ABSTRACT

ROBINSON, DOUGLAS MICHAEL. D.R. EVOL: Three Dimensional Realistic Evolution. (Advisor: Jeffrey Thorne)

Simplifying assumptions are necessary to model complex biological processes. Although some assumptions may make sense mathematically, they are often implausible when literally translated. This is especially true of the independence among codons assumption, which states that the evolutionary rate at one codon is independent of the evolutionary rate at surrounding codons. Sites within proteins must interact in order to form intricate three-dimensional binding sites and activation domains. This dissertation details the derivation of a procedure for statistical inference when independent change is not assumed.

The procedure is implemented in a Bayesian framework where Markov chain Monte Carlo methods permit approximation of posterior distributions. Analyses with the procedure on data sets with two and three taxa are explored and biologically plausible values of the solvent accessibility and pairwise interaction parameters are inferred. Via these analyses, we illustrate the chronological ordering of amino acid replacements and the detection of specific events to be positively selected. We also find spatial clustering of the amino acid replacements that have most affected sequence-structure compatibility during the evolution of primate eosinophil-derived neurotoxin proteins.

D.R. EVOL: THREE DIMENSIONAL REALISTIC EVOLUTION

by

Douglas Michael Robinson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

BIOINFORMATICS

Raleigh
2003

APPROVED BY:

Chair of Advisory Committee

DEDICATION

To the loving memory of my mom, Heidi D. Robinson. I will always know that you are proud of me.

PERSONAL HISTORY

Douglas Robinson was born on April 2, 1973, to Donald and Heidi Robinson, in Newton, Massachusetts. He spent his formative years growing up in Needham, Massachusetts, a small suburb located 10 miles southwest of Boston. While growing up, Doug enjoyed the outdoors, especially swimming, skiing, camping and hiking mountains in nearby New Hampshire and Vermont.

After graduating from Needham Highschool in 1991, Doug attended the University of Massachusetts at Amherst majoring in mathematics and pre-med. This gave him the opportunity to test his academic wings with both difficult math courses, such as differential and partial differential equations, and challenging science courses, such as organic chemistry, human anatomy and physiology and genetics. His desire to enter medical school led him to spend a tremendous amount of time and money on applications, as well as countless hours studying for the brutal MCAT exam. After receiving politely worded rejection letters from every single medical school to which he applied, thoughts concerning his future quickly changed, and the decision to attend the University of Vermont for graduate work was made.

It was during this time that two events occurred which would change his life forever. First was the release of Austin Powers: International Man of Mystery, in which Doug was introduced to a character known as Dr. Evil. Although the details of Dr. Evil's life were quite inconsequential, in a twisted sense of irony it seemed rather strange how Dr. Evil's childhood paralleled Doug's own: Summers in Rangoon, luge lessons, etc. It was in this vein that Dr. Evil would shape Doug's thoughts and make him want to major in World Domination. Because this was not a recommended graduate level major at UVM, Doug decided to continue along the mathematics route. The second event was meeting his future wife Heather at a Halloween party in October of 1997. Rumor has it that it was not until their third date that they would actually see each others true identity under their respective Halloween costumes. School was a major focus of Doug's life, but fortunately Heather had a large influence on Doug, by teaching him that there was more to life than just school....and thoughts of world domination, of course.

Doug's Master's thesis led him to enter the world of Protein Evolution, and to the understanding of a little band known as Phish. Between acquiring many live recordings of the band, as well as many scientific journal articles on protein evolution, he would decide that his education had to continue. Through the internet, Doug found North Carolina State University and a professor named Dr. Jeff Thorne who studied just that. Thus, a short 850 mile journey south and 4 + years later, Doug completed his Ph.D. with the creation of Three Dimensional Realistic Evolution, or D.R. EVOL for short. This phylogenetic software package may not dominate the entire world, but may at least dominate the world of molecular evolution.

In the future Doug has two goals: First, find as many ways as possible to dominate the world. Besides this, Doug also wants to go into the pharmaceutical industry where he hopes to make powerful drugs that will drive up the cost of health care. The idea being, that hopefully this cost will inadvertently put an enormous amount of pressure and strain on the medical schools to do something about the problem. To complete the vicious circle of life, Doug then dreams of the day when the medical schools contact him for relief, to which he will only sit back and laugh in their faces. Then calmly, he will remind them that this situation could have been avoided if they had only accepted his medical school applications when they had the chance so many years ago!!!

ACKNOWLEDGEMENTS

What a long strange trip its been. This dissertation is the culmination of effort of many people for which I am most grateful. I am especially grateful to my advisor Jeff Thorne who had what appeared to be an unending amount of patience, trust and moral support to guide me through a Ph.D that was enormous in scope. Also, I want to thank him for his immeasurable amount of confidence that I could accomplish such a difficult task. Jeff, I will always appreciate the many hours spent in the intellectual "stratosphere", discussing the sometimes minute intricacies of the model, especially those where gigantic leaps were made. I would also like to thank the other members of my committee, Bruce Weir, Spencer Muse, Bill Atchley, and Ed Buckler for their helpful comments and support. On a personal note, I want to thank Ed for his in depth insight into the future directions of my project, Spencer for his writing ability and for teaching me how to white water kayak, Bill for his humorous comments to help ease the tension, and Bruce for entrusting me with the Bioinformatics GSA and for allowing me the honor to brew beverages for the Bioinformatics Summer Institute. Besides my committee, none of this would be possible if were not for those who may have not been on my committee, but whose influence and assistance made this initial idea a reality. For instance David Jones and Nick Goldman, but also Hirohisa Kishino, whose statistical ingenuity and infinite generosity made going to, and living in Tokyo Japan an experience that I will remember and treasure for the rest of my life.

I am indebted to my family, especially my brother Chris, his wife Tara, and niece Jordan; for my in-laws Stephanie France, and Kevin Tyler and my sister in-

law Heidi Leong for their love and encouragement. I also want to thank my brother in-law Darryl Leong for his cutting sarcasm and for feigning poor golf skills. Don't worry Darryl, I won't tell anyone that it was not an act. The most important person however, is my lovely wife Heather. Heather, I could not have done this without your constant love of me during a time when my thoughts and ramblings were incomprehensible. I tried to warn you that it would be difficult, but I think we both learned the hard way that it was more than we bargained for. I promise you that my next dissertation will be on discovering why you ever said yes to me in the first place. I love you!

There are some other family members that I must thank and although they will not be able to read this, I want to thank my two cats, Boobah and Poopoos for their support during this process. By sleeping on the many stacks of papers around my office, they ensured that any sudden wind bursts would not negatively impact my progress. They were also instrumental in reminding me to take time for the truly important tasks, such as the occasional belly-rub and short snack breaks.

I am grateful to those in the program of Bioinformatics and Statistical Genetics who put up with my insanity. I will never understand why they listened to me as I rambled on about whatever N.P.R. was discussing during my ride into school. I want to especially thank Debbie Hibbard, whose daily discussions motivated me to get my butt to the gym and work out some of my tension on the rowing machine. Debbie, thanks for always being available to lend an ear to listen to my problems. I also want to thank Dahlia Nielsen for letting me interrupt her work on a regular basis and for being an endless bounty of knowledge.

I want to thank those people who graciously agreed to edit my thesis; my wife Heather, Terry Stigers, Stephane Aris-Brosou, Tae-kun Seo, Jeff Thorne, and especially my brother Chris, who patiently endured multiple revisions of my second manuscript. I also owe it to the Thorne working group namely, Stephane Aris-Brosou, Betsy Scholl, Tae-kun Seo and Jiaye Yu, who spent countless hours listening to me rehearse my presentations. To my many friends at the B.R.C., who I would like to collectively refer to as my answering service, I thank you for taking the time to write those messages and find me, despite my desk being the furthest from the phone. To Jimmy Doi, Errol Strain, Frank Mannino, Betsy Scholl, Stephane Aris-Brosou, David Aylor, Sunil Suchindran, Jack Liu and Josh Starmer, I want to thank you for your friendship and for succeeding to get me out of the office to play frisbee golf, real golf, or just go out to have a drink. You have no idea how much I appreciate your efforts. But I can not forget the person who I consider one of my best friend during this whole process, Andrea Johnson. Your offbeat comments, stories about ponies and almost constant barrages of my theories kept me in check and added a great deal of humor to my life. Andrea, we made it through this and we did it together.

Lastly, I want to thank the band Phish for providing the soundtrack to my dissertation. Through their endless jams, I was able to concentrate on my work and focus in on the many intricacies that my model always seemed to present. I also want to thank the coffee and tea growers around the world for their precious, precious caffeine rich crops. Furthermore, I have to thank the makers of TUMS, or any antacid for that matter, for helping my digestive system cope with the stress. I also want to thank the N.C. State crew team for allowing me valuable time on the rowing machines. This was especially true when I placed second during their

2000 meter erg time trials. They succeeded in making a 30 year old grad student feel like he could still compete with the undergrads. In conclusion, I want to thank my nephew Brandon Minervini, who at age 7 innocently asked me if the reason I was still in school was if I failed. Well Brandon, I finally passed!

Contents

List of Tables

List of Figures

Chapter 1

REVIEW

Introduction

The study of molecular evolution has had a rich history, yet is far from being completely resolved. While a great deal of progress has been made in recent years, there are still many facets that warrant explanation and questions that still need to be answered. With the introduction of high throughput sequencing, genome shot gunning and accelerated PCR methods, it is now possible to obtain massive amounts of high quality sequence data at a rapid pace. With the entire genome of many organisms now complete, in particular the human genome (International Human Genome Sequencing Consortium, 2001, Venter et al. 2001), interest in the field of molecular evolution has seen a dramatic increase. To meet this demand, new and more robust models for comparative sequence analysis and phylogenetic inference are necessary.

The relationship between geneticists and molecular evolutionists is mutually beneficial and cyclical in nature. Geneticists uncover minute components of the process of sequence change using the latest technologies. With this information, molecular evolutionists try to build robust, statistically founded models that discern the relationships between biologically motivated parameters to best fit the model system. Results from these models can then strengthen the support of currently held theories, or discover relationships that were previously unknown. Whatever the circumstance, the relationship between genetics and molecular evolution will provide avenues for future experimentation, as well the means for creating more realistic models and will inevitably broaden our understanding of sequence evolution.

To begin to model this complex phenomenon, researchers greatly simplify the process of sequence change by making several assumptions and considering only a few biologically realistic factors (Thorne, 2000). Most widely used models of sequence evolution exploit the assumption that individual sites evolve independently from one another. The independence assumption dictates that a substitution at one site does not influence the rate of substitution at surrounding sites. From a computational standpoint this assumption is quite attractive, for it makes statistical inference on evolutionary trees computationally tractable. This calculation is achieved via Felsenstein's pruning algorithm (Felsenstein, 1981) where the likelihood of an individual site in an alignment can be easily determined. The full likelihood is subsequently computed simply by taking the product of each individual site likelihood over all sites in the alignment.

The pruning algorithm requires an amount of computation proportional to the sequence length N, the number of internal nodes on the phylogeny (for bifurcating rooted topologies, this equals the number of taxa minus one) and the square of the number of characters states n that are allowed at each site. Typically, models of nucleotide substitution (e.g., Jukes and Cantor, 1969, Kimura, 1980, Felsenstein, 1981, Hasegawa et al., 1985, Felsenstein, 1989) have $n = 4$, whereas models of amino acid replacement (e.g., Dayhoff et al., 1978, Jones, Taylor and Thornton, 1992b) have $n = 20$ and models of codon change (e.g., Muse and Gaut, 1994, Goldman and Yang, 1994) have $n = 61$. Unfortunately, the computationally attractive assumption of evolutionary independence among sites is not biologically plausible for protein coding sequences. Protein sequences must adopt complex three dimensional folds and maintain specific activation sites and binding domains. Thus, evolution must occur between compatible residues in order to maintain protein functionality. In reality, the effect of a substitution might cascade through the protein, changing the substitution rates at other sites. This is especially true at sites in the protein core because of the densely packed nature of the folded protein structure.

Models are built upon assumptions that may be mathematically reasonable, but are not always considerate of the biological system to which they are applied. Consequently, relaxing certain assumptions may cause results to be computationally unobtainable. For instance, to relax the independent evolution among sites assumption, the notion of evolution occurring at individual sites or codons must be extended to the idea of entire sequence evolution. The form of the explicit rate matrix necessary to accomplish this is beyond the capabilities of modern day computers. Thus, to calculate likelihoods on phylogenetic trees, alternative methods have to be derived.

The field of molecular evolution is initially presented in a biological context. Without a clear understanding of the biological system under investigation, parameter estimates and subsequent data analyses have no foundation. A tour through some of the pioneering work that has driven the field of molecular evolution over the past forty years is then presented. Relaxing the independence assumption is a natural step in the long progression of statistical models of evolution and it will be interesting to observe the insights and ramifications that this work will have on the field.

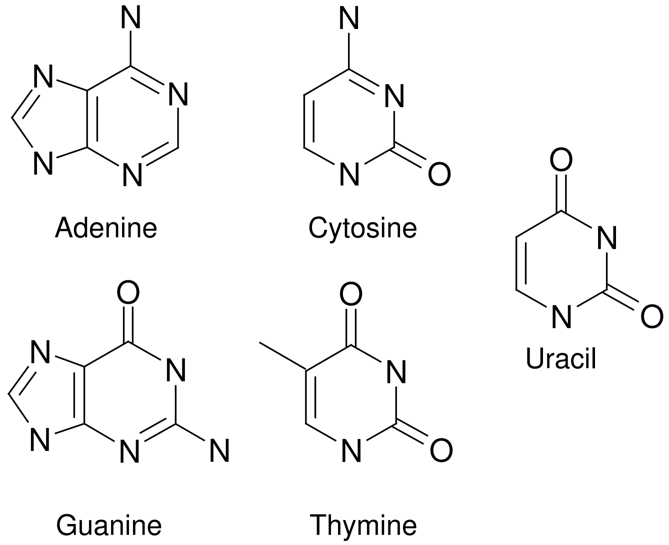


Figure 1.1: The four nucleotide bases, Adenine, Guanine, Cytosine and Thymine, plus the RNA nucleotide Uracil are shown. Adenine and Guanine are purines, while Cytosine, Thymine and Uracil are pyrimidines. Because of the structural similarity, purine-purine or pyrimidine-pyrimidine substitutions are much more likely than substitutions between groups.

Biological Background

The genetic makeup of most organisms is contained within DNA (Deoxyribonucleic acid). DNA is composed of four nucleotides: Adenine, Cytosine, Guanine and Thymine, abbreviated A, C, G and T, respectively. The nucleotides come in two varieties, the purines, which includes A and G and the pyrimidines, which includes C and T. All four nucleotides are unique, yet the structural similarity within purines or pyrimidines is substantially higher than that between the two groups (See Figure (1.1)). Just five short decades ago, Watson and Crick, (1953a) proposed the structural configuration of the DNA molecule as two antiparallel strands of DNA that combine through hydrogen bonding of nucleotide base pairs. The natural bonding pattern was determined to occur between a purine and a pyrimidine, namely A to T, with the formation of two hydrogen bonds and C to G, through the formation of three hydrogen bonds. This arrangement allows a semi-conservative mechanism of DNA replication whereby each parental strand would separate from the unwound double helix and serve as a template for the newly synthesized strand (Watson and Crick, 1953b; Messelson and Stahl, 1958).

DNA is transcribed to RNA (Ribonucleic acid) which is translated to amino acids, the building blocks of proteins. This is known as the central dogma of molecular biology.

$$\mathrm{DNA} \quad \overbrace{\Rightarrow}^{Transcription} \quad \mathrm{RNA} \quad \overbrace{\Rightarrow}^{Translation} \quad \mathrm{PROTEIN}$$

DNA and RNA are very similar to one another except for two points. Both DNA and RNA share Adenine, Cytosine and Guanine, but instead of the Thymine base, RNA contains Uracil (U) (See Figure 1). Also, both DNA and RNA bases are each connected to pentose sugar molecules called ribose, but RNA has a hydroxyl group bonded to the $2'$ carbon of its ribose while DNA has only a single hydrogen atom bonded at this position, hence the name *de-oxy-ribose*.

Three letter combinations of RNA nucleotides code for specific amino acids that are common to all organisms. The sixty-four possible words, or codons, comprise the (nearly) Universal Genetic Code (see Table 1.1). Exceptions to the universality of the genetic code are observed in yeast and certain protozoa, however the most visible is found in mammalian mitochondria. In particular, three noticeable differences include: UGA encoding tryptophan rather than a stop codon; AUA defining methionine instead of isoleucine; and AGA and AGG translating stop codons rather than arginine (e.g., Snustad and Simmons 2000). There are twenty distinct amino acid types in existence, which means that there is some degeneracy

The Universal Genetic Code

SECOND

FIRST	U		C		A		G		THIRD
U	UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine	**U**
	UUC	Phenylalanine	UCC	Serine	UAC	Tyrosine	UGC	Cysteine	**C**
	UUA	Leucine	UCA	Serine	UAA	**STOP**	UGA	**STOP**	**A**
	UUG	Leucine	UCG	Serine	UAG	**STOP**	UGG	Tryptophan	**G**
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	**U**
	CUC	Leucine	CCC	Proline	CAC	Histidine	CGC	Arginine	**C**
	CUA	Leucine	CCA	Proline	CAA	Glutamine	CGA	Arginine	**A**
	CUG	Leucine	CCG	Proline	CAG	Glutamine	CGG	Arginine	**G**
A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine	**U**
	AUC	Isoleucine	ACC	Threonine	AAC	Asparagine	AGC	Serine	**C**
	AUA	Isoleucine	ACA	Threonine	AAA	Lysine	AGA	Arginine	**A**
	AUG	Methionine	ACG	Threonine	AAG	Lysine	AGG	Arginine	**G**
G	GUU	Valine	GCU	Alanine	GAU	Aspartic Acid	GGU	Glycine	**U**
	GUC	Valine	GCC	Alanine	GAC	Aspartic Acid	GGC	Glycine	**C**
	GUA	Valine	GCA	Alanine	GAA	Glutamic Acid	GGA	Glycine	**A**
	GUG	Valine	GCG	Alanine	GAG	Glutamic Acid	GGG	Glycine	**G**

Table 1.1: The nearly Universal Genetic Code is used to decode all possible three nucleotide combinations into its proper amino acid. The table is set up such that the nucleotide in the left most column corresponds to the first position of the codon, the nucleotide in the top row of the table corresponds to the second position in the codon and the right most nucleotide corresponds to the third position in the codon. Notice the degeneracy of the genetic code as well as the existence of the three stop codons, namely UAA, UAG and UGA, which signal the ribosomes to terminate translation.

in the genetic code. For instance, there are six codons that each translate Leucine, while there are only two codons that translate Histidine. From Table 1.1, one can observe that there are sixty-one sense codons that translate true amino acids and three that translate STOP codons, namely UAA, UAG and UGA. When read, stop codons instruct the ribosomes to terminate the translation process and to separate into their component halves. In most cases, stop or nonsense codons signal the release of a fully translated protein sequence. Yet, in the case of a premature stop codon (i.e. a nonsense mutation), the resultant shortened sequence is most likely non-functional. In these cases, the protein is targeted by the cellular defense mechanisms for destruction. Depending on the role of the affected protein in the cell, a nonsense mutation may be lethal.

Each of the twenty amino acids share a common backbone, yet all have unique side chains which emanate from the central carbon atom, denoted C_α. Because of the composition and arrangement of the atoms within the various side chains, physical and biochemical properties are conferred on each residue. These properties include for example, overall size, polarity, hydrophobicity and charge which have been used by researchers as a basis of comparison (Grantham 1974; Taylor and Jones 1993; Koshi and Goldstein 1995). Because of the common backbone structure, any two amino acids can form a peptide bond. In this reaction, the negatively charged carboxyl terminal of one residue is attracted to the positively charged amino terminal of the next residue. Both are joined via a dehydration synthesis reaction with the expulsion of a single water molecule.

The precise sequence of amino acids is called the primary structure of a protein. Interactions between the side chains of the amino acids cause the linear sequence to contort and adopt unique conformations necessary for the protein to function properly. Changes in the identity of residues in a folded protein may negatively impact surrounding residues, which in a worst case scenario may force a change in the overall structure resulting in a complete loss of function (Purves, Orians and Heller 1992). The fact that any two amino acids can bond complicates the analysis of protein sequences, for in theory an amino acid sequence of length N has 20^N unique possible residue combinations. However, because of negative interactions only a minute fraction of the 20^N possible sequences ever encode viable proteins.

Certain sub-structures have been repeatedly observed in many proteins and have been classified as local secondary structures. Some examples include α-helices, β-pleated sheets, turns and loops. These secondary structures have been of great interest to molecular evolutionists for it has been proposed that the rate of amino acid replacement depends partly on local secondary structure (Overington et al., 1992; Koshi and Goldstein 1995). The amino acids themselves have differing affinities for each of the various secondary structures (e.g., Thorne, Goldman and Jones 1996) and certain amino acids, in particular Proline and Glycine, actually prevent the formation of the α-helix (Purves, Orians and Heller 1992). Secondary structures can stably interact with one another through hydrogen bonds, salt bridges, and Cysteine-Cysteine di-sulfide bonds, forming unique tertiary structures that confer functional specificity to the proteins enabling them to bind other proteins and DNA through the formation of activation and catalytic domains.

One way to describe sites within a folded protein structure is by local secondary structure. Some properties associated with the folded protein are percent solvent accessibility and pairwise interactions with surrounding residues. Given a highly resolved protein crystal structure, these quantities can be defined and fixed for each site in the protein. In general, the percent solvent accessibility is a measure of the degree to which a site is exposed to the surrounding solvent. Intuitively, those residues that lie on the surface of a protein have a large percent accessibility and are generally hydrophilic in nature. Likewise, those sites within the core of the protein have correspondingly low percent accessibility and are filled typically with hydrophobic residues. Hydrophobic residues aggregate similar to the behavior of oil droplets in water. This has led to the proposition that hydrophobicity drives the process of protein folding (Dill, 1990a, Dill, 1990b). In this hypothesis, the hydrophobic core of the protein forms first, while the proper folding of the external features of the protein takes place shortly thereafter.

Like gravity, the electrostatic force, or potential between residues within a protein decreases with the square of the three dimensional separation distance. Hence, pairwise interactions can only be defined between those residues that are in close proximity in the folded protein state. Interestingly, residues that are distantly separated along the primary structure of a protein may in fact be located close in three dimensions because of the complex folding pattern of the tertiary structure. In general, proteins are densely packed molecules and sites on the surface of the protein may generally interact with only a few sites, but those in the core inevitably interact with substantially more sites. Amino acid replacements in the core may consequently have greater influence on protein functionality, simply because of their effect on a higher proportion of sites.

It is with this biological background that molecular evolutionists build statistical models that try to mimic closely the complex evolutionary substitution process. Through an enormous amount of trial and error statistical assumptions are relaxed, allowing models to become increasingly more realistic and consequently, more computationally complex in the process. Over the years evolutionary models have themselves evolved to capture these observed behaviors, which is described in the following section.

Statistical Models

DNA is not a static entity. Over time, all organisms experience mutations in their DNA that have varying effects on the proteins they encode. The reasons for mutation events are varied and stem from normal cellular processes to interactions with the environment. Some causes include internal sequencing errors in DNA replication, repair mechanisms, insertions and deletions or more severe DNA damage such as the formation of pyrimidine dimers. Unfortunately, the word mutation conjures up images of freakish genetic experiments gone awry and in some cases, mutations can be deleterious to the host. Fortunately, over time these mutations are usually removed from a population through natural selection.

Mutations in which a nucleotide is substituted come in two varieties, namely transitions and transversions. Transitions occur when a nucleotide is replaced by a structurally similar nucleotide, that is Purine → Purine, Pyrimidine → Pyrimidine, while transversions occur when a nucleotide is replaced by a structurally dissimilar nucleotide, that is Purine → Pyrimidine, Pyrimidine → Purine. The ramification of such an event is realized when the codon containing the substituted nucleotide is translated to its corresponding amino acid residue.

Synonymous substitution events do not influence the identity of the encoded amino acid residue in the protein. However, nonsynonymous substitutions cause the translation of a different amino acid residue, yet because of the location of the change, or by possessing similar physical and biochemical properties, the protein may remain fully functional. Whether a substitution is synonymous or nonsynonymous depends on the identity of the nucleotides that comprise the codon unit at the instant of the substitution event. Unlike synonymous substitutions, nonsynonymous substitutions are not tolerated equally well across the protein and may be detrimental to overall functionality. In general, researchers have found that the rate of nonsynonymous substitutions on the surface of globular proteins is about twice that of residues buried in the core of the structure (Goldman, Thorne and Jones 1998). In rare instances mutations may impart a selective advantage, allowing an organism to better adapt to their surroundings. These mutations are often passed on to future generations and become fixed in the population.

The effects of nucleotide substitutions can be reflected in the creation of statistical models. Most models of molecular evolution rely upon theories of stochastic processes and Markov chains. By definition, a stochastic process is a collection of random variables governed by probabilistic laws defined on a common probability space, while a first order Markov chain is a specific type of stochastic process with the property that the next state of the system depends only on the current value (see Karlin, 1966). In this way, all states of the Markov chain prior to the current value have no influence on the probability of future events.

Nucleotide Models

In 1962, Zuckerkandl and Pauling recognized the fact that DNA sequence information could be used to classify species. Although at this time, not many sequences were available for analysis, homologous sequences from different species were aligned by hand, and the number of sites that differed were tabulated. This idea was formalized in the theory of parsimony (Edwards and Cavalli-Sforza 1963) which, like the notion of Occam's razor, proposed that the most likely explanation of sequence evolution was the one that minimized the total number of substitution events. In 1965, Zuckerkandl and Pauling proposed the concept of the molecular clock that stated the rate of evolution was constant over time. Consequently, the relationship of any two sequences could then be measured by simply counting the number of differences between them.

Using the statistical theories discussed above, Jukes and Cantor in 1969 altered the field by creating the first stochastic rate matrix, the JC69, which framed molecular evolution in the context of a stochastic process. Under this model, the rate of change to any nucleotide is equal to that of any other nucleotide. Consequently, not only is the probability of replacement the same, but the steady state, or limiting distribution of the four nucleotides is also equivalent and fixed at 0.25. In general, when building a model of molecular evolution, determining which sequence is ancestral is nearly impossible. For this reason, models are typically constructed with the statistical assumption of time reversibility which makes the ancestral sequence choice arbitrary. With this assumption, the probability of starting from nucleotide type i and changing to nucleotide type j in a time interval is the same as the probability of starting from j and going backwards to i in the same time duration. The models discussed herein (see Table 1.2) are all time reversible and contain between 1-9 parameters to estimate from the data.

Technological advances increased the availability of DNA sequences and limitations of Jukes and Cantor's JC69 model were more readily apparent. Through greater biological understanding, statisticians and geneticists used the JC69 methodology as a launch pad for their own models. For instance, researchers noticed that

JC69
$$\begin{pmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1 & \cdot & 1 \\ 1 & 1 & 1 & \cdot \end{pmatrix}$$

K2P
$$\begin{pmatrix} \cdot & 1 & \kappa & 1 \\ 1 & \cdot & 1 & \kappa \\ \kappa & 1 & \cdot & 1 \\ 1 & \kappa & 1 & \cdot \end{pmatrix}$$

F81
$$\begin{pmatrix} \cdot & \pi_C & \pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \pi_T \\ \pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \pi_C & \pi_G & \cdot \end{pmatrix}$$

F84
$$\begin{pmatrix} \cdot & \pi_C & \pi_G(1+\frac{\kappa}{\pi_{H(G)}}) & \pi_T \\ \pi_A & \cdot & \pi_G & \pi_T(1+\frac{\kappa}{\pi_{H(T)}}) \\ \pi_A(1+\frac{\kappa}{\pi_{H(A)}}) & \pi_C & \cdot & \pi_T \\ \pi_A & \pi_C(1+\frac{\kappa}{\pi_{H(C)}}) & \pi_G & \cdot \end{pmatrix}$$
$$\text{Where } \pi_{H(C)} = \pi_{H(T)} = \pi_C + \pi_T$$
$$\text{and } \pi_{H(A)} = \pi_{H(G)} = \pi_A + \pi_G$$

HKY85
$$\begin{pmatrix} \cdot & \pi_C & \kappa\ \pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\ \pi_T \\ \kappa\ \pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\ \pi_C & \pi_G & \cdot \end{pmatrix}$$

GTR
$$\begin{pmatrix} \cdot & a\ \pi_C & b\ \pi_G & c\ \pi_T \\ a\ \pi_A & \cdot & d\ \pi_G & e\ \pi_T \\ b\ \pi_A & d\ \pi_C & \cdot & f\ \pi_T \\ c\ \pi_A & e\ \pi_C & f\ \pi_G & \cdot \end{pmatrix}$$

Table 1.2: This table lists many of the time reversible models of nucleotide evolution. As one descends the table, the number of free parameters steadily increases from JC69 having one free parameter, to the most general time reversible model, GTR, which has nine free parameters. See the accompanying text for a full description of the models and the various parameters that are incorporated (see also Yang 1997).

transitions occurred more often than transversions which was incorporated into a model in 1980 called the K2P (Kimura 1980), and subsequently verified with an extensive empirical study (Brown and Simpson 1982). The K2P model is identical to the JC69 except for a new parameter κ, the transition / transversion rate parameter, which affords researchers the ability to estimate how much more likely transitions are in their dataset.

Shortly thereafter, Felsenstein proposed a pair of stochastic models, commonly referred to as the F81 (Felsenstein, 1981), and the F84 (see Felsenstein, 1989). Contradictory to the uniform steady state distribution hypothesized previously (Jukes and Cantor, 1969) Felsenstein recognized that in most cases there exists a nucleotide base composition bias in DNA sequence data. To alleviate this restriction three degrees of freedom were added to the F81 model by allowing the relative frequency of the nucleotides to vary. The instantaneous rate in this model is proportional to π_j, the relative frequency of nucleotide type j. Frequently, good estimates of the relative frequencies of the four nucleotides are easily derived from counts of the individual nucleotides in the dataset under investigation and used for the remainder of the likelihood calculation. It is reassuring to note that under the independence assumption, the estimates derived by this counting method are very similar to those obtained by estimation through likelihood procedures. The F84 model is an integral part of the evolutionary procedure presented in later chapters of this work, and thus will not be discussed now.

The idea of combining a model that allows the relative frequency of the nucleotides to vary with a transition-transversion rate parameter resulted in the formation of two models, namely the HKY85 (Hasegawa, Kishino and Yano, 1985) and what is commonly referred to as the F84 (Felsenstein, 1989). Only minute differences in how each measures a transition substitution distinguish these models. Similar to the K2P model of Kimura (1980), the HKY85 model captures this behavior using the transition-transversion rate parameter κ. However, the F84 model uses:

$$\frac{\kappa\pi_j}{\pi_{H(j)}} \qquad \text{where} \qquad \pi_{H(j)} = \begin{cases} \pi_A + \pi_G & \text{if j is a purine} \\ \pi_C + \pi_T & \text{if j is a pyrimidine} \end{cases} \tag{1.1}$$

which allows the rate of a change for transitions to be conditional on nucleotide group membership. The parameters in both models are estimated through maximum likelihood techniques.

The most general time reversible model contains at most nine parameters to estimate from the data; three for the relative nucleotide frequencies and up to six for each possible pairwise nucleotide substitutions. For this reason it is referred to as the General Time Reversible (GTR) model (e.g., Tavaré 1986; Yang 1994a; Zharkikh 1994). In order to obtain reasonable estimates for the large number of parameters considered, this model should not be used to analyze datasets with only a few, relatively short sequences. Thus for a particular dataset, in order to choose the optimal model two thoughts must be considered. Adding parameters the model makes it more general and increase its applicability. On the other hand, although more biologically realistic assumptions can be considered, the computational complexity of a model increases with extra parameters. This increases the error variance associated with each parameter estimate (e.g., Zharkikh, 1994). Using too simplified a model may also lead to biased results (e.g., Huelsenbeck, 1995;

Huelsenbeck and Rannala, 1997). The real trick then, is to strike the perfect balance between incorporating many of the rules that govern the process of evolution, yet still remain general enough to provide wide applicability and provide useful statistical inference without a tremendous increase in computational complexity.

Codon Models

Whether a substitution is synonymous or nonsynonymous depends on the identity of the nucleotides that comprise the codon unit. Underlying the process of nucleotide substitution is the nearly Universal Genetic Code (See Table 1.1), thus the process of nucleotide substitution in coding regions is inherently not independent among sites. By incorporating the dependence structure of the Genetic Code, models can be built to separate the biases in substitution patterns found at the nucleotide level, from selective constraints found at the amino acid level (e.g., Goldman and Yang 1994; Yang, Nielsen, and Hasegawa 1998).

The codon models that are frequently used are modifications of those originally proposed by Muse and Gaut (1994) and the Goldman and Yang (1994) (see Table 1.3). The explicit form of both matrices are quite sparse, for each disallows instantaneous changes at more than one codon position, as well as change to premature stop codons. In these Markov models the state space increases to 61 x 61 for the 61 sense codons of the universal genetic code. To model the evolution between two aligned codons that differ in more than one position, both models behave as though there are multiple distinct substitution events that evolve the codon in the given amount of time.

(a) Muse and Gaut, 1994

$$Q_{x,y} = \begin{cases} \pi_\beta\ \chi & \text{If event is synonymous} \\ \pi_\beta\ \phi & \text{If event is nonsynonymous} \\ 0 & \text{If codons differ at more than one posisiton} \end{cases}$$

(b) Goldman and Yang, 1994

$$Q_{x,y} \ = \begin{cases} u\ \pi_{c(y)}\ \kappa\ e^{-d(aa_x,aa_y)/V} & \text{If codons differ by a transition} \\ u\ \pi_{c(y)}\ e^{-d(aa_x,aa_y)/V} & \text{If codons differ by a transversion} \\ 0 & \text{If codons differ at more than one posisiton} \end{cases}$$

(c) Goldman and Yang, 1994 (Simplified)

$$Q_{x,y} \ = \begin{cases} u\ \pi_{c(y)} & \text{If event is a synonymous transversion} \\ u\ \pi_{c(y)}\ \kappa & \text{If event is a synonymous transition} \\ u\ \pi_{c(y)}\ \omega & \text{If event is a nonsynonymous transversion} \\ u\ \pi_{c(y)}\ \kappa\ \omega & \text{If event is a nonsynonymous transition} \\ 0 & \text{If codons differ at more than one position} \end{cases}$$

Table 1.3: Models of codon evolution. The Muse and Gaut codon model (1.3a) uses only 5 parameters. The dual parameterization of the parameter-rich Goldman and Yang codon rate matrix is also given in 1.3b,c. Most of the parameters in 1.3b,c are stationary codon frequencies. Explanation of the parameters used in each model are given in the accompanying text.

The Muse and Gaut model (Table 1.3(a)) accounts for the constraint at the nucleotide level through π_β, the equilibrium frequency of the substituted nucleotide type in the target codon accounting for stop codons. At the amino acid level, this model allows for heterogeneity of rates between synonymous and nonsynonymous substitutions through χ, the synonymous rate parameter and ϕ, the nonsynonymous rate parameter. This model has only five rate parameters which can be easily resolved from the data through maximum likelihood procedures.

To model the selective constraints at the nucleotide level, instead of using the equilibrium frequency of the target nucleotide as above, the equilibrium frequency of the target codon, denoted $\pi_{c(y)}$, can also be used (Goldman and Yang 1994). Because of the massive amount of data required to obtain reasonable parameter estimates of the codon frequencies, $\pi_{c(y)}$ is usually derived from the product of the equilibrium frequencies of the three nucleotides that comprise the codon. These frequencies are then normalized to account for the relative frequency of the three stop codons. Like the K2P (Kimura, 1980) and the HKY85 (Hasegawa, Kishino and Yano, 1985) models of nucleotide substitution, this model also includes the parameter κ, to account for the observance that transitions occur more often than transversions (e.g., Brown and Simpson, 1982).

The Goldman and Yang and Muse and Gaut models of codon substitution can be written to account for differences at the amino acid level using an amino acid distance matrix (Grantham, 1974) (see Table 1.3(b) for the Goldman and Yang parameterization). The Grantham matrix was derived by comparing physical and biochemical properties of the twenty amino acids such as volume, charge, and hydrophobicity. These values are denoted D_{aa_x,aa_y}, where aa_x denotes the amino acid that is translated by codon x in the model, and range from as small as 5 for (ILE - LEU) and as high as 210 for (CYS - PHE). The parameter V is a tuning parameter to allow the distance matrix to better fit the data. A rate normalization constant u is inserted to make the relation $\prod_{y=1}^{y=61} \pi_{c(y)} Q_{y,y} = 1$, possible (see Table 1.3(b)).

Through experimentation, Goldman and Yang found that this parameterization did not fit actual data very well, and a simplification involving the assumption of uniform distance between all amino acid pairs was imposed. This allowed the complex exponential term to be replaced by the single nonsynonymous / synonymous rate parameter ω (see Table 1.3(c)). Unlike the Muse and Gaut model, the Goldman and Yang model has 63 parameters to estimate from the data; sixty of which are the codon equilibrium frequencies. Lately, these methods have been expanded upon to incorporate the behavior seen in lentiviral evolution (Pedersen et al. 1998). This expanded model accounts for the disparity in relative frequencies of the four nucleotides at each of the three codon positions as well as selection against the CpG dinucleotide within codons.

Amino Acid Models

The nucleotide and codon models discussed are all parametric in nature, whereby various parameters are used to define relationships among pairs of nucleotides or codons. When considering models of amino acid evolution, the usual parameters used such as the transition / transversion rate parameter κ, or the nonsynonymous / synonymous rate parameter ω, lose meaning. Hence, many researchers have instead focused on empirical approaches (Dayhoff et al. 1978; Jones et al. 1992; Gonnet, 1992; Hennikoff and Hennikoff, 1992). Empirical models require a reference data set which, if small, will not provide a good fit to a given data set under investigation (Liò and Goldman, 1998). However, the advantage to using empirically derived matrices is that they may more realistically capture the unequal transition probabilities among amino acid pairs as seen in true proteins.

Because of the degeneracy of the genetic code, a synonymous substitution may be the result of several substitutions at the nucleotide level. For example, the codons AGU and UCC may differ at all three codon positions, yet when translated, both code for the same amino acid, Serine (see Table 1.1). When analyzing highly diverged sequences, models of amino acid replacement may be beneficial for they act as a filter for the noise of multiple substitutions at the DNA level (e.g., Goldman and Yang 1994).

An important method for handling protein evolution was that of Dayhoff, Schwartz and Orcutt in 1978 (see also Dayhoff et al. 1972). In this landmark work, highly similar sets of protein sequences were aligned by hand and subsequent counts of observed amino acid replacements were tabulated. The small genetic distance between the sequences allowed the assumption of multiple substitutions at a given site to be negligible. With these counts, the Point Accepted Mutation (PAM) matrix was derived. Because the counts came from global alignments of several protein sequences, the PAM matrix determines the probability for an amino acid replacement located in an average site within an average protein. Unfortunately, the definition of an average site within an average protein becomes difficult to verbalize with any precision, yet despite this apparent drawback, the Dayhoff method has gained wide acceptance and has been used successfully in making phylogenetic inference.

The Dayhoff method was duplicated on a more modern database using a substantially higher amount of sequence information to create the JTT amino acid substitution matrix (Jones, Taylor and Thornton, 1992b). The increase amount of data allowed a more robust estimate of the probability of change between rare amino acid pairs (e.g., Methionine vs. Tryptophan), which were nonexistent in the dataset used for the creation of the PAM matrices. Because very similar methodologies were employed to create the Dayhoff model, the JTT model suffers the same unavoidable consequence of also being defined for an average site in an average protein.

Through matrix multiplication, Dayhoff et al. (1978) were able to create amino acid transition matrices for various PAM distances. In particular, the commonly used PAM120 and PAM250 matrices were derived by multiplying the PAM matrix by itself 120 and 250 times, respectively, where larger numbers represent greater evolutionary distances. By decomposing the Dayhoff instantaneous rate matrix, the probability of change between amino acid residues can also be calculated for any amount of evolution T using standard matrix exponentiation procedures (e.g., Swofford et al. 1986, Kishino, Miyata and Hasegawa, 1990). This type of model also allowed phylogenetic inference using maximum likelihood, but for some unknown reason it has not enjoyed the same widespread use.

Lastly, instead of working with a large amino acid replacement matrix, researchers were able to represent all twenty residues by considering only a finite set of the most conserved physical and chemical characteristics (Koshi, Mindell and Goldstein 1997, 1999; Koshi and Goldstein 1998; Halpern and Bruno 1998). Considering these quantities as fixed, transition matrices based on functions of these properties were then derived. This significantly reduced the number of parameters in the model that may have allowed for accurate parameter estimation, yet may have also oversimplified the true complexities of the evolutionary process, giving results that are often difficult to interpret.

Models That Consider Protein Structure

Starting from a linear sequence of amino acids, the precise manner in which the sequence attains its resultant three dimensional structure is still unknown. Fortunately, there exist methods to discern the precise structures of proteins. In fact, as of January 1, 2003, the Brookhaven Data Bank (PDB) (Bernstein et al. 1977) contained the three dimensional coordinates of roughly 20,000 proteins. These coordinates were mainly derived from protein X-ray crystallography, but were also derived from nuclear magnetic resonance (NMR) techniques. Through precise protein conformations, functionality is conferred. Thus, any disruption in protein structure may easily render the protein non-functional. For example, enzymes are proteins that possess a high degree of specificity and function to catalyze precisely one reaction, which is why they are commonly referred to as lock-and-key proteins. For this reason, although it has long been observed that divergence in homologous protein sequences occurs rapidly, the corresponding change in their associated protein structure is much less severe because of the strong selective pressure to maintain functionality (e.g., Chothia and Lesk 1986; Flores et al. 1993; Russell et al. 1997).

One way in which researchers could explore the ramifications of amino acid replacements in proteins is with site-directed mutagenesis. However, using this method would not only be slow and tedious, but would inevitably be unsuccessful at testing the vast amount of amino acid replacement combinations. It was later hypothesized that by using the database of known sequences one could compile lists of protein characteristics in various local site environments to discern what properties had been conserved through evolutionary time (Koshi and Goldstein 1995). In order to maintain the overall conformation of the protein, it has been observed that amino acids that share physical and biochemical attributes tend to replace one another more often on average than those that are more disparate (e.g., Zuckerkandl and Pauling, 1965; Dayhoff et al. 1978; Parisi and Echave, 2001). Others have defined an explicit distance between amino acid residues based on measures of hydrophobicity, charge and side chain volume (e.g., Grantham, 1974; Taylor and Jones, 1993). Yet, it is not clear whether the forces behind evolutionary change are governed by these properties.

The idea that the rate of amino acid replacement depends on local secondary structure is not novel (Overington et al. 1992; Koshi and Goldstein, 1995). Furthermore, many researchers have understood the importance of incorporating structural and functional information into models to improve evolutionary and phylogenetic inference (e.g., Koshi and Goldstein, 1995, 1997; Thorne, Goldman and Jones 1996; Goldman, Thorne and Jones, 1998). Yet, exactly how each group has achieved this goal has been quite varied. However, researchers have been able to use large databases to analyze the relationships between amino acid residues and secondary structural elements with solved tertiary protein structures. For instance, researchers have analyzed the databases of aligned sequences to create amino acid transition matrices (Dayhoff et al. 1978, Jones et al. 1992). Unfortunately in these cases, only a single matrix was produced to explain the evolutionary history of all sites within a sequence regardless of the local environment of that site. Building on this idea, the structural databases were later analyzed to construct transition matrices for various secondary structural environments and surface accessibilities (Koshi and Goldstein, 1995, 1997; Goldman, Thorne and Jones 1998; Thorne et al. 1996). The results of Thorne et al. (1996) mainly showed that hydrophobicity was highly conserved, that the volume of a residue was slightly conserved in core regions because of tight internal packing constraints and surprisingly, that secondary structural elements were less conserved. Similar to the problems faced by the PAM and JTT matrices, Koshi and Goldstein (1998) problematically note that within each environment, amino acids are treated uniformly by the given matrix.

Logically following this idea even further, one may desire site-specific amino acid transition matrices. In order to obtain reasonable parameter estimates however, the tremendous amount of data necessary for their creation make this too daunting a task (Liò and Goldman, 1998). One way to handle this is to assume that sites can be classified in a limited number of structural classes with which one can create specific transition matrices by analyzing large databases of sequences (Thorne, Goldman and Jones, 1996; Goldman, Thorne and Jones, 1998; Li*ò* et al. 1998). However, in most cases it may be unreasonable to assume that only a finite number of structural classes can accurately account for the vast diversity of sites found in proteins and their associated structures.

One way to simplify the study of proteins is to classify individual sites into categories based on location (internal vs. external), on local secondary structures ($\alpha-$ helix, $\beta-$ sheet, loop, etc.), and percent solvent accessibility. Using various property combinations, large tables have been compiled to analyze the replacement patterns observed in aligned sequences (Overington et al., 1990, Lüthy et al., 1991, Topham et al., 1993, Wako and Blundell, 1994). Because these tables were derived regardless of evolutionary considerations, applications of such matrices to answer phylogenetic questions becomes difficult. Another way to simplify matters is to consider codon position and chemical properties of the amino acids themselves such as charge, acidity and hydrophobicity as surrogates for experimentally determined protein structures (Naylor and Brown, 1997). Because of the low number and type of structures analyzed, the applicability of their method to the broader range of globular proteins would most certainly be diminished.

Protein Threading and Pseudo–Energy Potentials

A different approach to discovering sequence structure correlations is through protein threading. The idea is to evaluate the fit of a sequence to a known protein structure based on what is seen in a structural database. There are many methods to evaluate sequence structure fit, but some of the more successful are based on Boltzmann physics. This theory often arises when discussing the energetics of compatibility of protein sequences and related protein structures. In the past Boltzmann physics has been used to describe the energetics of protein conformations (MacArthur and Thornton, 1991, Serrano et al. 1992), with the existence of internal cavities (Rashin et al. 1986), of ion pairs (Bryant and Lawrence, 1991), or with amino acid location preferences (Miller et al. 1987, Koshi and Goldstein, 1998, Koshi et al. 1997, 1999).

The components and exact form of the true forces that stabilize protein structure may never be known explicitly. One approach to calculate protein stability is to determine the Gibbs free energy associated with a true protein. However, these methods typically require intense computations that may explain the interest in faster approximation strategies. Creating pseudo-energy potentials is one way to approximate Gibbs free energy, and it has been successfully applied to protein threading. For instance, Jones, Taylor and Thornton (1992b) created a set of pseudo-energy potentials using the Inverse Boltzmann Principle (IBP) as well as the Thermodynamic Hypothesis (Anfinsen, 1973). The Thermodynamic Hypothesis postulates that the native conformation of a particular protein sequence is that which gives the lowest energy. As it is applied here, the IBP allows the approximation of the energy associated with solvation and pairwise interactions based on the frequency of occurrence in a database. Combined, these relations assume that if a certain conformation is more frequent, it must therefore be stable and thus will be given a low energy. Furthermore, by evaluating the fit of a particular sequence on all known conformations, the most likely native state is then assumed to be the one that gives the lowest energy (Jones, Taylor and Thornton, 1992b).

The potentials derived using the IBP are commonly referred to as knowledge-based potentials of mean force because they contain the information from the diverse assortment of crystal structures from which they were derived. Instead of an exact measure of protein stability, using the IBP determines a measure of how likely a sequence adopts the given fold based on how similar the interactions compare to those most often observed in the structural database of real proteins (Jones and Thornton 1996). To avoid introducing anomalies into the training set, the rarely crystallized non-globular transmembrane proteins, as well as those proteins with unusual prosthetic groups were excluded (Hendlich et al., 1990).

More data often translates to parameter estimates with lower variance. In this vein, the addition of unique crystal structures to the training set should increase the ability of a model to distinguish incorrect conformations from the true native state of the protein as well (Hendlich et al., 1990). Unfortunately, the true forces that stabilize protein structures are not known. Hence, even with an arbitrarily large database there is no guarantee that the calculated energies converge to the true stabilization forces of interaction. For a more complete discussion of the drawbacks of empirical energy potentials along with a test on a hypothetical lattice model protein see Thomas and Dill (1996a, 1996b).

One way in which researchers characterize sites within a protein structure is through measures of solvent accessibility. Because X-ray crystallography is sometimes a tedious and unsuccessful endeavor, many have focused their efforts on predicting solvent accessibility measures for sites in proteins whose structures have not been determined (e.g., Pascarella et al. 1998; Pollastri et al. 2002; Rost and Sander 1994; Thompson and Goldstein 1996, 1997). However, if the structure of one sequence in a dataset is known, a direct measure of solvent accessibility for each site can be found using the Dictionary of Secondary Structure of Proteins (DSSP) program (Kabsch and Sander, 1983). Assuming that alignment columns contain structurally homologous positions, the solvent accessibility measure can then be extrapolated across taxa.

To obtain the percent solvation for each site the raw solvation value from DSSP must be normalized using the associated amino acid's relative maximum accessibility. The maximum value for each residue is estimated using a fully extended pentapeptide GG(X)GG as the reference state mimics the configuration for which the residue attains its highest accessibility in a folded protein (see Appendix A) (Jones, Taylor and Thornton, 1992b). The normalized values are then subdivided into five solvation categories ranging from very low values representing sites that are completely buried in the protein core, to high values representing sites that are fully exposed (David Jones, personal communication). The cutoff values for each solvation category, γ_s, are given in Appendix B.

Within a protein crystal structure, the solvation category of each site is automatically fixed. When evaluating the fit of an non-native sequence to a structure, often times the amino acid residue at a site is not equivalent to the reference residue. Because of differences in side chain volume the exact percent solvation is likely to be slightly different than the given value. Because the solvation categories span an interval however, it seems reasonable to assume that the actual solvent accessibility value of the threaded residue will be contained in the given interval.

When examining protein structures, certain substructures, such as $\alpha-$helices and $\beta-$pleated sheets, are common occurrences even within diverse protein folds. Their high frequency corresponds to highly stable energy values as calculated using the IBP. If protein structure is considered in terms of pairwise inter-atomic distances, one can encompass all aspects of secondary structure as favorable interactions between pairs of amino acid residues that lie in close proximity in folded proteins. Several methods accomplish this by defining pairwise interactions among amino acids using pseudo–energy potentials (e.g., Hendlich et al., 1990, Jones, Taylor and Thornton, 1992b, Kocher et al., 1994, Miyazawa and Jernigan, 1996, Park et al., 1997). However, Sippl (1990) was the first to derive energy potentials for the interactions of all amino acid residue pairs as a function of interatomic Euclidean distance.

By analyzing the set of highly resolved, non-homologous protein crystal structures in the PDB database, counts of all possible amino acid residue pairwise interactions were tabulated (Jones, Taylor and Thornton, 1992b). In their procedure, two residues are defined to interact if their C_β carbon atoms are within 10 Angstroms of one another. Because Glycine is the only amino acid that does not have a true C_β atom, a fictitious C_β atom can be constructed using the spatial geometry of the other atoms in its backbone (David Jones personal communication). The Jones pairwise potentials were derived for five atom pairs, namely, the $C_\beta \Rightarrow C_\beta, C_\beta \Rightarrow N, C_\beta \Rightarrow O, O \Rightarrow C_\beta, N \Rightarrow C_\beta$.

The behavior of the pairwise energy potential is like that of gravitational force. If two residues lie in close proximity, there is only one potential defined for their interaction. The Nitrogen terminus has a slight positive charge, while the Carboxyl end has a slight negative charge, giving the overall residue an electrostatic gradient ($N^+ \Rightarrow COOH^-$). Because of this polarity, instead of making symmetric counts (Dayhoff et al. 1978) only observed counts between two residues if the initial residue fell prior to the second in the order along the protein chain were tabulated (Sippl, 1990). The equation used to derive all pairwise energy potentials used in GenTHREADER (Jones, 1999) is derived in nice detail in Hendlich et al. (1990).

The resultant counts for each atom pair are subdivided into three main categories according to their sequence separation along the protein chain. Short range interactions are those for which amino acids are separated by less than 12 residues along the chain and account for some secondary conformations such as α - helix formation. Medium range interactions are those for which amino acids are separated by between 12 and 22 residues, and account for substructures such as amphipathic helices and β - sheet motifs. Finally, long range interactions are those for which amino acid residue separation is greater than 22 residues, and account for the overall tertiary structure interaction and final protein conformation (Jones, Taylor and Thornton, 1992b). Although separated by substantial distances along the protein chain, long range interactions do occur, because of residues lying in close proximity in the final complex protein conformation.

Conclusion

The history given in the introduction chronicles the state of molecular evolution for protein coding genes as the topic itself has evolved over the years. The models considered assume independence among sites or codons, which is unfortunately not biologically plausible. In order to maintain functionality of the folded protein state, residues must interact. Relaxing the independence assumption comes with a tremendous increase in computation which in the not so distant past would have been intractable. Recently, computer processor speed and advances in Bayesian statistical modelling, such as Markov chain Monte Carlo methods, have increased to the point where such models can now be contemplated. The following chapters detail the derivation and implementation of one such model, whose general framework may have broad applicability in the field of molecular evolution.

References

Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science* **181**(96):223-30

Bernstein, F.C., T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Simanouchi, M. Tasumi (1977) The protein data bank: A computer based archival file macromolecular structures. *J. Mol. Biol.* **112**:535-542

Brown, G.G., and M.V. Simpson (1982) Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II gnes. *Proc. Natl. Acad. Sci.* **79**:3246-3250

Bryant, S.H., and C.E. Lawrence (1991) The frequency of ion pair substructures in proteins in quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. *Proteins* **9**:108-119

Chothia, C. and A.M. Lesk (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**:823-826

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. (1978) A model of Evolutionary Changes in Proteins. Pp. 345-352 in *Atlas of Protein Sequence and Structure* vol. **5**, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

Dill, K.A. (1990a) Dominant forces in protein folding. *Biochemistry* **29**(31):7133-7155

Dill, K.A. (1990b) The meaning of hydrophobicity. *Science* **250**(4978):297-298

Edwards, A.W.F., and L.L. Cavalli-Sforza (1963) The reconstruction of evolution. *Annals of Human Genetics* **27**:105-106

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368 - 376

Felsenstein, J. (1989) Phylogenetic Inference Package (PHYLIP), version 3.2. University of Washington, Seattle. *Cladistics* **5**:164-166

Flores, T.P., C.A. Orengo, D.S. Moss, and J.M. Thornton (1993) Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**:1811-1826

Goldman, N., J.L. Thorne, and D.T. Jones (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263**:196-208

Goldman, N., J.L. Thorne, and D.T. Jones (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**:445-458

Goldman, N., and Z. Yang. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol.Biol.Evol.* **11**(5):725-736

Gonnet, G.H., M.A. Cohen, and S.A. Benner (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**:1443-1445

Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science* **185**:862-864

Hasegawa, M., H. Kishino and T. Yano (1985) Dating of the human - ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **21**, 160 - 174

Hendlich, M., P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M.J. Sippl (1990) Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.* **216**:167-180

Henikoff S., and J.G. Henikoff (1992) Amino acid substitution matrices from protein blocks. *PNAS* USA. **8**, 10915 - 10919

Huelsenbeck, J.P. (1995) Performance of phylogenetic methods in sumulations. *Syst. Biol.* **44**(1):17-48

Huelsenbeck, J.P., and B. Rannala (1997) Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276**:227-231

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**:860-921

JONES, D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**:797-815

JONES DT, TAYLOR WR, THORNTON JM. (1992a) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* Jun; **8**(3):275-82

JONES, D.T., W.R. TAYLOR, AND J.M. THORNTON (1992b) A new approach to protein fold recognition. *Nature* **358**:86-89

JONES, D.T., AND J.M. THORNTON (1996) Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**:210-216

JUKES, T.H., AND CANTOR, C.R. (1969) Evolution of protein molecules. *Mammalian protein metabolism.* (Munro, H.N., ed Pp. 21 - 32 Academic Press, New York,

KABSCH, W., AND C. SANDER (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen - bonded and geometrical features. *Biopolymers* **22**(12):2577-2637

KARLIN, S. A first course on stochastic processes. Academic Press, New York: 1966

KIMURA, M. (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111 - 120

Kishino, H., T. Miyata, and M. Hasegawa (1990) Maximum likelihood inference of protein phylogeny and the origen of chloroplasts. *J. Mol. Evol.* **3**:151-160

Kocher, J.-P.A., M.J. Roomin, and S.J. Wodak (1994) Factors influencing the ability of knowledge based potentials to identify native sequence - structure matches. *J. Mol. Biol.* **235**:1598-1613

Koshi, J.M., D.P. Mindell and R.A. Goldstein (1997) Beyond mutation matrices: physical chemistry based evolutionary models. p.80-89 in S. Miyano and T. Takagi, eds. *Genome Informatics* (1997) Universal Academy Press, Tokyo

Koshi, J.M., D.P. Mindell and R.A. Goldstein (1999) Using physical chemistry based substitution models in phylogenetic analyses of HIV-1 subtypes. *Mol. Biol. Evol.* **16**(2):173-179

Koshi, J.M., and R.A. Goldstein (1995) Context dependent optimal substitution matrices. *Prot. Eng.* **8**:641-645

Koshi, J.M. and R.A. Goldstein (1997) Mutation matrices and physical - chemical properties: correlations and implications. *Proteins* **27**:336-344

Koshi, J.M., and R.A. Goldstein (1998) Models of natural mutations including site heterogeneity. *Proteins* **32**:289-295

Li, W.-H., C.-I.Wu, and C.-C. Luo (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**(2):150-174

Liò, P., N. Goldman, J.L. Thorne, and D.T. Jones (1998) PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* **14**(8):726-733

Liò, P. and N. Goldman (1998) Modles of molecular evolution on phylogeny *Genome Res.* **8**:1233-1244

Lüthy, R., A.D. McLachlan, and D. Eisenbery (1991) Secondary structure-based profiles: use of structure conserving scoring tables in searching protein sequence database for structurally similarities. *Proteins* **10**:229-239

MacArthur, M.W., and J.M. Thornton (1991) Influence of proline residues on protein conformation. *J. Mol. Biol.* **218** :397-412

Meselson, M., and F.W. Stahl (1958) The replication of DNA in E. coli. *Proc. Natl. Acad. Sci.* USA **44**:671-682

Miller, S., J. Janin, A.M. Lesk, and C. Chothia (1987) Interior and Surface of Monomeric Proteins. *J. Mol. Biol.* **196**:641-656

Miyazawa, S. and R.L. Jernigan (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**:623-644

Muse, S.V., and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715 - 724

Naylor, G. and W.M. Brown (1997) Structural biology and phylogenetic estimation. *Nature* **388**:527-528

Nei, M, and T. Gojobori (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**(5):418-426

Overington, J., M.S. Johnson, A. Sali, and T.L. Blundell (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structural prediction. *Proc. R. Soc. Lond. B* **241**:132-145

Overington, J., D. Donnelly, M.S. Johnson, A. Sali, and T.L. Blundell (1992) Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein. Sci.* **1**:216-226

Park, B.H., E.S. Huang and M. Levitt (1997) Factors affecting the ability of energy functions to discrimenate correct from incorrect folds. *J. Mol. Biol.* **256**:831-846

Parisi G. and J. Echave (2001) Structural Constraints and Emergence of Sequence Patterns in Protein Evolution. *Mol. Biol. Evol.* **18**(5):750-756.

Pascarella, S., R. De Persio, F. Bossa, and P. Argos (1998) Easy method to predict solvent accessibility from multiple protein sequence alignments. *Proteins* Aug 1; **32**(2): 190-199

Pedersen A.-M. K. and J.L. Jensen (2001) A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames. *Mol. Biol. Evol.* **18**(5):763-776.

Pedersen, A.-M.K., C. Wiuf, and R.B. Christiansen (1998) A codon - based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15**(8):1069-1081

Pollastri, G., P. Baldi, P. Farislelli and R. Casadio (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* May 1; **47**(2):142-153

Purves, W.K., G. H. Orians, and H.G. Heller. (1992) *Life: The science of biology. third edition.* Sinauer Associates, Inc. Sunderland, MA.

Rashin, A.A., M. Iofin, B.Honig (1986) Internal cavities and buried waters in globular proteins. *Biochem.* **25**: 3619-3625

Rost B. and C. Sander (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins* Nov; **20**(3): 216-226

Russell, R.B., M.A.S. Saqi, R.A. Sayle, P.A. Bates, and M.J.E. Sternberg (1997) Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269**:423-439

Serrano, L., J. Sancho, M. Hirshberg, A.R. Fersht (1992) Alpha Helix stability in Proteins: I. Empirical correlations concerning substitutions of side chains at the N and C-caps and the replacement of Alanine by Glycine or Serine at solvent-exposed surfaces. *J. Mol. Biol.* **227**:544-559

Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins *J. Mol. Biol.* **213**:859-883

Sippl, M.J. (1993) Recognition of errors in three - dimensional structures of proteins. *PROTEINS: Structure, Function and Genetics* **17**:355-362

Snustad, D.P. and M.J. Simmons (2000) *Principles of Genetics. Second Edition.* John Wiley and Sons, Inc. New York

Swofford, D.L., Olsen, G.J., Waddel, P. J. and Hillis, D. M. (1996),"Phylogenetic Inference", In *Molecular Systematics*, 2nd ed., (Hillis, D.M., Moritz, C. and Mable, B. K. eds.),407–514. Sinauer, Sunderland.

Tavare′, S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures in matematics inthe life sciences,* vol. 17 (ed. R.M. Miura), pp. 57-86, American Matematical Society, Providence, RI

Taylor, W.R. and D.T. Jones (1993) Deriving an amino acid distance matrix. *J. Theor. Biol.* **164**:65-83

Thomas, P.D. and K.A. Dill (1996a) Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **257**: 457-469

Thomas, P.D. and K.A. Dill (1996b) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci.* **93**(21):11628-11633

Thompson, M.J., and R.A. Goldstein (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins* May; **25** (1): 38-47

Thompson, M.J., and R.A. Goldstein (1997) Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Sci.* Sep; **6**(9) : 1963-1975

Thorne, J.L. (2000) Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* **10** :602-605

Thorne, J.L., and N. Goldman (2001) Probabilistic models for the study of protein evolution. *Handbook of Statistical Genetics.* ed. D.J. Balding et al. John Wiley and Sons, Ltd.

Thorne, J.L., N. Goldman and D.T. Jones (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**(5):666-673

Topham, C.M., A. McLeod, F. Eisenmenger, J.P. Overington, M.S. Johnson, and T.L. Blundell (1993) Fragment ranking in modelling of protein structure: conformationally constrained substitution tables. *J. Mol.*

Biol. **229**:194-220

VENTER, J.C. ET AL. (2001) The sequence of the human genome. SCIENCE **291**:1304-1351

WAKO, H., AND T.L. BLUNDELL (1994) Use of amino acid environment-dependent substitution tables and conformational properties in structural prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J. Mol. Biol.* **238**:682-692

WATSON, J.D., AND F.H.C. CRICK, (1953a) A structure for DNA. *Nature* **171**:737-738

WATSON, J.D., AND F.H.C. CRICK, (1953b) Genetic implications of the structure of DNA. *Nature* **171**:964-967

WHELAN, S., P. LIÒ, AND N. GOLDMAN (2001) Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends Genet.* **17**(5):262-272

YANG, Z. (1994a) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**:105-111

YANG, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**(5):555-556

YANG, Z., R. NIELSEN, AND M. HASEGAWA (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**(12):1600-1611

Yang, Z., R. Nielsen, N. Goldman, A.-M. K. Pedersen (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431-449.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39**:315-329

Zuckerkandl, E. and L. Pauling (1962) Molecular disease, evolution and genetic heterogeneity. In *Horizons in Biochemistry* (eds. M. Kasha and B. Pullman) Pp. 189-225. Academic Press, New York, N.Y.

Zuckerkandl, E. and L. Pauling (1965) Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (eds. V. Bryson and H. Vogel). Pp. 97-166. Academic Press, New York, N.Y.

Appendix A

Alanine	=	113	Leucine	=	179
Arginine	=	253	Lysine	=	215
Asparagine	=	167	Methionine	=	194
Aspartic Acid	=	167	Phenylalanine	=	226
Cysteine	=	140	Proline	=	151
Glutamic Acid	=	199	Serine	=	134
Glutamine	=	198	Threonine	=	148
Glysine	=	88	Tryptophan	=	268
Histidine	=	194	Tyrosine	=	242
Isoleucine	=	178	Valine	=	157

The relative solvent accessibility maximum values for each amino acid in a fully folded protein.

Appendix B

0	$\leq$	γ_1	$<$	12
12	$\leq$	γ_2	$<$	36
36	$\leq$	γ_3	$<$	44
44	$\leq$	γ_4	$<$	87
87	$\leq$	γ_5		

The five solvation categories range from fully buried to fully exposed. The form of the last solvation category is reflective of the fact that despite normalization, values may at times exceed 100 %, in particular at sites at the beginning and end of the protein sequence.

Chapter 2

PROTEIN EVOLUTION WITH DEPENDENCE AMONG CODONS DUE TO TERTIARY STRUCTURE

Robinson DM, DT Jones, H Kishino, N Goldman and JL Thorne (2003) Mol. Biol. Evol. 20(10):1692-1704.

Abstract

Markovian models of protein evolution that relax the assumption of independent change among codons are considered. With this comparatively realistic framework, an evolutionary rate at a site can depend both on the state of the site and on the states of surrounding sites. By allowing a relatively general dependence structure among sites, models of evolution can reflect attributes of tertiary structure. To quantify the impact of protein structure on protein evolution, we analyze protein-coding DNA sequence pairs with an evolutionary model that incorporates effects of solvent accessibility and pairwise interactions among amino acid residues. By explicitly considering the relationship between nonsynonymous substitution rates and protein structure, this approach can lead to refined detection and characterization of positive selection. Analyses of simulated sequence pairs indicate that parameters in this evolutionary model can be well estimated. Analyses of lysozyme c and annexin V sequence pairs yield the biologically reasonable result that amino acid replacement rates are higher when the replacements lead to energetically favorable proteins than when they destabilize the proteins. Although the focus here is evolutionary dependence among codons due to protein structure, the approach is quite general and could be applied to diverse cases of evolutionary dependence where surrogates for sequence fitness can be measured or modelled.

Introduction

Rates of amino acid replacement can vary both among protein positions (e.g., Yang, Nielsen and Hasegawa 1998) and among the types of amino acids involved in the replacement (e.g., Dayhoff, Schwartz and Orcutt 1978). Furthermore, rates of amino acid replacement at a protein site are likely to depend on the amino acids found at other positions in the protein. Although it is widely accepted that positions in a protein sequence do not evolve independently and although methods for detecting protein sites with correlated patterns of evolution have been proposed (e.g., Pollock, Taylor and Goldman 1999; Wollenberg and Atchley 2000), little progress has been made on incorporating dependence among sites into procedures for evolutionary inferences.

A notable exception to this lack of progress is found in studies by Pedersen and Jensen on evolutionary dependence among sites due to reading frame overlap (Jensen and Pedersen 2000; Pedersen and Jensen 2001). Here, we borrow ideas from Pedersen and Jensen, but focus on the evolutionary dependence among codons that is associated with protein tertiary structure. In addition, we build upon a recently proposed technique for simulating protein evolution (Parisi and Echave 2001). With this technique, the rate at which a site experiences change can be modified by substitutions at neighboring sites. Through simulations, Parisi and Echave convincingly demonstrated that their model incorporates information that is unobtainable with widely used models of protein evolution. For example, the simulation studies showed tendencies of certain amino acids to preferentially occupy certain sites in the left–handed β helix domain of UDP-N-acetylglucosamine acyltransferases. When a group of actual sequences with this helix domain was examined, qualitatively similar tendencies were observed.

Parisi and Echave began their simulations with a reference protein of known tertiary structure. They then selected a function to assign a distance between a protein sequence and the reference structure. Underlying the sequence–structure distance is the idea that protein tertiary structure evolves very slowly (Chothia and Lesk 1986; Flores et al. 1993). Therefore, the energy associated with the structure of an ancestral protein (e.g., the reference protein) and the energy associated with the structure of a descendant protein should be similar. The sequence–structure distance can be interpreted as a surrogate for the difference in energies between an ancestral and descendant protein.

When applying the Parisi–Echave procedure, most simulations that begin with an ancestral sequence will likely result in a descendant that differs from an observed descendant. Due to this apparent drawback, an alternative strategy is needed. Our approach differs from Parisi and Echave's work mainly because our goal is to perform statistical inference when studying the relationship between two observed sequences. It would be computationally inefficient to simulate sequence evolution and then discard all simulated descendant sequences that differ from those observed (though see Marjoram et al., submitted, for review and development of related methods).

Here, we explain our technique for statistical inference with evolutionary models that have dependence due to protein structure. With simulations, we show that this technique can obtain accurate parameter estimates. With analyses of actual sequence pairs, we show that our parameter estimates are biologically reasonable.

Although dependence due to protein structure is the focus here, slight modifications of our inferential procedure can be applied when the evolutionary dependence among sequence sites arises in other ways.

Modelling Protein Evolution Under Structural Constraints

Parameterization

Following Parisi and Echave (2001), we propose a Markov model to describe the evolution of proteins under structural constraints. This is accomplished with the creation of an instantaneous rate matrix, R, where the entry in row i and column j represents the rate of change from one sequence to another. Because we abandon the assumption of independent changes among codons, we cannot follow the conventional practice (e.g., Goldman and Yang 1994; Muse and Gaut 1994) by usefully expressing our model with a series of 61 x 61 rate matrices that each describe change at a specific codon location in the protein. For a DNA sequence of length N nucleotides and ignoring for now the possibility of stop codons, the dimensions of our rate matrix R are 4^N x 4^N. Our assumption that each substitution event changes only a single nucleotide residue reduces the maximum number of nonzero elements in each row of R to $3N + 1$. These $3N + 1$ elements comprise the diagonal entry that represents no change in sequence and one entry for each of the three possible substitutions at each of the N nucleotide sites.

The key motivation underlying our model is that nonsynonymous substitution rates should partially depend on whether the implied amino acid replacements would stabilize or destabilize the known and assumed fixed protein tertiary structure. To assess the effect of an amino acid replacement on protein stability, a measure is needed for how well the sequence fits the structure both before and after the replacement. If this measure indicates that the replacement would improve the sequence-structure fit, then the rate of the nonsynonymous change should be high. Likewise, if the sequence-structure fit would become more poor due to an amino acid replacement, then the corresponding nonsynonymous rate should be low.

Fortunately, systems for assessing the compatibility between sequence and structure have been developed for the purpose of protein fold recognition (e.g., see Jones and Thornton 1996). Our evolutionary model relies on a sequence-structure compatibility criterion that has been successfully applied to protein fold recognition by Jones, Taylor and Thornton (1992) and Jones (1999). This criterion can be split into two components, one component assessing solvent accessibility and the other assessing pairwise interactions between residues near to each other in 3-dimensional space. In our application we assume that protein tertiary structure is known and fixed, and consequently the values of these two components can be determined for a sequence by threading it onto the known structure. The solvent accessibility and pairwise measures of sequence-structure compatibility, respectively denoted by $E_s(i)$ and $E_p(i)$ for a Sequence i, are intended to be positively correlated with the free energy the protein has when folded into the known structure. Therefore, $E_s(i)$ and $E_p(i)$ have low (ideally negative) values when sequences and structures are relatively compatible. Rather than being actual energy potentials, the specific values of $E_s(i)$ and $E_p(i)$ are derived from statistical analysis of large numbers of known structures to assess the 'plausibility' of observing different amino acids at different degrees of solvent accessibility and observing different pairs of amino acids at different physical separations (Jones, Taylor and Thornton 1992; Jones 1999).

Except for the treatment of nonsynonymous rates, our parameterization is similar to that of widely used codon models. We include parameters π_A, π_G, π_C and π_T ($\pi_A + \pi_G + \pi_C + \pi_T = 1$ and all of these parameters are non-negative) so that mutations to the four nucleotide types need not be equally likely. Alternative modeling options that we have not yet pursued would have separate sets of these mutation frequency parameters for each of the three codon positions or separate parameters for each of the 61 codons (e.g., See Pedersen, Wiuf and Christiansen 1998). Our model contains the parameter $\kappa > 0$ because transitions and transversions may occur at different rates and contains the parameter u to scale the overall rate of change. To handle nonsynonymous rates, there are the parameters ω, s, and p which will be discussed in more detail below.

The instantaneous rate of change $R_{i,j}$ from Sequence i to Sequence j is set to 0 if Sequences i and j differ at more than one nucleotide or if Sequence j encodes a premature stop codon. For the other cases where Sequences i and j differ by exactly one nucleotide that has type h ($h \in \{A, G, C, T\}$) in Sequence j, our rate matrix entries are:

$$R_{i,j} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ \\ u\pi_h\kappa & \text{for a synonymous transition} \\ \\ u\pi_h\omega e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} & \text{for a nonsynonymous transversion} \\ \\ u\pi_h\kappa\omega e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} & \text{for a nonsynonymous transition} \end{cases}$$

When $s = p = 0$, our model simplifies to the sort of widely used codon model that has been studied by others (e.g., Muse and Gaut 1994; Goldman and Yang 1994). Biologically plausible values of both s and p are positive, for positive values of these parameters favor sequences with good fits to the known structure. The biologically unreasonable case where $s < 0$ and $p < 0$ would have evolution favoring sequences that do not fit the known structure well.

The s and p parameters reflect the contribution of nonsynonymous rates that comes from the sequence-structure fit, while the ω parameter is intended to capture contributions to nonsynonymous rates that are external to the protein of interest. A variety of external impacts on nonsynonymous rates can be envisioned. For example, the value of ω may be less than one if the protein being studied is part of a co-adapted system that might be disrupted when nonsynonymous changes cause the protein to function less well in the system. The value of ω could exceed one if, for example, nonsynonymous changes to the protein helped a pathogen evade the immune system of its host. We view this distinction between effects on nonsynonymous rates that are external to the protein (represented by ω) and effects on nonsynonymous rates that are internal to the protein (represented by

s and p) as being potentially useful for characterizing the process of molecular evolution. Although our implementation forces all codons to share the same ω value, it would be straightforward to adopt previously proposed strategies that allow ω to vary among sites (e.g., Yang et al. 2000).

Stationary Probabilities of Sequences

A nice feature of this model is the explicit form for the equilibrium distribution of each possible coding sequence of length N. For simplicity of notation, let θ represent all the parameters in the rate matrix R (i.e., $\theta = \{\kappa, \omega, s, p, u, \pi_A, \pi_C, \pi_G, \pi_T\}$) and use i_m to represent the nucleotide at position m in DNA sequence i. Sequence i has stationary probability

$$p(i|\theta) = \frac{e^{-2sE_s(i)-2pE_p(i)} \prod_{m=1}^{N} \pi_{i_m}}{\sum_k e^{-2sE_s(k)-2pE_p(k)} \prod_{n=1}^{N} \pi_{k_n}}, \tag{2.1}$$

where the sum in the denominator is over all possible sequences k that lack a premature stop codon. We note that the above formula resembles the Boltzmann distribution of statistical physics and the denominator resembles what is known in statistical physics as a partition function. Previously proposed models that have evolutionary independence among protein locations also take the Boltzmann form as a stationary distribution (e.g., Koshi, Mindell and Goldstein 1997; Koshi and Goldstein 1998; Koshi, Mindell and Goldstein 1999)

As with models that have independent evolution of codons (e.g., Goldman and Yang 1994; Muse and Gaut 1994), the stationary probability $p(i|\theta)$ does not depend on κ, ω, or u. Interestingly, if s and p are not both zero then the expected frequencies under stationarity of A, C, G, and T need not be close to π_A, π_C, π_G, and π_T. The model is also time reversible. The time reversibility property,

$$p(i|\theta)R_{i,j} = p(j|\theta)R_{j,i} \text{ for all } i \text{ and } j, \tag{2.2}$$

is computationally convenient because it allows the likelihood $p(i,j|\theta)$ for a data set with two sequences i and j to be computed with

$$p(i,j|\theta) = p(i|\theta)p(j|i,\theta) = p(j|\theta)p(i|j,\theta), \tag{2.3}$$

rather than by enumerating all possible common ancestral sequences of i and j (see Felsenstein 1981).

Sequence Path Densities

The relatively general dependence structure of our model does not facilitate calculation of $p(j|i,\theta)$. Transition probabilities for conventional models of sequence evolution are calculated by specifying the rate matrix among 61 codon states (or 4 nucleotide states or 20 amino acid states) and then exponentiating this matrix (see Swofford et al. 1996). The 4^N x 4^N dimension of our rate matrix is typically too large to make matrix exponentiation computationally tractable. Rather than matrix exponentiation, our strategy is to augment the (observed) sequence data i and j with a (unobserved) path ρ, or sequence of events, between the two sequences. Here, we will arbitrarily choose i to be ancestral and j to be the descendant. The sequence path ρ specifies how i is transformed to j on a branch of the evolutionary tree. The information within ρ includes both the times on the branch at which sequence changes occur and the nature of each sequence change event. Because there are actually an infinite number of possible sequence paths that could transform one sequence to another, our strategy is to randomly sample these sequence paths from an appropriate probability density.

Specifically, we adopt a Bayesian inference framework. We specify a prior density $p(\theta)$ for our parameters and then sample θ and ρ from their joint posterior density $p(\rho, \theta|i, j)$. This sampling of θ and ρ allows their joint posterior distribution to be approximated and thereby serve as the basis for Bayesian parameter estimation. We can express $p(\rho, \theta|i, j)$ as

$$p(\rho, \theta|i, j) = \frac{p(j, \rho|i, \theta)p(i|\theta)p(\theta)}{p(i, j)}. \tag{2.4}$$

The $p(i, j)$ term in the denominator is difficult to explicitly determine. However, $p(i, j)$ is not a function of θ and need not be calculated with the Markov chain Monte Carlo procedure described in the next section. The $p(j, \rho|i, \theta)$ term is governed by the rate matrix R. For a specific sequence path ρ from i to j, let q be the total number of nucleotide substitutions on the path and let $t(z)$ be the time of the z^{th} substitution where $t(0) = 0$ is the time at which the branch begins. For simplicity, the time at which the branch ends will be $t(q + 1) = 1$. Because the scaling parameter u in our model can be adjusted, the time at which the branch ends can always be considered to be 1. The sequence that exists immediately after the z^{th} substitution will be represented by $i(z)$. Let $i(0) = i$ and $i(q+1) = i(q) = j$. Therefore, the sequence path ρ is fully specified by $t(0), t(1), \ldots, t(q), t(q + 1)$ and $i(0), i(1), \ldots, i(q), i(q + 1)$.

Consider the rate of change away from a Sequence v to any other sequence of length N. This rate of change away from v will be denoted $R_{v,\bullet}$ where

$$R_{v,\bullet} = \sum_k R_{v,k},$$

and where the sum is over all sequences k that differ from v. In practice, $R_{v,\bullet}$ is feasible to calculate because $R_{v,k}$ must be zero unless k is one of the $3N$ sequences that differ from v at exactly one nucleotide. The time until v experiences a nucleotide substitution is exponentially distributed with parameter $R_{v,\bullet}$. Given that some substitution occurs, the probability that v is transformed to some sequence k is $R_{v,k}/R_{v,\bullet}$. Therefore,

$$\begin{aligned} p(j,\rho|i,\theta) &= (\prod_{z=1}^{q} \frac{R_{i(z-1),i(z)}}{R_{i(z-1),\bullet}} R_{i(z-1),\bullet} e^{-R_{i(z-1),\bullet}(t(z)-t(z-1))}) e^{-R_{i(q),\bullet}(t(q+1)-t(q))} \\ &= (\prod_{z=1}^{q} R_{i(z-1),i(z)} e^{-R_{i(z-1),\bullet}(t(z)-t(z-1))}) e^{-R_{i(q),\bullet}(t(q+1)-t(q))}, \qquad (2.5) \end{aligned}$$

where the final term $e^{-R_{i(q),\bullet}(t(q+1)-t(q))}$ represents the probability of no change in the time interval from the final substitution at $t(q)$ until the time $t(q+1)=1$ that the branch ends.

Metropolis-Hastings algorithm

Via the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970), we construct a Markov chain on the combined θ and ρ state space with stationary distribution $p(\rho,\theta|i,j)$. To implement our Markov chain Monte Carlo (MCMC) algorithm, we randomly pick initial state $(\rho^{(0)},\theta^{(0)})$ from the set of possibilities where $p(\rho,\theta|i,j)$ exceeds zero. As described below in detail, we then propose random new values ρ' and θ' conditional on the current values ρ and θ. The proposal density will be denoted $J(\rho',\theta'|\rho,\theta)$. With probability equal to the minimum of 1 and

$$r = \frac{p(\rho,\theta|i,j)\ J(\rho,\theta|\rho',\theta')}{p(\rho',\theta'|i,j)\ J(\rho',\theta'|\rho,\theta)}$$

$$= \frac{p(j, \rho' | i, \theta') \; p(i|\theta') \; p(\theta') \; J(\rho, \theta | \rho', \theta')}{p(j, \rho | i, \theta) \; p(i|\theta) \; p(\theta) \; J(\rho', \theta' | \rho, \theta)}, \tag{2.6}$$

we set the next state of our Markov chain ($\rho^{(1)}$, $\theta^{(1)}$) equal to the proposed state (i.e., $\rho^{(1)} = \rho'$, $\theta^{(1)} = \theta'$). Otherwise, $\rho^{(1)} = \rho^{(0)}$ and $\theta^{(1)} = \theta^{(0)}$. By repeatedly proposing ρ' and θ' and then randomly accepting the proposals with probabilities determined by Equation (2.6), a Markov chain with the desired stationary distribution $p(\rho, \theta | i, j)$ is formed.

Due to the sum over all sequences in the denominator of Equation (2.1), it is not computationally feasible to explicitly calculate the $p(i|\theta)$ or $p(i|\theta')$ terms in Equation (2.6). Therefore, we approximate the ratio $p(i|\theta')/p(i|\theta)$. Our approximation strategy relies on randomly sampling a group of M sequences of length N from the stationary distribution of sequences for parameter values $\theta^* = \{\kappa^*, \omega^*, s^*, p^*, u^*, \pi^*_A, \pi^*_C, \pi^*_G, \pi^*_T\}$. The M sequences are effectively independently sampled from this stationary distribution and will be denoted $\eta^{(1)}$, $\eta^{(2)}$, …, and $\eta^{(M)}$. The sampling of $\eta^{(h)}$ from $p(\eta^{(h)}|\theta^*)$ is achieved via construction of a Markov chain with a state space consisting of all sequences of length N. This Gibbs Sampling approach (Geman and Geman 1984) has the desired $p(\eta^{(h)}|\theta^*)$ as its stationary distribution and exploits the fact that the numerator of Equation (2.1) is straightforward to calculate even though the denominator may be computationally intractable. Consider four DNA sequences of length N that are identical except for the residue at one specific nucleotide site. The sequence with an A at this sole position will be denoted by v_A while the sequences with C, G, and T will be respectively denoted by v_C, v_G, and v_T. Because the denominator of Equation (2.1) need not be evaluated, it is easy to determine $p(v_\alpha|\theta^*)/(p(v_A|\theta^*)+p(v_C|\theta^*)+p(v_G|\theta^*)+p(v_T|\theta^*))$ for all $\alpha \in \{A, C, G, T\}$. Conditional upon $N-1$ of the N sequence positions, the residue at the sole position that is free to vary can therefore be randomly sampled according to its stationary probability.

The initial state of the Markov chain defined by our Gibbs Sampler is randomly selected from the set of all DNA sequences of length N that lack a stop codon. Thereafter, the Gibbs Sampler repeatedly cycles through two steps. The first step is to randomly select a site at which to propose a change. The second step is to fix the sequence at all positions except this site and then randomly choose the nucleotide to occupy the selected site according to the stationary probabilities of the four sequences that represent the four possible residues at this site. This Gibbs Sampler allows a sequence $\eta^{(h)}$ to be sampled with probability $p(\eta^{(h)}|\theta^*)$. Because we desire the M sequences $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(M)}$ to be effectively independent samples, it is important to simulate the Markov chain for a long time between the sampling of each of the M sequences.

An importance sampling argument (see Appendix 1) shows that when M is sufficiently large,

$$\begin{aligned}\frac{p(i|\theta')}{p(i|\theta)} &\doteq e^{-2(s'-s)E_s(i)-2(p'-p)E_p(i)} \left(\prod_{m=1}^{N} \frac{\pi'_{i_m}}{\pi_{i_m}}\right) \\ &\text{X} \left(\frac{\sum_{h=1}^{M} e^{-2(s-s^*)E_s(\eta^{(h)})-2(p-p^*)E_p(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}{\sum_{h=1}^{M} e^{-2(s'-s^*)E_s(\eta^{(h)})-2(p'-p^*)E_p(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi'_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}\right) \qquad (2.7)\end{aligned}$$

The quality of this approximation improves if θ^* is close to both θ' and θ. For this reason, the θ^* chosen for this approximation depends in our implementation on the values of θ and θ'. The ratio approximated in Equation (2.7) depends on the six parameters s, p, π_A, π_G, π_C, and π_T. Because $\pi_A + \pi_G + \pi_C + \pi_T = 1$, the values of these six parameters can be specified in a 5-dimensional space. We determine the possible values for θ^* by partitioning this 5-dimensional space into a grid. For a particular combination of θ and θ', we find their midpoint and then choose θ^* to be the grid point that is nearest this midpoint. Because not all grid points will be visited in a particular MCMC run, we save computation by not sampling the M sequences from $p(\eta^{(h)}|\theta^*)$ until a particular θ^* grid point is visited for the first time during the MCMC run.

Proposing θ

Our MCMC implementation actually consists of various proposal distributions $J(\rho', \theta'|\rho, \theta)$ and the Markov chain is formed by cycling through these different proposal distributions. Each proposal can result in only slight differences between (ρ, θ) and (ρ', θ'). Most of these proposal distributions lead to slight differences between θ' and θ but set $\rho' = \rho$. For example, one proposal distribution has $\rho' = \rho$ and $\theta' = \theta$ except that $\kappa' \neq \kappa$. We employ similar proposal steps that propose change to only s, only p, only u, or only ω. All of these proposal steps are Metropolis-Hastings schemes that involve sampling a proposed parameter value from a uniform distribution that is determined by the current parameter value, a prespecified "window" length surrounding the current parameter value, and any constraints on the parameters.

There is a separate proposal step for each of π_A, π_G, π_C, and π_T and these are also conventional in that the proposed value for π_A, π_G, π_C or π_T will involve sampling from some uniform distribution. However, these steps are slightly different from those for κ, s, p, u, and ω because the constraint that $\pi_A + \pi_G + \pi_C + \pi_T = 1$ necessitates that a change in one of these four parameters cannot be made without an accompanying total change of the opposite sign in the other three. Our implementation partitions this total change of the opposite sign among the other three parameters according to the relative size of the three parameters. For example, if $\pi'_A = \pi_A + \delta$ is obtained by sampling from a uniform distribution centered about π_A, then we set $\pi'_G = \pi_G - \delta\pi_G/(\pi_G + \pi_C + \pi_T)$, $\pi'_C = \pi_C - \delta\pi_C/(\pi_G + \pi_C + \pi_T)$, and $\pi'_T = \pi_T - \delta\pi_T/(\pi_G + \pi_C + \pi_T)$. If the three free parameters had been π'_A, $\pi'_C/(\pi'_C + \pi'_G + \pi'_T)$, and $\pi'_G/(\pi'_C + \pi'_G + \pi'_T)$ rather than π'_A, π'_C, and π'_G, then the proposal density $J(\rho', \theta'|\rho, \theta)$ would be the uniform density for δ because only π'_A would be changed with this proposal step. Because our parameterization is actually in terms of π'_A, π'_C, and π'_G, $J(\rho', \theta'|\rho, \theta)$ becomes the uniform density for δ multiplied by a factor $1/(1 - \pi'_A)^2$ that represents the Jacobian of the transformation from π'_A, $\pi'_C/(\pi'_C + \pi'_G + \pi'_T)$, and $\pi'_G/(\pi'_C + \pi'_G + \pi'_T)$ to π'_A, π'_C, and π'_G.

Proposing Site Paths

When proposing a new path ρ' different from ρ, all parameter values in θ are held constant. The sequence paths ρ and ρ' represent possible ways to transform an ancestral sequence i to a descendant j. A sequence path is fully specified by a series of nucleotide substitutions and the specific times at which the substitutions occur. For each possible sequence path, i must have a residue i_r at site r whereas

j must have a residue j_r at this site. A sequence path ρ can also be represented by a set of site paths ρ_1, ρ_2, $\ldots$, ρ_r, $\ldots$, ρ_N. A particular site path ρ_r specifies all nucleotide substitutions that change site r in path ρ, as well as the specific times of the changes. Our approach for proposing ρ' is to set ρ' equal to ρ with the exception of the site path at one randomly selected site. For this site, we base the sampling of the proposed site path ρ'_r on a simple model of evolution that assumes independence of nucleotide substitutions among sites. To emphasize that this simple model may have a parameterization that is quite different from the dependent sites model, ψ rather than θ will represent the independent sites model and its parameters. Our goal will be to sample a site path ρ_r for site r from $p(\rho_r|i_r, j_r, \psi)$.

Because of the relative ease of sampling the site paths ρ_r, we adopt what is commonly referred to as the Felsenstein 1984 model (see Felsenstein 1989). Nielsen (2001) has used a similar algorithm for sampling site paths from this model for a different application. Nielsen's algorithm, or an algorithm that simulates site paths beginning with i_r and then discards the paths unless they end with j_r, would be suitable for our application. However, our implementation uses a different algorithm for sampling site paths ρ_r from $p(\rho_r|i_r, j_r, \psi)$ (see Appendix 2). No matter what method is selected for sampling from $p(\rho_r|i_r, j_r, \psi)$, the resulting proposal ρ' should be accepted or rejected according to Equation (2.6).

Examples

Prior Densities and Implementation

In all analyses, all combinations of non-negative values for π_A, π_C, π_G, and π_T that satisfy $\pi_A + \pi_C + \pi_G + \pi_T = 1$ were treated as being equally likely *a priori*. Prior densities were uniform on the interval $(0, 4)$ for u, uniform on the interval $(0, 10)$ for κ, and uniform on the interval $(0, 10)$ for ω. Because so little is understood about the relationship between protein structure and protein evolution, the assignment of priors to the s and p parameters was somewhat but not completely arbitrary. Positive values of s and p represent the biologically reasonable scenario where evolution maintains sequence–structure compatibility. Negative values represent the scenario where evolution favors incompatibility. By centering prior distributions for s and p about zero, neither the compatibility nor the incompatibility of sequences and structures was favored *a priori*. With priors for s and p centered about zero, posterior distributions that are concentrated in the quadrant with s and p both exceeding 0 would be biologically reasonable and would therefore help to validate our approach. To set the endpoints of our uniform prior distributions for s and p, some exploration of posteriors was performed. We did not want the posterior distributions for s and p to be concentrated near the prior interval endpoints and, because the approximation of Equation (2.7) will be best for a finely meshed grid of θ^* values, we did not want the uniformly distributed intervals for s and p to be overly wide. We opted to analyze each pair of sequences with a prior for s that was uniform on the interval $(-2, 2)$ and a prior for p that was uniform on the interval $(-0.15, 0.15)$. This prior was chosen because posterior distributions of s and p were not concentrated near the endpoints of their respective prior intervals and because the grid mesh of θ^* proved fine enough to yield satisfactory approximations with Equation (2.7).

There is a tradeoff when determining the mesh size for the θ^* values. A mesh that is too wide will produce poor approximations with Equation (2.7). A mesh that is overly fine will be computationally infeasible. By selecting 9 points along the ranges of each of the s and p uniform priors to define the mesh size separating successive values of these parameters on the θ^* grid, we found an acceptable compromise between approximation accuracy and computational requirements. We performed some MCMC runs with a finer mesh and obtained similar results to those obtained when 9 points defined the grid for each of s and p. Posterior approximations when only 5 points defined the grid for each of s and p were not satisfactory. We did some exploratory analyses with a prior for s that was uniform on the interval $(-4, 4)$ and a prior for p that was uniform on the interval $(-0.3, 0.3)$. With this prior for s and p, 17 points were used to define the grid for each of s and p. We found that this latter prior for s and p yielded very similar results to those with the $(-2, 2)$ prior for s and $(-0.15, 0.15)$ prior for p (data not shown). The remaining three dimensions of the five dimensional θ^* grid represent π^*_A, π^*_C, and π^*_G. For these parameters, we have twelve points define the mesh along each of these dimensions. The grid points for these parameters are each evenly spaced in the interval from 0.14 to 0.36.

To enhance MCMC convergence, we have each MCMC cycle include one proposed update of s, p, u, κ, and ω but five proposed updates of site paths. Each $MCMC$ cycle also proposes an update to one of π_A, π_C, π_G, and π_T so that a proposed update to each of these parameters is made once per four cycles. Each analysis presented here consisted of 100,000 MCMC cycles. The first 5,000 of these cycles were treated as a "burn-in" period and did not contribute to the posterior approximations. For all cases, two separate MCMC runs were performed in order to check for convergence to the desired posterior distribution. Plots of sampled parameter values versus the cycle number at which samples were taken were also examined to check for convergence. The value of M, the number of sequences sampled from the stationary distribution represented by a particular grid point, was 1000 for the results presented here. The number of Gibbs iterations between successively sampled sequences was set to 10 multiplied by the length in nucleotides of the sequences.

It is conventional to express the amount of evolution separating two sequences in terms of the expected number of nucleotide substitutions per site. With our dependence model, the substitution rate at a site is determined by the residues occupying other sites, and can change when other sites change. To explore the prior distribution of the expected number of substitutions per site with our dependence model, we used a simulation procedure. For each simulation, values of the parameters that comprise θ were sampled from their prior distribution. Based on the resulting θ, the aforementioned Gibbs Sampling technique was employed to randomly select a sequence from its stationary distribution. This randomly sampled sequence was treated as being the ancestral sequence and then evolution was simulated according to our model to produce a descendant sequence. By computing the average number of changes per site for each simulated sequence path and by repeating the simulation procedure 10,000 times, the prior distribution for the expected number of changes per site was approximated to yield the results presented here.

Annexin V

We first illustrate our techniques by analysis of two different annexin V sequence pairs that each consist of 314 aligned codons. One pair includes the chicken and human annexin V sequences (Genbank Accession numbers M30971 and X12454) and the other pair consists of the mouse and rat sequences (Genbank Accession numbers NM_009673 and NM_013132). The two pairs of sequences represent nonoverlapping evolutionary lineages and substantially different divergence levels. The human-chicken pair is 78% identical at the amino acid level and 74% identical at the nucleotide level whereas the mouse–rat pair is 95% identical at the amino acid level and 93% identical at the nucleotide level.

Annexin V is a calcium dependent phospholipid binding protein that has potent vascular anticoagulant activity (Andree et al. 1990), is also an inhibitor of protein kinase C (Schlaepfer, Jones and Haigler 1992), and has been associated with the apoptotic pathway (Reutelingsperger and van Heerde 1997) and the initiation of the hepatitis B virus infection process (Neurath and Strick 1994). It was selected for this study as an unbiased, arbitrary example of a typical protein of biological interest. The tertiary structure of the chicken annexin V protein has been experimentally determined (Bewley et al. 1993) and we assume that the native conformations of all four annexin V sequences are essentially identical to this structure.

For these two annexin V sequence pairs, the posterior distributions of s and p are quite similar (Table 2.1). For both sequence pairs, the posteriors of s and p are concentrated in the $s > 0$ and $p > 0$ quadrant. This is the biologically plausible quadrant of the parameter space where evolution favors sequences that fold well into the known structure. The posterior distributions of κ, ω, and u were relatively unaffected by whether s and p were forced to be zero. This indicates that the information contributing to the s and p estimates is largely independent of the information contributing to the estimates of the κ, ω, and u parameters.

For biologically relevant values of s and p, the amino acid composition of sequences that are highly compatible with the tertiary structure may differ from the amino acid composition when $s = p = 0$. These amino acid composition differences may induce differences in expected nucleotide composition. This may explain why the estimates of π_A, π_C, π_G, and π_T were affected by whether or not s and p were forced to zero (Table 2.1).

To evaluate the potential performance of our estimation procedure, we simulated annexin V evolution using the posterior means of the model parameters as estimated from the human-chicken pair. For each of five simulated annexin V pairs, the values of $\kappa, \omega, u, \pi_A, \pi_C, \pi_G, \pi_T$, were set to their posterior means from our analysis of the original human-chicken data while the values of s and p varied among simulations. The posterior means of s and p for the human-chicken pair were approximately $s = 0.947$ and $p = 0.0282$. One pair of sequences was simulated for each of ($s = 0.947,\ p = 0.0282$), ($s = 0.947,\ p = -0.0282$), ($s = -0.947,\ p = 0.0282$), ($s = -0.947,\ p = -0.0282$), and ($s = 0,\ p = 0$). In each simulation, an ancestral sequence was selected via Gibbs Sampling from

		M. musculus vs. R. norvegicus		**G. gallus vs H. sapien**	
Param.	**Priors**	**($s = p = 0$)**	**($s \neq 0$, $p \neq 0$)**	**($s = p = 0$)**	**($s \neq 0$, $p \neq 0$)**
κ	5.0	3.414	3.360	1.788	1.587
	(0.25 , 9.75)	(1.803 , 6.220)	(1.803 , 5.779)	(1.279 , 2.467)	(1.124 , 2.189)
ω	5.0	0.135	0.147	0.0934	0.107
	(0.25 , 9.75)	(0.0694 , 0.229)	(0.0756 , 0.252)	(0.0651 , 0.129)	(0.0752 , 0.147)
s	0.0	0.0	0.881	0.0	0.947
	(-1.9, 1.9)	NA	(0.612 , 1.156)	NA	(0.704 , 1.199)
p	0.0	0.0	0.0375	0.0	0.0282
	(-0.143, 0.143)	NA	(0.0255 , 0.0510)	NA	(0.0156 , 0.0399)
u	2.0	0.148	0.151	1.579	1.604
	(0.1, 3.9)	(0.079 , 0.238)	(0.084 , 0.241)	(1.159 , 2.199)	(1.164 , 2.089)
π_A	0.25	0.272	0.290	0.300	0.310
		(0.246 , 0.300)	(0.259 , 0.322)	(0.274 , 0.325)	(0.284 , 0.339)
π_C	0.25	0.210	0.220	0.176	0.183
		(0.186 , 0.235)	(0.193 , 0.247)	(0.157 , 0.196)	(0.162 , 0.205)
π_G	0.25	0.279	0.294	0.250	0.263
		(0.252 , 0.306)	(0.265 , 0.326)	(0.226 , 0.274)	(0.239 , 0.290)
π_T	0.25	0.239	0.196	0.274	0.244
		(0.214 , 0.264)	(0.169 , 0.224)	(0.250 , 0.299)	(0.215 , 0.273)
BL	8.712	0.070	0.071	0.461	0.445
	(0.301 , 33.901)	(0.053 , 0.090)	(0.054, 0.091)	(0.371 , 0.566)	(0.361 , 0.537)

Table 2.1: Priors and Posteriors for the annexin V sequence comparisons. The second column shows prior means and 95% prior intervals for all parameters. The rightmost four columns show estimates of posterior means and 95% credibility intervals. Each sequence pair was analyzed both when s and p were allowed to vary and when both s and p were forced to be 0. The columns corresponding to each analysis are indicated. In the final rows of the table, "BL" represents branch lengths. Branch lengths are the expected number of changes per nucleotide site and were estimated as described in the text.

Param.	**Truth**	**Posterior Mean**	**95% Credibility Interval**
s	0	-0.045	(-0.127 , 0.226)
p	0	-0.002	(-0.00989 , 0.0135)
s	0.947	0.960	(0.718 , 1.205)
p	0.0282	0.0247	(0.0132, 0.0363)
s	-0.947	-0.926	(-1.105 , -0.757)
p	0.0282	0.0319	(0.0186 , 0.0449)
s	0.947	1.042	(0.807, 1.276)
p	-0.0282	-0.0352	(-0.0477, -0.0233)
s	-0.947	-0.908	(-1.079 , -0.740)
p	-0.0282	-0.0315	(-0.0445 , -0.0179)

Table 2.2: Posterior means and 95% credibility intervals for s and p with sequence pairs that were simulated according to our model and the annexin V tertiary structure.

the appropriate stationary distribution of sequences. A descendant sequence was then evolved according to our model and the resulting sequence pair was analyzed with the approach described here. For each of the five simulation scenarios, the posterior densities of s and p are concentrated close to the true values of these parameters (Table 2.2).

With our parameterization, a nonsynonymous nucleotide substitution that changes a Sequence i to a Sequence j will occur at the rate this substitution would have if it were synonymous multiplied by a factor of $\omega e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$. Therefore, if $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} > 1/\omega$, the change from Sequence i to Sequence j could be interpreted as involving positive selection. Likewise, the change from Sequence i to j could be interpreted as negatively selected if $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} < 1/\omega$. Among the nonsynonymous nucleotide substitutions that could occur at a particular codon in a particular sequence, some of these nonsynonymous substitutions may be positively selected and others may be negatively selected. The fact that some changes to a particular codon may be positively selected whereas others are negatively selected is a desirable property of our parameterization of nonsynonymous rates.

For both annexin V sequence pairs, we have explored variation due to protein structure among the possible nonsynonymous changes that could affect a sequence. For all 2058 possible sequences j that differ by one nonsynonymous substitution from the observed mouse sequence i, the rate factor associated with sequence–structure compatibility, $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$, was calculated. To produce the histogram of these 2058 values shown in Figure 2.1A, the posterior means of s and p from the mouse-rat comparison were used. A corresponding histogram generated with the 2069 possible nonsynonymous changes to the chicken sequence is similar and is not shown. Both histograms indicate that the nonsynonymous rate variation due to protein structure is substantial but both the mean and median effect of nonsynonymous substitutions on structural compatibility are deleterious. Although many of the possible nonsynonymous changes improve the sequence-structure compatibility (i.e., $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p} > 1$), this improved compatibility does not overcome the low posterior mean of ω by making any of these possible nonsynonymous changes positively selected.

To compare the impact of rates of solvent accessibility and pairwise interactions, Figure 2.1B plots $(E_s(i)-E_s(j))s$ versus $(E_p(i)-E_p(j))p$ for the 2058 values summarized in Figure 2.1A. This plot shows that relatively few nonsynonymous changes to the mouse sequence improve both the pairwise and solvent accessibility components of sequence–structure compatibility. The plot also indicates that sol-

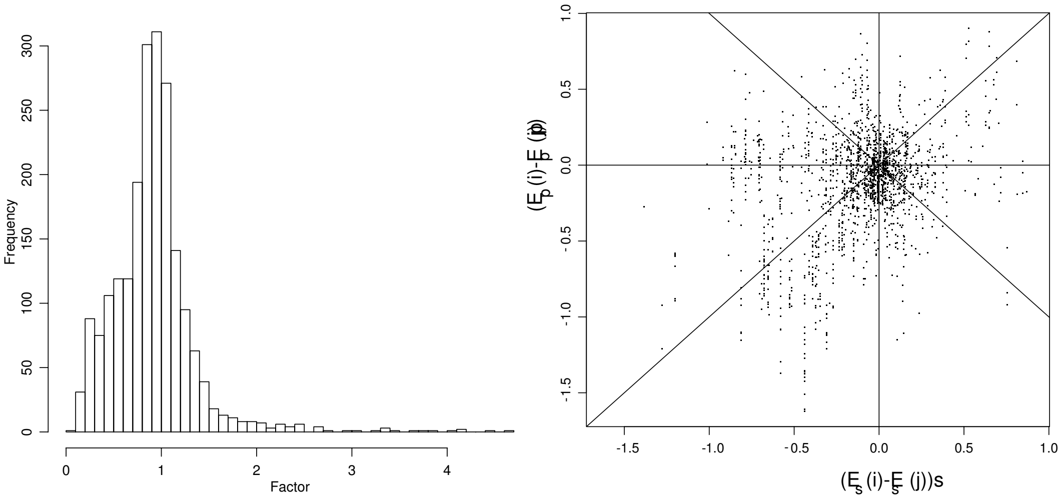


Figure 2.1: Figure 2.1A: Nonsynonymous rate variation due to structure in mouse annexin V. Letting mouse annexin V be sequence i, a histogram of the values of the factor $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$ is constructed from the 2058 sequences j that differ from i by exactly 1 nonsynonymous nucleotide substitution. The values of s and p used here are the posterior mean estimates from the mouse-rat comparison. Of the 2058 factors represented by the histogram, 713 factors exceed one. None of the 2058 factors exceeds 6.803, the inverse of the posterior mean of the ω parameter, and therefore none of the possible nonsynonymous substitutions is categorized as positive selection. Figure 2.1B: Comparison for mouse annexin V of effects of solvent accessibility and pairwise interactions on nonsynonymous rates. For each of the 2058 possible nonsynonymous changes to mouse annexin V that are summarized in Figure 2.1A, the value of $(E_s(i) - E_s(j))s$ is plotted versus the value of $(E_p(i) - E_p(j))p$. The $x = y$ and $x = -y$ lines are drawn to assist comparison between effects of pairwise interactions and solvent accessibility.

vent accessibility and pairwise interactions can both have substantial impacts on nonsynonymous rates.

Lysozyme

In their pioneering work on adaptation, Stewart and collaborators (e.g., Stewart, Schilling and Wilson 1987; Messier and Stewart 1997; see also Yang 1998; Yang and Nielsen 2002) demonstrated that the evolutionary lineage leading to the Colobine monkeys (e.g., *Colobus guereza*) has experienced an excess of nonsynonymous sub-

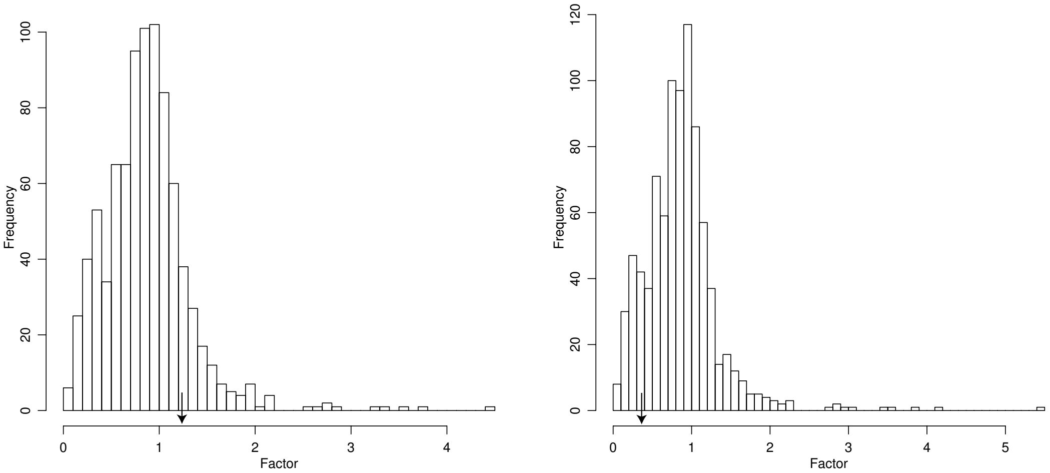


Figure 2.2: Figure 2.2A: Nonsynonymous rate variation due to structure in human lysozyme c. Letting human lysozyme c be sequence i, a histogram of the values of the factor $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$ is constructed from the 862 sequences j that differ from i by exactly 1 nonsynonymous nucleotide substitution. The values of s and p here are the posterior mean estimates from the human-callithrix comparison. Of the 862 factors represented by the histogram, 276 factors exceed one. The arrow at 1.233 shows the inverse of the posterior mean of the ω parameter. Therefore, the 115 possible nonsynonymous changes to the human sequence that exceed 1.233 can be categorized as positive selection. Figure 2.2B: Nonsynonymous rate variation due to structure in the Rhesus macaque lysozyme c. Letting Rhesus macaque lysozyme c sequence be sequence i, a histogram of the values of the factor $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$ is constructed from the 872 sequences j that differ from i by exactly 1 nonsynonymous nucleotide substitution. The values of s and p here are the posterior mean estimates from the Colobus-Rhesus comparison. Of the 872 factors represented by the histogram, 264 factors exceed one. The arrow at 0.377 shows the inverse of the posterior mean of the ω parameter. Therefore, the 754 possible nonsynonymous changes to the Rhesus macaque sequence that exceed 0.377 can be categorized as positive selection.

stitutions during lysozyme c evolution. The apparent explanation for this excess is that the Colobine monkeys have acquired a foregut in which bacteria ferment leaves. In the Colobine monkeys, lysozyme c is active at a lower PH and is more resistant to pepsin than in other primates. These properties of lysozyme c in Colobine monkeys may be adaptations that help with lysis of bacteria that pass from the foregut fermentation chamber into the true stomach (Stewart, Schilling and Wilson 1987; Messier and Stewart 1997).

To investigate lysozyme c evolution with our approach, we selected two pairs of lysozyme c sequences that represent nonoverlapping evolutionary lineages and that each consist of 130 aligned codons. The phylogenetic path separating the Rhesus macaque (*Macaca mulatta*) and *Colobus guereza* pair (Genbank accession numbers X60236 and U76916, 96% nucleotide identity and 91% amino acid identity) includes lineages that have previously been demonstrated (Stewart, Schilling and Wilson 1987; Messier and Stewart 1997; Yang and Nielsen 2002) to have experienced a strong excess of nonsynonymous substitutions. The phylogenetic path separating the marmoset (*Callithrix jacchus*) and human pair (Genbank accession numbers M19045 and U76923, 93% nucleotide identity and 87% amino acid identity) does not seem to have experienced such a strong excess of nonsynonymous change (Stewart, Schilling and Wilson 1987; Messier and Stewart 1997; Yang and Nielsen 2002), although our estimates indicate it has a ratio of nonsynonymous to synonymous rates that is substantially higher than for the annexin V pairs (see below). The tertiary structure of the human lysozyme c protein has been experimentally determined (Harata, Abe and Muraki 1998) and we assume that the native conformations of all four lysozyme c sequences are essentially identical to this structure.

As with the annexin V analyses, the posterior distributions of s and p are concentrated in the biologically plausible quadrant of the parameter space for both lysozyme c sequence pairs (Table 2.3). Figure 2.2A shows a histogram that summarizes the values of $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$ for the 862 sequences j that differ by exactly one nonsynonymous change from the human lysozyme c sequence. The values of s and p used to construct this histogram were those from the human-marmoset comparison.

		Human vs. Marmoset		**Colobus vs. Rhesus macaque**	
Param.	**Priors**	**($s = p = 0$)**	**($s \neq 0$, $p \neq 0$)**	**($s = p = 0$)**	**($s \neq 0$, $p \neq 0$)**
κ	5.0	4.630	4.274	4.599	3.958
	(0.25 , 9.75)	(1.998 , 9.027)	(1.813 , 8.385)	(1.620 , 9.154)	(1.351 , 8.622)
ω	5.0	0.741	0.811	2.675	2.653
	(0.25 , 9.75)	(0.302 , 1.577)	(0.349 , 1.683)	(0.626 , 8.026)	(0.717 , 7.081)
s	0.0	0.0	0.855	0.0	0.877
	(-1.9 , 1.9)	NA	(0.483 , 1.251)	NA	(0.498 , 1.272)
p	0.0	0.0	0.0495	0.0	0.0520
	(-0.1425 , 0.1425)	NA	(0.0326 , 0.0664)	NA	(0.0352 , 0.0708)
u	2.0	0.075	0.081	0.021	0.025
	(0.1 , 3.9)	(0.027 , 0.160)	(0.030 , 0.166)	(0.004 , 0.064)	(0.006 , 0.071)
π_A	0.25	0.307	0.340	0.325	0.365
		(0.265 , 0.352)	(0.286 , 0.405)	(0.282 , 0.370)	(0.311 , 0.430)
π_C	0.25	0.166	0.178	0.162	0.174
		(0.134 , 0.202)	(0.142 , 0.218)	(0.128 , 0.199)	(0.138 , 0.213)
π_G	0.25	0.279	0.299	0.270	0.289
		(0.236 , 0.323)	(0.254 , 0.347)	(0.230 , 0.312)	(0.246 , 0.334)
π_T	0.25	0.247	0.183	0.242	0.172
		(0.208 , 0.290)	(0.134 , 0.231)	(0.202 , 0.283)	(0.124 , 0.219)
BL	8.609	0.082	0.083	0.048	0.049
	(0.262 , 34.564)	(0.055 , 0.115)	(0.055 , 0.116)	(0.028 , 0.073)	(0.029 , 0.075)

Table 2.3: Priors and Posteriors for the Lysozyme c sequence comparisons.

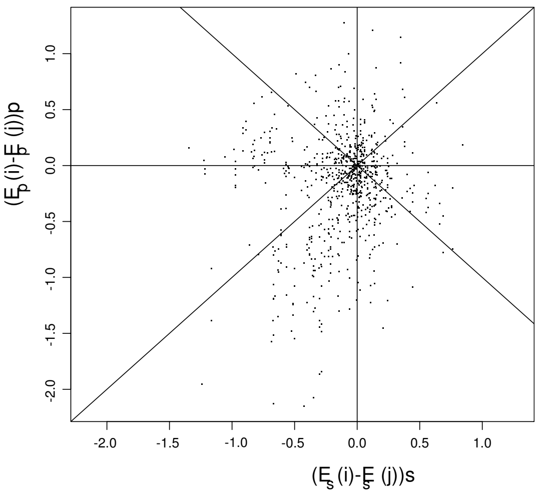


Figure 2.3: Figure 2.2C: Comparison for human lysozyme c of effects of solvent accessibility and pairwise interactions on nonsynonymous rates. For each of the 862 possible nonsynonymous changes to human lysozyme c that are summarized in Figure 2.2A, the value of $(E_s(i) - E_s(j))s$ is plotted versus the value of $(E_p(i) - E_p(j))p)$. The $x = y$ and $x = -y$ lines are drawn to assist comparison between effects of pairwise interactions and solvent accessibility.

Figure 2.2B shows a histogram that summarizes the values of $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$ for the 872 sequences j that differ by exactly one nonsynonymous change from the Rhesus macaque lysozyme c sequence.

The values of s and p used to construct this histogram were those from the colobus-rhesus sequence comparison. As indicated by the portions of these histograms to the right of the arrows, a substantial number of possible nonsynonymous substitutions would be positively selected. In comparison, none of the possible nonsynonymous annexin V substitutions represented in the histograms of Figure 2.1 would be positively selected. This disparity in fraction of positively selected nonsynonymous substitutions can be mainly attributed to the higher estimates of ω for the two lysozyme c sequence pairs than for those obtained from annexin V.

Figure 2.2C plots $(E_s(i) - E_s(j))s$ versus $(E_p(i) - E_p(j))p$ for the 862 possible nonsynonymous changes to human lysozyme c. This plot has a pattern that is qualitatively similar to the annexin V pattern in Figure 2.1B. The plot for the 872 possible nonsynonymous changes to the Rhesus macaque lysozyme c is also qualitatively similar to Figure 2.1B and is not shown.

Future Directions

An extension of our inferential procedure to data sets with more than two sequences should not be difficult. The main modification will be with how site paths are proposed. With more than two sequences, a site path will traverse multiple branches. Pupko et al. (2000) have introduced a fast algorithm for finding the optimal joint ancestral sequence reconstruction with simple independent sites models of evolution. A variation of this algorithm will allow joint ancestral sequence reconstructions to be sampled according to probabilities defined by the Felsenstein 1984 model. Given a set of reconstructed ancestral nodes for a site, the problem of proposing site paths for multiple sequences becomes the smaller problem of proposing site paths for each branch that comprises the tree. We have already addressed this smaller problem here.

An extension to multiple sequences may lead to improved methods for ancestral sequence reconstruction. With our approach, ancestral sequences that do not fold well into a tertiary structure are unlikely to be inferred. In addition, the paths from ancestral to descendant sequences may allow the order of adaptive nucleotide substitution events during protein evolution to be inferred.

Other methods for detecting positive selection (e.g., Yang et al. 2000; Yang and Nielsen 2002) treat each codon independently and ignore protein structure. In reality, the question of whether a particular nonsynonymous substitution confers a selective advantage should depend to some extent on the amino acids encoded by other codons. Our technique has the advantages of incorporating and quantifying this dependence.

Although the sequence–structure compatibility criterion employed here (Jones, Taylor and Thornton 1992; Jones 1999) has demonstrated considerable success when applied to protein fold recognition, other sequence–structure compatibility functions have been proposed (e.g., Singh, Tropsha and Vaisman 1996) and it will be interesting to explore which of these criteria is most appropriate for describing the process of molecular evolution. Incorporation of alternative criteria would necessitate few changes to our approach.

The sequence–structure compatibility criterion can be viewed as a surrogate for fitness. Measures of the fitness of a sequence that are not explicitly derived from protein structure could be incorporated into our approach without major modifications of our inference procedure. Such modifications might be of particular interest for studying bacteriophage or retrovirusses because these quickly evolving organisms are good systems for experimentally measuring the fitness consequences of particular nucleotide substitutions (e.g., Bull, Badget and Wichman 2000). For making inferences regarding history and evolutionary process, our statistical approaches could supplement more direct experimental information.

Acknowledgments

We thank Kevin Scott, Stefan Bekiranov, Reinhard Hopperger, and Elizabeth A. Thompson for their help. This work was supported by BIRD of the Japan Science and Technology Co. (H.K. and J.L.T.), J.S.P.S. (H.K.). N.S.F. INT 99-0934 (D.M.R. and J.L.T.), N.S.F. DEB-0120635 (J.L.T.), and a Wellcome Trust Senior Fellowship in Basic Biomedical Sciences (N.G.).

References

Andree, H. A. M., C. P. M. Reutelingsperger, R. Hauptmann, H. C. Hemker, W. Th. Hermens and G. M. Willems. 1990. Binding of vascular anticoagulant α (VAC α) to planar phospholipid bilayers. J. Biol. Chem. 265(9):4923-4928

Bewley, M. C., C. M. Boustead, J. H. Walker, D. A. Waller, and R. Huber. 1993. Structure of chicken annexin V at 2.25-A resolution. Biochemistry 32(15):3923-9

Bull, J. J., M. R. Badget, and H. A. Wichman. 2000. Big-benefit mutations in a bacteriophage inhibited with heat. Mol. Biol. Evol. 17:942-950

Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. EMBO. J. 5:519-527

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of Evolutionary Changes in Proteins. Pp. 345-352 in Atlas of Protein Sequence and Structure. vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376

Felsenstein, J. 1989. Phylogenetic Inference Package (PHYLIP), version 3.2. University of Washington, Seattle. Cladistics 5:164-166

Flores, T. P., C. A. Orengo, D. S. Moss, and J. M. Thornton. 1993. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 2:1811-1826

Geman S., and D. Geman. 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721-741.

Goldman, N. and Z. Yang. 1994. A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences. Mol. Biol. Evol. 11(5):725-736

Harata K., Y. Abe, and M. Muraki. 1998. Full-Matrix least squares refinement of lysozymes and analysis of anisotropic thermal motion. Proteins 30(3):232-243.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Jensen, J. L., and A. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32:499-517

Jones, D. T. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287:797-815

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. A new approach to protein fold recognition. Nature 358:86-89

Jones, D. T., and J. M. Thornton. 1996. Potential energy functions for threading. Curr. Opin. Struct. Biol. 6:210-216

Koshi, J. M., D. P. Mindell and R. A. Goldstein. 1997. Beyond mutation matrices: physical chemistry based evolutionary models. p.80-89 in S. Miyano and T. Takagi, eds. Genome Informatics 1997. Universal Academy Press, Tokyo

Koshi, J. M., D. P. Mindell and R. A. Goldstein. 1999. Using physical chemistry based substitution models in phylogenetic analyses of HIV-1 subtypes. Mol. Biol. Evol. 16(2):173-179

Koshi, J. M., and R. A. Goldstein. 1998. Mathematical models of natural amino acid site mutations. Proteins 32:289-295

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. (submitted).

Messier, W., and C. -B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature 385:151-154.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1092

Muse, S. V., and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. Mol. Biol. Evol. 11: 715 - 724

Neurath, A. R., and N. Strick. 1994. The putative cell receptors for hepatitis B virus (HBV), Annexin V, and Apolipoprotein H, bind to lipid components of HBV. Virology 204:475-477

Nielsen R. 2001. Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations. Genetics 159:401-411.

Parisi G. and J. Echave. 2001. Structural Constraints and Emergence of Sequence Patterns in Protein Evolution. Mol. Biol. Evol. 18(5):750-756.

Pedersen A. -M. K. and J. L. Jensen. 2001. A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames. Mol. Biol. Evol. 18(5):763-776.

Pedersen, A. -M. K., C. Wiuf, and R. B. Christiansen. 1998. A codon - based model designed to describe lentiviral evolution. Mol. Biol. Evol. 15(8):1069-1081

Pollock D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. 287(1):187-98.

Pupko, T., I. Pe'er, R. Shamir, and D. Grauer. 2000. A fast algorithm for joint reconstruction of ancestral amino–acid sequences. Mol. Biol. Evol. 17: 890-896

Reutelingsperger, C. P. M., and W. L. van Heerde. 1997. Annexin V. the regulator of phosphatidylserine-catalyzed inflammation and coagulation during apoptosis. Cell. Mol. Life Sci. 53:527-532

Schlaepfer, D. D., J. Jones, and H. T. Haigler. 1992. Inhibition of protein kinase C by Annexin V. Biochemistry 31:1886-1891

Singh, R. K., A. Tropsha, I. I. Vaisman. 1996. Delaunay tessellation of proteins. J. Comput. Biol. 2:213-221.

Stewart, C. -B., Schilling, J. W., and Wilson, A. C. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401-404.

Swofford, D. L., G. J. Olsen, P. J. Waddel, and D. M. Hillis. 1996. "Phylogenetic Inference", In Molecular Systematics, 2nd ed., Hillis, D. M., C. Moritz, and B. K. Mable, eds., pp. 407–514. Sinauer, Sunderland.

Wollenberg, K. R. and W. R. Atchley. 2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc. Nat. Acad. Sci. 97(7):3288-3291.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15(5):568-573

Yang, Z. and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908-917

Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15:1600-

1611.

Yang, Z., R. Nielsen, N. Goldman, A. -M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Appendix 1

The approximation of Equation (2.7) for $p(i|\theta')/p(i|\theta)$ relies on having a random sample of sequences $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(M)}$ from $p(\eta^{(h)}|\theta^*)$. The denominator of Equation (2.1) is

$$D(\theta) = \sum_k e^{-2sE_s(k)-2pE_p(k)} \prod_{n=1}^{N} \pi_{k_n}$$

and

$$\frac{p(i|\theta')}{p(i|\theta)} = e^{-2(s'-s)E_s(i)-2(p'-p)E_p(i)} \left(\prod_{m=1}^{N} \frac{\pi'_{i_m}}{\pi_{i_m}} \right) \frac{D(\theta)}{D(\theta')}. \tag{2.8}$$

The ratio of the denominators can be approximated via importance sampling,

$$\begin{aligned}
\frac{D(\theta)}{D(\theta')} &= \frac{\sum_k e^{-2sE_s(k)-2pE_p(k)} \prod_{n=1}^{N} \pi_{k_n}}{\sum_k e^{-2s'E_s(k)-2p'E_p(k)} \prod_{n=1}^{N} \pi'_{k_n}} \qquad (2.9) \\
&= \frac{\sum_k \frac{p(k|\theta^*)}{p(k|\theta^*)} e^{-2sE_s(k)-2pE_p(k)} \prod_{n=1}^{N} \pi_{k_n}}{\sum_k \frac{p(k|\theta^*)}{p(k|\theta^*)} e^{-2s'E_s(k)-2p'E_p(k)} \prod_{n=1}^{N} \pi'_{k_n}} \\
&\doteq \frac{\sum_{h=1}^{M} \frac{1}{p(\eta^{(h)}|\theta^*)} e^{-2sE_s(\eta^{(h)})-2pE_p(\eta^{(h)})} \prod_{n=1}^{N} \pi_{\eta_n^{(h)}}}{\sum_{h=1}^{M} \frac{1}{p(\eta^{(h)}|\theta^*)} e^{-2s'E_s(\eta^{(h)})-2p'E_p(\eta^{(h)})} \prod_{n=1}^{N} \pi'_{\eta_n^{(h)}}} \\
&= \frac{D(\theta^*) \sum_{h=1}^{M} e^{-2(s-s^*)E_s(\eta^{(h)})-2(p-p^*)E_p(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}{D(\theta^*) \sum_{h=1}^{M} e^{-2(s'-s^*)E_s(\eta^{(h)})-2(p'-p^*)E_p(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi'_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}} \\
&= \frac{\sum_{h=1}^{M} e^{-2(s-s^*)E_s(\eta^{(h)})-2(p-p^*)E_p(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}{\sum_{h=1}^{M} e^{-2(s'-s^*)E_s(\eta^{(h)})-2(p'-p^*)E_p(\eta^{(h)})} \prod_{n=1}^{N} \frac{\pi'_{\eta_n^{(h)}}}{\pi^*_{\eta_n^{(h)}}}}.
\end{aligned}$$

Substitution of this approximation for $D(\theta)/D(\theta')$ into Equation (2.8) yields Equation (2.7).

Appendix 2

To propose a sequence path ρ', we modify the current sequence path ρ by replacing the site path ρ_r at site r with a site path ρ'_r that is sampled from the posterior density $p(\rho'_r|i, j, \psi)$ as defined by the Felsenstein 1984 nucleotide substitution model (Felsenstein 1989). The Felsenstein 1984 model can be interpreted as being comprised of two separate processes: the within–group process, and the general process. For the general process, events to occur at a rate g. If a general event occurs, a nucleotide of type i is replaced by a nucleotide of type j with probability π_j. Besides general events, within–group events are also allowed by the Felsenstein 1984 model. These events occur at a rate w and the possible outcomes of a within–group event depend on whether the nucleotide occupying the site before an event is a purine or a pyrimidine. Let $H(\alpha)$ be R if nucleotide type α is a purine and Y if nucleotide type α is a pyrimidine. Also, define $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. If a within–group event occurs at a site currently occupied by type α, nucleotide type β replaces α with probability $\pi_\beta/\pi_{H(\beta)}$ if α and β are both purines or both pyrimidines and with probability 0 otherwise. In other words, within–group events either lead to transitions or to no change of nucleotide type.

Due to assumed independence of the within–group and general processes, the total number of events experienced at a site in time duration T has a Poisson distribution with parameter $(g + w)T$. This formulation of the Felsenstein 1984 model actually allows general or within–group events occur that result in the same nucleotide type occupying a site before and after an event. This sort of "hidden" event is algebraically convenient for determining how to sample ρ_r from $p(\rho_r|i, j, \psi)$

and it is not difficult to normalize so that the rate units are the expected number of nucleotide changes rather than the expected number of events.

With the Felsenstein 1984 model, the probability that a site is occupied by nucleotide type β after an amount of evolution T given that the site was originally occupied by type α is

$$p_{\alpha\beta}(T) = \begin{cases} (1-e^{-gT})\pi_\beta & \alpha \neq \beta, H(\alpha) \neq H(\beta) \\ (1-e^{-gT})\pi_\beta + e^{-gT}(1-e^{-wT})\frac{\pi_\beta}{\pi_{H(\beta)}} & \alpha \neq \beta, H(\alpha) = H(\beta) \\ (1-e^{-gT})\pi_\beta + e^{-gT}(1-e^{-wT})\frac{\pi_\beta}{\pi_{H(\beta)}} + e^{-gT}e^{-wT} & \alpha = \beta. \end{cases} \quad (2.10)$$

Each of the three terms found in these transition probabilities has an intuitive explanation. The probability of 0 general and 0 within–group events is $e^{-gT}e^{-wT}$. The $e^{-gT}(1-e^{-wT})\pi_\beta/\pi_{H(\beta)}$ term represents the case of 0 general events and at least one within–group event where the most recent within–group event results in the nucleotide type β occupying the site. The $(1-e^{-gT})\pi_\beta$ term is the probability that the site is occupied by type β and at least one general event occurs.

We employ three steps to sample a particular site path from the distribution $p(\rho_r|i,j,\psi)$. First, a random sample is obtained from the distribution of the number of general and within–group changes conditional on i_r, j_r and ψ. The second step, conditional upon the number of events that occur at site r, is to randomly sample the times of these events from a uniform distribution that spans the time period of length T.

The uniform distribution is used because, conditional upon the number of events, event times are uniformly distributed for a time–homogeneous Poisson process such as the one defined by the Felsenstein 1984 model. After the numbers of general and within–group events have been determined and the event times have been assigned, the third step is to appropriately assign the specific residue types that are yielded by each event.

The form of the transition probabilities facilitates sampling the number of general events and the number of within–group events at a site conditional upon the information that the site was occupied by type i_r at the beginning of the branch and by type j_r at the end of the branch. The most complicated situation is for $i_r = j_r$. In this situation, ρ_r has 0 general events and 0 within–group events with probability $e^{-gT}e^{-wT}/p_{i_r j_r}(T)$. The probability of 0 general events and at least one within–group event at the site for $i_r = j_r$ is $e^{-gT}(1 - e^{-wT})\pi_{j_r}/(\pi_{H(j_r)}p_{i_r j_r}(T))$. Given that this happens, the probability of exactly n_w within–group events on the path is $(wT)^{n_w}e^{-wT}/(n_w!(1-e^{-wT}))$ for $n_w \in \{1, 2, \ldots\}$. Therefore, the probability of 0 general events and n_w within–group events given that $i_r = j_r$ and $n_w \in \{1, 2, \ldots\}$ is $e^{-gT}(wT)^{n_w}e^{-wT}\pi_{j_r}/(n_w!\ \pi_{H(j_r)}\ p_{i_r j_r}(T))$. Likewise, the probability of $n_g > 0$ general events and n_w within–group events is $e^{-gT}(gT)^{n_g}e^{-wT}(wT)^{n_w}\pi_{j_r}/(n_g!\ n_w!\ p_{i_r j_r}(T))$ when $i_r = j_r$. Similar reasoning allows sampling from the probability distribution of n_g and n_w conditional upon the beginning nucleotide type i_r and the ending type j_r for the cases where $i_r \neq j_r$. After determining n_g and n_w, the times for each general or within–group event on a path can be randomly sampled from a uniform distribution along the branch.

The last step in randomly sampling a site path conditional upon the beginning type i_r and ending type j_r is to randomly determine which nucleotide type occupies a site after each event. The final event on a site path must always be to type j_r because this type occupies the site at the end of the branch. This final event may be of either the general or within–group variety.

If the last general event is not the final event on a site path then, in order to have type j_r at the end of the site path, the last general event must be to a nucleotide type included in group $H(j_r)$. Specifically, the last general event results in type k with probability $\pi_k/\pi_{H(j_r)}$ if k belongs to group $H(j_r)$. If k does not belong to $H(j_r)$, then the probability is 0 that type k is assigned to occupy the site after the last general event. Next, all general events that are not the last general event on the path at the site are handled. This is accomplished by having nucleotide type k occupy the site after one of these general events with probability π_k. When all general events have been assigned nucleotide types, the final step is to treat any remaining unassigned within–group events. Because the within–group events cannot result in transversions, the time intervals on a branch during which a site is occupied by a pyrimidine depend solely on whether the site is occupied by a purine or a pyrimidine at the beginning of the branch and on which general events result in purines and which result in pyrimidines. If an unassigned within–group event occurs during a "purine" interval, then the event results in A with probability π_A/π_R and G with probability π_G/π_R. If an unassigned within–group event occurs during a "pyrimidine" interval then the event results in C with probability π_C/π_Y and in T with probability π_T/π_Y. The procedure discussed to this point allows site paths to be sampled conditional upon the residues that begin and end the path, but some events on the sampled site path may be "hidden" in that the nucleotides before and after the event are identical. Because the goal is to propose a site path according to the Felsenstein 1984 model and then evaluate the resulting sequence path with a model that has dependence among codons and that is not parameterized so as to consider "hidden" events, all hidden events are pruned from the site path before it is considered in Equation (2.6). Because the Felsenstein 1984 model can also be parameterized in the conventional way that has all evolutionary events result in a changed sequence, the conventional parameterization is adopted for calculation of $J(\rho', \theta'|\rho, \theta)$ and $J(\rho, \theta|\rho', \theta')$.

Chapter 3

STOCHASTIC MAPPING WITH DEPENDENCE AMONG CODONS

Robinson DM, D Jones, H Kishino, N Goldman, and JL Thorne

Abstract

Mapping substitution events offers a unique view of the evolutionary history along each branch of a phylogeny. With this procedure, certain aspects of each event including the nature, timing and measure of its influence on the protein can be examined. Previous mapping strategies of protein-coding genes ignore constraints due to protein tertiary structure and instead assume that codons change independently. Although the independence assumption facilitates computation, it is biologically implausible. In earlier work, we introduced a statistical procedure for making evolutionary inferences from pairs of aligned protein sequences when one member had an experimentally determined structure. Here, the procedure is extended to three taxa and is applied to both lysozyme c and eosinophil-derived neurotoxin (EDN) genes. Substitution events inferred to be under purifying selection when independence among codons is assumed are estimated to be positively selected by our method due to the influence of pairwise interactions. When viewed structurally, spatial clusters of positively or negatively selected events are inferred for each branch of the EDN topology. The cluster formed by positively selected events along the human EDN branch is found to correspond with a region of enzymatic significance. Lastly, a temporal ordering between a structurally adjacent residue pair sheds light on possible evolutionary histories of a 13-fold increase in EDN RNase activity.

Introduction

When comparing protein-coding DNA sequences from different species, nonsynonymous differences are likely to be the result of mutations that occurred and were subsequently fixed in a population. Because observed nonsynonymous differences in a sample of homologous sequences are generated by the interaction between mutation, genetic drift, and natural selection, these nonsynonymous differences are said to be the result of nonsynonymous substitution events. Therefore, the rate at which a nonsynonymous substitution event occurs will be affected by both the mutation rate and the fitness of the mutation. Most nonsynonymous mutations are deleterious and will be removed from a population through purifying selection (see Li 1997). However, nonsynonymous mutations that impart a selective advantage are positively selected and may have higher fixation probabilities than if the events were synonymous.

Because of the difficulty in separating the effects of mutation and selection on substitution rates, one convenient practice is to classify a nonsynonymous substitution as being positively selected if and only if the rate of this substitution is higher than it would be if the change were synonymous rather than nonsynonymous. This definition of positive selection has proven valuable for characterizing evolution with codon-based models of protein sequence change (e.g., Yang 1998; Yang et al. 2000; Yang and Nielsen 2002; Yang and Swanson 2002). Although the details of these previously proposed models of codon change vary greatly, the key idea is to invoke a parameter that is typically referred to as ω. With these codon models, the rate of a nonsynonymous change is set equal to the product of ω and the rate the change would have if it were actually synonymous. Therefore, $\omega > 1$ corresponds to nonsynonymous changes being positively selected whereas $\omega < 1$ represents negative selection.

The ability to characterize positive selection and detect its presence has attracted widespread interest in employing codon-based models of evolution for better understanding the history and process of sequence change (e.g., Chang et al. 2002; Bielewski and Yang 2003). A problem with widely used codon-based models is their assumption that different codons in a sequence change independently. This independence assumption is not realistic, because translated amino acid residues must interact to form and stabilize complex tertiary structures such as binding sites and activation domains. Although most widely used models of sequence evolution assume independent change among sites or codons (see Felsenstein, 2003), this assumption is clearly violated (e.g., Pollock, Taylor and Goldman 1999; Wollenberg and Atchley 2000).

The rate of a nonsynonymous change should not depend solely on the codon that is modified by the change. Instead, the rate of a nonsynonymous change should also be affected by the compatibility of the resulting amino acid replacement and the amino acids specified by other codons in the protein-coding DNA sequence. Positive selection is not a property of codon in isolation but is a property of the entire ensemble of codons that determine a protein. A more refined and plausible treatment of positive selection would consider the exact nature of a nonsynonymous change as well as the entire context of the protein. For a given codon, most nonsynonymous substitutions may have an effectively neutral influence on the protein, however because of nonsynonymous events at surrounding codons, a particular nonsynonymous event may become highly advantageous due to complementary changes in pairwise interactions in the folded protein. On the other hand, circumstances may arise in which pairwise interactions cause nonsynonymous substitutions to be deleterious to the protein. These situations are examples of the "Dykhuizen – Hartl" effect (Dykhuizen and Hartl 1980) because the influence of the nonsynonymous substitution depends on the environment of the event (Zhang and Rosenberg 2002).

Here, we extend the approach of Robinson et al. (2003) and apply it to the characterization of positive and negative selection. This approach is based upon a model of protein-coding DNA sequence evolution that permits dependence among codons due to protein structure. Our procedure is built to account for the "Dykhuizen – Hartl" effect through our measure of sequence-structure compatibility (Jones, Taylor and Thornton 1992; Jones 1999), where we define the an environment of a site to be the collection of all amino acid residues whose C_β carbon atoms lie within ten Angstroms of the site in the folded protein. More importantly, the model also allows for nonsynonymous substitutions where the determination of positive or negative selection depends on which sequence is the ancestor and which sequence is the descendant. We particularly concentrate on mapping nonsynonymous changes to particular branches of a phylogeny and to particular portions of these branches. Methods for inferring the location of evolutionary events on a phylogenetic tree can be biologically illuminating and have been previously proposed (e.g., Swofford and Maddison 1987; Nielsen 2001, 2002; Huelsenbeck, Nielsen and Bollback 2003). Unlike our approach, the previous methods rely on the assumption that characters change independently.

Our method is illustrated with both lysozyme c and EDN genes. Summaries of event mappings allow the placement of substitution events with high posterior probability. In the EDN analysis, events that collectively are either positively or negatively selected, are inferred to form site clusters when viewed structurally. In particular, the site cluster formed by positively selected events along the human EDN branch is found to correspond with a region of enzymatic significance. Other site clusters inferred by our approach in the EDN protein may be shown to have functional significance if experimentally investigated. Although we measure the influence of nonsynonymous change according to sequence-structure compatibility, our approach is written in general and can accommodate other fitness measures. The next step is the extension of our method to datasets with greater than three taxa, which is currently being explored.

Evolutionary Model

To incorporate site dependencies, the idea of codon evolution must be extended to entire sequence evolution. A Markovian model of rate change is constructed, which defines the instantaneous rate $R_{i,j}$ from sequence i to sequence j, where i and j differ at precisely one nucleotide position. It is assumed that if sequence j has nucleotide type h at this position and does not translate a stop codon, then the instantaneous rate matrix has the following form:

$$R_{i,j} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ u\pi_h\kappa & \text{for a synonymous transition} \\ u\pi_h \ \Psi_{i,j} & \text{for a nonsynonymous transversion} \\ u\pi_h\kappa \ \Psi_{i,j} & \text{for a nonsynonymous transition} \\ 0 & \text{if } j \text{ contains a stop codon, or if } i \\ & \text{and } j \text{ differ in multiple positions} \end{cases} \tag{3.1}$$

where $\kappa > 0$ is the transition-transversion rate ratio and accounting for stop codons, π_h is the relative frequency of nucleotide type h with $h \in \{A, C, G, T\}$ and with restriction $\pi_A + \pi_C + \pi_G + \pi_T = 1$. The π_h and κ terms can be interpreted as reflecting mutational tendencies. In our procedure, the product of rate and time results in branch lengths that are measured in expected number of substitutions per nucleotide site. Because rates and times are confounded, time is arbitrarily set to one and the overall rate is captured by a parameter u for each branch.

Nonsynonymous substitution events influence how well a sequence adopts a particular fold. In our approach (Robinson et al. 2003), we utilize a sequence-to-structure compatibility function that has been successfully applied to protein fold recognition (Jones, Taylor and Thornton 1992; Jones 1999). This criterion approximates the fitness of any Sequence i in terms of solvent accessibility, denoted $E_s(i)$, and pairwise interactions, denoted $E_p(i)$. It is assumed that two amino acid residues interact if their C_β atoms are within ten Angstroms of one another in the folded protein (Jones, Taylor and Thornton 1992; Jones 1999). If $E_s(i)$ and $E_p(i)$ are low, then Sequence i is considered stable when folded onto the given structure. However, higher values $E_s(i)$ and $E_p(i)$ are associated with unstable sequences, and consequently have lower sequence-structure compatibility.

With our model, the rate of a nonsynonymous change from sequence i to sequence j is the rate the change would have were it synonymous multiplied by a factor

$$\Psi_{i,j} = \omega \ e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$$

When $\Psi_{i,j}$ exceeds one, the rate of the nonsynonymous change exceeds the rate it would have were it synonymous, and therefore, the change from i to j can be classified as being positively selected. The fitness of a particular protein-coding DNA sequence depends both on interactions among the amino acids that it encodes within the protein and on the external environment. The parameter ω is intended to capture the contributions to nonsynonymous rates that are external to the protein of interest. The $e^{(E_s(i)-E_s(j))s+(E_p(i)-E_p(j))p}$ component represents the effect on $R_{i,j}$ due to sequence-structure compatibility. The parameters s and p are respectively referred to as the solvent accessibility and pairwise interaction parameters. When $s = p = 0$, our model reduces to the standard codon models (e.g., Muse and Gaut 1994; Goldman and Yang 1994). Both s and p are expected to be positive, since positive values of these parameters favor sequences that fit the known structure well. The biologically unreasonable case where s and p are negative would have evolution favoring sequences that do not fit the known structure well.

An attractive feature of our previous work (Robinson et al. 2003) is the explicit form of the stationary probability of a given coding sequence i of length N. For simplicity of notation, let θ represent all the parameters in the instantaneous rate matrix R (i.e., $\theta = \{\kappa,\ \omega,\ s,\ p,\ u,\ \pi_A,\ \pi_C,\ \pi_G,\ \pi_T\}$) and use i_m to represent the nucleotide at position m in DNA sequence i. Sequence i has stationary probability:

$$p(i|\theta) = \frac{e^{-2sE_s(i)-2pE_p(i)} \prod_{m=1}^{N} \pi_{i_m}}{\sum_k e^{-2sE_s(k)-2pE_p(k)} \prod_{n=1}^{N} \pi_{k_n}}, \tag{3.2}$$

where the sum in the denominator is over all possible sequences k of length N that lack a premature stop codon.

Extension to three sequences

This simple model can be extended to three sequences by slightly modifying our earlier implementation of the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). Define a topology with three branches, where each branch b connects the unobserved internal node sequence i to an observed sequence j_b ($b \in \{1, 2, 3\}$). A conventional way to estimate the transition probability $p(j_b|i, \theta)$ for branch $b \in \{1, 2, 3\}$ is through spectral decomposition and matrix exponentiation (see Swofford et al. 1996). For nucleotide sequences of length N, the dimensions of our instantaneous rate matrix are on the order of $4^N \times 4^N$. Even for small values of N, matrices attain high dimensionality and explicit calculation of the transition probability $p(j_b|i, \theta)$ by matrix exponentiation becomes challenging computationally.

As an alternative to the explicit calculation of transition probabilities, we augment our data by mapping substitution events onto each branch of the evolutionary tree consistent with the observed sequences at the tip nodes. The series of events as well as their associated times are collectively referred to as a path. A path on branch b, denoted ρ_b, specifies how i is transformed to an observed descendant sequence j_b. Because we generally do not know the biologically correct evolutionary history or the precise substitution event times and because the number of possible paths are infinite, we instead sample paths from the appropriate probability density $p(j_b, \rho_b|i, \theta)$ to approximate the path posterior.

We implement this procedure in a Bayesian framework. In particular, we specify the prior density $p(\theta)$ and then sample from the joint posterior density, denoted $p(i, \rho_1, \rho_2, \rho_3, \theta|j_1, j_2, j_3)$. We can break the joint posterior density down further as follows,

$$\begin{aligned}
p(i, \rho_1, \rho_2, \rho_3, \theta|j_1, j_2, j_3) &= \frac{p(i, \rho_1, \rho_2, \rho_3, \theta, j_1, j_2, j_3)}{p(j_1, j_2, j_3)} \\
&= \frac{p(j_1, \rho_1|i, \theta)p(j_2, \rho_2|i, \theta)p(j_3, \rho_3|i, \theta)p(i|\theta)p(\theta)}{p(j_1, j_2, j_3)} \\
&= \frac{p(i|\theta)p(\theta)\prod_{b=1}^{3}[p(j_b, \rho_b|i, \theta)]}{p(j_1, j_2, j_3)} \qquad (3.3)
\end{aligned}$$

The denominator term $p(j_1, j_2, j_3)$ is difficult to determine explicitly. Fortunately, this term is not a function of $i, \rho_1, \rho_2, \rho_3$, or θ and thus does not need to be calculated with our Markov chain Monte Carlo (MCMC) procedure.

To compute $p(j_b, \rho_b|i, \theta)$ for any branch b, we must first define the rate of change from a sequence v to any other sequence k of length N as $R_{v,k}$. This term is computed using the instantaneous rate matrix in Equation (3.1). The term $R_{v,\bullet}$ will denote the rate away from sequence v where $R_{v,\bullet} = \sum_k R_{v,k}$ and where the sum is over all sequences k that differ from v. As with a Poisson process, the waiting time for a sequence v to experience a nucleotide substitution is exponentially distributed with parameter $R_{v,\bullet}$. The probability density that v is transformed to some sequence k given that a nucleotide substitution changes v

is $R_{v,k}/R_{v,\bullet}$. If we now define the rate away from a sequence i after incorporating substitution event x as $R_{i(x),\bullet}$, the probability of a series of q events in a path has the following form:

$$\begin{aligned} p(j_b, \rho_b|i, \theta) &= (\prod_{x=1}^{q} \frac{R_{i(x-1),i(x)}}{R_{i(x-1),\bullet}} R_{i(x-1),\bullet} e^{-R_{i(x-1),\bullet}(t(x)-t(x-1))}) e^{-R_{i(q),\bullet}(t(q+1)-t(q))} \\ &= (\prod_{x=1}^{q} R_{i(x-1),i(x)} e^{-R_{i(x-1),\bullet}(t(x)-t(x-1))}) e^{-R_{i(q),\bullet}(t(q+1)-t(q))}, \qquad (3.4) \end{aligned}$$

where the final term $e^{-R_{i(q),\bullet}(t(q+1)-t(q))}$ represents the probability of no change in the time interval from the final substitution at time $t(q)$ until the time $t(q+1) = 1$ that the branch ends.

A Markov chain is constructed on the joint i, ρ_1, ρ_2, ρ_3, and θ state space with stationary distribution $p(i, \rho_1, \rho_2, \rho_3, \theta|j_1, j_2, j_3)$. To initialize our MCMC algorithm for each branch, the state $(i^{(0)}, \rho_b^{(0)}, \theta^{(0)})$ is arbitrarily sampled from the set of possibilities where $p(i, \rho_b, \theta|j_b)$ exceeds zero. We then propose new random values i', ρ_b' and θ' conditional on the current values i, ρ_b and θ where the proposal density is denoted $J(i', \rho_1', \rho_2', \rho_3', \theta'|i, \rho_1, \rho_2, \rho_3, \theta)$. The Metropolis-Hastings probability that each branch accepts the proposed state is equal to the minimum of 1 and:

$$\begin{aligned} r &= \frac{p(i', \rho_1', \rho_2', \rho_3', \theta', j_1, j_2, j_3)}{p(i, \rho_1, \rho_2, \rho_3, \theta, j_1, j_2, j_3)} \text{ x } \frac{J(i, \rho_1, \rho_2, \rho_3, \theta|i', \rho_1', \rho_2', \rho_3', \theta')}{J(i', \rho_1', \rho_2', \rho_3', \theta'|i, \rho_1, \rho_2, \rho_3, \theta)} \\ &= \frac{p(i'|\theta')p(\theta')J(i, \rho_1, \rho_2, \rho_3, \theta|i', \rho_1', \rho_2', \rho_3', \theta')}{p(i|\theta)p(\theta)J(i', \rho_1', \rho_2', \rho_3', \theta'|i, \rho_1, \rho_2, \rho_3, \theta)} \text{ x } \prod_{b=1}^{3} \left[\frac{p(j_b, \rho_b'|i', \theta')}{p(j_b, \rho_b|i, \theta)}\right] \qquad (3.5) \end{aligned}$$

where if we accept the proposed state, the next state of the Markov chain, denoted ($\rho_1^{(1)}$, $\rho_2^{(1)}$, $\rho_3^{(1)}$, $\theta^{(1)}$, $i^{(1)}$), is set equal to the proposed state (i.e., $i^{(1)} = i', \rho_1^{(1)} = \rho_1', \rho_2^{(1)} = \rho_2', \rho_3^{(1)} = \rho_3', \theta^{(1)} = \theta'$). Otherwise, $i^{(1)} = i^{(0)}, \rho_1^{(1)} = \rho_1^{(0)}, \rho_2^{(1)} =$

$\rho_2^{(0)}, \rho_3^{(1)} = \rho_3^{(0)}$ and $\theta^{(1)} = \theta^{(0)}$. By repeatedly proposing $\rho_1', \rho_2', \rho_3', \theta'$ and i' and then randomly accepting the proposals with probabilities determined in Equation (3.5), a Markov chain with the desired stationary distribution is formed.

Proposing Ancestral Sequences

Our MCMC implementation actually consists of various proposal distributions for i, ρ_b and θ, and the Markov chain is formed by cycling through these different proposal distributions J. The proposal for each component of θ as well as the proposal of individual paths ρ_b on one branch have been described previously (see Robinson et al. 2003). An ancestral sequence proposal step however, is added to our MCMC routine to contend with uncertainty in the internal node.

We propose a new ancestral node sequence i' that is identical to i except for one randomly selected site z. We respectively represent the current and proposed nucleotides at site z of the ancestral node sequence by $i(z)$ and $i'(z)$. The residues at site z in the observed sequences are denoted $j_1(z), j_2(z)$ and $j_3(z)$. During this proposal step, the parameters in θ are held constant, i.e., $\theta' = \theta$, but the path on each branch must be updated to ensure that ρ_b' transforms i' to each observed descendant sequence. A sequence path can also be represented by site paths at each of the N nucleotide positions, namely $\rho_b(1)$, $\rho_b(2)$, ..., $\rho_b(z)$, ..., $\rho_b(N)$. The particular site path $\rho_b(z)$ specifies all nucleotide substitutions that change site z in path ρ_b, as well as the specific times of these changes.

Each site path is sampled from a model of evolution that assumes independence among nucleotides and is commonly referred to as the F84 model (see Felsenstein 1989). We use ψ to differentiate the parameters in the F84 model from the parameters in θ. Rather than proposing only the most probable nucleotide at site z, $i'(z)$ is randomly sampled from $p(i'(z)|j_1(z), j_2(z), j_3(z), \psi)$. We then propose $\rho'_b(z)$ by sampling from $p(\rho'_b(z)|i'(z), j_b(z), \psi)$ for each branch b. These two steps comprise the proposal for an ancestral node sequence, whose proposal density $J(i', \rho'_1, \rho'_2, \rho'_3|i, \rho_1, \rho_2, \rho_3)$ has the following form:

$$\begin{aligned}
J(i', \rho'_1, \rho'_2, \rho'_3|i, \rho_1, \rho_2, \rho_3) &= p(i'(z), \rho'_1(z), \rho'_2(z), \rho'_3(z)|j_1(z), j_2(z), j_3(z), \psi) \\
&= p(i'(z)|j_1(z), j_2(z), j_3(z), \psi) \\
&\qquad \mathrm{x}\; p(\rho'_1(z), \rho'_2(z), \rho'_3(z)|i'(r), j_1(r), j_2(r), j_3(r), \psi) \\
&= p(i'(z)|j_1(z), j_2(z), j_3(z), \psi)\; p(\rho'_1(z)|j_1(z), i'(z), \psi) \\
&\qquad \mathrm{x}\; p(\rho'_2(z)|j_2(z), i'(z), \psi)\; p(\rho'_3(z)|j_3(z), i'(z), \psi) \\
&= p(i'(z)|j_1(z), j_2(z), j_3(z), \psi) \prod_{b=1}^{3} p(\rho'_b(z)|j_b(z), i'(z), \psi) \qquad (3.6)
\end{aligned}$$

Combining this result with the general form of the Metropolis-Hastings acceptance probability in Equation (3.5), the MCMC algorithm accepts the proposed ancestral node and paths on each branch with probability equal to the minimum of 1 and:

$$\begin{aligned}
r &= \frac{p(i'|\theta)J(i, \rho_1, \rho_2, \rho_3|i', \rho'_1, \rho'_2, \rho'_3)}{p(i|\theta)J(i', \rho'_1, \rho'_2, \rho'_3|i, \rho_1, \rho_2, \rho_3)} \;\mathrm{x}\; \prod_{b=1}^{3} \left[\frac{p(j_b, \rho'_b|i', \theta)}{p(j_b, \rho_b|i, \theta)}\right] \\
&= \frac{p(i'|\theta)p(i(r)|j_1(r), j_2(r), j_3(r), \psi)}{p(i|\theta)p(i'(r)|j_1(r), j_2(r), j_3(r), \psi)} \;\mathrm{x}\; \prod_{b=1}^{3} \left[\frac{p(j_b, \rho'_b|i', \theta)p(\rho_b(r)|j_b(r), i(r), \psi)}{p(j_b, \rho_b|i, \theta)p(\rho'_b(r)|j_b(r), i'(r), \psi)}\right] \qquad (3.7)
\end{aligned}$$

One complication in updating the s, p or π parameters contained in θ is calculating the stationary distribution of the ancestral state, $p(i|\theta)$, given in Equation (3.2). However, when performing an ancestral node update, the computationally intense grid-based importance sampling approach as outlined in our previous paper (see Robinson et al. 2003) is unnecessary because the parameter values $\theta' = \theta$.

This significantly simplifies the ratio of the stationary distributions of the proposed and current sequences:

$$\frac{\Pr\left(i' \mid \theta'\right)}{\Pr\left(i \mid \theta\right)} = \frac{\Pr\left(i' \mid \theta\right)}{\Pr\left(i \mid \theta\right)} = \frac{\pi_{i'(z)}}{\pi_{i(z)}} \quad e^{-2s(E_s(i')-E_s(i))-2p(E_p(i')-E_p(i))} \tag{3.8}$$

where the pseudo-energy differential according to solvent accessibility and pairwise interactions between the current and the proposed sequences are multiplied by the specific parameter values s, p and π taken from the Markov chain at the time the update is made.

Data Analysis

For the analysis performed here, each MCMC run consists of 5.1 million cycles through the Markov chain where the first 100,000 cycles are treated as "burn-in" and are discarded from subsequent analyses. The posterior distribution is approximated by sampling all parameter values, the ancestral node and paths along each branch every 500 cycles of the chain, for a total of 10,000 samples.

For each branch and for each nucleotide position, the proportion of sampled paths that include at least one nonsynonymous substitution event at that position is computed. We refer to this proportion as the posterior probability of event placement on a branch. Unless otherwise noted, we discuss only substitutions that occur in at least 50% of paths sampled. To facilitate our analyses, although the probabilities are calculated for a particular nucleotide position, we only display the associated codon location of these events.

When independence among codons is assumed, the determination of whether a codon is positively selected depends solely on the estimation of the ratio of nonsynonymous to synonymous rates of change, denoted ω (e.g., Goldman and

Yang 1994). When viewed in a Bayesian framework, if ω is estimated to exceed one, these methods will consequently infer all codons that contain a nonsynonymous substitution event to be positively selected. These codons will also be positively selected with approximately the same proportion with which ω is estimated to exceed one despite codon type or branch location. Under our approach, whether an event is positively selected depends on the proportion that $\Psi_{i,j}$ exceeds one, denoted $p(\Psi_{i,j} > 1)$. Although the estimate of this proportion is influenced by the proportion of samples for which ω exceeds one, $p(\Psi_{i,j} > 1)$ also depends on the value of the sequence-to-structure compatibility criterion evaluated for each substitution event. Therefore, our treatment allows a range of positive selection probabilities to be inferred due to the influence of surrounding residues in the folded protein.

Applications

The evolutionary time direction can be inferred for a pair of sequences with the addition of an outgroup sequence. In general, the true root node should be positioned along the outgroup branch, yet its exact location is indeterminable for time reversible models due to the pulley principle (see Felsenstein 1981). Models of evolution that assume independence among codons (e.g., Muse and Gaut 1994; Goldman and Yang 1994) will infer codons to be positively selected despite temporal direction. By employing $\Psi_{i,j}$ however, sequence i is assumed to be the ancestral sequence and j is assumed to be the descendant. Although events are inferred to be positively selected under these assumptions, the same events may be negatively selected when the roles of i and j are reversed. Although this is not always the

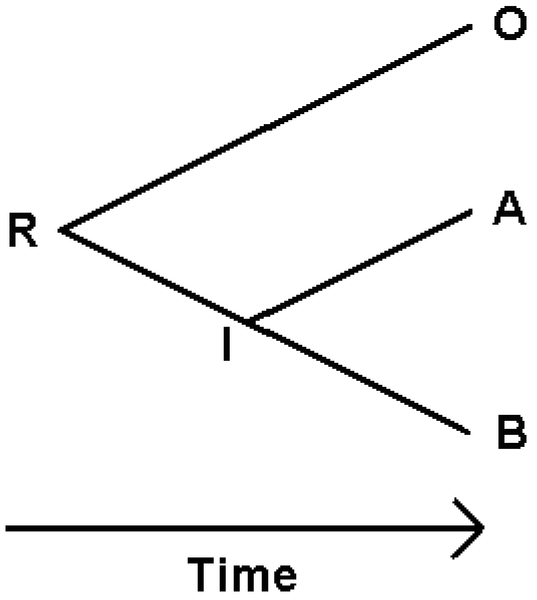


Figure 1: Given observed sequences A, B and outgroup sequence O, the phylogeny has a defined internal node sequence I and root node sequence R. On the I-O branch, the temporal direction of evolution changes at the root node R. That is, although the R-O component of the I-O branch is traversed forward in evolutionary time, the I-R component is traversed in reverse.

case, this behavior can be explained by observing the form of $\Psi_{j,i}$:

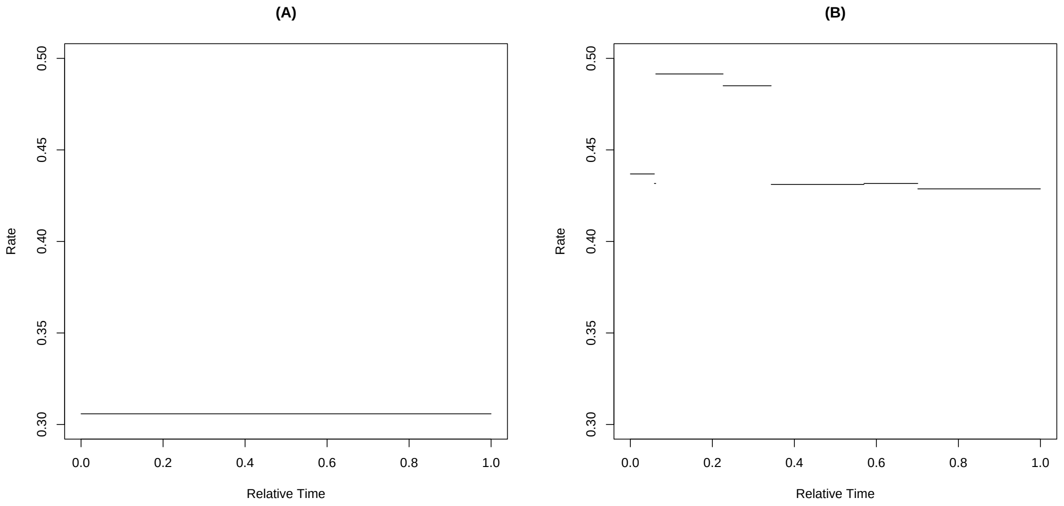
$$\Psi_{j,i} = \omega \; e^{(E_s(j)-E_s(i))s+(E_p(j)-E_p(i))p}$$

In this case, the difference in solvent accessibility $(E_s(j) - E_s(i))$ or pairwise interactions $(E_p(j) - E_p(i))$ sequence-structure compatibility may have a significant impact on $\Psi_{j,i}$ which depends on ancestral sequence choice. Although events simulated along the outgroup branch reverse evolutionary direction at the position of the true root node (see Figure 3.1), all inferences are made assuming the internal node is the root.

Lysozyme c protein

Langur monkeys have developed a foregut where leafy material is digested via bacterial fermentation. Bacteria that escape the foregut are immediately degraded by the high levels of lysozyme c in the stomach located downstream of the foregut (Stewart, Schilling and Wilson 1987; Messier and Stewart 1997). We focus on lysozyme c because it has experienced a large excess of positive selection during its evolutionary history (e.g., Stewart, Schilling and Wilson 1987; Messier and Stewart 1997; Yang 1998; Yang and Nielsen 2002; Robinson et. al 2003).

We illustrate our method using three lysozyme c sequences that each consist of 390 aligned nucleotides. These sequences are from the human (*Homo sapiens* : Genbank accession number U76923), the Rhesus macaque (*Macaca mulatta* : X60236), and *Colobus guerza* : U76916. The tertiary structure of lysozyme c has been experimentally determined for the human sequence (PDB Id: 1JSF) (Harata, Abe and Muraki 1998) and is assumed to be the native state conformation of all sequences analyzed. To differentiate the three species in our output, we will use



(C) Interactions with codon 61

Codon	Time	From AA	To AA
40	0.0586	D	N
52	0.0619	T	A
50	0.2262	K	R
52	0.3434	A	T
60	0.5698	D	N
67	0.7011	N	D

Figure 3.2: (A) The rate across a randomly sampled path from the analysis where codons evolve independently and identically ($i.e., s = p = 0$). (B) An extreme example of how the rate across a different randomly selected path from codon 61 can vary when codons are allowed to interact ($i.e., s \neq 0, p \neq 0$). (C) The nonsynonymous substitutions that occur on the sampled path that interact structurally with codon 61. The particular MCMC sampled parameter values for s and p associated with this path were 0.7746 and 0.0659, respectively.

“Mm” to represent the *Macaca mulatta*, “Cg” to represent the *Colobus guerza* and “Hs” to represent *Homo sapiens.*

The rate away from a codon equals the sum of the rate to each of the nine possible codons that differ by exactly one nucleotide, as defined by the instantaneous rate matrix in Equation (3.1). If no substitution event occurs at a codon, the rate away will remain constant across a path when independence among codons is assumed. However when site dependencies are considered, the rate away also depends on the identity and location of neighboring codons in the folded protein.

Thus, nonsynonymous substitution events at neighboring codons influence the rate away due to differences in calculated pairwise interactions. Even though the 61st aligned codon in our lysozyme c data is identical for all three sequences, Figure 3.2 illustrates how the rate away from codon 61 depends on nonsynonymous events elsewhere in the protein with our dependence model (see Figures 3.2B and 3.2C), but not with the corresponding independence model (Figure 3.2A). Although the contrast in rate away between independence and dependence models is particularly high for codon 61, other codons experience qualitatively similar behavior.

Although 10 MCMC runs were performed to check for convergence of the Markov chain, the results that are presented represent a single randomly selected run from this set. The posterior means and 95% credibility intervals are given for the model parameters and branch lengths in Table 3.1. Posterior estimates of s and p are both positive, because they represent the biologically plausible condition where evolution favors sequences that fold well onto the given structure. If the 95% credibility intervals of s and p contained zero, that would suggest that structure is not playing a role in the evolutionary process. The values of s and p may influence the amino acid composition of sequences that are compatible with the given structure. This may explain the observed differences in the distribution of relative frequencies of the four nucleotides (data not shown) as well as in the ω parameter posteriors (see Figure 3.3) inferred under both model conditions.

By summarizing the lysozyme c path posterior distribution for each branch, substitution event placement probabilities can be assigned to events when both site independence and dependencies are assumed. The codons listed in Table 3.2 represent only those positions whose place probabilities exceed 0.5. Except for

	$(s = p = 0)$	$(s \neq 0, p \neq 0)$
κ	3.6741 (1.6836 , 7.0617)	3.1603 (1.4381 , 6.2365)
ω	1.7047 (0.6590 , 3.9523)	1.8656 (0.7130 , 4.2493)
u_{MM}	0.0086 (0.0016 , 0.0239)	0.0095 (0.0017 , 0.0264)
u_{CG}	0.0223 (0.0063 , 0.0536)	0.0258 (0.0074 , 0.0604)
u_{HS}	0.0257 (0.0073 , 0.0599)	0.0292 (0.0086 , 0.0684)
s	0 NA	0.8515 (0.5036 , 1.2338)
p	0 NA	0.0515 (0.0359 , 0.0686)
π_A	0.3281 (0.2840 , 0.3749)	0.3661 (0.3114 , 0.4280)
π_C	0.1660 (0.1331 , 0.2008)	0.1604 (0.1243 , 0.2009)
π_G	0.2654 (0.2253 , 0.3070)	0.3011 (0.2565 , 0.3482)
π_T	0.2404 (0.2024 , 0.2803)	0.1725 (0.1273 , 0.2191)
BL_{Mm}	0.0143 (0.0040 , 0.0296)	0.0140 (0.0037 , 0.0297)
BL_{Cg}	0.0370 (0.0197 , 0.0600)	0.0379 (0.0199 , 0.0621)
BL_{Hs}	0.0427 (0.0237 , 0.0670)	0.0431 (0.0242 , 0.0678)

Table 3.1: Posterior means and 95% credibility intervals of all parameter values for the lysozyme c sequence alignment. The site independent $(s = p = 0)$ and site dependent $(s \neq 0, p \neq 0)$ cases are indicated at the top of each column. Branch Lengths (BL) are measured as expected number of changes per nucleotide site. Although most parameters are shared by each lineage, the two letter subscripts, as described in the text, are used as subscripts to indicate to which species the rate parameter u and branch lengths belong.

(A)	**Branch 1:** *Macaca mulatta* **(Mm)**		
Codon	**Independence**	**Dependence**	
position	**Placement**	**Placement**	$p(\Psi_{i,j} > 1)$
37	98.43	98.15	89.83
50	71.00	69.36	92.24
62	89.27	87.03	88.89
94	89.20	85.05	62.19

(B)	**Branch 2:** *Colobus guerza* **(Cg)**		
Codon	**Independence**	**Dependence**	
position	**Placement**	**Placement**	$p(\Psi_{i,j} > 1)$
14	98.56	98.96	95.10
21	98.65	98.36	91.65
23	98.78	97.89	86.07
41	100	100	89.72
50	62.18	62.93	77.26
87	98.84	98.59	88.87
113	98.45	99.21	96.08
114	99.67	99.77	97.98
126	99.54	99.70	84.30

(C)	**Branch 3:** *Homo sapiens* **(Hs)**		
Codon	**Independence**	**Dependence**	
position	**Placement**	**Placement**	$p(\Psi_{i,j} > 1)$
2	98.43	98.44	90.24
17	99.72	99.55	67.12
29	98.89	98.70	82.69
41	98.84	98.30	81.42
47	99.89	99.74	69.82
50	99.61	99.18	87.63
67	98.84	98.93	84.19
79	99.57	99.67	84.48
82	98.49	98.21	88.06
101	99.49	99.29	87.58
115	98.85	98.75	87.88
122	99.74	99.20	87.57

Table 3.2: Posterior probabilities of substitution event placement and of being under positive selection are presented for each branch of the lysozyme c analysis. The site independent and site dependent cases are indicated at the top of each column. When independence among codons is assumed, events are positively selected with approximately posterior probability of 0.8294. When site dependencies are considered, the posterior probability of being positively selected is given by the $p(\Psi_{i,j} > 1)$ column.

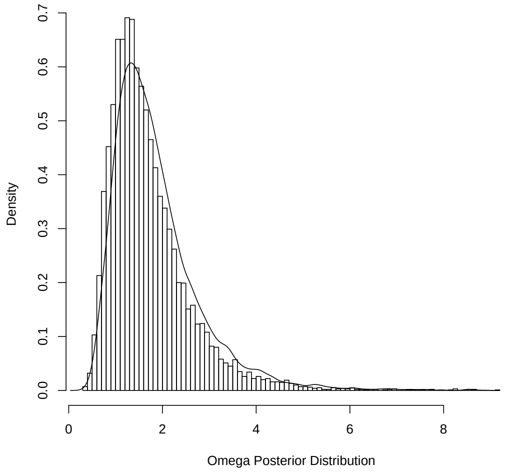


Figure 3.3: The posterior density of ω for the lysozyme c protein. The histogram represents the posterior density when independence is assumed, while the density for the site dependent approach is given as an overlayed line.

position 50 in the Mm and Cg branches, all events can be placed on their respective branches with at least 0.85 probability. By assuming independence among codons, or neglecting the structural component of $\Psi_{i,j}$, posterior probabilities of being positively selected can be estimated for all events by examining the posterior distribution of ω in Figure 3.3 for each model condition. The value of ω is estimated to exceed one in 82.94% of sampled values when independence is assumed and in 88.20% of samples when structure is considered. These posterior probabilities are going to be approximately identical for all nonsynonymous substitutions despite event type, or branch location.

Compared with the estimated probability of 0.8820, $p(\Psi_{i,j} > 1)$ can vary significantly due to pairwise interactions when site dependencies are considered (see Table 3.2, column 4). For instance, events with $p(\Psi_{i,j} > 1)$ exceeding 0.8820, such as residues 37, 50 and 62 in the Mm branch and residues 14, 21, 41, 87, 113 and

114 in the Cg branch, are positively selected through ω and increase sequence-structure compatibility. Conversely, events with $p(\Psi_{i,j} > 1)$ estimated less than 0.8820, such as residue 94 in the Mm branch, residues 23, 50 and 126 in the Cg branch and residues 17, 29, 41, 47, 50, 67, 79, 82, 101, 115, and 122 in the Hs branch may be positively selected through ω, but destabilize sequence-structure compatibility. The codons estimated to be positively selected along each branch tend to agree closely with what had been determined previously using alternate methods (Stewart, Schilling and Wilson 1987; Yang and Nielsen 2002).

Eosinophil-derived neurotoxin protein

Gene duplication has been hypothesized to be an important precursor for the development of novel gene function (e.g., Ohno 1970; Zhang, Zhang and Rosenberg 2002). In eosinophil leukocytes, duplication of the eosinophil-derived neurotoxin (EDN) and eosinophil-cationic protein (ECP) genes was estimated to have occurred 31 MYA (e.g. Zhang, Dyer and Rosenberg 2000; Zhang and Rosenberg 2002). *In vitro* studies have shown the human EDN protein to possess strong RNA antiviral properties that functions to decrease the ability of a virus to infect its host as seen in respiratory syncytial virus and HIV (Domachowske et al. 1998; Lee-Huang et al. 1999).

The protein sequences analyzed consist of 393 aligned nucleotide positions. To examine the interaction between codons 64 and 132 observed by Zhang and Rosenberg (2002), the EDN protein from the Squirrel monkey (*Saimiri sciureus* : AF4796321), the EDN protein from humans (*Homo sapiens* : M28129) and the ECP protein of the Orangutan (*Pongo pygmaeus* : U24101) were compared. The

Param.	$(s = p = 0)$	$(s \neq 0, p \neq 0)$
κ	1.7131 (1.0677 , 2.5699)	1.7363 (1.0809 , 2.6100)
ω	0.8615 (0.5299 , 1.3588)	0.9571 (0.5839 , 1.5077)
u_{Ss}	0.1490 (0.0823 , 0.2433)	0.1473 (0.0792 , 0.2419)
u_{Hs}	0.1125 (0.0584 , 0.1916)	0.1129 (0.05806 , 0.1919)
u_{Pp}	0.1258 (0.0669 , 0.2072)	0.1242 (0.06615 , 0.2065)
s	0 NA	0.6956 (0.3944 , 1.0154)
p	0 NA	0.0379 (0.0229 , 0.0533)
π_A	0.3152 (0.2779 , 0.3569)	0.3506 (0.3050 , 0.4013)
π_C	0.2468 (0.2038 , 0.2839)	0.2479 (0.2089 , 0.2901)
π_G	0.1861 (0.1499 , 0.2228)	0.1926 (0.1585 , 0.2258)
π_T	0.2518 (0.2138 , 0.2950)	0.2088 (0.1639 , 0.2522)
BL_{Ss}	0.1122 (0.0774 , 0.1532)	0.1132 (0.0768 , 0.1545)
BL_{Hs}	0.0840 (0.0532 , 0.1212)	0.0855 (0.0546 , 0.1231)
BL_{Pp}	0.0952 (0.0621 , 0.1334)	0.0957 (0.0632 , 0.1351)

Table 3.3: Posterior means and 95% credibility intervals of all parameter values and Branch Lengths for the EDN sequence comparison.

tertiary structure of the human EDN protein (PDB Id: 1HI2) (Mosimann et al. 1996) is assumed to be the native state conformation of all sequences analyzed. To differentiate the three species in this dataset, we will use “Ss” to represent the *Saimiri sciureus*, “Hs” to represent the *Homo sapiens*, and “Pp” to represent *Pongo pygmaeus*.

Homologous regions in each taxa were manually aligned to the structure using the resolved sequence alignment (Zhang and Rosenberg 2002). Unlike the lysozyme c sequences however, three gaps had to be introduced into the EDN alignment at codons 87, 88 and 117 which problematically translate to physical holes in the structure. Following previous methods (e.g., Swofford et al. 1996), codon positions containing gaps are removed from the alignment. To maintain accurate energy calculations, the list of solvent accessibility measures for each site is computed using the complete structure (Kabsch and Sanders 1983) and manually trimmed to account for the missing codons. Furthermore, to ensure consistent pairwise-energy calculations, a space holder is inserted at the position of each gap, which preserves the appropriate amino acid residue separation along the protein chain.

The posterior means and 95% credibility intervals are given for the model parameters and branch lengths in Table 3.3. As with the lysozyme c analysis, estimates of s and p both lie in the biologically plausible range and their credibility intervals do not contain zero. The effect of s and p on the amino acid composition of sequences selected by the approach appear to have a stronger influence on the EDN proteins than with the lysozyme c proteins and may explain the larger disparity in the estimated ω posterior distributions inferred under both model conditions (see Figure 3.4).

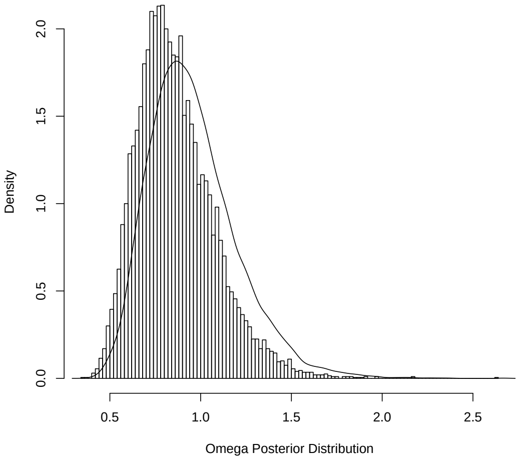
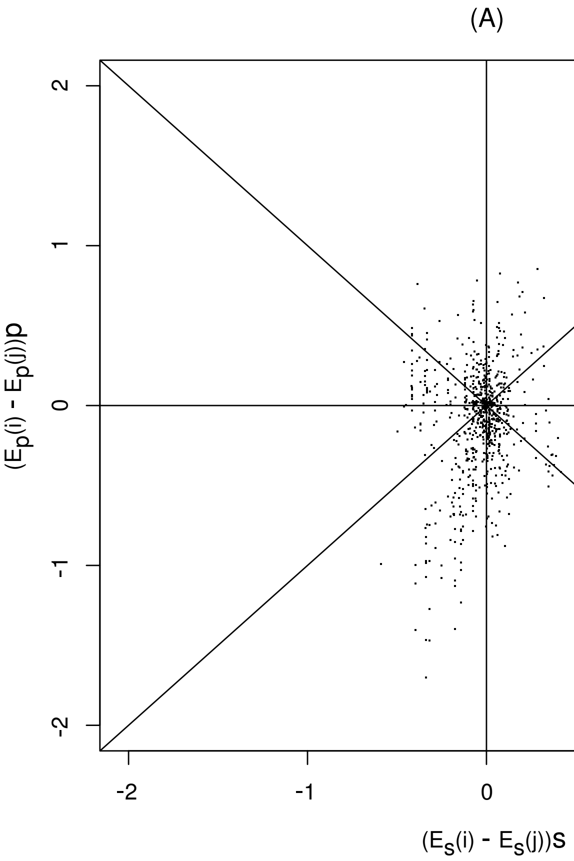


Figure 3.4: The posterior density of ω for the EDN protein. The histogram represents the posterior density when independence is assumed, while the density for the site dependent approach is given as an overlayed line.

For each event, solvent accessibility and pairwise interactions influence the nonsynonymous rate of change through the quantities $(E_s(i) - E_s(j))s$ and $(E_p(i) - E_p(j))p$ contained in $\Psi_{i,j}$. By randomly selecting a sample, the relationship between these quantities can be explored for all possible nonsynonymous substitution events from the associated ancestral node sequence (see Figure 3.5(A)). Negative values of either $(E_s(i) - E_s(j))s$ or $(E_p(i) - E_p(j))p$ indicate that the event, if accepted, would destabilize sequence-structure compatibility. From this plot, a significant number of points are observed to lie in the lower left quadrant, thus would reduce the compatibility measure through both quantities. However when derived from the events accepted by our approach along each path of the sample number, a qualitatively different pattern of these quantities is derived (see Figure 3.5(B)). Because events that are highly destabilizing are improbable under our approach,


(A)
2
1
0
-1
-2
$(E_p(i) - E_p(j))p$
-2
-1
0
1
2
$(E_s(i) - E_s(j))s$

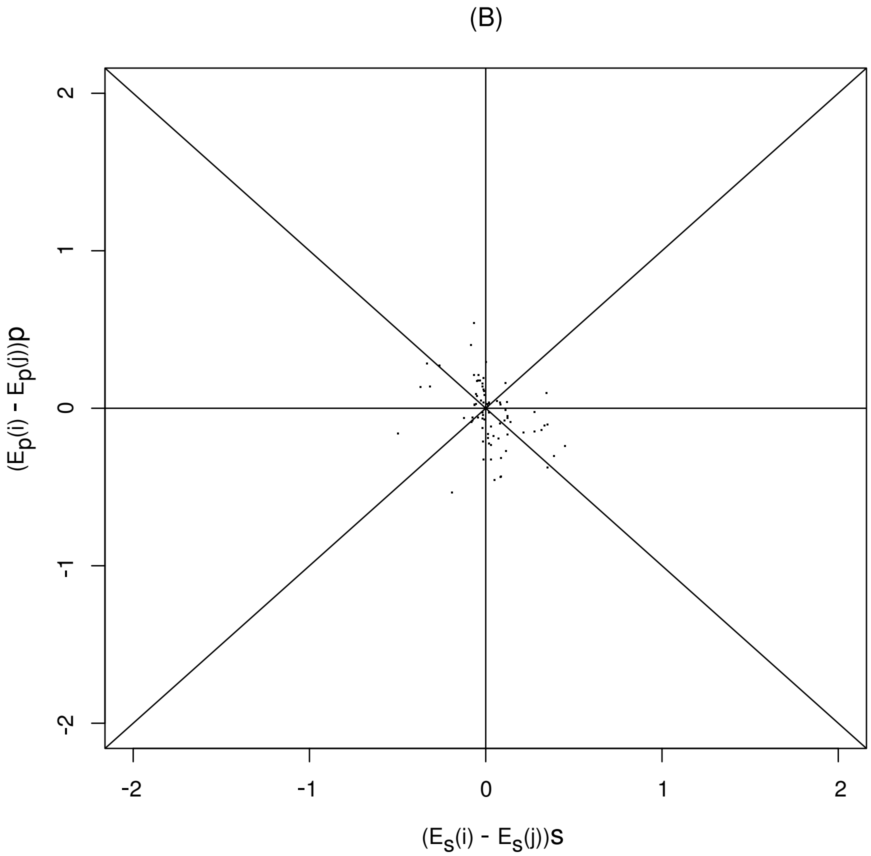


Figure 3.5: The values of $(E_s(i) - E_s(j))s$ and $(E_p(i) - E_p(j))p$ associated with (A) all possible nonsynonymous substitutions from the ancestral node sequence, and (B) the nonsynonymous events simulated on each branch, are shown for a randomly selected path of the EDN analysis with site dependencies.

this behavior is somewhat expected.

Increased evolutionary distance between taxa often translates to greater ambiguity in the ancestral node (e.g., Koshi and Goldstein 1996; Zhang and Nei 1997; Huelsenbeck and Bollback 2001). However with only a few exceptions, the vast majority of substitution events have high posterior placement probabilities (see Table 3.4). When independence among codons is assumed, or when the structural component of $\Psi_{i,j}$ is neglected, the posterior distributions of ω for each model condition can be compared. Estimated posterior probabilities of events to be positively selected will be approximately equal to the proportion with which ω is expected to exceed one (see Figure 3.4). When independence among codons is assumed, this proportion has estimated probability 0.2227, therefore all events will be inferred to be under purifying selection. Although when site dependencies are considered ω is estimated to exceed one in only 37.50% of sampled values, $p(\Psi_{i,j} > 1)$ infers several events to be positively selected due to the influence of surrounding codons (see Table 3.4, column 4).

When mapped onto the given structure, correlations between positively selected or negatively selected events can be estimated through the formation of substitution event site clusters. A site cluster is defined between events whose translated amino acids have C_β atoms within ten Angstroms of one another. However, site clusters are rarely observed when independence among codons is assumed since all events are positively selected with approximately the same posterior probability. When site dependencies are allowed, events inferred along the EDN branches can be categorized as being positively selected if $p(\Psi_{i,j} > 1)$ exceeds 0.5 probability. Yet, if this quantity is inferred with less than 0.375, the event is estimated to be under negative or purifying selection and to be destabilizing to sequence-structure compatibility as well.

For each branch, the events in Table 3.4 are parsed into both positively selected and negatively selected destabilizing categories, and are subsequently mapped onto the structure using RASMOL (Sayle and Milner-White 1995). For each branch of the phylogeny, these events are displayed in Figures 3.6(A-F). These events are also displayed in Table 3.5, which contains the identity of all site clusters shown in Figure 3.6. The site clusters in the table are displayed as codon positions in paren-

(A)	**Branch 1:** *Saimiri sciureus* **(Ss)**		
Codon	**Independence**	**Dependence**	
position	**Placement**	**Placement**	$p(\Psi_{i,j} > 1)$
6	97.34	96.46	42.48
12	99.12	98.75	58.67
17	99.90	99.96	52.44
19	96.26	99.93	18.28
21	65.36	86.23	6.54
22	95.35	99.81	37.04
25	96.66	95.13	7.38
29	96.18	96.31	29.56
32	97.64	98.26	62.93
34	98.16	97.23	15.26
45	66.20	71.01	65.07
50	95.99	95.91	43.87
58	97.75	97.81	64.60
66	98.84	98.74	27.16
67	97.45	96.35	41.33
76	69.72	71.24	54.76
81	96.62	95.24	7.82
82	96.18	97.01	77.59
91	98.13	96.87	21.97
92	97.89	96.26	15.10
97	98.96	99.16	51.92
99	98.30	99.55	57.97
100	98.74	99.28	67.61
102	97.72	97.49	43.72

(B)	**Branch 2:** *Homo sapien* **(Hs)**		
Codon	**Independence**	**Dependence**	
position	**Placement**	**Placement**	$p(\Psi_{i,j} > 1)$
3	97.30	97.15	48.75
12	97.67	97.90	57.03
13	94.09	93.97	16.65
20	94.14	94.38	50.29
21	92.49	92.89	60.73
28	94.82	94.97	39.15
45	71.80	70.65	17.60
64	96.21	96.08	55.51
66	59.27	63.66	78.31
68	96.16	97.20	91.75
69	99.84	99.87	69.75
75	91.94	92.16	64.13
76	98.81	98.87	86.45
89	96.24	94.58	59.11
97	60.53	58.37	42.24
108	95.16	93.96	38.47
116	96.52	97.12	25.38
120	94.86	94.74	28.30
132	99.96	99.76	19.20

(C)	**Branch 3:** *Pongo pygmaeus* **(Pp)**		
Codon	**Independence**	**Dependence**	
position	**Placement**	**Placement**	$p(\Psi_{i,j} > 1)$
7	97.25	98.59	79.01
12	76.17	83.00	92.33
16	93.53	93.49	25.97
17	95.04	93.51	31.27
18	96.61	96.48	62.86
19	97.92	96.25	28.34
21	59.58	85.88	81.49
25	96.72	95.18	25.18
39	94.04	92.86	19.63
45	63.12	59.95	43.76
60	96.18	95.59	53.93
66	65.26	62.91	27.50
73	93.78	99.63	45.41
76	76.16	73.56	25.45
81	96.92	94.59	11.93
90	97.60	96.11	20.69
97	65.82	66.63	49.91
100	99.19	98.54	65.79
101	97.03	96.00	18.41
102	97.82	97.00	82.39
103	96.60	95.76	9.25
104	99.87	99.78	35.69
105	97.33	97.34	23.36
122	94.87	93.37	26.41
133	95.58	94.26	11.28

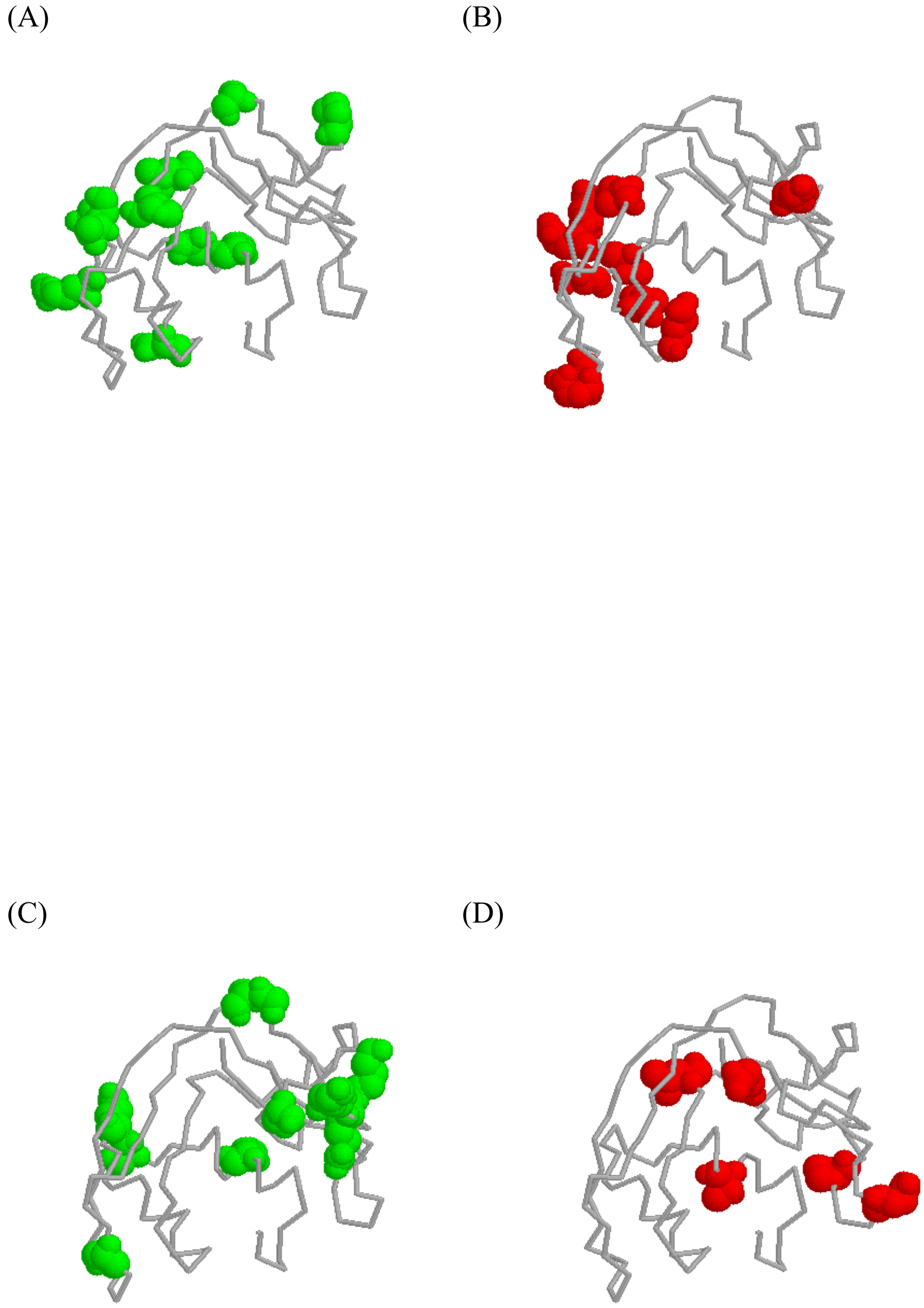
Table 3.4: Posterior probabilities of substitution event placement and of being under positive selection are presented for each branch of the EDN protein analysis. The site independent and site dependent cases are indicated at the top of each column. When independence among codons is assumed, events are positively selected with approximately 0.2227 posterior probability. When site dependencies are considered, the posterior probability of being positively selected is given by the $p(\Psi_{i,j} > 1)$ column.

Species	$p(\Psi_{i,j} > 1) > 0.5$	$p(\Psi_{i,j} > 1) < 0.375$
Ss	32, 45, 58, 76, 82, (12,17), (97,99,100)	66, (91,92), (19,21,22,25,29,34,81)
Hs	12, 89, (20, 21), (75, 76), (64, 66, 68, 69)	13, 45, 116, 120, 132
Pp	60, (7, 12), (18, 21, 100, 102)	39, 66, 90, 122 (16, 17, 19, 25) (76, 81, 101, 103, 104, 105, 133)

Table 3.5: Site clusters of both positively selected events, given by $p(\Psi_{i,j} > 1) > 0.5$, and negatively selected destabilizing events, given by $p(\Psi_{i,j} > 1) < 0.375$, formed along the three branches of the EDN protein sequence analysis. Single numbers represent codons that do not interact with the other codons in the set, while codons listed in parentheses form site clusters of various sizes.

theses, yet events that do not form clusters are shown as single codon positions. On the Ss branch, although two small positively selected site clusters given by (12,17) and (97,99,100) are formed, these events appear to be dispersed across the protein (see Figure 3.6(A)). However, the negatively selected destabilizing site cluster (19,21,22,25,29,34,81) forms a large band-like region on the back side of the EDN protein (see Figure 3.6(B)). Along the Hs branch, there are no site clusters formed between negatively selected events (see Figure 3.6(D)). Yet, although our method measures sequence-structure compatibility, the positively selected site cluster (64, 66, 68, 69) (see Figure 3.6(C)) corresponds precisely with the enzymatic region of the EDN protein (e.g., Zhang and Rosenberg 2002).

Unlike the behavior observed along the two branches above, substitution events estimated along the Pp branch form well defined site clusters when both the positively selected and negatively selected destabilizing events are considered. In


(A)
(B)
(C)
(D)

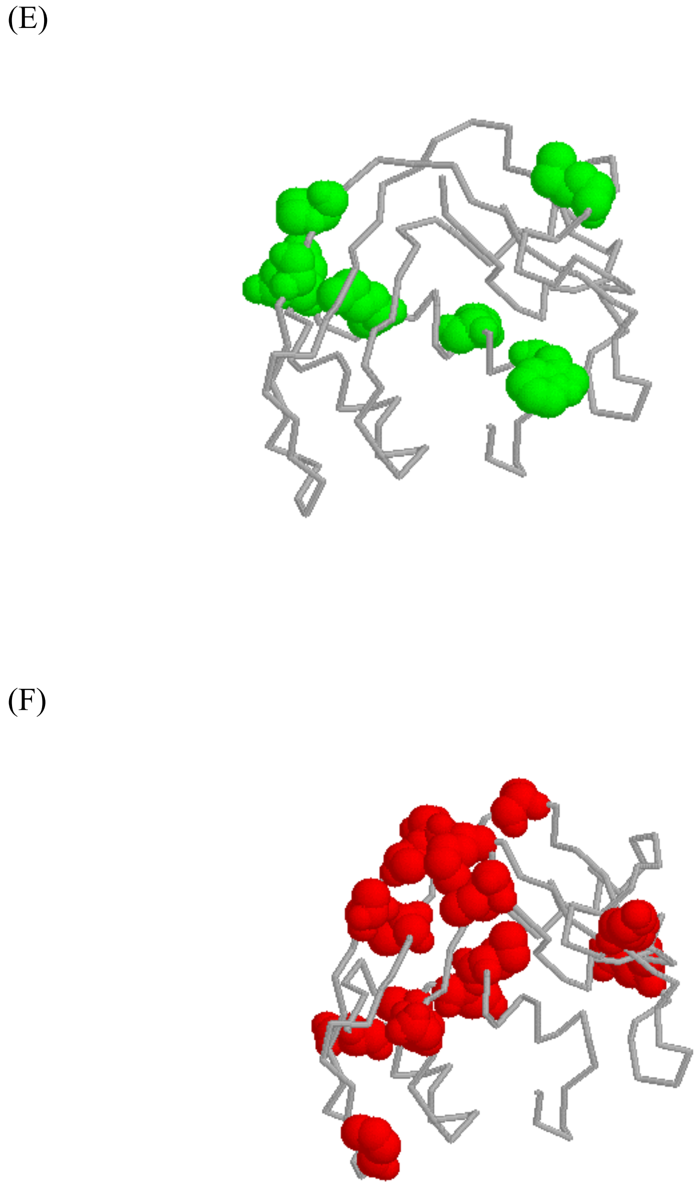


Figure 3.6: Using RASMOL (Sayle and Milner-White 1995), site clusters are visualized when both positively selected or negatively selected events are separately mapped onto the EDN structure. (A) Positively selected events inferred from *Saimiri sciureus*. (B) Negatively selected destabilizing events inferred from *Saimiri sciureus*. (C) Positively selected events from inferred *Homo sapiens*. (D) Negatively selected destabilizing events inferred from *Homo sapiens*. (E) Positively selected events inferred from *Pongo pygmaeus*. (F) Negatively selected destabilizing events inferred from *Pongo pygmaeus.*

particular, the positively selected site cluster (18, 21, 100, 102) is located in the upper left hand region of the protein and appears isolated from the other events selected in this group (see Figure 3.6(E)). However, the two negatively selected destabilizing site clusters given by (76, 81, 101, 103, 104, 105, 133) and (16, 17, 19, 25) appear as a tight grouping on top of the protein and as a horizontal band wrapping around the back of the protein, respectively in Figure 3.6(F). The undetermined biological significance of the site clusters inferred by this analysis deserve further experimental investigation.

Zhang and Rosenberg (2002) discovered that a 13-fold increase in EDN enzymatic activity could be attributed to the combined influence of nonsynonymous substitution events at codons 64 and 132, which map to adjacent sites in the protein structure. To examine this, they reconstructed the ancestral node protein with 64R and 132T using parsimony. Because the human sequence had 64S and 132R, they then synthesized the two intermediate proteins, namely I1:(64S , 132T) and I2:(64R , 132R), through site-directed mutagenesis and then tested each for RNase function (Zhang and Rosenberg 2002). They determined that I1 resulted in a 46% decrease in activity followed by a 24-fold increase, while I2 resulted in only a 21% decrease in activity followed by a 17-fold increase.

Reconstruction of the ancestral node using our approach also permits the derivation of the posterior distribution of paired codons 64 and 132. The most probable reconstruction is clearly CGT at codon 64 and ACC at codon 132, which is estimated with 0.8668 probability (see Table 3.6). The next three most probable reconstructions are also shown, and a category marked “Other” contains the remaining twenty reconstructions which collectively account for only 0.0195 posterior probability. One substitution at nucleotide position 190 in codon 64, as well as substitutions at nucleotides 395 and 396 in codon 132 are required to transform the ancestral node sequence to the human sequence. Thus, although Zhang and Rosenberg reconstructed two paths using proteins, by considering the associated DNA, a total of six unique path orders are determined to exist with our approach that transform CGT and ACC in the ancestral node to AGT and AGA in the observed human sequence.

From the 8,668 associated paths sampled, 1,009 contain multiple substitution events at these positions and are discarded for simplicity. Conditional on the paths that infer only one substitution event at each nucleotide position, the remaining paths are then investigated to account for the six possible orders (see Table 3.7). Using the human EDN sequence as a guide, each intermediate sequence defined within an order is manually constructed and evaluated for sequence-structure compatibility with our criterion. For each order, the pseudo-energy difference between the sequence before and after each substitution event is computed, where stabilizing substitutions result in negative pseudo-energy differences. This measure, as well as the estimated conditional posterior probability are shown in Table 3.7.

Orders I, II, III and IV collectively infer the nonsynonymous substitution event at codon 64 to occur prior to those in codon 132. By summing the conditional probabilities of these four orders, our approach estimates this result with estimated 0.6628 posterior probability. Conversely, orders V and VI infer the reverse scenario with combined 0.3372 posterior probability. The intermediate defined by I2 is formed by the second event along orders V and VI. Compared with the other intermediate sequences, our approach infers I2 to have the most unstable sequence-

Species	Codon 64	Codon 132
Human	AGT	AGA
Squirrel Monkey	CGT	ACC
Orangutan	CGT	ACC

	Codon 64	Codon 132	Probability
	CGT	ACC	86.68
Ancestral	CGT	ACA	4.27
Reconstruction	CGT	AGC	3.61
	AGT	ACC	3.49
Other			1.95

Table 3.6: Codons 64 and 132 are given for each taxon in the analysis. Using our dependence approach, the posterior distribution of the ancestral node is also shown for these positions in descending order of posterior probability, starting from the most probable. Twenty alternative reconstructions are included in the "Other" category.

Order	**Event 1**	**Event 2**	**Event 3**	**Conditional Probability**
Path I	190:A→C, (R→S)	395:C→G, (T→S)	396:C→A, (S→R)	16.24
Compatibility	− 2.96	+2.3	+6.26	
Path II	190:A→C, (R→S)	396:C→A, (T→T)	395:C→G, (T→R)	16.87
Compatibility	− 2.96	0	+8.56	
Path III	396:C→A, (T→T)	190:A→C, (R→S)	395:C→G, (T→R)	17.50
Compatibility	0	− 2.96	+8.56	
Path IV	395:C→G, (T→S)	190:A→C, (R→S)	396:C→A, (S→R)	15.67
Compatibility	+2.41	− 3.07	+6.26	
Path V	395:C→G, (T→S)	396:C→A, (S→R)	190:A→C, (R→S)	16.29
Compatibility	+2.41	+6.64	− 3.45	
Path VI	396:C→A, (T→T)	395:C→G, (T→R)	190:A→C, (R→S)	17.43
Compatibility	0	+9.05	− 3.45	

Table 3.7: For each of the three substitution events that comprise the six possible path orders, the nucleotide position, the associated nucleotide change and resultant amino acid change in parentheses are shown. The approximate influence of each substitution event on sequence-structure compatibility is also given where negative values represent stabilizing events. Conditional on the paths that infer only one substitution event at each nucleotide position, posterior probability estimates of each order are also shown.

structure compatibility measure. Although it appears to be difficult for our method to choose one of the six order with high posterior probability (see Table 3.7), by comparing the sum of the posterior probabilities our approach infers the event at codon 64 to occur before those in 132. In addition, our approach corroborates this general event order choice in Table 3.4(B), where the event at codon 64 is positively selected, while codon 132 is negatively selected and destabilizing to the protein sequence-structure compatibility.

Discussion

As a justification of our approach, estimates of positively selected events in the lysozyme c are found to agree with what has been determined previously (Stewart, Schilling and Wilson 1987; Yang and Nielsen 2002). Slight modifications of our method would allow a unique ω parameter to be estimated for each branch of the phylogeny. This could possibly isolate the branches in the lysozyme c topology that contain positively selected events. Similar to previously proposed maximum likelihood methods (e.g., Yang 1993), other extensions could be implemented to allow a discrete Gamma rate distribution to be estimated for ω. However, the benefit gained by these proposed modifications is difficult to determine.

In the EDN analysis, despite the absence of positive selection when independence among codons is assumed, several events on each branch were inferred to be positively selected due to the influence of surrounding codons. When mapped onto the structure, the formation of site clusters between events whose translated amino acids have C_β atoms within ten Angstroms of one another can often be inferred. In particular, the site cluster formed by positively selected events along the human EDN branch corresponds precisely with a region of enzymatic significance (e.g., Zhang and Rosenberg 2002). Both positively selected and negatively selected destabilizing event site clusters inferred by our method deserve further investigation experimentally.

When the alignment used by Zhang and Rosenberg (2002) is analyzed using parsimony, CGT and ACC are also inferred at positions 64 and 132, respectively (data not shown). However, because they worked with protein sequences and not the associated nucleotides, Zhang and Rosenberg lost some information that would have led to the discovery of alternative intermediary sequences. Using our approach, a total of six unique evolutionary path orders are determined. Although our method is not able to choose one particular order with high posterior probability, the substitution event at position 64 is estimated to occur prior to those of position 132 with 0.6628 posterior probability compared with only 0.3372 posterior probability for the reverse scenario. Furthermore, the event at site 64 is positively selected, while the events at site 132 are negatively selected and destabilizing to the protein. Thus, despite the greater decrease in enzymatic activity observed by Zhang and Rosenberg, our approach infers the event at position 64 prior to those of position 132.

A sequence-structure compatibility criterion (Jones, Taylor and Thornton 1992; Jones 1999) is used to approximate the fitness of a sequence, but because the method is written in general, other measures of sequence fitness can be accommodated. To ensure convergence of the Markov chain for three taxa, 5.1 million cycles are performed. How this increase will translate when more taxa are analyzed is difficult to determine. Although computationally intense, extending the capabilities of our approach to larger datasets is a top priority. To handle uncertainty in phylogenies with multiple ancestral nodes, algorithms that find the optimal joint ancestral sequence reconstruction will be explored (see Pupko et al. 2000). To contend with phylogenetic uncertainty, the phylogeny will most likely be determined by some external method, such as neighbor-joining (Saitou and Nei 1987), and assumed correct. Although Bayesian methods have been previously proposed (Huelsenbeck, Rannala and Masly 2000), significant modifications of our current implementation as well as improvements in computational speed will be necessary for this to be tractable.

References

Bielawski, J.P. and Z. Yang (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J. Struct. Func. Genomics* **3**:201-212.

Chang, B.S.W., K. Jonsson, M.A. Kazmi, M.J. Donoghue, and T.P. Sakmar (2002) Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19**(9):1483-1489.

Domachowske, J.B., K.D. Dyer, C.A. Bonville, and H.F. Rosenberg (1998) Recombinant human eosinophil-derived neurotoxin/RNase 2 functions as an effective antiviral agent against respiratory syncytial virus. *J. Infect. Dis.* **177**:1458-1464

Dykhuizen, D. and D. L. Hartl (1980) Selective neutrality of 6PGD allozymes i E. coli and the effects of geneetic background. *Genetics* **96**:801-817

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368 - 376

Felsenstein, J. (1989) Phylogenetic Inference Package (PHYLIP), version 3.2. University of Washington, Seattle. *Cladistics* **5**:164-166

Felsenstein, J. (2003) Inferring phylogenies. Sinauer, Sunderland, MA.

Goldman, N., and Z. Yang. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol.Biol.Evol.* **11**(5):725-736

Harata, K., Y. Abe, and M. Muraki (1998) Full - matrix least squares refinement of lysozymes and analysis of antisotropic thermal motion. *Proteins* **30**(3): 232-243

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

Huelsenbeck, J.P. and J.P. Bollback (2001) Empirical and hierarchical Bayesian estimation of ancestral states. *Syst. Bio.* **50**(3):351-366

Huelsenbeck, J.P., R. Nielsen and J.P. Bollback (2003) Stochastic mapping of morphological characters. *Syst. Biol.* **52**(2):131-158.

Huelsenbeck, J.P., B. Rannala and J.P. Masly (2000) Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **228**:2349-2350.

Jones, D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**:797-815

Jones, D.T., W.R. Taylor, and J.M. Thornton (1992) A new approach to protein fold recognition. *Nature* **358**:86-89

Kabsch, W., and C. Sander (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen - bonded and geometrical features. *Biopolymers* **22**(12):2577-2637.

Koshi, J.M., and R.A. Goldstein (1996) Probabilistic reconstructions of ancestral protein sequences. *J. Mol. Evol.* **42**:313-320.

Lee-Huang, S., P.L. Huang, Y. Sun, P.L. Huang, H.F. Kung, D.L. Blithe and H.C. Chen (1999) Lysozyme and RNase as anti-HIV components in β - core preparations of human chorionic gonadotropin. *Proc. Natl. Acad. Scio.* **96**: 2678-2681.

Li, W.-H. (1997) Molecular evolution. Sinauer, Sunderland, MA.

Messier, W., and C.-B. Stewart (1997) Episodic adaptive evolution of primate lysozymes. *Nature* **385**: 151-154.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**:1087-1092

Mosimann, S.C., D.L. Newton, R.J. Youle, and M.N. James (1996) X-ray crystallographic structure of recombinant eosinophil-derived neurotoxin at 1.83 A resolution. *J. Mol. Biol.* **260**: 540 - 552.

Muse, S.V., and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715 - 724.

Nielsen, R. (2001) Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics* **159**:401-411.

NIELSEN, R. (2002) Mapping mutations on phylogenies. *Syst. Biol.* **51**(5):729-739.

OHNO, S. (1970) Evolution by gene duplication (Springer-Verlag, Heidelberg, Germany)

POLLOCK DD, TAYLOR WR, GOLDMAN N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol.* Mar 19;**287** (1):187-98.

PUPKO, T., I. PE'ER, R. SHAMIR, D. GRAUER (2000) A fast algorithm for joint reconstruction of ancestral amino–acid sequences. *Mol. Biol. Evol.* **17**: 890-896

ROBINSON, D.M., D.T. JONES, H. KISHINO, N. GOLDMAN AND J.L. THORNE (2003) Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* **20**(10):1692-1704.

SAITO, N., AND M. NEI (1987) The neighbor joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406-425.

SAYLE, R.A. AND E.J. MILNER-WHITE (1995) RASMOL: biomolecular graphics for all. *Trends. Biochem. Sci.* **20**:374-376

STEWART, C.-B., SCHILLING, J.W., AND WILSON, A.C. (1987), Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**:401-404.

SWOFFORD, D.L., AND W.P. MADDISON (1987) Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* **87**:199-229.

Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis (1996) Phylogenetic inference. *In* “Molecular Systematics” (D.M. Hillis, C. Moritz, and B.K. Mable, Eds.), pp. 407-514. Sinauer, Sunderland, MA

Wollenberg, K.R. and W.R. Atchley (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *PNAS* vol. **97**. No. 7. Pp. 3288 - 3291.

Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**(6):1396-1401

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and applications to primate lysozyme evolution. *Mol. Biol. Evol.* **15**(5): 568-573.

Yang, Z. and R. Nielsen (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**(6):908-917.

Yang, Z., R. Nielsen, N. Goldman and A.-M.K. Pedersen (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431-449.

Yang, Z. and W.J. Swanson (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selection pressure among site classes. *Mol. Biol. Evol.* **19**(1):49-57.

Zhang, J. K.D. Dyer and H.F. Rosenberg (2000) Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc. Natl. Acad. Sci.* **97**(9): 4701 - 4706.

Zhang, J. and M. Nei (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood and distance methods. *J. Mol. Evol.* **44**(Suppl 1):S139-S146

Zhang, J. and H.F. Rosenberg (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc. Natl. Acad. Sci.* **99**(8): 5486-5491.

Zhang, J., Y.-p. Zhang and H.F. Rosenberg (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf eating monkey. *Nat. Genet.* **30**:411-415

Chapter 4

DISCUSSION

Introduction

Incorporating dependencies among sites offers many opportunities to explore the influence of protein structure on the substitution process. When structure is ignored (i.e., $s = p = 0$), the rate of change at a codon is not affected by the rate of change at surrounding codons. However, when structure is considered (i.e., $s \neq 0, p \neq 0$) a substantial amount of rate heterogeneity becomes apparent due to substitution events at codons that are in close physical proximity in the folded state. Although much insight is gained, certain assumptions required by our implementation, such as those concerning pseudo-energy potentials, have been considered drawbacks (see Thomas and Dill 1996). Extending this research to complex multi-taxa datasets is top priority, yet other possible future directions are also presented.

General Points

Many researchers have understood the importance of incorporating structural and functional information into models to improve evolutionary and phylogenetic inference (e.g., Koshi and Goldstein 1995, 1997; Thorne, Goldman and Jones 1996; Goldman, Thorne and Jones 1998; Liò et al. 1998). In particular, Fornasari, Parisi and Echave (2002) applied the method of Parisi and Echave (2001) to simulate amino acid changes on a protein with known structure. In their procedure, substitutions accepted under structural constraints were tabulated for six assumed site categories, from which probability matrices tailored to the protein of interest were constructed.

Although only six site categories were sufficient to model the protein under investigation, it seems difficult to assume that similar strategies would achieve the same level of success when when applied to alternate proteins. For instance, most proteins would require the creation of site specific rate matrices due to the existence of sites that do not fit into predefined categories. In addition, because the matrices are uniquely defined, the procedure must be repeated for each protein analyzed.

These drawbacks highlight many benefits of incorporating pseudo-energy potentials to measure the influence of nonsynonymous substitutions. Because the potentials are defined for residue pairs in general, they require no preliminary matrix calculations. This allows our approach to be immediately applicable to almost any protein of interest. Most importantly, since rates of change are tailored to each site depending on the location and identity of all residues in close physical proximity, broad rate category assumptions are never made.

Frozen Structure Assumption

In our approach, the experimentally determined protein structure of one taxon is assumed correct for the remaining taxa in the dataset. It is common for a particular dataset to contain only one taxon with a resolved structure because of the low success rate associated with crystalizing proteins. Despite this, the given structure may not be perfectly compatible with other sequences in the dataset due to differences in residue composition. Rather than the precise three dimensional atomic coordinates, it has been suggested that evolutionary models should use secondary structural elements (Mizuguchi and Go 1995). Yet, there is evidence that even secondary structure may be variable, as observed in different strains of at least one HIV-1 protein (Hansen et al. 1996). Protein structure evolves very slowly (e.g., Chothia and Lesk 1986; Flores et al. 1993; Russell et al. 1997), and although the frozen structure assumption is not perfect, it does allow the definition of a simple dependence function based on the physical separation of amino acid residues.

Assumptions with Pseudo-Energy Potentials

The most unique feature of our approach is incorporating pseudo-energy potentials to measure the influence of nonsynonymous substitution events on surrounding residues within a protein. These potentials were successfully applied to protein threading (Jones, Taylor and Thornton 1992; Jones 1999), but were derived using a variety of assumptions that are not without critics. The arguments stem mainly from the precise definition of a pairwise interaction and include from which point the distance between residues should be measured; how close residues must be to interact; and between which atom pairs should the potentials be defined.

To obtain robust pseudo-energy estimates, the database from which the potentials are derived must be sufficiently large. If not, the error associated with each energy potential would be substantial. Consequently, this would decrease the overall sensitivity of the potentials to discern the difference between valid structures or which substitution events are considered most plausible. At the time when Sippl (1990), Hendlich et al. (1990) and Jones, Taylor and Thornton (1992) performed their analyses, the number of solved crystal structures was limited. However each group correctly reasoned that their method would only improve when derived from a larger, more diverse training set of structures. The pseudo-potentials used in our procedure were borrowed from the GenTHREADER software package, which were derived from the unique set of structures available in the 1998-1999 PDB database (Jones 1999).

The origin from which pairwise interactions should be measured is biologically not known. This has a large affect on subsequent analyses for it concerns which residue pairs inevitably interact. Originally, the C_α atom was proposed to be the center point (Sippl 1990), but it was later determined that the C_β atom was a more sensitive probe of conformational preferences of the individual residues within proteins (Hendlich et al. 1990). Because the biochemical properties of each residue are derived form the unique side chain conformations, choosing the C_β atom as this reference point may agree more with biological considerations. However, the fact that a fictitious C_β atom must be fabricated for each glycine residue is an obvious drawback (David Jones, personal communication).

Once the reference point is chosen, the question of exactly how close two residues must be to interact can be addressed. In 1993, Sippl proposed that pairwise interactions should be defined within a hollow sphere measuring between four and fifteen Angstroms. This reasonable assumption neglects the possibility of extremely close contacts, which rarely occur due to the possibility of side chain overlap, and extremely distant contacts, which are mainly determined by large proteins in the database used to compile the potentials (Sippl 1993). However, due to the electrostatics of interaction as well as the fear of introducing unforseen side effects into the potentials, a more conservative upper bound of ten Angstroms was used in the creation of the potentials for our analyses (Jones, Taylor and Thornton 1992; Jones 1999).

In general, the potentials are defined between pre-specified atoms within interacting amino acid residue pairs. Compared with utilizing the $C_\alpha \Rightarrow C_\alpha$ pair (Sippl 1990) or the $C_\beta \Rightarrow C_\beta$ pair (Hendlich et al. 1990), the pseudo-potentials we incorporate are derived for five atom pairs, namely the $C_\beta \Rightarrow C_\beta, C_\beta \Rightarrow O, C_\beta \Rightarrow N, O \Rightarrow C_\beta, N \Rightarrow C_\beta$ (Jones, Taylor and Thornton 1992; Jones 1999). Using five atom pairs benefits the potentials by clearly defining the location and relationship between interacting residues without a comparable increase in computational demand.

Although this method makes a significant step towards incorporating dependence among codons, the pseudo-energies of interaction are computed only for residue pairs, which completely ignores that a third amino acid residue in close proximity may impart an unforseen influence on the pairwise energy score. However, the amount of observations necessary to estimate multi-residue effects with any accuracy is unfortunately prohibitive. Until the number of structures necessary to obtain reasonable estimates, or more realistic energy calculations become available, we are forced to assume that two residues behave as though in a vacuum, interacting independent of their surroundings.

Lastly, this type of modelling ignores all contextual effects and environmental constraints of pairwise interactions. For example, two residues in the same physical conformation will have identical pairwise-interaction energy score regardless of their location in the protein structure. The roots of this dilemma are from the Inverse Boltzmann Principle which functions to average over a wide range of interacting conformations as well as environmental conditions to compute the pseudo–energy potentials. Opponents hone in on these inherent weaknesses as being too unrealistic to be useful (see Thomas and Dill 1996). Although it is far from perfect, our procedure is written in general and can quickly approximate how well amino acid pairs interact no matter the protein under investigation. Lastly, because our model is adaptable, if a more robust set of pseudo-potentials become available, they can be easily accommodated by the method.

Future Directions

A number of different extensions to, or applications of, this approach are being discussed while others are actively being pursued. Rather than use event histories, one possible direction is to perform inference solely with path order. This can be achieved by integrating out the time component of a path leaving only the general event order. Other directions may include applications to serially sampled HIV sequences (Seo et al. 2002), to divergence times estimation (e.g., Thorne, Kishino and Painter 1998; Thorne and Kishino 2002; Aris-Brossou and Yang 2002), to transmembrane proteins (Liò and Goldman, 1999), or to protein families in the publicly available "Pandit" database (Whelan et al., in press). Before these ideas can be followed however, the first step is to extend the procedure to complex multi-taxa phylogenies. Although this improvement may not be of interest statistically, a dramatic increase in applicability is envisioned once this restriction is lifted. Furthermore, this will also allow our method to be applicable to the general scientific community, whose interest lies in the relationship between several taxa.

Extension to General Phylogenies

Datasets with many taxa require the simulation of paths on entire phylogenetic trees. The current implementation with three taxa treats all branches as independent entities who share a common ancestral node. Although it is applied to small datasets, this same algorithm can analyze greater than three sequences with the caveat that they evolve according to a star phylogeny. To study bifurcating or multifurcating tree topologies, internal nodes other than the root must be treated as descendant sequences as well as ancestral node sequences. To analyze large datasets, a modified version of the "Pupko algorithm" (Pupko et al. 2000) would be required to perform the simultaneous update of all ancestral nodes. Rather than estimate the probability of a single nucleotide at each node, slight modifications of the algorithm would allow the nucleotide choice to be stochastic.

Gamma Shape Parameter for ω

Our current implementation estimates a single ω parameter that is shared among all codons. This parameter is intended to capture the forces that affect nonsynonymous change which are external to the protein. However, it is probable that this force varies among codons. A large improvement in evolutionary models was achieved by allowing a discrete Gamma rate distribution to be estimated for ω (Yang 1993). This improvement was not only observed in how well models fit real sequence data, but also in enhanced reconstruction of more reliable evolutionary trees (Yang 1994, 1996). A similar strategy can be incorporated into our procedure, though how much improvement it will add is difficult to determine.

Integrating Time Out of a Path

A path is defined by a series of substitution events and their associated times. However, two paths that have the same substitution event order, yet differ only in event timing will be treated as distinct entities. Because time is a continuous quantity, there exists an infinite number of paths associated with each unique event order. A general expression can be obtained by integrating time out of the path equation, as long as the substitution event order is preserved:

$$\begin{aligned} \int_{t_1=0}^{T} \cdots \int_{t_q=t_{q-1}}^{T} & R_{1,2}\; e^{-R_{1,\bullet}\; t_1} \cdots R_{q,q+1}\; e^{-R_{q,\bullet} t_q}\; e^{-R_{q+1,\bullet}(T-t_q)} dt_q \cdots dt_1 \\ &= \left(\prod_{z=1}^{q} R_{z,z+1} \right) \sum_{\xi=1}^{q+1} \frac{e^{-R_{\xi,\bullet} T}}{\left[\prod_{\substack{\phi=1 \\ \phi \neq \xi}}^{q+1} (R_{\phi,\bullet} - R_{\xi,\bullet}) \right]} \end{aligned} \qquad (4.1)$$

where T is the normalized time of the path, q is the number of substitution events, $R_{z,z+1}$ is the instantaneous rate of change from sequence z to sequence $z+1$ as defined by our instantaneous rate matrix, and $R_{z,\bullet}$ is the rate of change away from sequence z to any other sequence of length N given by $R_{z,\bullet} = \sum_k R_{z,k}$, where the sum is over all sequences k that differ from z. (see Robinson et al. 2003). The proof of Equation (4.1) is given in Appendix A. If all rates $R_{z,\bullet}$ are equal, this equation reduces to a general Poisson process with rate parameter $R_{z,\bullet}$. Unfortunately, the formula becomes undefined if the rates $R_{\phi,\bullet}$ and $R_{\xi,\bullet}$ are equal. The general form of this equation for any number of rate equivalencies however, has yet to be determined.

Transmembrane Proteins

The applicability of our approach has been focused solely on globular proteins. Although hidden Markov models have been previously employed to model transmembrane proteins (Liò and Goldman, 1999), the application of our approach to this unique class of proteins has never been proposed. The functional constraints within the transmembrane versus globular protein residues appear almost contradictory, as observed in the mutation matrices calculated from membrane spanning proteins (Jones et al. 1994). Because the energy potentials of our approach were derived solely from globular proteins, the information coming from the external residues compared with those that come from residues that span the membrane would most likely conflict. This would have unforseen consequences on our method resulting mainly in misleading inference.

Conclusion

Relaxing the independence assumption is a natural step in the long progression of statistical models of evolution. As computer processor speed and the availability of protein crystal structures increase, so does the applicability of our approach. With two and three taxa datasets, biologically plausible parameter estimates are derived, and observations of the evolutionary history along a branch are inferred. In particular, events were place with high posterior probability and events inferred to be under purifying selection when independence among codons is assumed are estimated to be positively selected by our method due to the influence of pairwise interactions. In addition, through mapping either positively selected or negatively selected events onto the known EDN structure, spatial clusters of positively or negatively selected events can be inferred. Although extensions to greater than three sequences is the logical next step, the future directions of this approach are limitless.

References

Aris-Brosou, S. and Z. Yang (2002) Effects of models of rate evolution on estimation of devergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst. Biol.* **51**(5):703-714

Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. EMBO. J. 5:519-527

Flores, T. P., C. A. Orengo, D. S. Moss, and J. M. Thornton. 1993. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 2:1811-1826

Fornasari, M.S., G. Parisi, and J. Echave (2002) Site specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol. Biol. Evol.* **19**(3):352-356

Goldman, N., J.L. Thorne, and D.T. Jones (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263**:196-208

Goldman, N., J.L. Thorne, and D.T. Jones (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**:445-458

Goldman, N., and Z. Yang. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol.Biol.Evol.* **11**(5):725-736

Hansen, J.E., O. Lund, J.O. Nielsen, S. Brunak, and J.-E.S. Hansen (1996) Prediction of the secondary structure of HIV-1 gp120 *Proteins* **25**(1):1-11

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

Hendlich, M., P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M.J. Sippl (1990) Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.* **216**:167-180

Jones, D.T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**:797-815

Jones, D.T., W.R. Taylor, and J.M. Thornton (1992) A new approach to protein fold recognition. *Nature* **358**:86-89

Jones, D.T., W.R. Taylor, and J.M. Thornton (1994) A mutation data matrix for transmembrane proteins. *FEBS lett.* **339**(3):269-275

Koshi, J.M., and R.A. Goldstein (1995) Context dependent optimal substitution matrices. *Prot. Eng.* **8**:641-645

Koshi, J.M. and R.A. Goldstein (1997) Mutation matrices and physical - chemical properties: correlations and implications. *Proteins* **27**:336-344

Liò, P., N. Goldman, J.L. Thorne, and D.T. Jones (1998) PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* **14**(8):726-733

Liò, P. and N. Goldman (1999) Using protein structural information in evolutionary inference: Transmembrane proteins. *Mol. Biol. Evol.* **16**(12):1696-1710

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**:1087-1092

Mizuguchi, K. and N. Go (1995) Comparison of spatial arrangements of the secondary structure elements in proteins. *Protein Eng.* **8**:353-362

Muse, S.V., and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715 - 724

Parisi G. and J. Echave (2001) Structural Constraints and Emergence of Sequence Patterns in Protein Evolution. *Mol. Biol. Evol.* **18**(5):750-756.

Pupko, T., I. Pe'er, R. Shamir, D. Grauer (2000) A fast algorithm for joint reconstruction of ancestral amino–acid sequences. *Mol. Biol. Evol.* **17**: 890-896

Russell, R.B., M.A.S. Saqi, R.A. Sayle, P.A. Bates, and M.J.E. Sternberg (1997) Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269**:423-439

Seo, T.K., J.L. thorne, M. Hasegawa and H. Kishino (2002) A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* **18**(1):115-123

Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins *J. Mol. Biol.* **213**:859-883

Sippl, M.J. (1993) Recognition of errors in three - dimensional structures of proteins. *PROTEINS: Structure, Function and Genetics* **17**:355-362

Thomas, P.D. and K.A. Dill. (1996) Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **257**: 457-469

Thorne, J.L., N. Goldman and D.T. Jones (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**(5):666-673

Thorne, J.L., H. Kishino and I.S. Painter (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**(12):03-714

Thorne, J.L. and H. Kishino (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* **51**(5):689-702.

Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**(6):1396-1401

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**:306-314

Yang, Z. (1996) Among site rate variation and its impact on phylogenetic analysis. *TREE* **11**:367-372

Appendix A

THEOREM:

Given an ancestral sequence i and descendant sequence j separated by an evolutionary amount of time T, the infinite set of paths that share the same q substitution events with the same order can be calculated by integrating out time for each event with the following formula

$$\begin{aligned}\int_{t_1=0}^{T} \cdots \int_{t_q=t_{q-1}}^{T} & R_{1,2}\; e^{-R_{1,\bullet}\; t_1} \cdots R_{q,q+1}\; e^{-R_{q,\bullet} t_q}\; e^{-R_{q+1,\bullet}(T-t_q)} dt_q \cdots dt_1 \\ = & \left(\prod_{z=1}^{q} R_{z,z+1}\right) \sum_{\xi=1}^{q+1} \frac{e^{-R_{\xi,\bullet} T}}{\left[\prod_{\substack{\phi=1\\ \phi\neq\xi}}^{q+1} (R_{\phi,\bullet} - R_{\xi,\bullet})\right]}\end{aligned} \tag{4.2}$$

where $R_{z,z+1}$ is the instantaneous rate of change from sequence z to sequence $z+1$ as defined by our instantaneous rate matrix, and $R_{z,\bullet}$ is the rate of change away from sequence z to any other sequence of length N given by $R_{z,\bullet} = \sum_k R_{z,k}$, where the sum is over all sequences k that differ from z. (see Robinson et al. 2003).

PROOF BY INDUCTION:

If the rate away from the ancestral sequence is defined as $R_{1,\bullet}$, then the path over time interval T with zero substitution events is calculated as:

$$\Pr(q \;=\; 0) \;=\; e^{-R_{1,\bullet} T} \tag{4.3}$$

Notice that with a constant rate over the entire interval, the evolutionary path follows a simple Poisson process. However due to dependence among sites, this form does not hold because the rate of change usually differs after each substitution event. There exists two cases for the addition of one event. The first is for a path with a single substitution event:

$$\begin{aligned}\Pr(q \;=\; 1) = \quad & \textstyle\int_{t_1=0}^{T} R_{1,2}e^{-R_{1,\bullet}t_1}\ e^{-R_{2,\bullet}(T-t_1)}dt_1 \quad && = R_{1,2}\ e^{-R_{2,\bullet}T}\left(\frac{e^{-(R_{1,\bullet}-R_{2,\bullet})T}-1}{-(R_{1,\bullet}-R_{2,\bullet})}\right) \\ = \quad & R_{1,2}\left(\tfrac{e^{-R_{1,\bullet}T}}{R_{2,\bullet}-R_{1,\bullet}}+\tfrac{e^{-R_{2,\bullet}T}}{R_{1,\bullet}-R_{2,\bullet}}\right) && = R_{1,2}\ \sum_{\xi=1}^{2}\frac{e^{-R_{\xi,\bullet}T}}{\prod_{\substack{q=1\\q\neq\xi}}^{2}(R_{q,\bullet}-R_{\xi,\bullet})} \qquad (4.4)\end{aligned}$$

Without loss of generality, inserting an event in an interval that already contains q prior substitutions can be thought of as adding an event between the last substitution and final time T.

$$\begin{aligned}Pr(q=q+1,\tilde{T}=T-t_q) &= \int_{t_{q+1}=t_q}^{T} R_{q+1,q+2}e^{-R_{q+1,\bullet}t_{q+1}}e^{-R_{q+2,\bullet}(T-t_{q+1})}dt_{q+1} \\ &= R_{q+1,q+2}e^{-R_{q+2,\bullet}T}\left[\frac{e^{-(R_{q+1,\bullet}-R_{q+2,\bullet})t_{q+1}}}{-(R_{q+1,\bullet}-R_{q+2,\bullet})}\right]_{t_q}^{T} \\ &= R_{q+1,q+2}e^{-R_{q+2,\bullet}T}\ \frac{\left(e^{-(R_{q+1,\bullet}-R_{q+2,\bullet})T}-e^{-(R_{q+1,\bullet}-R_{q+2,\bullet})t_q}\right)}{-(R_{q+1,\bullet}-R_{q+2,\bullet})} \\ &= R_{q+1,q+2}\left(\frac{e^{-R_{q+1,\bullet}T}}{R_{q+2,\bullet}-R_{q+1,\bullet}}+\frac{e^{-R_{q+2,\bullet}T}\ e^{-(R_{q+1,\bullet}-R_{q+2,\bullet})t_q}}{R_{q+1,\bullet}-R_{q+2,\bullet}}\right) \qquad (4.5)\end{aligned}$$

By using induction, Equation (4.2) is assumed correct for q substitution events and will now be shown to hold for $(q+1)$ events. Hence, we start with the general equation for integrating out the time component for $q+1$ substitution events.

$$\begin{aligned}\int_{t_1=0}^{T}\cdots\int_{t_q=t_{q-1}}^{T}\int_{t_{q+1}=t_q}^{T} R_{1,2}\ e^{-R_{1,\bullet}\ t_1}\quad \cdots \quad & R_{q,q+1}\ e^{-R_{q+1,\bullet}(t_{q+2}-t_{q+1})}\ . \\ & R_{q+1,q+2}\ e^{-R_{q+2,\bullet}(T-t_{q+2})}\ dt_{q+1}\cdots dt_1\end{aligned}$$

By pulling all terms that do not have a t_{q+1} component through the rightmost integral, we arrive at:

$$\begin{aligned}= \quad \left(\prod_{z=1}^{q}R_{z,z+1}\right)\int_{t_1=0}^{T}\cdots\int_{t_q=t_{q-1}}^{T} & e^{-(R_{1,\bullet}-R_{2,\bullet})t_1}\cdots e^{-(R_{q,\bullet}-R_{q+1,\bullet})t_q} \\ & \left(\int_{t_{q+1}=t_q}^{T}R_{q+1,q+2}e^{-R_{q+1,\bullet}t_{q+1}}e^{-R_{q+2,\bullet}(T-t_{q+1})}dt_{q+1}\right)dt_q\cdots dt_1\end{aligned}$$

$$
\begin{aligned}
= \quad & \left(\prod_{z=1}^{q+1} R_{z,z+1}\right) \int_{t_1=0}^{T} \cdots \int_{t_q=t_{q-1}}^{T} e^{-(R_{1,\bullet}-R_{2,\bullet})t_1} \cdots e^{-(R_{q,\bullet}-R_{q+1,\cdot})t_q} \\
& \left(\frac{e^{-R_{q+1,\bullet}T}}{R_{q+2,\bullet}-R_{q+1,\bullet}} + \frac{e^{-R_{q+2,\bullet}T}\ e^{-(R_{q+1,\bullet}-R_{q+2,\bullet})t_q}}{R_{q+1,\bullet}-R_{q+2,\bullet}}\right) dt_q \cdots dt_1 \qquad (4.6)
\end{aligned}
$$

where Equation (4.6) is found by applying the result found in Equation (4.5) to the integral within the parentheses above. By distributing the integration signs across each term in the parentheses of Equation (4.6), we arrive at Equation (4.7)

$$
\begin{aligned}
= \quad & \frac{\left(\prod_{z=1}^{q+1} R_{z,z+1}\right)}{(R_{q+2,\bullet}-R_{q+1,\bullet})} \int_{t_1=0}^{T}\int_{t_2=t_1}^{T} \cdots \int_{t_q=t_{q-1}}^{T} e^{-(R_{1,\bullet}-R_{2,\bullet})t_1} \cdots e^{-(R_{q,\bullet}-R_{q+1,\bullet})t_q} e^{-R_{q+1,\bullet}T} dt_q \cdots dt_1 \\
+ \quad & \frac{\left(\prod_{z=1}^{q+1} R_{z,z+1}\right)}{(R_{q+1,\bullet}-R_{q+2,\bullet})} \int_{t_1=0}^{T}\int_{t_2=t_1}^{T} \cdots \int_{t_q=t_{q-1}}^{T} e^{-(R_{1,\bullet}-R_{2,\bullet})t_1} \cdots e^{-(R_{q-1,\bullet}-R_{q,\bullet})t_{q-1}} \\
& e^{-(R_{q,\bullet}-R_{q+2,\bullet})t_q} e^{-R_{q+2,\bullet}T} dt_q \cdots dt_1 \qquad (4.7)
\end{aligned}
$$

The form of the first integral in Equation (4.7) precisely defines the path with q events and is thus assumed to follow Equation (4.2). The second integral can also be integrated using Equation (4.2), except care must be given to properly account for the subscript (q+2). By combining like terms found in Equation (4.8), followed by subsequent algebraic simplifications of Equation(4.9), Equation (4.10) is derived via the calculation of a common denominator:

$$
\begin{aligned}
= & \frac{\left(\prod_{z=1}^{q+1} R_{z,z+1}\right)}{(R_{q+2,\bullet}-R_{q+1,\bullet})} \left(\left[\sum_{\xi=1}^{q} \frac{e^{-R_{\xi,\bullet}T}}{\left[\prod_{\substack{\phi=1\\ \phi\neq\xi}}^{q}(R_{\phi,\bullet}-R_{\xi,\bullet})\right](R_{q+1,\bullet}-R_{\xi,\bullet})}\right] + \frac{e^{-R_{q+1,\bullet}T}}{\prod_{\phi=1}^{q}(R_{\phi,\bullet}-R_{q+1,\bullet})}\right) \\
- & \frac{\left(\prod_{z=1}^{q+1} R_{z,z+1}\right)}{(R_{q+2,\bullet}-R_{q+1,\bullet})} \left(\left[\sum_{\xi=1}^{q} \frac{e^{-R_{\xi,\bullet}T}}{\left[\prod_{\substack{\phi=1\\ \phi\neq\xi}}^{q}(R_{\phi,\bullet}-R_{\xi,\bullet})\right](R_{q+2,\bullet}-R_{\xi,\bullet})}\right] + \frac{e^{-R_{q+2,\bullet}T}}{\prod_{\phi=1}^{q}(R_{\phi,\bullet}-R_{q+2,\bullet})}\right) \quad (4.8)
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \frac{\prod_{z=1}^{q+1} R_{z,z+1}}{(R_{q+2,\bullet}-R_{q+1,\bullet})} \sum_{\xi=1}^{q} \frac{e^{-R_{\xi,\bullet}T}}{\prod_{\substack{\phi=1\\ \phi\neq\xi}}^{q}(R_{\phi,\bullet}-R_{\xi,\bullet})} \left[\frac{1}{R_{q+1,\bullet}-R_{\xi,\bullet}} - \frac{1}{R_{q+2,\bullet}-R_{\xi,\bullet}}\right] \\
& + \frac{\left(\prod_{z=1}^{q+1} R_{z,z+1}\right)}{(R_{q+2,\bullet}-R_{q+1,\bullet})} \left[\frac{e^{-R_{q+1,\bullet}T}}{\prod_{\phi=1}^{q}(R_{\phi,\bullet}-R_{q+1,\bullet})} - \frac{e^{-R_{q+2,\bullet}T}}{\prod_{\phi=1}^{q}(R_{\phi,\bullet}-R_{q+2,\bullet})}\right] \quad (4.9) \\
= \quad & \left(\prod_{z=1}^{q+1} R_{z,z+1}\right) \sum_{\xi=1}^{q} \frac{e^{-R_{\xi,\bullet}T}}{\prod_{\substack{\phi=1\\ \phi\neq\xi}}^{q+2}(R_{\phi,\bullet}-R_{\xi,\bullet})}
\end{aligned}
$$

$$+\frac{\left(\prod_{z=1}^{q+1} R_{z,z+1}\right)}{(R_{q+2,\bullet} - R_{q+1,\bullet})}\left[\frac{e^{-R_{q+1,\bullet}T}}{\prod_{\phi=1}^{q}(R_{\phi,\bullet} - R_{q+1,\bullet})} - \frac{e^{-R_{q+2,\bullet}T}}{\prod_{\phi=1}^{q}(R_{\phi,\bullet} - R_{q+2,\bullet})}\right] \tag{4.10}$$

Because the pattern associated with the first component of Equation (4.10) is identical to that observed in the remaining terms, the final form is derived simply by extending the upper limit of the product in the denominator to include both the $(q+1)^{st}$ and $(q+2)^{nd}$ terms which proves the theorem as desired.

$$= \left(\prod_{z=1}^{q+1} R_{z,z+1}\right) \sum_{r=1}^{q+2} \frac{e^{-R_{\xi,\bullet}T}}{\prod_{\substack{\phi=1\\ \phi\neq\xi}}^{q+2}(R_{\phi,\bullet} - R_{\xi,\bullet})} \tag{4.11}$$