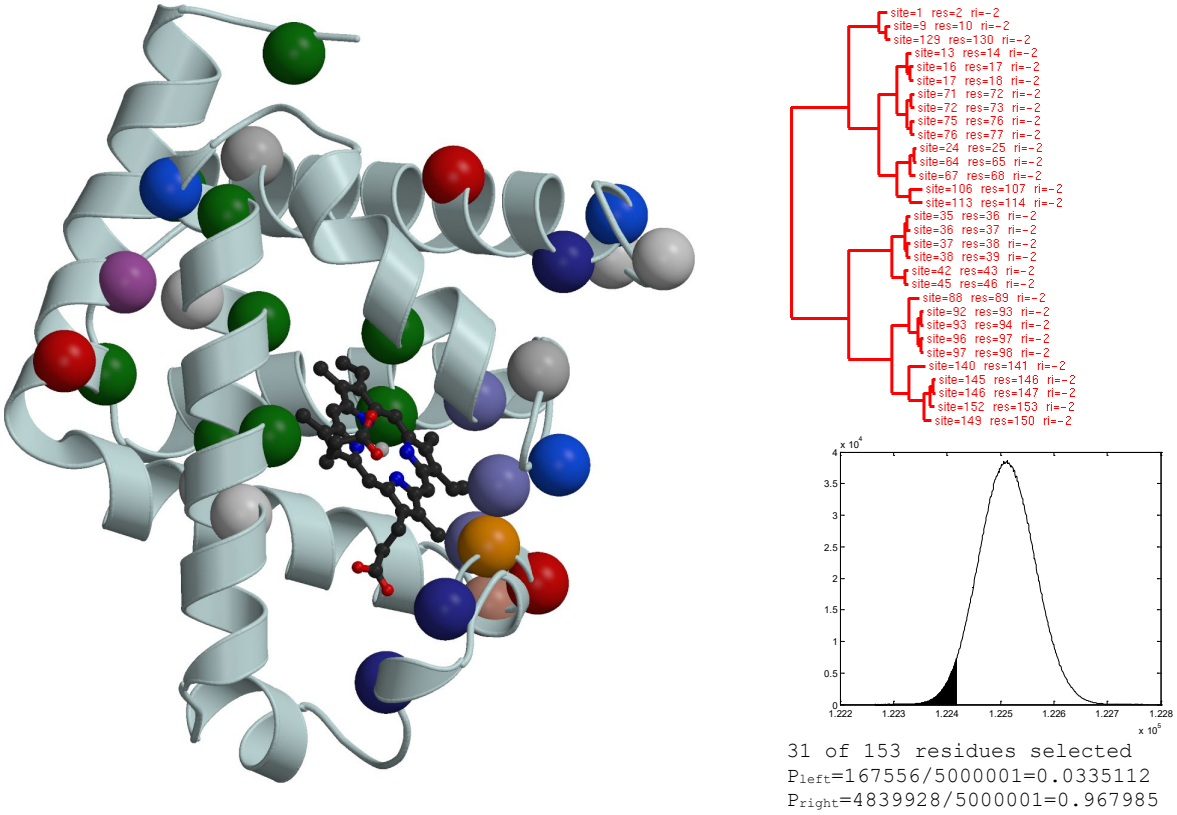


Figure 41. Myoglobin – Sites with No Change Highlighted

A significant clustering according to the p-values can be observed, but we can also see that the invariant residues seem to favor one region of the protein. There does not seem to be a connection between the cluster and the heme group.

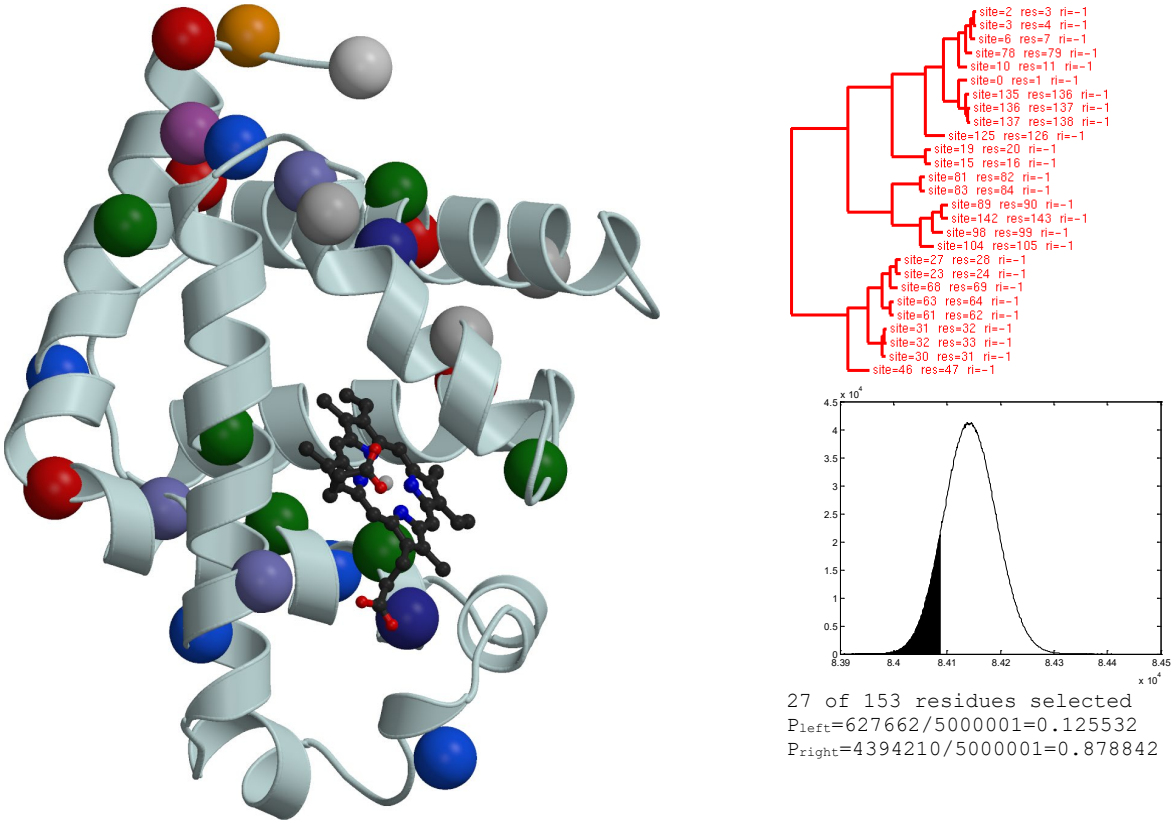


Figure 42. Myoglobin – Sites with RI = ∞ Highlighted

There is no apparent clustering of the residues with undefined RI values when considering the p-values or through visual inspection. The p-values decrease if the residues with max > 1 are removed.

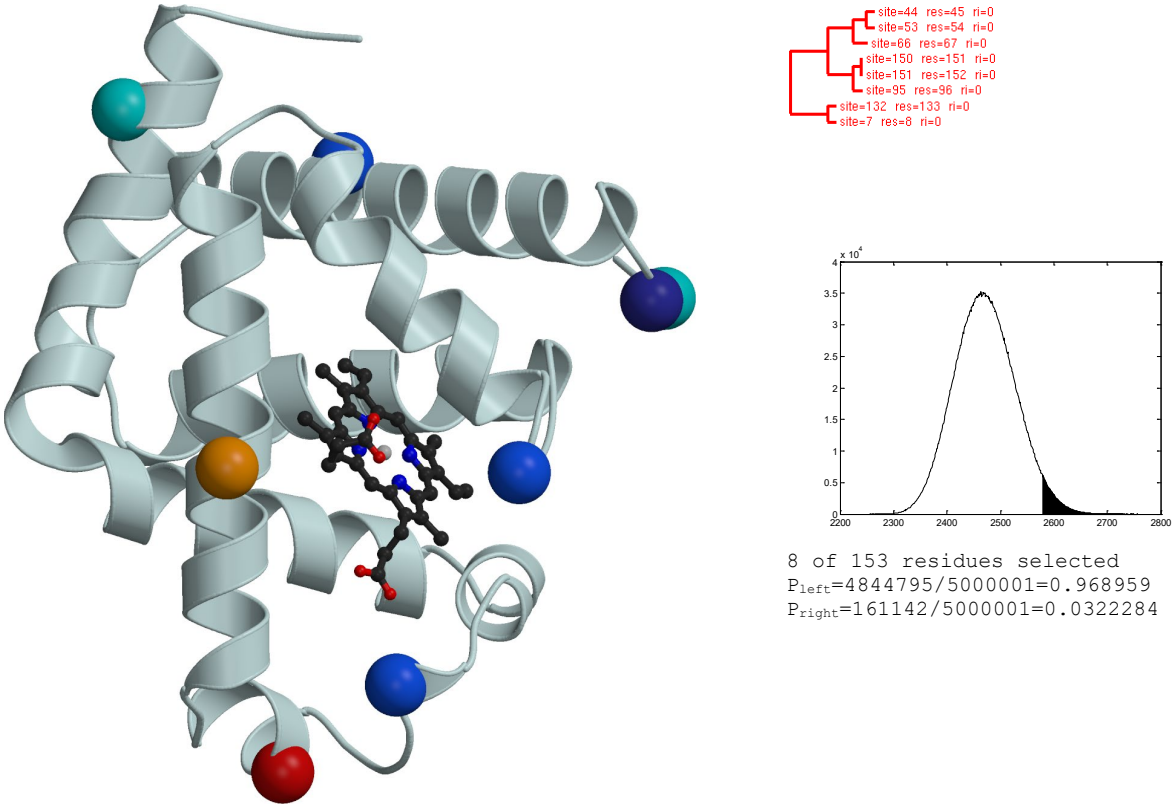


Figure 43. Myoglobin – Sites with True Tree RI = 0.0 Highlighted

Residues are distributed to a greater extent than would be expected by random. The significance increases as residues with non-zero RI values are added but decreases when the residues with RI=1 are considered.

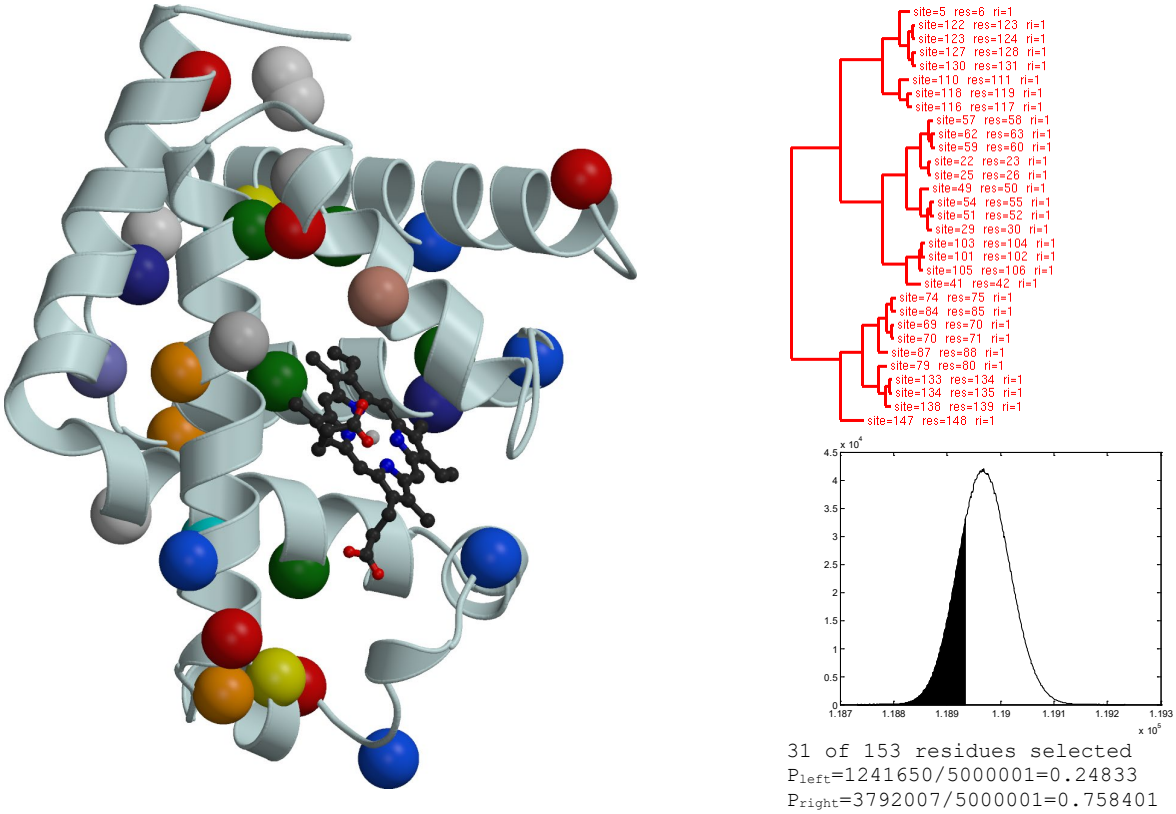
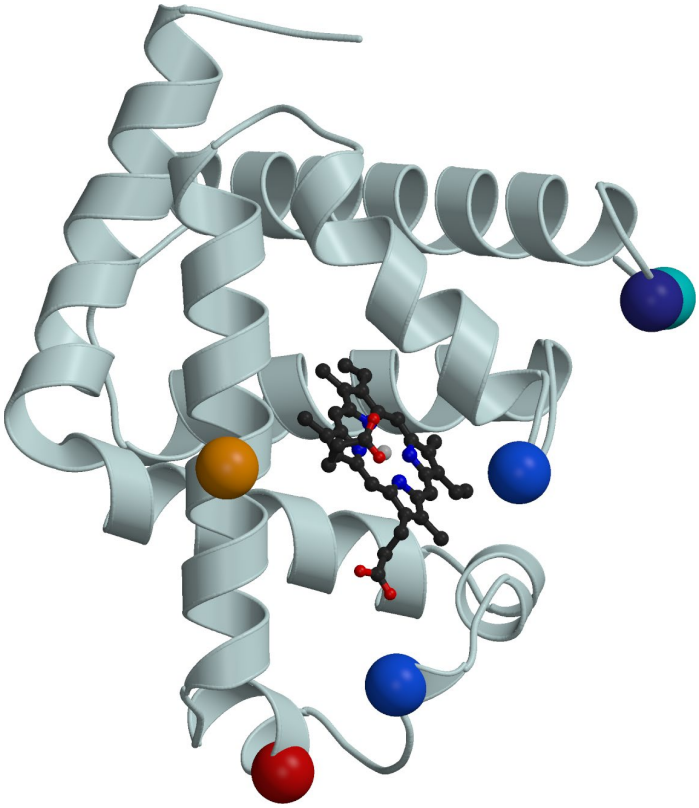
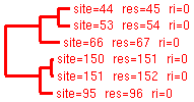
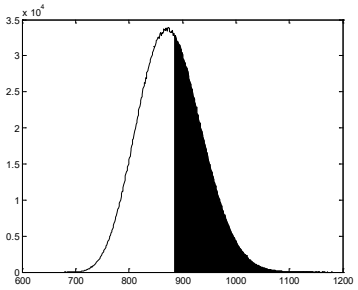


Figure 44. Myoglobin – Sites with True Tree RI = 1.0 Highlighted

No apparent clustering or patterns are observed. However, as mentioned when discussing the RI_{true}=0.0 category as additional non-zero residues are added the p-values indicate a shift to a more dispersed arrangement of the residues.







6 of 153 residues selected
P_{left}=2872110/5000001=0.574422
P_{right}=2160637/5000001=0.432127

Figure 45. Myoglobin – Sites with MPC Tree RI = 0.0 Highlighted

No apparent clustering or patterns are observed.

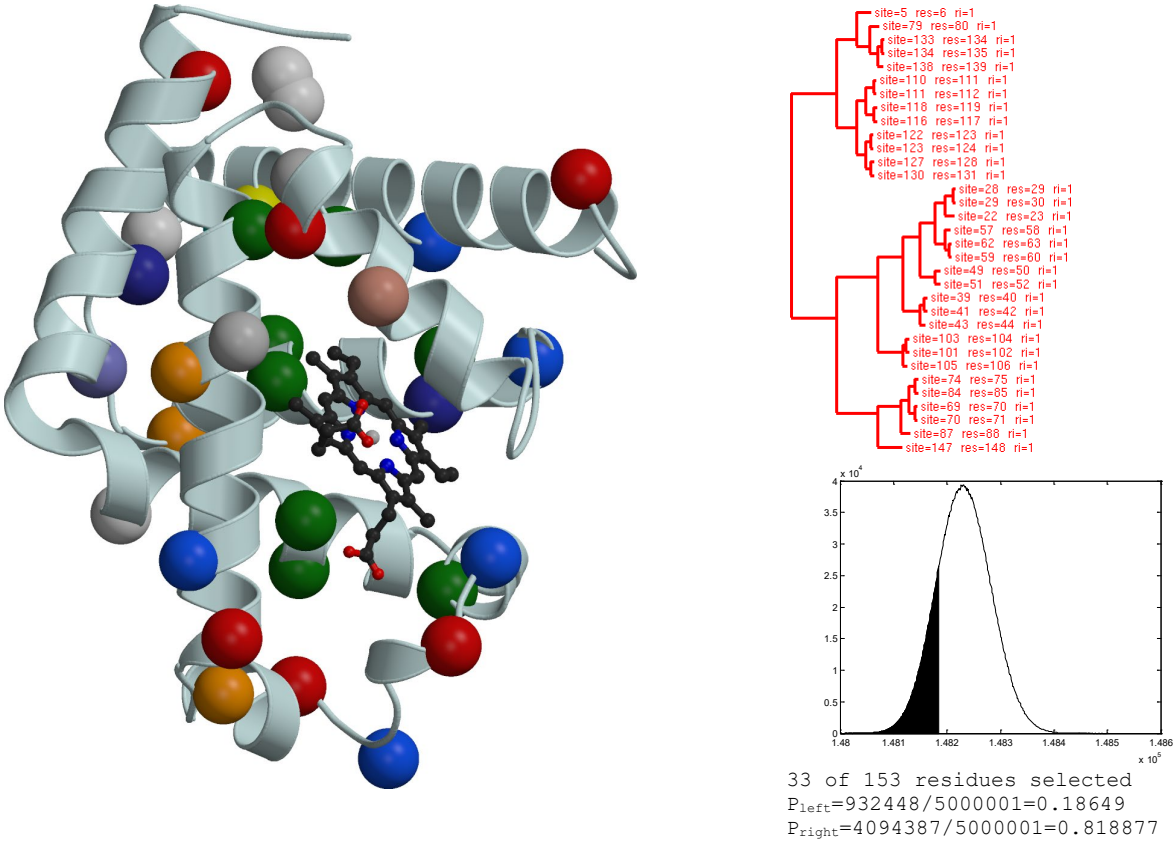


Figure 46. Myoglobin – Sites with MPC Tree RI = 1.0 Highlighted

No apparent clustering or patterns are observed.

Hemoglobin Dataset

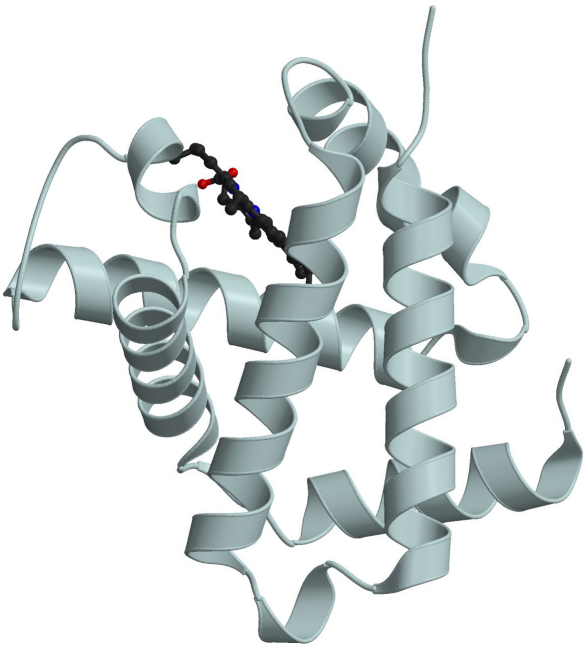


Figure 47. Hemoglobin α

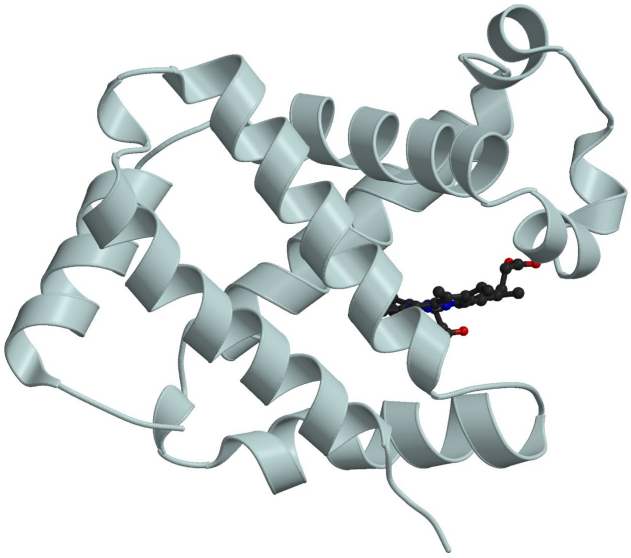


Figure 48. Hemoglobin β

Functional and Structural Description

Hemoglobin is related to myoglobin both functionally and structurally. Like myoglobin, hemoglobin binds oxygen (O_2). Hemoglobin transports oxygen from the oxygen rich environment of the lungs to tissues, exchanges oxygen for carbon dioxide waste, and returns to the lungs to once again trade the carbon dioxide for additional oxygen.

The structure of hemoglobin is a tetramer (four polypeptide chains) composed of two identical α chains and two identical β chains. The α and β chains are very similar with 141 and 146 amino acid residues, respectively, and both have eight α-helices. Each of the four chains fold to contain a site for binding oxygen called the heme pocket. The heme pocket is composed of carbon, nitrogen, and hydrogen surrounding a single iron ion. The iron ion is held in place by neighboring nitrogen atoms and its bonding to a histidine residue. Normally, histidine 87 is conserved in the α chain and histidine 92 in the β chain.

The binding properties of hemoglobin are affected by environmental influences such as pH, O_2, and CO_2 levels. Anyone who has run and felt the burn of lactic acid buildup in their muscles will not be surprised that tissues are in a more acidic environment than the lungs. This lower pH in the tissues compared to the lungs helps to trigger the exchange of oxygen and carbon dioxide at the correct times. The driving force of the exchange is called the Bohr effect, and is expressed as:

$$CO_2 + H_2O \leftrightarrow HCO_3^- + H^+$$

In the CO_2 rich tissues, carbon dioxide and water are reacting to form bicarbonate (HCO_3) and hydrogen ions (protons). This reaction increases the acidity of the surrounding tissues, which lowers hemoglobin's affinity for oxygen. During the release of the stored oxygen the protons and bicarbonate are captured ensuring higher support for the right hand side of the reaction. Back at the lungs the process reverses. In the presence of higher oxygen levels, hemoglobin's affinity shifts from proton carrying to oxygen. The protons are shed, reversing the above equation generating carbon dioxide as a gas (CO_2 is insoluble in the bloodstream).

Hemoglobin's affinity for oxygen is not linear. Hemoglobin exhibits a behavior known as cooperativity to bind oxygen. When in an environment of high oxygen levels, partially saturated hemoglobin has a disproportionally high affinity for oxygen while in a low oxygen environment, hemoglobin has a disproportionally low affinity for oxygen. The relationship is characterized by the Hill Equation:

$$Y_{O_2} = \frac{(pO_2)^n}{(p_{50})^n + (pO_2)^n}$$

where pO_2 is the partial pressure of O_2, p_{50} is the pO_2 of 50% saturation, and n is referred to as the Hill coefficient and is a measure of the cooperativity of the particular hemoglobin. A normal range of values for n is [2.8,3.0] and is related to the number of ligands simultaneously binding oxygen, and thus is limited by the number of subunits, namely four which would represent maximum cooperativity. This highly sensitive cooperativity of hemoglobin is assumed to be an evolved specialization of hemoglobin to reduce the volume needed to transport the same quantity of oxygen.

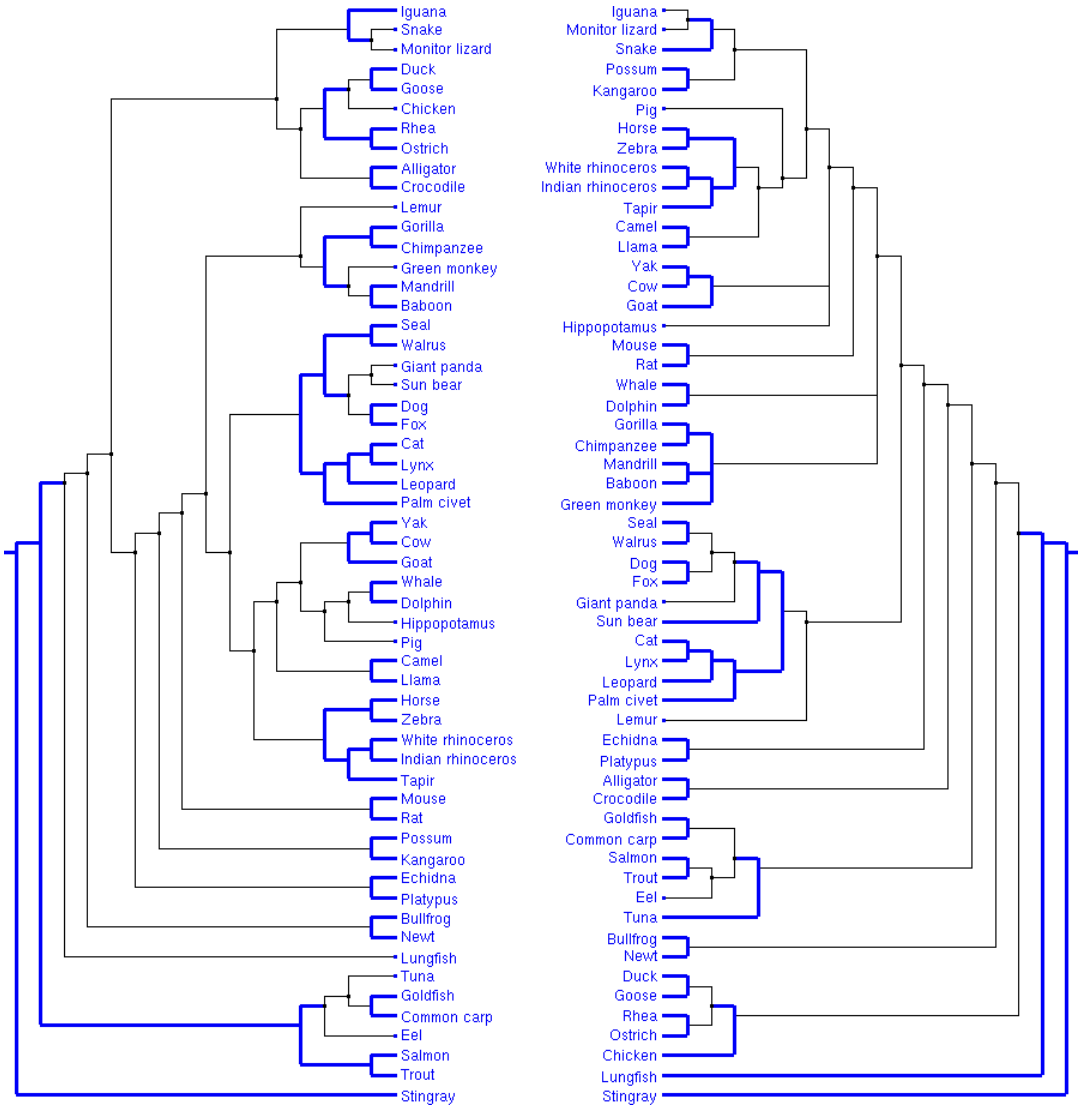


Figure 49. Hemoglobin α-Chain Phylogenetic Tree Comparison

Because the same set of organisms was used for both the hemoglobin α and hemoglobin β datasets, we are fortunate to be able to have the same true tree between datasets. An interesting observation that can be made by comparing the inferred trees for both of these datasets to the true tree is that the hemoglobin α dataset is able to more closely reconstruct the correct topology. We can place a qualitative measure on this observation by simply counting the number of clades that are correct (indicated by the highlighted lines) and comparing the values. Doing so we see that the hemoglobin α dataset has over twice as many correct clades as does the hemoglobin β dataset. The problems which are apparent in the hemoglobin α dataset include confusion about the relationship of the marsupials, snakes, and birds. Also, while the hemoglobin α dataset

allows reconstruction of several of the minor clades such as cats, birds, *etc.*, the fine detail within these clades is occasionally incorrect.

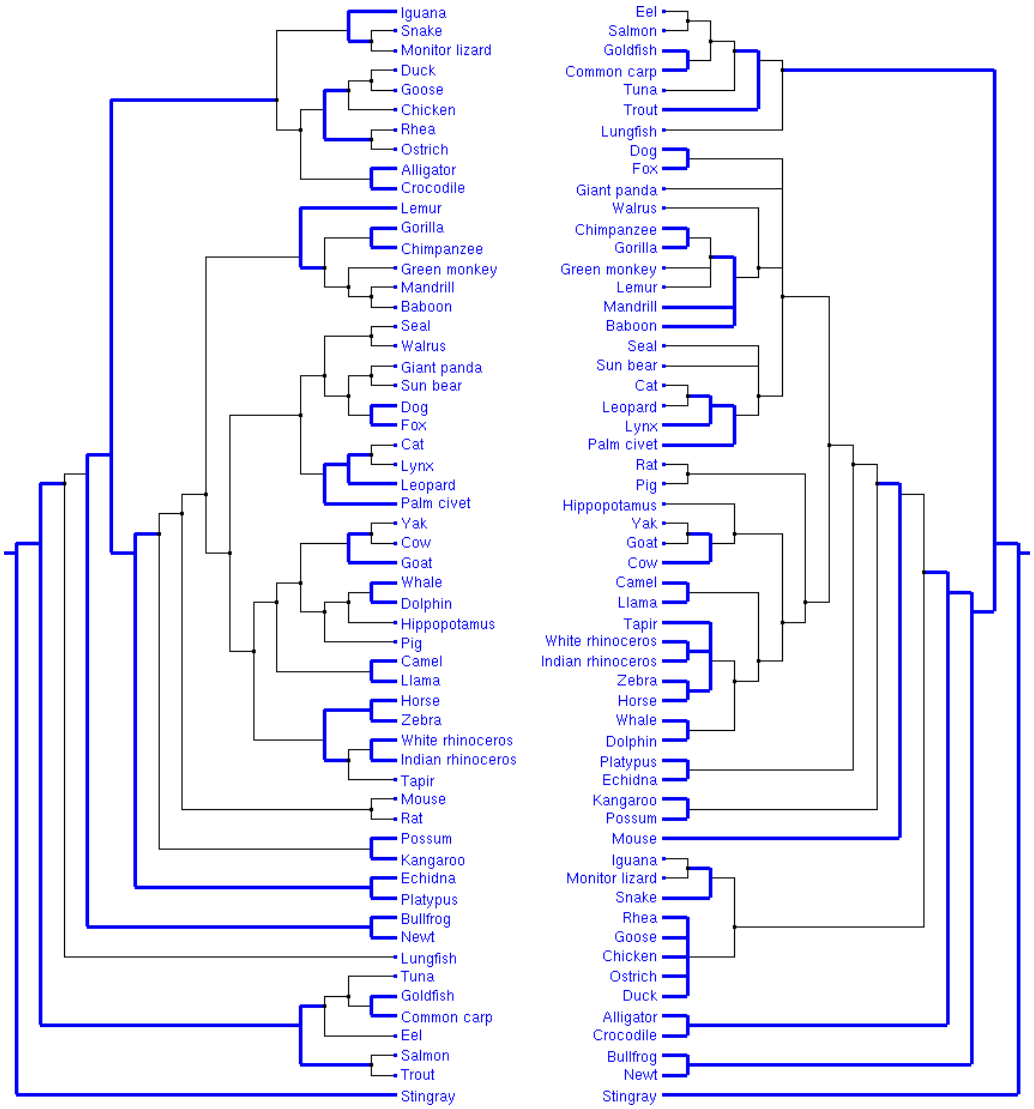


Figure 50. Hemoglobin β-Chain Phylogenetic Tree Comparison

Reconstruction of the evolutionary relationships given the hemoglobin β data was not as good as for hemoglobin α. Several polytomies exist, few of the minor clades are completely correct, and many are simply not present. This suggests that there are additional violations of the assumptions required by the phylogenetic inference procedure. This further suggests that constraints may exist in hemoglobin β that are not present in hemoglobin α.

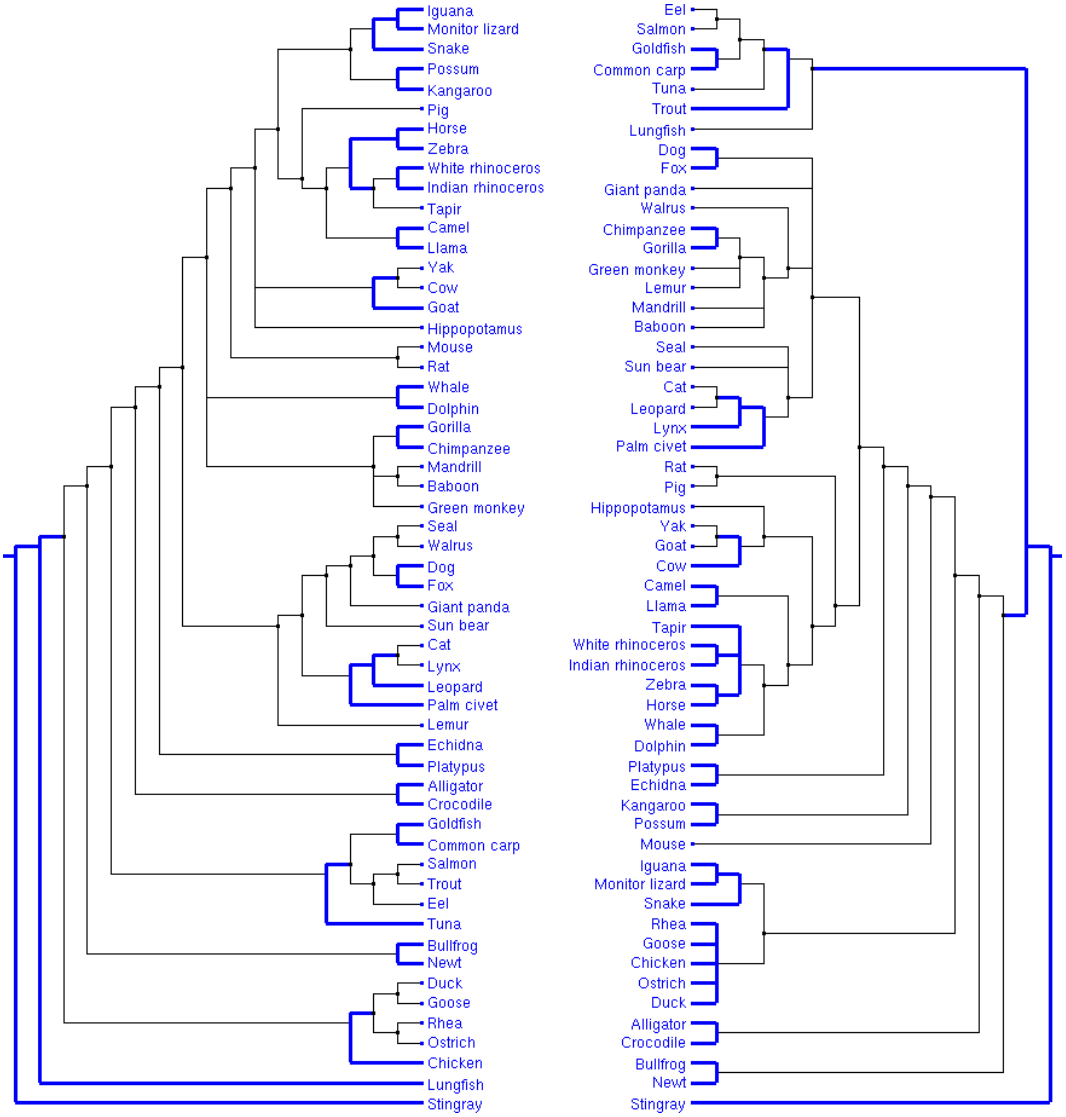


Figure 51. Hemoglobin α- β-Chain MPC Phylogenetic Tree Comparison

In the previous sections we have compared the true and MPC topologies. In the case of the hemoglobin α and hemoglobin β datasets, a further comparison is possible due to their close relationship and the same organisms having been used in both datasets. The MPC trees can be directly compared with each other just as the true and MPC trees were compared before. We can see that there are larger polytomies in the hemoglobin β and there is not a single major clade that is in complete agreement between the two trees.

α-chain Results

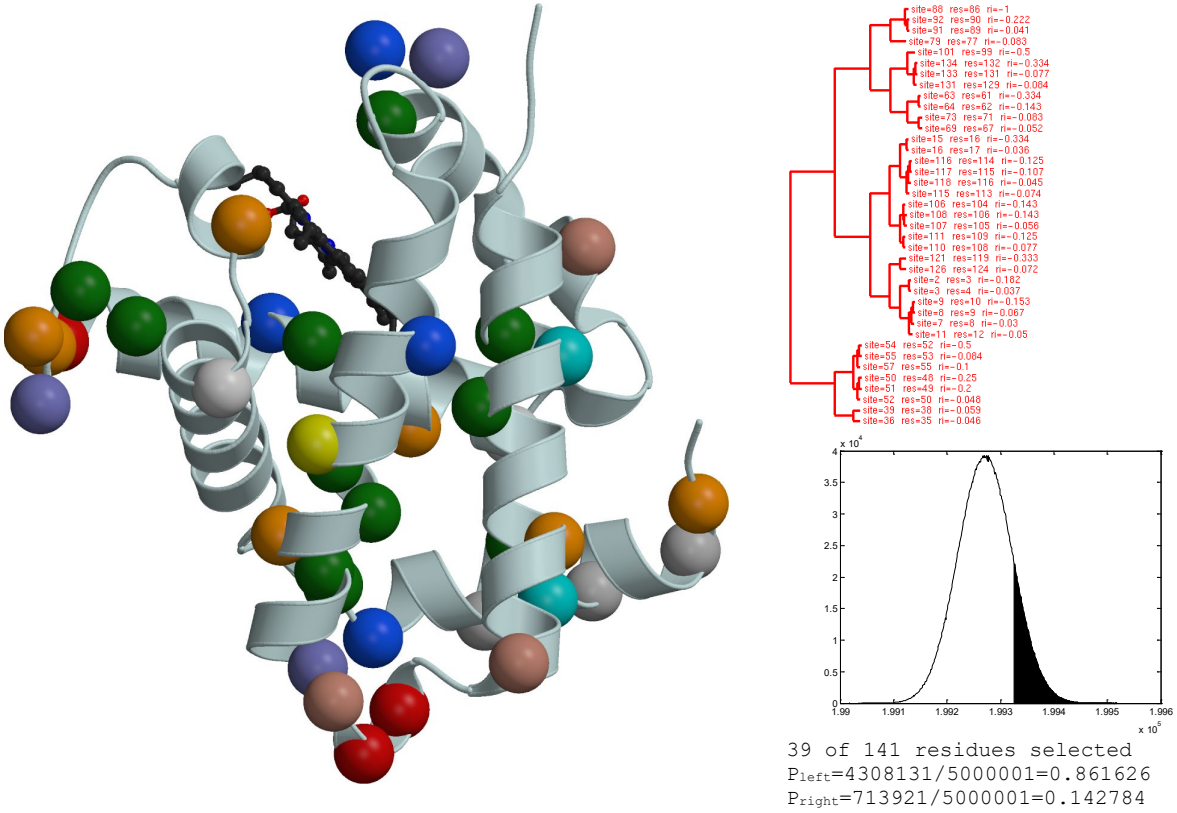


Figure 52. Hemoglobin α – RI Difference < 0.0 Highlighted

No apparent clustering or patterns are observed.

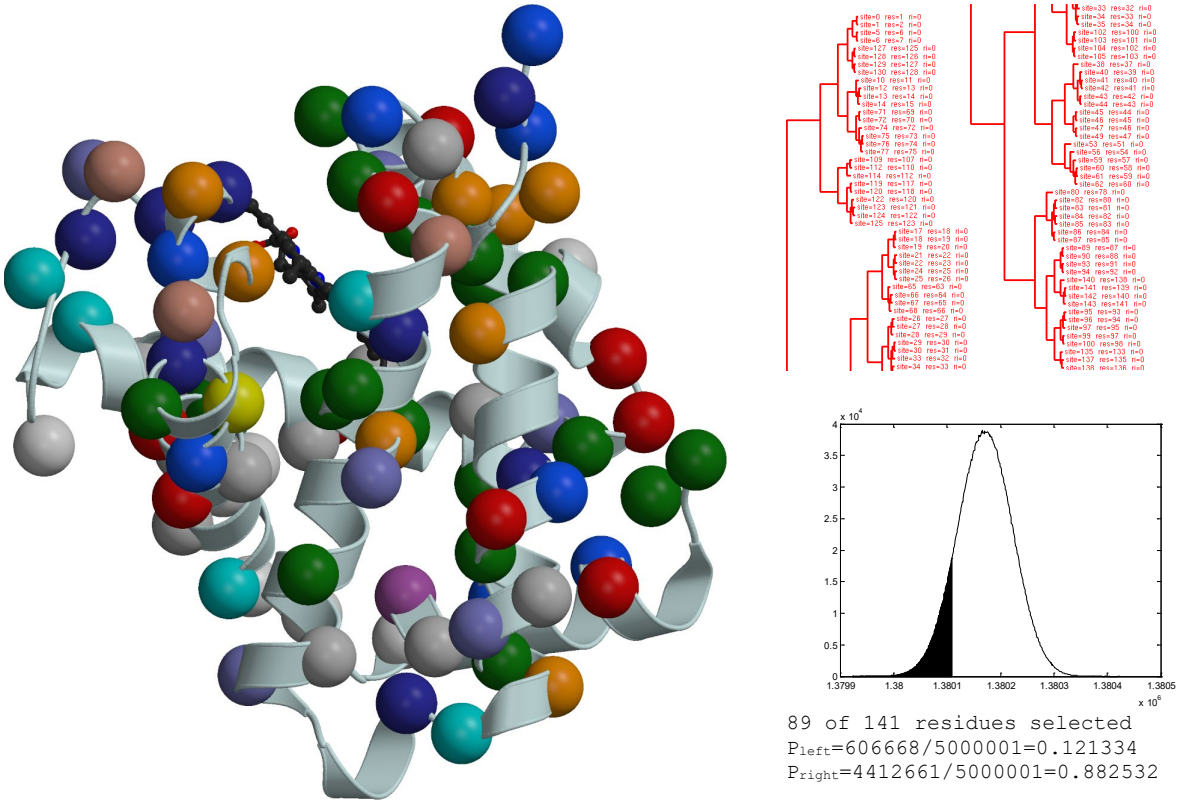


Figure 53. Hemoglobin α – RI Difference = 0.0 Highlighted

No apparent clustering or patterns are observed.

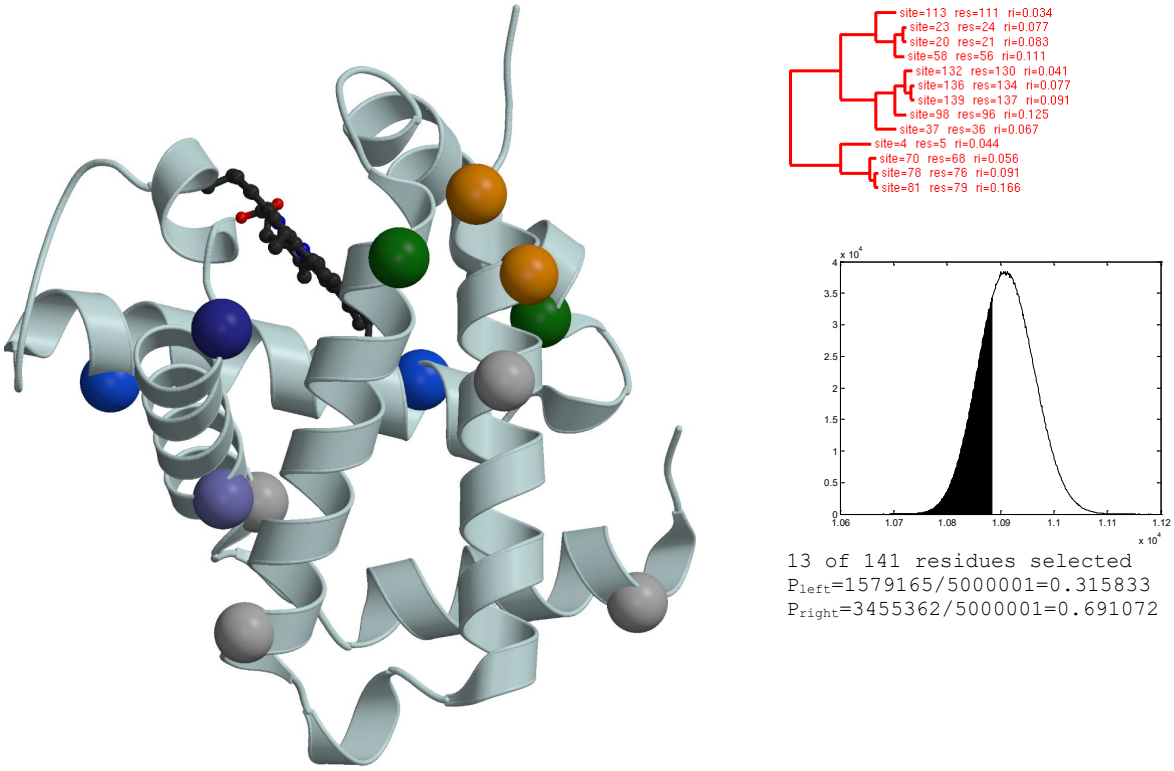


Figure 54. Hemoglobin α – RI Difference > 0.0 Highlighted

No apparent clustering or patterns are observed.

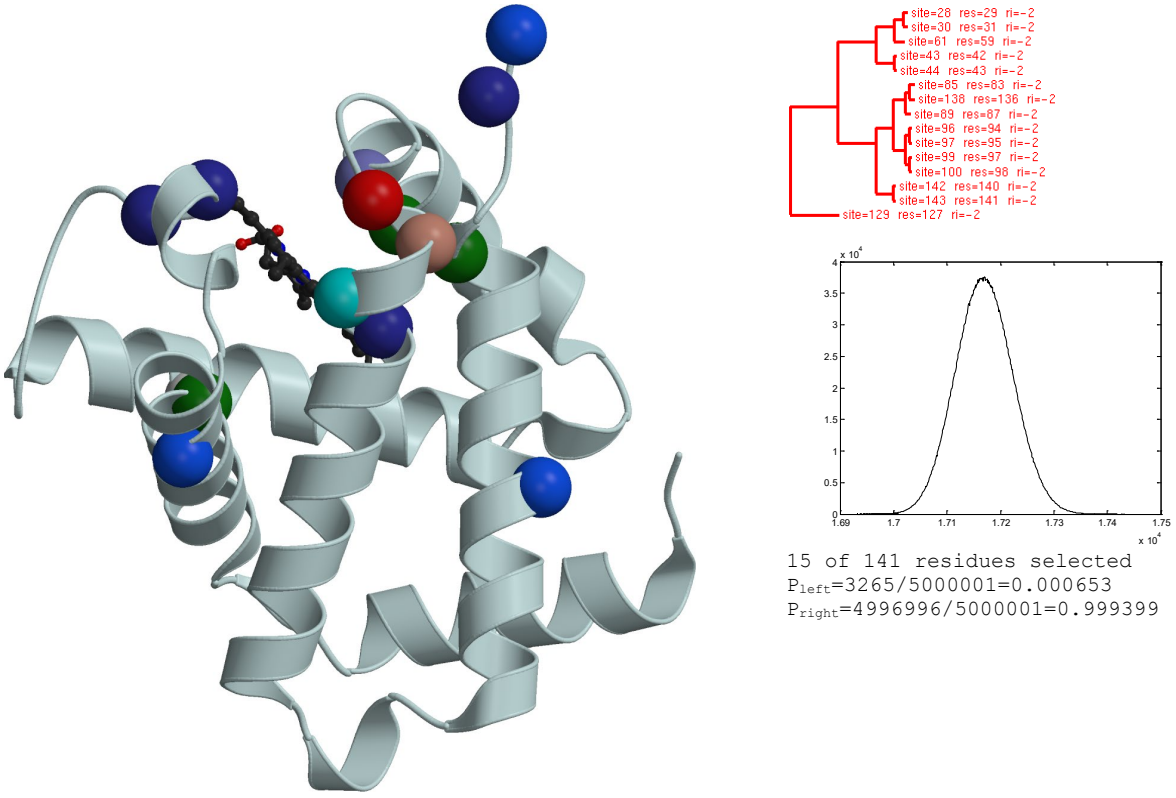


Figure 55. Hemoglobin α – Sites with No Change Highlighted

The invariant residues exhibit significantly tight clustering in an area close to the heme group. This is not surprising considering the chemical constraints required to hold the heme. However, this would not explain all the invariant residues. This is the same observation as with the hemoglobin β dataset.

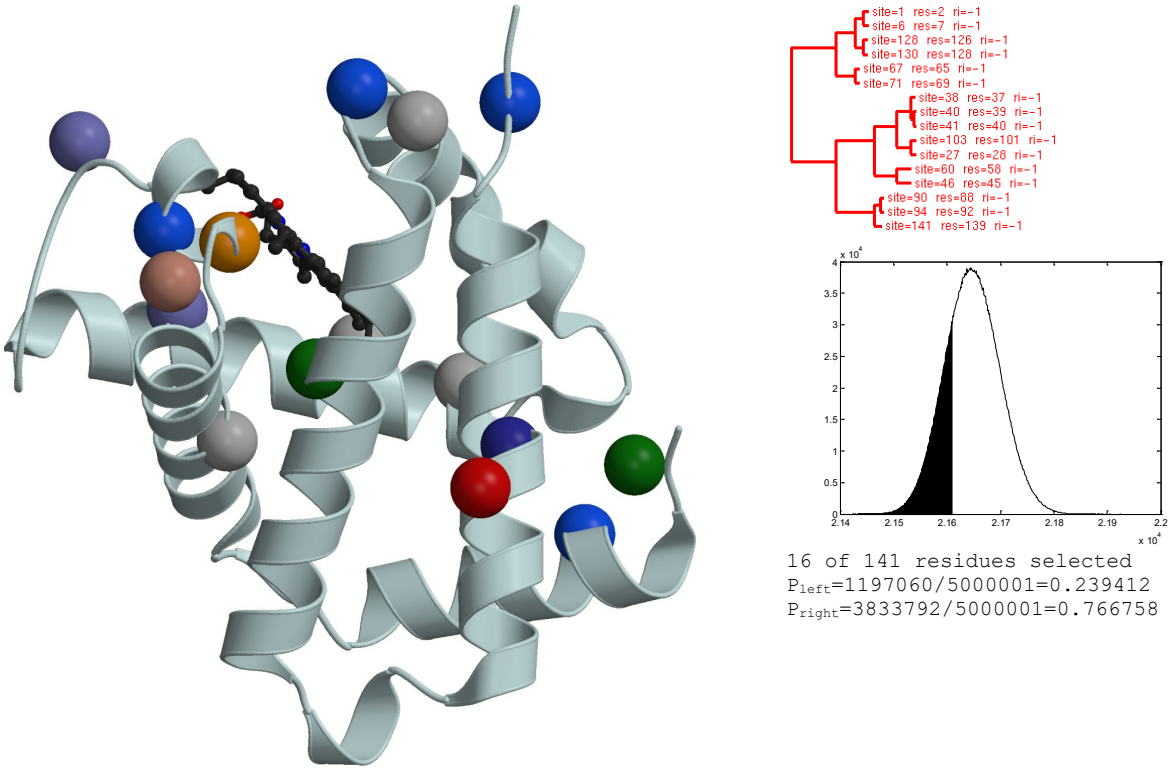


Figure 56. Hemoglobin α – Sites with RI = ∞ Highlighted

While according to the p-values there is no significant clustering apparent with the residues which undefined RI values, if one inspects the residues visually there does appear to be a clustering toward the side of the protein with the heme group.

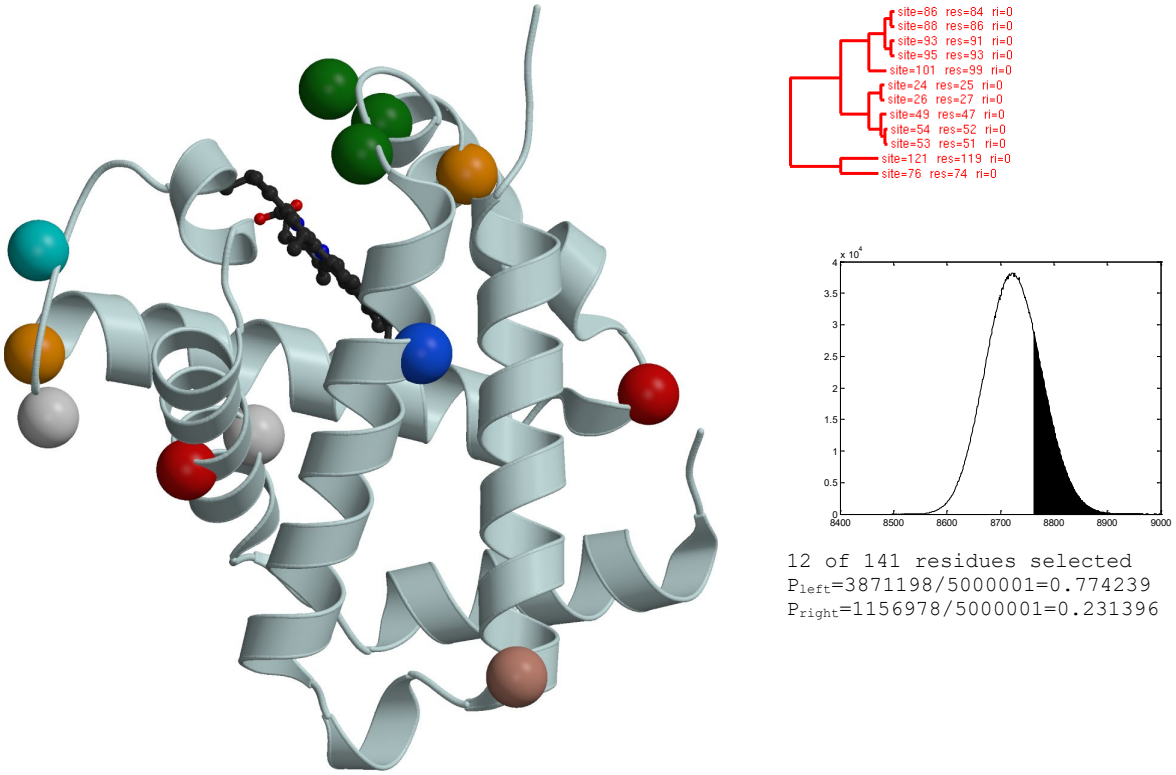


Figure 57. Hemoglobin α – Sites with True Tree RI = 0.0 Highlighted

Again, according to the p-values there is no significant clustering, but if inspected visually there does seem to be a clustering of the residues toward the side of the protein associated with the heme group.

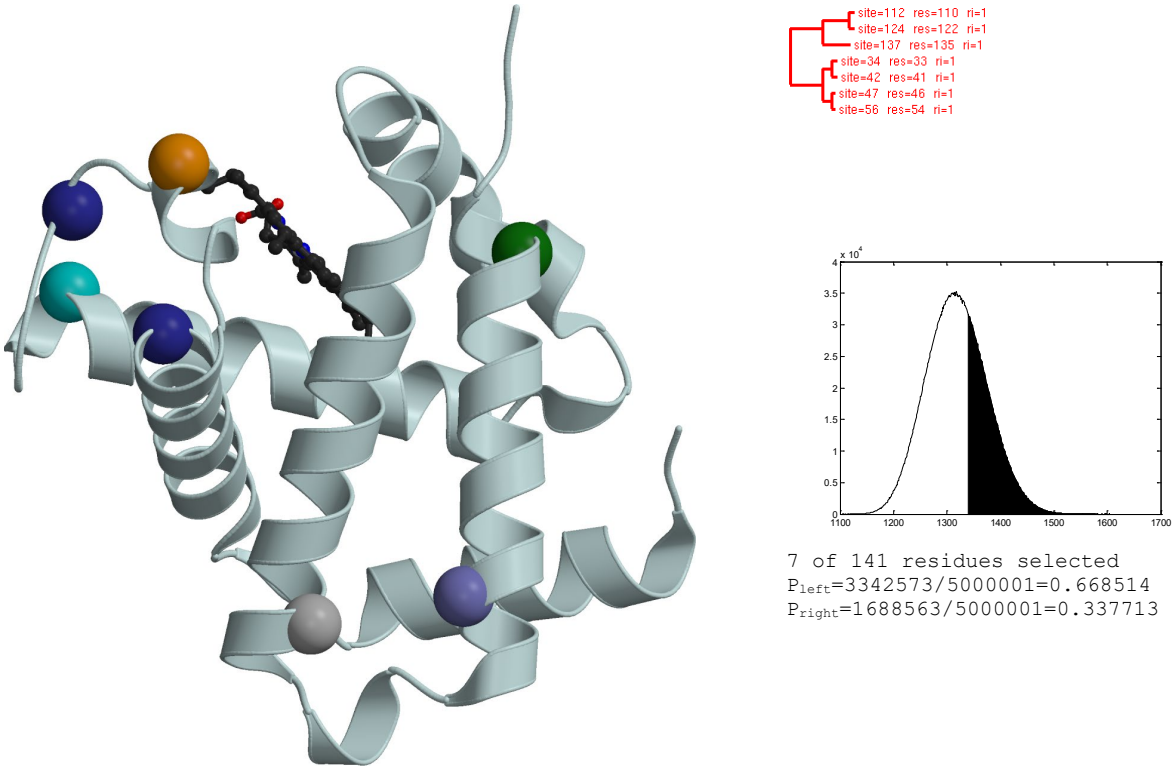


Figure 58. Hemoglobin α – Sites with True Tree RI = 1.0 Highlighted

Again, according to the p-values there is no significant clustering, but if inspected visually there does seem to be a clustering of the residues toward the side of the protein associated with the heme group.

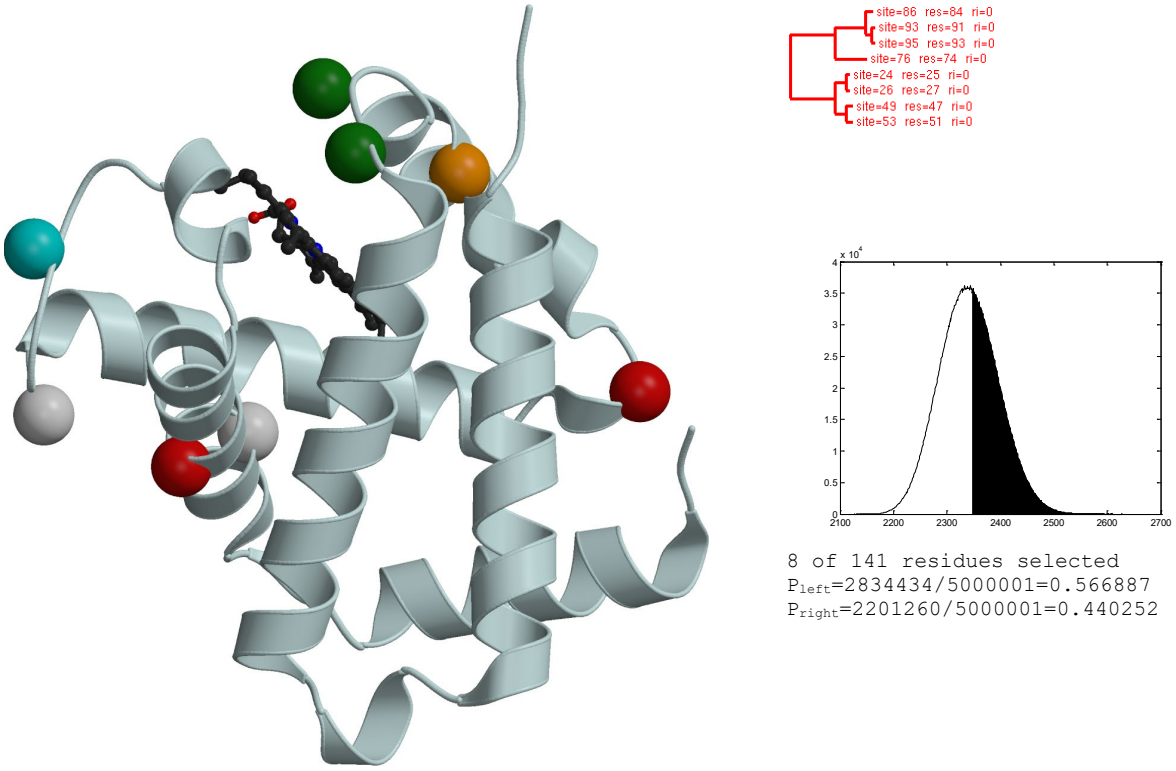


Figure 59 Hemoglobin α – Sites with MPC Tree RI = 0.0 Highlighted

Again, according to the p-values there is no significant clustering, but if inspected visually there does seem to be a clustering of the residues toward the side of the protein associated with the heme group.

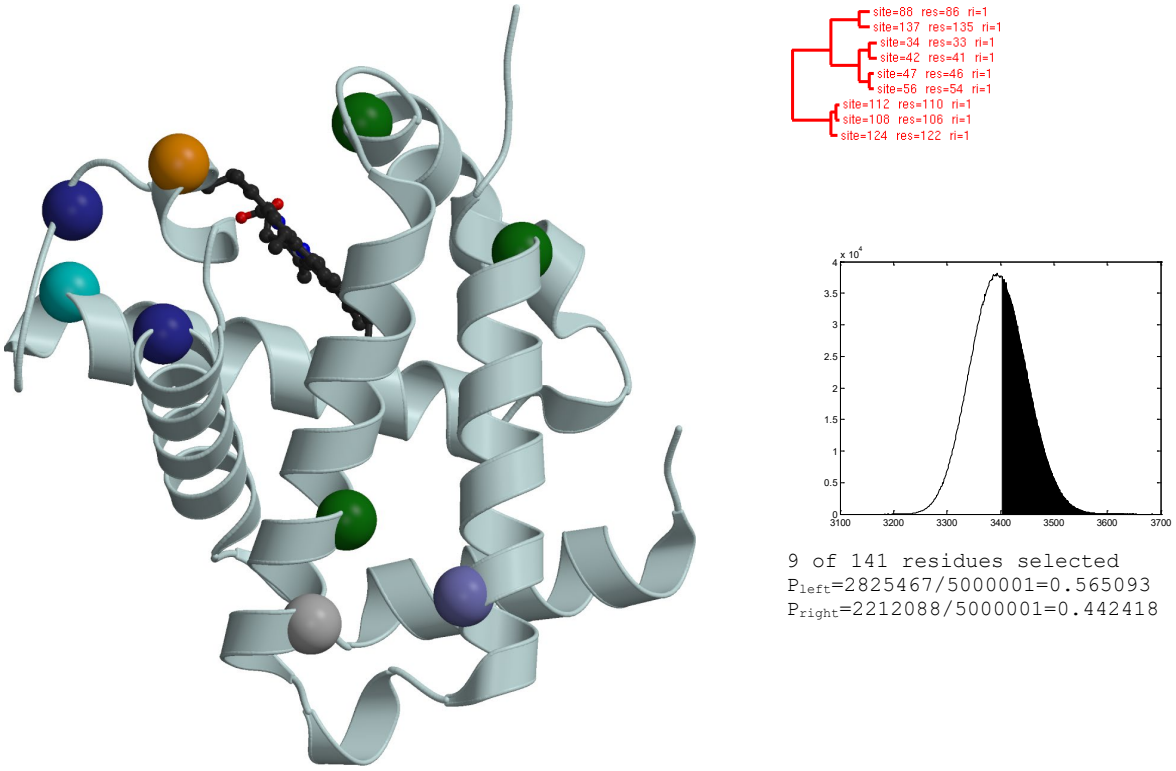


Figure 60. Hemoglobin α – Sites with MPC Tree RI = 1.0 Highlighted

No apparent clustering or patterns are observed.

β-chain Results

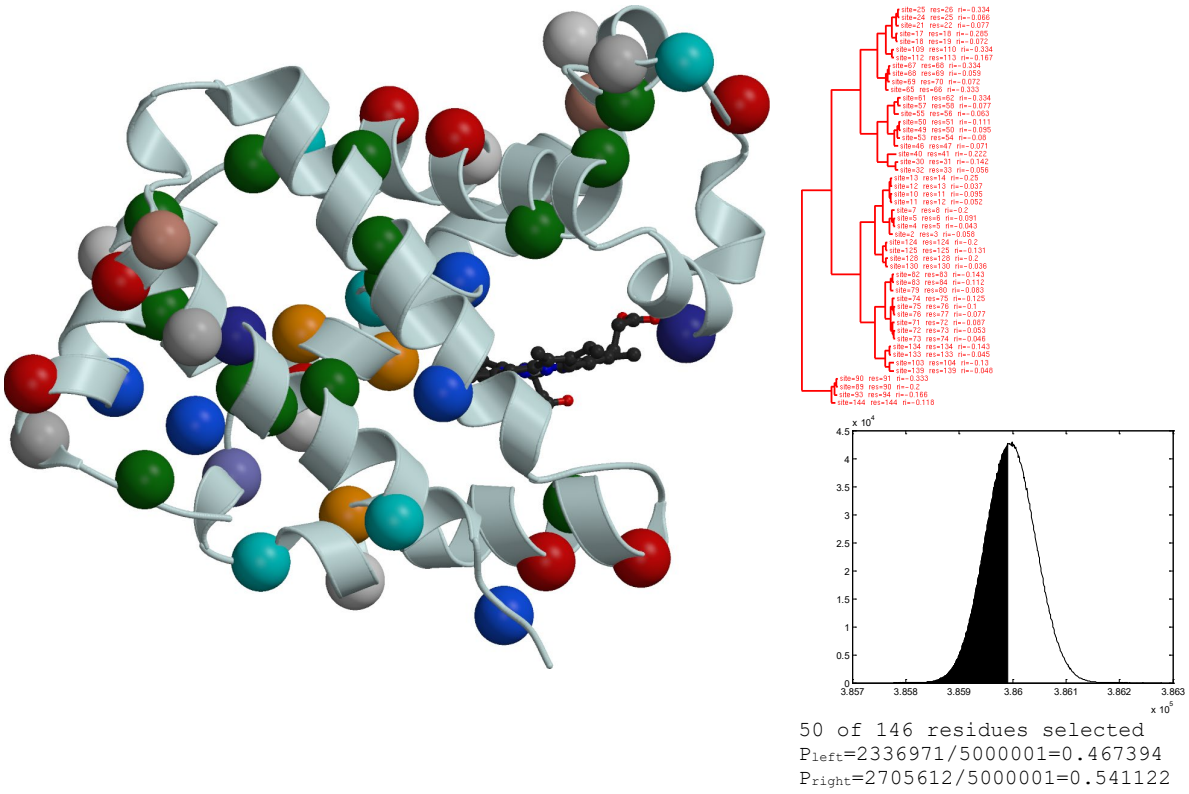


Figure 61. Hemoglobin β – RI Difference < 0.0 Highlighted

No apparent clustering or patterns are observed.

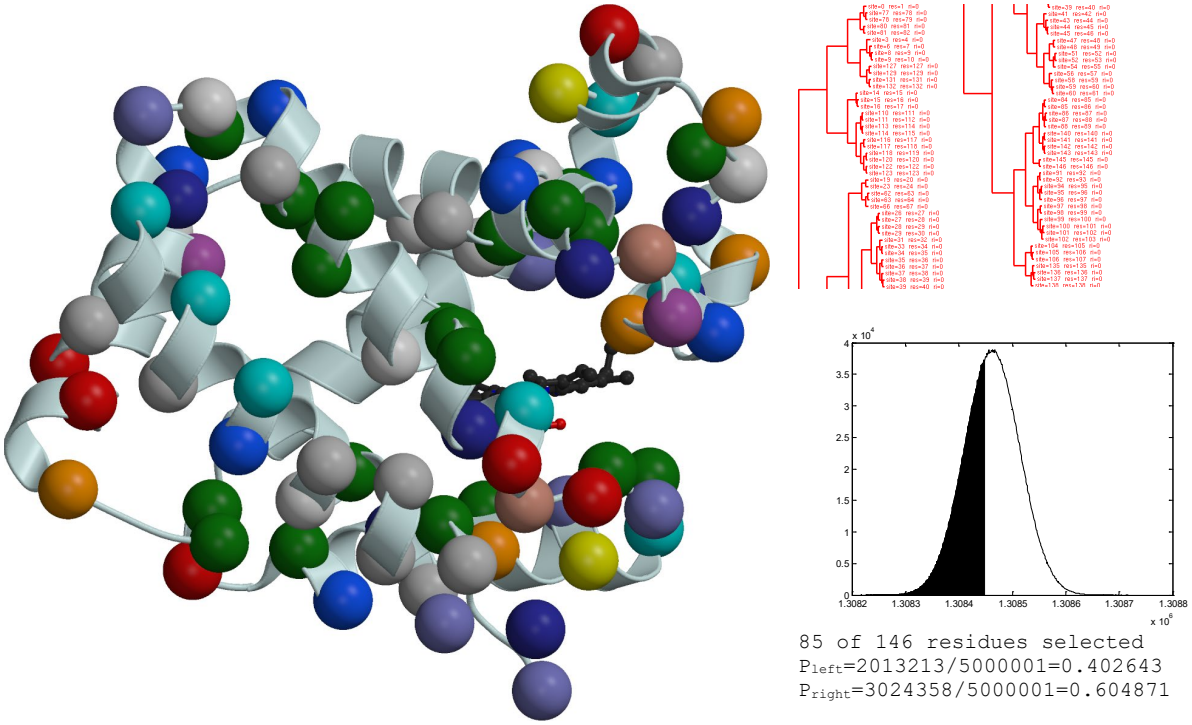


Figure 62. Hemoglobin β – RI Difference = 0.0 Highlighted

No apparent clustering or patterns are observed.

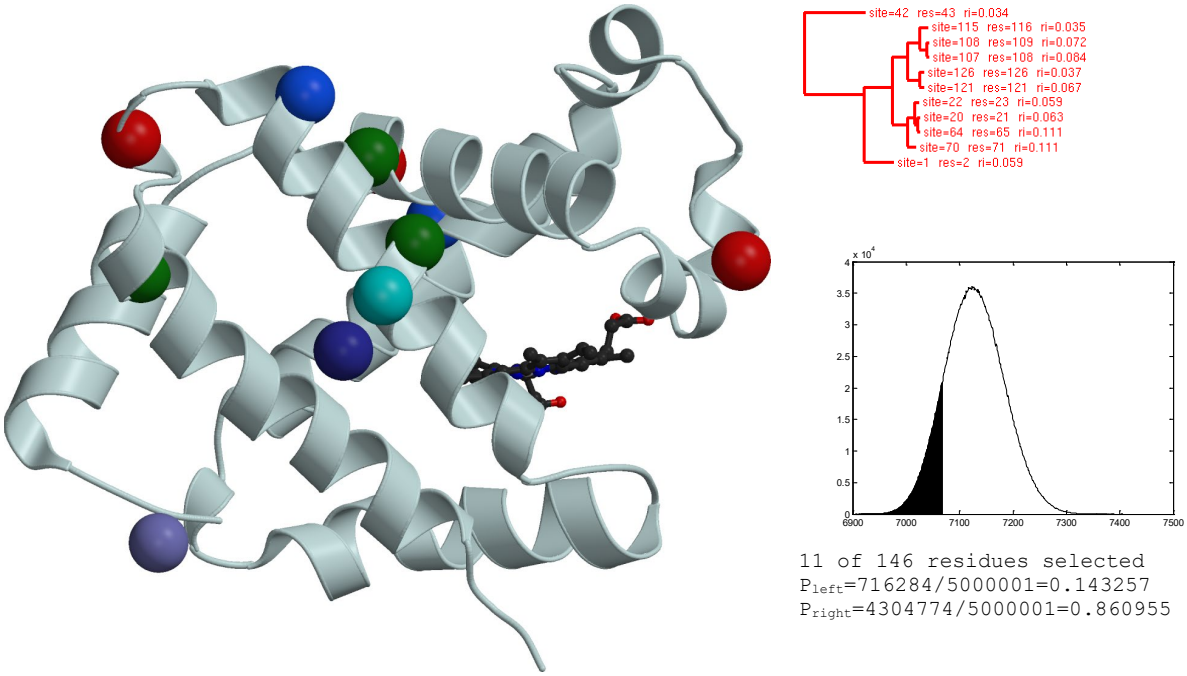


Figure 63. Hemoglobin β – RI Difference > 0.0 Highlighted

No apparent clustering or patterns are observed.

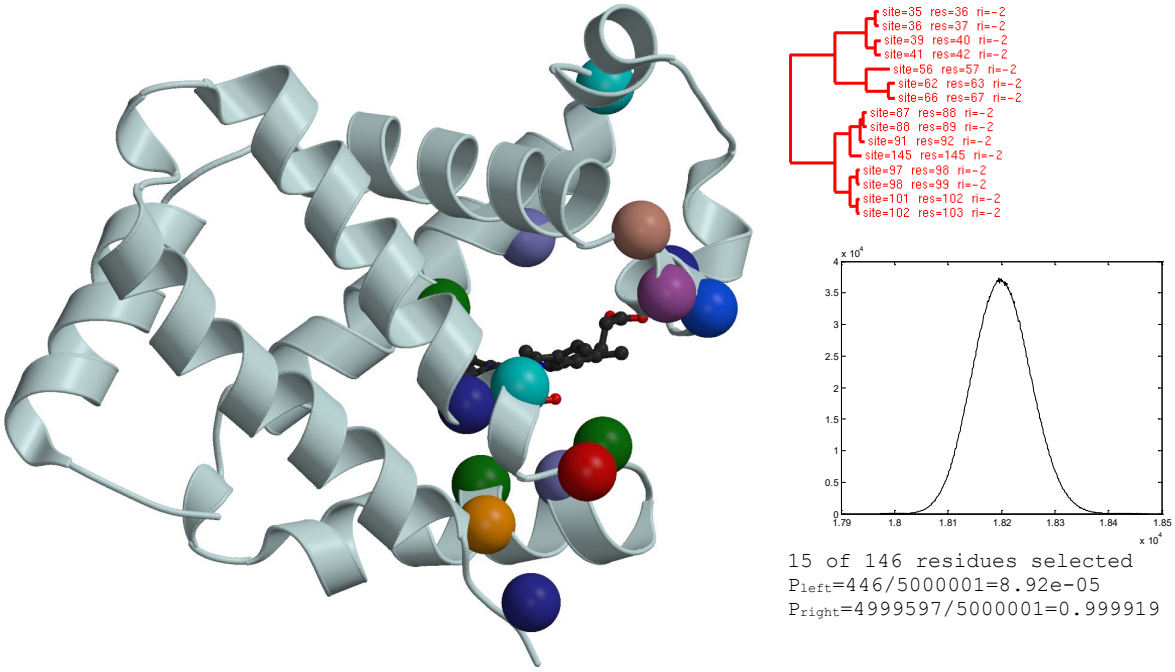


Figure 64. Hemoglobin β – Sites with No Change Highlighted

The invariant residues exhibit significantly tight clustering in an area close to the heme group. This is not surprising considering the chemical constraints necessary to hold the heme. However, this would not explain all the invariant residues. This is the same observation as with the hemoglobin α dataset.

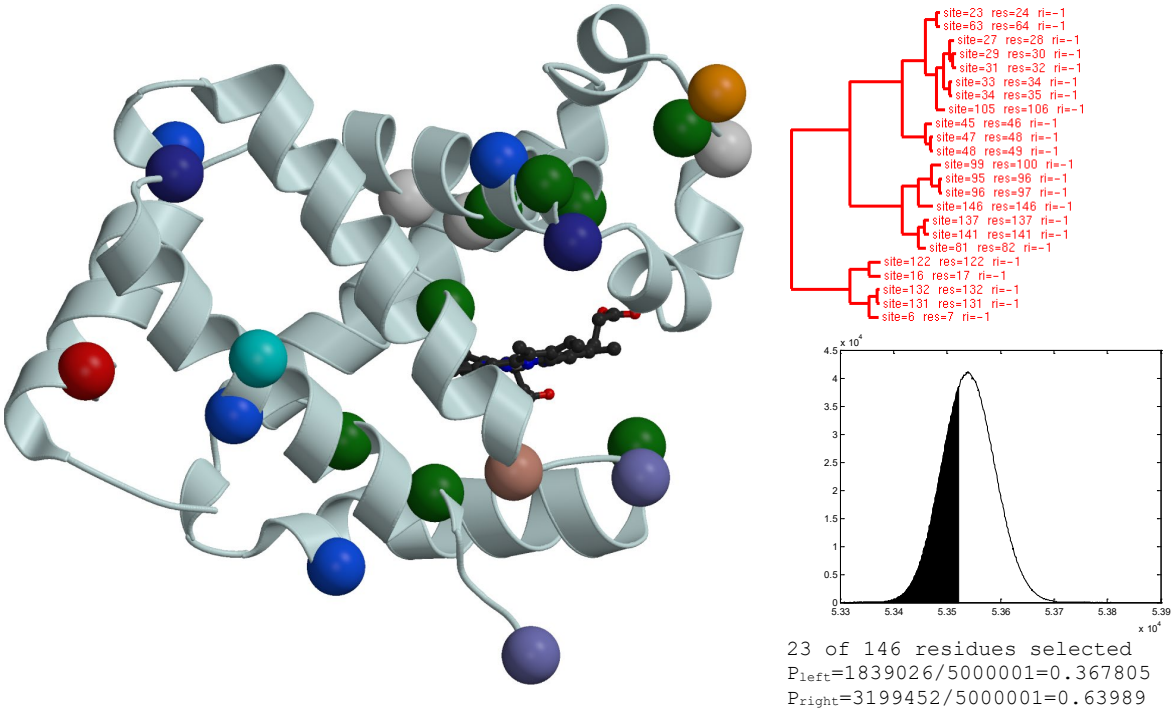


Figure 65. Hemoglobin β – Sites with RI = ∞ Highlighted

No apparent clustering or patterns are observed.

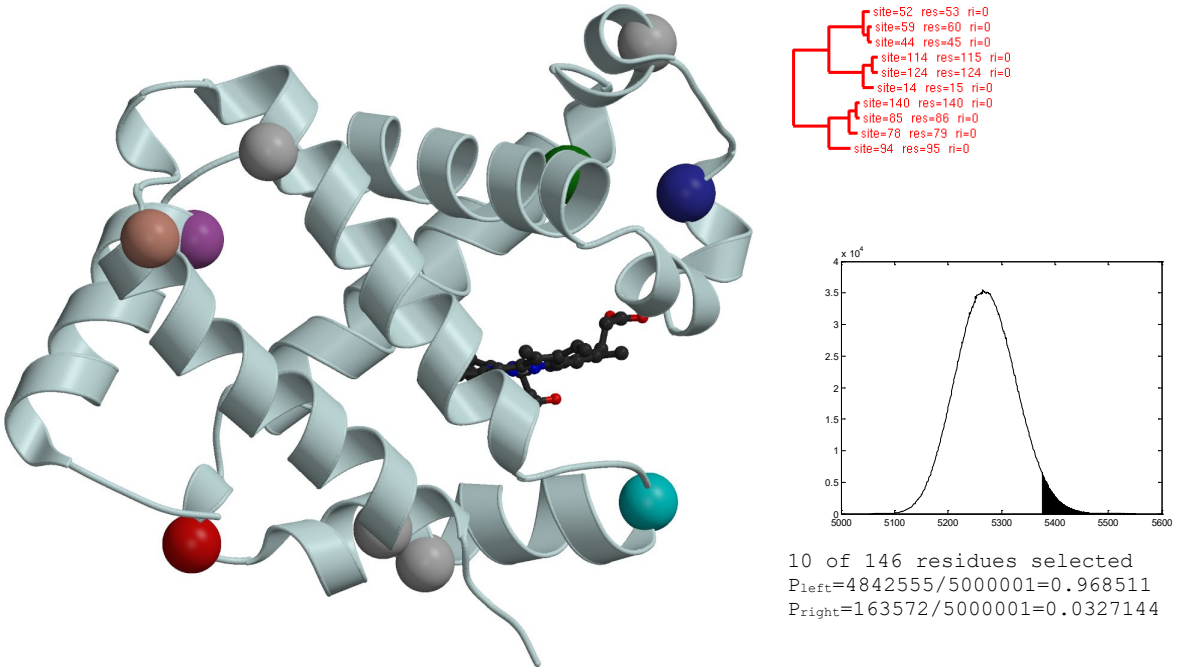


Figure 66. Hemoglobin β – Sites with True Tree RI = 0.0 Highlighted

Residues of the RI_{true}=0.0 category are more dispersed than one would expect by random.

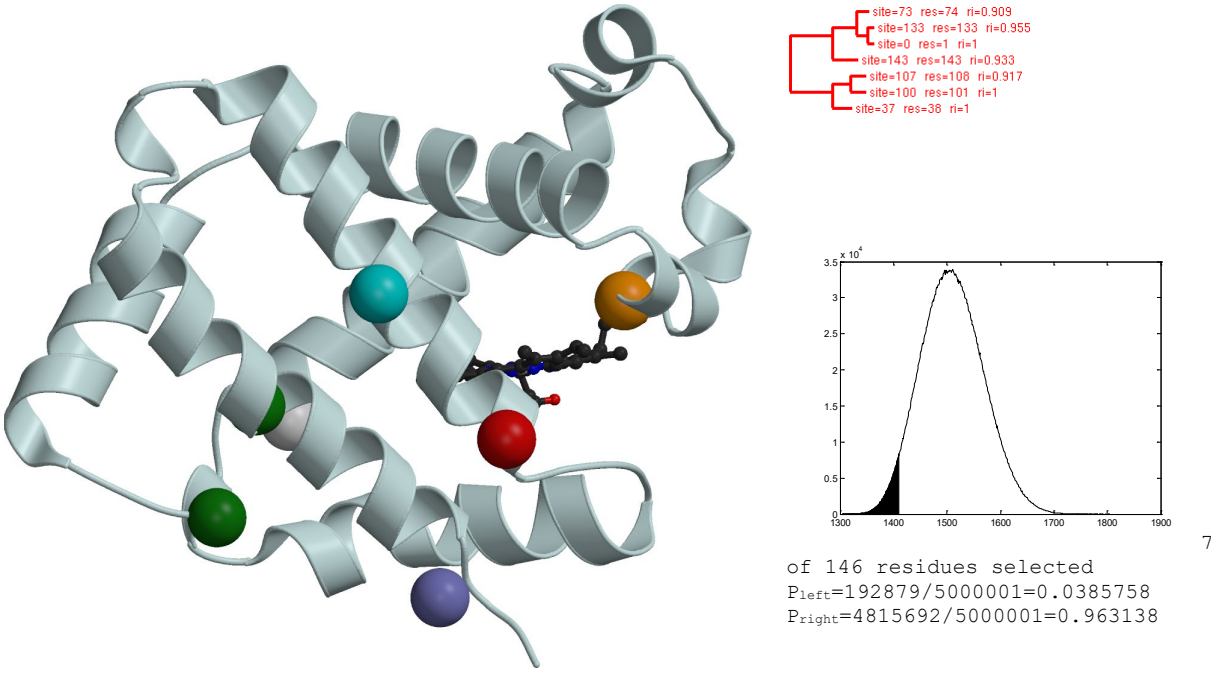


Figure 67. Hemoglobin β – Sites with True Tree RI >= 0.9 Highlighted

Because there were only three residues in the RI_{true}=1.0 category, additional residues were added to make the test more meaningful. All residues with a RI_{true} value of 0.9 or larger were considered, and in this case there appears to be clustering present around the heme group.

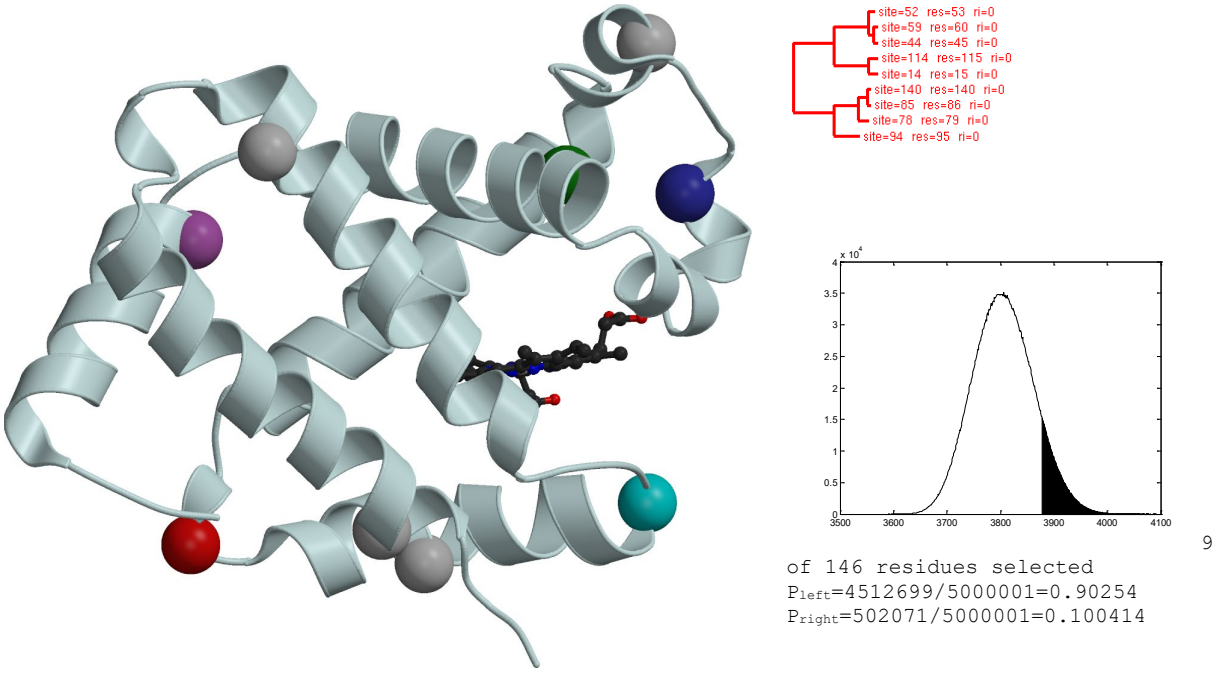


Figure 68. Hemoglobin β – Sites with MPC Tree RI = 0.0 Highlighted

No apparent clustering or patterns are observed. The highlighted residues are lightly dispersed, but not to a significant level according to the p-values.

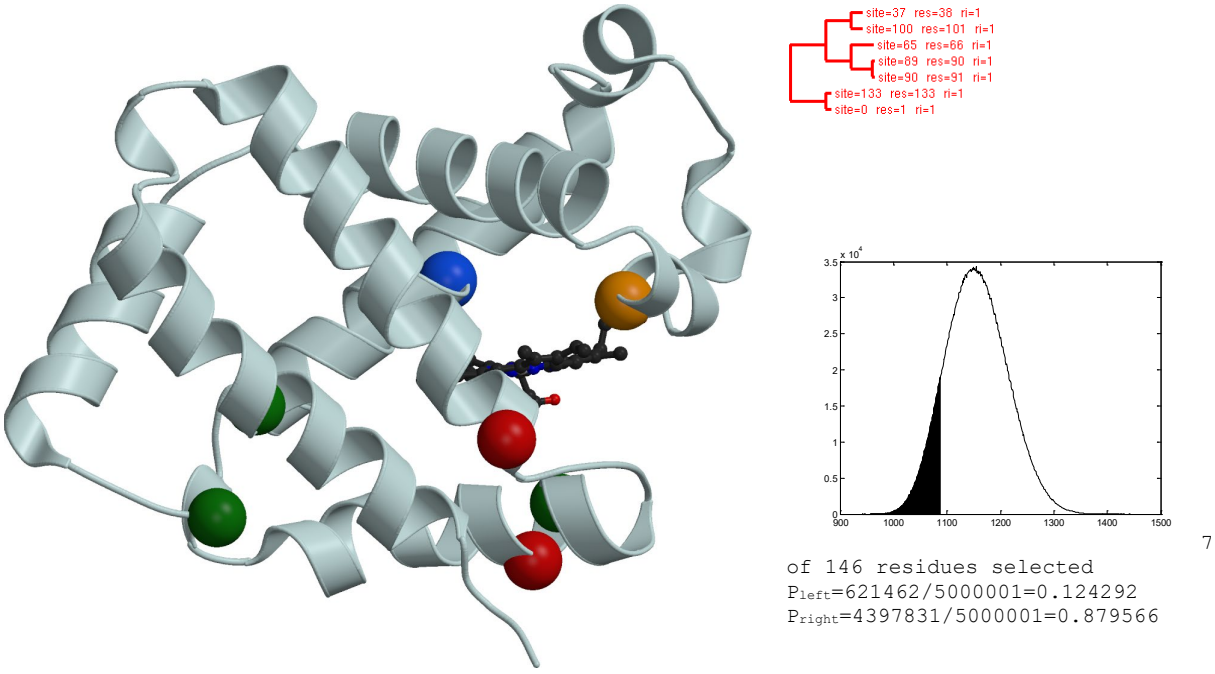


Figure 69. Hemoglobin β – Sites with MPC Tree RI = 1.0 Highlighted

No apparent clustering or patterns are observed.

Preliminary Joint α- and β-chains Results

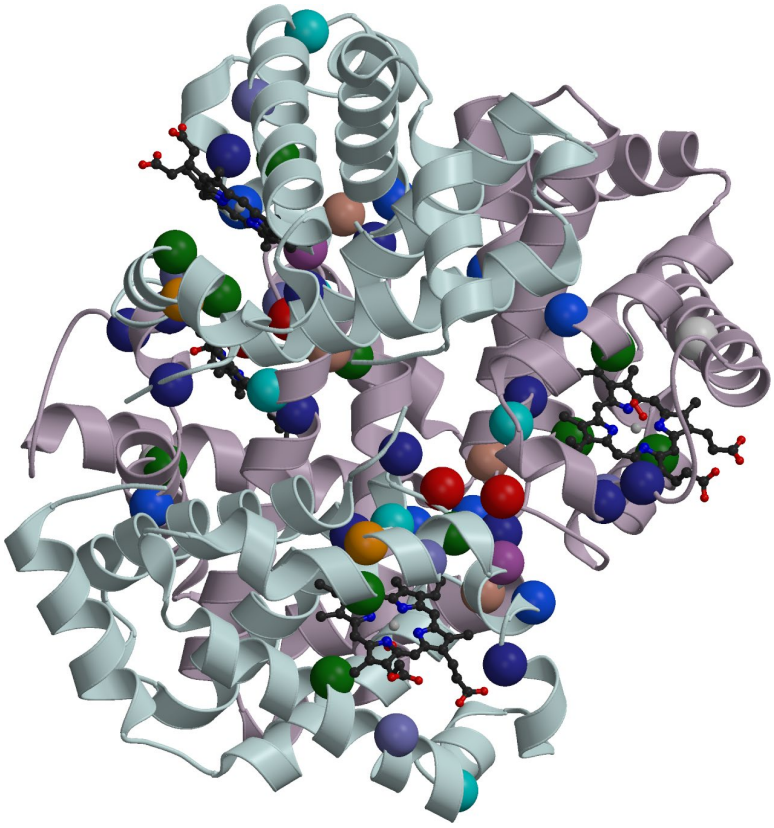


Figure 70. Hemoglobin α- and β-chains in Context with Invariant Residues Highlighted

We will consider the joint hemoglobin α- and β-chains data only briefly, mainly because a thorough analysis could not be performed or presented graphically in print form in a very clear fashion. Hemoglobin, as described in the functional and structural description section, is a complex of four chains – two α chains and two β-chains. Because these chains have been analyzed independently, considerable information may have been lost that would have been present if the native context of hemoglobin was maintained.

In the above figure we see hemoglobin as it is natively with all four chains present and the hemes displayed. The hemoglobin α chains are displayed in a thistle color while the hemoglobin β-chains are displayed in a light cyan. The residues that are highlighted are invariant residues which are the ones that seemed to exhibit the most obvious clustering when the chains were considered independently.

While clustering patterns were sometimes hard enough to detect when considering the chains independently, things become even more complicated when all the chains are shown. We know from previous examinations of the invariant residues that they clustered around the hemes. This is the same clustering that is present here. However, there does appear to be some favoring of the residues to be positioned near the borders of the α- and β-chain interactions. However, the hemes are in the same location so it is difficult to say if the clustering is a result of some constraint placed on both chains because of their proximity or because of constraints placed on the chain to hold the heme in place. One could argue that this clustering is indeed at the borders and the result of a constraint needed to hold the chains together. This could be supported by the lack of apparent clustering or the relationship of the invariant residues in myoglobin to the heme. Since myoglobin and the chains of hemoglobin are so similar, they may share similar constraints.

While only a single measure was considered here, namely if the residues were invariant, the various RI measures used in the previous sections could also be applied. This section was only provided to give the reader a glimpse of a possible future direction. Further research needs to be done to consider multimeric proteins such as hemoglobin. Statistical methods need to be developed to place a quantitative measure on the existence and significance of potential clusters or patterns resulting from interactions between chains.

Chapter 4. Discussion

Review of Goals

This project was undertaken to examine residues with misleading phylogenetic signals in the context of their protein structures and to develop a new method for predicting residues that may be of functional importance. These two goals become one as a result of a unique integration of sequence alignment, evolutionary history, and protein structure. The initial motivation of this project grew out of an interest in the spatial relationships of phylogenetically misleading residues and an interest in improving phylogenetic estimations through the incorporation of protein structural information. In fact, the use of this method for identification of functionally important residues only became apparent after careful examination of what might be causing a misleading phylogenetic signal.

As discussed in the introduction, phylogenetic inference procedures make certain assumptions about the data they are applied to. A summary of these assumptions is that sequence data must behave as a string of characters changing randomly at a stochastically constant pace without reversion. These assumptions are a consequence of our poor understanding of the process by which genes evolve at the level of individual residues. However, it is clear that there are restrictions on how a gene can change. There is variation in the rate of change at different sites in the gene, and reversions to preexisting states certainly must occur due to the limited alphabet. To complicate matters, deviations from the assumptions are present to varied degrees in different organisms.

The specific sites that cause failures of phylogenetic inference procedures can be identified if a true topology is known. The true topology can be compared to the generated topology and differences identified. In this project these differences were measured using the retention index (RI), which is a measure of how well a particular site exhibits hierarchical fit to a given topology under parsimony. If a site is under a constraint that prevents it from behaving according to the assumption, the site is unlikely to be very supportive of the tree and hence have a low RI value. If however, the site is without constraints and is free to evolve randomly then it is more likely to represent the evolution suggested by the topology. In this case, the RI for the site is a higher value. Examination of the sites that have the lowest RI values in context of the correct topology give some indication of residues that may be under constraints. However, the more extreme sites

may be identified by contrasting these values with corresponding values from the topology suggested by the entire alignment. This can be performed by simply subtracting corresponding RI values defining the RI_{diff} measurement.

Having identified residues that are under an evolutionary constraint and thus of possible functional significance is a worthy first step. However, it is possible to improve one's confidence that these particular residues are functionally significant by incorporating information about protein structures and the spatial relationships among misleading residues. One possible scenario causing residues to be misleading is if they work cooperatively. A suspicion of cooperation would be more credible knowing that the residues were physical neighbors. If however, they were distant, while still possibly cooperating, they must do so through a more complicated mechanism.

This project started as an attempt to improve phylogenetic inference procedures by incorporation of protein structure. Initially, the bioinformatics tool RI Compare was developed to explore possible relationships of residue fit to a given topology in context of protein structure. The tool is primarily a "hypothesis generator" since it helps a researcher develop ideas to test by providing a different means for exploration. Using RI Compare, it was noticed early in the project that in cytochrome b the most misleading sites seemed to form a tight cluster in one region of the protein. This led us to wonder if the clustering of misleading residues is a general property or particular to only this single example. If such clustering exists in all or most proteins, then perhaps it would be possible to extract this core of misleading sites before applying an inference procedure and generate improved trees.

While the clustering property of misleading residues that was initially observed in cytochrome b was later found to not be generalizable to all proteins, the identified residues by the RI_{diff} measure and the RI Compare tool may be able to be used to identify residues of functional importance and provide candidates for mutagenesis studies. While the results of this project do not provide universal results that can be applied in all contexts, interesting properties were observed. What follows is a review and interpretation of the results and suggestions for future work.

Review of Results

Having reviewed the goals of this project along with their implications and utility we now review the results of this project. While a complete presentation of the results is given in the Results section, some, of the more interesting observations are emphasized here.

The initial dataset that was subjected to the RI_{diff} measure and the RI Compare tool was the cytochrome b dataset. This dataset exhibits strong spatial clustering of the residues with RI_{diff} values < 0.0 in the transmembrane region. While there are a few additional residues in this category that prevent the p-value from being significant when subjected to a random sampling test, the clustering can not be ignored. The clustering of residues with RI_{diff} values < 0.0 is also a property shared by the rhodopsin dataset. While few residues again prevent the p-value from appearing significant, the clustering is fairly distinct. Both cytochrome b and rhodopsin are seven helix g-coupled transmembrane proteins, and perhaps this is significant. A difference between the two is that the cluster is positioned on the helices nearest the retinal ligand in rhodopsin while in cytochrome b (at least that with residues RI_{diff} values < 0.083) the cluster is on helices other than those near the heme ligand. While a cluster for RI_{diff} values < 0.0 does appear in both transmembrane proteins, the clusters do not always occur around what would initially be considered the active area of the protein. Contrasting these results with the members of the other main group of proteins examined, we see no significant clustering either by p-value or by visual inspection in the globin datasets.

The similarities unique to cytochrome b and rhodopsin also include the observation that residues with undefined RI values cluster on either side of the membrane in both proteins. Also, while the transmembrane region is fairly void of residues with undefined RI values, those that do exist are mainly along the helices that surround the ligand of the protein. Again, the clustering of residues with undefined RI values is a property that is shared only by the transmembrane proteins. The globular proteins were not observed to have this characteristic.

A strong banding of the residues of the RI_{diff}>0 category was present in the rhodopsin dataset. The bands crossed each of the three areas parallel to the membrane walls. The residues of this category in hemoglobin α also exhibited a favoring toward one side of the protein and around the heme. Clustering of the RI_{diff}>0 residues in myoglobin and hemoglobin β may exist, but it was weak. No apparent clustering of residues in the RI_{diff}>0 category was obvious in the cytochrome b dataset.

Interestingly, a property that was shared by all the protein datasets was that invariant residues (corresponding residues that do not change regardless of species) clustered. Generally, these

clusterings were very strong when considering their p-values. Interesting patterns were also noticed when considering each dataset individually.

Clustering of the invariant residues in cytochrome b, while visible in all three regions (both sides of the membrane as well as the transmembrane region), greatly favored those on one side of the membrane. The other side of the membrane did not have as many residues in this class by number, but perhaps by percentage. One side is composed mainly of helices packing a greater number of residues into a given area compared with the strands that are the main constituent of the other side. The transmembrane region only had a few invariant residues and those were mainly around one of the two heme ligands.

The rhodopsin dataset has a similar clustering pattern of invariant sites to cytochrome b. All three regions contain invariant residues with an apparent favoring of residues on the outside of the membrane. Again, as was in the cytochrome b dataset, the invariant residues that occur in the membrane are generally around the area that holds the ligand.

All of the globular proteins, myoglobin, hemoglobin α and hemoglobin β, also had significant clustering of invariant residues with respect to the p-values. While nothing appeared to be special about the particular clustering that existed in myoglobin, both hemoglobin α and hemoglobin β had tight clusters on the side of the protein around the heme. While this may seem obvious given the functional importance of the heme group, the fact that myoglobin is so similar but does not exhibit the same pattern is surprising. This supports the belief that the invariant residues in hemoglobin are as important in holding the chains together as they are in holding the heme in position.

Another property that was shared between all the datasets was the fact that the residues where $RI_{true}=0$ were very similar to those in the $RI_{MPC}=0$ category. The same was true when comparing those in the $RI_{true}=1$ and $RI_{MPC}=1$ categories. At most, only a handful of residues were different between these extreme RI categories in the true tree and the MPC tree. Even more interesting is how the residues that differ between the two categories are related spatially with several examples of pairing, suggesting cooperation between the residues in either a supportive or misguiding way.

When examining the $RI_{true}=1$ and $RI_{MPC}=1$ categories in the hemoglobin β, myoglobin, and rhodopsin datasets, each had a neighboring pair, and rhodopsin had two alternating pairs, of residues making up the differences. Also, myoglobin and rhodopsin each had a co-occurring

alternating pair of residues differing between these categories. (For a graphical explanation of these terms see the Alternations and Co-occurrence Types table below.)

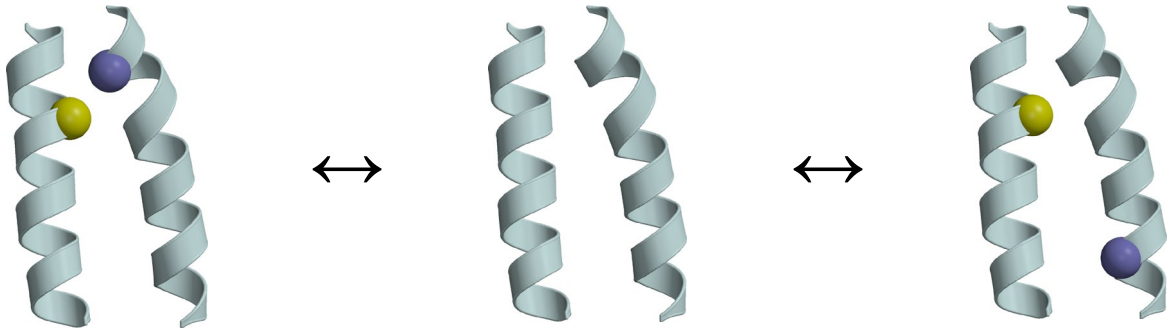
Comparison of the $RI_{true}=0$ and $RI_{MPC}=0$ categories in hemoglobin α showed that four residues differed by co-occurring but were not paired spatially and appeared only in the RI_{true} category. Hemoglobin α $RI_{true}=1$ and $RI_{MPC}=1$ comparison was similar with two residues but occurred only in the RI_{MPC} category.

Other differences existed between the $RI_{true}=0$ and $RI_{MPC}=0$ and the $RI_{true}=1$ and $RI_{MPC}=1$ categories. Other than the differences already mentioned, occasionally there would be a difference in one or two residues, but this was not pointed out here because they did not appear to be spatially paired or near a ligand.

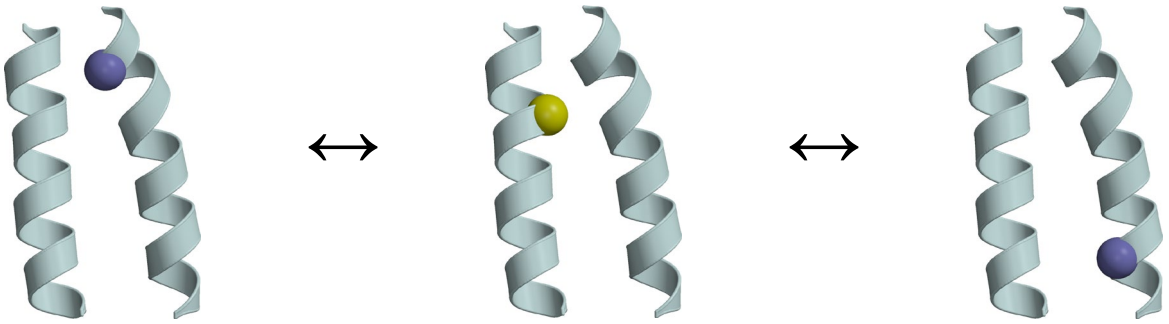
Unique to hemoglobin β was a more dispersed pattern than expected by random of residues with $RI_{true}=0$ and a significant clustering of residues where $RI_{true}>0.9$. Note this deviation in the cutoff value ($RI_{true}=1$ was used in the other datasets) was needed because hemoglobin β had such a small number of residues with $RI_{true}=1$.

	$RI_{diff}<0$	$RI_{diff}=0$	$RI_{diff}>0$	No Change	$RI=\infty$	$RI_{True}=0$	$RI_{True}=1$	$RI_{MPC}=0$	$RI_{MPC}=1$
Cytochrome b	X			X	X				
Rhodopsin	X		X	X	X				
Myoglobin			?	X					
Hemoglobin α			X	X					
Hemoglobin β			X	X		X	X		

Table 1. Summary of Significant Clusters



Co-occurring Pair Co-occurring but not Paired



Alternating Pair Alternating but not Paired

Table 2. Alternations and Co-occurrence Types

Interpretation of Results

The results do not suggest with overwhelming support that the RI_{diff} measure is informative for all datasets. It does seem possible, however, that the RI_{diff} measure may be of use for identifying residues of potential significance in transmembrane proteins. While clustering of residues where $RI_{diff}<0$ as well as residues that had undefined RI values existed in both cytochrome b and rhodopsin, no such clustering was present in the globin proteins.

The cytochrome b and rhodopsin datasets have similar clustering patterns, but this may be a consequence of being able to partition the protein in such a way as to elucidate the pattern. Transmembrane proteins have three obvious regions – the two regions on either side of the membrane and the region that spans the membrane. When considering the globin proteins the partitioning that is logical for the transmembrane proteins is not applicable, and it is not clear if a logical partitioning even exists. Perhaps if such a partitioning did exist and was applied to the results, then what currently appear as randomly dispersed residues would suddenly appear much more clustered.

Comparison of the RI_{true} and RI_{MPC} categories in hemoglobin α showed residues at the extremes of the retention index favoring the phylogenetic inference procedure constructing the correct tree. The categories of residues not supportive of the trees, namely $RI_{true}=0$ and $RI_{MPC}=0$ differed by extra residues in the $RI_{true}=0$ category. The residues supportive of the trees, namely the $RI_{true}=1$ and $RI_{MPC}=1$ categories differed by including extra residues in the RI_{MPC} category. This creates extra support for the failing tree and reduces hemoglobin α's ability as a carrier for a correct phylogenetic signal.

Effects of Individual Species and Entire Clades - Jackknifing

Since these experiments are based on sequence data from various species, it is important to consider the effects choice of species might have on the results. While an attempt was made to minimize this across datasets by selecting sequences from the same species for each dataset when possible, the desire to have a more complete dataset would occasionally force the inclusion

of unique species. Both the hemoglobin α and hemoglobin β datasets have sequences from precisely the same species, but the other datasets vary slightly with respect to each other. The datasets were also built by selecting those species where a fairly uncontroversial view of the evolutionary relationship exists. Ignoring the assumed small differences between selections of species between datasets, there is still another effect that choice of species can adversely affect results if different species differ in the degree to which they deviate from the assumptions of the inference model.

Such deviations might be associated within particular clades or species. This could happen if there has been a shift in how the protein functions requiring several residues to change in sync with each other, or perhaps the characteristics that make a particular species or clade distinct force a particular change in constraints for the entire group. One could imagine, for example, that birds might have genes that are under different constraints than terrestrial organisms. Whatever the reason for such a shift in mechanism or constraint, it could occur at any point in the tree including at an ancestral node separating the entire clade from the rest of the tree. When carried to an extreme, each clade or taxon could have its own peculiarities, rapidly complicating the interpretation of the results.

The degree to which particular taxa or clades are problematic can be tested using a Jackknife procedure in which each taxon or clade is systematically removed from the tree with replacement. In other words, each group depicted in the true topology is removed including those consisting of a single taxon, but the removed group is returned before removing the next node. The pruned dataset is passed to an inference procedure and the inferred tree compared against the true tree. If one finds the inferred topology matching the true topology after having removed a clade then it can be assumed that this clade is a main contributor to the original dataset yielding misleading results.

The Jackknife test was performed on the cytochrome b dataset with an additional measure added. By computing the RI_{diff} values and displaying the residues with $RI_{diff} < 0.0$ on the protein structure, one could quickly see if the clusters seen when considering the entire dataset were stable. As clades were removed and added back, we would see the highlighted residues change – sometimes clustering and other times appearing random. Even if there was no apparent clustering, even with significantly large clades removed, several of the same residues remained highlighted. This suggests that the cause of the misleading signal is something that is common regardless of the species and is likely significant to the function of the protein.

Soundness and Completeness of Results

Sampling of Residue Positions – C_α versus Residue C.O.M., *etc.*

The statistical test that is performed by RI Compare is a simple but powerful one since it does not depend on knowing the distribution *a priori* but instead constructs a null distribution through re-sampling. The estimate is sensitive to the set of data is being sampled. In particular RI Compare uses the positions of C_α atoms of the protein for the random sampling. However, each residue of the protein has a complex shape composed of several atoms only one of which is the C_α. It is possible that the statistical test would yield different results if a different combination of residue atoms were used such as the C_β atoms or the center of mass of the entire residue. Using either of these alternate positions has the advantage that it begins to take into account the orientation of the residue. Pairs of residues that appear to be equidistant from each other when considering the C_α atoms may be found to be of different distances when measurements are between the centers of masses. This could occur because of the size differences in residues or simply because residues are protruding away from the backbone toward or away from each other.

An experiment was performed to test whether the statistical results differed if the center of mass positions were used instead of the C_α positions. Comparison of these two datasets showed only slight differences which were judged insignificant.

While the sampling results do not appear to be affected by the choice of using the center of mass positions versus the C_α positions, there are other reasons to not use the center of mass positions. The source of protein structure used by RI Compare is the standard PDB file. The format of this file allows for detailing the positions of each atom of each residue of the protein and even multiple models and alternate positions for atoms and residues. While this flexibility exists, the available data is limited by the original submission. Often PDB files do not contain all the atoms of a residue or, because of authoring error, additional atoms are present. Absent atoms often include key atoms that define the shape of the residue, and very rarely, if ever, are the hydrogen atoms present. This imperfect data contributes to corruption of the center of mass values. Also there is the question of accuracy of position of the atoms other than the C_α atoms. The atoms are

normally positioned by computer software that solves equations that optimize the placement of residues and their atoms. Also, it should not be forgotten that even though a protein in a PDB file appears to be a static entity, proteins are flexible in biological systems and conformational changes are often required for a protein to perform its function. Such "flexing" could cause dramatic changes to orientation of and proximity of residues.

All the above factors influence the accuracy and reliability of all but the C_α atoms. Alternative ways to measure the distances between the residues could be devised, such as measuring the distances between all pairs of atoms in each pair of residues, but the same issues are raised. The structures contained in the PDB files are of relatively high resolution when viewed with tools such as RasMol, but the apparent clarity is deceptive. The positions are often crude and at best represent only the most likely position in a dynamic system. Developing more sophisticated means for measurement will not help, so we elected to restrict the analysis exclusively to C_α positions.

Chapter 5. Conclusions

Summary

The identification of residues that hold misleading phylogenetic signals and those that are of functional significance are intertwined. Advances in the one area can support the other mainly because misleading phylogenetic signals come from residues that are not evolving as a random process. The lack of a random process implies the existence of a constraint suggesting a possible functional importance. While the lack of clusters in all proteins when considering the RI Difference measure was somewhat discouraging, the presence of clusters in all transmembrane proteins when considering the RI Difference measure is interesting. Perhaps, the RI Difference measure is able to detect certain properties that are only present in transmembrane proteins. If this is true, then the availability of a tool for these proteins is an advancement. Further, determination of the essence that causes the RI Difference measure to find clusters of residues in the transmembrane proteins but not the globular proteins may lead to new advancements and understanding of the functional and evolutionary constraints of these two major classes of proteins.

While the RI Difference measure did not always appear to form distinguishable patterns as expected, if one assumes that a great deal of evolutionary constraint exists in the form of co-variation, there were interesting observations concerning invariant residues. Residues that remain constant across all species seemed to form fairly tight and obvious clusters in all the proteins that were considered. Ironically, this project started as an attempt to improve upon the methods to reconstruct phylogenetic trees which is an area that has very little interest in invariant residues because those residues contain no information about the evolutionary process. Interestingly, the residues that would normally be of little interest were the ones that exhibited such interesting properties in all the datasets.

What makes RI Compare an interesting tool is its unique integration of heterogeneous information. These data include protein structure and sequence along with evolutionary information including phylogenetic trees and a model of evolution. By contrasting different sources of evolutionary information (morphological tree versus inferred tree) we have developed a tool to identify residues responsible for misleading phylogenetic signals. It is this combination of heterogeneous data that has allowed us to gain insights into the evolutionary and functional

constraints of proteins that would not have been apparent if considering only a single source of data. Each data source provides a subtle hint but has limits to its explanatory power. If the sequences or the structure were considered independently by examining the biophysical properties of the residues present, we would have been unable to detect any constraints that are not present at that level. If only the alignment would have been considered then any unique spatial patterns would have been lost in the unnatural linear view of a protein that alignments create. If the evolutionary tree would have been ignored, then so would residues that have evolutionary constraints or alternate rates or perhaps one might have been misled into believing a pattern to be more significant than it actually is. The phylogenetic tree provides an alternative explanation for certain patterns. Only by combining all these sources of information were we able to extract the subtle patterns that have been discussed here.

Evolutionary forces along with functional constraints place complex restrictions on how a protein can change over time. If no such constraints existed the protein would be free to change randomly and uniformly. However, this would do nothing to preserve the essential function of the protein. While only a small portion of the known constraints (which is likely a very small portion of the total actual constraints) have been incorporated into the method reported on here resulting in a fairly crude measure, there appears to be some usefulness to the tool. The method has a strong dependence on the sources of data used especially the phylogenetic trees. These trees, both the morphological and assumed correct tree and the inferred parsimonious tree, are the results of the considerable effort by the researchers, but are still only approximation of the natural tree if the actual representation can actually even be represented as a tree. Ignoring the question of relationships present in the tree there is also the problem of branch lengths, or the time between speciation events, being essentially unknown.

While the method is crude, some interesting patterns have been observed. We believe that this bodes well for the future as further research is done addressing relationships between functional properties and how proteins evolve. It has been shown here how the integration of a few of the available heterogeneous sources of data can be powerful and as further sources become available so will the richness of the questions that can be answered. It is hoped that other researchers will be interested in extending this work either by integrating alternate data sources or methods, addition of pattern recognition methods to aid the user in find potentially interesting patterns, or by providing a more rigorous statistical framework for assessing the results.

Recommendations

Continued exploration of new datasets is needed before the RI Difference measure can be recommended for general use. As seen in the comparison of the transmembrane proteins and globular proteins there are apparently proteins for which the measure is applicable and others for which it is not. This does not appear to be a measure than can be applied in a general way to all proteins and identify residues which are under functional constraints.

The construction of new datasets is made quite easy in most cases by using the SP Parse construction tool developed during this work. While this tool is primary directed at constructing datasets from the SWISS-PROT and TrEMBL databases one could extend its functionality to include alternate data sources such as GenBank or propriety databases.

While the apparent lack of generality of the use of the RI Difference measure in functional constraint detection was apparent, there were interesting observations made concerning invariant residues. While it is not new that researchers pay close attention to invariant residues as possible functional active sites, by placing these residues in context with the protein structure additional information and confidence can be gained. An extension of the available categories may help researchers to explore other properties of proteins. For instance, it may be beneficial to examine clustering of polar residues, or perhaps cyclic residues, or any partitioning of the amino acids. A specific partitioning may be of particular significance for a particular protein or class of protein, but may not be applicable to all proteins.

Using the difference in retention index between two trees certainly identifies residues that are responsible for causing failures in phylogenetic inference procedures such as parsimony. At the moment appropriate trees of high quality as well as high quality sequence data are needed for this procedure to be effective. Little attention has been given to these issues in this work, but their importance should be clear considering the results are based on the supplied trees and sequence data. It may be possible to perform some prediction of the functionally significant residues without trees, some of which has been seen by examining invariant sites, but at least for the moment, trees are needed for the RI Difference measure. There is a chance that the sites that the RI Difference measure finds do have some functional significance since this analysis finds sites that are evolving at a non-random rate or are co-evolving. In the case of the proteins where these residues form clusters when plotted on the structure we have greater confidence, but even in the case of proteins that lack these clusters there may be some mechanism driving the

evolution of these residues. At the moment this tool should be used as an exploratory tool only, and care should be taken not to use it to "strengthen flawed reasoning."

Future Research

One area that could greatly benefit from future research is multimeric proteins. For the majority of this work proteins have been considered as individual chains without much consideration of the native context of the chain. Functional proteins are often composed of several chains or form large complexes through combinations of several individual proteins. For these complexes to be held in place, for signals to cascade across the chains, and for the preservation of function, the evolutionary constraints placed on the protein are likely much more complex than what would be present on a protein composed of a single chain. Methods need to be developed to help researchers assess the presence of patterns across chains in protein complexes. While RI Compare allows the user to view certain results from different chains together, the statistical measure is unaware of residues from neighboring chains.

This project focused mainly on examining the presence of clusters that form in proteins when considering residues which have varying support for two hypothetical views of the evolutionary relationships of the proteins. The two alternate views that were used included a tree which was assumed to be correct built from information gathered from fossils while the second tree was built using sequence data supplied to a phylogenetic inference procedure called parsimony. While parsimony is a common method for phylogenetic inference, it is certainly not the only one. Others include distance and maximum likelihood methods. Since these methods were all created to address the inaccuracies in other datasets, it is possible that one of these methods may be better suited to a particular dataset. Because of this, one could use these methods to generate alternate trees to be contrasted against the true tree, or another tree for that matter, in RI Compare. To limit the number of combinations that needed to be considered, no alternate inference methods were considered in this work, but this is an area that could use additional research.

The retention index is not the only method that is know which can be used to assess the support that a particular site has for a given phylogenetic topology. It was used in this work primarily because of its simplicity to understand and implement, but also its close relationship to parsimony which was the inference procedure used to generate the trees other than the true tree. The retention index is not perfect, having situations where it is undefined, and arguably is a

rather crude measure. A method such as maximum likelihood scoring would be more sensitive and also allow one to experiment with alternate models of evolution. The framework of RI Compare is extensible beyond what its name suggests, so alternate measures could easily be added. Actually, since RI Compare has the ability to interface with Paup, any of the measures present in Paup are also available to RI Compare. One item that prevented the maximum likelihood measure from being incorporated into this work is the fact that the values are not normalized like the RI scores are. This lack of normalization creates difficulties in comparison of the values across datasets and between sites of a single dataset.

Comparing the retention index at corresponding sites between trees certainly identifies residues that are responsible for causing failures in phylogenetic inference procedures such as parsimony. By removing these residues, one can remove the partition of the data that is suggestive of either of the topologies. In the case of comparing the true tree to one generated by parsimony, if the residues that were supportive of the parsimony tree and not supportive of the true tree were removed from the data and a new tree inferred using parsimony, the parsimony algorithm would find the true tree with perfect support. While this is somewhat circular as we are removing misleading data identified using the true tree to generate the true tree, it does suggest the power that the identification of these residues would give the inference procedure. What has yet to be seen is if there is a way to predict the failing sites without knowledge of the correct topology ahead of time. If this is possible and once these residues are identified, it should be possible to improve the inference software in addition to investigating these residues for possible functional significance. This is not an easy problem since the area of phylogenetic inference has existed for some time and no one has managed to identify misleading residues ahead of time despite the rewards. However, if the patterns observed in this paper are present in all transmembrane proteins, namely a tight clustering of misleading residues, perhaps there is some way to extract this misleading core.

Appendix A. Alignments

Cytochrome b

```
Fly                ---MHKPLRNSHPLFKIANNALVDLPAPINISSWWNFGSLLGLCLIIQILTGLFLAMHYT
Mosquito           ---MFKPIRKTHPLISIANNALVDLPAPSNISAWWNFGSLLGLCLMLQILTGLFLAMHYA
Lamprey            -SHQPSIIRKTHPLLSLGNSMLVDLPSPANISAWWNFGSLLSLCLILQIITGLILAMHYT
Blue Whale         ----MTNIRKTHPLMKIINDAFIDLPTPSNISSWWNFGSLLGLCLIVQILTGLFLAMHYT
Fin Whale          ----MTNIRKTHPLMKIVNDAFVDLPTPSNISSWWNFGSLLGLCLIMQILTGLFLAMHYT
Hippo              ----MTNIRKSHPLMKIINDAFVDLPAPSNISSWWNFGSLLGVCLILQILTGLFLAMHYT
Sheep              ----MINIRKTHPLMKIVNNAFIDLPAPSNISSWWNFGSLLGICLILQILTGLFLAMHYT
Cow                ----MTNIRKSHPLMKIVNNAFIDLPAPSNISSWWNFGSLLGICLILQILTGLFLAMHYT
Pig                ----MTNIRKSHPLMKIINNAFIDLPAPSNISSWWNFGSLLGICLILQILTGLFLAMHYT
White Rhino        ----MTNIRKSHPLIKIINHSFIDLPTPSNISAWWNFGSLLGICLILQILTGLFLAMHYT
Black Rhino        ----MTNIRKSHPLVKIINHSFIDLPTPSNISSWWNFGSLLGICLILQILTGLFLAMHYT
Donkey             ----MTNIRKSHPLIKIINHSFIDLPTPSNISSWWNFGSLLGICLILQILTGLFLAMHYT
Horse              ----MTNIRKSHPLIKIINHSFIDLPAPSNISSWWNFGSLLGICLILQILTGLFLAMHYT
Halicho            ----MTNIRKTHPLMKIINNSFIDLPTPSNISAWWNFGSLLGICLILQILTGLFLAMHYT
Seal Vitulina      ----MTNIRKTHPLMKIINNSFIDLPTPSNISAWWNFGSLLGICLILQILTGLFLAMHYT
Cat                ----MTNIRKSHPLIKIINHSFIDLPAPSNISAWWNFGSLLGVCLTLQILTGLFLAMHYT
Dog                ----MTNIRKTHPLAKIVNNSFIDLPAPSNISAWWNFGSLLGVCLILQILTGLFLAMHYT
Rat                ----MTNIRKSHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQILTGLFLAMHYT
Mouse              ----MTNMRKTHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYT
Myoxus             ----MTIIRKSHPLIKIINHSFIDLPTPSNISAWWIFGSLLGACLGIQILTGLFLAMHYT
Gibbon             ----MTPLRKTNPLMKLINHSLIDLPAPSNISMWWNFGSLLGACLILQIITGLFLAMHYT
Man                ----MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYS
Baboon             ----MTPMRKSNPIMKMINHSFIDLPTPSNISIWWNFGSLLATCLILQIITGLFLAMHYS
Platypus           ----MNNLRKTHPLIKIVNHSFIDLPTPSNISSWWNFGSLLGLCLIIQILTGLFLAMHYT
Possum             ----MTNIRKTHPLMKIINDSFIDLPTPSNISAWWNFGSLLGVCLIIQILTGLFLAMHYT
Kangaroo           ----MTNLRKSHPLIKIVNHSFIDLPAPSNISAWWNFGSLLGACLIIQILTGLFLAMHYT
Chicken            ---MAPNIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYT
Ostrich            ---MAPNIRKSHPLLKIINNSLIDLPSPSNISAWWNFGSLLGICLITQILTGLLLAMHYT
Crow               ---MGLNLRKNHPLLKIINNSLIDLPTPSNISAWWNFGSLLGLCLIMQIITGLLLAMHYT
Alligator          ---MTHQLRKSHPIIKLINRSLIDLPTPSNISAWWNFGSLLGLTLLIQILTGFFLMMHFS
Chrysem            ---MTMNHRKTHPLTKIINNSFIDLPSPSNISAWWNFGSLLGTCLILQTITGIFLAMHYS
Pelomed            ---MGTLHLKQNPLLKITNKSLINLPSPSNISAWWNFGSLLGMCLILQITTGIFLAMHYT
Carassi            ----MASLRKTHPLIKIANDALVDLPTPSNISAWWNFGSLLGLCLITQILTGLFLAMHYT
Carp               ----MASLRKTHPLIKIANDALVDLPTPSNISAWWNFGSLLGLCLITQILTGLFLAMHYT
Trout              ----MANLRKTHPLLKIANDALVDLPAPSNISVWWNFGSLLGLCLATQILTGLFLAMHYT
Salmon             ----MANLRKTHPLLKIANDALVDLPAPSNISVWWNFGSLLGLCLATQILTGLFLAMHYT
Smooth Dog Fish    ---MATNIRKTHPLLKIMNHALVDLPAPSNISLWWNFGSLMGLCLLIQILTGLFLAMHYT
Scyliorhinus       ---MATNIRKTHPLLKIVNHALIDLPAPSNISVWWNFGSLLGLCLIMQIITGLFLAMHYT
Spiny Dog Fish     ---MTTNIRKTHPLIKIVNHALVDLPSPSNISIWWNFGSLLGLCLIIQILTGLFLAMHYT
Skate              ---MTTNIRKTHPLFKIINSSLIDLPTPVNISIWWNYGSLLGLCLIIQILTGLFLAMHYT
Polypterus         ----MAIIRKTHPLAKIINSAFIDLPAPSNISSWWNMGSLLGLCLIAQIITGLFLAMHYV
Frog               --LMAPNIRKSHPLIKIINNSFIDLPTPSNISSLWNFGSLLGVCLIAQIITGLFLAMHYT
Lung Fish          ---MATNIRKTHPLLKIVNNSLIDLPTPSNISAWWNFGSLLGFCLITQILTGLFLAMHYT
Amphioxus          ---MSGPLRKHHPLLKVVNHSVIDLPVPSNISVMWNFGSLLGLCLVSQILTGLFLAMHYT
P Urchin           ---MLGPLRKEHPIFRILKSTFVDLPLPSKLSIWWKFGSLLGLCLMTQILTGLFLAMHYT
S Urchin           IKIMAAPLRKEHPIFRILKSTFVDLPLPSNLSIWWNSGSLLGLCLVVQMLTGMFLAMHYT
                           : :*:  : :  .::** * ::*  *  ***:.  *  *  **::* **:
```

```
Fly                ADVNLAFYSVNHICRDVNYGWLLRTLHANGASFFFICIYLHIGRGIYYGSYLFTPTWLVG
Mosquito           ADIETAFNSVNHICRDVNNGWFLRICHANGASFFFACLFIHVGRGVYYESYLYHMTWNTG
Lamprey            ANTELAFSSVMHICRDVNNGWLMRNLHANGASMFFICIYAHIGRGIYYGSYLYKETWNVG
Blue Whale         PDTMTAFSSVTHICRDVNYGWVIRYLHANGASMFFICLYAHMGRGLYYGSHAFRETWNIG
Fin Whale          PDTTTAFSSVTHICRDVNYGWIIRYLHANGASMFFICLYAHMGRGLYYGSYAFRETWNIG
Hippo              PDTLTAFSSVTHICRDVNYGWVIRYMHANGASIFFICLFTHVGRGLYYGSHTFLETWNIG
Sheep              PDTTTAFSSVTHICRDVNYGWIIRYMHANGASMFFICLFMHVGRGLYYGSYTFLETWNIG
Cow                SDTTTAFSSVTHICRDVNYGWIIRYMHANGASMFFICLYMHVGRGLYYGSYTFLETWNIG
Pig                SDTTTAFSSVTHICRDVNYGWVIRYLHANGASMFFICLFIHVGRGLYYGSYMFLETWNIG
White Rhino        PDTMTAFSSVAHICRDVNYGWIIRYLHANGASMFFICLFIHVGRGIYYGSYTFLETWNIG
Black Rhino        PDTTTAFSSVTHICRDVNYGWMIRYLHANGASMFFICLFIHVGRGLYYGSYTFLETWNIG
Donkey             SDTTTAFSSVTHICRDVNYGWIIRYLHANGASMFFICLFIHVGRGLYYGSYTFLETWNIG
Horse              SDTTTAFSSVTHICRDVNYGWIIRYLHANGASMFFICLFIHVGRGLYYGSYTFLETWNIG
Halicho            SDTTTAFSSVTHICRDVNYGWIIRYLHANGASMFFICLYMHVGRGLYYGSYTFTETWNIG
Seal Vitulina      SDTTTAFSSVTHICRDVNYGWIIRYLHANGASMFFICLYMHVGRGLYYGSYTFTETWNIG
Cat                SDTMTAFSSVTHICRDVNYGWIIRYLHANGASMFFICLYMHVGRGMYYGSYTFSETWNIG
Dog                SDTATAFSSVTHICRDVNYGWIIRYMHANGASMFFICLFLHVGRGLYYGSYVFMETWNIG
Rat                SDTMTAFSSVTHICRDVNYGWLIRYLQANGASMFFICLFLHVGRGLYYGSYTFLETWNIG
Mouse              SDTMTAFSSVTHICRDVNYGWLIRYMHANGASMFFICLFLHVGRGLYYGSYTFMETWNIG
Myoxus             SDTMTAFSSVTHICRDVNYGWLIRYMHANGASMFFICLFLHVGRGMYYGSYMFIETWNIG
Gibbon             PDASTAFSSVAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGLYYGSFLYLETWNIG
Man                PDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGLYYGSFLYSETWNIG
Baboon             PDTSSAFSSIAHITRDVNYGWTIRYLHANGASMLFICLFLHVGRGLYYGSYLLLKTWNIG
Platypus           SDTSTAFSSVAHICRDVNYGWLIRYMHANGASLFFMCIFLHIGRGLYYGSYTQTETWNIG
Possum             SDTLTAFSSVAHICRDVNYGWLIRNIHANGASMFFMCLFLHVGRGIYYGSYLYKETWNIG
Kangaroo           SDTLTAFSSVAHICRDVNYGWLIRNLHANGASMFFMCLFLHVGRGIYYGSYLYKETWNIG
Chicken            ADTSLAFSSVAHTCRNVQYGWLIRNLHANGASFFFICIFLHIGRGLYYGSYLYKETWNTG
Ostrich            ADTTLAFSSVAHTCRNVQYGWFIRNLHANGASFFFICIYLHIGRGLYYGSYLYKETWNTG
Crow               ADTSLAFASVAHMCRDVQFGWLIRNLHANGASFFFICIYLHIGRGFYYGSYLNKETWNIG
Alligator          SSDTLAFSSVSYTSREVWFGWLIRNLHTNGASLFFMFIFLHIGRGLYYTSYLHESTWNIG
Chrysem            PDISLAFSSVAHITRDVQYGWLIRNMHANGASLFFMCIYLHIGRGLYYGSYLYKETWNTG
Pelomed            PNITTAFSSVAHITRDVQYGWLLRGLHANGASIFFICLYFHIGRGIYYGSFLNKKTWYTG
Carassi            SDISTAFSSVTHICRDVNYGWLIRNIHANGASFFFICIYMHIARGLYYGSYLYKETWNIG
Carp               SDISTAFSSVTHICRDVNYGWLIRNVHANGASFFFICIYMHIARGLYYGSYLYKETWNIG
Trout              SDISTAFSSVCHICRDVSYGWLIRNIHANGASFFFICIYMHIARGLYYGSYLYKETWNIG
Salmon             SDISTAFSSVCHICRDVSYGWLIRNIHANGASFFFICIYMHIARGLYYGSYLYKETWNIG
Smooth Dog Fish    ADISMAFSSVVHICRDVNYGWLIRNIHANGASLFFICIYLHIARGLYYGSYLNKETWDIG
Scyliorhinus       ADISMAFSSVIHISRDVNYGWLMRNIHAYGASFFFICIYLHIARGLYYGSYLNKEAWNIG
Spiny Dog Fish     ADISTAFSSVVHICRDVNYGWLIRNIHANGASLFFICVYLHIARGLYYGSYLFKEAWNIG
Skate              PDIASAFSSVVHICRDVNYGWLIRNIHANGASLFFICIYIHMARGFYYGSYLNKETWNIG
Polypterus         SDINSAFSSVAHICRDVNYGWLIRNFHANGASLFFICIYLHIARGLYYGSYLYTETWNMG
Frog               ADTSMAFSSVAHICFDVNYGLLIRNLHANGLSFFFICIYLHIGRGLYYGSFLYKETWNIG
Lung Fish          ADTSTAFSSIAHIARDVNYGWLLRNIHANGASMFFICIYIHIGRGIYYGSFLYTETWNIG
Amphioxus          ADVNLAFSSVAHICRDVNYGWLLRNLHANGASFMFICLYMHIGRGLYYGSYFYRETWNIG
P Urchin           ADISLAFSSASHMCRDVNYGWLLRKVHAKGASLFFICMYCHMGRGLYYGGSNKMETWKVG
S Urchin           ADITLAFSSVMHILRDVNYGWFLRYVHAKGVSLFFICMYCHMGRGLYYGSYKKIETWKVG
                   ..   ** *  :   :*  *  :*  :: * *::*  :: *:.**.** .     :*  *
```

```
Fly                VIILFLVMGTAFMGYVLPWGQMSFWGATVITNLLSAIPYLGMDLVQWLWGGFAVDNATLT
Mosquito           VIILFLTMATGFLGYVLPWGQMSFWGATVITNLLSAVPYLGMDLVQWIWGGFAVDNATLT
Lamprey            VILFALTAATAFVGYVLPWGQMSFWGATVITNLISAMPYVGNDIVVWLWGGFSVSNATLT
Blue Whale         VILLFTVMATAFVGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Fin Whale          VILLFTVMATAFVGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Hippo              VILLLTTMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVEWIWGGFSVDKATLT
Sheep              VILLFATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATLT
Cow                VILLLTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATLT
Pig                VVLLFTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVEWIWGGFSVDKATLT
White Rhino        VILLFTLMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATLT
Black Rhino        IILLFTLMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTNLVEWIWGGFSVDKATLT
Donkey             IILLFTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Horse              IILLFTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Halicho            IILLFTIMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVQWIWGGFSVDKATLT
Seal Vitulina      IILLFTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYVGTDLVQWIWGGFSVDKATLT
Cat                IMLLFTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTELVEWIWGGFSVDKATLT
Dog                IVLLFATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVEWIWGGFSVDKATLT
Rat                IILLFAVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Mouse              VLLLFAVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Myoxus             IILLFTVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Gibbon             IILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAVPYIGTDLVQWVWGGYSVDNATLT
Man                IILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVQWIWGGYSVDSPTLT
Baboon             IMLLLMTMTTAFMGYVLPWGQMSFWGATVITNLLSAVPYIGTNLVQWVWGGPAIDNPTLM
Platypus           VVLLFTVMATAFVGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLT
Possum             VILLLTVMATAFVGYVLPWGQMSFWGATVITNLLSAIPYIGSTLVEWIWGGFSVDKATLT
Kangaroo           VILLLTVMATAFVGYVLPWGQMSFWGATVITNLLSAIPYVGTTLVEWIWGGFSVDKATLT
Chicken            VILLLTLMATAFVGYVLPWGQMSFWGATVITNLFSAIPYIGHTLVEWAWGGFSVDNPTLT
Ostrich            VILLLTLMATAFVGYVLPWGQMSFWGATVITNLFSAIPYIGQTLVEWAWGGFSVDNPTLT
Crow               VILLLTLMATAFVGYVLPWGQMSFWGATVITNLFSAIPYIGQTLVEWLWGGFSVDNPTLT
Alligator          VIMLLLLMATAFMGYVLPWGQMSFWGATVITNLLSATPYVGSTVVPWIWGGPSVDNATLT
Chrysem            IILLLLTMATAFMGYVLPWGQMSFWGATVITNLLSAIPFIGNTLVQWIWGGFSVDNATLT
Pelomed            IMLLFLTMATAFMGYILPWGQMSFWGATVITNLLSAIPYMGTNLVQWIWGGFSVDNATLT
Carassi            VVLLLLVMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYMGDMLVQWIWGGFSVDNATLT
Carp               VVLLLLVMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYMGDMLVQWIWGGFSVDNATLT
Trout              VVLLLLTMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYVGGALVQWIWGGFSVDNATLT
Salmon             VVLLLLTMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYVGGALVQWIWGGFSVDNATLT
Smooth Dog Fish    VILLFLLMATAFVGYVLPWGQMSFWGATVITNLLSAFPYIGNMLVQWIWGGFSVDNATLT
Scyliorhinus       VVLLFLLMATAFVGYVLPWGQMSFWGATVITNLLSAFPYIGNLLVQWIWGGFSVDNATLT
Spiny Dog Fish     VILLFLLMATAFVGYVLPWGQMSFWGATVITNLLSAFPYIGDMLVQWIWGGFSIDNATLT
Skate              VILLFLLMATAFVGYVLPWGQMSFWGATVITNLLSAFPYIGNILVEWIWGGFSVDNATLT
Polypterus         VILLLLTMMTAFVGYVLPWGQMSFWGATVITNLLSAIPYIGDTLVQWIWGGFSVDKPTLT
Frog               VILLFLVMATAFVGYVLPWGQMSFWGATVITNLLSAKPYIGNVLVQWSLGGFSVDNATLT
Lung Fish          VVLFLLTMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYLGDTLVQWIWGGFSVDNATLT
Amphioxus          VMLLVLTMATAFLGYVLPWGQMSFWGATVITNLFSAIPYLGPDLVQWLWGGFSVDNATLT
P Urchin           VILFLVTVLTAFVGYVLVWGRMSFWAATVIANLVTAVPCVGTTIVQWLWGGFSVDNATLT
S Urchin           VILFLVTILTAFMGYVLVWGQMSFWAATVITNLVSAIPYMGTIMVQWLWGGFSVDKATLT
                   ::::     *.*:**:* **:****.****:**.:* * :*  :* *  ** ::...**
```

```
Fly                RFFTFHFILPFIVLAMTMIHLLFLHQTGSNNPIGLNSNIDKIPFHPYFTFKDIVGFIVMI
Mosquito           RFFTFHFIFPFIILALMMIHLLFLHQTGSNNPLGLNSNVDKIPFHPYFIYKDIFGFIVFL
Lamprey            RFFTFHFILPFILAAMTMIHIMFLHQTGSSNPMGINSNLDKIQFHPYFSFKDILGFVILL
Blue Whale         RFFAFHFILPFIIMALAIVHLIFLHETGSNNPTGIPSDMDKIPFHPYYTIKDILGALLLI
Fin Whale          RFFAFHFILPFIILALAIVHLIFLHETGSNNPTGIPSDMDKIPFHPYHTIKDILGALLLI
Hippo              RFFAFHFILPFVITALAIVHLLFLHETGSNNPTGIPSNADKIPFHPYYTIKDILGILLLM
Sheep              RFFAFHFIFPFIIAALAMVHLLFLHETGSNNPTGIPSDTDKIPFHPYYTIKDILGAILLI
Cow                RFFAFHFILPFIIMAIAMVHLLFLHETGSNNPTGISSDVDKIPFHPYYTIKDILGALLLI
Pig                RFFAFHFILPFIITALAAVHLLFLHETGSNNPTGISSDMDKIPFHPYYTIKDILGALFMM
White Rhino        RFFAFHFILPFIIMALAITHLLFLHETGSNNPSGIPSNMDKIPFHPYYTIKDILGILLLI
Black Rhino        RFFAFHFILPFIILALAITHLLFLHETGSNNPSGIPSNMDKIPFHPYYTIKDILGALLLI
Donkey             RFFAFHFILPFIITALVIVHLLFLHETGSNNPSGIPSDMDKIPFHPYYTIKDILGLLLLV
Horse              RFFAFHFILPFIITALVVVHLLFLHETGSNNPSGIPSDMDKIPFHPYYTIKDILGLLLLI
Halicho            GFFAFHFILPFVVLALAAVHLLFLHETGSNNPSGIMPDSDKIPFHPYYTIKDILGALLLI
Seal Vitulina      RFFAFHFILPFVVLALDAVHLLFLHETGSNNPSGIMSDSDKIPFHPYYTIKDILGALLLI
Cat                RFFGFHFILPFIISALAGVHLLFLHETGSNNPSGITSDSDKIPFHPYYTIKDILGLLVLV
Dog                RFFAFHFILPFIIAALAMVHLLFLHETGSNNPSGITSDSDKIPFHPYYTIKDILGALLLL
Rat                RFFAFHFILPFIIAALAIVHLLFLHETGSNNPTGLNSDADKIPFHPYYTIKDLLGVFMLL
Mouse              RFFAFHFILPFIIAALAIVHLLFLHETGSNNPTGLNSDADKIPFHPYYTIKDILGILIMF
Myoxus             RFFAFHFILPFIIAALVMVHLLFLHETGSNNPSGLNSDTDKIPFHPYYTIKDILGLLLLI
Gibbon             RFFTFHFILPFIITALAALHLLFLHETGSNNPLGISSQPDKIAFHPYYTIKDILGLFLLL
Man                RFFTFHFILPFIIATLAALHLLFLHETGSNNPLGITSHSDKITFHPYYTIKDTLGLLLFL
Baboon             RFFTLHFILPFGIVALTIVHLLFLHETGSNNPCGISSDPDKITFHPYYTTKDILGVAPLL
Platypus           RFFAFHFILPFVIAALAVIHLLFLHETGSNNPSGLNSDPDKIPFHPYYSVKDLVGFFMTI
Possum             RFFAFHFILPFIILAMVVVHLLFLHETGSSNPTGLDPNSDKIPFHPYYTMKDILGLFLMI
Kangaroo           RFFAFHFILPFIITALVLVHLLFLHETGSNNPSGINPDSDKIPFHPYYTIKDALGLMLML
Chicken            RFFALHFLLPFAIAGITIIHLTFLHESGSNNPLGISSDSDKIPFHPYYSFKDILGLTLML
Ostrich            RFFALHFLLPFVIAGITLVHLTFLHESGSNNPLGIISHCDKIPFHPYFSLKDILGFTLMF
Crow               RFFAFHFLLPFVIAGLTLVHLTFLHETGSNNPLGIPSDCDKIPFHPYYSIKDLLGFALML
Alligator          RFTALHFLLPFALLASLITHLIFLHERGSFNPLGISPNADKIPFHPYFTMKDALGAALAA
Chrysem            RFFTLHFLLPFTIMGLTMVHLLFLHETGSNNPTGLNSNTDKIPFHPYFSYKDLLGVILML
Pelomed            RFFTLHFLTPFIISSLTTIHLLLLHEKGSNNPTGLNSNPDKIPFHPYFSYKDLLGVNLLM
Carassi            RFFAFHFLLPFIIAAATVIHLLFLHETGSNNPIGLNSDADKISFHPYFSYKDLLGFVIML
Carp               RFFAFHFLLPFVIAAATIIHLLFLHETGSNNPIGLNSDADKVSFHPYFSYKDLLGFVIML
Trout              RFFAFHFLFPFVIAAATVLHLLFLHETGSNNPAGINSDADKISFHPYFSYKDLLGFVAML
Salmon             RFFAFHFLFPFVIAAATVLHLLFLHETGSNNPAGINSDADKISFHPYFSYKDLLGFVAML
Smooth Dog Fish    RFFAFHFLLPFLIMALSIIHLLFLHESGSNNPLGINSDADKVSFHPYFSYKDLLGFFVMI
Scyliorhinus       RFFAFHFLLPFLILALSVIHILFLHETGANNPMGINSNTDKISFHPYFSYKDLFGFLIVI
Spiny Dog Fish     RFFAFHFLLPFLIVGLTLIHLLFLHETGSNNPMGLNSDMDKISFHPYFSYKDLLGFFLMI
Skate              RFFAFHFLFPFLIVALTLLHLLFLHEMGSNNPTGLNSNTDKIPFHPYFSYKDLLGFFILG
Polypterus         RFFAFHFILPFAIAAASLVHIVFLHETGSNNPVGINSDADQIPFHPYFTFKDLLGFIILL
Frog               RFFAFHFLLPFIIAGASILHLLFLHETGSTNPTGLNSDPDKVPFHPYFSYKDLLGFLIML
Lung Fish          RFFAFHFLLPFIISAMTAAHFLFLHETGSNNPTGLNSNLDKISFHPYFTMKDLLGFLMLA
Amphioxus          RFFAFHFFLPFMIAGLSVVHLLFLHQTGANNPTGLAGDVDKVPFHAYFSYKDVVGFVVLL
P Urchin           RFFAFHFLFPFIMAALAMIDLVFLHNSGANNPVGLKSNYDKAPFHIYYTTKDTVGFMALI
S Urchin           RFFPFHFLFPFMMAALAVMHLVFLHNSGANNPFAFKSNYDKAPFHIYFTTKDTVGFILLV
                    *  :**: ** :      .: :**: *: ** .:  . *:  ** *.  ** .*
```

```
Fly                 FILISLVLISPNLLGDPDNFIPANPLVTPAHIQPEWYFLFAYAILRSIPNKLGGVIALVL
Mosquito            WILVTFIWKFNYLLMDPENFIPANPLVTPVHIQPEWYFLFAYAILRSIPNKLGGVIALVL
Lamprey             GILFMISLLAPNALGEPDNFIYANPLSTPPHIKPEWYFLFAYAILRSVPNKLGGVVALAA
Blue Whale          LTLLMLTLFAPDLLGDPDNYTPANPLSTPAHIKPEWYFLFAYAILRSIPNKLGGVLALLL
Fin Whale           LILLMLTLFAPDLLGDPDNYTPANPLSTPAHIKPEWYFLFAYAILRSIPNKLGGVLALLL
Hippo               TTLLTLTLFAPDLLGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALAL
Sheep               LILMLLVLFTPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALIL
Cow                 LALMLLVLFAPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALAF
Pig                 LILLILVLFSPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVA
White Rhino         LALLALVLFSPDILGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVL
Black Rhino         LVLLILVLFFPDILGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALAF
Donkey              LLLLTLVLFSPDLLGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALIL
Horse               LLLLTLVLFSPDLLGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALIL
Halicho             LVLTLLVLFSPDLLGDPDNYIPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVL
Seal Vitulina       LVLTLLVLFSPDLLGDPDNYIPPNPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVL
Cat                 LTLMLLVLFSPDLLGDPDNYIPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVL
Dog                 LILMSLVLFSPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVF
Rat                 LFLMTLV--FPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVVALIL
Mouse               LILMTLVLFFPDMLGDPDNYMPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALIL
Myoxus              FLLMTLVLFSPDLLGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVF
Gibbon              LMLMSLVLFSPDLLGDPSNYTQANPLNTPPHIKPEWYFLFAYAILRSVPNKLGGVLALLL
Man                 LSLMTLTLFSPDLLGDPDNYTLANPLNTPPHIKPEWYFLFAYTILRSVPNKLGGVLALLL
Baboon              LALMTLTLFSPDLLNDPDNYTPADPLNTPPHIKPEWYFLFAYAILRSVPNKLGGVLALFL
Platypus            LVLLTLVLFTPDLLGDPDNYTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALVA
Possum              IILLSLAMFSPDLLGDPDNFTPANPLNTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLA
Kangaroo            FILLMLALFSPDMLGDPDNFSPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLA
Chicken             TPFLTLALFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALAA
Ostrich             IPLLSLAFFSPNLLGDPENFTPANPLATPPHIKPEWYFLFAYAILRSIPNKLGGVLALAA
Crow                IPLITLALFSPNLLGDPENFTPANPLATPPHIKPEWYFLFAYAILRSIPNKLGGVLALAA
Alligator           SSLLILALYLPALLGDPENFTPANSMITPTHIKPEWYFLFAYAILRSIPNKLGGVLAMFS
Chrysem             TLLLTLTLFSPNLLGDPDNFTPANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLL
Pelomed             IGLLTLTLFLPNLLTDPENFTPANPLSTPKHIKPEWYFLFAYAILRSIPNKLGGVLALLS
Carassi             LALTLLALFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Carp                LALTLLALFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Trout               LGLTSLALFAPNLLGDPDNFTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Salmon              LGLTSLALFAPNLLGDPDNFTPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Smooth Dog Fish     FLLALLALFLPNLLGDAENFIPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Scyliorhinus        TLLATLALFMPNLLGDAENFIPANPLVTPLHIQPEWYFLFAYAILRSIPNKLGGVLALLF
Spiny Dog Fish      ILLALLALFLPNLLGDAENFIPANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Skate               LLLTLLALFTPNLLGDTENFIPADPLLTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Polypterus          LIIIMLALLSPNLLNDPGNFTPANPLITPPHIKPEWYFLFAYAILRSIPNKLGGVLALLF
Frog                TALTLLAMFSPNLLGDPDNFTPANPLITPPHIKPEWYFLFAYAILRSM-NKLGGVLALVL
Lung Fish           SFLCLLALFSPNLLGDPENFTPANPLVTPTHIKPEWYFLFAYAILRSIPNKLGGVLALMA
Amphioxus           AGLVFIALFSPNLLTDPENYIPANPLVTPVHIQPEWYFLFAYAILRSIPNKLGGVVALAM
P Urchin            AALFVLALLFPCALKDPEKFIPANPLSHPPHMQPEWYFLFAYAILRSIPNKLGGVMALVA
S Urchin            AALFSLALLFPGALKDPEKFIPANPLVTPPHIQPEWYFLFAYAILRSIPNKLGGVIALVA
                      :  :       * :. ::  .:.:  * *::*********:****: ******:*:
```

```
Fly                SIAILMILPFYNLSKFRGIQFYPINQILFWSMLVTVILLTWIGARPVEEPYVLIGQILTI
Mosquito           SIAILLILPFTHSSKFRGLQFYPLNQILFWNMVIVASLLTWIGARPVEDPYILTGQILTV
Lamprey            AIMILLIIPFTHTSKQRGMQFRPLAQITFWILIADLALLTWLGGEPAEYPFILMTQIAST
Blue Whale         SILVLALIPMLHTSKQRSMMFRPFSQFLFWVLVADLLTLTWIGGQPVEHPYVIVGQLASI
Fin Whale          SILILAFIPMLHTSNQRSMMFRPFSQFLFWVLVADLLTLTWIGGQPVEHPYMIVGQLASI
Hippo              SILILALIPMLHTSKQRSLMFRPLSQCLFWALIADLLTLTWIGGQPVEHPFIIIGQVASI
Sheep              SILVLVIMPLLHTSKQRSMMFRPISQCMFWILVADLLTLTWIGGQPVEHPYIIIGQLASI
Cow                SILILALIPLLHTSKQRSMMFRPLSQCLFWALVADLLTLTWIGGQPVEHPYITIGQLASV
Pig                SILILILMPMLHTSKQRGMMFRPLSQCLFWMLVADLITLTWIGGQPVEHPFIIIGQLASI
White Rhino        SILTLLIIPFLHTSKQRSMMFRPLSQCMFWLLVADLLTLTWIGGQPVEHPFIIIGQLASI
Black Rhino        SILILLLIPYLHTSKQRSMMFRPLSQCMFWLLVADLLTLTWIGGQPVEHPFIIIGQLASI
Donkey             SILILALIPTLHMSKQRSMMFRPLSQCVFWLLVADLLTLTWIGGQPVEHPYVIIGQLASI
Horse              SILILALIPTLHMSKQRSMMFRPLSQCVFWLLVADLLTLTWIGGQPVEHPYVIIGQLASI
Halicho            SILILAIVPLLHTSKQRGMMFRPISQCLFWLLVADLLTLTWIGGQPVEHPYITIGQLASI
Seal Vitulina      SILVLAIMPLLHTSKQRGMMFRPISQCLFWFLVADLLTLTWIGGQPVEHPYITVGQLASI
Cat                SILVLAIIPILHTSKQRGMMFRPLSQCLFWLLVADLLTLTWIGGQPVEHPFITIGQLASI
Dog                SILILAFIPLLHTSKQRSMMFRPLSQCLFWLLVADLLTLTWIGGQPVEHPFIIIGQVASI
Rat                SILILAFLPFLHTSKQRSLTFRPITQILYWILVANLLVLTWIGGQPVEHPFIIIGQLASI
Mouse              SILILALMPFLHTSKQRSLMFRPITQILYWILVANLLILTWIGGQPVEHPFIIIGQLASI
Myoxus             SILILAILPVLQFSKQRSMMFRPLSQCPFWILTADLFTLTWIGGQPVEHPFIIIGQLASI
Gibbon             SILILAMIPALHTAKQQSMMFRPLSQLTYWLLVMNLLILTWIGGQPVSYPFITIGQVASA
Man                SILILAMIPILHMSKQQSMMFRPLSQSLYWLLAADLLILTWIGGQPVSYPFTIIGQVASV
Baboon             SILILAAIPMLHKSKQQSMMFRPLSQFLFWLLATTLLTLTWIGSQPVIQPLTTIGQVASM
Platypus           SILILILVPLLHTSYQRGLAFRPLTQMLFWILVTDLLTLTWIGGQPVEQPFIIIGQLASI
Possum             SILVLLIIPMLHTSTQRSMAFRPISQTLFWMLTANLIILTWIGGQPVEQPYITIGQWASI
Kangaroo           SILILLIIPLLHTSKQRSLMFRPISQTLFWILTANLITLTWIGGQPVEQPFIIIGQLASI
Chicken            SVLILFLIPFLHKSKQRTMTFRPLSQTLFWLLVANLLILTWIGSQPVEHPFIIIGQMASL
Ostrich            SVLILFLIPLLHKSKQRSMTFRPLSQLLFWFLVANLLILTWIGSQPVEHPFIIIGQVASF
Crow               SVLVLFLIPLLHVSKQRSMTFRPLSQILFWTLVADLLILTWVGSQPVEHPFIIIGQLASF
Alligator          SILVLFLMPALHTAKQQPMSMRPMSQLLFWALTLDFLLLTWIGGQPVNPPYILIGQTASL
Chrysem            SILVLFLMPALHTSKQRTTQFRPLTQTLFWSFIANLLVLTWIGGQPVENPFITIGQVASI
Pelomed            SVTILFIMPTLHTSKQRSATFRPFTQILFWSPTADLVILTWIGAQPVEDPFIMIGQTASV
Carassi            SILVLMVVPLLHTSKQRGLTFRPITQFLFWTLVADMIILTWIGGMPVEHPFIIIGQIASV
Carp               SILVLMVVPLLHTSKQRGLTFRPITQFLFWTLVADMIILTWIGGMPVEHPFIIIGQIASV
Trout              SILVLMVVPILHTSKQRGLTFRPLTQFLFWALVADMLILTWIGGMPVEHPFIIIGQVASV
Salmon             SILVLMVVPILHTSKQRGLTFRPLTQFLFWTLVADMLILTWIGGMPVEHPFIIIGQIASV
Smooth Dog Fish    SIFILLLVPLLHTSKQRSIIFRPLTQIFFWVLVANSIILTWIGGQPVEQPFIMVGQIASI
Scyliorhinus       SIFILLLVPLLHTSKLRSNIFRPLTQIFFWSLVTNAIILTWIGGQPVEQPFIMVGQIASV
Spiny Dog Fish     SIFILMLIPMLHTSKQRSNIFRPMTQFLFWTLVANAIILTWIGGQPVEQPFILVGQIASV
Skate              SILILMLVPMLHTSKQRSATFRPITQILFWTLLTNTIILTWIGGQPVEQPFIIIGQIASV
Polypterus         SILILMLVPLLHTSKIRSATFRPLFKITLWILAADVLILTWIGGQPVEDPYIIIGQAASI
Frog               SILILALMPLLHTSKQRSLMFRPFTQIMFWALVADTLILTWIGGQPVEDPYTMIGQLASV
Lung Fish          SILILFIIPFLHRAKQRTMSYRPLSQFMFWLLTADMLILTWIGGQPVEHPFILIGQIASA
Amphioxus          SIVVLFFMPFVHSSRQTSHNFRPLAQVLFWLMVVNVLLLTWLGGQPVEYPYIFLGQAASV
P Urchin           AMLVLFLMPLLKTSKKESNSFRPLSQATFWMLVATFFVLTWMGSQPVEQPFVLMGQMASL
S Urchin           AMLVLFLMPLLNTSKKESNSFRPLSQAAFWLLVAHLFMLTWMGSQPVEYPYVLLGQVASV
                   ::  *  :*  : :       *: :   *       ***:*. *.  *     *  :
```

```
Fly                 IYFLYYLIN-PLVTKWWDNLLN
Mosquito            LYFSYFIIN-PLLAKFWDKLLN
Lamprey             VYFMIFILVFPILGYLENKMLL
Blue Whale          LYFLLILVLMPVTSLIENKLMK
Fin Whale           LYFLLILVLMPVTSLIENKLMK
Hippo               LYFLLILVLMPVAGIIENKLLK
Sheep               MYFLIILVMMPVASIIENNLLK
Cow                 LYFLLILVLMPTAGTIENKLLK
Pig                 LYFLIILVLMPITSIIENNLLK
White Rhino         LYFTLILVLMPLAGIIENNLLK
Black Rhino         LYFSLILVLMPLAGIIENNLLK
Donkey              LYFSLILIFMPLASTIENNLLK
Horse               LYFSLILIFMPLASTIENNLLK
Halicho             LYFMILLVLMPIASIIENNILK
Seal Vitulina       LYFTILLVLMPIASIIENNILK
Cat                 LYFSTLLILMPISGIIENRLLK
Dog                 LYFTILLILMPTVSVIENNLLK
Rat                 SYFSIILILMPISGIVEDKMLK
Mouse               SYFSIILILMPISGIIEDKMLK
Myoxus              LYFSIILFFLPTFSLLENKLLK
Gibbon              LYFTTILVLMPAASLIENKMLK
Man                 LYFTTILILMPTISLIENKMLK
Baboon              VYFLTTLVLMPLAAQVENNLLK
Platypus            LYFLLITTLIPLTGLLENDLLK
Possum              SYFTIIIILMPLAGMLENYMLK
Kangaroo            SYFLLIIILMPLAGLFENYMLE
Chicken             SYFTILLILFPTIGTLENKMLN
Ostrich             TYFLILLVLFPAIAALENKMI-
Crow                AYFAIILILFPVVSALENKILK
Alligator           FYFIIILILMPMAGLLENKMVE
Chrysem             LYFSTLLILIPIAGVIENKML-
Pelomed             FYFTLILLLIPLAAILENKLLD
Carassi             LYFALFLVLFPLAGWLENKALK
Carp                LYFALFLIFMPLAGWLENKALK
Trout               IYFTIFLVLSPLAGWAEIKALQ
Salmon              IYFTIFLVLAPLAGWAENKALE
Smooth Dog Fish     SYFALFLIIMPFISWCENKILS
Scyliorhinus        AYFSLFLFVIPITSWCENKFLS
Spiny Dog Fish      TYFSLFLIIIPLTGWWENKMLN
Skate               IYFLLFLILLPLAGWWENKILN
Polypterus          LYFLIFLVLMPLSGWLENKMLN
Frog                IYFSIFIIMFPLMGWVENKLLN
Lung Fish           TYFLLFLLLFPLITSLENKLLY
Amphioxus           IYFVNILLLIPIVGYVENKLL-
P Urchin            LYFSLFMFGFPLVSSLEKKMMF
S Urchin            LYFSLFMFGFPMVSSMENKIMF
                     **       *        :
```

Rhodopsin

```
Japanese lamprey        MNGTEGDNFYVPFSNKTGLARSPYEYPQYYLAEPWKYSALAAYMFFLILVGFPVNFLTLF
Sea lamprey             MNGTEGENFYIPFSNKTGLARSPFEYPQYYLAEPWKYSVLAAYMFFLILVGFPVNFLTLF
Green anole             MNGTEGQNFYVPMSNKTGVVRNPFEYPQYYLADPWQFSALAAYMFLLILLGFPINFLTLF
Toad                    MNGTEGPNFYIPMSNKTGVVRSPFEYPQYYLAEPWQYSILCAYMFLLILLGFPINFMTLY
Frog                    MNGTEGPNFYVPMSNKTGIVRSPFEYPQYYLAEPWKYSVLAAYMFLLILLGLPINFMTLY
Salamander              MNGTEGPNFYVPFSNKSGVVRSPFEYPQYYLAEPWQYSVLAAYMFLLILLGFPVNFLTLY
Alligator               MNGTEGPDFYIPFSNKTGVVRSPFEYPQYYLAEPWKYSALAAYMFMLIILGFPINFLTLY
Chicken                 MNGTEGQDFYVPMSNKTGVVRSPFEYPQYYLAEPWKFSALAAYMFMLILLGFPVNFLTLY
Cow                     MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
Sheep                   MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Whale                   MNGTEGLNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSVLAAYMFLLIVLGFPINFLTLY
Dolphin                 MNGTEGLNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSVLAAYMFLLIVLGFPINFLTLY
Pig                     MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFMLIVLGFPINFLTLY
Dog                     MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Seal                    MNGTEGPNFYVPFSNKTGVVRSPFEFPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Mouse                   MNGTEGPNFYVPFSNVTGVGRSPFEQPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Rat                     MNGTEGPNFYVPFSNITGVVRSPFEQPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Hamster                 MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Rabbit                  MNGTEGPDFYIPMSNQTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
Blackmouth catshark     MNGTEGENFYVPMSNKTGVVRNPFEYPQYYLADHWMFAVLAAYMFFLIITGFPVNFLTLF
Spotted dogfish         MNGTEGENFYIPMSNKTGVVRSPFDYPQYYLAEPWKFSVLAAYMFFLIIAGFPVNFLTLY
Little skate            MNGTEGENFYVPMSNKTGVVRSPFDYPQYYLGEPWMFSALAAYMFFLILTGLPVNFLTLF
Goldfish                MNGTEGDMFYVPMSNATGIVRSPYDYPQYYLVAPWAYACLAAYMFFLIITGFPVNFLTLY
Common carp             MNGTEGPMFYVPMSNATGVVKSPYDYPQYYLVAPWAYGCLAAYMFFLIITGFPINFLTLY
Guppy                   MNGTEGPYFYVPMVNTTGIVRSPYEYPQYYLVSPAAYACLGAYMFFLILVGFPINFLTLY
Blind cave fish         MNGTEGPYFYVPMSNATGVVRSPYEYPQYYLAPPWAYACLAAYMFFLILVGFPVNFLTLY
                        ******  **:*: * :*: :.*:: *****     :. * ****:**: *:*:**:**:

Japanese lamprey        VTVQHKKLRTPLNYILLNLAMANLFMVLFGFTVTMYTSMNGYFVFGPTMCSIEGFFATLG
Sea lamprey             VTVQHKKLRTPLNYILLNLAVANLFMVLFGFTLTMYSSMNGYFVFGPTMCNFEGFFATLG
Green anole             VTIQHKKLRTPLNYILLNLAVANLFMVLMGFTTTMYTSMNGYFIFGTVGCNIEGFFATLG
Toad                    VTIQHKKLRTPLNYILLNLAFANHFMVLCGFTVTMYSSMNGYFILGATGCYVEGFFATLG
Frog                    VTIQHKKLRTPLNYILLNLAFANHFMVLCGFTITMYTSLHGYFVFGQTGCYFEGFFATLG
Salamander              VTIQHKKLRTPLNYILLNLAFANHFMVFGGFPVTMYSSMHGYFVFGQTGCYIEGFFATMG
Alligator               VTVQHKKLRSPLNYILLNLAVADLFMVLGGFTTTLYTSMNGYFVFGVTGCYFEGFFATLG
Chicken                 VTIQHKKLRTPLNYILLNLVVADLFMVFGGFTTTMYTSMNGYFVFGVTGCYIEGFFATLG
Cow                     VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Sheep                   VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Whale                   VTVQHKKLRTPLNYIPLNLAVANLFMVFGGFTTTLYTSLHAYFVFGPTGCNLEGFFATLG
Dolphin                 VTVQHKKLRTPLNYILLNLAVANLFMVFGGFTTTLYTSLHAYFVFGPTGCNLEGFFATLG
Pig                     VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Dog                     VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNVEGFFATLG
Seal                    VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Mouse                   VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Rat                     VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Hamster                 VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
Rabbit                  VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTTTLYTSLHGYFVFGPTGCNVEGFFATLG
Blackmouth catshark     VTIQNKKLRQPLNYILLNLAVANLFMVFGGFTTTLITSMNGYFVFGSTGCNLEGFFATLG
Spotted dogfish         VTIQHKKLRQPLNYILLNLAVADLFMIFGGFPSTMITSMNGYFVFGPSGCNFEGFFATLG
Little skate            VTIQHKKLRQPLNYILLNLAVSDLFMVFGGFTTTIITSMNGYFIFGPAGCNFEGFFATLG
Goldfish                VTIEHKKLRTPLNYILLNLAISDLFMVFGGFTTTMYTSLHGYFVFGRVGCNPEGFFATLG
Common carp             VTIEHKKLRTPLNYILLNLAISDLFMVFGGFTTTMYTSLHGYFVFGRIGCNLEGFFATLG
Guppy                   VTIEHKKLRTPLNYILLNLAVADLFMVFGGFTTTIYTSMHGYFVLGRLGCNLEGYFATLG
Blind cave fish         VTIEHKKLRTPLNYILLNLAVADLFMVFGGFTTTMYTSLNGYFVFGRLGCNLEGFFATFG
                        **:::**** ***** ***..:: **:: **. *: :*::.**::*   *  **:***:*
```

```
Japanese lamprey        GEVALWSLVVLAIERYIVICKPMGNFRFGNTHAIMGVAFTWIMALACAAPPLVGWSRYIP
Sea lamprey             GEMSLWSLVVLAIERYIVICKPMGNFRFGSTHAYMGVAFTWFMALSCAAPPLVGWSRYLP
Green anole             GEMGLWSLVVLAVERYVVICKPMSNFRFGETHALIGVSCTWIMALACAGPPLLGWSRYIP
Toad                    GEIALWSLVVLAIERYVVVCKPMSNFRFSENHAVMGVAFTWIMALSCAVPPLLGWSRYIP
Frog                    GEIALWSLVVLAIERYIVVCKPMSNFRFGENHAMMGVAFTWIMALACAVPPLFGWSRYIP
Salamander              GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVMMTWIMALACAAPPLFGWSRYIP
Alligator               GEVALWCLVVLAIERYIVVCKPMSNFRFGENHAIMGVVFTWIMALTCAAPPLVGWSRYIP
Chicken                 GEIALWSLVVLAVERYVVVCKPMSNFRFGENHAIMGVAFSWIMAMACAAPPLFGWSRYIP
Cow                     GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
Sheep                   GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
Whale                   GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGLALTWVMAMACAAPPLVGWSRYIP
Dolphin                 GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGLALTWIMAMACAAAPLVGWSRYIP
Pig                     GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGLALTWVMALACAAPPLVGWSRYIP
Dog                     GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLACWSRYIP
Seal                    GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVGFTWVMALACAAPPLVGWSRYIP
Mouse                   GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVVFTWIMALACAAPPLVGWSRYIP
Rat                     GEIGLWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
Hamster                 GEIALWSLVVLAIERYVVICKPMSNFRFGENHAIMGVVFTWIMALACAAPPLVGWSRYIP
Rabbit                  GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWIMALACAAPPLVGWSRYIP
Blackmouth catshark     GEISLWSLVVLAIERYVVVCKPMSNFRFGSQHAIAGVSLTWVMAMACAAPPLVGWSRYIP
Spotted dogfish         GEIGLWSLVVLAIERYVVVCKPMSNFRFGSQHAFMGVGLTWIMAMACAFPPLVGWSRYIP
Little skate            GEVGLWCLVVLAIERYMVVCKPMANFRFGSQHAIIGVVFTWIMALSCAGPPLVGWSRYIP
Goldfish                GEMGLWSLVVLAFERWMVVCKPVSNFRFGENHAIMGVVFTWFMACTCAVPPLVGWSRYIP
Common carp             GEMGLWSLVVLAFERWMVVCKPVSNFRFGENHAIMGVVFTWFMACTCAVPPLVGWSRYIP
Guppy                   GEIGLWSLVVLAVERWLVVCKPISNFRFSENHAIMGLVFTWIMANSCAAPPLLGWSRYIP
Blind cave fish         GINSLWCLVVLSIERWVVVCKPMSNFRFGENHAIMGVAFTWFMALACTVPPLVGWSRYIP
                        *  .**.****:.**::*:***:.****.. **  *:  :*.** :*: .** *****:*

Japanese lamprey        EGMQCSCGPDYYTLNPNFNNESYVVYMFVVHFLVPFVIIFFCYGRLLCTVKEAAAAQQES
Sea lamprey             EGMQCSCGPDYYTLNPNFNNESFVIYMFLVHFIIPFIVIFFCYGRLLCTVKEAAAAQQES
Green anole             EGMQCSCGVDYYTPTPEVHNESFVIYMFLVHFVTPLTIIFFCYGRLVCTVKAAAAQQQES
Toad                    EGMQCSCGVDYYTLKPEVNNESFVIYMFVVHFTIPLIIIFFCYGRLVCTVKEAAAQQQES
Frog                    EGMQCSCGVDYYTLKPEVNNESFVIYMFVVHFLIPLIIISFCYGRLVCTVKEAAAQQQES
Salamander              EGMQCSCGVDYYTLKPEVNNESFVIYMFLVHFTIPLMIIFFCYGRLVCTVKEAAAQQQES
Alligator               EGMQCSCGVDYYTLKPEVNNESFVIYMFVVHFAIPLAVIFFCYGRLVCTVKEAAAQQQES
Chicken                 EGMQCSCGIDYYTLKPEINNESFVIYMFVVHFMIPLAVIFFCYGNLVCTVKEAAAQQQES
Cow                     EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
Sheep                   QGMQCSCGALYFTLKPEINNESFVIYMFVVHFSIPLIVIFFCYGQLVFTVKEAAAQQQES
Whale                   EGMQCSCGIDYYTSRQEVNNESFVIYMFVVHFTIPLVIIFFCYGQLVFTVKEAAAQQQES
Dolphin                 EGMQCSCGIDYYTSRQEVNNESFVIYMFVVHFTIPLVIIFFCYGQLVFTVKEAAAQQQES
Pig                     EGLQCSCGIDYYTLKPEVNNESFVIYMFVVHFSIPLVIIFFCYGQLVFTVKEAAAQQQES
Dog                     EGMQCSCGIDYYTLKPEINNESFVIYMFVVHFAIPMIVIFFCYGQLVFTVKEAAAQQQES
Seal                    EGMQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIVIFFCYGQLVFTVKEAAAQQQES
Mouse                   EGMQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIVIFFCYGQLVFTVKEAAAQQQES
Rat                     EGMQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIVIFFCYGQLVFTVKEAAAQQQES
Hamster                 EGMQCSCGVDYYTLKPEVNNESFVIYMFVVHFTIPLIVIFFCYGQLVFTVKEAAAQQQES
Rabbit                  EGMQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPLIIIFFCYGQLVFTVKEAAAQQQES
Blackmouth catshark     EGLQCSCGIDYYTPKPEINNVSFVIYMFVVHFSIPLTIIFFCYGRLVCTVKAAAAQQQES
Spotted dogfish         EGMQCSCGIDYYTLKPEVNNESFVIYMFVVHFSIPLTIIFFCYGRLVCTVKEAAAQQQES
Little skate            EGLQCSCGVDYYTMKPEVNNESFVIYMFVVHFTIPLIVIFFCYGRLVCTVKEAAAQQQES
Goldfish                EGMQCSCGVDYYTRPQAYNNESFVIYMFIVHFIIPLVIFFCYGRLVCTVKEAAAQHEES
Common carp             EGMQCSCGVDYYTRAPGYNNESFVIYMFLVHFIIPLVIFFCYGRLVCTVKDAAAQQQES
Guppy                   EGMQCSCGVDYYTRAEGFNNESFVIYMFICHFCIPLIVVFFCYGRLLCAVKEAAAAQQES
Blind cave fish         EGMQCSCGIDYYTRAEGFNNESFVIYMFVVHFLTPLFVITFCYGRLVCTVKEAAAQQQES
                        :*:*****  *:*     :* *:*:***: **  *: :: ****.*: :** *** ::**
```

```
Japanese lamprey        ASTQKAEKEVTRMVVLMVIGFLVCWVPYASVAFYIFTHQGSDFGATFMTLPAFFAKSSAL
Sea lamprey             ASTQKAEKEVTRMVVLMVIGFLVCWVPYASVAFYIFTHQGSDFGATFMTVPAFFAKTSAL
Green anole             ATTQKAEREVTRMVVIMVISFLVCWVPYASVAFYIFTHQGSDFGPVFMTIPAFFAKSSAI
Toad                    ATTQKAEKEVTRMVIIMVVFFLICWVPYASVAFFIFSNQGSEFGPIFMTVPAFFAKSSSI
Frog                    ATTQKAEKEVTRMVVIMVIFFLICWVPYAYVAFYIFTHQGSEFGPIFMTVPAFFAKSSAI
Salamander              ATTQKAEKEVTRMVIIMVVAFLICWVPYASVAFYIFSNQGTDFGPIFMTVPAFFAKSSAI
Alligator               ATTQKAEKEVTRMVIIMVVSFLICWVPYASVAFYIFSNQGSDFGPVFMTIPAFFAKSSAI
Chicken                 ATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTNQGSDFGPIFMTIPAFFAKSSAI
Cow                     ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
Sheep                   ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKSSSV
Whale                   ATTQKAEKEVTRMVIIMVVAFLICWVPYASVAFYIFTHQGSDFGPIFMTIPSFFAKSSSI
Dolphin                 ATTQKAEKEVTRMVIIMVVAFLICWVPYASVAFYIFTHQGSDFGPIFMTIPSFFAKSSSI
Pig                     ATTQKAEKEVTRMVIIMVVAFLICWLPYASVAFYIFTHQGSDFGPIFMTIPAFFAKSASI
Dog                     ATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQGSDFGPIFMTLPAFFAKSSSI
Seal                    ATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQGSNFGPIFMTLPAFFAKAASI
Mouse                   ATTQKAEKEVTRMVIIMVIFFLICWLPYASVAFYIFTHQGSNFGPIFMTLPAFFAKSSSI
Rat                     ATTQKAEKEVTRMVIIMVIFFLICWLPYASVAMYIFTHQGSNFGPIFMTLPAFFAKTASI
Hamster                 ATTQKAEKEVTRMVILMVVFFLICWFPYAGVAFYIFTHQGSNFGPIFMTLPAFFAKSSSI
Rabbit                  ATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQGSNFGPIFMTIPAFFAKSSSI
Blackmouth catshark     ETTQRAEREVTRMVVIMVIGFLICWLPYASVALYIFNNQGSEFGPVFMTIPSFFAKSSAL
Spotted dogfish         ETTQRAEREVTRMVIIMVIAFLICWLPYASVAFFIFCNQGSEFGPIFMTIPAFFAKAASL
Little skate            ESTQRAEREVTRMVIIMVVAFLICWVPYASVAFYIFINQGCDFTPFFMTVPAFFAKSSAV
Goldfish                ETTQRAEREVTRMVVIMVIGFLICWIPYASVAWYIFTHQGSEFGPVFMTLPAFFAKTAAV
Common carp             ETTQRAEREVTRMVVIMVIGFLICWIPYASVAWYIFTHQGSEFGPVFMTVPAFFAKSAAV
Guppy                   ETTQRAEREVTRMVVIMVIGFLVCWIPYASVAWYIFTHQGSEFGPLFMTVPAFFAKSASI
Blind cave fish         ETTQRAEREVTRMVILMFIAYLVCWLPYASVSWWIFTNQGSEFGPIFMTVPAFFAKSSSI
                         :**:**:******::*.: :*:**.*** *: :** :** :* . ***:*:****::::

Japanese lamprey        YNPVIYILMNKQFRNCMITTLCCGKNPLGDDESGASTS-KTEVSSVSTSPVSPA
Sea lamprey             YNPIIYILMNKQFRNCMITTLCCGKNPLGDEDSGASTS-KTEVSSVSTSQVSPA
Green anole             YNPVIYILMNKQFRNCMIMTLCCGKNPLGDEETSAGT--KTETSTVSTSQVSPA
Toad                    YNPVIYIMLNKQFRNCMITTLCCGKNPFGEDDASSAATSKTEASSVSSSQVSPA
Frog                    YNPVIYIMLNKQFRNCMITTLCCGKNPFGDEDASSAATSKTEATSVSTSQVSPA
Salamander              YNPVIYIVLNKQFRNCMITTICCGKNPFGDDETTSAATSKTEASSVSSSQVSPA
Alligator               YNPVIYIVMNKQFRNCMITTLCCGKNPLGDDETATGSK--TETSSVSTSQVSPA
Chicken                 YNPVIYIVMNKQFRNCMITTLCCGKNPLGDEDTSAG-K--TETSSVSTSQVSPA
Cow                     YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTT------VSKTETSQVAPA
Sheep                   YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTT------VSKTETSQVAPA
Whale                   YNPVIYIMMNKQLRNCMLTTLCCGRNPLGDDEASTT------ASKTETSQVAPA
Dolphin                 YNPVIYIMMNKQFRNCMLTTLCCGRNPLGDDEASTT------ASKTETSQVAPA
Pig                     YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTT------TSKTETSQVAPA
Dog                     YNPVIYIMMNKQFRNCMITTLCCGKNPLGDDEASAS------ASKTETSQVAPA
Seal                    YNPVIYIMMNKQFRTCMITTLCCGKNPLGDDEVSAS------ASKTETSQVAPA
Mouse                   YNPVIYIMLNKQFRNCMLTTLCCGKNPLGDDDASAT------ASKTETSQVAPA
Rat                     YNPIIYIMMNKQFRNCMLTSLCCGKNPLGDDEASAT------ASKTETSQVAPA
Hamster                 YNPVIYIMMNKQFRNCMLTTLCCGKNILGDDEASAT------ASKTETSQVAPA
Rabbit                  YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASAT------ASKTETSQVAPA
Blackmouth catshark     YNPLIYILMNKQFRNCMITTLCCGKNPFEEEESTSASASKTEASSVSSSQVSPA
Spotted dogfish         YNPLIYILMNKQFRNCMITTICCGKNPFEEEESTSASASKTEASSVSSSQVAPA
Little skate            YNPLIYILMNKQFRNCMITTICLGKNPFEEEESTSASASKTEASSVSSSQVAPA
Goldfish                YNPCIYICMNKQFRHCMITTLCCGKNPFEEEEGASTTASKTEASSVSSSSVSPA
Common carp             YNPCIYICMNKQFRHCMITTLCCGKNPFEEEEGASTTASKTEASSVSSSSVSPA
Guppy                   YNPLIYICMNKQFRHCMITTLCCGKNPFEEEEGASTTASKTEASSVSSSSVSPA
Blind cave fish         YNPVIYICLNKQFRHCMITTLCCGKNPFEEEEGASTTASKTEASSVSS--VSPA
                        *** *** :***:* **: ::* *:* : :::  :       .:...:  *:**
```

Myoglobin

```
Green sea turtle      GLSDDEWNHVLGIWAKVEPDLTAHGQEVIIRLFQLHPETQERFAKFKNLTTIDALKSSEE
Map turtle            GLSDDEWHHVLGIWAKVEPDLSAHGQEVIIRLFQVHPETQERFAKFKNLKTIDELRSSEE
Alligator             ELSDQEWKHVLDIWTKVESKLPEHGHEVIIRLLQEHPETQERFEKFKHMKTADEMKSSEK
Lace monitor          GLSDEEWKKVVDIWGKVEPDLPSHGQEVIIRMFQNHPETQDRFAKFKNLKTLDEMKNSED
Human                 GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
Chimpanzee            GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
Gorilla               GLSDGEWQLVLNVWGKVEADISGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
Pig                   GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
Aardvark              GLSDAEWQLVLNVWGKVEADIPGHGQDVLIRLFKGHPETLEKFDRFKHLKTEDEMKASED
Tree shrew            GLSDGEWQLVLNVWGKVEADVAGHGQEVLIRLFKGHPETLEKFDKFKHLKTEDEMKASED
Rabbit                GLSDAEWQLVLNVWGKVEADLAGHGQEVLIRLFHTHPETLEKFDKFKHLKSEDEMKASED
Mouse                 GLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSED
Rat                   GLSDGEWQLVLNVWGKVEGDLAGHGQEVLIKLFKNHPETLEKFDKFKHLKSEDEMKGSED
Dog                   GLSDGEWQIVLNIWGKVETDLAGHGQEVLIRLFKNHPETLDKFDKFKHLKTEDEMKGSED
Fox                   GLSDGEWQLVLNIWGKVETDLAGHGQEVLIRLFKNHPETLDKFDKFKHLKTEDEMKGSED
Badger                GLSDGEWQLVLNVWGKVEADLAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKGSED
Otter                 GLSDGEWQLVLNVWGKVEADLAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKGSED
Muskrat               GLSDGEWQLVLHVWGKVEADLAGHGQDVLIRLFKAHPETLEKFDKFKHIKSEDEMKGSED
Deer                  GLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
Sheep                 GLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
Cow                   GLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
Horse                 GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED
Elephant              GLSDGEWELVLKTWGKVEADIPGHGEFVLVRLFTGHPETLEKFDKFKHLKTEGEMKASED
Whale                 VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
Dolphin               GLSDGEWQLVLNVWGKVEADLAGHGQDVLIRLFKGHPETLEKFDKFKHLKTEADMKASED
Seal                  GLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSED
Possum                GLSDGEWQLVLNAWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
Kangaroo              GLSDGEWQLVLNIWGKVETDEGGHGKDVLIRLFKGHPETLEKFDKFKHLKSEDEMKASED
Duckbill platypus     GLSDGEWQLVLKVWGKVEGDLPGHGQEVLIRLFKTHPETLEKFDKFKGLKTEDEMKASAD
Echidna               GLSDGEWQLVLKVWGKVETDITGHGQDVLIRLFKTHPETLEKFDKFKHLKTEDEMKASAD
Emperor penguin       GLNDQEWQQVLTMWGKVESDLAGHGHAVLMRLFKSHPETMDRFDKFRGLKTPDEMRGSED
Chicken               GLSDQEWQQVLTIWGKVEADIAGHGHEVLMRLFHDHPETLDRFDKFKGLKTPDQMKGSED
Common carp           ----HDAELVLKCWGGVEADFEGTGGEVLTRLFKQHPETQKLFPKFVGIAS-NELAGNAA
Yellowfin tuna        ----ADFDAVLKCWGPVEADYTTMGGLVLTRLFKEHPETQKLFPKFAGIAQ-ADIAGNAA
                       :   : . *:  *  ** .    *  ::  ::  **** . * :*  :     :  .
```

```
Green sea turtle      VKKHGTTVLTALGRILKQKNNHEQELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKHP
Map turtle            VKKHGTTVLTALGRILKLKNNHEPELKPLAESHATKHKIPVKYLEFICEIIVKVIAEKHP
Alligator             MKQHGNTVFTALGNILKQKGNHAEVLKPLAKSHALEHKIPVKYLEFISEIIVKVIAEKYP
Lace monitor          LKKHGTTVLTALGRILKQKGHHEAEIAPLAQTHANTHKIPIKYLEFICEVIVGVIAEKHS
Human                 LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP
Chimpanzee            LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLHSKHP
Gorilla               LKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHP
Pig                   LKKHGNTVLTALGGILKKKGHHEAELTPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
Aardvark              LKKHGTTVLTALGGILKKKGQHEAEIQPLAQSHATKHKIPVKYLEFISEAIIQVIQSKHS
Tree shrew            LKKHGNTVLSALGGILKKKGQHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
Rabbit                LKKHGNTVLTALGAILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIHVLHSKHP
Mouse                 LKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRHS
Rat                   LKKHGNTVLTALGGILKKKGQHAAEIQPLAQSHATKHKIPIKYLEFISEAIIQVLQSKHP
Dog                   LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIIQVLQSKHS
Fox                   LKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPVKYLEFISDAIIQVLQSKHS
Badger                LKKHGNTVLTALGGILKKKGHQEAELKPLAQSHATKHKIPVKYLEFISDAIAQVLQSKHP
Otter                 LKKHGNTVLTALGGILKKKGKHEAELKPLAQSHATKHKIPIKYLEFISEAIIQVLQSKHP
Muskrat               LKKHGBTVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPIKYLEFISEAIIHVLZSKHP
Deer                  LKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHP
Sheep                 LKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHP
Cow                   LKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHP
Horse                 LKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHP
Elephant              LKKQGVTVLTALGGILKKKGHHEAEIQPLAQSHATKHKIPIKYLEFISDAIIHVLQSKHP
Whale                 LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
Dolphin               LKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
Seal                  LRKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSKHP
Possum                LKKHGATVLTALGNILKKKGNHEAELKPLAQSHATKHKISVQFLEFISEAIIQVIQSKHP
Kangaroo              LKKHGITVLTALGNILKKKGHHEAELKPLAQSHATKHKIPVQFLEFISDAIIQVIQSKHA
Duckbill platypus     LKKHGGTVLTALGNILKKKGQHEAELKPLAQSHATKHKISIKFLEYISEAIIHVLQSKHS
Echidna               LKKHGGVVLTALGSILKKKGQHEAELKPLAQSHATKHKISIKFLEFISEAIIHVLQSKHS
Emperor penguin       MKKHGVTVLT-LGQILKKKGHHEAELKPLSQTHATKHKVPVKYLEFISEAIMKVIAQKHA
Chicken               LKKHGATVLTQLGKILKQKGNHESELKPLAQTHATKHKIPVKYLEFISEVIIKVIAEKHA
Common carp           VKAHGATVLKKLGELLKARGDHAAILKPLATTHANTHKIALNNFRLITEVLVKVMAEKAG
Yellowfin tuna        ISAHGATVLKKLGELLKAKGSHAAILKPLANSHATKHKIPINNFKLISEVLVKVMHEKAG
                      :  :* .*:. ** :** :. :   :  *: :**  **:.:: :. * : :  *:  :
```

```
Green sea turtle        SDFGADSQAAMKKALELFRNDMASKYKEFGFLG
Map turtle              SDFGADSQAAMRKALELFRNDMASKYKEFGFQG
Alligator               ADFGADSQAAMRKALELFRNDMASKYKEFGYQG
Lace monitor            ADFGADSQEAMRKALELFRNDMASRYKELGFQG
Human                   GDFGADAQGAMNKALELFRKDMASNYKELGFQG
Chimpanzee              GDFGADAQGAMNKALELFRKDMASNYKELGFQG
Gorilla                 GDFGADAQGAMNKALELFRKDMASNYKELGFQG
Pig                     GDFGADAQGAMSKALELFRNDMAAKYKELGFQG
Aardvark                GDFGADAQGAMSKALELFRNDIAAKYKELGFQG
Tree shrew              GDFGADAQAAMSKALELFRNDIAAKYKELGFQG
Rabbit                  GDFGADAQAAMSKALELFRNDIAAQYKELGFQG
Mouse                   GDFGADAQGAMSKALELFRNDIAAKYKELGFQG
Rat                     GDFGADAQGAMSKALELFRNDIAAKYKELGFQG
Dog                     GDFHADTEAAMKKALELFRNDIAAKYKELGFQG
Fox                     GDFHADTEAAMKKALELFRNDIAAKYKELGFQG
Badger                  GNFAAEAQGAMKKALELFRNDIAAKYKELGFQG
Otter                   GBFGADAQGAMKRALELFRNDIAAKYKELGFQG
Muskrat                 SBFGADVZGAMKRALELFRNDIAAKYKELGFQG
Deer                    SNFGADAQGAMSKALELFRNDMAAQYKVLGFQG
Sheep                   SNFGADAQGAMSKALELFRNDMAAEYKVLGFQG
Cow                     SDFGADAQAAMSKALELFRNDMAAQYKVLGFHG
Horse                   GDFGADAQGAMTKALELFRNDIAAKYKELGFQG
Elephant                AEFGADAQAAMKKALELFRNDIAAKYKELGFQG
Whale                   GDFGADAQGAMNKALELFRKDIAAKYKELGYQG
Dolphin                 AEFGADAQGAMNKALELFRKDIAAKYKELGFHG
Seal                    AEFGADAQAAMKKALELFRNDIAAKYKELGFHG
Possum                  GDFGGDAQAAMGKALELFRNDMAAKYKELGFQG
Kangaroo                GNFGADAQAAMKKALELFRHDMAAKYKEFGFQG
Duckbill platypus       ADFGADAQAAMGKALELFRNDMAAKYKEFGFQG
Echidna                 ADFGADAQAAMGKALELFRNDMATKYKEFGFQG
Emperor penguin         SNFGADAQEAMKKALELFRNDMASKYKEFGFQG
Chicken                 ADFGADSQAAMKKALELFRNDMASKYKEFGFQG
Common carp             LDAGG--QSALRRVMDVVIGDIDTYYKEIGFAG
Yellowfin tuna          LDAGG--QTALRNVMGIIIADLEANYKELGFSG
                            .    *: ..: :.  *: : ** :*: *
```

Hemoglobin α

```
Iguana                VLTEDDKNHIRAIWGHVDNNPEAFGVEALTR--LFLAYPATKTYFAHF-DLNPGSAQIKA
Monitor lizard        VLTEDDKNHVKGLWAHVHDHIDEIAADALTR--MFLAHPASKTYFAHF-DLSPDNAQIKA
Whale                 VLSPTDKSNVKATWAKIGNHGAEYGAEALER--MFMNFPSTKTYFPHF-DLGHDSAQVKG
Dolphin               VLSPADKTNVKGTWSKIGNHSAEYGAEALER--MFINFPSTKTYFSHF-DLGHGSAQIKG
Seal                  VLSPADKTNVKTTWDKLGGHAGEYGGEALER--TFTAFPTTKTYFPHF-DLSHGSAQVKA
Walrus                VLSPADKTNVKTTWDKLGGHAGEYGGEALER--TFMSFPTTKTYFPHF-DLSPGSAQVKA
Dog                   VLSPADKTNIKSTWDKIGGHAGDYGGEALDR--TFQSFPTTKTYFPHF-DLSPGSAQVKA
Fox                   VLSPADKTNIKSTWDKIGGHAGDYGGEALDR--TFQSFPTTKTYFPHF-DLSPGSAQVKA
Giant panda           VLSPADKTNVKATWDKIGGHAGEYGGEALER--TFASFPTTKTYFPHF-DLSPGSAQVKA
Sun bear              VLSPADKSNVKATWDKIGSHAGEYGGEALER--TFASFPTTKTYFPHF-DLSPGSAQVKA
Cat                   VLSAADKSNVKACWGKIGSHAGEYGAEALER--TFCSFPTTKTYFPHF-DLSHGSAQVKA
Lynx                  VLSAADKSNVKACWGKIGSHAGDYGTEALER--TFCSFPTTKTYFPHF-DLSHGSAQVKA
Leopard               VLSSADKNNVKACWGKIGSHAGEYGAEALER--TFCSFPTTKTYFPHF-DLSHGSAQVQA
Palm civet            VLSSADKNNIKATWDKIGSHAGEYGAEALER--TFISFPTTKTYFPHF-DLSHGSAQVKA
Lemur                 VLSPADKNNVKSAWNAIGSHAGEHGAEALER--MFLSFPPTKTYFPHF-DLSHGSAQIKT
Gorilla               VLSPADKTNVKAAWGKVGAHAGDYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAZVKG
Chimpanzee            VLSPADKTNVKAAWGKVGAHAGZYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAZVKG
Mandrill              VLSPADKKNVKAAWDKVGGHAGEYGAEALER--MFLSFPTTKTYFPHF-NLSHGSDQVKG
Baboon                VLSPDDKKHVKAAWGKVGEHAGEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSDQVNK
Green monkey          VLSPADKSNVKAAWGKVGGHAGEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAQVKG
Yak                   VLSAADKGNVKAAWGKVGGHAAEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAQVKG
Cow                   VLSAADKGNVKAAWGKVGGHAAEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAQVKG
Goat                  VLSAADKSNVKAAWGKVGGNAGAYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAQVKG
Hippopotamus          VLSANDKSNVKAAWGKVGNHAPEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSSQVKA
Pig                   VLSAADKANVKAAWGKVGGQAGAHGAEALER--MFLGFPTTKTYFPHF-NLSHGSDQVKA
Horse                 VLSAADKTNVKAAWSKVGGHAGEYGAEALER--MFLGFPTTKTYFPHF-DLSHGSAQVKA
Zebra                 VLSAADKTNVKAAWSKVGGNAGEFGAEALER--MFLGFPTTKTYFPHF-DLSHGSAQVKA
White rhinoceros      VLSPTDKTNVKTAWGHVGAQAGEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAQVKA
Indian rhinoceros     VLSPTDKTNVKTAWSHVGAHAGEYGAEALER--MFLSFPTTKTYFPHF-DLSHGSAQVKA
Tapir                 VLSPTDKTNVKAAWSKVGSHAGEYGAEALER--MFLGFPTTKTYFPHF-DLSHGSAQVQA
Camel                 VLSSKDKTNVKTAFGKIGGHAAEYGAEALER--MFLGFPTTKTYFPHF-DLSHGSAQVKA
Llama                 VLSSKDKANIKTAFGKIGGHAADYGAEALER--MFLGFPTTKTYFPHF-DLSHGSAQVKA
Mouse                 VLSGEDKSNIKAAWGKIGGHGAEYGAEALER--MFASFPTTKTYFPHF-DVSHGSAQVKG
Rat                   VLSADDKTNIKNCWGKIGGHGGEYGEEALQR--MFAAFPTTKTYFSHI-DVSPGSAQVKA
Possum                VLSANDKTNVKGAWSKVGGNSGAYMGEALYR--TFLSFPTTKTYFPNY-DFSAGSAQIKT
Kangaroo              VLSAADKGHVKAIWGKVGGHAGEYAAEGLER--TFHSFPTTKTYFPHF-DLSHGSAQIQA
Echidna               VLTDAEKKEVTSLWGKASGHAEEYGAEALER--LFLSFPTTKTYFSHM-DLSKGSAQVKA
Platypus              MLTDAEKKEVTALWGKAAGHGEEYGAEALER--LFQAFPTTKTYFSHF-DLSHGSAQIKA
Alligator             VLSMEDKSNVKAIWGKASGHLEEYGAEALER--MFCAYPQTKIYFPHF-DMSHNSAQIRA
Crocodile             VLSSDDKCNVKAVWSKVAGHLEEYGAEALER--MFCAYPQTKIYFPHF-DLSHGSAQIRA
Snake                 VLSEDDKNRVRTSVGKNPELPGEYGSETLTR--MFAAHPTTKTYFPHF-DLSSGSPNLKA
Duck                  MLTAEDKKLITQLWEKVAGHQEEFGSEALQR--MFLAYPQTKTYFPHF-DLHPGSEQVRG
Goose                 MLTADDKKLLAQLWEKVAGHQDEFGNEALQR--MFVTYPQTKTYFPHF-DLHPGSEQVRS
Rhea                  MLTADDKKLISQIWTKVAEHGGEFGGEALER--MFITYPQTKTYFPHF-DLHVGSEQVRG
Ostrich               MLTADDKKLIQQIWEKVGSHLEDFGAEALER--MFITYPQTKTYFPHF-DLHPGSEQIRG
Chicken               MLTAEDKKLIQQAWEKAASHQEEFGAEALTR--MFTTYPQTKTYFPHF-DLSPGSDQVRG
Bullfrog              SLSASEKAAVLSIVGKIGSQGSALGSEALTR--LFLSFPQTKTYFPHF-DLTPGSADLNT
Newt                  VLSAEEKALVVGLCGKISGHCDALGGEALDR--LFASFGQTRTYFSHF-DLSPGSADVKR
Goldfish              SLSDKDKAVVKALWAKIGSRADEIGAEALGR--MLTVYPQTKTYFSHWSDLSPGSGPVKK
Common carp           SLSDKDKAAVKGLWAKISPKADDIGAEALGR--MLTVYPQTKTYFAHWADLSPGSGPVKK
Tuna                  TLSDKDKSTVKALWGKISKSADAIGADALGR--MLAVYPQTKTYFSHWPDMSPGSGPVKA
Salmon                SLTARDKSVVNAFWGKIKGKADVVGAEALGR--MLTAYPQTKTYFSHWADLSPGSAPVKK
Trout                 SLTAKDKSVVKAFWGKISGKADVVGAEALGRDKMLTAYPQTKTYFSHWADLSPGSGPVKK
Eel                   SLTAKDKSLITGFWQKISSKADDLGAEALSR--MIVVFPATKVYFSHWPDLGPGSPSVKK
Stingray              VLSSQNKKAIEELGNLIKANAEAWGADALAR--LFELHPQTKTYFSKFSGFEACNEQVKK
Lungfish              RFSQDDEVLIKEAWG-LLHQIPNAGGEALAR--MFSCYPGTKSYFPHFGDFSANNEKVKH
                       ::  ::  :               : * *   :  .  :: **.:  ..   .  :.
```

```
Iguana                HGKKVVDALTQAVNNLDDIPDALAKLADLHAEKLRVDPVNFGLLGHCILVTIAAHNHGPL
Monitor lizard        HGKKVANALNQAVAHLDDIKGTLSKLSELHAQQLRVDPVNFGFLRHCLEVSIAAHLHDHL
Whale                 HGKKVADALTKAVGHMDNLLDALSDLSDLHAHKLRVDPANFKLLSHCLLVTLALHLPAEF
Dolphin               HGKKVADALTKAVGHIDNLPDALSELSDLHAHKLRVDPVNFKLLSHCLLVTLALHLPADF
Seal                  HGKKVADALTTAVAHMDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEF
Walrus                HGKKVADALTTAVAHIDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEF
Dog                   HGKKVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEF
Fox                   HGKKVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPNEF
Giant panda           HGKKVADALTTAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPAEF
Sun bear              HGKKVADALTTAAGHLDDLPGALSALSDLHAHKLRVDPVNFKFLSHCLLVTLASHHPAEF
Cat                   HGQKVADALTQAVAHMDDLPTAMSALSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPAEF
Lynx                  HGQKVADALTQAVAHIDDLPNALSALSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPAEF
Leopard               HGQKVADALTKAVAHINDLPNALSDLSDLHAYKLRVDPVNFKFLSHCLLVTLACHHPEEF
Palm civet            HGKKVADALTLAVGHLEDLPNALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPAEF
Lemur                 HGKKVADALTNAVNHIDDMPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPAEF
Gorilla               HGKKVAKALTBAVZHLDDMPNALSALSBLHAHKLRVBPVBFKLLNHCLLVTLAABFPSZF
Chimpanzee            HGKKVAKALSBAVZHLDDMPNALSALSBLHAHKLRVBPVBFKLLNHCLLVTLAABFPSZF
Mandrill              HGKKVADALTLAVGHVDDMPQALSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEF
Baboon                HGKKVADALTLAVGHVDDMPQALSKLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEF
Green monkey          HGKKVADALTLAVGHVDDMPHALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEF
Yak                   HGAKVAAALTKAVGHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDF
Cow                   HGAKVAAALTKAVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDF
Goat                  HGEKVAAALTKAVGHLDDLPGTLSDLSDLHAHKLRVDPVNFKLLSHSLLVTLACHLPNDF
Hippopotamus          HGKKVADALTKAVGHLDDLPGALSDLSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPSDF
Pig                   HGQKVADALTKAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDF
Horse                 HGKKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDF
Zebra                 HGKKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDF
White rhinoceros      HGKKVGDALTQAVGHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLALHHPQDF
Indian rhinoceros     HGKKVGDALTQAVGHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLALHNPQDF
Tapir                 HGKKVGDALTQAVGHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLALHHPDDF
Camel                 HGKKVGDALTKAADHLDDLPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTVAAHHPGDF
Llama                 HGKKVGDALTKAADHLDDLPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTVAAHHPGDF
Mouse                 HGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADF
Rat                   HGKKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDF
Possum                QGQKIADAVGLAVAHLDDMPTALSSLSDLHAHELKVDPVNFKFLCHNVLVTMAAHLGKDF
Kangaroo              HGKKIADALGQAVEHIDDLPGTLSKLSDLHAHKLRVDPVNFKLLSHCLLVTFAAHLGDAF
Echidna               HGKRVADALTTAAGHFNDMDSALSALSDLHAHKLRVDPVNFKLLAHCFLVVLARHHPAEF
Platypus              HGKKVADALSTAAGHFDDMDSALSALSDLHAHKLRVDPVNFKLLAHCILVVLARHCPGEF
Alligator             HGKKVFSALHEAVNHIDDLPGALCRLSELHAHSLRVDPVNFKFLAHCVLVVFAIHHPSAL
Crocodile             HGKKVFAALHEAVNHIDDLPGALCRLSELHAHSLRVDPVNFKFLAQCVLVVVAIHHPGSL
Snake                 HGKKVIDALDNAVEGLDDAVATLSKLSDLHAQKLRVDPANFKILSQCLLSTLANHRNPEF
Duck                  HGKKVAAALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLAQCFQVVLAAHLGKDY
Goose                 HGKKVAAALGNAVKSLDNISQALSELSNLHAYNLRVDPANFKLLSQCFQVVLAVHLGKDY
Rhea                  HGKKVVNALSNAVKNLDNLSQALAELSNLHAYNLRVDPVNFKLLSQCFQVVLAVHLGKEY
Ostrich               HGKKVANALGNAVKSLDNLSQALSELSNLHAYNLRVDPVNFKLLSQCFQVVLAVHMGKDY
Chicken               HGKKVLGALGNAVKNVDNLSQAMAELSNLHAYNLRVDPVNFKLLSQCIQVVLAVHMGKDY
Bullfrog              HGGKIINALAGAANHLDDLAGNLSSLSDLHAYNLRVDPGNFPLLAHIIQVVLATHFPGDF
Newt                  HGGKVLSAIGEAAKHIDSMDQALSKLSDLHAYNLRVDPGNFQLLSHCIQAVLAAHFPADF
Goldfish              HGKTIMGAVGDAVSKIDDLVGALSALSELHAFKLRIDPANFKILAHNVIVVIGMLFPGDF
Common carp           HGKVIMGAVGDAVSKIDDLVGGLAALSELHAFKLRVDPANFKILAHNVIVVIGMLYPGDF
Tuna                  HGKKVMGGVALAVTKIDDLTTGLGDLSELHAFKMRVDPSNFKILSHCILVVVAKMFPKEF
Salmon                HGGVIMGAIGNAVGLMDDLVGGMSGLSDLHAFKLRVDPGNFKILSHNILVTLAIHFPADF
Trout                 HGGIIMGAIGKAVGLMDDLVGGMSALSDLHAFKLRVDPGNFKILSHNILVTLAIHFPSDF
Eel                   HGKVIMAAVGDAVGKMNDLVGALSALSDLHAFKMRIDPGNFKTLSHNILVACAVNFPVDF
Stingray              HGKRVMNALADATHHLDNLHLHLEDLARKHGENLLVDPHNFHLFADCIVVTLAVNLQA-F
Lungfish              HGKKVVDAIGQGVQHLHDLSSCLHTLSEKHARELMVDPCNFQYLIEAIMTTIAAHYGEKF
                      :*  :  .:  ..  ...    :  *:  *. .: : *  *  : . .    .
```

```
Iguana                  KADVALSMDKFLTKVAKTLVAHYR
Monitor lizard          KASVIVSLDKFLEEVCKDLVSKYR
Whale                   TPSVHASLDKFLASVSTVLTSKYR
Dolphin                 TPSVHASLDKFLASVSTVLTSKYR
Seal                    TPAVHASLDKFFSAVSTVLTSKYR
Walrus                  TPAVHASLDKFFSTVSTVLTSKYR
Dog                     TPAVHASLDKFFAAVSTVLTSKYR
Fox                     TPAVHASLDKFFTAVSTVLTSKYR
Giant panda             TPAVHASLDKFFSAVSTVLTSKYR
Sun bear                TPAVHASLDKFFSAVSTVLTSKYR
Cat                     TPAVHASLDKFFSAVSTVLTSKYR
Lynx                    TPAVHASLDKFFSAVSTVLTSKYR
Leopard                 TPAVHASLDKFFSAVSTVLTSKYR
Palm civet              TPAVHSALDKFFSAVSTVLTSKYR
Lemur                   TPAVHASLDKFFAAVSTVLTSKYR
Gorilla                 TPAVHASVDKFLASVSTVLTSKYR
Chimpanzee              TPAVHASVDKFLASVSTVLTSKYR
Mandrill                TPAVHASLDKFLASVSTVLTSKYR
Baboon                  TPAVHASLDKFLASVSTVLTSKYR
Green monkey            TPAVHASLDKFLASVSTVLTSKYR
Yak                     TPAVHASLDKFLANVSTVLTSKYR
Cow                     TPAVHASLDKFLANVSTVLTSKYR
Goat                    TPAVHASLDKFLANVSTVLTSKYR
Hippopotamus            TPAAHASLDKFLANVSTVLTSKYR
Pig                     NPSVHASLDKFLANVSTVLTSKYR
Horse                   TPAVHASLDKFLSSVSTVLTSKYR
Zebra                   TPAVHASLDKFLSTVSTVLTSKYR
White rhinoceros        TPAVHASLDKFLSNVSTVLTSKYR
Indian rhinoceros       TPAVHASLDKFLSNVSTVLTSKYR
Tapir                   TPAIHASLDKFLSNVSTVLTSKYR
Camel                   TPSVHASLDKFLANVSTVLTSKYR
Llama                   TPAVDASLDKFLANVSTVLTSKYR
Mouse                   TPAVHASLDKFLASVSTVLTSKYR
Rat                     TPAMHASLDKFLASVSTVLTSKYR
Possum                  TPEIHASMDKFLASVSTVLTSKYR
Kangaroo                TPEVHASLDKFLAAVSTVLTSKYR
Echidna                 TPSAHAAMDKFLSRVATVLTSKYR
Platypus                TPSAHAAMDKFLSKVATVLTSKYR
Alligator               SPEIHASLDKFLCAVSAVLTSKYR
Crocodile               TPEVHASLDKFLCAVSSVLTSKYR
Snake                   GPAVLASVDKFLCNVSEVLESKYR
Duck                    SPEMHAAFDKFMSAVAAVLAEKYR
Goose                   TPEMHAAFDKFLSAVAAVLAEKYR
Rhea                    TPEVHAAYDKFLSAVASVLAEKYR
Ostrich                 TPEVHAAYDKFLTAVAAVLAEKYR
Chicken                 TPEVHAAFDKFLSAVSAVLAEKYR
Bullfrog                TAEVQAAWDKFLALVSAVLTSKYR
Newt                    TPQCQAAWDKFLAAVSAVLTSKYR
Goldfish                TPEVHMSVDKFFQNLALALSEKYR
Common carp             PPEVHMSVDKFFQNLALALSEKYR
Tuna                    TPDAHVSLDKFLASVALALAERYR
Salmon                  TPEVHIAVDKFLAALSAALADKYR
Trout                   TPEVHIAVDKFLAAVSAALADKYR
Eel                     TAEVHVAMDKFLAALGAALSDKYR
Stingray                TPVTHCAVDKFLELVAYELSSCYR
Lungfish                TPEINCAAEKCLGQIVHVLISLYR
                         .    : :* :  :   *   **
```

Hemoglobin β

```
Eel                  VEWTEDERTAIKSKWLKINIEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNDKV
Salmon               VDWTDAERSAIVGLWGKISVDEIGPQALARLLIVSPWTQRHFSTFGNLSTPAAIMGNPAV
Goldfish             VEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNPKV
Common carp          VEWTDAERSAIIALWGKLNPDELGPEALARCLIVYPWTQRFFASYGNLSSPAAIMGNPKV
Tuna                 VEWTQQERSIIAGFIANLNYEDIGPKALARCLIVYPWTQRYFGAYGDLSTPDAIKGNAKI
Trout                VEWTDAEKSTISAVWGKVNIDEIGPLALARVLIVYPWTQRYFGSFGNVSTPAAIMGNPKV
Dog                  VHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKV
Fox                  VHLTAEEKSLVTGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMGNAKV
Giant panda          VHLTGEEKAAVTGLWSKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKV
Walrus               VHLTADEKAAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKV
Seal                 VHLTGEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSADAIMGNPKV
Sun bear             VHLTGEEKSLVTGLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSADAIMNNPKV
Chimpanzee           VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
Gorilla              VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
Green monkey         VHLTPEEKTAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKV
Mandrill             VHLTPEEKTAVTTLWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKV
Baboon               VHLTPEEKNAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKV
Lemur                TFLTPEENGHVTSLWGKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPKV
Cat                  GFLTAEEKGLVNGLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSSADAIMSNAKV
Lynx                 GFLTAEEKGLVNGLWGKVNVDEVGGEALGRLLVVYPWTQRFFQSFGDLSSADAIMGNSKV
Leopard              SFLSAEEKNLVSGLWGKVNVDEVGGEALGRLLVVYPWTQRFFQSFGDLSSADAIMSNAKV
Palm civet           GFLTAEEKGLVNGLWGKVNVDEVGGEALGRLLVVYPWTQRFFQSFGDLSSADAIMHNSKV
Rat                  VHLTDAEKAAVNGLWGKVNPDDVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKV
Pig                  VHLSAEEKEAVLGLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSNADAVMGNPKV
Hippopotamus         VHLTAEEKDAVLGLWGKVNVQEVGGEALGRLLVVYPWTQRFFESFGDLSSADAVMNNPKV
Yak                  -MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSSADAVMNNPKV
Goat                 -MLTAEEKAAVTGFWGKVKVDEVGAEALGRLLVVYPWTQRFFEHFGDLSSADAVMNNAKV
Cow                  -MLSAEEKAAVTSLFAKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSSADAILGNPKV
Camel                VHLSGDEKNAVHGLWSKVKVDEVGGEALGRLLVVYPWTRRFFESFGDLSTADAVMNNPKV
Llama                VNLSGDEKNAVHGLWSKVKVDEVGGEALGRLLVVYPWTRRFFESFGDLSTADAVMNNPKV
Tapir                VELTGEEKAAVLALWDKVDEDKVGGEALGRLLVVYPWTQRFFDSFGDLSTAAAVMGNPKV
White rhinoceros     VELTAEEKAAVLALWDKVKEDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPAAVMGNAKV
Indian rhinoceros    VDLTAEEKAAVLALWGKVNEDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPAAVLGNAKV
Zebra                VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPAAVMGNPKV
Horse                VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKV
Whale                VHLTGEEKSGLTALWAKVNVEEIGGEALGRLLVVYPWTQRFFEHFGDLSTADAVMKNPKV
Dolphin              VHLTGEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMKNPNV
Platypus             VHLSGGEKSAVTNLWGKVNINELGGEALGRLLVVYPWTQRFFEAFGDLSSAGAVMGNPKV
Echidna              VHLSGSEKTAVTNLWGHVNVNELGGEALGRLLVVYPWTQRFFESFGDLSSADAVMGNAKV
Kangaroo             VHLTAEEKNAITSLWGKVAIEQTGGEALGRLLIVYPWTSRFFDHFGDLSNAKAVMGNPKV
Possum               VHLTSEEKNCITTIWSKVQVDQTGGEALGRMLVVYPWTTRFFGSFGDLSSPGAVMSNSKV
Iguana               VHWTAEEKQLITQVWGKIDVAQIGGETLACLLVVYPWTQRFFPDFGNLSNAAAICGNAKV
Monitor lizard       VHWTAEEKQLICSLWGKIDVGLIGGETLAGLLVIYPWTQRQFSHFGNLSSPTAIAGNPRV
Snake                VHWSAEEKQLITSLWAKVDVPEVGAATLGKMMVMYPWTQRFFAHFGNLSGPSALCGNPQV
Rhea                 VQWTAEEKQLITGLWGKVNVADCGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMV
Goose                VHWTAEEKQLITGLWGKVNVADCGAEALARLLIVYPWTQRFFSSFGNLSSPTAILGNPMV
Chicken              VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMV
Ostrich              VQWSAEEKQLISGLWGKVNVADCGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMV
Duck                 VHWTAEEKQLITGLWGKVNVADCGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMV
Mouse                VHFTAEEKAAITSIWDKVDLEKVGGETLGRLLIVYPWTQRFFDKFGNLSSAQAIMGNPRI
Bullfrog             ------GGSDVSAFLAKVDKRAVGGEALARLLIVYPWTQRYFSTFGNLGSADAISHNSKV
Alligator            ASFDAHERKFIVDLWAKVDVAQCGADALSRMLIVYPWKRRYFEHFGKMCNAHDILHNSKV
Crocodile            ASFDPHEKQLIGDLWHKVDVAHCGGEALSRMLIVYPWKRRYFENFGDISNAQAIMHNEKV
Newt                 -TFTNDESQHIHDVCGKIPVDQVGAEALGRLILVNPWTRRYFKSFGDLSSAEAIQHNPKV
Lungfish             VHWEDAEKQYIVSVFSKIDVDHVGANTLERVLIVFPWTKRYFNSFGDLSSPGAIKHNNKV
Stingray             VKLSEDQEHYIKGVWKDVDHKQITAKALERVFVVYPWTTRLFSKLQGLFSANDIG----V
                                  :     .:        ::   ::: **. * *     :  .  :     :
```

```
Eel                  AKHGTTVMGGLDRAIQNMDDIKNAYRQLSVMHSEKLHVDPDNFRLLAEHITLCMAAKFGP
Salmon               AKHGKTVMHGLDRAVQNLDDIKNAYTALSVMHSEKLHVDPDNFRLLADCITVCVAAKLGP
Goldfish             AAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFGP
Common carp          AAHGRTVEGGLMRAIKDMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFGP
Tuna                 AAHGVKVLHGLDRAVKNMDNINEAYSELSVLHSDKLHVDPDNFRILGDCLTVVIAANLG-
Trout                AAHGKVVCGALDKAVKNMGNILATYKSLSETHANKLFVDPDNFRVLADVLTIVIAAKFG-
Dog                  KAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Fox                  KAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Giant panda          KAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Walrus               KAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Seal                 KAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Sun bear             KAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Chimpanzee           KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-
Gorilla              KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Green monkey         KAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Mandrill             KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Baboon               KAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Lemur                KAHGKKVLSAFSEGLHHLDNLKGTFAQLSELHCVALHVDPENFKLLGNVLVIVLAHHFG-
Cat                  KAHGKKVLNSFSDGLKNIDDLKGAFAKLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-
Lynx                 KAHGKKVLNSFSDGLKNIDDLKGAFAKLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-
Leopard              KAHGKKVLNSFSDGLKNIDDLKGAFAKLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-
Palm civet           KAHGKKVLNSFSDGLKHVDDLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG-
Rat                  KAHGKKVINAFNDGLKHLDNLKGTFAHLSELHCDKLHVDPENFRLLGNMIVIVLGHHLG-
Pig                  KAHGKKVLQSFSDGLKHLDNLKGTFAKLSELHCDQLHVDPENFRLLGNVIVVVLARRLG-
Hippopotamus         KAHGKKVLDSFADGLKHLDNLKGTFAALSELHCDQLHVDPENFRLLGNELVVVLARTFG-
Yak                  KAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARHFG-
Goat                 KAHGKKVLDSFSNGMKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVVVLARHHG-
Cow                  KAHGKKVLDSFCEGLKQLDDLKGAFASLSELHCDKLHVDPENFRLLGNVLVVVLARRFG-
Camel                KAHGSKVLNSFGDGLNHLDNLKGTYAKLSELHCDKLHVDPENFRLLGNVLVVVLARHFG-
Llama                KAHGSKVLNSFGDGLSHLDNLKGTYAKLSELHCDKLHVDPENFRLLGNVLVVVLARHFG-
Tapir                KAHGKKVLHSFGDGVHHLDDLKVTFAQLSELHCDKLHVDPENFRLLGNVLVVVLAQQFG-
White rhinoceros     KAHGKKVLHSFGDGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLAKHFG-
Indian rhinoceros    KAHGKKVLHSFGDGVHNLDNLKGTYAALSELHCDKLHVDPENFRLLGNVLVVVLAQHFG-
Zebra                KAHGKKVLHSFGEGVHHLDNLKGTFAQLSELHCDKLHVDPENFRLLGNVLVVVLARHFG-
Horse                KAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHFG-
Whale                KKHGQKVLASFGEGLKHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVVVLARHFG-
Dolphin              KKHGQKVLASFGEGLKHLDDLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHFG-
Platypus             KAHGAKVLTSFGDALKNLDDLKGTFAKLSELHCDKLHVDPENFNRLGNVLIVVLARHFS-
Echidna              KAHGAKVLTSFGDALKNLDNLKGTFAKLSELHCDKLHVDPENFNRLGNVLVVVLARHFS-
Kangaroo             LAHGAKVLVAFGDAIKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNIIVICLAEHFG-
Possum               QAHGAKVLTSFGEAVKHLDNLKGTYAKLSELHCDKLHVDPENFKMLGNIIVICLAEHFG-
Iguana               KAHGKKVLTSFGDAVKNLDNIKDTFAKLSELHCDKLHVDPVNFRLLGNVMITRLAAHFG-
Monitor lizard       KAHGKKVLTSFGDAIKNLDNIKDTFAKLSELHCDKLHVDPTNFKLLGNVLVIVLADHHG-
Snake                RAHGKKVLTSFGEALKHLDNVKETFAKLSELHFDKLHVDPENFKLLGNVLIIVLAGHHG-
Rhea                 RAHGKKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAHFA-
Goose                RAHGKKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAHFA-
Chicken              RAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS-
Ostrich              RAHGKKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAHFT-
Duck                 RAHGKKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHVDPENFRLLGDILIIVLAAHFP-
Mouse                KAHGKKVLTSLGLAVKNMDNLKETFAHLSELHCDKLHVDPENFKLLGNMLVIVLSSYFG-
Bullfrog             LAHGQRVLDSIEEGLKHPZBLKAYYAKLSERHSGELHVDPANFYRLGNVLITVMARHFH-
Alligator            QEHGKKVLASFGEAVKHLDNIKGHFANLSKLHCEKFHVDPENFKLLGDIIIIVLAAHHP-
Crocodile            QAHGKKVLASFGEAVCHLDGIRAHFANLSKLHCEKLHVDPENFKLLGDIIIIVLAAHYP-
Newt                 ASHGAKVMHSIAEAVKHLDDLKAYYADLSTIHCKKLYVDPANFKLFGGIVSIVTGMHLG-
Lungfish             SAHGRKVLAAIIECTRHFGNIKGHLANLSHLHSEKLHVDPHNFRVLGQCLRIELAAALGF
Stingray             QQHADKVQRALGEAIDDLKKVEINFQNLSGKH-QEIGVDTQNFKLLGQTFMVELALHYK-
                       *.  *  .:     .   :      **  *   : **. **  :.  .    .
```

```
Eel                     TEFTADVQEAWQKFLMAVTSALARQYH
Salmon                  TVFSADIQEAFQKFLAVVVSALGRQYH
Goldfish                SGFNADVQEAWQKFLSVVVSALCRQYH
Common carp             SGFSPNVQEAWQKFLSVVVNALKRQYH
Tuna                    DAFTVETQCAFQKFLAVVVFALGRKYH
Trout                   ASFTPEIQATWQKFMKVVVAAMGSRYF
Dog                     KEFTPQVQAAYQKVVAGVANALAHKYH
Fox                     KEFTPQVQAAYQKVVAGVANALAHKYH
Giant panda             KEFTPQVQAAYQKVVAGVANALAHKYH
Walrus                  KEFTPQVQAAYQKVVAGVANALAHKYH
Seal                    KEFTPQVQAAYQKVVAGVANALAHKYH
Sun bear                KEFTPQVQAAYQKVVAGVANALAHKYH
Chimpanzee              KEFTPPVQAAYQKVVAGVANALAHKYH
Gorilla                 KEFTPPVQAAYQKVVAGVANALAHKYH
Green monkey            KEFTPQVQAAYQKVVAGVANALAHKYH
Mandrill                KEFTPQVQAAYQKVVAGVANALAHKYH
Baboon                  KEFTPQVQAAYQKVVAGVANALAHKYH
Lemur                   NDFSPQTQAAFQKVVTGVANALAHKYH
Cat                     HDFNPQVQAAFQKVVAGVANALAHKYH
Lynx                    HEFNPQVQAAFQKVVAGVANALAHKYH
Leopard                 HEFNPQVQAAFQKVVAGVASALAHRYH
Palm civet              KEFTPQVQAAYQKVVAGVASALAHRYH
Rat                     KEFTPCAQAAFQKVVAGVASALAHKYH
Pig                     HDFNPNVQAAFQKVVAGVANALAHKYH
Hippopotamus            KEFTPELQAAYQKVVAGVANALAHRYH
Yak                     KEFTPVLQADFQKVVVGVANALAHRYH
Goat                    SEFTPLLQAEFQKVVAGVANALAHRYH
Cow                     SEFSPELQASFQKVVTGVANALAHRYH
Camel                   KEFTPDLQAAYQKVVAGVANALAHRYH
Llama                   KEFTPDLQAAYQKVVAGVANALAHRYH
Tapir                   KAFTPELQAAYQKVVAGVANALAHKYH
White rhinoceros        KQFTPELQAAYQKVVAGVANALAHKYH
Indian rhinoceros       QEFTPELQAAYQKVVAGVANALAHKYH
Zebra                   KDFTPELQASYQKVVAGVANALAHKYH
Horse                   KDFTPELQASYQKVVAGVANALAHKYH
Whale                   KEFTPELQTAYQKVVAGVANALAHKYH
Dolphin                 KEFTPELQSAYQKVVAGVATALAHKYH
Platypus                KDFSPEVQAAWQKLVSGVAHALGHKYH
Echidna                 KEFTPEAQAAWQKLVSGVSHALAHKYH
Kangaroo                KEFTIDTQVAWQKLVAGVANALAHKYH
Possum                  KDFTPECQVAWQKLVAGVAHALAHKYH
Iguana                  KDFTPACHAAFQKLTGAVAHALARRYH
Monitor lizard          KEFTPAHHAAYQKLVNVVSHSLARRYH
Snake                   KEFTPSTHASFQKLVNVVAHALARRYH
Rhea                    KDFTPECQAAWQKLVRVVAHALARKYH
Goose                   KDFTPDCQAAWQKLVRVVAHALARKYH
Chicken                 KDFTPECQAAWQKLVRVVAHALARKYH
Ostrich                 KEFTPECQAAWQKLVRVVAHALARKYH
Duck                    KEFTPECQAAWQKLVRVVAHALARKYH
Mouse                   KEFTAEAQAAWQKLVVGVATALSHKYH
Bullfrog                EEFTPELQCALHSSFCAVGEALAKGYH
Alligator               EDFSVECHAAFQKLVRQVAAALAAEYH
Crocodile               KDFGLECHAAYQKLVRQVAAALAAEYH
Newt                    TDYTAQKQAAFEKFLHHVEAALATGYH
Lungfish                KEFTPERNAYFQKFMDVISHSLGREYH
Stingray                KTFRPKEHAAAYKFFRLVAEALSSNYH
                          :    :    .    :  ::   *.
```

Appendix B. Values

Cytochrome b

Note: 114 constant characters and 59 characters where $RI = \infty$ not shown.

Site	PDB Pos	Min Changes	Max Changes	True Tree Changes	Bad Tree Changes	Good RI	Bad RI	RI Difference
5	2	8	18	14	12	0.4	0.6	-0.2
6	3	10	19	13	13	0.667	0.667	0
7	4	5	15	8	8	0.7	0.7	0
8	5	3	18	11	12	0.467	0.4	0.067
11	8	6	23	15	16	0.471	0.412	0.059
12	9	1	4	2	2	0.667	0.667	0
14	11	1	4	3	3	0.333	0.333	0
15	12	6	33	17	18	0.593	0.556	0.037
16	13	2	4	3	3	0.5	0.5	0
17	14	3	7	6	6	0.25	0.25	0
18	15	6	21	14	13	0.467	0.533	-0.066
19	16	1	2	1	1	1	1	0
20	17	5	29	12	13	0.708	0.667	0.041
21	18	3	19	6	6	0.812	0.812	0
22	19	2	19	7	7	0.706	0.706	0
23	20	1	13	6	7	0.583	0.5	0.083
27	24	4	27	18	20	0.391	0.304	0.087
33	30	5	26	15	14	0.524	0.571	-0.047
42	39	2	3	3	3	0	0	0
43	40	6	26	15	16	0.55	0.5	0.05
46	43	6	13	11	11	0.286	0.286	0
47	44	6	27	17	16	0.476	0.524	-0.048
50	47	2	11	10	10	0.111	0.111	0
53	50	3	4	3	3	1	1	0
54	51	2	4	2	2	1	1	0
60	57	3	6	6	6	0	0	0
61	58	2	26	10	10	0.667	0.667	0
62	59	2	3	3	3	0	0	0
63	60	4	19	7	7	0.8	0.8	0
64	61	6	29	18	17	0.478	0.522	-0.044
65	62	3	16	9	9	0.538	0.538	0
70	67	2	4	4	4	0	0	0
71	68	7	28	12	13	0.762	0.714	0.048
73	70	2	5	3	4	0.667	0.333	0.334
74	71	4	9	6	6	0.6	0.6	0
76	73	2	3	2	2	1	1	0
78	75	3	8	3	3	1	1	0
79	76	2	4	4	4	0	0	0
82	79	5	20	8	8	0.8	0.8	0
83	80	2	9	5	4	0.571	0.714	-0.143
85	82	5	24	7	8	0.895	0.842	0.053
86	83	5	22	14	14	0.471	0.471	0

89	86	2	3	2	2	1	1	0
93	90	3	23	11	11	0.6	0.6	0
96	93	2	6	5	6	0.25	0	0.25
98	95	3	21	6	7	0.833	0.778	0.055
99	96	1	20	8	8	0.632	0.632	0
100	97	6	25	14	14	0.579	0.579	0
102	99	2	24	6	7	0.818	0.773	0.045
103	100	1	9	2	1	0.875	1	-0.125
106	103	4	12	11	11	0.125	0.125	0
111	108	3	8	7	7	0.2	0.2	0
112	109	7	21	10	9	0.786	0.857	-0.071
113	110	6	27	11	12	0.762	0.714	0.048
114	111	8	29	17	18	0.571	0.524	0.047
115	112	4	5	5	5	0	0	0
116	113	1	2	2	2	0	0	0
118	115	4	5	4	4	1	1	0
119	116	3	10	6	7	0.571	0.429	0.142
121	118	1	14	5	7	0.692	0.538	0.154
122	119	3	14	11	11	0.273	0.273	0
123	120	2	3	2	2	1	1	0
124	121	1	4	3	2	0.333	0.667	-0.334
125	122	3	22	11	11	0.579	0.579	0
126	123	4	27	8	8	0.826	0.826	0
127	124	3	28	12	13	0.64	0.6	0.04
129	126	4	10	5	6	0.833	0.667	0.166
133	130	2	23	8	8	0.714	0.714	0
138	135	1	2	1	1	1	1	0
146	143	1	2	1	1	1	1	0
154	151	3	7	4	4	0.75	0.75	0
157	154	5	16	10	10	0.545	0.545	0
160	157	3	15	11	10	0.333	0.417	-0.084
162	159	8	21	13	14	0.615	0.538	0.077
163	160	8	28	16	15	0.6	0.65	-0.05
164	161	3	4	4	4	0	0	0
166	163	3	24	7	7	0.81	0.81	0
168	165	4	11	7	7	0.571	0.571	0
172	169	2	4	3	3	0.5	0.5	0
173	170	1	3	2	2	0.5	0.5	0
174	171	1	2	2	2	0	0	0
176	173	2	23	5	5	0.857	0.857	0
177	174	1	6	4	4	0.4	0.4	0
184	181	3	10	6	7	0.571	0.429	0.142
185	182	1	6	3	3	0.6	0.6	0
188	185	2	19	5	4	0.824	0.882	-0.058
189	186	2	8	6	6	0.333	0.333	0
192	189	6	20	14	14	0.429	0.429	0
193	190	3	7	5	5	0.5	0.5	0
194	191	5	26	21	21	0.238	0.238	0
195	192	3	9	7	7	0.333	0.333	0
196	193	4	14	9	10	0.5	0.4	0.1
197	194	6	28	13	12	0.682	0.727	-0.045
198	195	6	32	24	24	0.308	0.308	0
199	196	5	23	13	14	0.556	0.5	0.056

201	198	2	4	4	4	0	0	0
202	199	4	10	6	6	0.667	0.667	0
206	203	2	6	2	2	1	1	0
207	204	4	8	6	6	0.5	0.5	0
209	206	1	4	3	3	0.333	0.333	0
210	207	3	4	4	4	0	0	0
213	210	8	31	17	18	0.609	0.565	0.044
215	212	2	18	8	9	0.625	0.562	0.063
216	213	8	25	13	14	0.706	0.647	0.059
217	214	2	5	4	5	0.333	0	0.333
218	215	3	18	13	11	0.333	0.467	-0.134
219	216	9	36	22	21	0.519	0.556	-0.037
222	219	2	6	5	5	0.25	0.25	0
223	220	4	12	6	6	0.75	0.75	0
226	223	2	3	2	2	1	1	0
228	225	2	21	5	5	0.842	0.842	0
229	226	2	18	6	8	0.75	0.625	0.125
230	227	6	25	13	13	0.632	0.632	0
233	230	4	22	9	10	0.722	0.667	0.055
234	231	2	7	4	5	0.6	0.4	0.2
236	233	4	26	12	13	0.636	0.591	0.045
237	234	7	27	17	19	0.5	0.4	0.1
238	235	6	21	14	13	0.467	0.533	-0.066
239	236	5	21	13	14	0.5	0.438	0.062
240	237	6	28	19	18	0.409	0.455	-0.046
241	238	7	21	17	16	0.286	0.357	-0.071
242	239	9	35	25	25	0.385	0.385	0
244	241	7	30	13	14	0.739	0.696	0.043
245	242	7	31	21	21	0.417	0.417	0
246	243	2	3	3	3	0	0	0
247	244	4	26	9	11	0.773	0.682	0.091
248	245	3	4	4	4	0	0	0
249	246	4	7	6	6	0.333	0.333	0
250	247	5	21	15	14	0.375	0.438	-0.063
252	249	5	23	5	5	1	1	0
253	250	3	7	5	5	0.5	0.5	0
255	252	4	7	6	6	0.333	0.333	0
257	254	2	4	2	2	1	1	0
258	255	3	18	8	8	0.667	0.667	0
259	256	1	2	1	1	1	1	0
260	257	1	22	3	4	0.905	0.857	0.048
261	258	3	15	5	6	0.833	0.75	0.083
264	261	1	2	2	2	0	0	0
267	264	5	30	15	15	0.6	0.6	0
270	267	5	9	8	7	0.25	0.5	-0.25
273	270	1	6	2	3	0.8	0.6	0.2
288	285	2	5	3	3	0.667	0.667	0
296	293	3	7	4	5	0.75	0.5	0.25
299	296	5	29	17	17	0.5	0.5	0
300	297	4	27	14	13	0.565	0.609	-0.044
301	298	1	3	2	2	0.5	0.5	0
302	299	2	6	3	3	0.75	0.75	0
303	300	5	8	5	5	1	1	0

304	301	2	16	12	12	0.286	0.286	0
306	303	5	29	13	13	0.667	0.667	0
307	304	5	24	16	14	0.421	0.526	-0.105
308	305	3	25	15	15	0.455	0.455	0
310	307	7	29	21	19	0.364	0.455	-0.091
311	308	3	4	4	3	0	1	-1
312	309	3	4	4	4	0	0	0
313	310	7	12	10	9	0.4	0.6	-0.2
314	311	1	3	3	3	0	0	0
316	313	4	6	4	4	1	1	0
317	314	3	7	4	4	0.75	0.75	0
318	315	3	16	11	11	0.385	0.385	0
319	316	6	22	16	14	0.375	0.5	-0.125
320	317	6	25	12	11	0.684	0.737	-0.053
322	319	1	2	1	1	1	1	0
324	321	3	17	12	13	0.357	0.286	0.071
325	322	4	19	8	8	0.733	0.733	0
327	324	8	33	17	17	0.64	0.64	0
328	325	6	15	12	11	0.333	0.444	-0.111
329	326	2	5	3	3	0.667	0.667	0
331	328	8	32	23	21	0.375	0.458	-0.083
332	329	3	5	4	3	0.5	1	-0.5
333	330	4	14	13	11	0.1	0.3	-0.2
334	331	5	9	9	8	0	0.25	-0.25
335	332	4	18	11	13	0.5	0.357	0.143
336	333	6	16	12	11	0.4	0.5	-0.1
337	334	5	16	10	10	0.545	0.545	0
338	335	4	27	10	10	0.739	0.739	0
342	339	3	5	4	4	0.5	0.5	0
344	341	2	9	5	5	0.571	0.571	0
345	342	3	7	3	3	1	1	0
348	345	3	4	3	3	1	1	0
349	346	6	21	14	15	0.467	0.4	0.067
351	348	2	17	11	11	0.4	0.4	0
352	349	3	10	8	8	0.286	0.286	0
353	350	4	21	15	14	0.353	0.412	-0.059
354	351	4	11	8	8	0.429	0.429	0
357	354	6	27	15	15	0.571	0.571	0
358	355	1	2	1	1	1	1	0
359	356	1	2	1	1	1	1	0
360	357	6	23	15	15	0.471	0.471	0
361	358	7	22	19	19	0.2	0.2	0
364	361	6	32	22	23	0.385	0.346	0.039
365	362	4	26	15	15	0.5	0.5	0
366	363	4	24	7	7	0.85	0.85	0
367	364	3	8	6	7	0.4	0.2	0.2
368	365	4	26	17	16	0.409	0.455	-0.046
369	366	6	14	10	10	0.5	0.5	0
370	367	5	19	13	13	0.429	0.429	0
372	369	6	25	19	19	0.316	0.316	0
373	370	7	25	18	18	0.389	0.389	0
374	371	3	23	15	15	0.4	0.4	0
375	372	10	35	17	16	0.72	0.76	-0.04

376	373	7	28	14	14	0.667	0.667	0
377	374	1	2	1	1	1	1	0
378	375	3	6	4	4	0.667	0.667	0
379	376	4	15	8	9	0.636	0.545	0.091
380	377	4	25	15	15	0.476	0.476	0
381	378	3	6	4	4	0.667	0.667	0
382	379	8	18	12	12	0.6	0.6	0

Rhodopsin

Note: 175 constant characters and 43 characters where $RI = \infty$ not shown.

Site	PDB Pos	Min Changes	Max Changes	True Tree Changes	Bad Tree Changes	Good RI	Bad RI	RI Difference
7	7	4	10	6	7	0.667	0.5	0.167
8	8	3	7	4	4	0.75	0.75	0
11	11	1	5	5	5	0	0	0
13	13	1	12	5	4	0.636	0.727	-0.091
16	16	5	8	6	6	0.667	0.667	0
19	19	2	5	4	4	0.333	0.333	0
20	20	2	3	2	2	1	1	0
22	22	1	2	2	2	0	0	0
24	24	1	5	2	2	0.75	0.75	0
25	25	1	4	3	3	0.333	0.333	0
26	26	3	5	3	4	1	0.5	0.5
32	32	2	4	3	2	0.5	1	-0.5
33	33	4	6	5	5	0.5	0.5	0
36	36	3	12	6	6	0.667	0.667	0
37	37	1	10	4	4	0.667	0.667	0
38	38	2	5	3	3	0.667	0.667	0
39	39	4	17	7	7	0.769	0.769	0
46	46	2	12	3	4	0.9	0.8	0.1
49	49	3	16	5	5	0.846	0.846	0
50	50	3	9	4	5	0.833	0.667	0.166
52	52	1	2	2	2	0	0	0
54	54	1	9	5	6	0.5	0.375	0.125
57	57	1	2	1	1	1	1	0
60	60	1	5	3	3	0.5	0.5	0
63	63	1	12	3	3	0.818	0.818	0
64	64	1	4	1	1	1	1	0
70	70	2	4	2	2	1	1	0
81	81	3	6	3	3	1	1	0
82	82	1	3	2	2	0.5	0.5	0
83	83	1	9	5	4	0.5	0.625	-0.125
84	84	1	3	1	1	1	1	0
88	88	1	7	5	4	0.333	0.5	-0.167
89	89	3	5	3	3	1	1	0
92	92	1	2	2	2	0	0	0
93	93	4	6	5	5	0.5	0.5	0
95	95	2	13	5	5	0.727	0.727	0
96	96	1	3	1	1	1	1	0
97	97	1	3	3	3	0	0	0
99	99	1	11	4	4	0.7	0.7	0
100	100	1	10	4	4	0.667	0.667	0
101	101	1	2	1	1	1	1	0
104	104	1	3	3	3	0	0	0
105	105	1	2	2	2	0	0	0
107	107	6	11	6	6	1	1	0
108	108	5	7	6	6	0.5	0.5	0
109	109	1	2	1	1	1	1	0
111	111	2	6	3	3	0.75	0.75	0
112	112	4	13	11	11	0.222	0.222	0
123	123	3	8	7	6	0.2	0.4	-0.2
124	124	2	10	7	6	0.375	0.5	-0.125
127	127	1	3	3	3	0	0	0
133	133	2	5	4	4	0.333	0.333	0

136	136	1	4	1	1	1	1	0
137	137	3	8	6	6	0.4	0.4	0
139	139	1	4	3	2	0.333	0.667	-0.334
143	143	2	3	2	2	1	1	0
144	144	2	3	2	2	1	1	0
149	149	1	2	2	2	0	0	0
150	150	2	5	2	3	1	0.667	0.333
151	151	2	6	3	2	0.75	1	-0.25
155	155	2	3	3	3	0	0	0
157	157	1	4	3	2	0.333	0.667	-0.334
158	158	4	12	11	9	0.125	0.375	-0.25
159	159	3	7	5	4	0.5	0.75	-0.25
162	162	2	12	7	7	0.5	0.5	0
165	165	3	8	5	5	0.6	0.6	0
166	166	2	7	6	6	0.2	0.2	0
169	169	3	8	6	6	0.4	0.4	0
173	173	3	7	6	6	0.25	0.25	0
183	183	1	3	3	3	0	0	0
189	189	3	13	7	7	0.6	0.6	0
194	194	4	10	6	6	0.667	0.667	0
195	195	6	10	6	6	1	1	0
196	196	2	6	4	5	0.5	0.25	0.25
197	197	3	6	3	3	1	1	0
198	198	4	11	8	8	0.429	0.429	0
209	209	2	6	6	5	0	0.25	-0.25
213	213	7	16	13	14	0.333	0.222	0.111
214	214	2	3	3	3	0	0	0
216	216	2	6	3	3	0.75	0.75	0
217	217	5	11	8	7	0.5	0.667	-0.167
218	218	1	11	7	6	0.4	0.5	-0.1
225	225	2	12	2	2	1	1	0
227	227	1	3	2	2	0.5	0.5	0
228	228	1	11	1	1	1	1	0
232	232	2	3	3	3	0	0	0
236		1	3	2	2	0.5	0.5	0
241	241	1	7	2	1	0.833	1	-0.167
242	242	1	3	2	2	0.5	0.5	0
245	245	1	7	2	1	0.833	1	-0.167
248	248	1	8	3	2	0.714	0.857	-0.143
255	255	1	8	6	4	0.286	0.571	-0.285
256	256	1	4	3	3	0.333	0.333	0
259	259	1	8	7	6	0.143	0.286	-0.143
260	260	3	13	8	7	0.5	0.6	-0.1
263	263	1	5	3	3	0.5	0.5	0
266	266	3	12	6	7	0.667	0.556	0.111
270	270	2	4	3	3	0.5	0.5	0
273	273	3	6	3	3	1	1	0
274	274	2	3	3	3	0	0	0
277	277	4	6	6	6	0	0	0
278	278	1	8	5	4	0.429	0.571	-0.142
282	282	2	13	5	6	0.727	0.636	0.091
285	285	1	2	1	1	1	1	0
286	286	4	9	7	7	0.4	0.4	0
290	290	2	15	7	6	0.615	0.692	-0.077
292	292	1	3	2	2	0.5	0.5	0
297	297	2	6	6	6	0	0	0
298	298	1	7	6	5	0.167	0.333	-0.166
299	299	1	12	5	6	0.636	0.545	0.091
300	300	2	9	4	5	0.714	0.571	0.143
304	304	3	8	5	5	0.6	0.6	0

308	308	3	13	5	5	0.8	0.8	0
309	309	1	5	3	3	0.5	0.5	0
315	315	2	5	2	2	1	1	0
318	318	2	9	3	3	0.857	0.857	0
321	321	1	4	4	4	0	0	0
325	325	1	2	1	1	1	1	0
328		1	10	3	1	0.778	1	-0.222
329		1	7	2	1	0.833	1	-0.167
330		1	8	3	2	0.714	0.857	-0.143
331		1	11	5	5	0.6	0.6	0
332		1	5	4	4	0.25	0.25	0
333		4	14	5	6	0.9	0.8	0.1
334	334	3	11	5	5	0.75	0.75	0
335	335	2	16	5	4	0.786	0.857	-0.071
336	336	3	13	5	5	0.8	0.8	0
337		2	7	3	3	0.8	0.8	0
338		3	7	3	3	1	1	0
343	337	2	8	4	5	0.667	0.5	0.167
345	339	2	12	2	2	1	1	0
346	340	1	11	1	1	1	1	0
347	341	1	11	1	1	1	1	0
348	342	1	9	4	2	0.625	0.875	-0.25
350	344	2	4	2	2	1	1	0
352	346	1	13	3	3	0.833	0.833	0

Myoglobin

Note: 31 constant characters and 27 characters where $RI = \infty$ not shown.

Site	PDB Pos	Min Changes	Max Changes	True Tree Changes	Bad Tree Changes	Good RI	Bad RI	RI Difference
5	5	5	10	7	7	0.6	0.6	0
6	6	1	2	1	1	1	1	0
8	8	5	8	8	7	0	0.333	-0.333
9	9	5	9	7	6	0.5	0.75	-0.25
12	12	5	13	8	6	0.625	0.875	-0.25
13	13	5	16	8	8	0.727	0.727	0
15	15	2	4	3	3	0.5	0.5	0
19	19	4	12	10	9	0.25	0.375	-0.125
21	21	5	18	11	13	0.538	0.385	0.153
22	22	5	15	10	10	0.5	0.5	0
23	23	4	5	4	4	1	1	0
26	26	4	7	4	5	1	0.667	0.333
27	27	4	9	8	8	0.2	0.2	0
29	29	1	4	3	1	0.333	1	-0.667
30	30	3	5	3	3	1	1	0
34	34	3	11	8	5	0.375	0.75	-0.375
35	35	9	18	14	15	0.444	0.333	0.111
40	40	2	7	3	2	0.8	1	-0.2
41	41	2	8	5	5	0.5	0.5	0
42	42	2	8	2	2	1	1	0
44	44	3	6	4	3	0.667	1	-0.333
45	45	1	2	2	2	0	0	0
48	48	2	9	6	5	0.429	0.571	-0.142
49	49	2	4	3	3	0.5	0.5	0
50	50	2	3	2	2	1	1	0
51	51	2	15	8	6	0.538	0.692	-0.154
52	52	4	6	4	4	1	1	0
53	53	4	10	6	5	0.667	0.833	-0.166
54	54	3	6	6	6	0	0	0
55	55	2	4	2	3	1	0.5	0.5
56	56	2	5	4	4	0.333	0.333	0
57	57	4	16	7	6	0.75	0.833	-0.083
58	58	1	2	1	1	1	1	0
59	59	1	4	2	2	0.667	0.667	0
60	60	3	5	3	3	1	1	0
61	61	3	6	4	5	0.667	0.333	0.334
63	63	2	3	2	2	1	1	0
66	66	6	19	13	11	0.462	0.615	-0.153
67	67	1	2	2	2	0	0	0
70	70	2	3	2	2	1	1	0
71	71	2	3	2	2	1	1	0
74	74	8	16	10	11	0.75	0.625	0.125
75	75	1	2	1	1	1	1	0
78	78	3	7	4	4	0.75	0.75	0
80	80	1	2	1	1	1	1	0
81	81	5	14	11	8	0.333	0.667	-0.334

83	83	2	6	4	4	0.5	0.5	0
85	85	2	3	2	2	1	1	0
86	86	2	14	4	7	0.833	0.583	0.25
87	87	3	6	4	5	0.667	0.333	0.334
88	88	1	3	1	1	1	1	0
91	91	4	8	5	5	0.75	0.75	0
92	92	1	4	3	3	0.333	0.333	0
95	95	2	6	4	4	0.5	0.5	0
96	96	2	3	3	3	0	0	0
100	100	2	4	3	3	0.5	0.5	0
101	101	2	13	11	9	0.182	0.364	-0.182
102	102	2	4	2	2	1	1	0
103	103	2	6	3	3	0.75	0.75	0
104	104	1	2	1	1	1	1	0
106	106	2	3	2	2	1	1	0
108	108	2	4	3	3	0.5	0.5	0
109	109	1	9	6	4	0.375	0.625	-0.25
110	110	3	11	6	5	0.625	0.75	-0.125
111	111	1	2	1	1	1	1	0
112	112	3	8	4	3	0.8	1	-0.2
113	113	4	21	9	8	0.706	0.765	-0.059
115	115	2	11	4	4	0.778	0.778	0
116	116	3	18	8	7	0.667	0.733	-0.066
117	117	4	12	4	4	1	1	0
118	118	1	3	2	2	0.5	0.5	0
119	119	2	3	2	2	1	1	0
120	120	3	12	8	8	0.444	0.444	0
121	121	3	17	10	8	0.5	0.643	-0.143
122	122	2	8	7	7	0.167	0.167	0
123	123	1	2	1	1	1	1	0
124	124	2	3	2	2	1	1	0
125	125	1	3	2	2	0.5	0.5	0
127	127	3	8	4	4	0.8	0.8	0
128	128	1	2	1	1	1	1	0
129	129	4	19	11	10	0.533	0.6	-0.067
131	131	1	2	1	1	1	1	0
132	132	5	23	11	10	0.667	0.722	-0.055
133	133	2	4	4	3	0	0.5	-0.5
134	134	1	2	1	1	1	1	0
135	135	1	2	1	1	1	1	0
139	139	1	2	1	1	1	1	0
140	140	4	8	5	5	0.75	0.75	0
142	142	2	17	5	6	0.8	0.733	0.067
144	144	2	11	5	5	0.667	0.667	0
145	145	5	10	7	7	0.6	0.6	0
148	148	1	3	1	1	1	1	0
149	149	2	9	5	5	0.571	0.571	0
151	151	1	2	2	2	0	0	0
152	152	4	6	6	6	0	0	0

Hemoglobin α

Note: 17 constant characters and 16 characters where $RI = \infty$ not shown.

Site	PDB Pos	Min Changes	Max Changes	True Tree Changes	Bad Tree Changes	Good RI	Bad RI	RI Difference
1	1	4	14	6	6	0.8	0.8	0
3	3	1	12	6	4	0.545	0.727	-0.182
4	4	7	34	15	14	0.704	0.741	-0.037
5	5	8	31	17	18	0.609	0.565	0.044
6	6	2	5	3	3	0.667	0.667	0
8	8	7	40	21	20	0.576	0.606	-0.03
9	9	7	22	13	12	0.6	0.667	-0.067
10	10	2	15	10	8	0.385	0.538	-0.153
11	11	9	14	12	12	0.4	0.4	0
12	12	6	26	18	17	0.4	0.45	-0.05
13	13	7	34	16	16	0.667	0.667	0
14	14	4	6	5	5	0.5	0.5	0
15	15	7	27	16	16	0.55	0.55	0
16	16	3	6	5	4	0.333	0.667	-0.334
17	17	4	32	12	11	0.714	0.75	-0.036
18	18	7	19	13	13	0.5	0.5	0
19	19	8	28	20	20	0.4	0.4	0
20	20	6	16	13	13	0.3	0.3	0
21	21	7	19	13	14	0.5	0.417	0.083
22	22	5	28	14	14	0.609	0.609	0
23	23	4	19	15	15	0.267	0.267	0
24	24	7	20	11	12	0.692	0.615	0.077
25	25	2	3	3	3	0	0	0
26	26	6	17	12	12	0.455	0.455	0
27	27	1	3	3	3	0	0	0
30	30	7	20	11	11	0.692	0.692	0
34	32	2	18	6	6	0.75	0.75	0
35	33	2	6	2	2	1	1	0
36	34	10	32	22	22	0.455	0.455	0
37	35	7	29	15	14	0.636	0.682	-0.046
38	36	2	17	5	6	0.8	0.733	0.067
40	38	5	22	7	6	0.882	0.941	-0.059
43	41	3	4	3	3	1	1	0
46	44	2	14	8	8	0.5	0.5	0
48	46	4	9	4	4	1	1	0
49		3	5	5	5	0	0	0
50	47	2	3	3	3	0	0	0
51	48	3	7	6	5	0.25	0.5	-0.25
52	49	5	10	7	6	0.6	0.8	-0.2
53	50	5	26	11	10	0.714	0.762	-0.048
54	51	3	5	5	5	0	0	0
55	52	1	3	3	2	0	0.5	-0.5
56	53	5	17	10	9	0.583	0.667	-0.084
57	54	5	10	5	5	1	1	0
58	55	2	12	9	8	0.3	0.4	-0.1
59	56	3	12	6	7	0.667	0.556	0.111

60	57	6	27	14	14	0.619	0.619	0
63	60	4	12	7	7	0.625	0.625	0
64	61	4	7	6	5	0.333	0.667	-0.334
65	62	1	8	4	3	0.571	0.714	-0.143
66	63	6	23	11	11	0.706	0.706	0
67	64	5	22	12	12	0.588	0.588	0
69	66	2	9	5	5	0.571	0.571	0
70	67	6	25	13	12	0.632	0.684	-0.052
71	68	8	44	20	22	0.667	0.611	0.056
73	70	2	10	6	6	0.5	0.5	0
74	71	9	33	20	18	0.542	0.625	-0.083
75	72	5	13	7	7	0.75	0.75	0
76	73	4	25	16	16	0.429	0.429	0
77	74	3	6	6	6	0	0	0
78	75	2	9	4	4	0.714	0.714	0
79	76	3	14	7	8	0.636	0.545	0.091
80	77	8	20	11	10	0.75	0.833	-0.083
81	78	8	27	16	16	0.579	0.579	0
82	79	5	11	9	10	0.333	0.167	0.166
83	80	1	4	3	3	0.333	0.333	0
84	81	5	9	8	8	0.25	0.25	0
85	82	8	30	19	19	0.5	0.5	0
87	84	1	2	2	2	0	0	0
88	85	3	13	6	6	0.7	0.7	0
89	86	1	2	2	1	0	1	-1
92	89	5	29	10	9	0.792	0.833	-0.041
93	90	4	13	7	5	0.667	0.889	-0.222
94	91	1	2	2	2	0	0	0
96	93	1	2	2	2	0	0	0
99	96	5	13	8	9	0.625	0.5	0.125
102	99	4	6	6	5	0	0.5	-0.5
103	100	4	17	10	10	0.538	0.538	0
105	102	6	15	11	11	0.444	0.444	0
106	103	3	9	5	5	0.667	0.667	0
107	104	4	11	6	5	0.714	0.857	-0.143
108	105	3	21	8	7	0.722	0.778	-0.056
109	106	5	12	6	5	0.857	1	-0.143
110	107	3	5	4	4	0.5	0.5	0
111	108	3	16	6	5	0.769	0.846	-0.077
112	109	5	13	11	10	0.25	0.375	-0.125
113	110	1	2	1	1	1	1	0
114	111	10	40	20	21	0.667	0.633	0.034
115	112	3	7	5	5	0.5	0.5	0
116	113	7	34	18	16	0.593	0.667	-0.074
117	114	4	12	6	5	0.75	0.875	-0.125
118	115	10	38	27	24	0.393	0.5	-0.107
119	116	6	28	13	12	0.682	0.727	-0.045
120	117	2	9	3	3	0.857	0.857	0
121	118	5	7	6	6	0.5	0.5	0
122	119	1	4	4	3	0	0.333	-0.333
123	120	5	27	11	11	0.727	0.727	0
124	121	5	13	11	11	0.25	0.25	0
125	122	6	7	6	6	1	1	0

126	123	6	11	7	7	0.8	0.8	0
127	124	1	15	5	4	0.714	0.786	-0.072
128	125	6	21	12	12	0.6	0.6	0
132	129	2	14	4	3	0.833	0.917	-0.084
133	130	6	30	13	14	0.708	0.667	0.041
134	131	8	34	18	16	0.615	0.692	-0.077
135	132	2	5	4	3	0.333	0.667	-0.334
136	133	4	14	9	9	0.5	0.5	0
137	134	7	20	8	9	0.923	0.846	0.077
138	135	4	9	4	4	1	1	0
140	137	5	16	7	8	0.818	0.727	0.091
141	138	3	12	4	4	0.889	0.889	0

Hemoglobin β

Note: 15 constant characters and 24 characters where $RI = \infty$ not shown.

Site	PDB Pos	Min Changes	Max Changes	True Tree Changes	Bad Tree Changes	Good RI	Bad RI	RI Difference
1	1	4	7	4	4	1	1	0
2	2	9	26	13	14	0.765	0.706	0.059
3	3	2	19	5	4	0.824	0.882	-0.058
4	4	3	15	11	11	0.333	0.333	0
5	5	7	30	16	15	0.609	0.652	-0.043
6	6	6	17	11	10	0.545	0.636	-0.091
8	8	5	10	7	6	0.6	0.8	-0.2
9	9	9	41	21	21	0.625	0.625	0
10	10	9	26	14	14	0.706	0.706	0
11	11	2	23	5	3	0.857	0.952	-0.095
12	12	10	29	20	19	0.474	0.526	-0.052
13	13	6	33	21	20	0.444	0.481	-0.037
14	14	4	12	11	9	0.125	0.375	-0.25
15	15	4	5	5	5	0	0	0
16	16	6	19	16	16	0.231	0.231	0
18	18	2	9	7	5	0.286	0.571	-0.285
19	19	6	20	12	11	0.571	0.643	-0.072
20	20	7	16	11	11	0.556	0.556	0
21	21	8	24	14	15	0.625	0.562	0.063
22	22	6	19	15	14	0.308	0.385	-0.077
23	23	4	21	8	9	0.765	0.706	0.059
25	25	2	17	7	6	0.667	0.733	-0.066
26	26	6	9	8	7	0.333	0.667	-0.334
27	27	1	5	3	3	0.5	0.5	0
29	29	4	18	7	7	0.786	0.786	0
31	31	3	10	7	6	0.429	0.571	-0.142
33	33	2	20	7	6	0.722	0.778	-0.056
38	38	1	2	1	1	1	1	0
39	39	4	9	7	7	0.4	0.4	0
41	41	4	13	8	6	0.556	0.778	-0.222
43	43	8	38	18	19	0.667	0.633	0.034
44	44	6	16	15	15	0.1	0.1	0
45	45	2	3	3	3	0	0	0
47	47	3	17	8	7	0.643	0.714	-0.071
50	50	3	24	14	12	0.476	0.571	-0.095
51	51	1	28	12	9	0.593	0.704	-0.111
52	52	9	29	18	18	0.55	0.55	0
53	53	1	2	2	2	0	0	0
54	54	2	27	8	6	0.76	0.84	-0.08
55	55	7	16	11	11	0.556	0.556	0
56	56	4	20	14	13	0.375	0.438	-0.063
58	58	5	18	15	14	0.231	0.308	-0.077
59	59	5	10	6	6	0.8	0.8	0
60	60	1	2	2	2	0	0	0
61	61	5	20	9	9	0.733	0.733	0
62	62	4	7	6	5	0.333	0.667	-0.334

65	65	7	16	10	11	0.667	0.556	0.111
66	66	3	6	4	3	0.667	1	-0.333
68	68	5	8	7	6	0.333	0.667	-0.334
69	69	9	43	17	15	0.765	0.824	-0.059
70	70	2	16	6	5	0.714	0.786	-0.072
71	71	2	11	3	4	0.889	0.778	0.111
72	72	8	31	13	11	0.783	0.87	-0.087
73	73	5	24	13	12	0.579	0.632	-0.053
74	74	2	24	4	3	0.909	0.955	-0.046
75	75	4	28	14	11	0.583	0.708	-0.125
76	76	8	18	11	10	0.7	0.8	-0.1
77	77	3	29	13	11	0.615	0.692	-0.077
78	78	5	12	7	7	0.714	0.714	0
79	79	3	4	4	4	0	0	0
80	80	3	15	10	9	0.417	0.5	-0.083
81	81	2	18	4	4	0.875	0.875	0
83	83	6	20	12	10	0.571	0.714	-0.143
84	84	4	13	9	8	0.444	0.556	-0.112
85	85	2	13	6	6	0.636	0.636	0
86	86	5	6	6	6	0	0	0
87	87	9	35	20	20	0.577	0.577	0
90	90	5	10	6	5	0.8	1	-0.2
91	91	5	8	6	5	0.667	1	-0.333
93	93	3	9	4	4	0.833	0.833	0
94	94	6	12	8	7	0.667	0.833	-0.166
95	95	3	5	5	5	0	0	0
101	101	6	12	6	6	1	1	0
104	104	3	26	12	9	0.609	0.739	-0.13
105	105	4	7	6	6	0.333	0.333	0
107	107	1	5	2	2	0.75	0.75	0
108	108	4	16	5	6	0.917	0.833	0.084
109	109	6	20	9	10	0.786	0.714	0.072
110	110	4	13	10	7	0.333	0.667	-0.334
111	111	5	20	7	7	0.867	0.867	0
112	112	4	35	9	9	0.839	0.839	0
113	113	3	9	6	5	0.5	0.667	-0.167
114	114	5	8	6	6	0.667	0.667	0
115	115	2	3	3	3	0	0	0
116	116	10	39	13	14	0.897	0.862	0.035
117	117	7	12	8	8	0.8	0.8	0
118	118	3	12	11	11	0.111	0.111	0
119	119	6	11	8	8	0.6	0.6	0
121	120	8	17	13	13	0.444	0.444	0
122	121	7	22	18	19	0.267	0.2	0.067
124	123	4	13	11	11	0.222	0.222	0
125	124	4	9	9	8	0	0.2	-0.2
126	125	10	33	18	15	0.652	0.783	-0.131
127	126	9	36	16	17	0.741	0.704	0.037
128	127	2	7	4	4	0.6	0.6	0
129	128	5	10	7	6	0.6	0.8	-0.2
130	129	5	8	7	7	0.333	0.333	0
131	130	4	32	15	14	0.607	0.643	-0.036
134	133	3	25	4	3	0.955	1	-0.045

135	134	4	11	6	5	0.714	0.857	-0.143
136	135	11	24	15	15	0.692	0.692	0
137	136	5	20	7	7	0.867	0.867	0
139	138	5	11	7	7	0.667	0.667	0
140	139	6	27	15	14	0.571	0.619	-0.048
141	140	1	2	2	2	0	0	0
143	142	4	9	7	7	0.4	0.4	0
144	143	5	20	6	6	0.933	0.933	0
145	144	5	22	13	11	0.529	0.647	-0.118

Appendix C. Scripts

Example PAUP Script to Compute RI Values

The following script is a modified version of the PAUP scripts generated by RI Compare while working with the Rhodopsin dataset. The modifications were only cosmetic such as spacing and truncation of the alignment strings to save space on the page. These modifications will not affect the intent of the example.

```
#NEXUS
begin taxa;
       dimensions ntax=26;
       taxlabels
Alligator
Chicken
Toad
Frog
Salamander
Blackmouth_catshark
Spotted_dogfish
Little_skate
Goldfish
Common_carp
Guppy
Blind_cave_fish
Cow
Sheep
Whale
Dolphin
Pig
Dog
Seal
Mouse
Hamster
Rat
Rabbit
Green_anole
Japanese_lamprey
Sea_lamprey
;
end;
Begin trees;
tree Tree =
((((Alligator,Chicken),(((Toad,Frog),Salamander),(((Blackmouth_catshark,Spotted_dogfish),Little_s
kate),(((Goldfish,Common_carp),Guppy),Blind_cave_fish))),(((Cow,Sheep),((Whale,Dolphin),Pig),((Do
g,Seal),((Mouse,Hamster),Rat))),Rabbit)),Green_anole),(Japanese_lamprey,Sea_lamprey))
;
end;
begin characters;
       dimensions nchar=354;
       format missing=? gap=- datatype=protein;
       options gapmode=missing;
       matrix
Alligator             MNGTEGPDFYIPFSNKTGVVRSPFEYPQYYLAEPWKYSALAAYMFMLIILGFPINFLTLYVTVQHKKLRSP
Chicken               MNGTEGQDFYVPMSNKTGVVRSPFEYPQYYLAEPWKFSALAAYMFMLILLGFPVNFLTLYVTIQHKKLRTP
Toad                  MNGTEGPNFYIPMSNKTGVVRSPFEYPQYYLAEPWQYSILCAYMFLLILLGFPINFMTLYVTIQHKKLRTP
Frog                  MNGTEGPNFYVPMSNKTGIVRSPFEYPQYYLAEPWKYSVLAAYMFLLILLGLPINFMTLYVTIQHKKLRTP
```

```
Salamander              MNGTEGPNFYVPFSNKSGVVRSPFEYPQYYLAEPWQYSVLAAYMFLLILLGFPVNFLTLYVTIQHKKLRTP
Blackmouth_catshark     MNGTEGENFYVPMSNKTGVVRNPFEYPQYYLADHWMFAVLAAYMFFLIITGFPVNFLTLFVTIQNKKLRQP
Spotted_dogfish         MNGTEGENFYIPMSNKTGVVRSPFDYPQYYLAEPWKFSVLAAYMFFLIIAGFPVNFLTLYVTIQHKKLRQP
Little_skate            MNGTEGENFYVPMSNKTGVVRSPFDYPQYYLGEPWMFSALAAYMFFLILTGLPVNFLTLFVTIQHKKLRQP
Goldfish                MNGTEGDMFYVPMSNATGIVRSPYDYPQYYLVAPWAYACLAAYMFFLIITGFPVNFLTLYVTIEHKKLRTP
Common_carp             MNGTEGPMFYVPMSNATGVVKSPYDYPQYYLVAPWAYGCLAAYMFFLIITGFPINFLTLYVTIEHKKLRTP
Guppy                   MNGTEGPYFYVPMVNTTGIVRSPYEYPQYYLVSPAAYACLGAYMFFLILVGFPINFLTLYVTIEHKKLRTP
Blind_cave_fish         MNGTEGPYFYVPMSNATGVVRSPYEYPQYYLAPPWAYACLAAYMFFLILVGFPVNFLTLYVTIEHKKLRTP
Cow                     MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTP
Sheep                   MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Whale                   MNGTEGLNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSVLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Dolphin                 MNGTEGLNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSVLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Pig                     MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFMLIVLGFPINFLTLYVTVQHKKLRTP
Dog                     MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Seal                    MNGTEGPNFYVPFSNKTGVVRSPFEFPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Mouse                   MNGTEGPNFYVPFSNVTGVGRSPFEQPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Hamster                 MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Rat                     MNGTEGPNFYVPFSNITGVVRSPFEQPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Rabbit                  MNGTEGPDFYIPMSNQTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTP
Green_anole             MNGTEGQNFYVPMSNKTGVVRNPFEYPQYYLADPWQFSALAAYMFLLILLGFPINFLTLFVTIQHKKLRTP
Japanese_lamprey        MNGTEGDNFYVPFSNKTGLARSPYEYPQYYLAEPWKYSALAAYMFFLILVGFPVNFLTLFVTVQHKKLRTP
Sea_lamprey             MNGTEGENFYIPFSNKTGLARSPFEYPQYYLAEPWKYSVLAAYMFFLILVGFPVNFLTLFVTVQHKKLRTP
;
end;
begin paup;
        log file=ri_tmp.log replace=yes start;
        set criterion=parsimony;
        set taxlabels=full;
        describetrees / diag=yes plot=none chglist=no;
        log stop;
        quit;
end;
```

Appendix D. Residue Properties, Codes, and Colors

Aspartic acid (D, Asp) – Acidic, acyclic, charged, medium, negative, polar, and surface
Glutamic acid (E, Glu) – Acidic, acyclic, charged, large, negative, polar, and surface

Lysine (K, Lys) – Acyclic, basic, charged, large, polar, positive, and surface
Arginine (R, Arg) – Acyclic, aliphatic, buried, hydrophobic, neutral, and small

Phenylalanine (F, Phe) – Aromatic, buried, cyclic, hydrophobic, large, and neutral
Tyrosine (Y, Tyr) – Aromatic, cyclic, hydrophobic, large, neutral, and surface

Glycine (G, Gly) – Acyclic, aliphatic, hydrophobic, neutral, small, surface

Alanine (A, Ala) – Acyclic, aliphatic, buried, hydrophobic, neutral, and small

Histidine (H, His) – Aromatic, basic, charged, cyclic, large, neutral, polar, positive, and surface

Cystine (C, Cys) – Acyclic, buried, medium, neutral, and polar
Methionine (M, Met) – Acyclic, buried, hydrophobic, large, and neutral

Serine (S, Ser) – Acyclic, neutral, polar, and surface
Threonine (T, Thr) – Acyclic, medium, neutral, polar, and surface

Asparagine (N, Asn) – Acyclic, medium, neutral, polar, and surface
Glutamine (Q, Gln) – Acyclic, large, neutral, polar, and surface

Isoleucine (I, Ile) – Identical to Leucine
Leucine (L, Leu) – Acyclic, aliphatic, buried, hydrophobic, large, and neutral
Valine (V, Val) – Acyclic, aliphatic, buried, hydrophobic, medium, and neutral

Tryptophan (W, Trp) – Aromatic, buried, cyclic, hydrophobic, large, and neutral

Proline (P, Pro) – Cyclic, hydrophobic, medium, neutral, and surface

Default

The above properties, codes, and colors are based on those used by RasMol.

Appendix E. Imagery

All of the molecular images in this document were made using Molscript (Kraulis 1991) and Raster3D (Merritt *et al.* 1997). While RI Compare does have a built in molecular viewer, the quality that this pair of tools generates is more suited for publications. Molscript is a tool which, given a PDB file and various options from the user, can generate a Raster3D script. This is generally done in two steps. First, `molaulto` is used to generate a rough Molscript script. This script positions the camera, lighting, *etc.*, but most importantly it describes how the molecule should appear by specifying which chains and ligands of the PDB file to show and also where secondary structures exist in the structure. The secondary structure assignment is determined either by reading the assignment from the PDB file or by estimation based on the bond angles and known properties of secondary structures. The assignment is important for high quality imagery, including strands, turns, helices, etc, and without it the entire protein would be rendered as a tube passing through the C_α positions. This initial script was used as a template and modified by adding commands to highlight residues that the RI Compare program had identified. Once a Molscript script was ready it could be processed by `molscript` generating a Raster3D script. While the Molscript files are quite small and easily edited by hand, the Raster3D scripts are much larger and it would be difficult to edit much more than the headers. The Molscript script establishes in a high level language how the geometry will appear and is parameterized by the PDB file. The Raster3D script is basically a collection of raw geometric primitives with fixed positions. Raster3D is used to generate the final graphic file at a high resolution (1500x1500 pixels was used), which is important for quality printing. Since the resolution of a printer is typically quite a bit higher than that of the screen, what appears to be large on the screen is small relative to the output of the printer forcing the image to be scaled up in size. Any scaling can cause distortion, but scaling up can be especially noticeable since solid blocks of a single color often manifest making the image "blocky." The images generated by Raster3D often have a rather large empty border which was cropped off using the auto crop or trim functionality of Adobe Photoshop.

The histograms were generated using The MathWorks' MATLAB. When performing a statistical test with RI Compare a temporary MATLAB `.m` file would be written containing the histogram data itself, already binned, and code to plot the histogram. Since there was no function for plotting the histograms in the fashion presented here additional code was included for drawing the outline of the graph and then filling in either the left or right portion of the graph

indicating which tail of the distribution contained the observed value. These images where exported as `.emf` files (Extended Meta Files) from MATLAB. This format is a vector graphic format and allows for clean scaling.

The phylogenetic trees shown came from screenshots of RI Compare itself. While the single trees could easily be made by a standard tool such as Phylip (Felsenstein, J. 1993, 1989), no tool at the time of this writing was known that could be used to generate the graphs comparing clades from different topologies highlighting identical clades. The screenshots occasionally had to be stitched back together in Photoshop if the tree was larger than what could be displayed on the screen at a single time.

References

Archie, J.W. 1989. Homoplasy excess ratios: New indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. Systematic Zoology. 38(3):253-269.

Bacon, D.J., Anderson, W.F. 1988. A fast algorithm for rendering space-filling molecule pictures. Journal of Molecular Graphics. 6:219-220.

Bairoch A., Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research. 28:45-48.

Clarke, N.D. 1995. Co-variation of residues in the homeodomain sequence family. Protein Science. 4(11):2269-2278.

Crandall, K.A., Hillis, D.M. 1997. Rhodopsin evolution in the dark. Nature. 387:667-8

Dickerson, R.E. 1983. Hemoglobin: structure, function, evolution, and pathology.

Elcock, A.H. 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. Journal of Molecular Biology. 312:885-896.

Farris, J.S. 1989. The retention index and the rescaled consistency index. Cladistics. 5:417-419.

Felsenstein, J. 1989. PHYLIP -- Phylogeny inference package (Version 3.2). Cladistics. 5: 164-166.

Felsenstein, J. 1993. PHYLIP (Phylogeny inference package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Foley, J.D., Van Dam, A., Feiner, S.K., Hughes, J.F. 1990. Fundamentals of interactive computer graphics. Second edition. Addison-Wesley. ISBN 0-201-12110-7.

Hwa, J. *et al.* 1999. Structure and function in rhodopsin. Proceedings of the National Academy of Science. 96:1932-1935.

Hwa, J. *et al.* 2001. Structure and function in rhodopsin. Proceedings of the National Academy of Science. 98:4872-4876.

Kim, H.W., Shen, T.J., Sun, D.P., Ho, N.T., Madrid, M., Tam, M.F., Zou, M., Cottam, P.F., Ho, C. 1994. Restoring allosterism with compensatory mutations in hemoglobin. Proceedings of the National Academy of Science. 91:11547-11551.

Klassen, G.J., Mool, R.D., Locke, A. 1991. Consistency indices and random data. Systematic Zoology. 40(4):446-457.

Kluge, A.G., Farris, J.S. 1969. Quantitative phyletics and the evolution of anurans. Systematic Zoology. 18:1-32.

Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. Journal of Applied Crystallography. 24:946-950.

Korber, B.T.M., Farber, R.M., Wolpert, D.H., Lapedes, A.S. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proceedings of the National Academy of Science. 90:7176-7180.

Lapedes, A.S., Giraud, B.G., Liu, L.C., Stormo, G.D. 1997. Correlated mutations in protein sequences: phylogenetic and structural effects. Proceedings of the AMS/SIAM Conference on Statistics in Molecular Biology. Seattle, WA July 1997.

Maddison, D.R., Maddison, W.P. 1996. The Tree of Life: A multi-authored, distributed Internet project containing information about phylogeny and biodiversity. Internet address: http://phylogeny.arizona.edu/tree/phylogeny.html

Merritt, E.A., Murphy, M.E.P. 1994. Raster3D version 2.0: A program for photorealistic molecular graphics. Acta. Cryst. D50:869-873.

Merritt, E.A., Bacon, D.J. 1997. Raster3D: Photorealistic molecular graphics. Methods in Enzymology. 277:505-524.

Neider, J.L., Davis, T.R., Woo M. 1993. OpenGL programming guide. OpenGL Architecture Review Board, Addison-Wesley, ISBN 0-201-63274-8.

Pesola, G. *et al.* 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. Journal of Molecular Evolution. 48:427-434.

Palczewski, K. 2000. Crystal structure of rhodopsin: A G protein-coupled receptor. Science. 289:739-745.

Pichaud, F., Briscoe, A., Desplan, C. 1999. Evolution of color vision. Current Opinion in Neurobiology. 9:622-627.

Rongey, S.H., Paddock, M.L., Feher, G., Okamura, M.Y. 1993. Pathway of proton transfer in bacterial reaction centers: Second-site mutation Asn-M44 → Asp restores electron and proton transfer in reaction centers from the photosynthetically deficient Asp-L213 → Asn mutant of *Phodobacter sphaeroides*. Proceedings of the National Academy of Science. 90:1325-1329.

Shannon, C.E. 1948. A mathematical theory of communication. The Bell System Technical Journal. 27:379-423, 27:623-656.

Smith, J.L. 1998. Secret life of cytochrome bc1. Science. 281:58-59.

Teichmann, S.A. *et al.* 2001. Determination of protein function, evolution, and interactions by structural genomics. Current Opinion in Structural Biology. 11:354-363.

Todd, A.E. *et al.* 2001. Evolution of function in protein superfamilies, from a structural perspective. Journal of Molecular Biology. 307:1113-1143.

Xia, D. *et al.* 1997. Crystal structure of the cytochrome bc_1 complex from bovine heart mitochondria. Science. 277:60-66.