ABSTRACT

CHOI, SANG CHUL. Statistical Inference and Biological Interpretation via Comparatively Realistic Models of Molecular Evolution. (Under the direction of Professor Jeffrey L. Thorne).

Recently, advances in statistical inference techniques have allowed analyses of molecular evolution to proceed without the biologically implausible assumption of independent change among DNA sequence sites. These techniques permit incorporation of molecular phenotypes such as RNA secondary and protein tertiary structure directly into the models of DNA sequence evolution, and they thereby facilitate assessment of the impact of molecular phenotype on the rates of sequence evolution. Our analysis of 1,195 non-redundant protein-coding sequences suggests that solvent accessibility and pairwise interactions among amino acids have important and roughly comparable impacts on the rates of evolution. We show how solvent accessibility and pairwise amino acid interactions can be used with protein-coding single nucleotide polymorphism (SNP) data to predict which SNP allele is ancestral and which is derived. Our analysis of 142 non-synonymous SNPs indicates that ancestral alleles are more selectively advantageous with respect to tertiary structure than are derived alleles. In other work, we show how recently developed models of molecular evolution with dependent change among sites can be adapted to generate stationary distributions that match a desired variable length Markov model or profile hidden Markov model for protein sequence organization. Departures between a neutral model for protein evolution and the variable length Markov model or profile hidden Markov model are attributed to natural selection. We show how these departures lead to a crude approximation of the product of effective population size and the difference in relative fitnesses between sequences.
Statistical Inference and Biological Interpretation via Comparatively Realistic Models of Molecular Evolution

by

Sang Chul Choi

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, North Carolina

2007

Approved By:

______________________________  ______________________________
Brian M. Wiegmann                Eric A. Stone

______________________________  ______________________________
David M. Bird                    Jeffrey L. Thorne
Chair of Advisory Committee
Dedication

To my parents
Biography

Sang Chul Choi was born in Pusan, Korea on November 12 of 1974 in the solar calendar (September 29, 1974 in the lunar calendar), to Young Choon Choi, and Jae Ok Kim. He spent his earliest years growing up in Munhyeon-dong, located in middle of Pusan. He spent his kindergarten and his first grade of elementary school in Sajik-dong, located in eastern Pusan. During his second grade of elementary school, he lived in Beomil-dong, which is located in the middle of Pusan. He spent his third grade of elementary school up to the first grade of middle school in Gwaebeop-dong, located in western Pusan. From his second grade of middle school to high school, he lived in Daesin-dong, located in southwest of Pusan.

After graduating from Daedong high school in 1992, Sang Chul attended Pohang Institute of Science and Technology at Pohang (POSTECH), located 100 miles north of Pusan, and majored in Life Science. People talked about biotechnology in the then-coming 21st century. He was introduced to modern biology during his undergraduate course work. He admired a few genius friends who majored in Electronic and Electrical Engineering, and was intrigued by electronic engineering courses, which led him to co-major in Electronic and Electrical Engineering. He was also intrigued by computer gurus, who talked about operating systems (e.g., BeOS, NEXT, UNIX, Microsoft DOS) and command-line-based programs (e.g., VI editor, gcc, make). This was the time when he “coined” his almost unique email ID, goshng, which means run as fast as winds with zooming sound in Korean.

He used to love singing grass-roots activist songs, and used to play guitar and the piano by teaching himself. He also had contact with a local music band group at Pohang. He led an intramural club called Hanapae, which means a band that performs sorrowful music about the modern history of the Korean Peninsula. Because of antagonism toward the grass-roots activist songs in the university, undergraduate students did not like to join the club, and support of the university to the club was poorer than those of other musical clubs of the university. This made him meet and negotiate with the officer of academic affairs of the university, who was half-persuaded to support financially the club for the years to come.

Two requisites for his graduation were the requirement of Bachelor’s degree in Electronic and Electrical Engineering to do undergraduate research and the requirement of Bachelor’s degree in Life Science to complete an undergraduate thesis. The flower development laboratory that he joined to do undergraduate thesis work was located a mile
northeast of the university, near Pohang Accelerator Laboratory. The security and information technology laboratory that he joined to do undergraduate research was located a half mile southwest of the university, near the Research Institute of Science and Technology. His fifth and last year of undergraduate life was not fruitful or enjoyable because he would spend days doing genetic transformation of chrysanthemums on the workbench, and nights working on mysterious cryptography in front of the computer. Despite the brutal undergraduate research environment, he tried to learn a more advanced level of mathematics, and attempted to involve himself in biological research.

After completing his five-year-long undergraduate career, he joined a private company called Geonong, which means humongous agriculture. Although the name of the company sounds like it is related to bio-tech, it was a small software company aiming to serve big corporations such as telecommunication companies, education broadcasting systems, and the like. He served longer than three years to fulfill his compensating compulsory military duty. He organized team-based software development rather than programming work, and cultivated his skill for software development project leader.

He then attended an intensive course of bioinformatics which was hosted by the Korea National Genome Research Institute for several months. This was essentially the first time that he was introduced to the broad area of bioinformatics. Unlike usual software development companies, bioinformatics related companies and research centers alike were not thought to provide well-paid jobs. However, he struggled searching for jobs in both academia and industry for almost two years. This was not a bad time for him because he found his true love, Jina, and this was the second moment when he was thrilled in his life.

This period turned his view toward studying abroad in the United States of America. After receiving numerous politely-worded rejection letters from graduate schools to which he applied, one letter from North Carolina State University gave him a beam of hope under the long dark tunnel of finding this next career. A friend of his and he tried to find a person who would be his advisor, and found a professor named Dr. Jeffrey L. Thorne at the graduate school. He decided to study evolution of biological sequences right away, and emailed Dr. Thorne that he wanted to pursue his doctoral degree under the supervision of Dr. Thorne. This was quite opposite to the way Douglas Robinson found his advisor.

The only thing that he thought he needed to learn was statistics. The Bioinformatics program at North Carolina State University looked like the place where he would spend years. Unfortunately, the acceptance letter from Dr. Barbara Sherry did not contain
anything about financial support for the graduate program. Someone let him know that
there might be a scholarship for him. It was the Graduate Study Abroad Scholarship of-
erered by Korean Science and Engineering Foundation (KOSEF). There were two critical
restrictions that could have hampered him applying for the scholarship: a weird GPA score
in percentage, and age limit. He was a little older than others who would study abroad
because of his industry career period and he was not an excellent student although he won
almost full-year undergraduate scholarship. The undergraduate university was so generous
to provide scholarship, and yes it was quite right to say that it was a “fine” school. The two
limitations fortunately turned out to be lifted because KOSEF considered him as the most
aged applicant so that he had his last chance to apply for the scholarship, and the academic
records service of POSTECH informed him of the fact that his grade point average (GPA)
looked low, but his academic performance met the requirement of the scholarship applica-
tion. He felt like that he already won the scholarship. Indeed, he was one of the almost 300
scholarship winners. This was his third moment in his life that he was ever thrilled, and
the first moment was back into 1993 when he was one of the 300 who were admitted to his
undergraduate university.

Four years have passed since he touched his foot on Raleigh, North Carolina. He
has been armed with a little mathematical statistics that he had wanted. He has a goal at
the present, that is to make his American dream come true. It is being a homeowner with
his beloved girl friend, Jina, so that he will be a good father for his two baby girls, Jena and
Juliana. He will be much happier if he can make it through staying at academia and being
a professor like Dr. Jeffrey Thorne. He is pretty sure that he cannot be like his advisor, but
he will mimic his advisor during his life.
Acknowledgements

Jeffrey Thorne has my gratitude for his patience, guidance, and friendship through the most beautiful life ever of my graduate study. I also thank faculty members in Bioinformatics Research Center, especially Bruce Weir for his kind consideration of possible financial support, Dahlia Nielsen for her always welcoming my sporadic interruption, Zhao-Bang Zeng for his service of Bioinformatics Program, Spencer Muse for his coordination of Bioinformatics Retreat to Great Smoky Mountain, Steffen Heber for his help in computer prediction techniques, and Eric Stone for his creative and thoughtful help. I thank William Atchley and Philip Awadalla for their lively conversation with Jeffrey Thorne so that I could overhear. I thank Barbara Sherry for her letter that said that I could start my graduate study.

I thank Juliebeth Briseno for her practical help with my graduate study. I thank Leena Kathuria and Alex Rogers for their accountant service. I thank Stanton Martin, Christopher Smith, and Gary Howard for their technical support of computer resource.

My Ph.D. committee members and the graduate representative spent their precious time for me. I thank David Bird, Brian Wiegmann, and Eric Stone, and I thank Jim Burton.

Jeff’s group members have helped me to start a good academic career. Douglas Robinson helped me to start research right away. Jiaye Yu helped me understand the Robinson model better. I benefited from more intelligent people: Stéphane Aris-Brosou, Tae-Kun Seo, Asger Hobolth, Benjamin Redelings, and Reed Cartwright. They might not know that I did not see what they were talking about. I deeply thank Hirohisa Kishino and David Jones for their unseen guidance.

Many friends among Bioinformatics graduate students helped me in various ways. I apologize if I forgot acknowledging their helps in my graduate study. I enjoyed my lunch group with Jian Li, Jixin Deng, and Jiaye Yu. I enjoyed playing racquetball with Xiaohua Gong. I thank Frank Mannino for giving me football tickets and allowing Jina and me to watch a real football game at Carter-Finley Stadium. I thank David Aylor for being a friend. I always thought that I would have lunch or dinner with him, but I could not. I thank Brian Howard for the ride to Bioinformatics Retreat to Atlantic Ocean beach. I thank Jonathan Keebler for reading my thesis. I do not know why, but I thank Elizabeth Scholl, Samuel Dickson, Christine Duarte, Youfang Liu, Jessica Maia, Lisa McFerrin, Kate McGee, Sunil Suchindran, Li Zhang, and Sihui Zhao. I thank Rachel Myers for adoring my children when they visited Bioinformatics Research Center. I thank Jihye Kim and Yunjung
Kim for being friends whom I can talk.

This work was impossible without the financial support from N.I.H. grant GM070806 and N.S.F. grant D.E.B-0445180, and additionally from Korea Science and Engineering Foundation grant M06-2003-000-10086-0.

Although some people did not help directly in my graduate study, they helped me shape who I am before I came to Raleigh. I thank them: my best friend, Woo Seong Jeong; my high school teacher, Myeong Sool Koo; my undergraduate mentor, Gynheung An; my undergraduate advisor, Jong-Seong Jeon; my friends, Jongmin Nam and Joonyul Kim; my roommate in Seoul, Seung Hwan Lee; my friends in my company career, Eunkyung Kim and Jin Soo Kim.

Living in different world would have made me homesick without my Korean friends in Raleigh, especially the Hyobin family, the Suwon family, and the Seori family. I thank them and many other neighboring Korean friends.

I am a member of a relatively big family. There are simply too many family members whom I thank for their material and emotional support. I thank my uncle Seong Gyu for my basic algebra. I thank my grandparents on my father’s and mother’s side alike for loving me. I deeply thank my parents for giving me every opportunity to do what I love to do. My two daughters, Jena (Yena) and Juliana (Yebin), would spend many nights talking to me and keeping me awake into night. I thank them, too. My wife basically made all this possible, and I thank Jina.
# Contents

List of Figures xi

List of Tables xii

1 Introduction 1

Introduction 2

Advances in Models of Sequence Evolution 4

- Models of DNA sequence evolution 4
- Models of amino acid sequence evolution 5
- Among-site rate variation 5

Toward More Realistic Models 6

- Biological reality of site-dependency 6
- More realistic but less tractable 7
- Relaxation of independent-site assumption 8

Pursuit of Biological Interpretation with Statistical Inference 9

- Impact of protein structure 9
- Model comparison 10
- Linking biology to rates of evolution 10
- Genotype-phenotype mapping of antigenicity 11
- Variable length memory of Markov model 11
- Biological interpretation of rate of evolution 12
- Insertion and deletion process 13
- Interrogation of Robinson’s model 13

<table>
<thead>
<tr>
<th>Literature Cited</th>
<th>14</th>
</tr>
</thead>
</table>

2 Quantifying the impact of protein tertiary structure on molecular evolution 23

Abstract 24

Introduction 24

Evolutionary model 26

Data preparation 29

Methods 30
3 Designing evolutionary rates to yield desired stationary distributions of sequences
   Introduction ................................................................. 62
   Variable Length Markov Models ..................................... 63
   Neutral model for codon substitution .............................. 65
   The VLMM rate matrix ............................................... 67
   VLMM example ......................................................... 69
   Profile Hidden Markov Models ...................................... 72
   Neutral model for insertions and deletions ...................... 72
   The profile HMM rate matrix ....................................... 75
   Profile HMM example .................................................. 77
   Discussion .............................................................. 78
   Literature Cited ....................................................... 82

4 Probability of single nucleotide polymorphism ancestry via an evolutionary model that incorporates protein tertiary structure
   Abstract ......................................................................... 86
   Introduction ................................................................. 86
   Materials and Methods .................................................. 89
   Data preparation ......................................................... 89
   Probability of an allele being ancestral ............................ 90
   Dependent-sites model of protein-coding DNA sequence evolution ... 90
   Population genetic interpretation of phenotype impact parameters 91
   Probability of an allele being ancestral under selection .......... 92
   Results ......................................................................... 95
   Comparison of frequency of ancestral alleles under selection and neutrality .................... 95
   Molecular fitness and allele frequency .............................. 98
   Ancestral alleles are more advantageous with respect to molecular fitness .................... 98
   Probability that an allele is ancestral with molecular fitness .... 100
   Discussion ..................................................................... 102
   Acknowledgments ......................................................... 105
5 Summary

Followup Research of the Robinson Model .................. 112
Sequence Path with Insertion and Deletion .................. 113
Genotype-Phenotype Mapping Techniques ...................... 115
Merger of Interspecific Evolution and Population Genetics .... 116
Literature Cited ............................................. 117
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A pairwise sequence alignment of parts of two genes</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Distributions of protein tertiary structure impact parameters</td>
<td>34</td>
</tr>
<tr>
<td>2.2</td>
<td>Scatter plots of sequence length against protein tertiary structure impact posterior means</td>
<td>36</td>
</tr>
<tr>
<td>2.3</td>
<td>Distribution of Bayes factor estimates among proteins</td>
<td>38</td>
</tr>
<tr>
<td>2.4</td>
<td>Average selective effects of nonsynonymous changes</td>
<td>40</td>
</tr>
<tr>
<td>2.5</td>
<td>Boxplots of $\hat{\Delta}_i$ estimates for the 23 Gene Ontology level-1 terms</td>
<td>41</td>
</tr>
<tr>
<td>2.6</td>
<td>Boxplots of $\hat{\Delta}_i$ estimates for the Cellular Component terms</td>
<td>48</td>
</tr>
<tr>
<td>2.7</td>
<td>Boxplots of $\hat{\Delta}_i$ estimates for the Molecular Function terms</td>
<td>49</td>
</tr>
<tr>
<td>2.8</td>
<td>Boxplots of $\hat{\Delta}_i$ estimates for the Biological Process terms</td>
<td>50</td>
</tr>
<tr>
<td>3.1</td>
<td>Distribution of $2N_e s$ estimates using VLMM</td>
<td>70</td>
</tr>
<tr>
<td>3.2</td>
<td>Distribution of the mean $2N_e s$ estimate per human gene using VLMM</td>
<td>71</td>
</tr>
<tr>
<td>3.3</td>
<td>Estimates of $2N_e s$ for substitutions using profile HMM</td>
<td>79</td>
</tr>
<tr>
<td>3.4</td>
<td>Estimates of $2N_e s$ for single codon deletions using profile HMM</td>
<td>80</td>
</tr>
<tr>
<td>3.5</td>
<td>Estimates of $2N_e s$ for single codon insertions using profile HMM</td>
<td>81</td>
</tr>
<tr>
<td>4.1</td>
<td>Frequency of minor alleles being ancestral of the human SNPs</td>
<td>97</td>
</tr>
<tr>
<td>4.2</td>
<td>Scaled selection coefficient and allele frequency</td>
<td>99</td>
</tr>
<tr>
<td>4.3</td>
<td>Proportion of non-synonymous changes to the ancestral sequence</td>
<td>101</td>
</tr>
<tr>
<td>4.4</td>
<td>Probability of minor allele being ancestral</td>
<td>103</td>
</tr>
<tr>
<td>5.1</td>
<td>A pairwise sequence with gaps and its partition</td>
<td>113</td>
</tr>
<tr>
<td>5.2</td>
<td>A site path between two entities of C- and CG</td>
<td>114</td>
</tr>
</tbody>
</table>
List of Tables

4.1 Four possible scenarios of allele ancestry and molecular fitness . . . . . . . . . . 99
Chapter 1

Introduction
Introduction

A living organism carries its own genetic information in a stable form that is inherited from its antecedents, and it will bequeath to its descendants the secret of life before its death. The genetic information is not permanent but changeable via mutation every generation of life. Mutations induce unnoticeable alterations in the blueprint of life. The inconceivably long evolutionary time has allowed organisms to “transform” to different forms of life — evolution happens (Darwin 1859). The relatively short evolutionary distance since the split of human and chimpanzee differentiates the two species in about 1% of their genomes (Chimpanzee Sequencing and Analysis Consortium 2005). During recent human evolution, human beings acquired novel phenotypes such as upright posture. Human evolution also yielded undesirable phenotypes such as diseases that are the complex final product of genotypes. Proteins are a kind of phenotype at the molecular level, and the micro-level molecular phenotypes may affect macro-level phenotypic traits whether they are desirable or not. The importance of the relationship between diseases and proteins can be seen in the list of the top ten causes of death during 2002, as published by the World Health Organization. Cardiovascular diseases claimed 7.2 million people with coronary heart disease and 5.5 million people with stroke or other cerebrovascular disease among the estimated 57 million deaths. Although it is controversial whether the protein called C-Reactive Protein is a causative agent of cardiovascular disease, an abnormal C-Reactive Protein may be associated with cardiovascular diseases (Ridker 2003). The relationship between C-Reactive Protein and cardiovascular disease is one aspect of numerous disease-related proteins, and biologists and medical scientists who investigate mechanisms of diseases with proteins are often interested in evolutionary relationships among homologous genes that encode the proteins of interest.

The evolutionary relationships among homologous genes that are sampled from different species can be represented by phylogenetic trees. We have a variety of methods to infer phylogenetic trees from inter-specific sequence data. These methods are often based on the measurement of evolutionary distance between two homologous genes (i.e., two genes that share a common ancestor). Descendant genes may be formed via speciation events (orthology) or via gene duplication events (paralogy). After the formation of two genes from their common ancestor, a series of evolutionary events can give rise to similar but different gene functions. The very degree of sequence dissimilarity serves as a surrogate of
evolutionary distance between the genes. For example, a hypothetical part of the pairwise alignment between genes is shown in figure 1.1. The two sequences are different at 9 sites, and the alignment length is 30, which means that the sequence dissimilarity of the two parts of the genes is 30%. The parsimony criterion claims that there are 9 substitutions between the two sequences. Parsimony tends to under-estimate substitution amounts because there might have been revertants (e.g., A to G to A) or multiple substitutions (e.g., G to C to A). A correction to account for multiple substitutions and reversions was made by Jukes and Cantor (1969) in their chapter titled “Evolution of protein molecules.” Their model of DNA sequence evolution is called the “Jukes-Cantor” or JC69 model. The JC69 model has been a foundation of mathematical stochastic models of DNA sequence evolution but it is overly simplistic. We will overview the advances toward more realistic models after we discuss evolutionary distances and rates of evolution.

Models of molecular evolution are often characterized by rates of substitution events. We can use the rates to estimate how many substitution events occurred on a branch of an evolutionary tree. If we observed all of the evolutionary events and we knew the evolutionary time separating two biological sequences, we could easily infer rates of the events. Although we are neither able to observe the evolutionary events nor to pinpoint the evolutionary time of the events, we can estimate the number of events separating two biological sequences by assuming a certain mathematical model of molecular evolution. For example, if we assume that there are equal rates of substitution among DNA bases, maximum likelihood with the JC69 model estimates 0.38 as the expected number of substitutions per site in the data of the two sequences in figure 1.1. The average number of events per site is the evolutionary distance, which inextricably depends on the rate of events and the length of time. Evolutionary distances may be very long either because rates of evolution are fast or because evolutionary time is long. Although it is not easy to disentangle evolutionary distance to time and rate, there are methods available (e.g., Thorne and Kishino 2005).
Advances in Models of Sequence Evolution

Mathematical models of molecular evolution are advancing by relaxing assumptions that have been made for mathematical convenience and computational tractability. We do not attempt here to comprehensively describe the advances in models of molecular evolution, but we briefly summarize them in order to introduce a model on which we based our work in the subsequent chapters.

Models of DNA sequence evolution

Models of DNA sequence evolution are usually continuous-time Markov models. They range from the simplicity of the JC69 model to the slightly more elaborate general time-reversible model (Tavaré 1986) and beyond (e.g., see Swofford et al. 1996, Liò and Goldman 1998 and Felsenstein 2004). Kimura (1980) considered the difference in the rates of transition and transversion (Kimura’s two-parameter model). Felsenstein (1981) added three free parameters representing the nucleotide equilibrium frequency to the JC69 model. There are also widely-used models that combine the model of Felsenstein (1981) and Kimura’s two-parameter model in slightly different ways: the HKY model (Hasegawa, Kishino, and Yano 1985), and the F84 model (introduced in the phylogeny package PHYLIP since 1984, and described by Kishino and Hasegawa 1989). Tamura and Nei (1993) proposed their TN93 model to slightly generalize the F84 and HKY models by distinguishing between purine-purine transitions and pyrimidine-pyrimidine transitions. Tavaré (1986) considered the most general time-reversible model (GTR) of DNA sequence evolution among those that have independent and identical evolutionary substitutions among sites. The GTR model has 9 free parameters. The time-reversibility property is that the stationary probability of a state \(i\) multiplied by the rate from \(i\) to \(j\) is equal for all \(i\) and \(j\) to the stationary probability of \(j\) multiplied by the rate from \(j\) to \(i\). This is a very convenient mathematical property because it means phylogenies need not be rooted when doing phylogenetic inference. However, we need extra information (e.g., outgroup sequences) to root trees with time-reversible models because we cannot find the root of trees with only extant sequences using time-reversible models. Blaisdell (1985) proposed a non-reversible model of DNA sequence evolution that is based on Kimura’s two-parameter model.
Models of amino acid sequence evolution

Amino acids have diverse physico-chemical properties and the rates of amino acids replacement may be more biologically informative and interpretable than rates of nucleotide substitution. Models of amino acid replacement have been studied since about the time when the first model of DNA sequence evolution appeared. The JC69 model of DNA evolution was published by Jukes and Cantor (1969) and the Dayhoff model of amino acid sequence evolution appeared in 1968 (Dayhoff and Eck 1968). As opposed to the development of mostly parametric models in DNA sequence evolution, the driving force of model development in protein sequence evolution has been empirical. Dayhoff and Eck (1968) derived empirical models of amino acid sequence evolution from groups of closely related amino acid sequences that they had compiled. They later updated parameter estimates for this model without changing the basic structure of the model (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978). Dayhoff and her coworkers used a protein database of relatively small size. Later, the database of protein sequences became larger, and Jones, Taylor, and Thornton (1992) developed an efficient way of estimating relative rates of amino acid replacement from the larger database (see also Gonnet, Cohen, and Benner 1992). Dayhoff’s mutation data matrices were based on a maximum parsimony technique for inferring ancestral sequences. Whelan and Goldman (2001) improved the empirical model through a maximum-likelihood approach. Advantages of maximum-likelihood methods include having less systematic error than parsimony and the basis maximum likelihood provides for inferring more accurate ancestral sequences (Yang, Kumar, and Nei 1995).

Among-site rate variation

Rates of evolution can be different from site to site along a sequence. Disparate rates among sequence positions may be due to the variable tolerability to evolutionary changes. The sites that are not changing at all might be critical for biological function, and any changes of those sites would drastically affect the organismal fitness. Sites might allow a certain class of DNA bases or amino acid residues because changes to residues within the class might not affect the organismal fitness. Yang (1993) assumed that substitution rates among sites have a gamma distribution, and a computationally tractable approach for incorporating rate heterogeneity was published by Yang in 1994. Multiple genes or parts from different regions
in genomes may be combined and analyzed within a single model. Combining the parts and analyzing with one model may be inappropriate when the different parts have different evolutionary processes. Cao et al. (1994) used multiple models that were appropriate to different parts of the combined data. Yang et al. (1995) and Felsenstein and Churchill (1996) employed the method of hidden Markov models (Baum and Petrie 1966) to allow for different sites to have different categories of rates.

Among-site rate variation has also been studied directly at the protein level by Thorne, Goldman, and Jones (1996) who devised different models of amino acid substitution depending on different categories of protein secondary structure that are stochastically determined according to a hidden Markov model. Goldman, Thorne, and Jones (1998) extended the technique of Thorne, Goldman, and Jones (1996) by splitting the loop region into the turn and coil and by adding two more categories of buried and exposed as consideration of the solvent accessibility. Their models were retrospectively used to predict protein secondary structure (Goldman, Thorne, and Jones 1996). Koshi and Goldstein (1998) parameterize the rates of amino acid evolution as exponential functions of the physico-chemical properties of amino acid residues, and avoid the \textit{a priori} categorization of protein structure. Lartillot and Philippe (2004) deal with the number of categories as a free parameter to infer in the Bayesian framework, and these authors have developed the CAT model of amino acid replacement to explain variation of replacement rates among sites by variation of preferred amino acids among sites.

\section*{Toward More Realistic Models}

\subsection*{Biological reality of site-dependency}

While models of sequence evolution advance with more realistic parameterization of rates at the one-character level, many assume that a substitution event at one site is independent of substitution events at other sites. This “independent-site” assumption is made for the sake of mathematical and computational convenience and is often violated. For example, Coulondre et al. (1978) observed that high mutation rates occurred at 5-methylcytosine sites in \textit{Escherichia coli}, and suggested that these mutational hotspots might be due to the deamination of 5-methylcytosine to thymine. The substitution rate of Cytosine when adja-
cent to a Guanine (i.e., a CpG) is higher in mammals than are other substitution rates (e.g., Cooper and Youssoufian 1988; Nachman and Crowell 2000). The tendency for a higher rate of substitution of Cytosine in CpG is genome-wide in mammals (e.g., Chimpanzee Sequencing and Analysis Consortium 2005). As another genome-wide biological phenomenon, DNA sequences are packed in chromosomes when they are inactivated and not transcribed, and the way they are packed may require periodical arrangement of nucleotides (Segal et al. 2006). This may constrain DNA sequence patterns, and may induce dependency of site substitutions by other site states. At the gene level, DNA sequences can be transcribed to messenger RNA (mRNA) sequences, which can be in turn translated to amino acid sequences. As opposed to mRNA sequences, structural RNA sequences (e.g., ribosomal RNA and transfer RNA sequences) may form three-dimensional conformations to be functional. RNA base pairs in the stem regions of structural RNA sequences may constrain substitutions of their counterpart RNA bases. Loop regions of RNA secondary structures may be functionally important, and multiple loops may interact with one another, which can constrain substitutions of RNA bases at other sites. Protein sequences form three-dimensional conformations, and they can have functionally important regions (e.g., active sites). Like RNA sequences but more profoundly, amino acid sites that are far apart in one-dimension may be nearby in three-dimensions, and the interactions among those sites may favor certain pairs of amino acids. These two macromolecular types are encoded by DNA sequences and may affect rates of evolutionary events by having a substitution event at one site be influenced by substitutions at other sites.

More realistic but less tractable

Felsenstein (1973) placed the inference procedures of evolutionary history on sound statistical footing by developing a likelihood approach. The calculation of the likelihood of a tree topology and model parameters given sequence data was thought to be intractable because the enumeration of all combinations of the internal node character states of a tree might be computationally infeasible. Felsenstein (1981) introduced the pruning algorithm that dramatically decreases the computational time of likelihood calculations and makes the inference of phylogenies practical. Many models with the independent-site assumption have been used for phylogeny estimation. Phylogenetic inference procedures can benefit from more efficient algorithms of tree reconstruction (e.g., Guindon and Gascuel 2003).
Markovian models of sequence evolution are characterized by rates among evolutionary entities such as nucleotides, amino acids, codons, or sequences. Transition probabilities among the entities are the probability that one entity changes to another during a specified time duration. Finding transition probabilities is essential to applying the pruning algorithm in likelihood-based phylogenetic inference. There exist analytic solutions of transition probabilities for simple substitution models. On the other hand, we have to find transition probabilities numerically by exponentiating instantaneous rate matrices for more realistic models. The computing cost of the numerical calculation of transition probabilities can be prohibitively expensive. As an alternative, the statistical technique of uniformization can be employed to facilitate likelihood-based inference with complicated models of sequence change (Mateiu and Rannala 2006, and see also Lartillot 2006). The biological reality of dependence among sites gives rise to much more complex models, and can be incorporated into inference by adopting the ideas of Jensen and Pedersen (2000). They generate intermediate (hidden) mutation events between two tip sequences, and then employ the statistical technique of Markov chain Monte Carlo to analyze pairwise sequence alignments of HIV-1 genes.

Relaxation of independent-site assumption

Models that reflect the biological reality of dependence among bases or amino acid residues by definition cannot assume that a site evolves independently of all other sites. Therefore, they cannot take advantage of Felsenstein's pruning algorithm in most cases. Accordingly, there have been other proposed likelihood calculation techniques for making evolutionary inferences (Jensen and Pedersen 2000; Hwang and Green 2004; Siepel and Haussler 2004; Christensen, Hobolth, and Jensen 2005). Fornasari, Parisi, and Echave (2002) maintained computational tractability by using simulation to find independent-site evolutionary models that approximate dependent-site evolutionary models. Influenced by simulation techniques that incorporate sequence-structure compatibility (Parisi and Echave 2001; Fornasari, Parisi, and Echave 2007), and the statistical inference approaches with dependent models of sequences changes (Jensen and Pedersen 2000; Pedersen and Jensen 2001), Robinson et al. (2003) introduced a model of protein sequence evolution with dependence among codons due to the constraint of protein tertiary structure. Robinson’s dependent-sites model assesses how non-synonymous substitutions affect sequence-structure compatibility. The
substitutions that improve sequence-structure compatibility will have accelerated rates of evolution, and the substitutions that worsen sequence-structure compatibility will have decelerated rates of substitution. Initially designed for pairwise sequence analysis, Robinson’s model was soon extended to inference with more than two sequences (Robinson 2003; Rodriguez et al. 2005). Yu and Thorne (2006) used a similar statistical approach to study the constraints RNA secondary structure places on sequence change.

Pursuit of Biological Interpretation with Statistical Inference

We base our research in this thesis on the Robinson model (Robinson et al. 2003; Robinson 2003). Here, we outline our motivation for what is studied in the subsequent chapters.

Impact of protein structure

The Robinson model extends codon models (Goldman and Yang 1994; Muse and Gaut 1994) by incorporating phenotypic impact parameters that are related to protein tertiary structure. Inspired by Jensen and Pedersen (2000), Robinson and his coworkers proposed to sample “sequence paths” to overcome the computational difficulties that would arise by trying to handle dependence among codons due to protein tertiary structure with Felsenstein’s pruning algorithm. The Robinson approach can be classified as a data augmentation technique (Tanner and Wong 1987). The sequence path contains the information about when and where substitution events have occurred. Incorporating the sequence path as an auxiliary variable, one needs to evaluate the probability of an ending sequence and a sequence path given a starting sequence and model parameters (see Robinson et al. 2003, and also Rodrigue et al. 2005; Yu and Thorne 2006). The sequence path approach is too computationally demanding to investigate the evolutionary impact of protein tertiary structure for a large and diverse set of protein families. If we are interested in the model parameters relating to evolutionary impact of tertiary structure but not in the rest of the model parameters, then we do not necessarily need to incorporate the sequence path into the estimation procedure because the stationary probability of a single sequence includes information about the phenotypic impact parameters. Accordingly, from a single sequence with protein tertiary structure, we can estimate the phenotypic impact parameters. In Chapter
2, we exploit the stationary probability of sequences according to the Robinson model. We do this to study the distribution of phenotypic impact parameters among a large number of protein families.

Model comparison

The models that best fit data at hand are not necessarily the most desirable. One obvious complication is the compromise between statistical fit and number of model parameters. Comparing nested models, one of which is a special case of another, we can employ likelihood ratio tests to determine if the general model explains data better than the special model (e.g., Posada and Crandall 1998). Akaike’s Information Criterion (AIC; Akaike 1973) and the Bayesian Information Criterion (BIC; Schwarz 1978) can be used to compare models that are not nested. Both information criteria give more weight to models that have larger likelihoods, and penalize models that are rich in number of parameters. A Bayesian approach of model comparison is the Bayes factor (Kass and Raftery 1995), which is the marginal likelihood ratio of two models that are compared. A few different approaches to calculating Bayes factors have been pursued in phylogenetic methods: the Savage-Dickey ratio (Verdinelli and Wasserman 1995; Suchard, Weiss, and Sinsheimer 2001), the reversible jump Markov chain Monte Carlo (Green 1995; Huelsenbeck, Larget, and Alfaro 2004), and thermodynamic integration (Lartillot and Philippe 2004; Rodrigue, Philippe, and Lartillot 2006). As another approach for calculating the Bayes factors, we employ the marginal likelihood calculation of Chib and Jeliazkov (2001) using Markov chain Monte Carlo output (see also Chib 1995) in Chapter 2.

Linking biology to rates of evolution

Unquestionably, rates of molecular evolution are affected by biology. Which biological aspects dictate them? It is both important and difficult to find causative or associative agents of evolutionary rates (e.g., see Pál, Papp, and Lercher 2006). For example, less important proteins may tend to evolve relatively faster (Hirsh and Fraser 2001) because mutants of these proteins may be less subject to selective pressure. Proteins that work together may tend to evolve more slowly than do other proteins (Fraser et al. 2002). Proteins that have high expression level may tend to evolve more slowly than low-expression proteins (Pál,
In general, higher phenotypic impacts of proteins lead to stronger selection pressure on proteins. Dispensable, low-expressed proteins with few interacting partners may not have a large phenotypic impact on fitness. Translational selection, or the number of translational events of genes has been claimed to be a dominant determinant of protein evolutionary rates (Drummond, Raval, and Wilke 2006). The structure of a protein, or the solvent accessibility of amino acids, was also tested to be another dominant determinant of protein evolutionary rates (Lin et al. 2007). The models of evolution that incorporate phenotypic properties should quantify the impact of phenotype on rates of evolution, and the impact of phenotype may be associated with protein annotation. In the Gene Ontology (Harris et al. 2004) analysis of Chapter 2, we test the hypothesis that gene annotation is correlated with the impact of protein structure on rates of evolution.

**Genotype-phenotype mapping of antigenicity**

Protein tertiary structure and RNA secondary structure are one aspect of molecular phenotype. A central effort in computational biology is to predict phenotype from genotype. There are a great number of computational methods for predicting a variety of biological phenomena. For example, viruses need to escape from their host’s immune response. Antigenicity prediction is of great interest to the immunology community. If we developed a technique that adequately predicted the antigenicity of viral sequences, we could study viral evolution in the statistical framework of the dependent-sites models described by Robinson et al. (2003). Although the prediction of epitope antigenicity continues to receive attention (Nielsen et al. 2003; Bui et al. 2005; Peters et al. 2006), our attempts at measuring the antigenicity of a long sequence did not prove successful. As a result, our goal of quantifying the impact of the host immune system on viral evolution will have to wait for improved antigenicity prediction systems.

**Variable length memory of Markov model**

Probabilistic models of sequence organization can assist in functional annotation by assigning new sequences to groups of sequences with known function. Building models given a group of sequences can be accomplished via Markov models (e.g., Durbin et al. 1998). The probability of a single sequence can be written as the product of marginal probability of
the first character and a product over the remaining characters of the probabilities of each character conditional on all preceding characters. Building a model that has the probability of the next character depend on all previous character states is intractable because the number of previous states grows exponentially. The biological reality of dependence on the previous states calls for more realistic models that are often intractable. On the other hand, the probability of a character at a site may depend not on all the previous states, and but on a limited number of previous states. One simple model that is often employed is to consider only one step backward, which is a first-order Markov model. This simple model may not be realistic. A reasonable compromise between realism and tractability can be made by a construct called a variable length Markov model (also known as variable order Markov models and variable length Markov chains). In Chapter 3, we use variable length Markov models (VLMM) to study protein sequence organization. We used a probabilistic suffix tree technique (Ron, Singer, and Tishby 1996; Bejerano and Yona 2001) as implemented in software that learns models of human protein-coding DNA sequences (Bejerano et al. 2004). We contrast the VLMM model of human genes with a neutral model of sequence change by reconciling population genetic theory with interspecific evolution (Thorne et al. 2007).

**Biological interpretation of rate of evolution**

Evolution happens at the population level, but the models used in phylogenetics rarely have an explicit connection to population genetic theory. Population genetics has a longer history than molecular evolution. While theories have driven population genetic research, research in interspecific evolution has been more motivated by newly emerging data. The disconnect between population genetics and molecular evolution is beginning to be remedied (Halpern and Bruno 1998; Nielsen and Yang 2003; Berg, Willmann, and Lässig 2004; Sella and Hirsh 2005) by attaching population genetic interpretations to rates in models designed for studying interspecific sequence comparisons. Robinson’s model has a weak connection with population genetic theory although its application to phylogeny inference is not practical at the moment (Thorne et al. 2007). Thorne and his coworkers connected the fixation probability of diffusion theory (Kimura 1962) to rates of evolution. The product of effective population size and selection coefficient was then approximated via rate parameters of models of sequence evolution and the genotype-phenotype mapping of a protein tertiary structure prediction scheme.
Insertion and deletion process

The processes of nucleotide substitution and amino acid replacement are not the only forces shaping molecular evolution. Insertion and deletion processes are beginning to be studied in a tractable statistical framework (Lunter et al. 2005; Redelings and Suchard 2005). The dependent-sites models developed so far have extended models of substitution and replacement processes, and it is desirable to improve the substitution-only-based model to reflect the insertion and deletion process. In Chapter 3, we show that a dependent-sites model with insertion and deletion events can be time reversible, and can attain desired stationary distributions of sequences. We modify the substitution-only model of dependent change, and employ profile hidden Markov models (e.g., Eddy 1996). The Pfam database is a comprehensive protein family database, in which trained profile hidden Markov models and sequence alignments are provided for particular protein families (Sonnhammer, Eddy, and Durbin 1997). Our modified model has evolutionary rates that yield a trained profile hidden Markov model as the stationary distribution for a particular protein family of interest. We also use the same technique we used for VLMM to make crude population genetic inferences about natural selection.

Interrogation of Robinson’s model

Mutations introduce new alleles that are different than their parental alleles. The new mutant alleles may be selectively neutral, in which case their phenotypes are neither “superior” nor “inferior” to those of their parental genotypes in terms of fitness. In each generation, randomly selected individual alleles are allowed to contribute to gene pools of the next generation — the gene pool of a locus is stochastically determined by genetic drift. The new phenotypes that result from most mutants is likely to be inferior to original phenotypes in fitness. Natural selection might immediately purify deleterious alleles out of the population. Lucky deleterious alleles may reside on near neighboring “good” alleles, and may maintain a low frequency without dying out. Or, they may defy the odds and be chosen to contribute to the next generation gene pool. Rarely, newly arisen alleles may be beneficial, and natural selection will favor them over the original alleles.

Single nucleotide polymorphisms (SNPs) are the most vast resource of human genetic variation available to date (Sherry et al. 2001). SNPs are categorized depending
SNPs may have two alleles (dimorphic) or more (polymorphic), and many SNPs seem to have only two alleles. The inference of ancestral states of human SNPs may be made with the parsimony criterion and a sister species genome such as the chimpanzee genome. More frequent alleles are likely to be ancestral, and less frequent alleles are likely to be derived (Watterson and Guess 1977). Although macro-level phenotypes may interact with environment, and environmental changes may shift the fitness of phenotypes of alleles, molecular phenotype may be largely robust to the external environment. If molecular phenotypes do not change over time, then the molecular phenotype of derived alleles may be mostly less beneficial than that of ancestral alleles. The conjecture that ancestral alleles must be the ones that have higher average fitness (Donnelly and Kurtz 1999) may be empirically testable. In Chapter 4, we study SNPs with the technique that we developed in Chapter 2 and 3.

Models of molecular evolution are increasingly becoming complicated in order to reflect the biological reality. Although they are not practical for inferring phylogenies at the moment, more realistic models will find their applications as computing resources increase and more powerful inference techniques are developed. This effort to make models more realistic will continue as more diverse biological data accrue.

Literature Cited


Chapter 2

Quantifying the impact of protein tertiary structure on molecular evolution

Abstract

To investigate the evolutionary impact of protein structure, the experimentally determined tertiary structure and the protein-coding DNA sequence were collected for each of 1,195 genes. These genes were studied via a model of sequence change that explicitly incorporates effects on evolutionary rates due to protein tertiary structure. In the model, these effects act via the solvent accessibility environments and pairwise amino acid interactions that are induced by tertiary structure. To compare the hypotheses that structure does and does not have a strong influence on evolution, Bayes factors were estimated for each of the 1,195 sequences. Most of the Bayes factors strongly support the hypothesis that protein structure affects protein evolution. Furthermore, both solvent accessibility and pairwise interactions among amino acids are inferred to have important roles in protein evolution. Our results also indicate that the strength of the relationship between tertiary structure and evolution has a weak but real correlation to the annotation information in the Gene Ontology database. Although their influences on rates of evolution vary among protein families, we find that the mean impacts of solvent accessibility and pairwise interactions are about the same.

Introduction

Most widely used models of sequence evolution exploit the assumption that individual sites or codons evolve independently from one another. The independence assumption dictates that a substitution at one site or codon does not influence substitution rates at other sites or codons. Unfortunately, the assumption of evolutionary independence among sites is not biologically plausible. In reality, the effect of a change at a particular site might cascade throughout the sequence, causing changes in substitution rates at other sites or codons. Nonsynonymous substitutions (i.e., those nucleotide substitutions that cause an amino acid replacement) are especially likely to affect rates at other positions, because the amino acid that occupies one site in a protein may have interactions with amino acids that are located nearby in the protein structure.

Despite strong evidence of a relationship between protein structure and protein change, evolutionary models of protein-coding genes tend to neglect protein structure. Protein tertiary structure changes very slowly over time (Chothia and Lesk 1986; Flores et al. 1993; Russell et al. 1997). If a nucleotide substitution results in an amino acid replacement
that disrupts protein structure, it is likely to be selectively deleterious, and the rate at which such a substitution occurs should be low. Likewise, a nucleotide substitution should have a relatively high rate if it improves compatibility between protein sequence and structure. This is the intuition underlying a recently developed procedure for making evolutionary inferences when, owing to protein tertiary structure, codons do not evolve independently (Robinson et al. 2003).

Protein tertiary structure is phenotype whereas protein-coding DNA is genotype. Evolutionary models that lack dependence induce sterile adaptive landscapes where each codon or sequence position can be separately optimized. In contrast, models with dependence have the prospect of yielding biologically plausible fitness landscapes. Allowing dependent change among sequence positions is important because it permits a reasonable treatment of the effects of phenotype on evolution of genotype. A central goal of the newly emerging field of computational biology is to make in silico predictions of aspects of phenotype from genotype. The statistical strategies used in the study reported here can incorporate these predictive products of computational biology into evolutionary models and thereby better reveal the impact of the phenotype on evolution of the genotype.

Despite the obvious and important role for protein tertiary structure in shaping the evolution of protein-coding DNA sequences, tools for quantifying the evolutionary impact of tertiary structure remain primitive. One major difficulty is that natural selection to maintain protein tertiary structure induces evolutionary dependencies among sequence positions and these dependencies complicate statistical inference. Felsenstein’s pruning algorithm (Felsenstein 1981) is the conventional mechanism for transforming continuous time Markov models of sequence evolution into a basis for statistical inference, but adapting the pruning algorithm to dependent changes among positions is problematic.

Recently, strategies have been developed for making inferences with models that have evolutionary dependence among sites (Jensen and Pedersen 2000; Pedersen and Jensen 2001; Hwang and Green 2004; Pedersen et al. 2004; Siepel and Haussler 2004; Christensen, Hobolth, and Jensen 2005). Protein tertiary structure as a source of dependent change among positions was the focus of Robinson et al. (2003). The inference procedure of Robinson et al. (2003) was limited to data sets with two aligned protein-coding DNA sequences, but this was soon extended to data sets with three sequences (Robinson 2003) and data sets with three or more sequences (Rodrigue et al. 2005). Although these inference procedures require the restrictive assumption that all protein sequences being analyzed share a single
known tertiary structure, their biggest limitation is probably that it can be computationally prohibitive to analyze a large number of highly diverged sequences.

It would be desirable to assess the impact of protein structure on protein evolution across a large set of protein families. This sort of assessment would shed light on how variable this impact is among protein families, and it might help determine if certain features of a protein family are correlated with the relationship between protein structure and evolution. Here, we show that information about the relationship between protein structure and evolution can be extracted from data sets consisting only of a single sequence per known protein structure.

Analyzing data sets with single sequences requires much less computation per data set than does analyzing multiple sequence data sets. As a result, we have been able to quantify the relationship between protein structure and evolution for 1,195 protein families. We find that incorporating tertiary structure into models of protein evolution almost always produces a much better fit to data. We explore how the influences on evolutionary rates of solvent accessibility and pairwise interactions vary among protein families. When averaged among protein families, we find that the sizes of the solvent accessibility and pairwise interaction impacts are similar. We also examine whether there is a correlation between annotation information in the Gene Ontology database and the strength of the structure–evolution relationship. We conclude that there is a weak but real correlation.

Evolutionary model

We adopt and summarize the Robinson et al. (2003) model of protein evolution, referred to here as the dependent–sites model. It is designed so that the rate of change $R_{ij}$ from sequence $i$ to sequence $j$ depends both on the mutational process and on natural selection. Some parameters in the model are intended to capture the effects of mutation, while others are intended to reflect natural selection. The natural selection parameters are further subdivided into those that relate tertiary structure to evolution and those that reflect phenomena external to protein structure.

In the complete absence of natural selection, the probability of a neutrally evolving sequence is governed solely by the mutational process. In this neutral scenario, the frequencies of the four nucleotide types would be $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$, where $\pi_A + \pi_C + \pi_G + \pi_T = 1$. 

26
To account for the differential occurrence of transition and transversion mutations, we add the transition/transversion ratio parameter $\kappa$ to the neutral model.

Effects on evolutionary rates that cannot be attributed to protein structure are captured by the parameter $\omega$. This parameter represents the factor by which a nonsynonymous rate differs from what the rate would be if the change were actually synonymous. If $\omega$ were the only parameter to differentiate between neutral evolution and natural selection, the dependent–sites model would have no important distinction from other simple codon substitution models (e.g. Goldman and Yang 1994; Muse and Gaut 1994).

To represent the influence of tertiary structure, the dependent–sites model has two additional parameters. One parameter, $s$, captures the effects on evolutionary rates of solvent accessibility. The other, $p$, reflects the effects of pairwise amino acid interactions. When $s$ and $p$ are both zero, protein structure does not alter evolutionary rates. When both are positive, proteins evolve so as to be compatible with tertiary structure. It is formally possible that $s$ or $p$ can be negative. However, negative values for these parameters are not biologically plausible because they represent the situation where a protein is evolving so as to poorly fit its three-dimensional structure.

To measure or score how compatible a particular sequence is with a particular structure, a scoring system that was empirically defined for protein fold recognition is used (Jones, Taylor, and Thornton 1992; Jones and Thornton 1996). This scoring system is designed for measuring how well a specific protein sequence folds into a specific structure. Rather than using the sequence–structure compatibility measure to find the best structure, the dependent–sites model assumes a known protein structure that is invariant during evolution and the sequence-structure compatibility measure links rates of protein–coding gene evolution to tertiary structure.

The sequence–structure compatibility measure is designed for globular proteins and has two components. One assesses solvent accessibility and the other evaluates pairwise amino acid interactions. The solvent accessibility score assigned by folding an amino acid sequence $i$ (or a translated DNA sequence $i$) into the known tertiary structure is denoted by $E_s(i)$, whereas the pairwise amino acid interaction score is $E_p(i)$. Low values of $E_s(i)$ and $E_p(i)$ represent better sequence–structure compatibility than high values. The scoring systems are akin to free energies in that the values of $E_s(i)$ and $E_p(i)$ should be negative when sequence $i$ is folded into its actual structure and positive otherwise.

The dependent–sites model does not allow instantaneous changes from a sequence
i to a sequence j unless these sequences differ at only 1 nucleotide position. The rate $R_{ij}$ is zero if $i$ and $j$ differ at more than one position. Likewise, changes that introduce stop codons are not allowed. Letting $h \in \{A,C,G,T\}$ be the nucleotide in sequence $j$ at the sole position where $i$ and $j$ differ and letting $u$ be a rate scaling factor, the rates of change are

$$R_{ij} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ u\pi_h\kappa & \text{for a synonymous transition} \\ u\pi_h\omega e^{s(E_s(i)-E_s(j))+p(E_p(i)-E_p(j))} & \text{for a nonsynonymous transversion} \\ u\pi_h\kappa\omega e^{s(E_s(i)-E_s(j))+p(E_p(i)-E_p(j))} & \text{for a nonsynonymous transition} \end{cases}$$ (2.1)

The length in nucleotides of sequence $i$ will be $L$. Having $i_m$ and $k_n$ respectively, be the $m^{th}$ nucleotide of sequence $i$ and the $n^{th}$ nucleotide of sequence $k$, this system of rates yields the following stationary distribution of sequences

$$p(i|\pi_A, \pi_C, \pi_G, \pi_T, \kappa, \omega, s, p) = \frac{e^{-2sE_s(i)-2pE_p(i)} \prod_{m=1}^L \pi_{i_m}}{\sum_k e^{-2sE_s(k)-2pE_p(k)} \prod_{n=1}^L \pi_{k_n}},$$ (2.2)

where the sum in the denominator is over all possible sequences $k$ that have length $L$ and that lack a premature stop codon (Robinson et al. 2003). The fact that the $s$, $p$, and $\pi$ parameters appear in equation (2.2) means these parameters can be estimated without the computational burden that stems from analyzing multiple–sequence data sets. Because $\kappa$ and $\omega$ do not appear in equation (2.2),

$$p(i|\pi_A, \pi_C, \pi_G, \pi_T, \kappa, \omega, s, p) = p(i|\pi_A, \pi_C, \pi_G, \pi_T, s, p).$$

This means estimation of $\kappa$ and $\omega$ requires two or more sequences.
Data preparation

There were 32,987 experimentally determined protein tertiary structures deposited at the Protein Data Bank (PDB) database as of December 2, 2005 (Berman et al. 2000). Some protein families in this database are much more heavily represented than others. Our objective is to characterize a large and diverse sample of protein structures in order to begin to understand how the influence of tertiary structure varies in the protein universe. For this reason, we decided to partially avoid redundancy in the PDB by concentrating on the PDBselect list (Hobohm et al. 1992) available on September 15, 2006. This list identifies PDB protein chains with less than 25% sequence identity. For PDB structures with multiple protein models, we selected the representative conformation supplied by the OLDERADO database (Kelley and Sutcliffe 1997) at http://pqs.ebi.ac.uk/pqs-nmr.html. For those proteins not included in OLDERADO, the first structure conformation entry in PDB was selected.

We employed a resource (Martin 2005) that maps PDB structures to their corresponding protein sequences in the Swiss-Prot database (Boeckmann et al. 2003). Subsequently, we retrieved the DNA sequences corresponding to the protein sequences using the public Nucleotide Sequence Database (Kanz et al. 2005). This procedure provided us with a nonredundant data set of protein-coding DNA sequences that each had a corresponding protein tertiary structure in the PDB database. Each of the protein sequences was processed to produce output from the DSSP program (Kabsch and Sander 1983), from which solvent accessibility of protein chains can be generated. The DSSP data were used to determine the solvent accessibility scores $E_s(i)$ for a sequence $i$ folded into a known structure. To be able to compute the pairwise interaction score $E_p(i)$, the three-dimensional coordinates of the C–α, C–β, O, and N atoms for each amino acid were extracted from the PDB data.

Because there is excessive uncertainty in doing evolutionary inference with short proteins, we eliminated protein-coding sequences with fewer than 200 nucleotides. We also removed the longest protein for reasons of computational feasibility. All results reported here are based on the remaining 1,195 proteins.

For each of these 1,195 protein families, the amino acid sequence information resulting from the translated DNA exactly matches the corresponding amino acid sequence information in the PDB database. Our requirement of exact matching serves to help eliminate PDB entries that are not naturally occurring. Because there are a variety of data
collection issues that can cause terminal gaps, protein families were not removed from our
data set if the alignments between the amino acid sequences from translated DNA and the
PDB entries exhibited terminal gaps. However, only the DNA and structure information
corresponding to the region of exact matching was used in our analyses.

Methods

Our estimation approach is a simplified version of that described by Robinson et al. (2003).
Because our data sets consist only of single sequences, we need not consider the nucleotide
substitutions that separate a group of homologous protein–coding DNA sequences. Instead,
our Bayesian inference procedure is based entirely on equation (2.2) and on the prior dis-
tributions of parameters. By having prior distributions of $s$ and $p$ that are independent of
one another and independent of the nucleotide frequency prior, the posterior density for a
sequence $i$ is

$$
p(s, p, \pi_A, \pi_C, \pi_G, \pi_T | i) = \frac{p(i | s, p, \pi_A, \pi_C, \pi_G, \pi_T) p(s) p(p) p(\pi_A, \pi_C, \pi_G, \pi_T)}{p(i)}. \tag{2.3}
$$

To simplify notation, the vector $\theta = \{s, p, \pi_A, \pi_C, \pi_G, \pi_T\}$ will be employed so that

$$
p(s, p, \pi_A, \pi_C, \pi_G, \pi_T | i) = p(\theta | i). \tag{2.4}
$$

We select a prior distribution for $s$ that is uniform between $-5$ and $5$ and a prior
distribution for $p$ that is uniform between $-0.3$ and $0.3$. Although the biologically plausible
values of $s$ and $p$ are positive, we intentionally specified the prior distributions so that
positive values of $s$ and $p$ were not favored a priori. We therefore view posterior densities
concentrated on the positive values of these parameters as partial validations of our model.
The endpoints of the uniform priors for $s$ and $p$ were chosen from our prior experience with
the dependent–sites model. We wanted prior intervals that were symmetric about 0 and
that were as short as possible while still yielding very low posterior density near the interval
endpoints. The incentive for having short prior intervals stems from the grid-based Gibbs
sampler that we employed (see below and also see Robinson et al. 2003). The joint prior distribution of \( \pi_A, \pi_C, \pi_G, \) and \( \pi_T \) was uniform for all analyses reported here. Because the sum of these four parameter values must be one, this prior on the \( \pi \) parameters is equivalent to a Dirichlet distribution with 4 categories that has the hyperparameter corresponding to each category set to one.

To approximate the posterior density, we employ a Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). The parameter values after \( t \) steps \((t > 0)\) of the Markov chain will be denoted \( \theta(t) \). The algorithm begins with any initial set of parameter values \( \theta(0) \) that has nonzero posterior density. Thereafter, new states of the Markov chain are determined by randomly proposing new parameter values and then randomly accepting or rejecting the proposed state with the appropriate probability. Let the proposed parameter values at step \( t + 1 \) of the Markov chain be denoted \( \theta'(t) \). These proposed values are obtained by randomly sampling from a probability density \( J(\theta'(t) | \theta(t)) \) that is conditional upon the current state of the Markov chain. A new proposal \( \theta' \) is accepted with probability equal to the minimum of 1 and \( r \), where

\[
r = \frac{p(\theta'|i)J(\theta(t)|\theta')}{p(\theta(t)|i)J(\theta'|\theta(t))} = \frac{p(i|\theta')p(\theta')J(\theta(t)|\theta')}{p(i|\theta(t))p(\theta(t))J(\theta'|\theta(t))} \\
= \frac{(e^{-2s_E(i)} - 2p_E(i) \prod_{m=1}^{L} \pi_{i,m}^{'})}{(e^{-2s_E(i)} - 2p_E(i) \prod_{m=1}^{L} \pi_{i,m})} \\
\times \frac{\left( \sum_k e^{-2s_E(k)} - 2p_E(k) \prod_{m=1}^{L} \pi_{k,m} \right) p(\theta') J(\theta(t)|\theta')}{\left( \sum_k e^{-2s_E(k)} - 2p_E(k) \prod_{m=1}^{L} \pi_{k,m} \right) p(\theta(t)) J(\theta'|\theta(t))}
\]

(2.5)

and where \( J(\theta(t)|\theta') \) is the probability density of proposing \( \theta(t) \) if the current parameter values had actually been \( \theta' \). If the proposal is accepted \( \theta(t+1) = \theta' \), and otherwise \( \theta(t+1) = \theta(t) \).

The sums in the numerator and denominator of equation (2.5) are challenging to evaluate because they range over all DNA sequences that have the same length as sequence \( i \). We approximate the ratio of these two sums via the grid-based Gibbs sampler of Robinson et al. (2003). Pilot experiments indicated that our implementation of this grid-based Gibbs sampler yielded satisfactory results (data not shown). Rather than simultaneously proposing changes to all parameters represented by \( \theta \), our implementation has one proposal that changes only \( s \), another that changes only \( p \), and a third that changes the \( \pi \) parame-
ters. The proposed solvent accessibility parameter \( s' \) is sampled from a uniform distribution between \( s - 1 \) and \( s + 1 \). When \( s - 1 \) is less than \(-5\) and when \( s + 1 \) is greater than \( 5 \) (i.e., when there is a risk of proposing values beyond the boundaries of the uniform prior distribution), the lower or upper limits of the proposal density are shortened so that proposed values do not have prior densities of zero. The proposal density for \( p' \) is similar in nature to that for \( s' \), except that the uniform proposal distribution for \( p' \) is between \( p - 0.1 \) and \( p + 0.1 \). The proposal density for the \( \pi \) parameters is a Dirichlet distribution of order four with parameters \( \pi_A \times 100 \), \( \pi_C \times 100 \), \( \pi_G \times 100 \), and \( \pi_T \times 100 \). The proposal acceptance proportions varied among the 1,195 protein families but about 31\% of \( s \) proposals, 15\% of \( p \) proposals, and 24\% of \( \pi \) proposals were accepted.

We implemented the Metropolis-Hastings algorithm by cycling through the \( s \), \( p \), and \( \pi \) proposals. We refer to one set of these 3 proposals as a cycle. To design a Metropolis-Hastings strategy that would reliably approximate the posterior distribution of \( \theta \) and that would simultaneously be feasible for analyzing 1,195 single-sequence data sets, we arbitrarily selected 10 of the data sets for particularly careful examination. Each of the 10 different data sets was analyzed by initializing each of 5 different Metropolis-Hastings runs at a different set of initial parameter values \( \theta^{(0)} \).

With these pilot analyses, we obtained very similar estimates of the posterior distribution for the 5 different runs when we discarded the first 10,000 Metropolis-Hastings cycles and then sampled every 100 cycles until 1,000 samples had been obtained (data not shown). The Tracer program calculates an “effective sample size” from Metropolis-Hastings output (Rambaut and Drummond 2003). We found that the effective sample size tends to be almost the same as the actual sample size of 1000 when Metropolis-Hastings runs had the burn-in of 10,000 cycles and the sampling frequency of 100 cycles as described above. Similarly, we found that consecutive Metropolis-Hastings samples have low correlation. The Metropolis-Hastings settings that we selected yielded potential scaling reduction statistic values (Gelman et al. 2003) that were less than 1.001 for all runs with the 10 data sets. A value of this statistic that is close to 1 indicates that increasing the number of Metropolis-Hastings cycles may not greatly improve posterior density approximations. Because our Metropolis-Hastings implementation seems well-behaved for the posterior density approximations of the 10 selected data sets, we elected to use the same Metropolis-Hastings settings to analyze the entire collection of 1,195 single-sequence data sets.
The distribution of tertiary structure impact among proteins

The Metropolis-Hastings analyses yield samples from the posterior density \( p(\theta|i) \) for each of the 1,195 protein-coding sequences \( i \). These samples can then be used to estimate the posterior densities of individual parameters. The posterior means of \( s_i \) and \( p_i \) for a protein family represented by sequence \( i \) can be approximated with the mean of the Metropolis-Hastings samples for that protein family. The estimated posterior means \( \hat{s}_i \) and \( \hat{p}_i \) will have little Monte Carlo error when the number of approximately independent samples from the joint posterior density of \( s_i \) and \( p_i \) is large. As noted above, our pilot experiments suggest our Metropolis–Hastings run lengths are long enough to yield posterior density estimates with relatively small Monte Carlo errors.

The predominantly positive values for \( \hat{s}_i \) and \( \hat{p}_i \) in figure 2.1 can be viewed as a partial validation of our model because they represent situations where sequences evolve so as to be compatible with their tertiary structure. These positive values cannot be attributed to the prior distributions for \( s \) and \( p \) because the prior distributions were uniform and centered at 0. An important source of variation between the estimates \( \hat{s}_i \) and \( \hat{p}_i \) and the true values \( s_i \) and \( p_i \) is the limited amount of information in a single-sequence data set. Therefore, credibility intervals for \( s_i \) and \( p_i \) can be wide. Because of this variation, a histogram of the 1,195 posterior mean estimates of \( s \) (or \( p \)) is expected to be more variable than the distribution of the 1,195 actual but unknown values of \( s \) (or \( p \)). Also, unlike the distribution of actual parameter values, the distribution of the 1,195 posterior mean estimates will be influenced by prior distributions for parameters.

To obtain a maximum-likelihood estimate of the distribution of actual values, we assign the distribution a parametric form. Specifically, we use independent normal densities to describe the variability of \( s \) and \( p \) values among proteins. The independence assumption for \( s \) and \( p \) seems reasonable because, among the 1,195 protein families, the correlation between the estimated posterior means for \( s \) and \( p \) was only about 0.035. The details of our maximum likelihood procedure can be found in Appendix A.

Figure 2.1A shows a histogram for the solvent accessibility parameter of the 1,195 posterior means \( \hat{s}_i \). The estimated normal density of true \( s \) values among proteins is a dashed line that is superimposed on the histogram. This estimated distribution has mean 0.735 whereas the mean of the histogram values is 0.771. The standard deviation of the estimated normal density (0.138) is less than the standard deviation of the histogram values.
Figure 2.1: **Distributions of protein tertiary structure impact parameters.** Maximum-likelihood estimates of the distributions among genes of $s$ and $p$ are superimposed on the histograms of the $\hat{s}_i$ and $\hat{p}_i$ estimates. (A) The solvent accessibility parameter $s$. (B) The pairwise interaction parameter $p$.

(0.292) because the posterior means are error-prone estimates of the true values of $s$.

Only 11 of the 1,195 posterior means of $s$ (i.e., about 0.9%) are negative. Because some of these posterior means may be negative due to estimation error, it is not surprising that an even smaller percentage (less than 0.01%) of the estimated normal density for $s$ is less than zero. Of these 11 protein families with negative posterior means, 9 are actually integral membrane or membrane–associated proteins. The negative $s$ estimates for these 9 families are not surprising because our sequence-structure compatibility measure (Jones, Taylor, and Thornton 1992; Jones and Thornton 1996) was empirically derived from globular proteins. Thus, this measure was not intended for evaluating integral membrane and membrane–associated proteins. The other two proteins with negative posterior means for $s$ are a lipid–binding protein and a hypothetical protein.

The histogram with the 1,195 posterior means of the $p$ parameter (fig. 2.1B) has an average of 0.038 and a standard deviation of 0.022, whereas the estimated normal density for $p$ has a mean of 0.034 and a standard deviation of 0.013. The percentage of the estimated normal density for $p$ that is less than 0 is about 0.5%. There are 12 (about 1.0%) protein families for which the estimated posterior mean of $p$ was less than 0. We have not identified
a common attribute among these 12 families.

The smaller standard deviation for the estimated normal density of $p$ than for the 1,195 posterior means is expected because the variability among posterior means will be inflated as a result of estimation error. For the same reason, we expected the percentage of the estimated normal density below 0 to be substantially smaller than the percentage of posterior means below 0. The percentages are closer than we expected and this is evidently attributable to the conflict between the asymmetry in the positively skewed histogram of posterior means for $p$ and the symmetry of the assumed normal density for the distribution of $p$ among protein families. A more flexible distributional form than the normal density may be warranted.

The posterior means for the solvent accessibility parameter estimates are more variable among short than long protein-coding sequences but the distribution of posterior means of $s$ among short sequences seems to be centered at about the same value as the distribution for long sequences (fig. 2.2A). This is presumably because short sequences contain less information about $s$. The posterior means for $p$ are also more variable for short than long sequences (fig. 2.2B). However, the pattern for $p$ differs from that for $s$. Long sequences tend to yield positive but relatively low posterior means for $p$. Short sequences tend to yield either low positive values or higher positive values for $p$. We are unsure how to explain this pattern but it seems to be responsible for inducing the positive skew in the histogram of $p$ posterior means (fig. 2.1B).

**Bayes factors**

Rodrigue, Philippe, and Lartillot (2006) have performed a careful examination of how a wide range of evolutionary models fit three different multiple-sequence data sets. They conclude that adding sequence-structure compatibility measures substantially improves model fit but that available sequence-structure measures are not sufficient by themselves to produce a good model fit. To supplement the sequence-structure compatibility measures, these authors find that further model improvements can be achieved via permitting additional rate variation among sites (Yang 1994; Yang, Nielsen, and Hasegawa 1998; Goldman and Whelan 2002) and additional protein-specific variation of amino acid composition (Cao et al. 1994; see).

The single-sequence nature of our data sets and the associated computational
demands prevent us from being able to carefully examine as many evolutionary model
variants as did Rodrigue, Philippe, and Lartillot (2006). Nevertheless, we were interested
for our 1,195 data sets to look at the improvement to model fit that is contributed by
sequence-structure compatibility measures. We compared a model \((M_1)\) that sets both the
\(s\) and \(p\) parameters to 0 with a more general model \((M_2)\) that does not constrain the \(s\)
and \(p\) values. We compare models by approximating Bayes factors. In our case, the Bayes
factor for a protein family \(k\) with sequence \(i_k\) is

\[
\frac{p(i_k|M_2)}{p(i_k|M_1)} = \frac{\int_s \int_p \int_{\pi} p(i_k|s, p, \pi, M_2)p(s, p, \pi|M_2)d\pi dp ds}{\int_{\pi} p(i_k|\pi, M_1)p(\pi|M_1)d\pi}.
\] (2.6)

In the above equation, the numerator integrand is the stationary distribution of sequence
\(i_k\) multiplied by the prior density of the \(s\), \(p\), and nucleotide frequency parameters. The
denominator integrand is the special case of equation (2.2), where \(s = 0\) and \(p = 0\) multiplied
by the prior density of the nucleotide frequency parameters.

Rodrigue, Philippe, and Lartillot (2006) employed a promising technique known
as thermodynamic integration (Lartillot and Philippe 2006) for Bayes factor approximation.
In contrast, we implemented the technique of Chib and Jeliazkov (2001; see also Chib 1995) to approximate Bayes factors. Details of our implementation strategy are in Appendix B.

Our Bayes factor approximations show that incorporating $s$ and $p$ improve model fit for the vast majority of our 1,195 protein families (fig. 2.3A). The approximation is less than 1 and thereby favors the more simple evolutionary model without protein structure (i.e., $M_1$) for only 12 of the 1,195 data sets (about 1.0%). Twice the natural logarithm of the Bayes factor is between 0 and 10 for an additional 61 data sets (about 5.1%). The remaining 1,122 protein families have a Bayes factor approximation with twice the natural logarithm exceeding 10. These results are fully consistent with the idea that tertiary structure should be incorporated into evolutionary models. As the proteins get longer, the Bayes factors tend to increase (fig. 2.3B and 2.3C). The positive correlation (0.71) between length and twice the natural logarithm of the Bayes factor is presumably attributable to the fact that longer sequences represent more data.

**The effects of a nonsynonymous change**

Because evolutionary rates from sequence $i$ to sequence $j$ are determined by $s(E_s(i) - E_s(j))$ and $p(E_p(i) - E_p(j))$ rather than solely by $s$ and $p$ (see eq. 2.1), we cannot conclude that solvent accessibility affects protein evolution more than pairwise interactions, simply because our estimates of $s$ tend to be bigger than our estimates of $p$. We examined the value of $s(E_s(i) - E_s(j))$ and $p(E_p(i) - E_p(j))$ for each possible nonsynonymous substitution that could change an observed sequence $i$. If natural selection has played a strong role in shaping an observed sequence $i$, most of these values should be negative.

For the observed protein-coding DNA sequence $i$, suppose there are $T_i$ possible nonsynonymous substitutions. Let $j_{i,1}$, $j_{i,2}$, ..., $j_{i,T_i}$ be the set of protein–coding sequences that are different from $i$ by exactly 1 nonsynonymous change. One assessment of the degree of optimization by natural selection for sequence $i$ is

$$\Delta_i = \Delta_{s,i} + \Delta_{p,i} \tag{2.7}$$

where

$$\Delta_{s,i} = \frac{1}{T_i} \sum_{m=1}^{T_i} s(E_s(i) - E_s(j_{i,m})) \tag{2.8}$$

37
Figure 2.3: **Distribution of Bayes factor estimates among proteins.** (A) A histogram of twice the natural logarithm of the Bayes factor estimates. The dashed line separates the protein families for which twice the natural logarithm of the Bayes factor is less than ten from those for which it exceeds 10. (B) Scatter plot of the nucleotide length of protein-coding DNA sequences against twice the natural logarithm of the Bayes factor estimate. (C) Scatter plot of the nucleotide length against the ratio of twice the natural logarithm of the Bayes factor estimate and nucleotide length.
\[ \Delta_{p,i} = \frac{1}{T_i} \sum_{m=1}^{T_i} p(E_p(i) - E_p(j_{i,m})) \]  

(2.9)

If protein tertiary structure has a strong evolutionary impact, the observed sequence \( i \) should tend to be more compatible with its structure than neighboring sequences. This would mean that rates of changes to neighboring sequences will tend to be small and the values of \( \Delta_i \), \( \Delta_{s,i} \), and \( \Delta_{p,i} \) should be less than 0 (see eq. 2.1).

Our estimates \( \hat{\Delta}_i \), \( \hat{\Delta}_{s,i} \), and \( \hat{\Delta}_{p,i} \) are obtained via equations (2.7), (2.8), and (2.9) except that the estimated posterior means \( \hat{s}_i \) and \( \hat{p}_i \) for protein family \( i \) replace \( s \) and \( p \).

The mean of the \( \hat{\Delta}_{s,i} \) among the 1,195 protein families is about −0.083, the mean for \( \hat{\Delta}_{p,i} \) is about −0.091, and the mean for \( \hat{\Delta}_i \) is about −0.174. Figure 2.4 shows the variation of \( \hat{\Delta}_i \), \( \hat{\Delta}_{s,i} \), and \( \hat{\Delta}_{p,i} \) among protein families. Some of this variation is due to estimation errors. A more direct impression of the variation among protein families in the actual values of \( \Delta_i \), \( \Delta_{s,i} \), and \( \Delta_{p,i} \) might be obtained through Stein–type shrinkage estimation (Stein 1956). The maximum likelihood estimates of the true distributions of \( s \) and \( p \) values among proteins (see fig. 2.1A) could serve as the prior distributions for the \( s_i \) and \( p_i \) values of each protein family \( i \). The resulting posterior mean estimates of \( s_i \) and \( p_i \) could then be regarded as empirical Bayes estimates. Plugging the empirical Bayes estimates of \( s_i \) and \( p_i \) values into equations (2.7), (2.8), and (2.9) would yield approximations for \( \Delta_i \), \( \Delta_{s,i} \), and \( \Delta_{p,i} \). Instead, we decided to focus on variation among the \( \Delta_i \) values with the permutation procedures that are described below.

Gene Ontology, tertiary structure, and evolution

One possibility is that certain protein attributes are associated with particularly strong constraints on evolution due to tertiary structure. To explore this, we employed the Gene Ontology (i.e., GO) database (Ashburner et al. 2000; Camon et al. 2004). The three fundamental units of this database are “Cellular Component”, “Molecular Function”, and “Biological Process.” Each unit is hierarchically organized into annotation terms and the terms within each hierarchy that apply to a given protein family are stored in the GO database. To lessen the impact of individual protein families on our investigation of GO terms, we decided to only consider terms for which at least 6 of the 1,195 protein families have been annotated with the term and for which at least 6 have not been annotated with
Figure 2.4: **Average selective effects of nonsynonymous changes.** (A) Histograms of $\hat{\Delta}_{s,i}$, $\hat{\Delta}_{p,i}$, and $\hat{\Delta}_i$. Alternate histogram categories are labelled and labels represent the central value of data points in the category. For example, the histogram category labelled “0” represents all points in the interval from $-0.025$ to $0.025$. The leftmost histogram category (labelled “*”) represents all points that are less than $-0.475$. (B) Scatter plot of $\hat{\Delta}_{s,i}$ versus $\hat{\Delta}_{p,i}$. 

\[\text{Figure 2.4: Average selective effects of nonsynonymous changes. (A) Histograms of } \hat{\Delta}_{s,i}, \hat{\Delta}_{p,i}, \text{ and } \hat{\Delta}_i. \text{ Alternate histogram categories are labelled and labels represent the central value of data points in the category. For example, the histogram category labelled “0” represents all points in the interval from } -0.025 \text{ to } 0.025. \text{ The leftmost histogram category (labelled “*”) represents all points that are less than } -0.475. \text{ (B) Scatter plot of } \hat{\Delta}_{s,i} \text{ versus } \hat{\Delta}_{p,i}. \]
Figure 2.5: Boxplots of $\hat{\Delta_i}$ estimates for the 23 Gene Ontology level-1 terms. The vertical lines indicate the median $\hat{\Delta_i}$ value among the 456 protein families that were annotated with at least one of the 209 Gene Ontology terms. (A) Cellular Component terms. (B) Molecular Function terms. (C) Biological Process terms.

For the 209 terms that satisfied this criterion in the GO database on October 11 of 2006, the numbers in the “Cellular Component”, “Molecular Function”, and “Biological Process” divisions were respectively 35, 65, and 109. The GO hierarchy places 23 of these 209 terms at the most general “Level 1” rank and the distributions of the $\hat{\Delta_i}$ values for these 23 terms is summarized in figure 2.5. A figure summarizing the distributions for all 209 GO terms is available in figures 2.6, 2.7 and 2.8 and the Supplementary Material online.

For each of these 209 GO terms, we calculated the sample median of the $\hat{\Delta_i}$ among protein families associated with the GO term. We refer to these sample medians as optimization measures of the GO terms. We expect the optimization measure associated with a GO term to be negative because most possible nonsynonymous substitutions lessen sequence–structure compatibility.

A null hypothesis is that GO attributes are independent of the strength of the structure–evolution relationship. As a test statistic, we use the sample variance among
the 209 optimization measures. We expect more variation among optimization measures if annotation and evolutionary influence of structure are related than if they are unrelated. The null distribution of the test statistic is simulated by fixing the GO attributes for each protein family and then calculating the sample variance of the optimization measures after permuting the $456 \hat{\Delta}_i$ among protein families that were assigned to at least one of the 209 GO terms. This permutation procedure preserves the complex hierarchical structure among GO terms while simultaneously being consistent with the null hypothesis. We simulated the null distribution of the test statistic by performing the permutation procedure 10,000 times. The proportion of the simulated test statistic values that were greater than the observed value was quite small (0.001). A small proportion (0.001) was also observed when the test statistic was the variance among GO terms of the mean $\hat{\Delta}_i$ for each GO term. These small proportions suggest that there is an association between GO attributes and optimization measures.

However, the association does not appear to be strong. The null hypothesis that GO attributes are independent of the strength of the structure–evolution relationship could be violated either if certain attributes tend to have unusually high optimization measure values or if certain ones tend to have unusually low optimization values. These two violations have different implications. The alternative hypothesis that tertiary structure has a particularly strong impact on evolution of proteins with certain attributes would lead to unusually low optimization values. Corresponding to another alternative hypothesis, unusually high optimization measures might be generated either if tertiary structure is not particularly important for proteins with some GO attributes or if our evolutionary model is not very appropriate for proteins with some attributes.

Because of the different implications of unusually high versus unusually low optimization measures, we did two additional hypothesis tests. For the alternative hypothesis that leads to low values of optimization measures, our test statistic was the lowest optimization measure value among the 209 GO attributes. The lowest value was about $-0.266$ and was achieved by “carrier activity.” Although this test statistic differs from the previous test that used the variance among optimization measures, our null hypothesis has not changed and we therefore applied the same permutation procedure. The low value for “carrier activity” does not seem particularly noteworthy because more than 29% of the null distribution for the test statistic was even lower.

To examine the remaining alternative hypothesis, our test statistic was the highest
optimization measure value among the GO attributes. This highest value was about -0.071 and was achieved by the GO term “extracellular region part.” The proportion of simulated test statistic values that were greater than the observed “extracellular region part” value was 0.0077. Because only 6 protein families were annotated with “extracellular region part” and because there was considerable variation among the \( \hat{\Delta}_i \) estimates for these 6 families, we also looked at a modified test statistic that was the maximum among GO terms of the means of the \( \hat{\Delta}_i \) values for each term. When we used the mean rather than the median \( \Delta_i \) estimate for the GO terms, “extracellular region part” had a mean of -0.09844 and was not the GO term with the highest value. Using the maximum of the 209 means rather than the maximum of the 209 medians as the test statistic, 0.1269 of the permuted data sets yielded a higher value than -0.09844. Therefore, we are hesitant to make strong conclusions about the “extracellular region part” GO term.

The \( \hat{\Delta}_i \), \( \hat{\Delta}_{s,i} \), and \( \hat{\Delta}_{p,i} \) values are computed by comparing observed sequences to sequences that are one nonsynonymous change different. These values therefore assess the local adaptation of an observed sequence relative to its sequence neighborhood. In contrast, our Bayes factor estimates are measures of global adaptation because they represent the overall evidence that tertiary structure influences protein evolution. Twice the natural logarithms of these Bayes factor estimates have an approximately linear relationship with sequence length (fig. 2.3B) and can be normalized by dividing by the sequence length (fig. 2.3C). Similar permutation tests to those described above were performed for each GO term when the test statistic was the median of twice the natural logarithm of the Bayes factor normalized for sequence length. The results using these global measures of adaptation were qualitatively similar to those based upon the \( \hat{\Delta}_i \), \( \hat{\Delta}_{s,i} \), and \( \hat{\Delta}_{p,i} \) measures of local adaptation and are therefore not detailed here.

We also examined possible connections between GO terms and evolution by forming a \( 2 \times 2 \) contingency table for each of the 209 GO attributes. Protein families were assigned to contingency table columns according to whether their \( \hat{\Delta}_i \) value was above or below the median \( \hat{\Delta}_i \) value among the 456 protein families that were annotated with at least one of the 209 GO terms. Protein families were assigned to rows according to whether they were (or were not) annotated with the GO term. The null hypothesis that protein annotation is independent of the \( \hat{\Delta}_i \) value was investigated by applying Fisher’s exact test. We concentrated on the two-tailed alternative hypothesis that a GO term may be associated either with high or low \( \hat{\Delta}_i \) values.
The result was 209 $P$-values. Because some of these $P$-values will be low by chance, we selected the lowest of the 209 values as a test statistic for evaluating the null hypothesis that the evolutionary impact of protein structure is independent of annotation. We approximated the sampling distribution of this test statistic with a permutation procedure similar to that described above. We found that 5% of the sampling distribution is less than 0.001165. The four GO terms in the actual data set with a $P$-value lower than 0.001165 are “intracellular” ($P$-value of 0.001161), “carbohydrate metabolism” (0.000371), “organelle” (0.000429), and “intracellular organelle” (0.000429). The $P$-values for “organelle” and “intracellular organelle” are necessarily the same because each is represented by exactly the same 65 families in our data set. These two GO terms and “intracellular” yield a relatively low median $\hat{\Delta}_i$ values while “carbohydrate metabolism” has a median $\hat{\Delta}_i$ value that is relatively high. Despite the statistical evidence of a relationship between the evolutionary effect of protein structure and these GO terms, we cannot offer a compelling biological explanation for why these terms stand out. Moreover, the strong statistical evidence for a relationship should not be misconstrued as evidence for a strong relationship. The lack of dramatic variation of median $\hat{\Delta}_i$ values among GO terms argues against a strong relationship (see fig. 2.5 and the Supplementary material).

Discussion

The sequence–structure compatibility method that we adopt was designed for protein fold recognition and it combines solvent accessibility and pairwise interaction information (Jones, Taylor, and Thornton 1992; Jones and Thornton 1996). We find strong evidence that incorporating this structural information improves evolutionary models. We also find substantial variation among protein families in the evolutionary impacts of solvent accessibility and pairwise interactions. This suggests that protein fold recognition might be improved if the contributions of solvent accessibility and pairwise interaction information are allowed to vary among protein folds.

Melo, Sánchez, and Sali (2002) conclude that both pairwise amino acid interactions and solvent accessibility can make important contributions to successful protein fold recognition, but that pairwise interactions are more important. Prior to this study, importance in evolution of pairwise interactions versus solvent accessibility had been largely
unexamined. The slightly more extreme mean value of $\hat{\Delta}_{p,i}$ among the 1,195 protein families hints that pairwise interactions tend to be more important in evolution than does solvent accessibility. However, $\hat{\Delta}_{s,i} = -0.083$ and $\hat{\Delta}_{p,i} = -0.091$ have comparable magnitude and the most important message emerging from these analyses is that pairwise interactions and solvent accessibility seem to have roughly comparable impact on protein evolution.

Although there seems to be a real association between the evolutionary importance of tertiary structure and GO terms, the tendency is not strong. A less parametric study (Yu and Thorne 2006b) addressed whether the degree of spatial clustering of amino acid replacements on protein tertiary structure is associated with GO attributes. Strong association was not found in this previous study. We cannot exclude the possibility that associations between evolutionary impact of tertiary structure and GO terms are strong albeit subtle and can only be found by employing more realistic evolutionary models.

The evolutionary model used in this study would be a very simple codon substitution model except for the $s$ and $p$ parameters that have been added to incorporate protein tertiary structure. Simple codon substitution models have been improved by relaxing the assumption that all codons share the same value of the $\omega$ parameter (see eq. 2.1) and therefore the same ratio between nonsynonymous and synonymous rates (Nielsen and Yang 1998; Yang and Nielsen 1998). These approaches allow different codons to be associated with different $\omega$ values via a variety of schemes (Yang et al. 2000; Yang and Nielsen 2002; Yang and Swanson 2002). The dependent–sites model could also be modified to permit variation of $\omega$ values among codons. The stationary distribution of sequences (eq. 2.2) would not be affected by this modification. This means there is a wide range of codon substitution models that would all yield identical estimates of the $s$ and $p$ parameters from single–sequence data sets.

As with variation of $\omega$ values among codons, adding variation of evolutionary rates among codons via the discretized gamma technique of Yang (1994) would not alter the stationary distribution of sequences. Rate variation parameters and $\omega$ parameters do not have the close connections to phenotype possessed by the $s$ and $p$ parameters. For interspecific studies, there has been surprisingly little work on explicitly incorporating aspects of phenotype into models of molecular evolution. This is disappointing because the connection between phenotype and sequence change is central to evolution. Only slight modifications to the statistical inference procedure adopted here are necessary for studying the link between evolution and other aspects of phenotype (e.g., see Yu and Thorne 2006).
Because the stationary distribution of sequences is affected by mutation as well as natural selection, one of many worthwhile model improvements would be a more careful treatment of mutation. Recently, several techniques for making evolutionary inferences with context-dependent mutation patterns have been developed (Jensen and Pedersen 2000; Pedersen and Jensen 2001; Hwang and Green 2004; Pedersen et al. 2004; Siepel and Haussler 2004; Christensen, Hobolth, and Jensen 2005). It would be straightforward to incorporate context-dependent mutation into models of protein change.

There is a tension between mutation and natural selection. Mutation is a possible explanation for why observed sequences might not be phenotypically optimal. Another — and not mutually exclusive possibility — is finite evolutionary history. Evolution may not yet have achieved a stationary distribution of sequences. This is particularly likely if the nature of the evolutionary process experiences fundamental changes over time. For example, the optimal phenotypic value might change. For our inferences, we have ignored the possibility that evolution has not achieved a stationary distribution of sequences. If we are correct to ignore this possibility, lack of optimal sequence-structure compatibility would be explained by the balance between mutation and natural selection. It is this assumed balance that allows us to estimate the $s$ and $p$ parameters. Data sets consisting of multiple homologous sequences would contain information that might let the stationarity assumption be evaluated. The question would then be whether the observed sequences are characteristic of the kinds of changes that are inferred to have occurred since the most recent common ancestral sequence existed.

A remaining explanation for why observed sequences may not seem to be phenotypically optimal is simply that the evolutionary model is flawed. Obviously, measures of sequence-structure compatibility could be improved. More importantly, phenotype is not simply a matter of protein structure. Phenotypic effects of a protein also depend on the environment and on the amount and pattern of protein expression.

Differences in the relative rates of specific nonsynonymous changes seem easier to explain with tertiary structure than through levels of protein expression. However, there is a strong association between the expression rate of a protein and its overall rate of nonsynonymous change (Drummond, Raval, and Wilke 2005). Moreover, expression rates in yeast seem to be more strongly correlated with the overall nonsynonymous rate of a gene than does protein structure (Bloom et al. 2006).

A satisfactory evolutionary model should reflect the fact that protein expression
is phenotype whereas the regulatory regions that influence expression are genotype. Evolutionary models that connect DNA sequences of regulatory regions and expression phenotypes are possible. Promising strategies for describing evolution of regulatory elements have already been proposed (Berg, Willmann, and Lässig 2004; Mustonen and Lässig 2005). An improvement upon the probabilistic model described here would have both sequence-structure compatibility measures and in silico predictions of protein expression simultaneously influence the evolution of the coding and regulatory regions of a gene.

Certainly, the evolutionary model used here could be improved in a variety of ways. However, we view the current status of probabilistic evolutionary models in an optimistic light. With model–based evolutionary inference, the evolutionary relevance of different biological features can be statistically evaluated and the most important features can be incorporated into improved evolutionary models. Now that statistical procedures exist for making inferences when phenotype and context-dependent mutation affect evolutionary rates, development of usefully more realistic evolutionary models should proceed at a rapid pace.

Acknowledgments

We thank D. Jones for guidance. We also thank A. von Haeseler and two anonymous reviewers for their input. This research was supported by N.I.H. grant GM070806 and N.S.F. grant D.E.B-0445180. S. C. Choi was additionally supported by Korea Science and Engineering Foundation grant M06-2003-000-10086-0. A. Hobolth was additionally supported by the Danish Research Council grant 21-04-0375. H. Kishino was supported by J.S.P.S. (SR B-16300086).

Supplementary Material

Supplementary materials (fig. 2.6, 2.7, 2.8) can be found online at the Molecular Biology and Evolution website (http://www.mbe.oxfordjournals.org).
Figure 2.6: Boxplots of $\hat{\Delta}_i$ estimates for the 35 Gene Ontology Cellular Component terms. The vertical lines indicate the median $\hat{\Delta}_i$ value among the 456 protein families that were annotated with at least one of the 209 Gene Ontology terms.
Figure 2.7: Boxplots of $\hat{\Delta}_i$ estimates for the 65 Gene Ontology Molecular Function terms. The vertical lines indicate the median $\hat{\Delta}_i$ value among the 456 protein families that were annotated with at least one of the 209 Gene Ontology terms.
Figure 2.8: Boxplots of $\hat{\Delta}_i$ estimates for the 109 Gene Ontology Biological Process terms. The vertical lines indicate the median $\hat{\Delta}_i$ value among the 456 protein families that were annotated with at least one of the 209 Gene Ontology terms.
Literature Cited


Appendices

In the appendix A, we explain our maximum likelihood estimation procedure for inferring the distribution of solvent accessibility and pairwise interaction impact among proteins. In the appendix B, we describe our Bayes factor approximation.
Appendix A

To explain our maximum likelihood estimation procedure for inferring the distribution of $s$ and $p$ values among proteins, we introduce additional notation. The 1,195 sequences will be denoted as $i$ whereas the 1,195 true values of $s$ and $p$ will be $s = \{s_1, s_2, \ldots, s_{1195}\}$ and $p = \{p_1, p_2, \ldots, p_{1195}\}$. Here, $s_i^{(s)}$ and $p_i^{(s)}$ will refer to the $g^{th}$ of $G$ approximately independent samples from the joint posterior density of $s_i$ and $p_i$. For the 1195 single-sequence analyses, $G$ was 1,000. The mean and variance of the normal distribution of $s$ values among proteins will be collectively denoted $\alpha$, whereas the mean and variance of the normal distribution for $p$ values among proteins will collectively be $\beta$. The uniform prior distributions for $s$ and $p$ will be referred to as $\alpha_0$ and $\beta_0$.

The goal of the procedure is to find the values of $\alpha$ and $\beta$ that maximize the likelihood $p(i|\alpha, \beta) = p(i_1, i_2, \ldots, i_{1195}|\alpha, \beta)$. The likelihood approximation exploits the assumption that the probability of observing a sequence for a specific protein family depends on the $s$ and $p$ values of the family but, conditional upon the $s$ and $p$ values for the family, the probability of observing a sequence is independent both of the prior density for $s$ and $p$ and of the distribution of $s$ and $p$ among protein families. In other words,

$$p(i|s, p) = p(i|s, p, \alpha, \beta) = p(i|s, p, \alpha_0, \beta_0). \quad (2.10)$$
The likelihood approximation is justified as follows:

\[
p(i|\alpha, \beta) = \int_s \int_p p(i, s, p|\alpha, \beta) \, dp \, ds
\]

\[
= \int_s \int_p p(i|s, p, \alpha, \beta)p(s, p|\alpha, \beta) \, dp \, ds
\]

\[
= \int_s \int_p p(i|s, p)\, p(s, p|\alpha, \beta) \, dp \, ds
\]

\[
= \int_s \int_p p(i|s, p, \alpha_0, \beta_0)p(s, p|\alpha, \beta) \, dp \, ds
\]

\[
= \int_s \int_p p(s, p|i, \alpha_0, \beta_0)p(i|\alpha, \beta) \, dp \, ds
\]

\[
= p(i|\alpha_0, \beta_0) \int_s \int_p \frac{p(s, p|\alpha, \beta)}{p(s, p|\alpha_0, \beta_0)} p(s, p|i, \alpha_0, \beta_0) \, dp \, ds
\]

\[
= p(i|\alpha_0, \beta_0) E \left[ \frac{p(s, p|\alpha, \beta)}{p(s, p|\alpha_0, \beta_0)} \bigg| \alpha_0, \beta_0, i \right]
\]

\[
= p(i|\alpha_0, \beta_0) E \left[ \frac{p(s|\alpha)p(p|\beta)}{p(s|\alpha_0)p(p|\beta_0)} \bigg| \alpha_0, \beta_0, i \right]
\]

\[
= p(i|\alpha_0, \beta_0) E \left[ \frac{p(s|\alpha)}{p(s|\alpha_0)} \bigg| \alpha_0, i \right] E \left[ \frac{p(p|\beta)}{p(p|\beta_0)} \bigg| \beta_0, i \right]
\]

\[
= p(i|\alpha_0, \beta_0) E \left[ \prod_{k=1}^{1195} \frac{p(s_k|\alpha)}{p(s_k|\alpha_0)} \bigg| \alpha_0, i_k \right] E \left[ \prod_{k=1}^{1195} \frac{p(p_k|\beta)}{p(p_k|\beta_0)} \bigg| \beta_0, i_k \right]
\]

\[
\simeq p(i|\alpha_0, \beta_0) \prod_{k=1}^{1195} \frac{1}{G} \sum_{g=1}^{G} \left\{ \frac{p(s_k^{(g)}|\alpha)}{p(s_k^{(g)}|\alpha_0)} \right\} E \left[ \prod_{k=1}^{1195} \frac{p(p_k|\beta)}{p(p_k|\beta_0)} \bigg| \beta_0, i_k \right]. \tag{2.11}
\]

Numerical optimization routines can find the \( \alpha \) and \( \beta \) that maximize the above approximate likelihood. These routines need not consider the factor \( p(i|\alpha_0, \beta_0) \) because it is not a function of \( \alpha \) or \( \beta \). For our analyses, the simplex routine of Nelder and Mead (1965) as implemented by the GNU Scientific Library (Galassi et al. 2005) has proven satisfactory for numerical optimization.

The uniform prior densities specified by \( \alpha_0 \) and \( \beta_0 \) constrain \( s \) and \( p \) to a finite range while the normal densities specified by \( \alpha \) and \( \beta \) have no such constraint. This difference in
support between \( \alpha_0 \) and \( \beta_0 \) versus \( \alpha \) and \( \beta \) has the potential to affect inference. However, the difference in support has no important consequences for our analyses because the posterior densities for \( s \) and \( p \) have little mass near the boundaries of the uniform priors that are set by \( \alpha_0 \) and \( \beta_0 \).

**Appendix B**

In this appendix, we describe how to approximate the Bayes factor using Chib (1995) and Chib and Jeliazkov (2001). Recall from equation (2.6) that the Bayes factor is given by

\[
\frac{p(i|M_2)}{p(i|M_1)} = \frac{\int_s \int_p \int_\pi p(i|s, p, \pi, M_2)p(s, p, \pi|M_2)d\pi dp ds}{\int_\pi p(i|\pi, M_1)p(\pi|M_1)d\pi}.
\]

Chib (1995) notes that

\[
p(\theta|i, M) = \frac{p(\theta, i|M)}{p(i|M)} = \frac{p(i|\theta, M)p(\theta|M)}{p(i|M)},
\]

so that for any \( \theta \) we have

\[
p(i|M) = \frac{p(i|\theta, M)p(\theta|M)}{p(\theta|i, M)}.
\]

Chib (1995) assumes the two terms in the numerator of equation (2.12) are easy to calculate and shows how to estimate the denominator from output of a Gibbs sampler. Chib and Jeliazkov (2001) extend Chib (1995) to allow estimating the denominator from Metropolis-Hastings output.

Our case is more complicated because the first term in the numerator, the stationary distribution of a sequence, cannot be calculated directly. However, ratios of stationary distributions can be approximated using Robinson et al. (2003). If we let \( \theta_1 = \pi_1 \) and \( \theta_2 = (s_2, p_2, \pi_2) \) and apply equation (2.12), we obtain the Bayes factor

\[
\frac{p(i|M_2)}{p(i|M_1)} = \frac{p(i|\theta_2, M_2)}{p(i|\theta_1, M_1)} \cdot \frac{p(\theta_2|M_2)}{p(\theta_1|M_1)} \cdot \frac{p(\theta_1|i, M_1)}{p(\theta_2|i, M_2)}.
\]
The first term is a ratio of stationary distributions and can be approximated using Robinson et al. (2003). The second term, the ratio of the priors of the two models, is also easily calculated. The last term is approximated using the Metropolis-Hastings output, as described in Chib and Jeliazkov (2001).
Chapter 3

Designing evolutionary rates to yield desired stationary distributions of sequences

Sang Chul Choi, Benjamin Redelings, and Jeffrey L. Thorne
Introduction

Evolutionary analysis of interspecific DNA or protein sequence data often proceeds by adopting explicit stochastic models of sequence change. Typically, values of parameters in these models determine the relative instantaneous rates by which each possible sequence changes to each other possible sequence. For widely used evolutionary models, computationally feasible likelihood-based inference can be performed by applying the pruning algorithm of Felsenstein (Felsenstein 1981).

In order to calculate likelihoods, Felsenstein’s pruning algorithm requires transition probabilities (i.e., the probability of observing a sequence state at the end of a branch on an evolutionary tree conditional upon the sequence state at the beginning and upon the values of model parameters). Simple models of sequence change treat DNA or protein sequences as a concatenated series of units that all evolve on the same phylogeny but that independently experience changes. This allows the likelihood of an entire data set to be decomposed into a product of likelihoods for each of the independently evolving units. These units might represent nucleotides with 4 possible states, amino acids with 20 possible states, or codons with 61 possible states.

When the number of possible character states per unit is small (e.g., 4 or 20 or 61), exponentiation of instantaneous rate matrices to obtain transition probabilities is computationally tractable and Felsenstein’s pruning algorithm is feasible. When sequences are not easily decomposed into a series of independently evolving units, it can be difficult to convert instantaneous rate matrices to transition probabilities because the large number of possible sequences makes the dimension of the instantaneous rate matrix prohibitively large.

A strategy pioneered by Jensen and Pedersen (Jensen and Pedersen 2000; Pedersen and Jensen 2001) exists for performing likelihood-based evolutionary inference when dependence among changes at different sequence positions makes the pruning algorithm intractable. Variants of the Jensen-Pedersen inference technique have been applied when evolutionary dependence among sequence positions is attributable to context-dependent mutation (Hwang and Green 2004; Siepel and Haussler 2004; Christensen, Hobolth, and Jensen 2005) and when it is explained by natural selection (Robinson et al. 2003; Rodrigue et al. 2005; Yu and Thorne 2006). In this chapter, we consider dependence due to natural selection but we do not focus on parameter inference. Instead, we consider how to param-
eterize evolutionary rates so that the induced evolutionary process is time–reversible and has a desired stationary distribution.

We concentrate on two cases. In the first, protein sequence data are used to train a variable length Markov model (VLMM) and then evolutionary rates are parameterized to produce an evolutionary model with a stationary distribution of amino acid sequences that matches the distribution of sequences according to the VLMM. As a consequence, the rate of a particular codon substitution at a particular sequence position will be affected by the amino acids specified by nearby codons.

In the second case, the stationary distribution of amino acid sequences follows a profile hidden Markov model (see Durbin et al. 1998) that has been trained on sequences from a protein family of interest. This sort of stationary distribution is particularly appealing because variation in sequence lengths is possible with a profile hidden Markov model (HMM). As a result, the evolutionary model permits insertions and deletions as well as codon substitutions.

We contrast both the VLMM and profile HMM evolutionary models to corresponding models that would result if all sequence changes were neutral. The contrasts permit the evolutionary rates in the VLMM and profile HMM models to be interpreted as being proportional to the product of a mutation rate and a probability that the mutation is fixed. Following the pioneering approach of Halpern and Bruno (1998), this interpretation allows us to obtain a crude estimate of the effective population size multiplied by the difference in relative fitnesses of the two sequences involved in a change.

Variable Length Markov Models

Although proteins have a tertiary structure, protein sequences are one-dimensional information and can be expressed via the order of amino acids from the amino-terminus to the carboxyl-terminus. The one-dimensional information about proteins can therefore be summarized as a sequence of discrete states with each state representing one of twenty possible amino acid types. The ability to summarize protein sequences in this way naturally leads to protein sequences being modelled as discrete–state discrete–time Markov chains.

We let $L(i)$ and $L(I)$ respectively represent the number of codons in protein-coding sequence $i$ and the number of amino acids in protein sequence $I$. Although the VLMM does
not allow the possibility of variation of sequence lengths, we explicitly condition upon sequence lengths for the neutral model formulae because we will later incorporate length variation into the neutral model.

Consider a protein sequence $I$ with $L(I)$ residues. The amino acid occupying position $k$ of the protein will be denoted by $I_k$ whereas the subsequence consisting of the first $k$ residues of the protein will be $I^k$. Taking liberty with conventional probabilistic notation by not distinguishing between random variables and their values, a 0th order discrete–state discrete–time Markov model for protein sequence organization would have

$$P(I) = P(I^{L(I)}) = \prod_{k=1}^{L(I)} P(I_k)$$

(3.1)

whereas a Markov model of order $r \geq 1$ would have

$$P(I) = P(I^r) \prod_{k=r+1}^{L(I)} P(I_k|I_{k-1}, \ldots, I_{k-r}).$$

(3.2)

Well–developed statistical techniques are available for analysis of data generated according to discrete–state discrete–time Markov chains (e.g., see Guttorp 1995). One decision when employing discrete–time Markov models is the choice of order for the Markov chain. The advantage of Markov models with low orders is that they are simple and have few parameters to estimate. A disadvantage is that models of low order can be relatively unrealistic and may not fit data well. For example, protein sequences tend to have stretches of consecutive hydrophilic and stretches of consecutive hydrophobic amino acid residues and a Markov model of order 0 would not reflect this tendency. Higher order Markov models are more flexible and can be more realistic, but they can suffer from being overly parameter–rich. A homogeneous discrete–state discrete–time Markov model of order $r$ has $20^r \times (20 – 1)$ parameters for the case where the discrete states each represent one of 20 possible amino acid types. Unless the order $r$ is quite small, this means that the number of parameters to be estimated will be very large and it also means that data sets will not yield much information about the values of most parameters except when the data set is very large.

Variable length Markov models can offer a parameterization advantage over those with fixed order. Consider the transition probability $P(I_k|I_{k-1}, \ldots, I_{k-r})$ that is associated
with a Markov model of fixed order \( r \) and with a subsequence matching \( I_{k-1}, \ldots, I_{k-r} \). If this subsequence is rare in the data set used to train the Markov model, then estimates of the transition probability \( P(I_k|I_{k-1}, \ldots, I_{k-r}) \) are apt to be unreliable. In such a situation, one may instead desire to find the biggest integer \( l \) satisfying \( l < r \) for which \( P(I_k|I_{k-1}, \ldots, I_{k-l}) \) can be well estimated for all \( I_k \). Likewise, if subsequences matching \( I_{k-1}, \ldots, I_{k-r} \) are abundant in the training data, then it may be desirable to find the largest integer \( l > r \) for which \( P(I_k|I_{k-1}, \ldots, I_{k-l}) \) can be well estimated for all \( I_k \). The key insight of VLMMs is to realize that \( P(I) \) can be expressed as

\[
P(I) = P(I_1) \prod_{k=2}^{L(I)} P(I_k|I_{k-1}, \ldots, I_1),
\]

(3.3)

and that \( P(I_k|I_{k-1}, \ldots, I_1) \) can be approximated by \( P(I_k|I_{k-1}, \ldots, I_{k-l}) \), the transition probability where the value of \( l \) will vary among possible subsequences that immediately precede \( I_k \). The value of \( l \) should be selected by balancing the advantages of improved model reality that could accompany large values with the disadvantages of uncertainty in transition probability estimates that can also accompany large values. Achieving an optimal balance between these advantages and disadvantages is a somewhat subjective endeavor but a sensible approach and a nice software implementation for training VLMMs from protein sequence data is available (Bejerano and Yona 2001; Bejerano 2004). Our goal here is not to provide alternative techniques for training VLMMs from protein sequence data. Instead, we are motivated by the potentially realistic descriptions of protein sequences by VLMMs and focus on parameterizing evolutionary models of protein-coding DNA so that the evolutionary models yield a stationary distribution of protein sequences which matches a desired VLMM.

Neutral model for codon substitution

The F81 model of nucleotide substitution (Felsenstein 1981) has sequence positions evolve independently and identically. The substitution rate at a position to a nucleotide type \( h \) (\( h \in \{A, C, G, T\} \)) is assumed to be proportional to \( \pi_h \) where \( 0 \leq \pi_h \leq 1 \) and where \( \pi_A + \pi_C + \pi_G + \pi_T = 1 \). The Hasegawa-Kishino-Yano (HKY) model is a slight generalization of the F81 model that has transition substitutions occurring at a rate proportional to \( \kappa \pi_h \).
and transversion substitutions occurring at a rate proportional to $\pi_h$ (Hasegawa, Kishino, and Yano 1985). The stationary distribution of a DNA sequence $i$ of length $L$ for both the F81 and HKY models is

$$P_{\text{HKY}}(i|\pi, \kappa) = P_{\text{F81}}(i|\pi) = \prod_{k=1}^{L} \pi_{i_k},$$

(3.4)

where $i_k$ represents the nucleotide type at position $k$ of sequence $i$ and where $\pi$ collectively represents the parameters $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$.

If the HKY model is coupled to the assumption that all mutations are selectively neutral, then the mutation rate must be proportional to $\kappa \pi_h$ for transitions and to $\pi_h$ for transversions. This perspective of the HKY model can easily be converted to a model for protein-coding DNA evolution where all point mutations that introduce stop codons are lethal and where all other point mutations are selectively neutral. This very simple model of codon substitution is similar to those explored by Muse and Gaut (1994) and Goldman and Yang (1994), except it does not differentiate between synonymous and nonsynonymous substitutions. The stationary probability $P_0(i|\pi, \kappa)$ of a protein-coding DNA sequence $i$ with $L(i)$ codons (and $3L(i)$ nucleotides) for neutral model is

$$P_0(i|\pi, \kappa) = (1/Z)^{L(i)} \prod_{k=1}^{L(i)} \pi_{i_{k1}} \pi_{i_{k2}} \pi_{i_{k3}},$$

(3.5)

where $i_{k1}$, $i_{k2}$, and $i_{k3}$ respectively refer to the nucleotide types at the first, second, and third position of the $k^{th}$ codon. The $(1/Z)^{L(i)}$ term accounts for the fact that sequences containing stop codons will not arise with this evolutionary model. For the universal genetic code, the stop codons are $\text{TAA}$, $\text{TAG}$, and $\text{TGA}$, and $Z = 1 - \pi_T \pi_A \pi_A - \pi_T \pi_A \pi_G - \pi_T \pi_G \pi_A$.

At the protein level, the stationary probability $P_0(I|\pi, \kappa)$ of amino acid sequence $I$ for this neutral model will be the sum of $P_0(i|\pi, \kappa)$ over all DNA sequences $i$ that yield $I$ when translated. Throughout this chapter, lowercase letters denote the nucleotide level and uppercase letters denote the corresponding amino acid information. When necessary, the notation $T(i)$ is employed to represent the amino acids that result from translating the DNA sequence $i$. Henceforth, we will use $i_k$ to refer to the $k^{th}$ codon of DNA sequence $i$. 

66
rather than to the $k^{th}$ position of $i$. With this notation, we have

$$ P_0(I|\pi, \kappa) = P_0(I|\pi) = \sum_{i:T(i)=I} P_0(i|\pi, \kappa) = (1/Z)^L(I) \prod_{k=1}^{L(I)} \left( \sum_{i_k:T(i_k)=I_k} \pi_{i_{k1}} \pi_{i_{k2}} \pi_{i_{k3}} \right) $$

(3.6)

The VLMM rate matrix

We (Robinson et al. 2003; Yu and Thorne 2006) have recently been working with models of DNA sequence change that have evolutionary rates

$$ R_{i,j} = \begin{cases} u\pi_h e^{(E(i)-E(j))f} & \text{transversion} \\ u\pi_h \kappa e^{(E(i)-E(j))f} & \text{transition} \end{cases} $$

(3.7)

when sequences $i$ and $j$ differ at exactly one site and when sequence $j$ has nucleotide type $h$ at this site. Rates of change to sequences that differ at more than one site are set to 0. The values of $E(i)$ and $E(j)$ can represent scores of phenotypes encoded by $i$ and $j$ whereas the parameter $f$ can convert the phenotypic effects induced by a change from $i$ to $j$ into an effect on evolutionary rate (Robinson et al. 2003; Rodrigue et al. 2005; Yu and Thorne 2006). These models are time–reversible and yield a stationary distribution for a protein–coding DNA sequence $i$ that is

$$ P(i|\pi, \kappa) = P(i|\pi) = \frac{e^{-2fE(i)} \prod_{k=1}^{L(i)} \pi_{i_{k1}} \pi_{i_{k2}} \pi_{i_{k3}}}{\sum_c e^{-2fE(c)} \prod_{k=1}^{L(i)} \pi_{c_{k1}} \pi_{c_{k2}} \pi_{c_{k3}}} $$

(3.8)

The sum in the denominator is over all sequences $c$ with the same length as $i$. We note that the stationary distribution for the neutral case of equation (3.5) results when $fE(i)$ is 0 for all DNA sequences $c$.

More importantly, the stationary distribution of equation (3.8) can match any desired VLMM for amino acid sequence organization. To represent the probability of amino acid sequence $I$ for the VLMM, we will write $P_{\text{VLMM}}(I)$. The stationary distribution of the
evolutionary model can be matched to the VLMM when \( f_E(i) \) is set to

\[
f_E(i) = -\frac{1}{2} \log \frac{P_{VLMM}(J)}{P_0(I|\pi, \kappa)}.
\]

We again assume \( i \) and \( j \) differ solely at one position that has type \( h \) in sequence \( j \). We let \( I_H \) (\( J_H \)) be the amino acid specified at the codon in \( i \) (\( j \)) with the single position where \( i \) and \( j \) differ. Also, we use \( v \) and \( w \) to represent nucleotide triplets and \( T(v) \) and \( T(w) \) to be the amino acids specified by \( v \) and \( w \). By combining the rates of equation (3.7) with equation (3.9), we can write

\[
R_{i,j} = \begin{cases} 
  \frac{u\pi_h}{P_{VLMM}(I)/\left( \sum_{w:T(w)\neq I_H} \pi_{w1}\pi_{w2}\pi_{w3} \right)} & \text{synonymous transversion} \\
  \frac{u\pi_h,\kappa}{P_{VLMM}(J)/\left( \sum_{v:T(v)\neq J_H} \pi_{v1}\pi_{v2}\pi_{v3} \right)} & \text{synonymous transition} \\
  \frac{u\pi_h}{P_{VLMM}(J)/\left( \sum_{v:T(v)\neq J_H} \pi_{v1}\pi_{v2}\pi_{v3} \right)} & \text{nonsynonymous transversion} \\
  \frac{u\pi_h,\kappa}{P_{VLMM}(I)/\left( \sum_{w:T(w)\neq I_H} \pi_{w1}\pi_{w2}\pi_{w3} \right)} & \text{nonsynonymous transition.}
\end{cases}
\]

Let the relative fitness of allele (sequence) \( i \) be \( w_i = 1 \) and the relative fitness of allele (sequence) \( j \) be \( w_j = 1 + s \). We assume multiplicative fitnesses so that the fitness of a genotype is the product of the fitnesses of the alleles that constitute it. Following Halpern and Bruno’s pioneering work (1998) and the recent work of others (Berg, Willmann, and Lässig 2004; Knudsen and Miyamoto 2005; Sella and Hirsh 2005), we recently showed that an approximation of \( 2N_e(w_j - w_i) = 2N_es \) is \( f \times (E(i) - E(j)) \) where \( N_e \) is the effective population size (Thorne et al. 2007). The approximation is crude and requires a host of population genetic assumptions (e.g., constant effective population size) but it has the virtue of connecting interspecific evolution with population genetic theory.

Because equation (3.9) specifies the values of \( f_E(i) \) for each protein-coding DNA
sequence $i$, the values of $f \times (E(i) - E(j))$ implied by a VLMM can be calculated for any sequences $i$ and $j$. This allows us to investigate how much natural selection is inherent in a VLMM trained from a data set of interest.

**VLMM example**

As an example, we trained a VLMM on protein sequence data from the human genome. We did this by downloading all of the annotated human protein sequences from NCBI human genome build 36.1 (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). We used the protein sequences to train a variable length Markov model (VLMM) with software that incorporates probabilistic suffix trees (Bejerano and Yona 2001; see also Bejerano 2004). Using the notation employed by this software (Bejerano and Yona 2001; Bejerano 2004), we elected to train with settings of $L = 9$, $P_{min} = 0.00005$, $r = 1.5$, $\gamma_{min} = 0.0000001$, and $\alpha = 0$.

We also needed estimates of the nucleotide frequency parameters $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$ that are derived from human genome data. If DNA experiences solely neutral evolution according to the HKY model, then stationary distributions of sequences obey equation (3.4) and nucleotide frequency parameters can be estimated by the proportions of nucleotide types found in the DNA. The vast majority of the human genome consists of DNA with no known biological function (International Human Genome Sequencing Consortium 2001) and, for the sake of this analysis, we make the assumption that nucleotide frequencies in the human genome can produce reasonable estimates of the nucleotide frequency parameters. About 41% of the genome consists of GC base-pairs and the remaining 59% consists of AT base-pairs (International Human Genome Sequencing Consortium 2001). We therefore set the nucleotide frequency parameters to be: $\pi_A \doteq 0.59 / 2 = 0.295$, $\pi_C \doteq 0.41 / 2 = 0.205$, $\pi_G \doteq 0.41 / 2 = 0.205$, and $\pi_T \doteq 0.59 / 2 = 0.295$. These crude treatments of nucleotide frequency parameters could certainly be improved by allowing values to vary among genes so as to incorporate regional and strand differences in mutation patterns, but we do not pursue these improvements here.

By combining the VLMM trained from human data with these nucleotide frequency estimates, we can use

\[ 2N_e s \doteq f(E(i) - E(j)) \] (3.10)

to estimate the value of $2N_e s$ (Thorne et al. 2007) for a nonsynonymous change from
Figure 3.1: Distribution of $2N_e s$ estimates among possible nonsynonymous changes in human genome using variable length Markov model.
Figure 3.2: Distribution of the mean $2N_e$s estimate per gene among human genes using variable length Markov model.
sequence $i$ to $j$ for one of the protein-coding genes in the human genome. From the NCBI human genome build 36.1, we downloaded 25,925 human messenger RNA sequences. For each possible nonsynonymous change to each of these sequences, we estimated $2N_e$ (fig. 3.1). The mean and standard deviation of these $2N_e$ estimates are -0.211 and 0.796. Of the total of 89,486,730 possible nonsynonymous mutations, 33,317,376 yielded a positive estimate of $2N_e$. We can also find the average $2N_e$ estimate for each protein-coding gene in the human genome and investigate how these vary (fig. 3.2). The mean and standard deviation of these average $2N_e$ values are -0.148 and 0.102, and none of the averages were positive.

**Profile Hidden Markov Models**

**Neutral model for insertions and deletions**

Explicit models of sequence change that account for insertion and deletion events can be challenging to employ when making evolutionary inferences, but some progress has been made (Fleissner, Metzler, and von Haeseler 2005; Lunter et al. 2005; Redelings and Suchard 2005; 2007). We first describe a model with codon substitutions and insertions and deletions that is intended to represent a neutral process of sequence change, and we then describe how the neutral process can be modified to yield a stationary distribution of protein sequences that matches the probability distribution specified by a profile hidden Markov model. The interpretation will be that departures from the neutral process can be attributed to natural selection.

To represent the neutral sequence changes due to insertions and deletions, we consider a modification of the TKF92 insertion-deletion model (Thorne, Kishino, and Felsenstein 1992). This model was not originally framed at the codon level, but we present it as a codon-level model here. We make the convenient but restrictive assumption that insertions and deletions can only insert or delete entire codons and that insertions can only occur between codon boundaries. This is done to prevent the occurrence of mutations that introduce stop codons or frameshift mutations. We suspect that more general and realistic insertion and deletion processes could also be considered but we reserve this possibility for future research.
The TKF92 model has the insertion and deletion processes operate independently of the substitution processes. Conditional upon the length of a newly inserted subsequence, the TKF92 model assumes that the inserted residues are sampled from the stationary distribution of the substitution process (see eq. 3.5). For a subsequence $q$ with length $L(q)$ codons, this means that the probability of $q$ given $L(q)$ would be

$$P_0(q|\pi, L(q)) = \frac{1}{Z} \prod_{k=1}^{L(q)} \pi_{q_1} \pi_{q_2} \pi_{q_3}.$$  \hspace{1cm} (3.11)

The TKF92 model describes the birth and death of entities termed “links” and it considers a type of link termed “immortal” and another termed “normal.” Each normal link is associated with one or more consecutive codons. The normal link and its associated codons are referred to as a fragment. The death rate per normal link is $\mu$. When a normal link dies, the entire fragment with which it is associated is deleted from the sequence. At the extreme 5’ end of each DNA sequence is the immortal link. The immortal link is not associated with any codons and is not subject to death. Both the immortal link and the normal links can become parents and they each experience births at rate $\lambda$. The newborn link and its associated fragment are assumed to be inserted directly to the right (i.e., 3’ end) of the parental fragment.

With this model, the stationary distribution of the number of links in a sequence is geometric (Thorne, Kishino, and Felsenstein 1991) and the stationary distribution for the number of codons per sequence depends on the probability distribution of the number of codons per fragment. It is computationally convenient to have the number of codons per fragment be geometrically distributed so that the probability a fragment has $k$ codons is $(1 - r)r^{k-1}$ where $k$ is an integer that is greater than or equal to 1. The resulting stationary distribution for the number of codons in a sequence $i$ is

$$P_0(L(i)|\lambda, \mu) = P_0(L(I)|\lambda, \mu) = \begin{cases} 
  1 - \frac{\lambda}{\mu} & L(i) = 0 \\
  (1 - \frac{\lambda}{\mu}) \frac{\lambda}{\mu} (1 - r) & L(i) = 1 \\
  (1 - \frac{\lambda}{\mu}) \frac{\lambda}{\mu} (1 - r) (\frac{\lambda}{\mu} (1 - r) + r)^{L(i)-1} & L(i) > 1 
\end{cases}$$  \hspace{1cm} (3.12)
A major flaw of the TKF92 model is that fragment boundaries are not allowed to change over time. If two consecutive codons are inserted together, then neither codon can later be deleted from the sequence unless both are deleted. Here, we consider a variant of the TKF92 models that rectifies the flaw of unchanging fragment boundaries. Our variant differs from the TKF92 model because it does not consider fixed fragment boundaries. A more general but slightly different improvement upon the TKF92 model has been outlined by Miklős, Lunter, and Holmes (2004).

Let $D_0(i, c, q)$ be the rate at which sequence $i$ experiences a deletion event that begins at the $c^{th}$ codon of $i$ and that removes from $i$ the subsequence $q$ which has a total of $L(q)$ codons (i.e., the codons in positions $c, c+1, \ldots, c+L(q)-1$ of $i$ are removed due to the deletion). Obviously, $D_0(i, c, q) = 0$ if $c + L(q) - 1 > L(i)$ or if the subsequence beginning at codon $c$ and having length $L(q)$ codons does not actually match subsequence $q$. For the TKF92 model, the rate of a deletion event specified by $D_0(i, c, q)$ would be 0 unless there is a fragment that begins at codon $c$ and ends at codon $c + L(q) - 1$. In the event where fragment boundaries permit a particular deletion, the TKF92 model has the rate of that deletion being $\mu$.

Let $I_0(i, c, q)$ be the rate at which $i$ experiences an insertion event between the codon at position $c$ and position $c + 1$ (if $0 \leq c < L(i)$) or, in the case of $c = L(i)$, that begins 3’ of the codon at position $L(i)$. For the TKF92 model, the rate of the insertion event specified by $I_0(i, c, q)$ is 0 unless the fragment boundaries of $i$ are such that an insertion is possible. If an insertion is possible, the TKF92 model has the insertion rate be $\lambda(1 - r)\mu^{L(q)-1}P_0(q|L(q), \pi)$.

With the TKF92 model, there can be many possible ways of assigning fragment boundaries to a sequence of length $L(i)$, and the probability of length $L(i)$ in equation (3.12) is the sum over all these possible fragmentations of the probability of each fragmentation. Because of the multiple ways sequences can be fragmented, different sequences of the same length that all evolve according to the TKF92 model may experience different overall insertion and deletion rates and also different insertion and deletion rates at individual sequence locations. Our variant is much the same as the TKF92 model except that each sequence of a particular length experiences the same insertion and deletion rates as each other. Specifically, the insertion rates $I_0(i, c, q)$ and deletion rates $D_0(i, c, q)$ for this variant model are obtained by averaging the TKF92 model rates over all possible fragmentations of
a sequence of length $L(i)$. We note that the probability of observing a fragment boundary between any two codons is $\frac{\lambda(1-r)}{\mu(1-r)+r}$. Some algebra then shows that our rates are

$$D_0(i, c, q) = \begin{cases} 
\frac{\lambda}{\mu} \frac{L(i) - 1}{L(i) + 1} & c = 1 \quad \text{and} \quad L(i) = L(q) \\
\frac{\lambda}{\mu} \frac{(1-r) L(i) - 1}{(1-r) + r} & (c = 1 \text{ and } L(q) < L(i)) \\
\frac{\lambda^2}{2(\mu(1-r)+r)^2} & 1 < c < L(i) - L(q) + 1.
\end{cases}$$

(3.14)

These rates can be shown to yield a time-reversible insertion-deletion model with a stationary distribution of sequence lengths that is identical to equation (3.12). A nice feature of the TKF92 model is the availability of explicit transition probabilities for transforming one sequence into another. Our modification to it does not share these transition probabilities, but our purposes here do not require explicit transition probabilities.

**The profile HMM rate matrix**

If codon substitutions and insertion and deletion events are exclusively neutral, then the stationary probabilities of $i$ and $I$ would be

$$P_0(i|\pi, \lambda, \mu) = P_0(i|\pi, L(i))P_0(L(i)|\lambda, \mu)$$

(3.15)
and

\[ P_0(I|\pi, \lambda, \mu) = P_0(I|\pi, L(I)) P_0(L(I)|\lambda, \mu) . \]  

(3.16)

However, we want to consider departures due to natural selection from these neutral probabilities. For a change from sequence \( i \) to sequence \( j \), the rate \( R_{ij} \) with natural selection will again be assumed to be \( e^{(E(i)-E(j))f} \) multiplied by the neutral rate. This means that the rates of point mutations are given by equation (3.7) and the insertion and deletion rates are

\[ \mathcal{I}(i, c, q) = e^{(E(i)-E(j))f} \mathcal{I}_0(i, c, q) \]  

(3.17)

and

\[ \mathcal{D}(i, c, q) = e^{(E(i)-E(j))f} \mathcal{D}_0(i, c, q) . \]  

(3.18)

With insertions and deletions, sequence lengths are not fixed and equation (3.8) needs to be modified,

\[ P(i|\pi, \lambda, \mu) = \frac{e^{-2fE(i)}P_0(i|\pi, \lambda, \mu)}{\sum_c e^{-2fE(c)}P_0(i|\pi, \lambda, \mu)} , \]  

(3.19)

where the sum is now over all possible sequences for all possible lengths.

We would like the rates of equation (3.19) to yield a stationary distribution for which the stationary probability of an amino acid sequence \( I \),

\[ P(I|\pi, \lambda, \mu) = \sum_{i:T(i)=I} P(i|\pi, \lambda, \mu) , \]  

(3.20)

is equal to the probability of \( I \) according to a profile HMM. We will write this latter probability as \( P_{\text{HMM}}(I) \). To do this, we set

\[ fE(i) = \frac{1}{2} \log \frac{P_{\text{HMM}}(I)}{P_0(I|\pi, \lambda, \mu)} \]  

\[ = \frac{1}{2} \log \frac{P_{\text{HMM}}(I)}{P_0(L(I)|\lambda, \mu)P_0(I|\pi, L(I))} . \]  

(3.21)

Now, we employ reasoning similar to that of Thorne et al. (2007) and to that of the VLMM case in this chapter. We again assume the relative fitness of \( i \) is 1 and the relative fitness
of \( j \) is \( 1 + s \). This yields

\[
2N_e s = f(E(i) - E(j)) \\
= \frac{1}{2} \log \frac{P_{\text{HMM}}(J)P_0(L(I)|\lambda, \mu)P_0(I|\pi, L(I))}{P_{\text{HMM}}(I)P_0(L(J)|\lambda, \mu)P_0(J|\pi, L(J))}.
\]

(3.22)

In the above equation, the terms \( P_0(L(I)|\lambda, \mu) \) and \( P_0(L(J)|\lambda, \mu) \) depend on \( \lambda \) and \( \mu \). Rather than estimate \( \lambda \) and \( \mu \), we reason that protein sequences tend to be relatively long. This means that the ratio of \( P_0(L(I)|\lambda, \mu) \) and \( P_0(L(J)|\lambda, \mu) \) is close to 1 if \( L(I) \) is not too small and if the lengths \( L(I) \) and \( L(J) \) are not too different. The approximation to \( 2N_e s \) then becomes

\[
2N_e s \approx \frac{1}{2} \log \frac{P_{\text{HMM}}(J)P_0(I|\pi, L(I))}{P_{\text{HMM}}(I)P_0(J|\pi, L(J))}.
\]

(3.23)

Profile HMM example

As an example, we used a profile HMM that was trained from members of the envelope surface glycoprotein protein family. The profile HMM was obtained from the Pfam database (Sonnhammer, Eddy, and Durbin 1997). We obtained a human immunodeficiency virus 1 protein-coding DNA sequence that belongs to this family (accession number is NC_001802.1: Martoglio et al. 1997). We considered all 3,233 possible nonsynonymous point mutations to this sequence and we estimated \( 2N_e s \) for each of them by applying equation (3.23).

To apply this equation, we need estimates for the nucleotide frequency parameters \( \pi \). We assumed that the neutral point mutation process in HIV would yield the same nucleotide frequencies as in humans (i.e., \( \pi_A = \pi_T = 0.295, \pi_C = \pi_G = 0.205 \)) so that we could adopt the nucleotide frequency estimates used for the VLMM example. This treatment of the \( \pi \) parameters is simple, but it is admittedly suspect. In the future, we plan more careful treatments of the \( \pi \) parameters.

A histogram of \( 2N_e s \) estimates for the possible nonsynonymous point mutations is depicted in figure 3.3. The mean and standard deviation of these \( 2N_e s \) values for the point mutations were -1.239 and 0.916. Of the total of 3,232 possible nonsynonymous mutations, only 228 yielded a positive estimate of \( 2N_e s \). We then considered the 483 possible single-codon deletions to this coding sequence (fig. 3.4). The mean and standard deviation of these \( 2N_e s \) values for the deletions were -3.246 and 1.587, and 30 of the estimates were positive.
Finally, we considered all possible single–codon insertions to the coding sequence (fig. 3.5). There are a total of 29,524 possible single–codon insertions and they yield a mean $2N_e s$ estimate of -2.289 and a standard deviation of 1.089. There were 1,786 of the 29,524 total possible insertions with a positive $2N_e s$ estimate. Figure 3.5 is a histogram of these $2N_e s$ estimates for insertions.

Discussion

Computational biologists have devised diverse useful probabilistic descriptions for the organization of molecular sequence data. VLMMs and profile HMMs are prominent examples of such probabilistic descriptions. Natural selection and mutation are forces that shape molecular sequence data but the connections between these forces and the computational biology techniques for analyzing sequence data are often weak. Our goal here has been to help lay the groundwork for establishing such connections. Reconciliation of VLMM and profile HMM descriptions of sequence data with their population genetic implications seems to us to be a desirable endeavor.

A major shortcoming of most models of sequence evolution is that they can inappropriately assign much probability to sequences that are selectively deleterious. Computational biologists have had substantial motivation to create probabilistic descriptions of sequence data that assign high probability to the sequences that seem most biologically plausible. These probabilistic descriptions have the potential to assist with model–based evolutionary inference.

In this study, we employ the VLMM and profile HMM models mainly to obtain $N_e s$ estimates. In the future, inference techniques that use these or similar models could be developed for analyzing sets of homologous DNA sequences. Inference with the VLMM model could be performed via relatively straightforward modifications of previously published statistical techniques (e.g., see Robinson et al. 2003; Rodrigue et al. 2005). Inference with the profile HMM model would be more difficult due to the possibility of insertion and deletion events and this is a challenge that we are currently considering.
Figure 3.3: Estimates of $2N_e s$ for possible nonsynonymous substitutions to gene for envelope surface glycoprotein gp120 using profile hidden Markov model.
Figure 3.4: Estimates of $2N_e s$ for possible single codon deletions to gene for envelope surface glycoprotein gp120 using profile hidden Markov model.
Figure 3.5: Estimates of $2N_e s$ for possible single codon insertions to gene for envelope surface glycoprotein gp120 using profile hidden Markov model.
Literature Cited


Chapter 4

Probability of single nucleotide polymorphism ancestry via an evolutionary model that incorporates protein tertiary structure

Sang Chul Choi, Eric A. Stone, Hirohisa Kishino, and Jeffrey L. Thorne
Abstract

We consider the inference of which of two alleles is ancestral when the two alleles exhibit a single nonsynonymous difference and when natural selection acts via protein tertiary structure. Whereas the probability that an allele is ancestral under neutrality is equal to the allele frequency, under selection it depends on allele frequency and the magnitude and direction of selection pressure. Although allele frequencies can be well estimated from intraspecific data, small fitness differences have a large evolutionary impact but can be difficult to estimate with only intraspecific data. Methods for predicting aspects of phenotype from genotype can supplement intraspecific sequence data. Recently developed statistical techniques can assess effects of phenotypes, such as protein tertiary structure on molecular evolution. While these techniques were initially designed for comparing protein-coding genes from different species, the resulting interspecific inferences can be assigned population genetic interpretations to assess the effect of selection pressure, and we use them here along with intraspecific allele frequency data to estimate the probability that an allele is ancestral. We focus on nonsynonymous single nucleotide polymorphisms in proteins with known tertiary structures. We detect a weak signal that protein sequences from common alleles better conform to protein tertiary structure than sequences from rare alleles. This supports the expectation that ancestral alleles are selectively more advantageous than derived alleles.

Introduction

When species are distantly related, inference of the ancestral DNA sequences that relate them is conventionally done without regard to intraspecific polymorphism (e.g., Felsenstein 2004; Yang, Kumar, and Nei 1995). When reconstructing ancestral allele types from intraspecific data, information about the frequencies of extant allele types becomes highly relevant. When variation among alleles is selectively neutral, the probability an allele type is the ancestral state is equal to its frequency in the population (Watterson and Guess 1977). Watterson and Guess (1977) noted that if mutation rates are not negligible, the most frequent allele is less likely to be oldest. Likewise, if genetic variation is not neutral, the probability that an allele is ancestral is no longer the allele frequency (e.g., Slade 2000; Fearnhead 2002; Taylor 2007). New mutations are more likely to be deleterious than advantageous and new deleterious mutations are unlikely to reach high or even moderate
allele frequencies before being lost. Accordingly, the probability of deleterious alleles being ancestral is less than would be expected under neutrality. Donnelly and Kurtz (1999) made the sensible conjecture that the expected fitness of a common ancestral allele will exceed that of a randomly selected allele from the population. Slade (2000) and Fearnhead (2002) corroborated the conjecture with simulation and analytical treatments, respectively, using the ancestral selection graph technique of Krone and Neuhauser (1997). Because non-synonymous SNPs (nsSNPs) are expected to be more deleterious than synonymous SNPs (sSNPs), these results coincide with empirical findings such as the tendency of nsSNPs to have lower allele frequencies than sSNPs (Cargill et al. 1999).

SNPs are the largest resource of genetic variation (Sherry et al. 2001) and there are growing protein structure databases (Berman et al. 2000). The concurrent outgrowth of genetic variation and protein structural data has motivated the integration of coding SNPs into protein structure (Chasman and Adams 2001; Sunyaev, Lathe, and Bork 2001a; Sunyaev et al. 2001b; Wang and Moul 2001; Ramensky et al. 2002; Hughes et al. 2003; Krishnan and Westhead 2003; Yue, Li, and Moul 2005; Yue and Moul 2006). These techniques attempt to associate genetic variation of SNPs with phenotypic consequences of abnormal protein structural changes and diseases in order to predict which nsSNPs are likely to affect protein function. Although these studies connect genotypic changes with phenotype, they typically do not provide a population genetic interpretation about differences in relative fitnesses among alleles. Over the last half century, phylogenetic inference with interspecific sequence data has experienced steady methodological advance. At the cornerstone of the progress are models of molecular evolution. Recently, protein tertiary structure information has become much more abundant. This benefits our knowledge about protein-coding evolution (e.g., Marsden et al. 2006). Despite the surge of molecular phenotype information (e.g., protein tertiary structure) and the advance of genotype-phenotype mapping techniques (e.g., protein structure prediction), many models of molecular evolution treat sequence changes without considering their phenotypic consequences. Therefore, many models treat natural selection superficially.

Efforts to better handle natural selection in models of interspecific evolution emerged with Halpern and Bruno (1998), who related rates in interspecific models of sequence change to fixation probabilities of population genetic diffusion theory (Kimura 1962). More recently, Nielsen and Yang (2003) studied codon models and estimated the distribution of the product of effective population size ($N$) and selection coefficient ($s$), a product referred to as...
the scaled selection coefficient. The codon model used by Nielsen and Yang (2003) has the independent-site assumption that each codon evolves independently of each other, and there is no genotype-phenotype mapping built into the model. Robinson et al. (2003) relaxed the independent-site assumption of codon models by explicitly incorporating molecular phenotype (e.g., protein tertiary structure) into a codon model. Hereafter, we refer to Robinson et al. (2003) model as the dependent-sites model. Although the dependent-sites model was designed to study interspecific sequence data, Thorne et al. (2007) assigned population genetic interpretations to the parameters relating to phenotype in the model. These interpretations exploit techniques for predicting phenotype from genotype and can approximate the distribution of scaled selection coefficients among possible nonsynonymous changes. In so doing, a population genetic interpretation of the dependent-sites model yields estimates of the magnitude and direction of selection pressure.

As noted above, there have been advances in adding natural selection to determining the probability that an allele is ancestral under selection (Fearnhead 2002; Taylor 2007). There has also been a surge of high-quality data pertaining to human genetic variation (Sherry et al. 2001; Livingston et al. 2004) and protein three-dimensional structure (Berman et al. 2000). Despite these advances, there is a paucity of studies that combine molecular phenotype and genetic variation to pursue the probability that an allele is ancestral. Here, we attempt to quantify the probability that an allele is ancestral by considering natural selection due to protein structure.

We first analyze allele frequency information of SNPs in the human genome. With a genome-wide analysis, we confirm that the frequency of ancestral alleles under putative purifying selection is less than that for alleles under neutrality. In subsequent sections, we derive the probability that an allele is ancestral given allele frequency and given a measure of compatibility between protein sequence and tertiary structure (see Jones, Taylor, and Thornton 1992). Second, we infer the probability of ancestry for 142 nsSNPs generated by the National Institute of Environmental Health Science (Livingston et al. 2004). We find a weak tendency for more-frequent alleles to fit protein structure better than less-frequent alleles and a weak tendency for ancestral alleles to be more selectively advantageous than derived alleles with respect to protein tertiary structure. However, we do not find large differences of ancestral SNP allele probabilities when comparing those calculated with the selection model to those calculated with the neutral model.
Materials and Methods

Data preparation

For the genome-wide comparison of the frequency of ancestral alleles between synonymous and non-synonymous SNPs, we obtained chimpanzee DNA that is orthologous to human SNP sites. We did this via the University of California Santa Cruz (UCSC) Genome Browser table called “snp126orthoPanTro2Rhemac2” (Thomas et al. 2007). Via the web interactive application of “Galaxy” (Giardine et al. 2005), we used the table that was available as of April 13th, 2007. SNP frequency data of an African population from the International HapMap Project were downloaded from the dbSNP database with functional class annotations (Sherry et al. 2001).

To estimate the probability that an allele is ancestral, we faced two challenges in preparing data for the analysis of all nsSNPs in the human genome. First, finding all DNA sequences and protein structures for all nsSNPs is not a simple task. Second, SNPs discovered by the International HapMap Project suffer from ascertainment bias in SNP genotyping processes (Nielsen and Signorovitch 2003). Rather than directly sequencing all individual chromosomes, SNP discovery usually proceeds by locating SNP candidates in a small set of individuals, and then continues by typing the located SNPs from a larger sample of chromosomes. The National Institute of Environmental Health Science (NIEHS) SNP project produces quality SNPs in genes via direct sequencing (Livingston et al. 2004) and this has lower ascertainment bias (Clark et al. 2005). We therefore invested our efforts in the NIEHS SNP data.

We obtained SNP data for 612 genes from the NIEHS SNPs database (Livingston et al. 2004). We searched the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCBS-PDB) database (Berman et al. 2000) for protein tertiary structures with amino acid sequences that are similar to those of the genes in the NIEHS data set. We did this by using BLAST (Altschul et al. 1997) without internal gaps. We retained the protein structures that yielded the top match, and kept only DNA sequences that were with at least 200 nucleotides long. The resulting data set consists of genes with DNA sequences for both the major and minor SNP alleles along with major and minor allele frequencies. Each gene also includes an accompanying protein tertiary structure. There are a total of 142 non-synonymous SNPs from 74 genes in the data set.
Probability of an allele being ancestral

We briefly outline the dependent-sites model of protein-coding DNA sequence evolution (Robinson et al. 2003) and a population genetic interpretation of it (Thorne et al. 2007). Subsequently, we will consider the probability of an allele being ancestral when applying a population genetic interpretation of the dependent-sites model.

Dependent-sites model of protein-coding DNA sequence evolution

The dependent-sites model explicitly takes protein tertiary structure into account to relax the assumption that each site evolves independently of each other (see Robinson et al. 2003). Assume we have DNA sequences of $i$ and $j$. When $i$ and $j$ differ at more than one site or when $j$ encodes a premature stop codon, we set the rate of change from $i$ to $j$ to 0. For sequences $i$ and $j$ that differ only at a site where $j$ has nucleotide type $h$, the instantaneous rate $R_{ij}$ of change from sequence $i$ to sequence $j$ is given by

$$
R_{ij} = \begin{cases} 
  u\pi_h & \text{for a synonymous transversion} \\
  u\pi_h\kappa & \text{for a synonymous transition} \\
  u\pi_h\omega e_f s \times (E_s(i)-E_s(j)) + f_p \times (E_p(i)-E_p(j)) & \text{for a nonsynonymous transversion} \\
  u\pi_h\kappa\omega e_f s \times (E_s(i)-E_s(j)) + f_p \times (E_p(i)-E_p(j)) & \text{for a nonsynonymous transition.}
\end{cases}
$$

(4.1)

The $\pi$ parameters allow mutations to the four nucleotide types to not be equally likely. Transitions and transversions are allowed to happen at different rates via $\kappa$. We scale the overall rate of change through parameter $u$. The non-synonymous treatment involves two more kinds of parameters: $\omega$ to capture contributions to non-synonymous rates that are external to the protein under analysis, and $f_s$ and $f_p$ to capture the effects of tertiary structure. The $f_s$ parameter incorporates solvent accessibility effects while the $f_p$ parameter reflects pairwise amino acid interactions. We adopt the sequence-structure compatibility system of Jones et al. (1992). We respectively use $E_s(i)$ and $E_p(i)$ to represent solvent
accessibility compatibility and pairwise interaction compatibility. Lower scores of $E_s(i)$ and $E_p(i)$ are better in terms of sequence-structure compatibility. A biologically plausible scenario is the case where $f_s$ and $f_p$ both are positive. The model reduces to a conventional codon model that ignores protein structure (e.g., Goldman and Yang 1994; Muse and Gaut 1994) if $f_s$ and $f_p$ are set to 0. Sequence $i$ has stationary probability

$$p(i|f_s, f_p, \pi) = \frac{e^{-2f_s \times E_s(i) - 2f_p \times E_p(i)} \prod_{m=1}^{M} \pi_{i_m}}{\sum_k e^{-2f_s \times E_s(k) - 2f_p \times E_p(k)} \prod_{n=1}^{M} \pi_{k_n}} \quad (4.2)$$

where $i_m (k_n)$ is the nucleotide type at position $m (n)$ in DNA sequence $i (k)$ and the sum in the denominator is over all possible sequences $k$ of length $M$ that lack a premature stop codon.

### Population genetic interpretation of phenotype impact parameters

Statistical procedures for inferring phylogenies and characterizing interspecific evolution tend not to be explicitly connected to population genetics despite the fact that evolution happens at the population level. Thorne et al. (2007) placed a population genetic interpretation upon phenotype impact parameters of the dependent-sites model (e.g., $f_s$ and $f_p$) by extending the pioneering work of Halpern and Bruno (1998). The interpretation assumes that there is a diploid organism with constant effective population size $N$ and with a multiplicative relative fitness scheme (genotype fitness: $w_{ii} = 1$, $w_{ij} = 1 + s$, and $w_{jj} = (1 + s)^2$ for two allele types $i$ and $j$) where $s$ is the selection coefficient. Parameters of the model of protein-coding evolution can then be assigned a population genetic interpretation (Thorne et al. 2007) as follows:

$$2Ns \doteq f_s \times (E_s(i) - E_s(j)) + f_p \times (E_p(i) - E_p(j)) \quad (4.3)$$

where sequence $j$ can be generated by a single non-synonymous substitution to sequence $i$ at one site. Another more direct population genetic interpretation is possible through the conversion of the selection coefficient $s$ to $a_s(E_s(i) - E_s(j)) + a_p(E_p(i) - E_p(j))$ where $a_s$ and $a_p$ are new parameters related to solvent accessibility and pairwise interaction aspects of sequence-structure compatibility. Then, the scaled selection coefficient can be also given.
by
\[ 2N_s = 2Na_s \times (E_s(i) - E_s(j)) + 2Na_p \times (E_p(i) - E_p(j)) \] (4.4) (Thorne et al. 2007). In addition to being able to indirectly approximate scaled selection coefficients through equation (4.3), we also can directly estimate \(2N_s\) by estimating \(2Na_s\) and \(2Na_p\). However, when estimates are based on only a single sequence and its stationary distribution as they will be here, equation (4.3) and equation (4.4) yield the same estimate of \(2N_s\) (Thorne et al. 2007).

### Probability of an allele being ancestral under selection

Fearnhead (2002) and Taylor (2007) give rigorous mathematical treatments of the probability that an allele is ancestral with selection. Fearnhead (2002) leverages the ancestral selection graph technique (Krone and Neuhauser 1997) to find the stationary distribution of the common ancestral type of a sample of genes with consideration of a non-neutral population genetics model. Taylor (2007) uses the structured coalescent process (Kaplan, Darden, and Hudson 1988) to characterize the common ancestor process for a Wright-Fisher diffusion, and analytically calculates the stationary distribution of the common ancestral type for a two-allele model. The above two methods of finding the stationary probability of a common ancestor are more general than considered here. We assume that segregating SNPs are the result of a low mutation rate, which means that they should be rare. The low mutation assumption could be especially problematic for SNPs located in mutation hotspots.

We consider a gene for which the protein tertiary structure is available. The gene is assumed to have a biallelic SNP site with known allele frequencies. We designate \(\rho_i\) and \(\rho_j\) where \(\rho_i + \rho_j = 1\) for the allele frequencies of variant \(i\) and the variant \(j\), respectively. We use the same multiplicative relative fitness scheme of three genotypes as in the previous subsection: \(w_{ii} = 1\), \(w_{ij} = 1 + s\), and \(w_{jj} = (1 + s)^2\). Note that allele \(j\) is selectively advantageous compared to allele \(i\) when \(s\) is positive. Either allele \(i\) or \(j\) will be assumed to be ancestral, and no further mutation is assumed to have happened since the formation of the SNP. We use a diffusion theory approach that takes selection and genetic drift but not mutation into account (Kimura 1962).

We consider two mutually exclusive scenarios of allele ancestry. One scenario is
that allele \( i \) is ancestral. If allele \( i \) is ancestral, then we know that allele \( j \) must be the result of a new mutation and that the frequency of \( j \) has reached at least \( \rho_j \) since the origin of allele \( j \). Note that we assume the sample frequency of \( j \) is equal to the population frequency of \( j \). We refer to the possibility that allele \( i \) is ancestral to allele \( j \) as scenario \( I \).

We also define a scenario termed \( J \) that has allele \( j \) being ancestral. If scenario \( J \) is true, we know that allele \( i \) must be the result of a new mutation and that the frequency of \( i \) has reached at least \( \rho_i \) since the origin of allele \( i \). We seek the probability of scenario \( I \) given the assumption that exactly one of scenarios \( I \) and \( J \) is true:

\[
p(I|\text{I or } J) = \frac{p(I)}{p(I) + p(J)}. \tag{4.5}
\]

The probability density of scenario \( I \) can be expressed as the product of 3 factors. The first is the probability that a population is fixed for allele \( i \) immediately before the new mutation of \( i \) to \( j \). The second is the probability density (i.e., the instantaneous mutation rate) of a mutation from \( i \) to \( j \) at some instant. The third is the probability that new mutant allele \( j \) ever reaches the observed frequency \( \rho_j \) given that \( j \) has initial frequency \( \epsilon \). The first probability is given by stationary distribution of sequence \( i \) of the dependent-sites model of equation (4.2). The third probability is a generalization (Ewens 2004) of the fixation probability of Kimura (1962),

\[
\frac{1 - e^{-4Ns\epsilon}}{1 - e^{-4Ns\rho_j}}. \tag{4.6}
\]

Accordingly, the probability of scenario \( I \) is

\[
p(I) = p(i|f_s, f_p, \pi) \times \pi_{j(h)^K} \times \frac{1 - e^{-4Ns\epsilon}}{1 - e^{-4Ns\rho_j}}. \tag{4.7}
\]

Similarly, the probability of scenario \( J \) is

\[
p(J) = p(j|f_s, f_p, \pi) \times \pi_{i(h)^K} \times \frac{1 - e^{-4N(-s)\epsilon}}{1 - e^{-4N(-s)\rho_i}}. \tag{4.8}
\]

Substituting equations (4.7) and (4.8) in equation (4.5), we have the probabilities that we
are seeking,

\[
\Pr(I|I \text{ or } J) = \frac{1 - e^{4Ns \rho_i}}{1 - e^{4Ns}} \tag{4.9}
\]

\[
\Pr(J|I \text{ or } J) = \frac{1 - e^{-4Ns \rho_j}}{1 - e^{-4Ns}} = \frac{1 - e^{4N(-s)\rho_j}}{1 - e^{4N(-s)}}. \tag{4.10}
\]

We obtain these results by assuming that \( \epsilon = 1/2N \) and \( 1 - e^{-2s} \simeq -(1 - e^{2s}) \), and we apply equation (4.3) of the population genetic interpretation. Simply, the probability of a minor allele being ancestral to a major allele given the data and assuming that exactly one of two SNP alleles is ancestral is

\[
\frac{1 - e^{4Ns \rho}}{1 - e^{4Ns}} \tag{4.11}
\]

where

\[
2Ns \approx f_s \times (E_s(\text{i}_{\text{minor}}) - E_s(\text{i}_{\text{major}})) + f_p \times (E_p(\text{i}_{\text{minor}}) - E_p(\text{i}_{\text{major}})) \tag{4.12}
\]

and where \( \rho \) is the minor allele frequency. The quantity \( 2Ns \) is the scaled selection coefficient for a change from the minor allele sequence to the major allele sequence, and we refer to equation (4.12) as minor/major scaled selection coefficient. A positive value of a minor/major scaled selection coefficient means that the major allele has higher molecular fitness than the minor allele.

The probability of scenario \( I \) given that either scenario \( I \) or \( J \) is true (eq. 4.9) is, interestingly, the same as the probability (third term of eq. 4.8) that allele \( i \) eventually gets fixed given its contemporary allele frequency \( \rho_i \) (see also Taylor 2007; for more rigorous mathematical treatment). This result is due to the time-reversibility of the dependent-sites model. Halpern and Bruno (1998) separate rates of substitution into mutation rate and fixation probability (see also Thorne et al. 2007). Let us denote the rate of mutation from \( i \) to \( j \) by \( \mu_{ij} \). We call \( Z_j(\epsilon \rightarrow \rho_j) \) the probability that allele \( j \) with initial frequency \( \epsilon \) ever reaches frequency \( \rho_j \). Staring with the time-reversibility condition of the dependent-sites
model (Robinson et al. 2003),

\[ p(i|f_s, f_p, \pi)R_{ij} = p(j|f_s, f_p, \pi)R_{ji} \]

we come to equation (4.13) that relates the probability densities of scenarios $I$ and $J$ to the fixation probabilities. We can use equations (4.5) and (4.13) to show that the probability of scenario $I$ given the assumption that either $I$ or $J$ is true is the same as the fixation probability of allele $i$ with initial frequency $\rho_i$ as follows:

\[
p(I|I \text{ or } J) = \frac{p(I)}{p(I) + p(J)} = \frac{1}{1 + p(J)/p(I)}
\]

\[
= \frac{1}{1 + Z_j(\rho_j \rightarrow 1)/Z_i(\rho_i \rightarrow 1)}
\]

\[
= \frac{Z_i(\rho_i \rightarrow 1)}{Z_i(\rho_i \rightarrow 1) + Z_j(\rho_j \rightarrow 1)} = \frac{Z_i(\rho_i \rightarrow 1)}{1}.
\]

The denominator of equation (4.14) must be 1 because two events that allele $i$ or $j$ gets fixed are mutually exclusive.

**Results**

**Comparison of frequency of ancestral alleles under selection and neutrality**

Following the Chimpanzee Sequencing and Analysis Consortium (2005), our parsimony inference of ancestral alleles at biallelic SNP sites in the human genome uses the chimpanzee genome as an outgroup to plot the frequency of ancestral alleles of the human SNPs. We use the same technique in order to compare the frequency of ancestral alleles under putative selection (e.g., nsSNPs) with that for alleles under putative neutrality (e.g., sSNPs). We used SNP data from the African population of the HapMap data, and subdivided the SNP
data into intron, synonymous, and non-synonymous SNPs. We ignore a small number of SNPs for which neither human SNP allele matches the outgroup chimp. In figure 4.1, we bin all SNPs with similar minor SNP allele frequencies together, and then plot the minor SNP allele frequency (x-axis) versus the proportion of minor SNP alleles that are ancestral when using the chimpanzee as outgroup (y-axis). Under neutrality, the probability that an allele is ancestral is equal to its allele frequency, and the dashed line of figure 4.1 is this theoretical line with slope 1 and intercept 0. We call the dashed line the “neutral line.”

We confirmed that nsSNPs had relatively smaller frequencies of ancestral minor alleles than sSNPs and intron SNPs. Intron and synonymous SNP plots might be expected to follow the neutral line, if they are thought to be approximately neutral. Lines of intron and synonymous SNPs follow the neutral line of ancestral allele frequency from 0.5 to 0.2, and then begin deviating upward from the neutral line (see also Chimpanzee Sequencing and Analysis Consortium 2005). The deviation from the neutral line is much larger in European and Asian populations (data not shown). The Chimpanzee Sequencing and Analysis Consortium (2005) listed several possibilities for why this deviation from the neutral lines might occur. The possibilities include incorrect ancestral allele assignments, bottlenecks during human evolution, and selective sweeps. The Chimpanzee Sequencing and Analysis Consortium (2005) concluded that bottlenecks during human history might be the main causes of the deviation. Reich et al. (2001) also found a larger linkage-disequilibrium in non-African populations than in the African population, which might be associated with different demographic histories of non-African populations from that of the African population. The large linkage-disequilibrium in the genome of non-African populations might imply that there were many young, emerging alleles after non-African populations had separated from African populations. If there were many recurrent mutations back to ancestral alleles, the empirical frequency of alleles that are ancestral would have been larger than expected under neutrality. Another possible explanation might be attributable to hitchhiking with rare recombination, which would push derived alleles of intermediate frequency toward either very low or high frequency (Fay and Wu 2000).
Figure 4.1: **Frequency of minor alleles being ancestral against minor allele frequency.** Black lines with circles are for intron SNPs, blue lines with diamonds are for synonymous SNPs, and red lines with triangles are for non-synonymous SNPs. The dashed line has slope 1 and intercept 0, and represents the neutral prediction. Error bars extend twice the estimated standard deviation above and below points. The lengths of error bars are estimated via a normal approximation to a binomial distribution (i.e., the standard deviation is \( \sqrt{P_c(1 - P_c)/n_c} \) where \( P_c \) is the frequency of ancestral alleles and \( n_c \) is the number of SNPs with minor allele frequencies grouped together in bin \( c \)). (A) The International HapMap data of African ethnicity and (B) NIEHS data.
Molecular fitness and allele frequency

For 95 of 142 SNPs, the major allele was assessed to be selectively more advantageous than the minor allele. For the remaining 47 SNPs, the minor allele was judged to be selectively more advantageous than major allele (table 4.1, and fig. 4.2). A binomial one-sided test of the alternative hypothesis that the true proportion is greater than 0.5 gave a $P$-value less than 0.001. At the molecular level, the result of more advantageous major alleles we had more SNPs whose major alleles seem to be more advantageous supports, if weakly, the idea that low fitness of alleles would result in low allele frequencies. We may expect that the selective advantage of major alleles is weaker among the SNPs where the minor alleles have reached a high frequency. Rather than assuming a linear relationship between SNP frequency and the minor/major scaled selection coefficient, we employed a nonparametric measure of correlation, Kendall’s rank correlation coefficient. The correlation coefficient between minor allele frequency and minor/major scaled selection coefficient was -0.110 with a $P$-value of 0.044 (one-sided test).

We had a handful of SNPs whose minor alleles were ancestral: the average minor allele frequency was 0.051, and thus it was expected that minor alleles were ancestral in about 7 SNPs out of 142. Among 133 SNPs whose major alleles were ancestral, the major alleles were selectively more advantageous in about 2/3 (92) of them. Among the 9 SNPs whose minor alleles were ancestral, minor alleles were selectively more advantageous in 2/3 (6). Assuming that the molecular phenotype of a protein structure does not change over time, we may measure ancestral phenotypic advantage via the phenotype prediction technique we use. Whether or not major or minor alleles are ancestral, ancestral alleles seem to be more selectively advantageous. The null hypothesis of interest, that allele ancestry and molecular fitness of major/minor alleles are independent, was tested via the Fisher’s Exact test and yielded a $P$-value of 0.036. The weak significance is consistent with the idea that ancestral alleles usually have higher fitness.

Ancestral alleles are more advantageous with respect to molecular fitness

For each SNP, we can measure the proportion of possible non-synonymous changes to the ancestral sequence that are more deleterious than the actual derived allele. Derived alleles with low frequencies might tend to be more deleterious than derived alleles that reach
Table 4.1: Four possible scenarios of allele ancestry and molecular fitness

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Major allele is ancestral</th>
<th>Minor allele is ancestral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major allele is advantageous</td>
<td>92</td>
<td>3</td>
<td>95</td>
</tr>
<tr>
<td>Minor allele is advantageous</td>
<td>41</td>
<td>6</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>133</td>
<td>9</td>
<td>142</td>
</tr>
</tbody>
</table>

The major allele is a SNP allele with frequency is greater than 0.5, and the minor allele has frequency smaller than 0.5.

Figure 4.2: Scatter plot of minor/major scaled selection coefficient against minor allele frequency. A positive value along the y-axis represents a SNP where the major allele is judged to have a higher fitness than the minor allele. Bullets are for SNPs whose major alleles are ancestral to minor alleles, and cross marks are for SNPs whose minor alleles are ancestral to major alleles. The dashed horizontal line is at a y-axis value of 0.
moderate or high frequency. As a result, one might expect a positive correlation between the derived allele frequency and the proportion of possible non-synonymous changes that are more deleterious than the change that yielded the actual derived allele. We can contrast this alternative hypothesis to the null hypothesis that derived allele frequency is independent of the proportion of possible non-synonymous changes that are more deleterious than the derived allele. These proportions were plotted versus derived allele frequency (fig. 4.3). When we applied a linear regression model to the proportion as a response variable and the derived allele frequency as an explanatory variable, we could not even find a weakly significant correlation (fig. 4.3). Rather than assuming any specific distribution for the data, we used the nonparametric measure of correlation as before. Kendall’s correlation coefficient between the proportion and derived allele frequency was about 0.092 (P-value 0.074). When taken together, the regression result and the Kendall’s correlation result cannot be viewed as more than extremely weak support for the alternative hypothesis.

Replacing allele $i$ with the ancestral allele and allele $j$ with the derived allele in equation (4.3), we can measure the direction and magnitude of selective advantage of each derived allele $j$. The scaled selection coefficient ($2Ns$) of each derived allele $j$ against ancestral allele $i$ may be used as a criterion by which the derived allele $j$ is called neutral, deleterious, or advantageous. When the absolute value of $2Ns$ of an allele is less than 1, the allele might be called neutral (Ohta 1992). Out of 142 SNPs, there were 120 SNPs with the derived alleles that would be categorized as neutral by this operational criterion.

**Probability that an allele is ancestral with molecular fitness**

The probability that an allele is ancestral, with selection taken into account, is shown in equation (4.11). This probability is a function of both the scaled selection coefficient and the allele frequency. The probability of a minor allele being ancestral with selection becomes less than expected under neutrality as selection against the minor alleles becomes stronger, and vice versa. For each SNP, we calculate the values of the minor/major scaled selection coefficient (eq. 4.12). We then plug the calculated minor/major scaled selection coefficient and the minor allele frequencies for the SNPs into equation (4.11) in order to calculate the probability that an allele is ancestral with selection. These values of the calculated probability that an allele is ancestral with selection were plotted against minor allele frequency (fig. 4.4). When considering the weak significance of the analyses of previous
Figure 4.3: Scatter plot of the proportion of possible non-synonymous changes to the ancestral sequence that are more deleterious than the actual derived allele versus the derived allele frequency. The dashed line is the regression line of the data. The line has a slope of 0.042, and an intercept of 0.52264. We cannot reject the null hypothesis that the slope is 0 (ANOVA F-test $P$-value is 0.68).
sections, we did not expect any significant results. We hoped that the probability of a minor allele being ancestral for minor-allele-ancestral SNPs would be above the dashed neutral line, and that for major-allele-ancestral SNPs would be below the neutral line in figure 4.4.

However, we did not observe any striking patterns. Most minor-allele SNPs were rare, and 86% of SNPs have a minor allele frequency less than 0.1. Among these SNPs with rare minor alleles, there seemed to be many SNPs with ancestral major alleles.

Discussion

A variety of phenotypic properties of an organism contribute to fitness. A hierarchical structure of a phenotype may be envisioned from the molecular phenotype (e.g., RNA secondary structure) to the expression of a trait (e.g., blood type). At the tip of the phenotype hierarchy is the molecular phenotype that might determine fate of the genotype, and this induces the molecular fitness of a genotype. We explored the association of fitness and genetic variation by attempting to relate sequence-structure compatibility to fitness. Although we did not directly measure molecular fitness, we gauged the approximate difference of molecular fitnesses of two alleles using phenotype prediction tools from computational biology. We performed our association study of molecular fitness and genetic variation through scaled selection coefficients and allele frequencies.

A weak point of our analysis is the small sample size of protein-coding SNPs with known tertiary structure. There are far fewer protein structures available than sequences available, and the analyses with protein structures are often therefore limited in their sample size. We imposed strict requirements when compiling our data set that further reduced our sample size.

Our analysis also suffers from the evolutionary model that we employed. This assumes that molecular phenotype (i.e., protein tertiary structure) does not change over time. We hope to measure molecular fitness of contemporary alleles, one of which is ancestral. The fixed-structure assumption should not be extremely bad because protein structures do not change as much as do sequences (Chothia and Lesk 1986; Flores et al. 1993; Russell et al. 1997), but small changes of structure during evolutionary time could reverse sequence-structure compatibility scores of ancestral and descendant alleles, and of minor and major alleles.
Figure 4.4: Scatter plot of estimated probability of minor allele being ancestral \textit{versus} minor allele frequency. Ancestral probabilities are inferred by accounting for selection due to protein structure. Bullets are for SNPs with major alleles that are ancestral to minor ones, and cross marks are for SNPs with minor alleles that are ancestral to major ones. The dashed diagonal line indicates the expectation under neutrality.
Another flaw of our analysis stems from our procedure for collecting tertiary structure information. We chose protein structures of genes that had been chosen for SNP studies in the NIEHS data. Many protein structures may have been determined using protein sequences that are more similar to the major allele protein sequences than to minor alleles. This may induce a tendency for major allele proteins to fit the tertiary structures in our data set better than minor allele proteins. The degree to which this may affect our analyses is unclear.

We used the parsimony criterion for determining the ancestral allele of a biallelic SNP based on the pairwise sequence alignment of human and chimpanzee genome. Our assumption was that a single mutation induced the formation of a SNP in the human gene, and no other mutations at the SNP position occurred on the evolutionary tree relating the chimp allele with the minor and major human alleles. Mutation was assumed to be rare because of the relatively short evolutionary distance between humans and chimpanzees. Violation of the parsimony assumption could cause misidentification of ancestral alleles. A potential correction is available via a technique proposed by Hernandez, Williamson, and Bustamante (2007).

Our analysis was performed by assuming that the possibility of multiple SNP positions per protein was negligible. Multiple SNP positions did occur in our data set and, contrary to our analysis assumptions, natural selection might not independently act upon these variable positions. This issue could be handled if SNP databases made haplotype information easily available. Unfortunately, we were not able to obtain haplotype information. Last but not least, SNP genotyping is a complex ascertainment process, and it is not easy to discover rare SNPs whose allele frequency is extremely small. This incurs “ascertainment bias” in SNP genotyping (Nielsen and Signorovitch 2003; Clark et al. 2005). Failure to detect all SNPs properly, or imbalanced discovery of SNPs for various range of allele frequency would mislead the analysis, and estimates of empirical frequencies of ancestral alleles could suffer from ascertainment bias.

We can apply the technique we use in this work to other SNP types. Whereas there have been advances in the integration of functional nsSNPs discovery and protein structure prediction schemes (Chasman and Adams 2001; Sunyaev, Lathe, and Bork 2001a; Sunyaev et al. 2001b; Wang and Moult 2001; Ramensky et al. 2002; Hughes et al. 2003; Krishnan and Westhead 2003; Yue, Li, and Moult 2005; Yue and Moult 2006), it has been a couple of years since the integration of functional regulatory SNPs (rSNPs) discovery and
transcription factor binding site prediction in regulatory regions caught the attention of researchers (Stepanova et al. 2006). In the short evolutionary time since the split of human and chimpanzee, the phenotypic impact of regulatory regions might be stronger than that of protein structure. Mutations in regulatory regions may affect mRNA expression level, which will in turn affect protein expression level even when the amino acid sequences of the expressed proteins are the same. There might be a much stronger signal of association of molecular fitness and genetic variation in regulatory regions than in protein structure. The recent increase of rSNPs in terms of database size and development of transcription factor binding site prediction schemes might offer an avenue for molecular association studies in regulatory regions (Stepanova et al. 2006).

Different disease susceptibilities amongst individuals are due to their distinct genetic and environmental backgrounds. We can envision emergence of disease-associated alleles due to genetic and environmental changes. Rare Mendelian diseases seem to often be caused by deleterious derived alleles, and complex diseases are speculated to sometimes be due to ancestral alleles that are deleterious owing to environmental shift to favoring new alleles (Di Rienzo and Hudson 2005). It may be interesting using molecular phenotypes to assess how often SNPs whose ancestral alleles are advantageous are associated with rare Mendelian diseases, and how often SNPs whose ancestral alleles are deleterious are associated with complex diseases.

Here, we suggested a method to apply genotype-phenotype mapping to a molecular population genetic study. Genotype-phenotype mapping is central to modern biology. The process whereby genotypes are passed from generation to generation can be fully comprehended only if the way phenotypes affect genotypes is understood. As genotype-phenotype mapping techniques become refined, more applications of these techniques will arise. Furthermore, more diverse genetic variation data and methods using more advanced genotype-phenotype mapping techniques would provide an avenue of population genetic study at the molecular level.

Acknowledgments

We thank Philip Awadalla, Kate McGee, Asger Hobolth, Benjamin Redelings, and Reed Cartwright for their help. This work was supported by NSF grant DEB-0445180 and by
NIH grant GM070806.

Literature Cited


Chapter 5

Summary
A common theme among the three preceding chapters is the use of models of molecular evolution that exhibit dependent change of sequence sites due to natural selection. These models can be employed to quantify the impact of phenotype on rates of evolution and they can also serve as a basis for estimating population genetic parameters from interspecific data. In Chapter 2, we characterize the impact of globular protein tertiary structure on the rates of protein-coding gene evolution. The Robinson model (Robinson et al. 2003) fits most data of protein-coding genes better than a competing independent-site model that ignores tertiary structure. However, our analyses fail to find a strong connection between the importance of tertiary structure for the evolution of a protein family and the biological function of the protein family. The weak links that we do find seem to call for more realistic evolutionary models and improved genotype-phenotype mapping techniques. In fact, the four years subsequent to the publication of the Robinson et al. model have been a period during which related lines of research have appeared. In this chapter, we summarize these other research lines as well as speculate on related future directions.

**Followup Research of the Robinson Model**

While Robinson et al. (2003) initially implemented their statistical procedure for analyzing two sequences, Robinson (2003) and Rodrigue et al. (2005) extended it to more than two sequences. Rodrigue et al. (2005) also directly applied it to amino acid sequence evolution models. Yu and Thorne (2006) borrowed the statistical approach of Robinson et al. (2003), but they employed ribosomal RNA secondary structure as the surrogate of molecular fitness. Rodrigue et al. (2006) took the thermodynamic integration technique (Lartillot and Philippe 2006) to compare models with and without phenotypic dependency due to tertiary structure. Kleinman et al. (2006) retrospectively inferred parameters of a statistical potential under the maximum likelihood principle.

There remains much further investigation in the line of research. These includes the relaxation of fixed protein structure assumption; the creation of better and optimized energy potentials for protein and RNA structures; and the employment of other genotype-phenotype mapping techniques such as trans-membrane and mRNA structure (see Robinson 2003; Yu 2005). There have also been recent advances in computational strategies for inference that can drastically accelerate Markov chain Monte Carlo analyses with Robinson-
Sequence Path with Insertion and Deletion

Inspired by Jensen and Pedersen (2000), the Robinson model proposed sampling the history of nucleotide substitutions along a branch separating a sequence pair. This data augmentation technique (see Tanner and Wong 1987) permits inference with complicated models of molecular evolution. The hidden substitution history is called the sequence path, and it specifies when, where, and what mutation events occur along the branch. Given a pairwise sequence alignment without gaps, we may generate intermediate sequences randomly by sampling site paths for each site with an independent-site model. Implicitly, we assume that the intermediate sequences do not have gaps. If we want to relax the no-gap assumption with sequence alignment with gaps, then site paths are not fixed any more and sampling sequence paths becomes more challenging. For example, we have a pairwise sequence alignment of two sequences in figure 5.1.

\[
\begin{align*}
\text{sequence (A)} & \rightarrow \text{C-TAG--A-} \\
\text{sequence (B)} & \rightarrow \text{ACG-ACAT-T}
\end{align*}
\]

Figure 5.1: A pairwise sequence with gaps and its partition.

Let us define the length of states within a partition as the number of characters. We do not count gaps, but the length of a state is 1 if there is a single gap. We partition the alignment into columns where the starting entity has its length of 1 as above. We need to generate a "site" path between two entities. Each entity is of variable length characters. For example, we might need to sample a site path between C- and CG for the second partition of the above pairwise alignment. Because we have three general events (i.e., substitution, insertion, and deletion), the site path is not a substitution history any more.

This is a procedure to generate a sequence path with gaps for one selected partition among the partitions of the alignment. In the partition, we could adopt the following procedure. Firstly, we begin with the starting entity of the partition (e.g., C- in the second partition). We choose a character to which we apply one of the three general events. If the current state has length that is longer than 2, we randomly select a character among them. If
the current state has length of 1 (i.e., one character or single gap), then we use the character or gap to which we apply an event. Next, we choose a type of event for the selected character depending on rates of substitution, insertion, and deletion. We also sample the time to the next event from an exponential distribution. If we choose the substitution event type, we sample the next character using an independent-site model. If we choose the deletion event type, then we sample a deletion length from an insertion/deletion gap length distribution (e.g., Cartwright 2006), and delete the characters of the chosen size including the selected character and characters to the right-hand side of the selected character. If the deletion size is larger than the remaining right-hand side of the characters, we delete all the right-hand side of characters. If we choose the insertion event type, we sample a length of insertion from an insertion/deletion gap length distribution, and we randomly select characters from the stationary distribution of the characters according to an independent-site model. We then insert them at the right-hand side of the selected character. Once we sample the next state of a current state, we move on to the next state as the current one. We stop and redo the procedure unless the sample state is consistent with the ending state. We go back to the step where we choose a character to which we apply an event. We repeat the above procedure until we end up with an appropriate ending entity where the partition of the alignment and the simulated path are consistent and we can apply rejection sampling (von Neumann 1951; Nielsen 2001). For example, a possible scenario of site path from C to CG is shown in figure 5.2.

<table>
<thead>
<tr>
<th>time</th>
<th>state</th>
<th>event</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>C-</td>
<td>C is chosen and insert G</td>
</tr>
<tr>
<td>0.1</td>
<td>CG</td>
<td>C is chosen and substitute T</td>
</tr>
<tr>
<td>0.3</td>
<td>CT</td>
<td>T is chosen and substitute G</td>
</tr>
<tr>
<td>0.7</td>
<td>CG</td>
<td>no event until time 1.0</td>
</tr>
</tbody>
</table>

Figure 5.2: A site path between two entities of C- and CG.

Another complication of a sequence path with gaps is the calculation of the rate at which the sequence changes to a different sequence. Unless gaps are allowed, there are a fixed number of possible sequences that can result from a change to a sequence (e.g., $3L$ sequences can result from a point mutation to a DNA sequence of length $L$). If we take gaps into account in the calculation, the number of possible sequences that can be generated from a change is infinite because of the infinite possible ways of insertions. A mundane solution
would be to restrict possible changes by setting rates of overly long insertions or deletions to 0. With this restriction, the rates at which sequences change to other sequences might be approximately calculated. This is not perfect but it may be a starting point for generating sequence paths with consideration of insertion and deletion.

Improvement and better understanding of insertion and deletion models with independent-site models (Fleissner et al. 2005; Lunter et al. 2005; Redelings and Suchard 2005) may be helpful to develop the inference procedure of a dependent-sites model with insertion and deletion. One interesting topic involves measuring how alignment flaws affect the reconstruction of trees. What kind of features of alignment would affect the reconstruction of trees? This understanding might be relevant to the more practical issues of joint estimation of tree and alignment.

**Genotype-Phenotype Mapping Techniques**

If there is a good technique of phenotype prediction of genotype, we can immediately apply a Robinson-type model to study sequence evolution with dependence among sites. Unfortunately, phenotype prediction techniques other than molecular structure predictions do not yet seem to be well developed. Two other phenotype prediction techniques of interest are antigenicity prediction and protein expression prediction.

Although the development of antigenicity prediction (usually prediction of epitopes, which trigger host immune response and are 8 to 13 amino acids long) dates back to the time when protein structure prediction schemes were starting to be developed, antigenicity prediction methods are still not good enough to be relied upon. This is partly because experimental techniques of measuring antigenicity seem to still be poor. There is a centralized effort in collecting experimental results (Peters et al. 2005), and to provide robust antigenicity prediction tools, but the prediction tools are used to assess antigenicity of epitopes rather than of a whole sequence. Although it is possible to assess antigenicity of epitopes, it is difficult to combine all antigenicity scores of epitopes along the sequence to get a good sequence-level measure of antigenicity. The difficulty is due to the complicated biology of host immune system responses to the invasion of viral sequences. Some portions of a viral sequence may be subject to the host immune surveillance system, and other portions may not be because of different positions of epitopes in protein tertiary structures.
and different cuts by cleavage proteins before presentation to immune cells. The current methods of antigenicity prediction do not take long sequences into account but only deal with small segments of them. Just as Kleinman et al. (2006) investigated an optimized statistical potential of protein structure, there might be a way to retrospectively estimate antigenicity using currently available viral sequences and known phylogenies.

Protein expression levels have spatio-temporal variation. Rates of protein evolution are believed by Drummond et al. (2006) to be largely controlled by protein expression levels. Although tissue-specific genes are expressed differentially, housekeeping genes are expressed relatively constantly among cell types. Mammalian housekeeping genes show lower protein sequence conservation than tissue-specific genes (Farré et al. 2007). Just as protein or RNA structure constrains possible sequences, there may be evolutionary constraints on regulatory regions of tissue-specific genes. If the relatively constant protein expression level is predicted from protein sequences, then we may be able to study a dependent sequence change with expression-level prediction systems with a kind of the Robinson model. To my knowledge, there seems to be a paucity of gene expression data resources that link regulatory sequence to gene expression and that could thereby be employed to predict gene expression from regulatory sequence.

**Merger of Interspecific Evolution and Population Genetics**

Population genetics plays a key role in understanding evolution. The development of population genetics has usually been decoupled from that of interspecific evolution. This is unfortunate. Interspecific data should be able to inform the study of evolution within populations just as population genetic theory should be able to inform methods for analyzing interspecific sequence data. Approaches that combine intraspecific and interspecific considerations are promising (e.g., Kuhner and Felsenstein 1994; Hey and Nielsen 2007). We think such combinations will soon become a major emphasis of evolutionary genetics.
Literature Cited


