Abstract

YU, JIAYE. Dependence among sites in protein and RNA evolution. (Under the direction of Dr. Jeffrey L. Thorne)

Widely used models of molecular evolution assume independent change among sequence sites. This assumption facilitates computation but it is biologically unrealistic. RNA secondary structure and protein tertiary structure both change more slowly over time than do the encoding DNA sequences. The constraints upon sequence evolution that serve to maintain structure induce dependent change among molecular sequence positions. The object of this thesis is to characterize the impact of structure on sequence evolution.

The dependence among sites in protein evolution is first studied. Two simple and not very parametric hypothesis tests are introduced to study the spatial clustering of amino acid replacements within protein tertiary structure. Results of applying these tests to 273 protein families support the expectation that spatial clustering of amino acid replacements within tertiary structure is a ubiquitous phenomenon. More importantly, patterns of amino acid replacements do not seem to be solely attributable to spatial clustering of sequence positions that are independently evolving and have high rates of change. Instead, application of the newly introduced simple hypothesis tests yields evidence for dependent change among spatially clustered protein positions. This portion of the thesis work thereby casts doubt upon widely used methods for phylogeny inference.

The second focus of this thesis is the impact of RNA secondary structure on RNA evolution. A model of RNA evolution incorporating RNA secondary structure is developed. The model introduces dependence among sites in RNA evolution via

the effects of sequence changes on the approximate free energy of the resulting RNA secondary structure. This approximate free energy information can be thought as surrogate of fitness that serves as a link between genotype and phenotype in the model. Analysis of eukaryotic 5S ribosomal RNA sequences with this model shows the importance of RNA secondary structure on evolution. This analysis also confirms the value of the new model for studying adaptive evolution and for inferring ancestral sequences.

Dependence among sites in protein and RNA evolution

BY

Jiaye Yu

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

2005

APPROVED BY:

William R. Atchley	Carla Mattos
Jeffrey L. Thorne CHAIR OF ADVISORY COMMITTEE	Bruce S. Weir

献给我的父母和妻子

To my parents and my wife

Biography

Jiaye Yu was born as the only child to his parents Zhi Yu and Jianhua Zhou in Wenshan, Yunnan Province, China in November 1976. He was admitted to Shanghai JiaoTong University (SJTU) in 1993 and received a Bachelor of Engineering degree in Biochemical Engineering in July 1997. He then received a Master of Science degree in Biochemistry and Molecular Biology in March 2000 from SJTU. In August 2000, Jiaye came to North Carolina State University (NCSU) to pursue a PhD degree in Bioinformatics. Under the direction of Dr. Jeffrey L. Thorne, Jiaye studied dependence among sites in evolution of protein and RNA sequences. He will continue to work in the area of molecular evolution as a postdoc in University of Copenhagen, Denmark.

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisor Dr. Jeffrey L. Thorne for his thoughtful guidance to me and his unbelievable patience with me in the last several years. Without him, it will be much more difficult for me to make progress and finish as scheduled. I have learned a lot from Jeff both academically and personally. I feel very fortunate to have such an incredible advisor. I also thank Dr. Atchley, Dr. Mattos and Dr. Weir for contributing their time to serve as the members of my advisory committee.

I really appreciate the helps from the current and past members Jeff Thorne's group, Stéphane Aris-Brosou, Sang-Chul Choi, Asger Holboth, Douglas Robinson and Tae-Kun Seo. I enjoy the fruitful discussions with them a lot.

I thank Kejun Liu for sharing his thoughts on computer programming with me. I also thank my friends in North Carolina State University, Jixin Deng, Xiaoyi Gao, Weichun Huang, Jian Li, Xiang Yu and Wei Zou for their friendship.

Finally, I thank my wife Xiaobei Zhao for her love and patience during my hard times. She always takes good care of me. I am also indebted to my parents Zhi Yu and Jianhua Zhou for their endless love and their supports to me over years. To them I dedicate this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Review

1.1 Introduction

The study of evolution has a long history, dating back to Darwin in the nineteenth century. Phylogenetics, the reconstruction of evolutionary history, lies in the center of all evolutionary studies. Before the rapid development of modern molecular biology in the 1970's, phylogenetic reconstruction was mainly based on the morphological data analysis. The number of reliable homologous morphological characters is obviously limited (e.g., in microorganisms). The use of molecular sequence data greatly changed this situation. Even when the amount of sequence data was not as large as it is today, Zuckerkandl and Pauling (1962) foresaw the possibility of using DNA sequence information to classify species.

After successful applications of parsimony and distance methods (e.g., Eck and Dayhoff 1966; Fitch and Margoliash 1967) to reconstruct the evolutionary history from sequence data, more statistical methods were introduced into this area and gradually became popular (see Holder and Lewis 2003). With the completion of various sequencing projects, more sequence data are available than at any time before. Phylogenetic analyses of huge amounts of genomic data incurs much more interest (Delsuc et al. 2005) than before. Mathematical models of sequence evolution are keys to phylogenetic analysis. Since Jukes and Cantor (1969) proposed the first stochastic model for DNA evolution, a large number of variant models have been developed.

In this chapter, I will briefly review the development of models of sequence evolution and especially emphasize the problems that I have been studying.

1.2 Statistical models of sequence evolution

Evolution of molecular sequences is a complex biological process and perfect mathematical models are unavailable. In practice, assumptions have to be made to propose a model. Only a few biologically realistic factors will be taken into account and the others will be greatly simplified. Although it is possible to incorporate arbitrarily many possible parameters into a model to represent reality, an over-parameterized model tends to be not only mathematically difficult, but also computationally impractical. The balance between model complexity and computational feasibility is always a big concern of researchers. Until now, a large number of models have been developed for different types of molecules. Although versatile, they do share common assumptions.

Probably the most widely used assumption for models of sequence evolution is that each site evolves independently in a sequence. Under this assumption, the substitution events at one site do not affect the substitution events at neighboring sites. This assumption is statistically convenient and computationally feasible. Under this assumption, evolution of sequence can be partitioned into evolution of character states at single sites. For example, in the common maximum-likelihood procedure to reconstruct phylogenetic trees (Felsenstein 1981), the total likelihood can be written as the product of all site likelihoods. Although it is not often explicitly claimed, parsimony and distance methods also tend to assume independence among sites.

When the problem of sequence evolution is turned into a problem of evolution of character states at individual sequence positions, the size of the problem is greatly simplified. The number of possible character states per sequence position is limited; it is 4 for DNA sequences and is generally 20 for protein sequences if we do not consider

rare amino acid variants. A Markov process for changes among character states is a further typical assumption. The Markov assumption means that the probability of starting with character state i at time t_0 and ending with state j at time $t_0 + \Delta t$ does not depend on any information of character states before time t_0, see Equation 1.1 where $s(t)$ represents the character state at time t.

$$p_{ij}(\Delta t) = P(s(t_0 + \Delta t) = j | s(t_0) = i) \tag{1.1}$$

In other words, the transition probability $p_{ij}(\Delta t)$ is only dependent on the starting and ending character states and the time interval Δt. The starting time t_0 is not important in this process. Knowing all possible character states, the transition probability $p_{ij}(t)$ can be tabulated as a square matrix $P(t)$. It can be calculated by

$$P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!} \tag{1.2}$$

according to the theory of continuous time Markov process (Karlin and Taylor 1975). The square matrix Q is known as the instantaneous rate matrix, which represents the substitution rate from one character state to another in an infinitesimal time period. If insertions and deletions are not considered, the instantaneous rate matrix Q determines a Markovian model of character state changes. Transition probabilities can be calculated by matrix exponentiation as described above.

Time reversibility is a property associated with some Markov processes. For any pair of character states i and j, time reversibility means that

$$\pi_i \times p_{ij}(t) = \pi_j \times p_{ji}(t), \tag{1.3}$$

where π_i represents the stationary probability of character state i. Given an ancestral and a descendant sequence, time reversibility indicates that we cannot determine which sequence is ancestral and which is descendant on the sole basis of sequence information.

In practice, what we can estimate is the evolutionary distance between sequences. This is because the substitution rate and evolutionary time are effectively confounded. This means that, without external information, it is impossible to distinguish between the situation of Qt and $(2Q)(\frac{1}{2}t)$ because the resulting distances are the same. Because of the confounding effect of rate and time, the rate matrices Q are usually rescaled so that the average rate of change is 1 according to Equation 1.4,

$$\mu \sum_i \pi_i \sum_{j \neq i} q_{ij} = 1. \tag{1.4}$$

Then, the amount of evolution separating two sequences can be measured by the expected number of changes that separate them.

1.2.1 Models of DNA substitution

Jukes and Cantor (1969) proposed the first model of DNA substitution (JC69). The basic assumption of their model is that all possible substitution events have identical rates. If one nucleotide changes, it will change to any of the other three possible nucleotide with probability of $\frac{1}{3}$. Thus, the stationary probability of each nucleotide is 0.25. Although it seems oversimplified, the JC69 model well presents the essence of stochastic modeling of sequence evolution. Most of the alternative models have the same framework.

One biological fact that JC69 ignores is that not all types of substitutions among nucleotides are equally probable. There are two main chemical groups among the four nucleotides. Adenine and Guanine are purines and Cytosine and Thymine are pyrimidines. Substitution events within groups are called transitions whereas substitutions between groups are called transversions. In practice, transitions occur much more frequently than transversions. Kimura (1980) incorporated a new parameter κ, the transition/transversion rate ratio, to differentiate the two kinds of substitutions. We refer to this as the K2P or Kimura 2-Parameter model. Like the JC69 model, when stationarity is reached, it is still equally probable for a nucleotide to be A, C, G or T under K2P. The substitution rate matrices of JC69 and K2P model and some other models are shown in Table 1.1.

It is recognized that there is nucleotide frequency bias in DNA sequences and it is not necessary to force them to be equal in evolutionary models. Felsenstein (1981) incorporated nucleotide frequencies as free parameters into his model by setting the substitution rate from nucleotide i to j as begin proportional to π_j, the base frequency of destination nucleotide. We refer to this as the F81 model.

In their HKY85 model, Hasegawa, Kishino and Yano (1985) combined the ideas of both K2P and F81 models by proposing a model with both unequal nucleotide frequencies and transition/transversion rate differences. Felsenstein has long used a very similar model known as F84 in his software package PHYLIP (Felsenstein 1993), but he never published it (Felsenstein 2004). Tamura and Nei (1993) further modeled the difference between two types of transitions (A$\leftrightarrow$G and C$\leftrightarrow$T) and two transition/transversion rate ratios are introduced into their TN93 model.

The most general model of DNA substitution for a single site where all sites change

independently but according to the same process is an unrestricted one with 12 free parameters, which can be referred to as the UNREST model. It is not a widely used Markovian model and some useful properties in evolutionary analysis are lost (e.g., time reversibility). In common practice of molecular evolution, the most general widely used model is the General Time Reversible (GTR) model (Lanave et al. 1984). Under this model, the nucleotide frequencies are free parameters and the differences among all possible substitutions are taken into account. Obviously, the GTR model is a special case of the UNREST model with the assumption of time reversibility. Except for the UNREST model, the other models that we have mentioned are all special cases of the GTR model.

With these early models of sequence evolution, substitutions occur independently and also identically among sites. This means that all sites evolve at the same rate. Identical rates among sites is rarely the truth in reality (e.g., Yang 1993; Yang 1994). The rate of change at different sites can be rescaled so that the rate at the average site is 1. For each site i, the scaling factor γ_i that determines its rate can be modeled as an independent draw from a gamma distribution with the shape parameter $\alpha = 1/\beta$. It is pretty computationally infeasible to integrate over a continuous gamma distribution of rates in phylogenetic likelihood calculations. Yang (1994) approximated the continuous gamma distribution with the "discrete gamma distribution" by using several categories of rates with equal probabilities in each category. This substantially reduced the computational time needed.

1.2.2 Models of amino acid replacement

Nucleotide models work well when the divergences among sequences are not large. When the species are distantly related, using DNA sequences might cause some problems. A good alignment based solely on highly diverged DNA sequences is not easy to obtain and it is possible that the nucleotide frequencies in species are significantly different from each other. In this case, the usage of protein sequences is desirable (e.g., Hasegawa and Hashimoto 1993; Liò and Goldman 1998; Whelan et al. 2001). Quite different from the models for nucleotide substitution, models of amino acid replacement generally do not have analytic forms for their replacement rate matrices and were traditionally empirically obtained from data. The most influential model was developed by Dayhoff and collaborators (Dayhoff et al. 1972; Dayhoff et al. 1978).

Dayhoff et al. (1978) collected a large number of the protein sequences available at that time and relied upon 71 groups of closely related sequences from data. A parsimony method was used to reconstruct the phylogenetic tree for each group and to infer ancestral sequences for each group. The number of replacements were then tabulated into an Accepted Point Mutation (PAM) matrix. Evolutionary distances can be represented by the number of accepted point mutations. By definition, the Dayhoff distance of 1 PAM means 1 accepted point mutation per 100 amino acids. The transition probabilities for the 1 PAM distance can be derived the matrix tabulated by Dayhoff et al. (1978). Transition probabilities for other distances can be obtained by self-multiplication of the 1 PAM matrix. The transition probability matrices can be transformed into scoring matrices, which are widely used to align protein sequences (Dayhoff et al. 1978). Among them, the most well known is 250 PAM matrix. It was formerly widely adopted for alignment of distantly related sequences but has now

been largely replaced by the BLOSUM62 matrix (Henikoff and Henikoff 1992).

Jones et al. (1992b) built a new model (the 'JTT model') with similar methodology to Dayhoff's but with a larger database. A slight difference is that Jones, Taylor and Thornton did not infer ancestral sequences, instead they only utilized pairwise comparisons between observed protein sequences to estimate the number of amino acid replacements. Jones et al. (1992b) thereby avoided some of the inaccuracy caused by parsimony-based ancestral reconstruction.

In the procedure of either Dayhoff et al. (1978) or Jones et al. (1992b), transition probability but not rate matrices are directly produced. Kishino et al. (1990) devised a method to obtain the instantaneous rate matrix Q from a transition probability matrix $P(t)$ by taking advantage of the relationship of the eigen systems of $P(t)$ and Q, see Equation 1.2. The recovered Q can then be used to generate transition probabilities at any real-numbered evolutionary distance. Here, Q is a 20×20 matrix and it can be decomposed to the product of two matrices S and Π by $Q = S \times \Pi$. S is the symmetric exchangeability matrix and the non-zero diagonal elements of Π are amino acid frequencies. The exchangeabilities can be fixed to the values that were inferred from the original data of Dayhoff et al. (1978) or Jones et al. (1992b), but the amino acid frequencies π can be estimated solely from protein family or families being analyzed (Cao et al. 1994). When the frequency π are directly estimated in this ways but Dayhoff or JTT exchangeabilities are used, we conventionally denote the resulting models by Dayhoff+F or JTT+F.

Besides the nuclear genome, mitochondrial and chloroplast genomes encode some proteins too. Because these genomes evolve differently, some models have been specifically for specific genomes. Based on vertebrate mitochondrial proteins, Adachi and

Hasegawa (1996) developed the specialized mtREV model. They demonstrated that mtREV outperforms the Dayhoff/JTT model when applied to mitochondrial proteins. Similar work was done on chloroplast proteins (Adachi et al. 2000). For protein sequences from different sources, it is important that an appropriate model be selected for reliable phylogenetic inferences. Another noticeable feature of the procedures adopted by Adachi and Hasegawa is that the maximum likelihood approach is used instead of the simple counting method in the work of Dayhoff et al. and Jones et al.. Yang et al. (1998) adopted a similar maximum likelihood procedure. Compared with the relatively fast counting method, the maximum likelihood method is slow but is more statistically solid. Whelan and Goldman (2001) combined good properties of both method and estimated a new model (WAG) from a large database by ML approximation in a reasonable amount of time.

Recently, Kosiol and Goldman (2004) demonstrated that the common procedure of generating Q from $P(t)$ based on simple counting methods is not accurate enough. They proposed to calculate Q directly from the original data of Dayhoff et al. (1978) by the number of observed changes $n_{ij,\,i\neq j}$, the mutability of amino acids m_i and the amino acid frequencies f_i. They named the model as DCMut and suggested it to be adopted as a standardized alternative for the Dayhoff/JTT type of model.

1.3 Evolutionary dependence among sites

In the models of DNA and amino acid substitution mentioned above, dependence among sites is not taken into account. This makes computation feasible even though

the assumption of independence among sites is absolutely not true. In a protein-coding region, due to the dependence within triplet codons, it is more appropriate to model evolution with the unit of codon, which incorporates the information within a triplet. In DNA sequences that code for non-coding RNAs (e.g., tRNA or rRNA), the impact of secondary structure cannot be easily ignored. Efforts to incorporate secondary structure information into evolutionary models have been consistently made during the last ten years. The attempts of different researchers to overcome the assumption of independent substitution events among sites have been made for years. The simplest cases include the dependence structure within a triplet codon and the dependence among two nucleotides due to the impact of conserved RNA secondary structure.

The models of mononucleotide and amino acid substitution described above assume that substitutions are completely independent of each other. In reality, this is hardly the truth. For example, both protein and RNA molecules need to form specific in higher order structures to be functionally active. This simple fact provides the basis of relaxing the assumption of site independence.

1.3.1 Models of codon substitution

In protein coding regions, the dependence among neighboring sites within a triplet codon was considered by Goldman and Yang (1994). They proposed a model focusing on the substitutions at the level of codons (triplets) instead of the traditional mononucleotide models (e.g., JC69 or HKY85). Instead of modeling the substitution events among four possible nucleotide types, the substitutions among 61 possible non-stop codons are modeled. It makes using distantly related DNA sequences for

evolutionary study more appropriate than before.

This leads to a bigger rate matrix and makes computations more demanding, but it is a more realistic model for protein coding region compared with nucleotide models. When the nucleotides within a triplet are taken as a whole unit of evolution, the assumption of independence among these sites is naturally relaxed. The dependence among sites within a triplet codon will be taken into account and the different substitution rates among codons can be modeled explicitly. Goldman and Yang (1994) used the distance matrices derived from the comparisons of physiochemical properties among 20 amino acids by Grantham (1974) to represent the differences among amino acids in their original model. Muse and Gaut (1994) also proposed a codon approach. These authors modeled equilibrium frequencies of nucleotides, instead of codons. Codon-based models have the advantage of differentiating between nucleotide changes that do not affect the encoded amino acid (synonymous changes) and nucleotide changes that do affect the encoded amino acid (nonsynonymous changes).

For a nonsynonymous change, depending on the difference between the resulting amino acid and the original amino acid in terms of physical and chemical properties, the protein can be fully or partially functional, or lose the function. Most nonsynonymous substitutions are slightly deleterious and they have higher probability of fixation in small populations than in large ones (Ohta 1995). In some genes, the nonsynonymous substitution rates are higher than the synonymous substitution rates (e.g., Hughes and Nei 1988; Lee, Ota and Vacquier 1995). In these genes, nonsynonymous substitutions tend to be beneficial to the overall functionality of protein and they tend to be fixed by selection. The development of codon models provides the possibility to study the selection pressure on the protein coding regions.

Yang and Nielsen (2000) simplified the original model of Goldman and Yang (1994) by introducing the parameter ω, which represents the ratio of nonsynonymous to synonymous substitution rate, so the substitution rate from codon i to j ($i \neq j$) can be written in Equation 1.5,

$$q_{ij} = \begin{cases} \pi_j & \text{synonymous transversion} \\ \kappa\pi_j & \text{synonymous transition} \\ \omega\pi_j & \text{nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{nonsynonymous transition} \\ 0 & i \text{ and } j \text{ differ at more than one site} \end{cases} \tag{1.5}$$

Generally, we can assume that nonsynonymous substitutions are free of selection pressure because they do not alter the resulting proteins. When a nonsynonymous substitution has higher rate than it would if it were a synonymous one ($\omega > 1$), it indicates that this substitution is beneficial and is more likely to be fixed into the progeny. When $\omega < 1$, it suggests that the nonsynonymous substitution is deleterious and will be purified by selection. This parameter ω is in fact a counterpart of the conventionally used d_N/d_S ratio in the counting method of Nei and Gojobori (1986). Because of its solid statistical interpretation, the approach of the model-based ω estimate has become a standard procedure to detect positive selection in protein coding regions.

1.3.2 Protein structure and evolution

Although primary sequences of protein can contain all the necessary information to fold into the correct tertiary structure (Anfinsen 1973), it is known that protein tertiary structure is more conserved and evolves more slowly than do protein sequences. This is because a relatively stable structure is crucial for a protein to be functionally active (Chothia and Lesk 1986; Flores et al. 1993; Russell et al. 1997).

Although it is not easy to directly combine protein tertiary structure into models of protein sequence evolution, researchers realized that the information of secondary structure is also worthwhile to be considered (e.g., Thorne et al. 1996; Goldman et al. 1998). By classifying sites into different categories according to the types of secondary structure to which they belong, Thorne et al. (1996) modeled the substitution events for each category separately. Compared with the approach adopted by Dayhoff et al. to model the substitution process of "average" protein over all available proteins, Thorne et al. (1996) concentrated on the difference among this process at different sites due to different secondary structure ($\alpha-$helix, $\beta-$sheet and loop). Goldman et al. (1998) further estimated the impact of secondary structure and solvent accessibility on the substitution rates of protein evolution.

There have been several recent studies incorporating evolutionary dependence among sites (e.g., Jensen and Pedersen 2000; Pedersen and Jensen 2001; Hwang and Green 2004;Pedersen et al. 2004) by taking into account local dependence among sequence positions. Robinson et al. (2003) developed the model among entire sequences that incorporates protein tertiary structure. The sequence-structure compatibility measured by pseudo-energy potential (Jones et al. 1992; Jones 1999), which is often used in protein threading to predict protein structures, is adopted to represent

the effect of structure over substitution rates. The statistical approach adopted in Robinson et al. (2003) is quite general and can be applied to different types of general dependence structures, which are commonly found in the study of molecular evolution.

1.3.3 Structure of RNA molecules

There are different kinds of RNA molecules and these RNA molecules have diverse biological functions. For RNAs that do not code for protein sequences (ribosomal RNA and transfer RNA are the classic textbook examples), a single RNA sequence normally folds into a specific structure in three dimensional space to be functionally viable. Although it is the tertiary structure that determines the functionality of RNAs, previous studies demonstrated that the tertiary structure of RNA is basically determined by its secondary structure (e.g., Celander and Cech 1991; Doherty and Doudna 1997) although some exceptions do exist where the formation of tertiary structure greatly changes the secondary structure (e.g., Wu and Tinoco 1998). A typical RNA secondary structure is shown in Figure 1.4, which is taken and modified from the website of Signal Recognition Particle Database (SRPDB, `http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html`). Basic elements are labeled accordingly, including helix and different types of loops.

RNA molecules have the potential to form base pairs within the sequence when two parts of the sequence are complementary because of the interaction of hydrogen bonds. It is common to find Watson-Crick type base pairs CG and AU. Although less stable, the GU pair is also significantly more frequent than the other base pairs. Isolated pairs are usually not stable, so it is common to find helical regions containing several

stacked base pairs although the number of continuous base pairs is rarely more than 10 (Higgs 2000). The stability of a helical region is much from the stacked interactive attractions between successive base pairs. It is usually assumed that free energy of a helical region follows the nearest neighbor model, which means that only the free energy between two neighboring base pairs will be considered. A fact about RNA secondary structure is the relative stability compared with the primary sequence.

In the last two decades, methods to determine RNA secondary structure from a single sequence has been greatly improved. Besides the direct approach of experimental determination, the predictive methods based on energy-minimization or comparative sequence analysis have gained much success. Nussinov et al. (1978) applied dynamic programming to predict secondary structure by finding the secondary structure for an RNA sequence with the maximum number of base pairs. Zuker and Stiegler (1981) extended this algorithm to the situation that the minimum free energy is the target. It is possible that the minimum free energy structure is not the "true" structure in reality. Instead of predicting a single secondary structure from sequence, Zuker (1989) improved the previous method by predicting multiple suboptimal secondary structures within a range of energy values. All of these suboptimal secondary structures have reasonably low free energy and it is more probable to get the "true" structure among them. The predictive methods work reasonably well when the sequences are not long. There are some limitations in the early versions of dynamic programming algorithms applied to predict RNA structure (e.g., pseudo-knots are ignored). Recently, new methods based on energy approximations were proposed to incorporate pseudo-knots (e.g., Lyngsø et al. 1999). Currently, there are several well known software packages available for RNA structure prediction, including Vienna

RNA (Hofacker et al. 1994; Hofacker 2003) and RNAStructure (Mathews et al. 1999; Mathews et al. 2004). For large RNA molecules, the accuracy of prediction based on energy minimization is not high. In these cases, comparative analysis is commonly used. The basic ideas are well introduced in Durbin et al. (1998) and Eddy (2004). Zuker (2000) had a good review about the energy-minimization based methods for secondary structure prediction and Gardner and Giegerich (2004) had a comprehensive comparison of different methods of predicting RNA secondary structure.

Both energy and entropy changes of helix formation can be measured in experiments with short nucleotide sequences (e.g., Freier et al. 1986; SantaLucia and Turner 1997). The single-stranded regions that occur between helices can be named in various ways depending on their locations (e.g., hairpin loop, bulge, internal loop and multi-branched loop). Some loop free energies have already been experimentally estimated although these loop parameters are less accurate than the helix parameters in general (SantaLucia and Turner 1997). For special cases like multi-branched loops, there are no thermodynamic data available. It is the number of unpaired bases instead of the types of these bases that determines the free energies of this region. Among different loops, tetraloops are somehow unique. They are particular sequences of four single stranded bases (e.g., GNRA, where N could be any base and R represents a purine) that occur frequently in length-four hairpin loops, and that have increased thermodynamic stability due to interactions between the unpaired bases. For the thermodynamic parameters that cannot be experimentally determined (e.g., energy contribution from a multibranch loop), reasonable estimates are available. The total free energy of a complete molecular structure is usually estimated by summing over all free energy terms coming from the different parts of a secondary structure.

According to the second law of thermodynamics, it is intuitive to expect that the lower the free energy, the more stable the structure. It seems obvious that real sequences should have lower free energy than random sequences. However, this issue has been the basis of some controversy. Seffens and Digby (1999) concluded that free energy of actual sequences is significantly lower than random sequences. Workman and Krogh (1999) subsequently cast doubt on the study of Seffens and Digby by pointing out that Seffens and Digby (1999) did not consider the dinucleotide frequencies when performing the permutation tests used in their study. In terms of biology, it is not surprising that mRNA does not have low free energy. Different from tRNA or rRNA, mRNA will be degraded soon after the translation has completed and the requirement of a stable structure is not necessary. Workman and Krogh (1999) further concluded that even for tRNA or rRNA sequences, the free energies are not significantly lower than the ones of random sequence. Recently, Clote et al. (2005) demonstrated from large scale data analysis that free energies of tRNA/rRNA sequences are much lower than random sequences. This is a more biologically reasonable result. Furthermore, Clote (2005) showed that the number of locally optimal structures is significantly less for real structural RNA sequences than for random RNA sequences.

1.3.4 RNA structure and evolution

For both protein and RNA, there is a hierarchy of primary, secondary and tertiary structure. However, there are radical differences between the secondary structure of protein and RNA. Protein secondary structure is simply the local conformation of the backbone of polypeptidyl chain. The β-sheet and α-helix secondary structure conformations are widespread in proteins because these are conformations where free

energy often be minimized. Taking into account the knowledge of protein secondary structure indeed helps to improve the models of protein change (e.g., Thorne et al. 1996), but the difference between fits of models that do and do not incorporate protein secondary structure does not tend to be dramatic. One reasonable explanation for why models of protein evolution are only moderately improved by adding secondary structure is that evolutionary dependence among sites in proteins is only partially associated with protein secondary structure.

RNA secondary structure is a different case. The secondary structure of RNA contains much more information pertaining to tertiary structure than does protein secondary structure does. In some sense, RNA secondary structure is a not-very-accurate approximation of RNA tertiary structure. The dependence among sites in RNA evolution is relatively strong. It is also relatively specific in that much of this dependence is associated with base pairing in helical regions of RNA secondary structure.

Schöniger and von Haeseler (1994) proposed jointly modeling substitution events in RNA helical regions. With the assumption of a fixed RNA secondary structure, all sites can be classified into two categories, helical region or loop region. The substitutions in loop region can be modeled with the conventional nucleotide models while the unit of evolution in helical regions will be doublets (base pairs). There are totally sixteen possible base pair combinations and the dimension of the instantaneous rate matrix will be 16×16. This model can be considered as a general extension of F81 model to doublet case. Muse (1995) then proposed three different models by using mononucleotide frequencies instead of doublet frequencies. A new parameter λ is introduced to represent the effect of forming or destructing a base pair over the

substitution rates. The most general one of his models have substitution rate from doublet i to j as

$$q_{ij} = \begin{cases} \kappa\pi_t & \text{transition, pairing unchanged} \\ \pi_t & \text{transversion, pairing unchanged} \\ \kappa\pi_t\lambda & \text{transition, unpaired} \rightarrow \text{paired} \\ \pi_t\lambda & \text{transversion, unpaired} \rightarrow \text{paired} \\ \kappa\pi_t/\lambda & \text{transition, paired} \rightarrow \text{unpaired} \\ \pi_t/\lambda & \text{transversion, paired} \rightarrow \text{unpaired} \\ 0 & \text{two nucleotide differences between } i \text{ and } j \end{cases} \tag{1.6}$$

Here, π_t is frequency of the mononucleotide that differ in i and j, κ is common transition/transversion rate ratio. The new parameter λ reflects the effect of base pairing over substitution rates. Ideally, λ will be a number that larger than 1, which means a substitution from unpaired bases to paired bases is favored by evolution. In contrast, the substitution rate from a paired doublet to an unpaired doublet will be slow under this model.

In the Schöniger and von Haeseler model, all sixteen possible combinations of base pairs are considered. There are several variants of this type of models developed by several other researchers. Tillier (1994) only considered the six possible matching pairs among A, C, G, U and the rate matrix is 6×6 in her model. It is also possible to treat all mismatch pairs as one type of state to obtain a 7×7 rate matrix (Tillier and Collins 1998). Instead of using the doublet frequencies, mononucleotide frequencies can be applied into the model (Muse 1995) or these frequencies can be simply assumed to be identical as with the JC69 and K2P models (Muse 1995; Rzhetsky

1995). Because almost all of these doublet models are nested, likelihood ratio tests can be easily applied to compare these models. Savill et al. (2001) compared these models and many more other possible models under likelihood framework and concluded that the most general models perform best, which is not a surprising result. Recently, Smith et al. (2004) used the idea of an empirical rate matrix, which is usually adopted for protein data, to summarize rate matrix for RNA sequence from a large number of sequences. It is possibly a feasible approach when more biological knowledge accumulated and it becomes more difficult to develop a parameter-rich model to incorporate all information available.

For several widely available RNA molecules (e.g., transfer RNA, ribosomal RNA and ribonuclease P RNA), secondary structures are relatively more conserved over time but the primary sequences vary considerably (e.g., Dixon and Hillis 1993; Kirby, Muse and Stephan 1995, Gutell 1996; Gutell et al. 2002; Hofacker, Stadler and Stocsits 2004). This indicates that stable secondary structure is essential to maintain the functions of these RNA molecules. The mechanism of compensatory mutations is widely accepted as the explanation as the conserved secondary structure since it is proposed by Kimura (1983). Stephan (1996) studied the rate of compensatory mutations under different conditions of selection pressure by simulations. Higgs (1998) extended the study by allowing reversible mutations. The dependence among substitution events at different sites, especially within the helical region of secondary structure, is obvious. If the secondary structure can be assumed as known and unchanged, the helical and loop regions can be analyzed with different models, respectively. The unit of substitutions within helical regions is doublets (base pairs) instead of single nucleotides. The number of doublet states is 16 (4×4) if all the possible combinations of base

pairs are taken into account.

1.3.5 Genotype, phenotype and fitness

At the level of molecular sequences, the traditional concepts of genotype and phenotype in genetics can be applied to the same molecule. The primary sequence can be referred as genotype and the functionally active tertiary structure can be considered as the phenotype. Until now, although much of the effort has been put on developing different kinds of models for evolution of molecular sequences, little is done to take into account genotype-phenotype interactions.

The structure of general dependence among sequence positions extensively exists in nature. In the case of RNA, it has been shown that the stabilizing forces maintaining a tertiary structure mainly come from the formation of appropriate secondary structure and the tertiary structure is normally an arrangement of secondary structure in three-dimensional space (Grüner et al. 1996). It is natural to utilize RNA secondary structure to represent the concept of phenotype. Free energy of RNA secondary structure can be effectively treated as a surrogate of "fitness" information that connects genotype and phenotype. The availability of procedures to approximate free energy of RNA secondary structure can be especially useful when this surrogate for "fitness" will be used.

With the tendency to maintain a structure with low free energy for an RNA molecule, free energy can not only be used for predicting RNA secondary structure but also be interpreted as the surrogate of fitness information when the concepts of genotype and phenotype are applied to primary sequence and secondary structure of

RNA, respectively. Free energy is a kind of mapping between sequence and structure. The lower the free energy, the more stable the secondary structure and the higher the "fitness". It is natural to expect that the substitution rate away from an RNA sequence with low free energy will be low and the rate away from an unstable RNA sequence will be noticeably higher. Using the information of free energy of RNA secondary structure, genotype and phenotype can be effectively connected and this provides the possibility to study the scenario of adaptive evolution. Combining with reliable secondary structure information from comparative analysis, free energy information can be used as an reasonable approximate measure of fitness.

Dependence among sites extensively exists in molecular sequences. The most obvious ones are RNA secondary structure and protein tertiary structure. At the level of molecular sequences, the traditional concepts of genotype and phenotype can be applied to primary sequences and corresponding structures (e.g., RNA secondary structure or protein tertiary structure). Until now, evolutionary models of molecular sequences have been extensively developed, most of the efforts were put on the evolution of primary sequences (i.e., genotypes) while the information about structures (i.e., phenotype) is more or less ignored. There are versatile reasons for the lack of consideration of phenotype information into an evolutionary model, probably the most important one is the computational complexity that will be introduced by taking into account the general impact of phenotype on genotypic evolution.

1.4 Conclusion

The development of models of molecular sequence evolution is crucial to the study of molecular evolution. Although a large number of models have been developed and elaborated in the last several decades, there are still many aspects to improve. Statistically convenient assumptions, especially independence among sites, are computationally feasible but biologically flawed.

After we have sophisticated models under the assumption of independence among sites, it is a natural step to relax this assumption and to attempt to develop more realistic models. We can do this because of the much more powerful computing facilities available today. Actually, we hope the work described here helps to pave the lanes for the study of relationship between genotype and phenotype during evolution. The work can also assist the study of adaptive evolution.

The following two chapters present two independent but related works. The first one is the study of relationship between protein tertiary structure and protein evolution. The second one develops and applies a model of RNA evolution that incorporates RNA secondary structure.

1.5 References

Adachi, J. and M. Hasegawa. 1996. Models of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

Adachi, J., P. J. Waddell, W. Martin and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348–358.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. Science **181**:223–230.

Cao, Y., J. Adachi, A. Janke, S. Paabo and M. Hasegawa. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. J. Mol. Evol. **39**:519–527.

Celander, D. W. and T. R. Cech. 1991. Visualizing the higher order folding of a catalytic RNA molecule. Science **4992**:401–7.

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

Clote, P. 2005. RNALOSS: a web server for RNA locally optimal secondary structures. Nucl. Acids Res. **33**:W600–W604.

Clote, P., F. Ferré, E. Kranakis and D. Krizanc. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA **11**:578–591.

Dayhoff, M. O., R. V. Eck and C. M. Park. 1972. In Atlas of protein sequence and structure, vol. 5, 89–99. National Biomedical Research Foundation, Washington, D.C.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt. 1978. A model of evolutionary change in proteins. In Atlas of protein sequence and structure, vol. 5, 345–352. National Biomedical Research Foundation, Washington, D.C. Suppl. 3.

Delsuc, F., H. Brinkmann and H. Phillippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Rev. Genet. **6**:361–375.

Dixon, M. T. and D. M. Hillis. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetics analysis. Mol. Biol. Evol. **10**:256–267.

Doherty, E. A. and J. A. Doudna. 1997. The P4-P6 domain directs higher order folding of the Tetrahymena ribozyme core. Biochemistry **36**:3159–3169.

Durbin, R., S. R. Eddy, A. Krogh and G. Mitchison. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, chap. 10. Cambridge University Press, Cambridge, United Kingdom.

Eck, R. V. and M. O. Dayhoff. 1966. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, MD.

Eddy, S. R. 2004. How do RNA folding algorithms work? Nature Biotech. **22**:1457–1458.

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. University of Washington, Seattle.

———. 2004. Inferring phylogenies. Sinauer, Sunderland, MA.

Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science **155**:279–284.

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

Freier, S. M., R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Nielson and D. H. Turner. 1986. Improved free energy parameters for prediction of RNA duplex stability. Proc. Natl. Acad. Sci. USA **83**:9373–9377.

Gardner, P. P. and R. Giegerich. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics **5**:140.

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

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

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

Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. Schuster and P. F. Stadler. 1996. Analysis of RNA sequence structure maps by exhaustive enumeration. Monatsh. Chem. **125**:167–188.

Gutell, R. R. 1996. Comparative sequence analysis and the structure of 16S and 23S RNA. In R. A. Zimmermann and A. E. Dahlberg, eds., Ribosomal RNA: structure, evolution, processing and function in protein biosynthesis, 15–27. CRC Press, Boca Raton, FL.

Gutell, R. R., J. C. Lee and J. J. Cannone. 2002. The accuracy of ribosomal RNA comparative structure models. Curr. Opin. Struct. Biol. **13**:301–310.

Hasegawa, M. and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature **361**:23.

Hasegawa, M., H. Kishino and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

Henikoff, S. and J. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915–10919.

Higgs, P. G. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica **102/103**:91–101.

———. 2000. RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **33**:199–253.

Hofacker, I. L. 2003. Vienna RNA secondary structure server. Nucl. Acids Res. **31**:3429–3431.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**:167–188.

Hofacker, I. L., P. F. Stadler and R. R. Stocsits. 2004. Conserved RNA secondary structures in viral genomes: a survey. Bioinoformatics **20**:1495–1499.

Holder, M. and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature Rev. Genet. **4**:275–284.

Hughes, A. L. and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection. Nature **335**:167–170.

Hwang, D. G. and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA **101**:13994–14001.

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

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

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

———. 1992b. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. In H. N. Munro, ed., Mammalian protein metabolism, 21–32. Academic Press, New York.

Karlin, S. and H. W. Taylor. 1975. A first course in stochastic processes. Academic Press, New York, NY.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, NY.

Kirby, D. A., S. V. Muse and W. Stephan. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92**:9047–9051.

Kishino, H., T. Miyata and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

Kosiol, C. and N. Goldman. 2004. Different versions of the Dayhoff rate matrix. Mol. Biol. Evol. **22**:193–199.

Lanave, C., G. Preparata, C. Saccone and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20**:86–93.

Lee, Y. H., T. Ota and V. D. Vacquier. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12**:231–238.

Liò, P. and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res. **8**:1233–1244.

Lyngsø, R. B., M. Zuker and C. N. S. Pedersen. 1999. Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics **15**:440–445.

Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker and D. H. Turner. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. USA **101**:7287–7292.

Mathews, D. H., J. Sabina, M. Zuker and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**:911–940.

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139**:1429–1439.

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

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

Nussinov, R., G. Pieczenik, J. R. Griggs and D. J. Kleitman. 1978. Algorithm for loop matchings. SIAM J. Appl. Math. **35**:68–82.

Ohta, T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. **40**:56–63.

Pedersen, A.-M. K. and J. L. Jensen. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. **18**:763–776.

Pedersen, J. S., R. Forsberg, I. M. Meyer and J. Hein. 2004. An evolutionary model for protein-coding regions with conserved RNA structure. Mol Biol Evol **21**:1913–1922.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**:1692–1704.

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

Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. Genetics **141**:771–783.

SantaLucia, J., Jr. and D. H. Turner. 1997. Measuring the thermodynamics of RNA secondary structure formation. Biopolymers **44**:309–319.

Savill, N. J., D. C. Hoyle and P. G. Higgs. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157**:399–411.

Schöniger, M. and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3**:240–247.

Seffens, W. and D. Digby. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucl. Acids Res. **27**:1578–1584.

Smith, A. D., T. W. H. Lui and E. R. M. Tillier. 2004. Empirical models for substitution in ribosomal RNA. Mol. Biol. Evol. **21**:419–427.

Stephan, W. 1996. The rate of compensatory evolution. Genetics **144**:419–426.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

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

Tillier, E. R. M. 1994. Maximum likelihood with multi-parameter models of substitution. J. Mol. Evol. **39**:409–417.

Tillier, E. R. M. and R. A. Collins. 1998. High apparent rate of simultaneous compensatory basepair substitutions in ribosomal RNA. Genetics **148**:1993–2002.

Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**:691–699.

Whelan, S., P. Liò and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in Genetics **17**:262–272.

Workman, C. and A. Krogh. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucl. Acids Res. **27**:4816–4822.

Wu, M. and I. Tinoco, Jr. 1998. RNA folding causes secondary structure rearrangement. Proc. Natl. Acad. Sci. USA **95**:11555–11560.

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

———. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

Yang, Z. and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

Yang, Z., R. Nielsen and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15**:1600–1611.

Zuckerkandl, E. and L. Pauling. 1962. Molecular disease, evolution and genetic heterogeneity. In M. Kasha and B. Pullman, eds., Horizons in biochemistry, 189–225. Academic Press, New York, NY.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. Science **244**:48–52.

———. 2000. Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol. **10**:303–310.

Zuker, M. and P. Stiegler. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. **9**:133–148.

Table 1.1: Instantaneous rate matrices of mononucleotide substitution models

JC69
$$Q = \begin{pmatrix} . & \beta & \beta & \beta \\ \beta & . & \beta & \beta \\ \beta & \beta & . & \beta \\ \beta & \beta & \beta & . \end{pmatrix}$$

K80
$$Q = \begin{pmatrix} . & \beta & \kappa\beta & \beta \\ \beta & . & \beta & \kappa\beta \\ \kappa\beta & \beta & . & \beta \\ \beta & \kappa\beta & \beta & . \end{pmatrix}$$

F81
$$Q = \begin{pmatrix} . & \pi_C & \pi_G & \pi_T \\ \pi_A & . & \pi_G & \pi_T \\ \pi_A & \pi_C & . & \pi_T \\ \pi_A & \pi_C & \pi_G & . \end{pmatrix}$$

HKY85
$$Q = \begin{pmatrix} . & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & . & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & . & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & . \end{pmatrix}$$

TN93
$$Q = \begin{pmatrix} . & \pi_C & \kappa_1\pi_G & \pi_T \\ \pi_A & . & \pi_G & \kappa_2\pi_T \\ \kappa_1\pi_A & \pi_C & . & \pi_T \\ \pi_A & \kappa_2\pi_C & \pi_G & . \end{pmatrix}$$

REV(GTR)
$$Q = \begin{pmatrix} . & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & . & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & . & \pi_T \\ c\pi_A & e\pi_C & \pi_G & . \end{pmatrix}$$

UNREST
$$Q = \begin{pmatrix} . & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & . & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & . & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & . \end{pmatrix} = \begin{pmatrix} . & a & b & c \\ d & . & e & f \\ g & h & . & i \\ j & k & l & . \end{pmatrix}$$

For diagonal elements, $q_{ii} = -\sum_{i \neq j} q_{ij}$

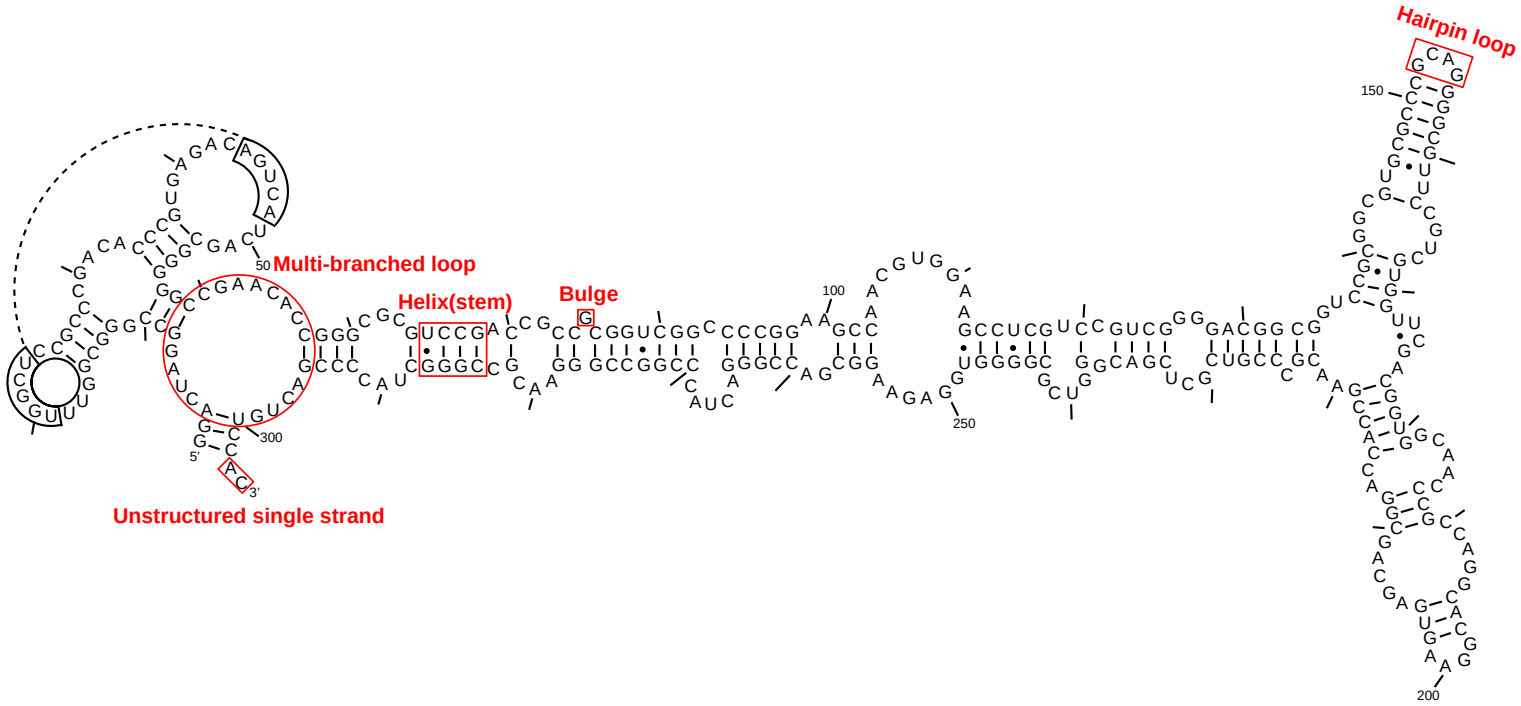


Figure 1.1: SRP RNA of Halobacterium halobium

Chapter 2

Testing for spatial clustering of amino acid replacements within protein tertiary structure

Jiaye Yu and Jeffrey L. Thorne[1]

Keywords: Protein tertiary structure, Protein evolution, Spatial clustering

Accepted by Journal of Molecular Evolution

[1]Corresponding author; *Address*: Campus Box 7566, North Carolina State University, Raleigh, NC 27695-7566 *Phone*: (919)515-1946 *Fax*: (919)515-7315 *Email*: thorne@statgen.ncsu.edu

2.1 Abstract

Widely used models of protein evolution ignore protein structure. Therefore, these models do not predict spatial clustering of amino acid replacements with respect to tertiary structure. One formal and biologically implausible possibility is that there is no tendency for amino acid replacements to be spatially clustered during evolution. An alternative to this is that amino acid replacements are spatially clustered and this spatial clustering can be fully explained by a tendency for similar rates of amino acid replacement at sites that are nearby in protein tertiary structure. A third possibility is that the amount of clustering exceeds that which can be explained solely on the basis of independently evolving protein sites with spatially clustered replacement rates. We introduce two simple and not very parametric hypothesis tests that help distinguish these three possibilities. We then apply these tests to 273 homologous protein families. The null hypothesis of no spatial clustering is rejected for 112 of 273 families. The explanation of spatially clustered rates but independent change among sites is rejected for 47 families. These findings need to be reconciled with the common practice of basing evolutionary inferences on models that assume independent change among sites.

2.2 Introduction

Protein tertiary structures evolve more slowly than do protein sequences (e.g., Chothia and Lesk 1986; Flores et al. 1993). Experimental studies (e.g., Oosawa and Simon 1986; Lim et al. 1992) have demonstrated the importance of interactions between residues on protein stability. Therefore, there is no biological basis for expecting that protein sites evolve independently.

Unfortunately, widely used models of protein evolution assume that amino acids (e.g., Dayhoff et al. 1978; Jones et al. 1992; Adachi and Hasegawa 1996) or codons (e.g., Goldman and Yang 1994; Muse and Gaut 1994) change independently. This assumption is computationally convenient but biologically flawed. Although no evolutionary model is completely realistic, the impact of biologically implausible assumptions on evolutionary inference needs to be characterized. As a first step, a biologically implausible assumption can be treated as a null hypothesis and can then be tested to determine whether the amount of information in a data set allows the null hypothesis to be rejected in favor of a more realistic alternative. While the failure to reject a null hypothesis does not necessarily indicate it is true, hypothesis test results can reflect the amount of pertinent information in a data set.

Diverse techniques for identifying coevolution or covariation between sites in a protein have been previously proposed. For example, Shindyalov et al. (1994) introduced a method to detect pairs of protein sites that experienced correlated amino acid replacements over time. After studies on multiple protein families, they found a weak but significant tendency for the correlated residue pairs to be close to each other in tertiary structure.

Pollock et al. (1999) subsequently adopted a likelihood-based method to identify

the coevolving protein sites. In their approach, the twenty possible amino acids at a site were converted into a simpler two-state system. A likelihood ratio test was then used to assess whether pairs of sites were undergoing independent or dependent replacements. Applying their method to tetrapod myoglobin sequences, they found a tendency for replacement aggregation in tertiary structure.

Coevolving sites need not be near one another in the protein structure, but it is reasonable to expect these sites to be spatially grouped. Our goal here is not to identify specific pairs of covarying sites. The goal is instead to examine whether sites that are near one another in a protein tertiary structure have similar and possibly non-independent patterns of change.

Some previous studies support the expectation that protein structure impacts patterns of amino acid replacement. For example, Dean and Golding (2000) developed a maximum likelihood method to identify regions of a protein that experience amino acid replacements at a particularly high or low rate. They concluded that solvent accessibility and distance from the catalytic site explain most of the rate variation among sites in eubacterial isocitrate dehydrogenases. Also, Goldman and his collaborators estimated in globular proteins that sites which are relatively exposed to solvent change at about twice the rate of sites that are more inaccessible to solvent (Goldman et al. 1998). Because the amounts of solvent accessibility at nearby protein sites are positively correlated, these previous studies indicate a spatial structuring of evolutionary rates among sites.

Here we introduce a simple test of the null hypothesis that there is no spatial element to the distribution of amino acid replacements. Perhaps more interestingly, we introduce a hypothesis test to examine whether spatial heterogeneity of rates among

independently evolving sites is sufficient to explain the pattern of amino acid replacements. An alternative is that spatial heterogeneity of replacement rates may or may not exist but that changes at nearby protein sites do not always occur independently. The two hypothesis tests that we introduce here were both designed with the intent that they not be very parametric. Highly parametric model-based treatments of evolutionary dependence among sites are increasingly available (e.g., Pedersen and Jensen 2001; Robinson et al. 2003; Siepel and Haussler 2004; Hwang and Green 2004; Rodrigue et al. 2005) and these treatments have many advantages. However, highly parametric treatments can sometimes be misleading due to sensitivity to violations of assumptions. More parametric procedures would be likely to have higher power for rejecting null hypotheses but we opt here for less reliance upon assumptions. We view the not very parametric approaches described here as being complementary to more parametric and potentially more powerful approaches.

2.3 Methods

2.3.1 The Irrelevant Structure Hypothesis

Consider the formal possibility that we term the Irrelevant Structure (IS) hypothesis,

$A0$: There is no spatial clustering among substitution events in protein sequence and there is homogeneity of replacement processes among sites. The alternative to the IS null hypothesis is,

$A1$: There is spatial clustering among substitution events. To examine the IS null hypothesis, we design a criterion to measure spatial clustering of amino acid replacements on protein tertiary structure. Our criterion is simply one of many

reasonable assessments of amino acid replacement aggregation.

To explain our criterion, assume that there is a known evolutionary tree topology relating all protein sequences being analyzed. Imagine also that each amino acid replacement that occurred during sequence divergence has been identified and mapped to the specific branch of the known evolutionary tree topology on which it occurred. Uncertainty about which amino acid replacements have occurred and on which branch they have occurred is an important practical issue that we consider below. For now, assume the uncertainty does not exist.

We focus on data sets of aligned and homologous amino acid sequences that have a protein family member with experimentally determined tertiary structure. Because tertiary structure changes much more slowly over evolutionary time than does protein sequence (e.g., Chothia and Lesk 1986; Flores et al. 1993), we assume that all aligned protein sequences share the same experimentally determined structure. Therefore, the amino acid replacements that are mapped to specific branches of the evolutionary tree can also be mapped to the protein tertiary structure and can then be used to summarize the amount of spatial clustering of amino acid replacements on each branch. The assumption of a "frozen" tertiary structure over time is formally incorrect because tertiary structure does evolve. However, considering the much slower rate of structural evolution and complexities that would arise if structural change over time was incorporated, the frozen assumption seems to be a good starting point.

For each amino acid replacement on a branch of the phylogeny, we define a sphere with a 10Å radius centered on the α-carbon atom of the amino acid in the known protein structure. All amino acids with an α-carbon atom within the 10Å sphere are considered to be in the "neighborhood" of the amino acid at the center of the

sphere. Although the 10Å neighborhood definition has yielded successful protein folding recognition procedures (e.g., Jones et al. 1992), it is somewhat arbitrary. Other ways to define neighborhoods of protein sites have been proposed (e.g., Larson et al. 2000; Pritchard et al. 2001) and could be adapted to our procedure. Our next step is to count on each branch the number of sites within each 10Å neighborhood that have experienced a replacement. The central site of the sphere is not included in this count so that, if the only replacement within this sphere is at the central site, the count is zero. Only if there are amino acid replacements elsewhere in the sphere that are on the same branch will the count for the sphere exceed zero. As an example, if there are four amino acid replacements along a branch then there will be four 10Å balls for that branch. Imagine that the numbers of amino acid replacements within the four balls (excluding the centers) are 2, 2, 2 and 0. In this case, the average number of replacements within a ball would be 1.5.

For branch i, define R_i as the total number of amino acid replacements on the branch and N_i as the average number of replacements among the R_i 10Å balls corresponding to the branch. Our testing strategy measures how unusual it would be to observe the value of N_i if both the null hypothesis were true and R_i replacements occurred upon branch i. The strategy then combines these measures across branches.

The pattern of amino acid replacements affecting a protein family can be summarized as a matrix with the number of rows equal to the total number of branches on the tree and the number of columns equal to the number of sites in the protein. The matrix has a 1 in row i of column j if site j is different at the beginning and ending of the i^{th} branch, and has a 0 otherwise. We refer to this binary matrix as the replacement matrix.

The distribution of N_i conditional upon R_i under the IS null hypothesis can be simulated by permuting the entries of row i in the replacement matrix. A permuted row will have the same row total as the original matrix but the actual entries in the permuted row can differ from the row entries in the original matrix. The average number of replacements per 10Å ball when branch i is permuted can be calculated. Because the permutation procedure will be performed many times to approximate the distribution of N_i under the null hypothesis, we use N_i^t to be the average number of replacements per ball on branch i for the t^{th} permuted data set. With a total of T permuted matrices, we can calculate the sample mean for branch i

$$\overline{N_{i(S)}} = \frac{\sum_{t=1}^{T} N_i^t}{T} \tag{2.1}$$

where S denotes simulated (permuted). We can also determine the sample variance

$$\widetilde{\sigma_{i(S)}}^2 = \sum_{t=1}^{T} \frac{(N_i^t - \overline{N_{i(S)}})^2}{T-1}. \tag{2.2}$$

This allows the value of N_i to be normalized,

$$Z_i = \frac{N_i - \overline{N_{i(S)}}}{\widetilde{\sigma_{i(S)}}}. \tag{2.3}$$

If the IS null hypothesis is correct, values of Z_i will tend to be near 0. If the alternative $A1$ hypothesis of spatial clustering is instead correct, values of Z_i will tend to be positive. The biologically implausible scenario where replacements are over-dispersed in space would lead to generally negative values of Z_i. After discarding branches on which less than 2 replacements occur or for which $\widetilde{\sigma_{i(S)}} = 0$, assume there are I

remaining branches. To summarize the Z_i for all I branches, we concentrate on the summary statistic

$$Z = \frac{\sum_{i=1}^{I} Z_i}{\sqrt{I}}. \tag{2.4}$$

To determine whether the observed value of Z warrants rejection of the IS null hypothesis, we need to know what the distribution of Z would be according to this null hypothesis. Rather than relying on any assumption of a specific form for the null distribution of Z (e.g., a standard normal distribution), we approximate this distribution. The approximation is straightforward to obtain by calculating Z as in the procedure described above, except that N_i^t values from permutations are substituted for N_i in Equation 2.3. This generates T simulated values of Z under the IS null hypothesis. When comparing the IS null hypothesis with the alternative $A1$, a p-value can be approximated by adding 1 to the number of simulated Z values that exceed the observed value and by then dividing by T+1.

2.3.2 The Structural Independence Among Sites hypothesis

To further investigate any relationship between protein structure and evolution, we compare two more refined hypotheses. These hypotheses concern whether spatially organized rate heterogeneity is sufficient to explain the pattern of amino acid replacements on tertiary structure. With this comparison, the null hypothesis allows different sites to have different rates of evolution and it allows this rate heterogeneity to be spatially organized, but it assumes amino acid replacements occur independently. The null hypothesis is,

B0: The spatial clustering of substitution events can be fully explained by the

clustering of substitution rates. We will refer to null hypothesis $B0$ as the Structural Independence Among Sites (SIAS) hypothesis. The SIAS null hypothesis is violated if some amino acid replacements are compensatory or occur in some other dependent fashion. The alternative to the SIAS null hypothesis is,

$B1$: Spatial clustering cannot be solely explained by clustered substitution rates among independently evolving sites.

Our test of the null hypothesis $B0$ uses the test statistic Z that is defined in Equation 2.4. For now, we again assume the replacement matrix is known with certainty. The difference between the IS and SIAS hypothesis tests is the way in which the null distribution of Z is approximated. For the IS hypothesis test, entries within rows of the inferred replacement matrix are permuted so as to maintain row totals. For the test of the SIAS null hypothesis, we only consider simulated matrices where both the row and column totals are identical to those in the inferred replacement matrix. Column totals are fixed so that quickly evolving sites in the observed data are quickly evolving in the permuted data while slowly evolving sites in the observed data are slowly evolving in the permuted data. The permutation procedure thereby accounts for rate heterogeneity among sites.

To generate simulated matrices with the same row and column totals as the actual replacement matrix, we employ the permutation procedure of Roff and Bentzen (1989). In the actual replacement matrix, all entries will be either 0 or 1. In contrast, the computationally convenient Roff and Bentzen procedure only constrains row and column totals. Ideally for our purposes, all entries in the simulated matrices would be either 0 or 1. Unfortunately, it is possible with this Roff and Bentzen permutation procedure to generate matrices with one or more entries that exceed 1. Our *ad*

hoc treatment for an entry in row i and column j of a permuted matrix that has some value $K > 1$ is to center K spheres about site j for branch i. The number of replacements within each of these K 10Å balls is set to K-1 plus the sum of entries corresponding to the other protein sites on branch i that are within 10Å of site j. Other treatments of entries that exceed 1 are possible, but we selected this one because of its conservative nature. We only reject the SIAS null hypothesis for large values of Z and this treatment serves to shift the null distribution of Z toward higher values than if entries in simulated matrices were not allowed to exceed one.

2.3.3 Inferred replacement matrix

Although we assume the sequence data are aligned and related by a known evolutionary tree topology, we have not yet described a procedure for constructing the inferred replacement matrix. A simple way to infer replacements and map them to branches on the phylogeny is to use the parsimony criterion (e.g., see Felsenstein 2004). Parsimony has the advantages of being intuitive and straightforward, but it does not resolve the issue of which of many possibly equally parsimonious reconstructions to choose. In addition, parsimony does not account for the fact that the true history may not be the most parsimonious. Rather than parsimony, we prefer to jointly sample sets of ancestral sequences according to their probability density conditional upon the observed sequences, the estimated branch lengths and topology, and values of parameters that define a simple model of amino acid replacement.

Our procedure for stochastically sampling sets of ancestral sequences is a slight modification of the algorithm for ancestral sequence reconstruction by Pupko et al. (2000). The unmodified Pupko 2000 algorithm was designed to find the optimal set

of ancestral sequences whereas our modification randomly samples a set according to its probability. The details of our modification are described in the Appendix.

An important assumption of the original Pupko 2000 algorithm and our modification is that all sites evolve independently and identically. Although Pupko and collaborators introduced a subsequent algorithm that relaxes the assumption of identical processes among sites but maintains the independence assumption (Pupko et al. 2002), here we consider only the stochastic version of their earlier algorithm (Pupko et al. 2000). With the assumption that sites evolve independently, ancestral sequences can be jointly reconstructed by successive joint reconstruction at individual alignment columns.

To incorporate the uncertainty of ancestral sequences, we base our hypothesis test on a large number of sets (e.g., J=1000) of ancestral sequence reconstructions. This corrects the artifacts that would be introduced by using a single most parsimonious reconstruction. Let $Z_i^{(j)}$ be the statistic defined in Equation 2.3 for branch i with the j^{th} sampled set of ancestral sequences. Then,

$$Z^{(j)} = \frac{\sum_{i=1}^{I_j} Z_i^{(j)}}{\sqrt{I_j}}, \tag{2.5}$$

where I_j is the number of branches with at least 2 changes for the j^{th} sampled set of ancestral sequences. Conditional upon each reconstruction, we can estimate a p-value $p^{(j)}$ as described earlier for the case where ancestral sequence uncertainty is neglected. An unconditional estimate of the p-value is

$$\tilde{p} = \frac{\sum_{j=1}^{J} p^{(j)}}{J}. \tag{2.6}$$

To account for uncertainty of ancestral sequences when testing hypotheses, we employ $\tilde{p}$ rather than the conditional p-value estimates. Because ancestral sequences are sampled by assuming a model with independent amino acid replacements among sites, our approach is not completely nonparametric. An effect of the independence among sites assumption might be to yield too little evidence for spatial clustering due to dependent changes. The magnitude of this effect would increase as the amount of uncertainty about ancestral sequences rises.

2.3.4 Examples

We applied our two hypothesis tests to 273 protein families. To select these families, non-homologous single chain proteins with experimentally determined structures were identified via the PDB_SELECT database (Hobohm and Sander 1994, updated October 2004; Berman et al. 2000). For each protein structure, aligned and homologous amino acid sequences were taken from the HSSP database (Sander and Schneider 1991). The protein sequence in each family that has the experimentally determined reference structure will be denoted the "master" sequence. In each alignment, all sequences with low weighted similarity (<60%) or long indels (>20%) compared to the master sequence were discarded. The weighted similarities were determined according to Sander and Schneider (1991). Following removal of the sequences with long indels or low weighted similarity relative to the master sequence, all protein families that had less than 4 or more than 200 remaining sequences were eliminated from further analysis.

For each aligned protein family, pairwise maximum likelihood distances were inferred via the software TREE-PUZZLE (Schmidt et al. 2002) and the WAG model

of amino acid replacement (Whelan and Goldman 2001). Next, these distances were the basis for inferring a neighbor-joining tree (Saitou and Nei 1987) with the NEIGHBOR program of the PHYLIP software package (Felsenstein 1993). On the inferred topology, we used the software PAML (Yang 1997) to obtain maximum likelihood estimates of branch lengths for the WAG model of amino acid replacement and then jointly sampled sets of ancestral sequences via the modified Pupko algorithm under the same WAG model (see Appendix). For each protein family analyzed, $J = 1000$ sets of ancestral sequences were sampled. We also inferred sets of ancestral sequences via parsimony with the ACCTRAN setting of software package PAUP* (Swofford 2002) and then performed the IS and SIAS tests using the resulting set of optimal ancestral sequences. We elected to also employ parsimony so that we could contrast results obtained with a single set of ancestral sequences to those obtained by accounting for uncertainty in ancestral sequence reconstruction.

2.4 Results

A histogram of the p-values obtained by testing the IS null hypothesis on 273 protein families is shown in Figure 2.1. The IS null hypothesis was rejected at a significance level of 0.05 for 112 of 273 protein families. If the IS null hypothesis was actually true for all 273 families and if the test was not conservative, the number of families for which the IS null hypothesis could be rejected at significance level 0.05 would be a realization from a binomial distribution with parameters 273 and 0.05. The expected number of families for which the IS null hypothesis could be rejected would then be 13.65. The probability that such a binomial distribution would yield 112 or

more rejections of the IS null hypothesis is less than 10^{-70}. Because the hypothesis test is conservative, we should actually expect fewer than 13.65 rejections if the null hypothesis were true for all protein families.

To characterize which sorts of protein families lead to rejection of the IS null hypothesis, we calculated a rough measure of the amount of evolutionary information in each data set. We did this by estimating the expected number of replacements per protein site (i.e., the sum of the branch lengths) and then multiplying this by the number of sites per protein. We assess the amount of evolution on a per data set basis rather than on a per site basis because long protein sequences have more potential to display spatial clustering patterns than do short sequences. Figure 2.2 shows a plot for the 273 protein families of the logarithm of this measure versus the logarithm of the $\tilde{p}$-values for testing the IS null hypothesis.

The histogram of $\tilde{p}$-values obtained by testing the SIAS null hypothesis is shown in Figure 2.3. The SIAS null hypothesis was rejected at a significance level of 0.05 for 47 of 273 protein families. If the SIAS null hypothesis was true for all 273 protein families and if the test were not conservative, the probability of rejecting it for 47 or more protein families would be less than 10^{-12}. As with the IS test, the SIAS test is conservative and this would make finding 47 significant results even more unlikely if the null hypothesis were true for all protein families. Figure 2.4 is a counterpart to Figure 2.2 but it contrasts the amount of evolution to the p-value estimates obtained by testing the SIAS null hypothesis.

One possibility is that certain biological attributes are associated with protein families for which the IS or SIAS null hypotheses could be rejected. To explore this, we used the gene ontology annotation database (Camon et al. 2004) to connect PDB

entries of each of the 273 protein families with information in the gene ontology (i.e., GO) database (The Gene Ontology Consortium 2000). The categories *biological process*, *cellular component* and *molecular function* are at the root of the hierarchically organized GO database. These three categories have been further divided into 28 subcategories. A protein family can be a member of more than one of these 28 subcategories. For 7 of the 28 subcategories, none of the 273 protein families are members. For the remaining 21 subcategories, the Wilcoxon rank sum test (e.g. Mood et al. 1974) was applied to test whether estimated p-values from the IS test were correlated with membership in the subcategory. The null hypothesis is that the $\tilde{p}$-values are independent of whether a protein family belongs to the subcategory being examined. The two-tailed version of the test was adopted because it seems plausible that a given biological attribute could be associated with either unusually high or unusually low amount of spatial clustering of replacements. The Wilcoxon rank sum test was also applied to $\tilde{p}$-values from the SIAS test for each of the 21 subcategories. Results are shown in Table 2.1.

2.5 Discussion

Figure 2.2 and Figure 2.4 show that there is a rough negative correlation between the $\tilde{p}$-value for a protein family and the amount of evolution separating sequences in a data set. This general pattern holds for both the IS and SIAS tests. The amount of evolution is not a perfect surrogate for the amount of potential spatial clustering in a data set. Excessive amounts of evolution may make it difficult to confidently map specific amino acid replacements to specific branches on an evolutionary tree.

Nevertheless, these figures are consistent with the notion that the null hypotheses are false for many data sets but that they were not rejected due to limited power of the hypothesis tests.

At a significance level of 0.05, the biological subcategory attributes 'virion' and 'antioxidant activity' were associated with spatial clustering for both the IS and SIAS tests (see Table 2.1). However, the IS and SIAS tests were each performed for 21 different subcategories. In light of these multiple tests, we cannot conclude there is evidence of spatial clustering being strongly associated with any of the subcategories. This lack of evidence for association of spatial clustering with any particular biological attribute implies that spatial clustering and evolutionary dependence are ubiquitous phenomena in protein evolution.

Generally, there is a tendency that the SIAS null hypothesis is harder to reject than the IS null hypothesis (Figure 2.5). This is reasonable because the SIAS null hypothesis is more general than the IS null hypothesis. The SIAS null hypothesis allows the possibility of spatial clustering of replacements simply due to the spatial clustering of substitution rates. Although we devised the SIAS test with the purpose of detecting evolutionary dependence among sites, rejection of the SIAS null hypothesis does not indicate that there is a causal relationship among replacement events. It could be that some force external to a protein such as natural selection causes the substitution rates of all positions in certain regions in the protein to all simultaneously increase or decrease in some lineage. The SIAS null hypothesis does not permit region-specific variation of evolutionary rates over time.

We studied the general effects of ancestral sequence reconstruction methods on our hypothesis tests. The logarithm of 273 $\tilde{p}$-values for the IS and SIAS hypothesis tests

with parsimony and probabilistic ancestral sequence reconstructions were averaged, respectively. The results are summarized in Table 2.2. As expected, the probabilistic method, taking uncertainty into account, makes the average logarithm of the $\tilde{p}$-value slightly larger than the logarithm of the $\tilde{p}$-values we obtained from parsimony method. Table 2.2 is consistent with the SIAS null hypothesis being harder to reject than the IS null hypothesis.

We found ample evidence for spatial clustering of amino acid replacements, but we do not find this evidence for all protein families. The not very parametric nature of our hypothesis tests undoubtedly reduces their power. Nevertheless, our results from analyzing 273 protein families indicate that protein structure does play a role in protein evolution. Unfortunately, possible effects of protein structure are ignored by widely used models for studying sequence change. Recently, there has been progress on explicit evolutionary models that incorporate protein structure and evolution (e.g., Robinson et al. 2003; Rodrigue et al. 2005), but procedures useful for computationally intensive tasks such as phylogeny inference remain unavailable. For these intensive tasks, models that reflect protein structure but that have the computationally attractive assumption of independent change among sites (e.g., Fornasari et al. 2002) are particularly appealing.

2.6 Appendix

In order to jointly sample a combination of amino acids for all internal nodes, an algorithm slightly modified from Pupko et al. (2000) is used. The original version of Pupko algorithm was developed for computationally efficient maximum-likelihood

reconstruction of all ancestral amino acid sequences in a given phylogenetic tree. Here, we make a probabilistic modification to sample reconstructions. The algorithm is presented as if only the reconstructed nucleotides at a single sequence position were of interest. For independently evolving sequence positions, the entire set of sampled ancestral sequences is obtained by successively applying the modified algorithm to individual sequence positions (i.e., alignment columns).

To help explain the modified algorithm, some notation is introduced. On an unrooted tree, an interior node R is selected to be the root. For any other node N on the arbitrarily rooted tree, its parent node will be denoted by N_p. For a particular alignment column of interest, S_N represents the residue type at node N and S_N^D represents the collection of residues at the descendant nodes of N. If N is a tip node, S_N^D represents the empty set. The transition probability of residue type j at the end of a branch given that type i occupies the site at the beginning of the branch is $p_{ij}(t)$ where t represents the expected amount of evolution on the branch. To calculate the transition probability, a simple amino acid replacement model with independent changes among sites (e.g., Whelan and Goldman 2001) is assumed.

The simple tree in Figure 2.6 is used to further illustrate the modified Pupko algorithm. In this tree, nodes X and Y are children of node N. With the modified Pupko algorithm, we need to be able to calculate $P_N(j,i) = Pr(S_N = j, S_N^D = s_N^D | S_{N_p} = i)$ for all non-root nodes N and for all combinations i and j of residue types. If N is a tip node, $P_N(j,i) = P_{ij}(t_N)$. For an interior node N that is not the

root,

$$
\begin{aligned}
P_N(j,i) &= Pr(S_N = j, S_N^D = s_N^D | S_{N_p} = i) \\
&= Pr(S_N = j, S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_{N_p} = i) \\
&= Pr(S_N = j | S_{N_p} = i) \\
&\quad \times Pr(S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_N = j, S_{N_p} = i) \\
&= p_{ij}(t_N) \times Pr(S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_N = j, S_{N_p} = i) \\
&= p_{ij}(t_N) \times Pr(S_X = k, S_X^D = s_X^D | S_N = j) \times Pr(S_Y = l, S_Y^D = s_Y^D | S_N = j) \\
&= p_{ij}(t_N) \times P_X(k,j) \times P_Y(l,j).
\end{aligned}
\tag{2.7}
$$

Now we consider the probability of sampling residue type j at node N. For a bifurcating tree, this only depends on the residue types at nodes N_p, X, and Y. This

probability is given by

$$
\begin{aligned}
&Pr(S_N = j | S_X = k, S_Y = l, S_{N_p} = i) \\
&= Pr(S_N = j | S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D, S_{N_p} = i) \\
&= \frac{Pr(S_N = j, S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_{N_p} = i) \times Pr(S_{N_p} = i)}{Pr(S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_{N_p} = i) \times Pr(S_{N_p} = i)} \\
&= \frac{Pr(S_N = j, S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_{N_p} = i)}{\sum_{m=1}^{20} Pr(S_N = m, S_X = k, S_Y = l, S_X^D = s_X^D, S_Y^D = s_Y^D | S_{N_p} = i)} \\
&= \frac{p_{ij}(t_N) \times Pr(S_X = k, S_X^D = s_X^D | S_N = j) \times Pr(S_Y = l, S_Y^D = s_Y^D | S_N = j)}{\sum_{m=1}^{20} p_{im}(t_N) \times Pr(S_X = k, S_X^D = s_X^D | S_N = m) \times Pr(S_Y = l, S_Y^D = s_Y^D | S_N = m)} \\
&= \frac{p_{ij}(t_N) \times P_X(k, j) \times P_Y(l, j)}{\sum_{m=1}^{20} p_{im}(t_N) \times P_X(k, m) \times P_Y(l, m)} \\
&= \frac{P_N(j, i)}{\sum_{m=1}^{20} P_N(m, i)}.
\end{aligned}
\tag{2.8}
$$

If node N is the root node R with children X, Y and Z, the above equation is instead

$$
\begin{aligned}
&Pr(S_R = j | S_X = k, S_Y = l, S_Z = n) \\
&= \frac{\pi_j \times P_X(k, j) \times P_Y(l, j) \times P_Z(n, j)}{\sum_{m=1}^{20} \pi_m \times P_X(k, m) \times P_Y(l, m) \times P_Z(n, m)}.
\end{aligned}
\tag{2.9}
$$

Following the Pupko strategy (Pupko et al. 2000), the modified algorithm visits

all nodes on the rooted tree. The tip nodes are visited first. Next, non-root interior nodes are visited if they have not yet been visited but all of their offspring nodes have already been visited. When all non-root nodes have been visited, the root node is visited.

At each non-root node N that is visited, we set two kinds of quantities. The first type of quantity is denoted $C_N(i)$ and its value is one of the possible residue types. If $C_N(i) = j$, this means that node N will be assigned residue type j in the reconstructed set of residues if its parental node N_p is assigned residue type i. When the values of $C_N(i)$ have been assigned for all non-root nodes N, then the entire reconstructed set of residues at ancestral nodes on the tree can be determined simply by knowing the residue type of the root node. The other quantities that are determined at each node N that is visited are the $P_N(j, i)$ terms, as defined in Equation (2.7). These quantities are necessary for randomly sampling the value of $C_N(i)$ with the appropriate probability.

When a tip node T is visited, we set $P_T(j, i) = p_{ij}(t_T)$ where j is the observed residue type at node T. We also set $C_T(i) = j$ because residue type j is directly observed at tip node T. This is done for all residue types i.

When an interior non-root node Z is visited, $P_Z(j, i)$ is calculated by applying Equation (2.7) for each possible combination of i and j. Then, for each possible residue type i, we randomly pick a residue type by using Equation (2.8) to sample from $Pr(S_Z = j|S_X = C_X(j), S_Y = C_Y(j), S_{Z_p} = i)$. For the selected residue type j, we set $C_Z(i) = j$.

For the root node R, we choose its residue type S_R by using Equation (2.9) to sample from $Pr(S_R = j|S_X = C_X(j), S_Y = C_Y(j), S_Z = C_Z(j))$. As stated above,

when the values of $C_N(j)$ have been defined for all non-root nodes N and the residue type of the root node has been sampled, the entire set of ancestral residues at the internal nodes has been determined.

2.7 Acknowledgments

We thank Stéphane Aris-Brosou, Steffen Heber, Hirohisa Kishino, Tae-Kun Seo and Dmitri Zaykin for their help. This research was supported by N.S.F grants D.E.B.-0089745, D.E.B.-0120635, and D.E.B-O445180.

2.8 References

Adachi J, Hasegawa M (1996) Models of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459–468

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucl Acids Res 28:235–242

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucl Acids Res 32:D262–266

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:519–527

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In Atlas of protein sequence and structure, volume 5, 345–352. National Biomedical Research Foundation, Washington, D.C.

Dean AM, Golding GB (2000) Enzyme evolution explained (sort of). In Pacific Bioinformatics Symposium 2000

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. University of Washington, Seattle

Felsenstein J (2004) Inferring phylogenies. Sinauer, Sunderland, MA

Flores TP, Orengo CA, Moss DS, Thornton JM (1993) Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 2:1811–1826

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

Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458

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

Hobohm U, Sander C (1994) Enlarged representative set of protein structures. Protein Sci 3:522–524

Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101:13994–14001

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Larson SM, Nardo AAD, Davidson AR (2000) Analysis of covariation in an sh3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 303:433–446

Lim WA, Farruggio DC, Sauer RT (1992) Structural and energetic consequences of disruptive mutations in a protein core. Biochemistry 31:4324–4333

Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics, chapter XI. McGraw-Hiss, 3rd edition

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. Mol Biol Evol 11:715–724

Oosawa K, Simon M (1986) Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in escherichia coli. Proc Natl Acad Sci USA 83:6930–6934

Pedersen AMK, Jensen JL (2001) A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol 18:763–776

Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287:187–198

Pritchard L, Bladon P, Mitchell J, Dufton M (2001) Evaluation of a novel method for the identification of coevolving proteins residues. Protein Eng 14:549–555

Pupko T, Pe'er I, Graur D, Hasegawa M, Friedman N (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. Bioinformatics 18:1116–1123

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

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein

evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692–1704

Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217

Roff DA, Bentzen P (1989) The statistical analysis of mitochondrial DNA polymorphisms - Chi Square and the problem of small samples. Mol Biol Evol 6:539–545

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7:349–358

Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468–488

Swofford D (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sinauer Associates

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genet 25:25–29

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13:555–556

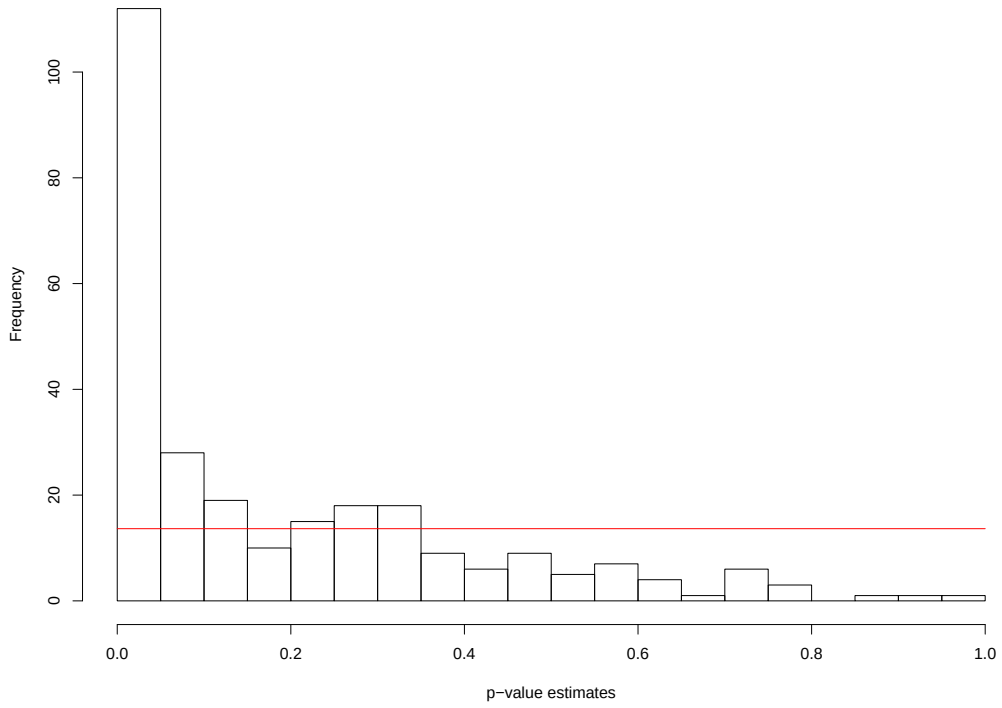
Table 2.1: Wilcoxon rank sum tests for second level GO terms

Ontology ID	GO Subcategory	GO Category	$P_{(IS)}$	$P_{(SIAS)}$	Partition
GO:0005623	cell	C	0.1009	0.1184	112,161
GO:0008372	unknown	C	0.2574	0.0323	4,269
GO:0031012	extracellular matrix	C	0.3709	0.7692	7,266
GO:0005576	extracellular region	C	0.6967	0.3203	30,243
GO:0043226	organelle	C	0.0572	0.0511	49,224
GO:0043234	protein complex	C	0.2087	0.0720	18,255
GO:0019012	virion	C	0.0318	0.0280	4,269
GO:0016209	antioxidant activity	F	0.0248	0.0372	5,268
GO:0005488	binding	F	0.1685	0.0938	145,128
GO:0003824	catalytic activity	F	0.6654	0.6598	133,140
GO:0030234	enzyme regulator activity	F	0.8320	0.9675	16,257
GO:0005554	unknown	F	0.4270	0.3735	4,269
GO:0004871	signal transducer activity	F	0.0687	0.5159	24,249
GO:0005198	structural molecule activity	F	0.4659	0.5915	12,261
GO:0030528	transcription regulator activity	F	0.0618	0.0980	12,261
GO:0005215	transporter activity	F	0.3442	0.3211	43,230
GO:0007610	behavior	P	0.7532	0.9067	5,268
GO:0009987	cellular process	P	0.9918	0.1692	198,75
GO:0007275	development	P	0.7117	0.7648	24,249
GO:0007582	physiological process	P	0.2873	0.9525	221,52
GO:0050789	regulation of biological process	P	0.1797	0.1039	33,240

Categories: Cellular Component (C), Biological Process (P) and Molecular Function (F). $P_{(IS)}$ and $P_{(SIAS)}$ represent the p-values of the Wilcoxon rank sum tests that are evaluating whether the IS or SIAS test results are associated with membership of a protein family in a GO subcategory. The column labeled 'Partition' lists the number of protein families that are and that are not associated with the GO subcategory.

Table 2.2: Average logarithm of $\tilde{p}$-values for IS and SIAS hypothesis tests with ancestral sequences reconstructed by parsimony and probabilistic methods

$\overline{log(\tilde{p})}$	Parsimony	Probabilistic
IS	-3.3525	-2.9459
SIAS	-2.0987	-1.6993



The horizontal line of $y = 13.65$ represents the expected number of protein families that would be rejected by chance. It also indicates the expected number of observations in each bin of the histogram, at a significance level of 0.05.

Figure 2.1: Histogram of $\tilde{p}$ when testing the IS hypothesis

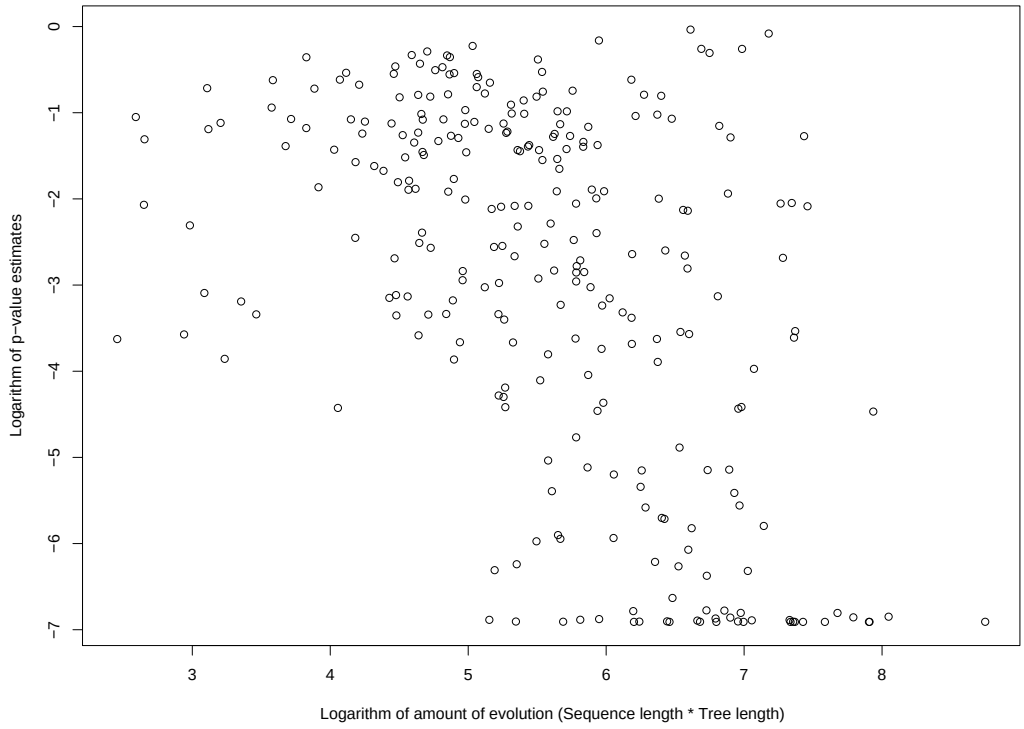
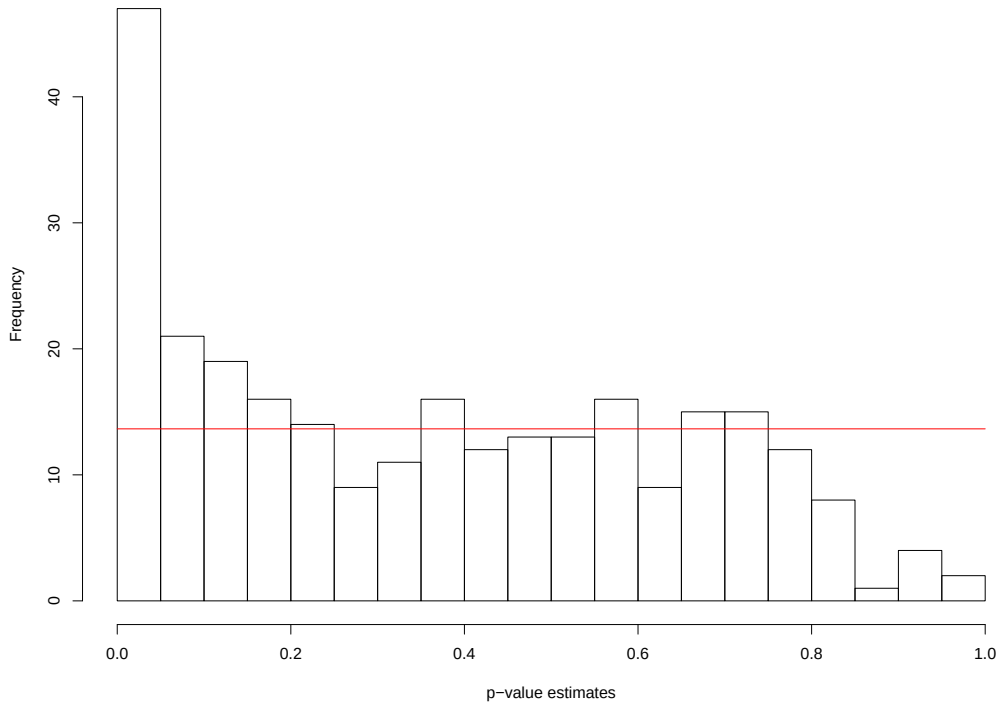


Figure 2.2: Comparison of the logarithm of $\tilde{p}$ with the logarithm of the amount of evolution when testing the IS hypothesis



The horizontal line of $y = 13.65$ represents the expected number of protein families that would be rejected by chance at a significance level of 0.05.

Figure 2.3: Histogram of $\tilde{p}$ when testing the SIAS hypothesis

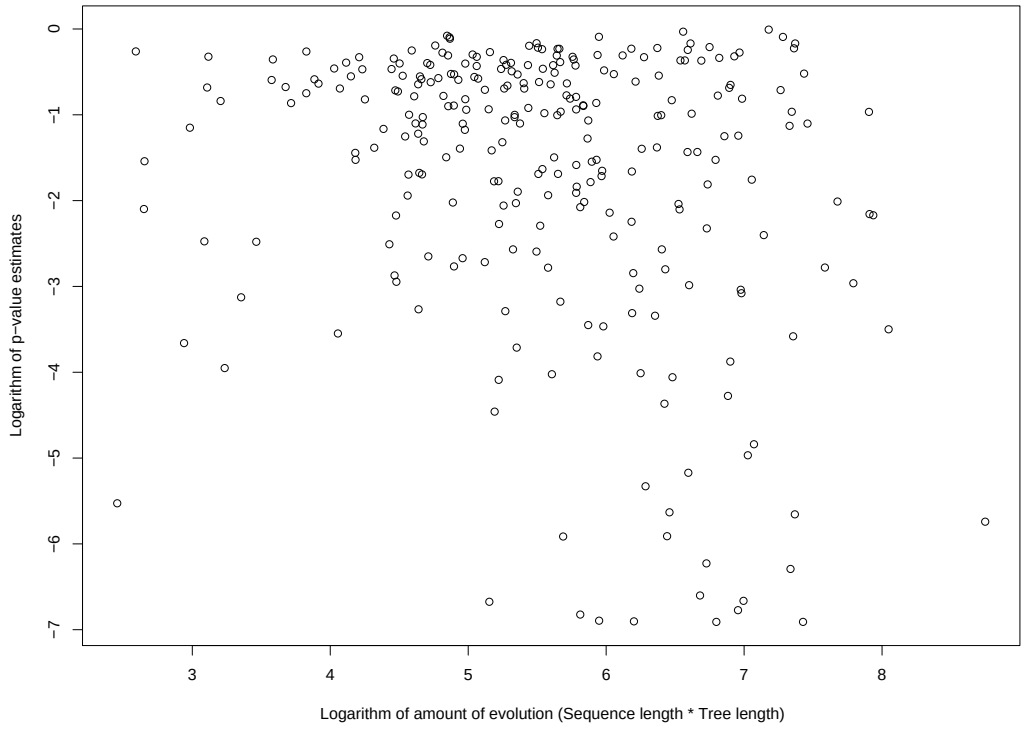
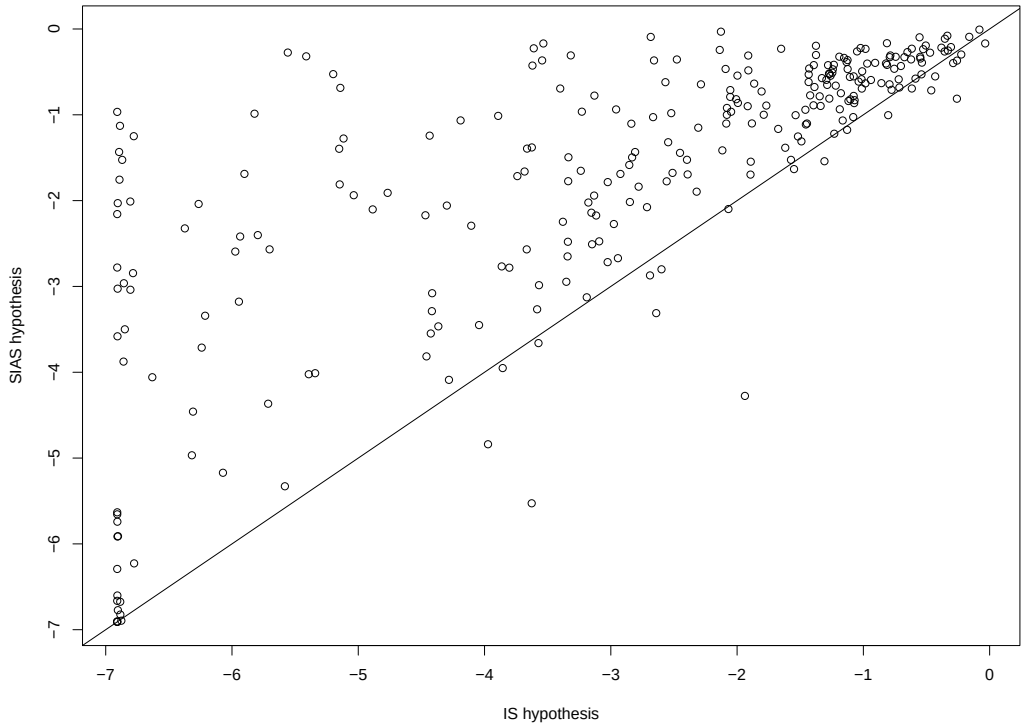


Figure 2.4: Comparison of the logarithm of $\tilde{p}$ with the logarithm of the amount of evolution when testing the SIAS hypothesis



The line of $y = x$ is drawn for visual assistance.

Figure 2.5: Comparison of the logarithm of p-value estimates for the IS and SIAS tests

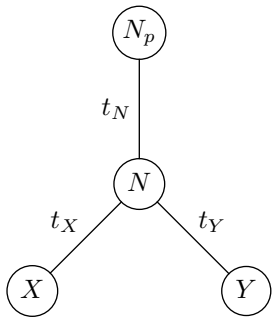


Figure 2.6: Example tree

Chapter 3

Dependence among sites in RNA evolution

Jiaye Yu and Jeffrey L. Thorne[1]

Keywords: RNA secondary structure, Free energy, Dependence among sites

[1]Corresponding author; *Address*: Campus Box 7566, North Carolina State University, Raleigh, NC 27695-7566 *Phone*: (919)515-1946 *Fax*: (919)515-7315 *Email*: thorne@statgen.ncsu.edu

3.1 Abstract

In recent decades, evolutionary models of molecular sequences have been extensively developed and widely used. However, although models of genotype (e.g., DNA sequence) evolution have been greatly elaborated, less attention has been paid to the effect of phenotype on the evolution of the genotype. Here we proposed an evolutionary model aimed at fill this gap. In this model, RNA encoding regions are considered with RNA secondary structure taken into account. It is well known that RNA sequences can vary considerably while RNA secondary structure may be much more conserved. Compensatory mutations maintaining conserved secondary structure indicate that it is biologically implausible to adopt the widely used assumption of independence among sites for phylogeny analyses. Global dependence among sites at the level of the entire sequence is introduced to our model as a term of free energy of RNA secondary structure. More importantly, the free energy information can be treated as surrogate of fitness that links genotype and phenotype. We have implemented this model under a Bayesian framework and generalized it to any number of sequences connected by a known phylogeny. For our implementation, previous knowledge about RNA secondary structure is necessary and the secondary structure is allowed to slightly change over time. Analyses of eight eukaryotic 5S ribosomal RNA sequences show that RNA secondary structure has a strong impact over the substitution rates and that rates are higher if the substitutions lead to a sequence with lower free energy than if they raise the free energy. Analyses on simulated sequences show that parameters in our model can be well estimated. We also show the potential applications of this model, including improved ancestral sequence reconstruction and location of functionally interesting sites. The attempt of connecting

genotype and phenotype we have made here will pave the lanes to future study of adaptive evolution.

3.2 Introduction

The connection of genotype to phenotype is central to evolution. Ideally, models of nucleotide substitution would fully reflect the impact of phenotype on the change over time of the underlying genotype. In practice, models of sequence change are usually constructed with little regard to the phenotype.

Notable exceptions to the usual practice are models of nucleotide substitution that incorporate information about RNA secondary structure. RNA secondary structures change more slowly than do the sequences that fold into them (e.g., Dixon and Hillis 1993; Gutell 1996; Higgs 2000). A variety of nucleotide substitution models have been constructed that exploit the slow change of ribosomal and transfer RNA secondary structures (e.g., Schöniger and von Haeseler 1994; Tillier 1994; Rzhetsky 1995; Muse 1995; Tillier and Collins 1995; Tillier and Collins 1998; Smith et al. 2004). An assumption of these models, the assumption is that RNA secondary structure is known and invariant over time. This assumption allows all residues in a RNA sequence to be classified on the basis of whether they pair with another site in the sequence. For unpaired sites, conventional models of nucleotide substitution are adopted. For paired sites in a helical region of RNA secondary structure, evolution at the two sites is jointly modeled. The joint modeling of paired sites is performed because substitution events that restore or preserve hydrogen bonding can have higher rates than those that disrupt hydrogen bonding. These models that incorporate RNA secondary structure permit correlated nucleotide substitutions at paired sites, but they require independent changes at unpaired sites. In addition, two sites that are paired are assumed by these models to evolve independently of other site pairs.

In contrast, the most successful techniques for approximating free energy of RNA

secondary structure do not independently handle different residue pairs in an RNA helix. Instead, the approximated free energy of an RNA secondary structure depends on the 'stacking' interactions between residue pairs that are adjacent in a RNA helix. Because RNA secondary structure tends to be preserved during evolution and because the phenomena that assist energy-based RNA secondary structure prediction are likely to affect evolutionary patterns and rates, it seems worthwhile to incorporate these phenomena into evolutionary models.

Stacking interactions between adjacent residue pairs in an RNA helix induce an evolutionary dependence among sites that cannot be handled by conventional evolutionary inference procedures. The assumption by conventional procedures of independent evolution among sites allows the likelihood for an entire data set of aligned sequences to be expressed as a product of individual site likelihoods and each site likelihood can be determined by applying the pruning algorithm of Felsenstein (1981). The pruning algorithm relies on the ability to calculate the probability of observing a specific residue type at the end of a branch given the parameter values of the evolutionary model. With more general forms of evolutionary dependence among sites, transition probability calculations can become intractable.

Starting with the pioneering work of Jensen and Pedersen (Jensen and Pedersen 2000; Pedersen and Jensen 2001), there has been much recent effort to make statistical inferences about evolution when sequence sites do not change independently (e.g., Robinson et al. 2003; Hwang and Green 2004; Pedersen et al. 2004; Siepel and Haussler 2004; Rodrigue et al. 2005). Here, we modify the sequence path approach of Robinson et al. (2003) to formulate and explore a possibility where the relative rate of sequence evolution is affected by approximate free energy of RNA secondary

structure. We discuss the strengths and weaknesses of our approach as well as some of its potential applications.

3.3 Methods and materials

3.3.1 Parameterization

We propose a Markovian model for evolution of RNA encoding regions with constraints due to secondary structure. In terms of both parameterization and statistical inference, this study closely parallels the protein evolution work of Robinson et al. (2003). Because we do not use the assumption of independent changes among sites, we define an instantaneous rate matrix R that specifies rates of change from each possible sequence to each other possible sequence. All sequences in the matrix are assumed to have length N. The entry of row i and column j represents the substitution rate from one sequence with length N to another with the same length. We make the assumption that no more than one position in a sequence can change in a particular instant. This means that all rates R_{ij} equal 0 if sequences i and j differ at more than one position. As a result, the rate matrix R tends to be sparse. In each matrix row i, the diagonal entry can be negative and the entries for the $3N$ neighboring sequences j that are different from i at exactly one position can be positive. The remaining $4^N - (3N + 1)$ entries in row i must be 0.

The values of the $3N$ entries in each row of R that can be positive depend on how the corresponding substitutions affect the approximate free energy. The biologically plausible expectation is that a substitution rate should be relatively high if the substitution improves the stability of RNA secondary structure. Likewise, if the secondary structure becomes less stable due to a specific substitution event, then the rate should be low. To assess the stability of a sequence folded into a specific secondary structure, we approximate the free energy of the sequence. This approximate free energy will

be denoted $E(i)$ and details for how it is calculated in our implementation will be provided below.

Otherwise, our parameterization is similar to that of widely used nucleotide models. We include parameters π_A, π_G, π_C and π_T ($\pi_A + \pi_G + \pi_C + \pi_T = 1$ and these parameters are all non-negative) so that mutation rates to the four nucleotide types need not be equal. Here we are studying the evolution of DNA sequences that encode RNAs. Although all Thymines are transcribed to Uracils. Our model describes evolution at the DNA level and we therefore use π_T rather than π_U. The parameter κ differentiates between transitions and transversions. The instantaneous substitution rate R_{ij} is set to 0 if sequences i and j differ at more than one site. For the cases where sequences i and j differ by exactly one site that has type h ($h \in \{A, G, C, T\}$) in sequence j, the rate matrix entries are:

$$R_{ij} = \begin{cases} u\pi_h \kappa e^{s(E(i)-E(j))} & \text{transition} \\ u\pi_h e^{s(E(i)-E(j))} & \text{transversion} \\ 0 & i \text{ and } j \text{ differ at more than one site.} \end{cases} \tag{3.1}$$

When the parameter s is zero, secondary structure does not affect the substitution rates and our model reduces to the widely used 'HKY' independent-site model (Hasegawa, Kishino and Yano 1985). The biologically reasonable s value is positive. Positive s values mean that low free energy of RNA secondary structure is favored by evolution. It is also formally possible to have a negative value for s. This would indicate that higher free energy and unstable secondary structure is favored by evolution. This scenario is biologically implausible for most situations. A partial validation of our approach would be to determine whether data support positive values for s.

3.3.2 Stationary probabilities of sequences

Under our time reversible model, the stationary probability of a sequence i with length N is

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

where i_m is the nucleotide type at position m of sequence i and where k_n is the n^{th} position of a sequence k that has n total residues (see Robinson et al. 2003). The denominator of the above equation is the summation over all possible sequences with length N. When $s = 0$, the denominator is 1 and the stationary distribution becomes the product of nucleotide frequencies. If s is substantially above zero, the stationary distribution is concentrated among sequences with a particularly good fit between sequence and secondary structure. Likewise, if s is substantially below zero, the stationary distribution is characterized by a relatively small number of sequences with particularly high free energies. The dependence of this stationary distribution on s means that s can be estimated from a single sequence. Therefore, it is not necessary to have a data set with two or more related sequences to get information about the impact of RNA secondary structure on evolution. This fact will be exploited below.

3.3.3 Sequence path density

For conventional models of sequence evolution, the dimension of the substitution rate matrix is manageable (e.g., a codon-based model with a 61×61 rate matrix) and common matrix exponentiation can be applied to calculate a transition probability.

With the dependence structure adopted here, the rate matrix R is $4^N \times 4^N$ and it is not feasible to exponentiate R unless N is extremely small. To overcome this high dimensionality, we employ a sequence path approach (Jensen and Pedersen 2000; Pedersen and Jensen 2001; Nielsen 2002; Robinson et al. 2003). We do this by augmenting the observed sequence data at tips of a phylogenetic tree with a sequence path. For every branch on the tree, the sequence path on a branch contains information about how the sequence at the beginning of the branch is transformed by substitution events to the sequence at the end of the branch. The sequence path also specifies the exact times of these substitution events.

Assume an unrooted tree topology that relates k observed sequences i^1, i^2, ..., i^k at nodes 1, 2, ..., k. There will be I internal nodes on this tree and they will be numbered $k+1$, $k+2$, ..., $k+I$. The unobserved sequences at these nodes will be denoted i^{k+1}, i^{k+2}, ..., i^{k+I}.

Because our dependence model is time reversible, any node can be selected to root the tree. We use node 1 as the root node and then a relative time ordering can be imposed on all nodes of the tree and the nodes that begin and end a branch can therefore be designated. Except for node 1, each node on the tree ends exactly one branch. A sequence path ρ on a tree is the set of sequence paths on the different branches of the tree. The sequence path on the branch that ends at node a will be denoted ρ^a and the corresponding branch will be named branch a. Because node 1 serves as the root, it will be the only node that does not end a sequence path. Therefore, the sequence path on a tree (i.e., ρ) consists of the collection of ρ^2, ρ^3, ..., ρ^{k+I}.

Letting $B(a)$ refer to the parental node of node a and letting θ represent all

parameters in the dependence model as well as the tree topology, we have

$$p(\theta, \rho | i^1, i^2, \ldots, i^k) = p(\theta, \rho^2, \rho^3, \ldots, \rho^{k+I} | i^1, i^2, \ldots, i^k) = \frac{p(i^1|\theta)p(\theta)\prod_{a=2}^{k+I} p(\rho^a | i^{B(a)}, \theta)}{p(i^1, i^2, \ldots, i^k)}. \tag{3.3}$$

In the above posterior distribution, $p(\theta)$ represents the prior density for the vector θ of parameters. Rather than specifying the time duration of each branch as part of θ, we choose to let branch lengths vary on the tree by assigning a different value of the rate scaling factor u to each branch on the tree. The rate scaling parameters for the branch ending at node a is a component of θ and will be denoted u^a.

If the sequence at node $B(a)$ and the sequence path ρ^a from $B(a)$ to a are known, then the sequence at a is also known. This means

$$p(\rho^a | i^{B(a)}, \theta) = p(i^a, \rho^a | i^{B(a)}, \theta). \tag{3.4}$$

Suppose there are q substitutions along the branch path ρ^a and let $t^a(z)$ be the time of the z^{th} substitution on this branch. At the beginning of the branch, we set $t^a(0) = 0$. The time the branch ends will be $t^a(q+1)$. Although different branches on a tree can obviously have different time durations, evolutionary rates and chronological times are confounded when only sequence data are available because only the product of rate and time can be estimated. In our case, we can set times of all branches to be 1 because we let rate scaling parameters vary among branches. For this reason, we set $t^a(q+1) = 1$. The sequence after the z^{th} substitution on branch a is defined as $i^a(z)$. For convenience, $i^a(0) = i^{B(a)}$ and $i^a(q+1) = i^a(q) = i^a$.

The rate at which a specific sequence v changes to a specific sequence k is R_{vk}.

By summing over all sequences k that differ from v with one single nucleotide, we can calculate $R_{v\bullet}$, the rate at which sequence v changes to some different sequence. We have

$$R_{v\bullet} = \sum_{k,k\neq v} R_{vk}. \tag{3.5}$$

Because at most $3N$ of $4^N - 1$ rates being summed above can be positive, the value of $R_{v\bullet}$ can be calculated without overly much computation. The time interval between two consecutive substitution events is exponentially distributed with the rate parameter $R_{v\bullet}$. Given that there is a substitution affecting sequence v at some time point, the probability that v changes to k is $R_{vk}/R_{v\bullet}$. The sequence path density for branch a is then,

$$\begin{aligned} p(i^a, \rho^a | i^{B(a)}, \theta) &= \left(\prod_{z=1}^{q} \frac{R_{i^a(z-1)i^a(z)}}{R_{i^a(z-1)\bullet}} R_{i^a(z-1)\bullet} e^{-R_{i^a(z-1)\bullet}(t(z)-t(z-1))} \right) e^{-R_{i^a(q)\bullet}(t(q+1)-t(q))} \\ &= \left(\prod_{z=1}^{q} R_{i^a(z-1)i^a(z)} e^{-R_{i^a(z-1)\bullet}(t(z)-t(z-1))} \right) e^{-R_{i^a(q)\bullet}(t(q+1)-t(q))}, \end{aligned} \tag{3.6}$$

(see Robinson et al. 2003). The final term $e^{-R_{i^a(q)\bullet}(t(q+1)-t(q))}$ is needed because there is no substitution event during the last time period [t(q),1]. The sequence path density over a phylogeny can be simply calculated by applying Equation 3.6 to each branch and then obtaining the product of all branch path densities.

3.3.4 Metropolis-Hastings algorithm

Given the observed multiple sequence data $i^1, i^2, \ldots, i^k$, we construct a Markov chain on the state space of θ and ρ via the Metropolis-Hastings algorithm (Metropolis

et al. 1953; Hastings 1970). The stationary distribution of this Markov chain is $p(\theta, \rho | i^1, i^2, \ldots, i^k)$, the posterior distribution of interest. Samples from this Markov chain can be used to approximate the posterior density. We initiate the chain at a randomly selected combination of $(\theta^{(0)}, \rho^{(0)})$. Then, we propose random new values θ' and ρ'. The probability density of proposing θ' and ρ' from θ and ρ will be denoted $J(\theta', \rho' | \theta, \rho)$. With probability equal to the minimum of 1 and r, where

$$r = \frac{J(\rho, \theta | \rho', \theta') p(\theta') p(i^{1'} | \theta') \prod_{a=2}^{k+I} p(i^{a'}, \rho^{a'} | i^{B(a)'}, \theta')}{J(\rho', \theta' | \rho, \theta) p(\theta) p(i^1 | \theta) \prod_{a=2}^{k+I} p(i^a, \rho^a | i^{B(a)}, \theta)}, \tag{3.7}$$

we set the next state $(\theta^{(1)}, \rho^{(1)})$ along our Markov chain to be the proposed state (i.e., $\theta^{(1)} = \theta'$, $\rho^{(1)} = \rho'$). Otherwise, $\theta^{(1)} = \theta$, $\rho^{(1)} = \rho$. By repeating this procedure, a Markov chain with stationary density $p(\theta, \rho | i^1, i^2, \ldots, i^k)$ is formed.

In Equation (3.7), the $\prod_{a=2}^{k+I} p(i^a, \rho^a | i^{B(a)}, \theta)$ and $\prod_{a=2}^{k+I} p(i^{a'}, \rho^{a'} | i^{B(a)'}, \theta)$ terms can be calculated via Equation 3.6. The $J(\rho, \theta | \rho', \theta')$ and $J(\rho', \theta' | \rho, \theta)$ terms are proposal densities that are described below. The remaining terms $p(i^1 | \theta)$ and $p(i^{1'} | \theta')$ are more difficult to handle because the denominator of the stationary distribution in Equation 3.2 is a sum over all possible sequences with length N. We approximate their ratio $p(i^{1'} | \theta') / p(i^1 | \theta)$ with the grid-based Gibbs sampling approach of Robinson et al. (2003).

3.3.5 Proposing θ

Our Markov chain Monte Carlo (MCMC) implementation actually consists of various proposal distributions $J(\theta', \rho'|\theta, \rho)$ and the Markov chain is formed by cycling through these proposal distributions. Each proposal can result in only slight differences between (θ, ρ) and (θ', ρ'). For example, one proposal distribution can have $\rho' = \rho$ and $\theta' = \theta$ except that $s' \neq s$. We employ similar proposal steps that propose change to only κ or to only the rate scaling factor for a branch. All of these proposal steps are Metropolis-Hastings schemes that involve sampling a proposed parameter value from a uniform distribution that is determined by the current parameter value, a pre-specified "window" length surrounding the current parameter value and any constraints on the parameters. Our technique for proposing new values of π_A, π_C, π_G and π_T is the same as that of Robinson et al. (2003).

3.3.6 Proposing sequence paths

Our current sequence path ρ and our proposed sequence path ρ' will differ only by the site path at a randomly selected site m. The proposed site path ρ'_m is generated via a simple nucleotide substitution model that has independent changes among sites. Our implementation used the 'HKY' independent change model (Hasegawa, Kishino and Yano 1985). The topology and HKY parameters will be represented by ψ. To

sample site paths, we use a strategy similar to that of Rodrigue et al. (2005),

$$
\begin{aligned}
& p(\rho_m | i_m^1, \ldots, i_m^k, \psi) \\
& \quad = p(\rho_m^2, \ldots, \rho_m^{k+I}, i_m^{k+1}, \ldots, i_m^{k+I} | i_m^1, \ldots, i_m^k, \psi) \\
& \quad = p(i_m^{k+1}, \ldots, i_m^{k+I} | i_m^1, \ldots, i_m^k, \psi) p(\rho_m^2, \ldots, \rho_m^{k+I} | i_m^1, \ldots, i_m^k, i_m^{k+1}, \ldots, i_m^{k+I}, \psi) \\
& \quad = p(i_m^{k+1}, \ldots, i_m^{k+I} | i_m^1, \ldots, i_m^k, \psi) \prod_{a=2}^{k+I} p(\rho_m^a | i_m^a, i_m^{B(a)}, \psi).
\end{aligned}
\tag{3.8}
$$

The $p(i_m^{k+1}, \ldots, i_m^{k+I} | i_m^1, \ldots, i_m^k, \psi)$ term in Equation 3.8 represents the probability of the interior node residues at site r conditional upon ψ and the observed tip residues. To sample internal node residues from the conditional density, we slightly modify the ancestral sequence reconstruction algorithm of Pupko et al. (2000). Our modification is described in Yu and Thorne (2005). Each of the remaining factors in Equation 3.8 represents the probability of a site path for a specific branch conditional upon ψ and upon the residues that begin and end the branch. To sample these site paths for each branch, we adopt the "forward simulation" procedure of Nielsen (2002) but our implementation does not include his suggestion for improving computational feasibility for cases where a branch length is small but the beginning and ending residues differ for the site on the branch. Instead, we simply set the branch lengths specified in ψ so that none are extremely small. Our implementation somewhat reduces computation by storing site paths where the ending residue in the forward simulation procedure does not match the ending residue i_m^a that is desired. When an already stored site path has the beginning and ending residues needed at some later time, the cached site path can then be used.

3.3.7 Inference from a single sequence

Information about parameters π_A, π_C, π_G and π_T is contained in Equation 3.2, the formula for stationary probabilities of sequences. More interestingly, the stationary probability formula also contains information about the structural impact factor s. This means that inferences about s can be made solely based on the stationary probability of a single sequence. In the case of a single sequence, $\theta = \{s, \pi_A, \pi_C, \pi_G, \pi_T\}$. For a single sequence i, the posterior distribution $p(\theta|i) = p(i|\theta)p(\theta)/p(i)$, can be approximated by sampling θ via a Metropolis-Hastings technique similar to, but much simpler than the one described above. Accordingly, the acceptance ratio in Equation 3.7 is modified to become

$$r = \frac{p(\theta'|i)J(\theta|,\theta')}{p(\theta|i)J(\theta'|\theta)} = \frac{p(i|\theta')p(\theta')J(\theta|\theta')}{p(i|\theta)p(\theta)J(\theta'|\theta)}. \tag{3.9}$$

The ratio of $p(i|\theta')/p(i|\theta)$ again can be approximated by the grid-based Gibbs sampler (Robinson et al. 2003) and the other terms are easy to compute.

3.3.8 Calculating RNA free energy

To approximate the free energy of an RNA molecule folded into a pre-specified structure, computer subroutines were adopted from the Vienna RNA software package (Hofacker et al. 1994). The default energy parameters in the Vienna RNA package (Mathews et al. 1999) were used in our study. Many potential factors can contribute to an RNA secondary structure energy approximation, but some of these factors are rare and we neglect them. For example, we ignore the coaxial stacking energy of adjacent helices in multi-loops (this is the "d2" option in the Vienna RNA package).

We use the canonical secondary structure, normally determined from comparative sequence analysis, as a reference structure to indicate which sites might be paired and which would not. Although the canonical structure is used to indicate which sites might be paired, the sites that might be paired are not forced to pair. For example, consider a case where two sites are paired in the canonical RNA structure. If one of these sites is occupied by an A and the other is occupied by a U, then the energy of the sequence will be evaluated by pairing the two sites. In contrast, the sites will not be paired if one is occupied by an A and the other is occupied by a G.

3.4 Analyses

3.4.1 Prior Densities

In all analyses, all combinations of non-negative values for π_A, π_C, π_G and π_T that satisfy the $\pi_A+\pi_C+\pi_G+\pi_T = 1$ were treated as being equally likely a *priori*. Prior densities were uniform on the interval (0,1) for all u's and uniform on the interval (0,4) for κ. Although biologically reasonable values of s are positive because positive values of s favor evolution of stable secondary structure, we center a uniform prior distribution for s about zero. By centering the s priors about zero, we can examine whether the sequence data have sufficient information to yield posterior densities for s that are concentrated in the positive values. Specifically, we set the prior distribution for s to be uniform on the interval (-1,1). The width of this interval was chosen after some pilot experiments. The goal was to use a prior where the posterior distribution for s was not concentrated near either the lower or the upper endpoint of the interval. We did not want to make the interval too wide because the amount of computation

needed for the grid-based Gibbs sampling can quickly grow as intervals grow in width.

To avoid possible problems in convergence, the results presented here force the κ and s value to be shared among branches. For each MCMC cycle, we propose one update for the κ, s and each of the u parameters. Each cycle includes 10 separate random selections of sites at which to propose site path updates to all branches. One of the four base frequency parameters is also the focus of a proposed update during each MCMC cycle. After examining substantial MCMC output to assess convergence of the Markov chain, we settled upon 150,000 MCMC cycles in each analysis with the first 30,000 of these cycles treated as a "burn-in" period that is not included in the posterior approximation. At least two independent runs with different starting points were performed to check convergence for each analysis.

3.4.2 eukaryotic 5S ribosomal RNAs

Ribosomal 5S RNA (5S rRNA) is an integral component of the large ribosomal subunit in almost all known organisms. The conserved nature of its secondary structure, the existence of 5S rRNA sequence level variation and the relatively short sequence length combine to make it a good choice for exploring the techniques introduced here.

Primary sequences and secondary structure

We analyzed a data set with eight eukaryotic 5S rRNA sequences, each 119 nucleotides in length. They are collected from the 5S ribosomal RNA database (Szymanski et al. 2002). Their sequence identifications are C_paradox, M_polymor, M_sativa, P_silvest, T_violea, U_hordei, S_pombe and Z_mays. The eight sequences were intentionally selected so as to avoid gaps in the alignment that relates them. Their canonical

structure has been determined by combining knowledge from comparative analysis and experiments and is depicted in Figure 3.6. Several 5S rRNA sites are almost identical among all eukaryotic 5S rRNAs. With the numbering scheme of Figure 3.6, these sites are 11, 41, 66, 74, 75, 76, 77, 90 and 99.

Tree topology

The neighbor-joining method (Saitou and Nei 1987) was employed to infer a tree topology that relates these eight sequences (Swofford 2002). The topology is in accordance with conventional hypotheses about eukaryotic relationships (Benson et al. 2000) and is shown in Figure 3.2. This topology was assumed when analyzing the eight sequences with the dependent-sites model.

Importance of RNA secondary structure

It is interesting to know the importance of helical and loop regions for stabilizing RNA secondary structure. We examined this via several permutation schemes (Figure 3.6). For each of the 1,000,000 sequences generated according to the simplest permutation scheme "Complete Permutation" (Figure 3.6A), the permuted sequences had exactly the same number of each of the four nucleotide types but these types were arranged in a different order than in the actual M_polymor sequence that was permuted. Order of nucleotides does not matter with conventional evolutionary models that have independent changes among sites. It is clear that if we ignore the secondary structure, the free energy of actual sequence is significantly lower than that of random sequences and this shows the importance of secondary structure to actual RNA sequences.

From the result of "Loop-only Permutation", we find that the contribution from

interactions among nucleotides in loop regions to the free energy of actual sequence is significant. In other words, the order of nucleotides in loop regions seems to be important despite the fact that conventional evolutionary models ignore the order. The "Helix-only Permutation" indicates how much information is lost by dinucleotide models that ignore stacking interactions (e.g., Schöniger and von Haeseler 1994; Rzhetsky 1995; Muse 1995; Tillier and Collins 1995; Tillier and Collins 1998). Our "Helix-only Permutation" results indicate that stacking interactions are not particularly important. The permutation procedures also yield a large variance for the distribution of approximate free energies of random sequences. It is worthwhile to mention here that, these permutation tests we adopted here all maintain the same stationary distribution as the original sequence. The results of permutation tests reflect the impact of secondary structure over evolutionary model only in terms of stationary distribution. The impact of secondary structure over substitution rates, which is a more interesting question, cannot be simply answered by these permutation tests. This large variance may also indicate room of improvement with dinucleotide models. We think our model presented here incorporating more information from the general dependence structure and hopefully it is better than dinucleotide models. It is not immediately straightforward to know how to compare our model with the dinucleotide models more thoroughly. The result from "Mixed complete permutation" is shown for comparison.

Inferences on multiple sequences

When the value of s is set to zero, our model reduces to the conventional HKY independence-site model (Hasegawa, Kishino and Yano 1985). As a check of our

software implementation, we elected to analyze the 5S rRNA sequence data with the s value forced to be 0. Via the *baseml* program of the PAML software (Yang 1997), we compared our results to maximum likelihood results with the HKY model. We also analyzed the 5S rRNA data with our program when s was not forced to be 0.

The maximum likelihood estimates from the *baseml* program are 2.6193 for κ and 0.2320, 0.2701, 0.2625 and 0.2354 respectively for π_A, π_C, π_G and π_T. The nucleotide frequency estimates when $s = 0$ are quite similar to the *baseml* estimates, but the κ estimate of 2.2892 from our implementation is noticeably lower than that from maximum likelihood method. When applying our estimation program to longer simulated sequences (e.g., length 595), the difference in the κ estimates from our program and from PAML became substantially smaller (data not shown). This is an indication that differences between the PAML maximum likelihood results and our Bayesian results are due to effects of prior distributions. With our program, a uniform prior distribution for κ is specified. Our program also places uniform prior distributions on the rate scaling "parameters" u. With our program, the rate scaling and κ prior distributions combine to induce a prior distribution on the expected number of substitutions per branch (i.e., the branch length). This means that κ and the branch length are *a priori* strongly and positively correlated with prior structure. Our κ estimate of 2.2892 appears to be less than the *baseml* estimate because the branch lengths on the 5S rRNA tree are relatively short. With our implementation, relatively short branch lengths indicate *a priori* that the value of κ is likely to be small. As a simple check, we analyzed the 5S rRNA data with our program but we fixed all parameters except κ at their maximum likelihood estimates from *baseml*. When we did this, we found that our posterior mean estimate of κ was 2.6268 and

therefore much closer to the *baseml* estimate of 2.6193. In the future studies, we plan to investigate more modifications to our prior distributions so that the analyses are more comparable with those of other software packages.

The posterior estimates when $s = 0$ and $s \neq 0$ are both tabulated in Table 3.1. The posterior distribution of s is concentrated in the biologically plausible $s > 0$ region (Table 3.1). This is the region of parameter space where evolution favors low free energy and stability of secondary structure. The values of π_A, π_C, π_G and π_T estimated under the full dependence model (i.e., $s \neq 0$) are quite different from those estimated when $s = 0$. The differences probably arise because the frequencies of nucleotides in RNA sequences depend on both mutation rates and fixation probabilities. One interpretation of the parameterization of our model (see Equation 3.1) is that the relative rate of mutations to A, C, G and T are proportional to π_A, π_C, π_G and π_T. With this interpretation, observed frequencies of the four nucleotide types can be expected to differ from π_A, π_C, π_G and π_T because certain nucleotide types may tend to be found more often in the sequences with low free energies. When $s > 0$, mutations that generate low free energies have relatively high fixation probability. This association between nucleotide types and fixation probabilities can generate departures between observed relative frequencies of the nucleotide types and π_A, π_C, π_G and π_T.

For validation purposes, 5S rRNA evolution was simulated according to the dependence model. The simulation assumed the tree topology that was used to analyze the actual data. For the actual data, the posterior mean of s is estimated to be approximately 0.2749. One set of sequences was simulated for each case of $s = 0.2749$, $s = 0$ and $s = -0.2749$. For all other parameters, the posterior means from the

actual data were used to simulate. For each of the simulation scenarios, the posterior densities of s are concentrated close to their true values (Table 3.2).

Inferences on a single sequence and sequence pairs

Inferences based on a single sequence were also performed. For all eight 5S rRNA sequences, the posterior estimates of parameter s are depicted in Table 3.3. The results are roughly consistent with what we expected. It is not surprising that the 95% credibility intervals from the single sequence analysis are wider than the credibility interval from the multiple sequence analysis because the amount of information contained in one sequence is significantly less than that contained in multiple sequences. With these single sequence analyses, there is a negative correlation between the free energy of the sequence being evaluated and the posterior mean estimated s value. This negative correlation suggests the favored direction of evolution is towards more stable RNA secondary structure and lower free energy.

To serve as an intermediate between the single-sequence and eight-sequence analyses, sequence pairs were also considered. For each of the 28 pairwise combinations of the eight sequences, two MCMC runs were performed. In addition, eight data sets of sequence pairs were formed by having two identical copies of the 5S rRNA sequences. With maximum likelihood analyses, the lack of variation between identical sequences means that the expected amount of evolution separating sequences will be estimated to be zero and other parameter estimates will be the same for single-sequence analysis and analyses of two identical sequences. This is not necessarily the situation for Bayesian analyses. It is possible that two homologous sequences had the opportunity to diverge but no nucleotide substitutions occurred since their common

ancestry. Because posterior means average over all possible evolutionary scenarios rather than just reflecting the maximum likelihood one, the posterior mean of the expected amount of evolution separating two sequences will exceed zero even if the sequences are identical.

Results from the pairwise analyses are shown in Table 3.4. There is an interesting pattern regarding the estimated values of s. The estimates tend to be highest for the single sequence analyses and are lowest for eight-sequence estimate. The cause of the pattern is unclear to us but we suspect it is related to inadequacies of our model. Specifically, the single-sequence s estimates depend solely on the stationary distribution of sequences and the prior distribution of s. The estimates of s from multiple sequences also depend on the sequence path relating them.

Although information about the s parameters comes from both the sequence stationary distribution and the sequence path, limitations of our model need not equally affect these two sources of information. For example, one limitation of our model is that variation of nucleotide substitution rates among sites is solely caused by RNA secondary structure together with the s parameter. Other factors affecting variation of rates among sites are not allowed by our model. Some ways of generalizing our model to allow increased rate heterogeneity among sites would not affect the stationary distribution of Equation 3.2 but would change the sequence path densities of Equation 3.6, our use of an overly simplistic model might therefore mean that the stationary distribution information about s and the sequence path information about s would not point to the same value for s. In future work, we plan to examine this possibility in more detail.

Impact of secondary structure over substitution rates

With our model, the factor

$$A_{ij} = e^{s(E(i)-E(j))} \tag{3.10}$$

represents the effects of natural selection on nucleotide substitution rates. If $A_{ij} > 1$, the substitution rate from sequence i to j would be higher than it would be without selection pressure. Therefore, $A_{ij} > 1$ could be interpreted as positive selection whereas $A_{ij} < 1$ could be interpreted as negative selection.

Because the 5S rRNA sequences have length 119 nucleotides, there are $357 = 119 \times 3$ sequences that differ from a given sequence i at a single position. For each of these 357 possible sequences j and for a particular value of s, the rate factor can be calculated. Using the posterior mean estimate of s from the eight sequence analysis, Figure 3.5 depicts histograms of A_{ij} values. Most A_{ij} values are less than 1. This makes sense because natural selection has surely shaped the compatibility of rRNA sequence and secondary structure. Substitutions where A_{ij} exceeds one are of particular interest. We then study the relationship between the potential changing sites and the known invariant sites. It is somehow surprising that for some invariant sites (e.g., site 74, 76, 77, 99) that have been observed in almost all available species, our model predicts that the substitution will lead to higher fitness for the entire sequences. Other references indicate that the conflict between facts and our predictions mainly occurs in the interesting loop E region of 5S rRNA. This loop E region is known to be important to several RNA-protein interactions (Wimberly, Varani and Tinoco 1993) and it is also the only significant RNA-RNA interaction region in 5S rRNA (Szymanski et al. 2003). This fact does show that our model is not

perfect we have used only partial information about RNA secondary structure while it also indicates that it is possible to find functionally interesting sites by applying our model.

Ancestral sequence free energy

We expect that the ancestral sequences reconstructed under our model are more realistic than those from independence-site models. The reason is that the free energy information of RNA secondary structure is taken into account in our model and the addition of structural information will definitely lead to better ancestral sequence reconstruction. One natural prediction is that the ancestral sequences estimated from our model have lower free energy values than those from independence-site models.

Under the Bayesian framework we have used, it is possible to obtain a distribution of estimated free energies of ancestral sequences on internal nodes. A reasonable expectation is that the free energy estimates of ancestral sequences associated with the dependence model (when $s \neq 0$) are lower than the ones associated with independence-site model (when $s = 0$). Here we present the estimated posterior distribution of free energies of internal node sequences on the 5S tree (see Figure 3.6). It is obvious that, for deeper internal nodes (e.g., node 1 and 2), the ancestral sequences inferred from our dependence model tend to have lower free energy than the ones inferred from independence-site model, which is reasonable and what we expected. It is also not surprising that for internal nodes closely related to extant sequences, we cannot find significant separation between observed sequence data and the distribution of free energy distribution of ancestral sequences inferred from different model. Probably because of the extremely short branch lengths (e.g., the branches leading to node

5 and 6), either a dependence-site or independence-site model will predict ancestral sequences reasonably well for some internal nodes.

3.5 Discussion

Using the general statistical approach developed by Robinson et al. (2003), we have developed an approach to model the evolution of RNA sequences that incorporates RNA secondary structure. Furthermore and independently of Rodrigue et al. (2005), we extended the Robinson et al. (2003) work to the multiple sequence case. To a limited extent, we also relaxed the assumption of a "frozen" structure.

The free energy of an RNA sequence might be a rough measure of the extent that a specific sequence is favored by evolution. There is substantial knowledge that the real RNA structures do not necessarily hold the configuration with minimum free energy. It is also as a general tendency that the more stable the secondary structure, the lower the free energy.

Seffens and Digby (1999) concluded that the free energies of mRNA secondary structures are significantly lower than those of random sequences. This conclusion contrasted with the belief that, because of the short life of mRNA, it is not absolutely necessary to maintain a relative stable secondary structure for mRNA. Workman and Krogh (1999) later pointed out that random simulated sequences should reflect counts of consecutive nucleotide pairs ("dinucleotide") that are found in actual RNA sequences. Dinucleotide counts are important because they contain information about the stacking interactions between adjacent pairs of bases in an RNA helical region. In contrast to Seffens and Digby (1999), Workman and Krogh (1999) further concluded

that, even for tRNA and rRNA sequences, the free energy is not significantly lower than random sequences. The conclusion is a little surprising because, unlike mRNA, the stability of secondary structure is vital for tRNA and rRNAs to be functionally active. Recent work of Clote et al. (2005), based on large scale analysis, indicates that the free energy is indeed much lower than that of random sequences for most structural RNAs. Even smaller RNA molecules such as the precursor of microRNA have lower free energy than random sequences (Bonnet et al. 2004). In all these works, it should be remembered that the free energy of a sequence was obtained from secondary structure prediction algorithms based on energy minimization. This means that the predicted secondary structure with minimized free energy might be substantially different from the canonical secondary structure. Ideally one would evaluate the free energy of a sequence over its real structure, or at least over structures similar to its real structure.

Models that allow correlated changes between paired sites in an RNA helix fit data better than those that ignore RNA secondary structure (e.g., Muse 1995). Because our approach incorporates stacking interactions and other factors influencing free energy, it may fit data better than models that incorporated secondary structure only by discriminating between sites in helices and sites in loops. As part of our future planned work, we intend to perform rigorous statistical comparisons of our model and alternatives. One way to do these model comparisons in a Bayesian framework is by estimating Bayes factors. There are a variety of techniques for estimating Bayes factors and we are particularly interested in the "thermodynamic integration" procedure of Lartillot and Philippe (2004).

One limitation of the model introduced here is that it does not take into account

RNA tertiary structure. An effect of this limitation is that energetic contributions from pseudoknots are not taken into account. Algorithms exist for approximating the energy of RNA structures with pseudoknots (e.g., Lyngsø and Pedersen 2000). These algorithms tend to be computationally demanding for predicting RNA structure, but extending our approach to pseudoknots should be computationally feasible for cases where the canonical structure is known.

Here, we have developed an evolutionary model that incorporates the effect of RNA secondary structure over substitution rates. The evolutionary dependence among sites is naturally integrated in our model. In a more general sense, we developed a model to effectively combine genotype and phenotype. In this specific case of RNA, the DNA encoding the RNA sequence is genotype and the RNA secondary structure is phenotype. Because of the generality of the approach we adopted here, we can extend the statistical procedure to many other situations where phenotype interacts with genotype during evolution.

3.6 Acknowledgments

We thank Stéphane Aris-Brosou, Sang-Chul Choi, Hirohisa Kishino and Tae-Kun Seo for their help.

3.7 References

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp and D. L. Wheeler. 2000. Genbank. Nucl. Acids Res. **28**:15–18.

Bonnet, E., J. Wuyts, P. Rouze and Y. Van de Peer. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics **20**:2911–2917.

Clote, P., F. Ferré, E. Kranakis and D. Krizanc. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA **11**:578–591.

Dixon, M. T. and D. M. Hillis. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetics analysis. Mol. Biol. Evol. **10**:256–267.

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

Gutell, R. R. 1996. Comparative sequence analysis and the structure of 16S and 23S RNA. In R. A. Zimmermann and A. E. Dahlberg, eds., Ribosomal RNA: structure, evolution, processing and function in protein biosynthesis, 15–27. CRC Press, Boca Raton, FL.

Hasegawa, M., H. Kishino and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

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

Higgs, P. G. 2000. RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **33**:199–253.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**:167–188.

Hwang, D. G. and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA **101**:13994–14001.

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

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. **21**:1095–1109.

Lyngsø, R. B. and C. N. Pedersen. 2000. RNA pseudoknot prediction in energy-based models. J. Comput. Biol. **7**:409–427.

Mathews, D. H., J. Sabina, M. Zuker and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**:911–940.

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

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139**:1429–1439.

Nielsen, R. 2002. Mapping mutations on phylogenies. Syst. Biol. **51**:729–739.

Pedersen, A.-M. K. and J. L. Jensen. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. **18**:763–776.

Pedersen, J. S., R. Forsberg, I. M. Meyer and J. Hein. 2004. An evolutionary model for protein-coding regions with conserved RNA structure. Mol Biol Evol **21**:1913–1922.

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

Robinson, D. M., D. Jones, H. Kishino, N. Goldman and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**:1692–1704.

Rodrigue, N., N. Lartillot, D. Bryant and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene **347**:207–217.

Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. Genetics **141**:771–783.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Schöniger, M. and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3**:240–247.

Seffens, W. and D. Digby. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucl. Acids Res. **27**:1578–1584.

Siepel, A. and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol **21**:468–488.

Smith, A. D., T. W. H. Lui and E. R. M. Tillier. 2004. Empirical models for substitution in ribosomal RNA. Mol. Biol. Evol. **21**:419–427.

Swofford, D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sinauer Associates.

Szymanski, M., M. Z. Barciszewska, V. A. Erdmann and J. Barciszewski. 2002. 5S ribosomal RNA database. Nucl. Acids Res. **30**:176–178.

Szymanski, M., B. M. Z and B. J. Erdmann V. A. 2003. 5S rRNA: structure and interactions. Biochem J. **371**:641–651.

Tillier, E. R. M. 1994. Maximum likelihood with multi-parameter models of substitution. J. Mol. Evol. **39**:409–417.

Tillier, E. R. M. and R. Collins. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12**:7–15.

Tillier, E. R. M. and R. A. Collins. 1998. High apparent rate of simultaneous compensatory basepair substitutions in ribosomal RNA. Genetics **148**:1993–2002.

Wimberly, B., G. Varani and I. Tinoco. 1993. The conformation of loop E of eukaryotic 5S ribosomal RNA. Biochemistry **32**:1078–1087.

Workman, C. and A. Krogh. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucl. Acids Res. **27**:4816–4822.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics **13**:555–556.

Yu, J. and J. L. Thorne. 2005. Testing for spatial clustering of amino acid replacements within protein tertiary structure. J. Mol. Evol. Accepted.

Table 3.1: Posterior estimates for the eukaryotic 5S rRNA data set.

Parameter	Priors	Posteriors (s $\neq$ 0)	Posteriors (s = 0)
u_1	0.5	0.1607	0.0781
(1 $\rightarrow$ 2)	(0.025,0.975)	(0.0574,0.3137)	(0.0055,0.1991)
u_2	0.5	0.3334	0.2995
(1 $\rightarrow$ 3)	(0.025,0.975)	(0.1764,0.534)	(0.1553,0.4884)
u_3	0.5	0.386	0.2928
(1 $\rightarrow$ C_paradox)	(0.025,0.975)	(0.2126,0.6366)	(0.1606,0.4847)
u_4	0.5	0.1503	0.0922
(2 $\rightarrow$ 4)	(0.025,0.975)	(0.0529,0.2859)	(0.0193,0.1903)
u_5	0.5	0.2572	0.2166
(2 $\rightarrow$ S_pombe)	(0.025,0.975)	(0.138,0.4185)	(0.1082,0.3578)
u_6	0.5	0.1614	0.118
(4 $\rightarrow$ U_hordei)	(0.025,0.975)	(0.0699,0.2907)	(0.0478,0.2156)
u_7	0.5	0.1861	0.1389
(4 $\rightarrow$ T_violea)	(0.025,0.975)	(0.0885,0.3176)	(0.0643,0.2379)
u_8	0.5	0.0536	0.0322
(3 $\rightarrow$ 5)	(0.025,0.975)	(0.0031,0.139)	(0.001,0.1012)
u_9	0.5	0.1416	0.1277
(3 $\rightarrow$ M_polymor)	(0.025,0.975)	(0.0591,0.2511)	(0.0537,0.2238)
u_{10}	0.5	0.0666	0.0583
(5 $\rightarrow$ 6)	(0.025,0.975)	(0.0203,0.1344)	(0.0185,0.1174)
u_{11}	0.5	0.0288	0.0204
(5 $\rightarrow$ P_silvest)	(0.025,0.975)	(0.0021,0.0823)	(8e-04,0.0646)
u_{12}	0.5	0.0689	0.056
(6 $\rightarrow$ Z_mays)	(0.025,0.975)	(0.0246,0.136)	(0.0198,0.11)
u_{13}	0.5	0.0338	0.0268
(6 $\rightarrow$ M_sativa)	(0.025,0.975)	(0.0053,0.0851)	(0.0037,0.0682)
κ	2.0	2.0432	2.2892
	(0.1,3.9)	(1.3728,2.9221)	(1.5476,3.2795)
s	0.0	0.2749	0
	(-0.95,0.95)	(0.2183,0.3322)	NA
π_A	0.25	0.2963	0.2326
		(0.2374,0.3569)	(0.1816,0.2889)
π_C	0.25	0.2352	0.2789
		(0.1823,0.293)	(0.2204,0.3409)
π_G	0.25	0.1978	0.2586
		(0.1519,0.2469)	(0.2021,0.3203)
π_T	0.25	0.2708	0.2298
		(0.2149,0.3316)	(0.1777,0.286)

Below the posterior (prior) means, 95% credibility (95% prior) intervals are indicated in parentheses. The subscripts on the rate scaling u parameters correspond to the branches as numbered on Figure 3.2.

Table 3.2: Posterior means and 95% credibility intervals for parameter s from simulated sequence data.

Truth	Posterior mean	95% Credibility Interval
0.2749	0.2617	(0.2021,0.3223)
0	0.0096	(-0.1022,0.1169)
-0.2749	-0.2388	(-0.3379,-0.1432)

Sequences were simulated according to our model, the eukaryotic 5S rRNA canonical structure and the topology of Figure 3.2.

Table 3.3: Posterior estimates of s based upon individual eukaryotic 5S rRNA sequences.

Sequence	Posterior	Free energy (kcal/mol)
C_paradox	0.4213 (0.3022,0.5558)	-46.9
M_polymor	0.3934 (0.2824,0.5210)	-41.3
U_hordei	0.3916 (0.2815,0.5193)	-39.7
S_pombe	0.3965 (0.2833,0.5243)	-37.6
T_violea	0.3997 (0.2853,0.5309)	-33.9
P_silvest	0.3449 (0.2396,0.4594)	-33.4
M_sativa	0.3384 (0.2330,0.4530)	-28.8
Z_mays	0.3421 (0.2344,0.4560)	-28.4

Posterior means are followed in parentheses by 95% credibility intervals.

Table 3.4: Posterior means of s for eukaryotic 5S rRNA sequence pairs

Ancestor \ Descendant	M_sativa	C_paradox	M_polymor	Z_mays	S_pombe	T_violea	U_hordei	P_silvest
M_sativa	0.3343	0.3204	0.3017	0.3214	0.3046	0.296	0.2947	0.3174
C_paradox	0.3201	0.4205	0.3405	0.3214	0.3392	0.3455	0.347	0.3382
M_polymor	0.3001	0.3441	0.3882	0.3059	0.3322	0.3168	0.3203	0.3158
Z_mays	0.3206	0.3165	0.3069	0.3377	0.3124	0.3041	0.2967	0.314
S_pombe	0.3025	0.3529	0.3346	0.3285	0.3948	0.326	0.3254	0.3034
T_violea	0.2924	0.3567	0.322	0.2965	0.3417	0.3961	0.352	0.3006
U_hordei	0.2987	0.3413	0.3176	0.3003	0.3292	0.3535	0.3905	0.3116
P_silvest	0.3189	0.3234	0.3165	0.312	0.314	0.3043	0.3061	0.3421

The entries in the upper and lower triangle is not perfectly symmetric because the order of ancestor and descendant is different. The approach of sequence path augmentation leads to the close estimates instead of exactly the same one.

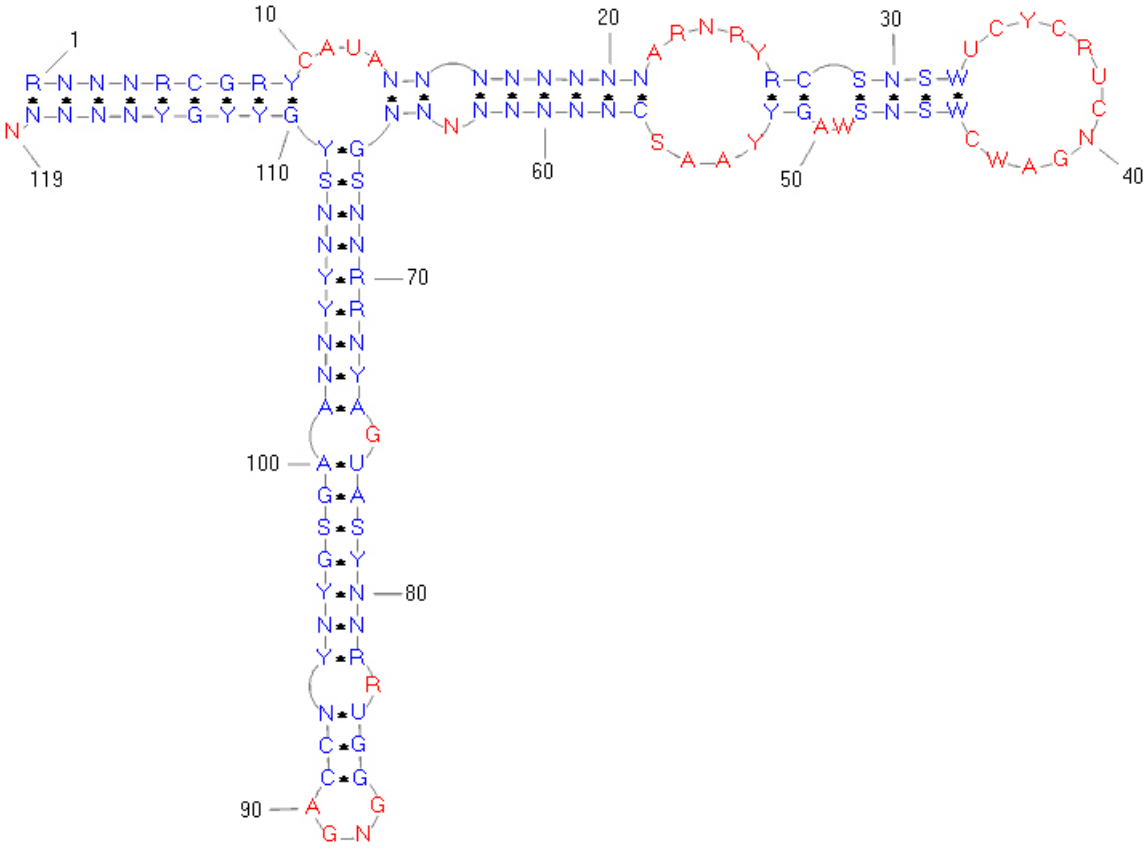


Figure 3.1: Canonical secondary structure of eukaryotic 5S rRNA

IUPAC ambiguity codes for RNA are used to indicate positions where there is substantial sequence variation among eukaryotes.

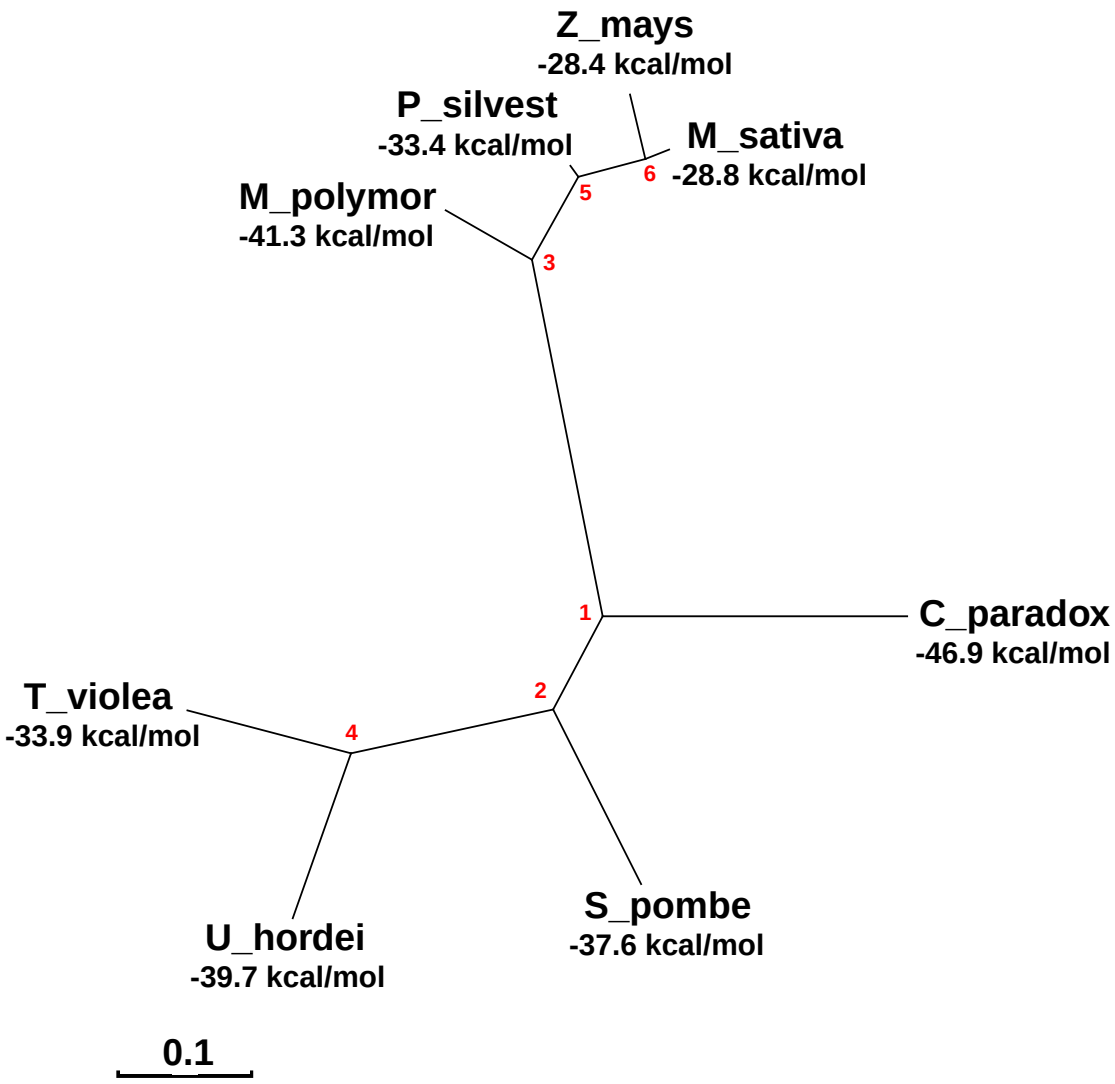


Figure 3.2: Phylogeny of eight eukaryotic 5S rRNA sequences
Approximate free energies of the eight sequences are also listed.

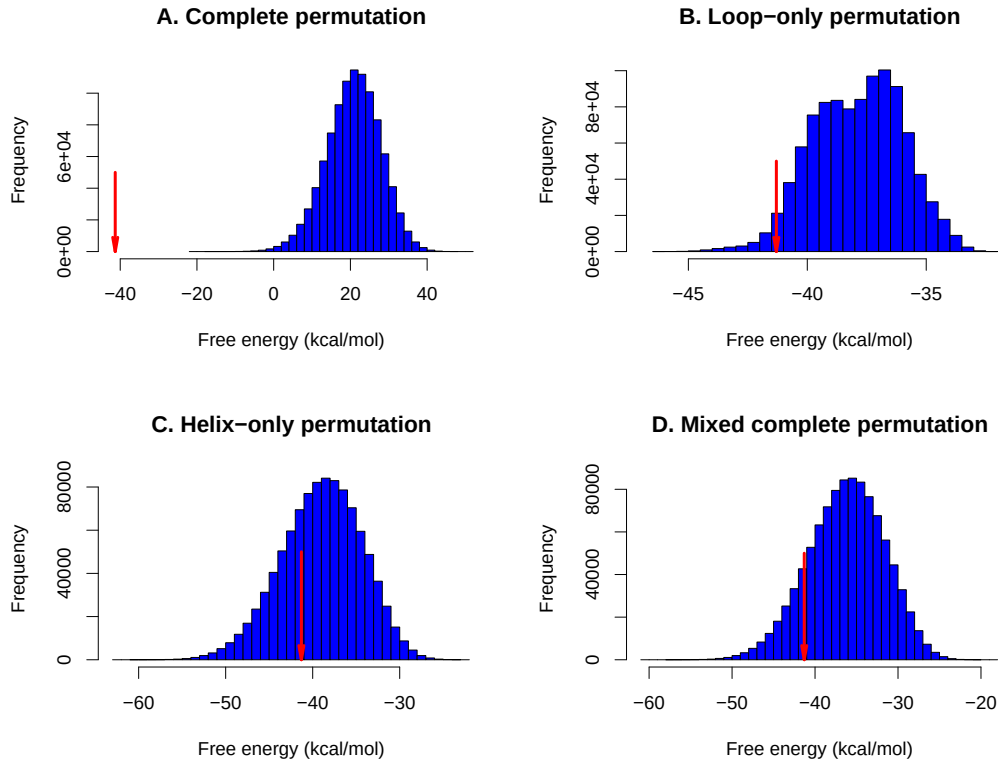


Figure 3.3: Permutation tests on sequence M_polymor

Each histogram shows approximate free energies from 1,000,000 permuted M_polymor sequences. Arrows represent the approximate free energy of the actual M_polymor sequence.

A) The entire M_polymor sequence was permuted. B) Only positions of the M_polymor sequence in loops of the canonical structure were permuted. Positions corresponding to helical regions were not involved in the permutation and therefore the permuted sequences were identical to the M_polymor sequence at these positions. C) Only positions of the M_polymor sequence in helical regions were permuted. Sites corresponding to loops were not included in the permutation. Paired sites were not disrupted by the permutations. This means that helical regions in the permuted sequences had exactly the same counts of AU, CG and GU pairs as does the actual M_polymor sequence. D) Helical and loop regions of the M_polymor sequence were separately permuted.

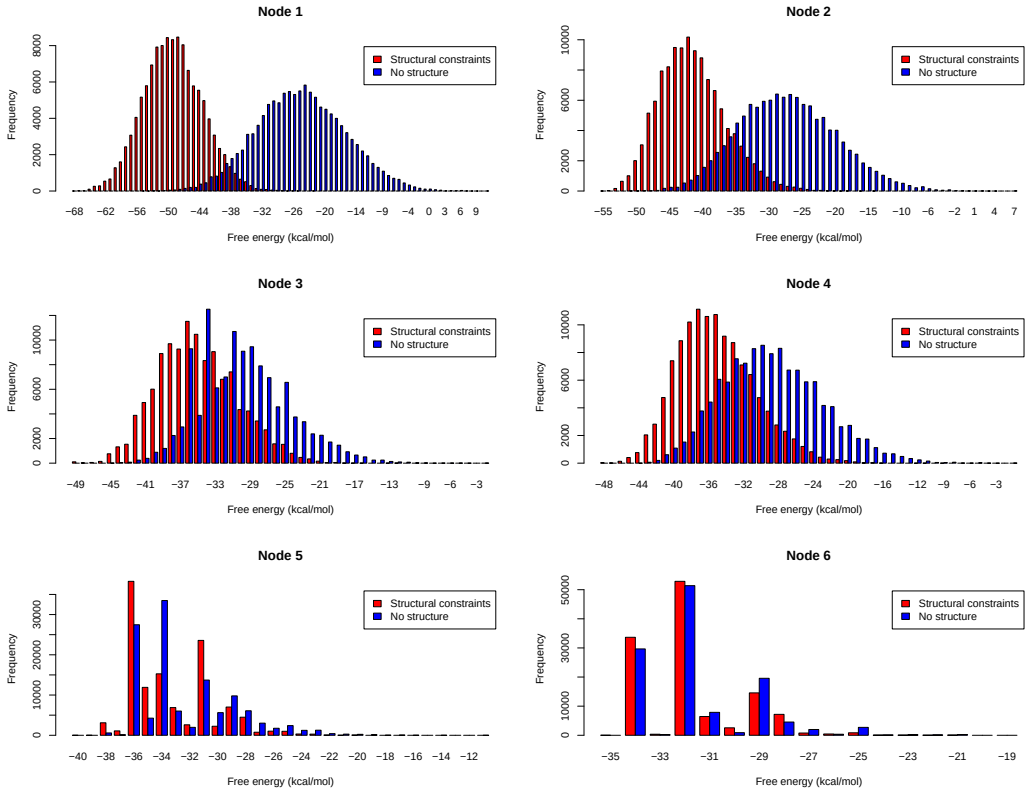


Figure 3.4: Free energy distribution of internal node sequences

The numbering of internal nodes is consistent with the phylogeny in Figure 3.2. The 'structural constraints' categories of the histogram correspond to an analysis where the s parameter was free to vary. The 'No structure' categories to an analysis where the s parameter was forced to 0.

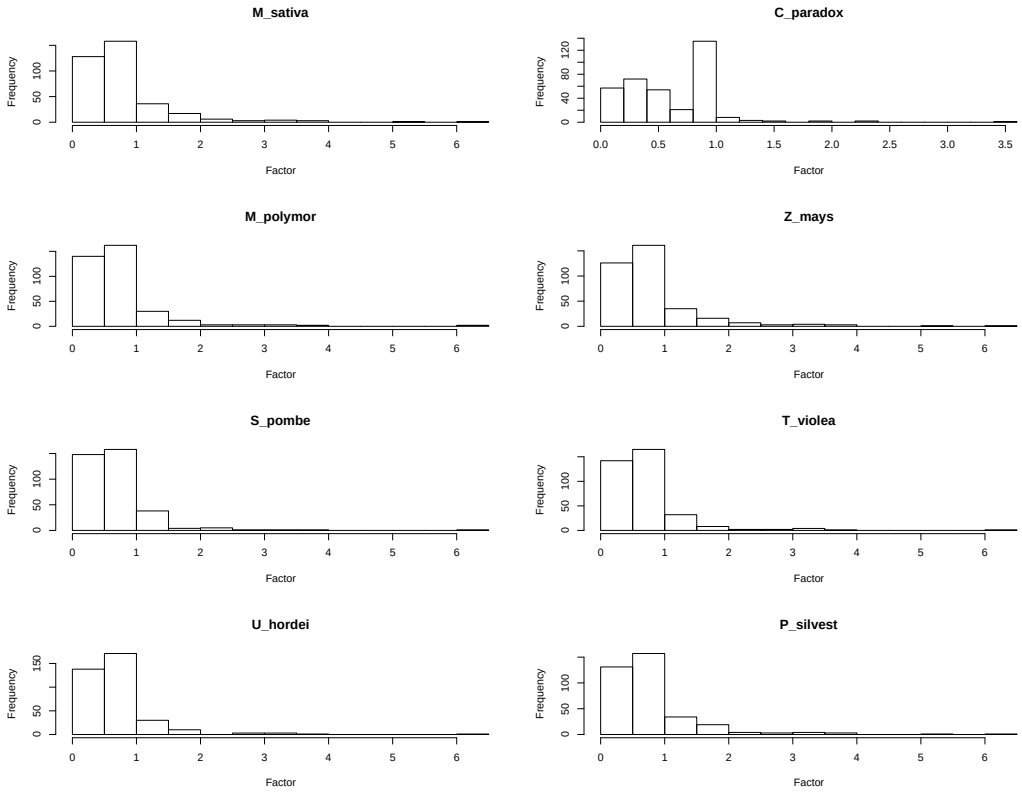


Figure 3.5: Histogram of rate factor A_{ij}

Chapter 4

Conclusion and future direction

4.1 Introduction

Until recently, most models for molecular evolution have ignored important biological facts. The widely adopted assumption of evolutionary independence among sites is statistically convenient and computationally feasible, but not biologically realistic. Great advances in our understanding of molecular evolution have been made with this independence assumption but it is now natural to develop more realistic models that abandon this assumption.

Cases for allowing dependence due to protein or RNA structure are especially strong. Most proteins and non-coding RNAs have to maintain certain structures to perform their functional roles. As a result, protein and RNA evolve much more slowly over time than DNA sequences (e.g., Chothia and Lesk 1986; Dixon and Hillis 1993; Flores et al. 1993; Higgs 2000).

Especially in RNA evolution, the free energy of RNA secondary structure can be adopted as a surrogate of evolutionary "fitness". If the free energy of a RNA secondary structure is low, this means that the structure is stable and further indicates high "fitness". It has been shown in the previous chapter that substitution rates in RNA sequences can be affected by approximate free energy of RNA secondary structure.

The effort to incorporate genotype and phenotype in the area of molecular evolution is still in its infancy. Because of the generality of the statistical approaches we adopted here, they can be applied to many other areas with the situations with interplay between genotype and phenotype. Here, we will point out several future directions to improve this line of research. Some of these lines are quite general and others are rather technical and specific.

4.2 Expected free energy and structural change over time

Although structures of molecular sequences change much more slowly than do sequences over time, they do change. It seems that modeling structural changes over time is much more difficult than modeling sequence changes. It is even more difficult to incorporate changing protein tertiary structures into evolutionary models. Although difficult, possible solutions are still available. For the case of RNA, these solutions might be more straightforward. In the approaches we described in the previous chapter, the calculation of free energy is based on the information of a known canonical structure. If this information is not available, the situation might be more complicated. One reasonable alternative is not to base energy calculation on only one structure. Instead, the expected free energy of a sequence can use used.

The most widely used algorithms for RNA secondary structure prediction not only predict the RNA secondary structure with the lowest free energy for a given RNA sequence, but also a set of suboptimal secondary structures (e.g., Zuker 1989; McCaskill 1990). Previously, we used the notation $E(s)$ to refer to the approximate free energy of a sequence s when folded into a single known secondary structure. Now, imagine the secondary structure of s is unknown but there are M possible secondary structures. The energy of secondary structure y for sequence s will be $E_y(s)$. Because the Boltzmann distribution describes the expected proportion of time that a sequence spends in a specific conformation among the M possible conformations, the Boltzmann distribution can be used to estimate the expected free energy $\overline{E(s)}$ of sequence

s where the expectation is over all M conformations.

$$\overline{E(s)} = \sum_{y=1}^{M} E_y(s) \times \frac{e^{\frac{E_y(s)}{kT}}}{\sum_{y=1}^{M} e^{\frac{E_y(s)}{kT}}} = \frac{1}{\sum_{y=1}^{M} e^{\frac{E_y(s)}{kT}}} \times \sum_{y=1}^{M} E_y(s) e^{\frac{E_y(s)}{kT}}. \tag{4.1}$$

In the above equation, k is the Boltzmann constant and T is the temperature (normally 310.15K). It is much more difficult to directly model the change of structure over time compared with that of molecular sequence. By adopting expected free energy, the problem of structural change over time could be effectively overcome.

4.3 Numerical optimization

When applying our Markov chain Monte Carlo estimation procedure, it is common to evaluate free energy values for billions of sequences. Although the calculation of free energy given a known secondary structure can be done in linear time compared with the $O(n^3)$ time complexity of an RNA secondary structure prediction algorithm, it still takes a long time to compute for so many intermediate sequences.

Currently, a computer subroutine slightly modified from *RNAeval* program of the Vienna RNA software package (Hofacker et al. 1994) is used to evaluate the free energy values of RNA secondary structures in our implementation. This subroutine is designed for general purpose usage and has not yet been optimized for our special situations.

In our estimation procedure, there are indeed some aspects that could be exploited to reduce computation. Each time when the general rate away from a specific sequence of length N is evaluated, the sequences being considered are exactly one

nucleotide different from the original one. These sequences are otherwise identical to the original sequence. If we can avoid redundant computation with an improvement to the *RNAeval* subroutine, our estimation program will definitely obtain increased speed.

The RNA secondary structure can be stored in a "linear tree" data structure proposed by Schmitt and Waterman (1994), see also Rastegari (2004). Each base pair can be represented by a node in the tree and the children nodes represent the inner base pairs in a bracket-form representation of RNA secondary structure. The stacking energy of interactions between two consecutive base pairs can be well represented by the branch length between the two nodes. With this data structure, when only one nucleotide change occurs from the current sequence (as described previously), we can save considerable amount of computation with keeping all the calculation we have done previously for the unchanged part of the secondary structure. We plan to implement this improvement in a future version of our estimation program.

4.4 Secondary structure of mRNA

Currently, we have applied our model to tRNA/rRNA encoding regions. However, there are more general potential applications. Although previous studies (e.g., Workman and Krogh 1999) indicate that mRNAs do not have significantly lower free energy than random sequences, it is still possible that mRNA forms secondary structures (e.g., Chartrand et al. 1999; Rocha et al. 1999; Katz and Burge 2003). Incorporating the mRNA secondary structure information into the evolution of protein encoding region is another possible future direction. One big difference is that mRNAs contain

the information of encoded proteins, so the difference between synonymous and non-synonymous substitutions need to be considered, the substitution rate from sequence i to j is then,

$$R_{ij} = \begin{cases} u\pi_h \kappa e^{s(E(i)-E(j))} & \text{synonymous transition} \\ u\pi_h e^{s(E(i)-E(j))} & \text{synonymous transversion} \\ u\pi_h \omega\kappa e^{s(E(i)-E(j))} & \text{nonsynonymous transition} \\ u\pi_h \omega e^{s(E(i)-E(j))} & \text{nonsynonymous transversion} \\ 0 & i \text{ and } j \text{ differ at more than one site.} \end{cases} \tag{4.2}$$

Both RNA secondary structure and protein tertiary structure can be conceptualized as phenotype in this case. The substitution rate is not only affected by corresponding RNA secondary structure, but also protein tertiary structure. Because the dependence term related to sequence-structure compatibility measure at the level of protein sequence takes the similar mathematical form, it is possible to combine directly the information of the two types of dependence terms. Incorporating our work here with that of Robinson et al. (2003), we let pairwise amino acid interaction scores for folding a translated sequence i into a known protein structure be $E_p(i)$. The $E_s(i)$ will represent a corresponding solvent accessibility score. Also, we let s_r, s_s and s_p respectively be parameters that capture evolutionary impacts of RNA secondary structure, pairwise amino acid interactions and solvent accessibility. This notation

yields a rate matrix such as

$$R_{ij} = \begin{cases} u\pi_h \kappa e^{\Delta F} & \text{synonymous transition} \\ u\pi_h e^{\Delta F} & \text{synonymous transversion} \\ u\pi_h \omega\kappa e^{\Delta F} & \text{nonsynonymous transition} \\ u\pi_h \omega e^{\Delta F} & \text{nonsynonymous transversion} \\ 0 & i \text{ and } j \text{ differ at more than one site,} \end{cases} \tag{4.3}$$

where $\Delta F = e^{s_r[E(i)-E(j)]+s_s[E_s(i)-E_s(j)]+s_p[E_p(i)-E_p(j)]}$.

4.5 Rate variation among sites

Rate variation among sites is a well recognized phenomenon and has been modeled reasonably well with a discretized Gamma distribution (Yang 1994). It is interesting to compare the rate variation in our model with the conventionally adopted Gamma distribution. We can find out that how much rate variation can be attributed to site dependence that can be represented by free energy information and potentially quantify the amount of rate variation that can not be fully explained by our dependence model. It is also possible to incorporate Gamma distribution to represent rate variation among sites.

4.6 Site specific rate matrix incorporating site dependence

If we assume that RNA secondary structure does not change over time, then the whole secondary structure can be partitioned into a set of doublets (base pairs) for helical regions and a set of mononucleotides for loop regions. We can still apply the Equation (3.1) to compute the substitution rate at each site (doublet or mononucleotide). The site-specific substitution rate is different from the rate based on doublet models of Tillier's type (e.g., Tillier 1994; Tillier and Collins 1995; Tillier and Collins 1998) because the information about the whole secondary structure is included in the form of free energy. We can expect that it will perform better than the conventional 16×16 models. Compared with the full dependence model, the models of site-specific substitutions should be much easier to implement and will not suffer from the slowness of computations.

4.7 Population genetics process

Until now, the studies of sequence evolution basically do not have information about population involved. Halpern and Bruno (1998) developed a codon-based model with site-specific residue frequencies. These authors justified their method for interspecific sequence comparison within a population genetics framework. The mathematical formulation in Halpern and Bruno (1998) is similar to ours. It seems possible for us to take advantage of this and to extend our model of RNA evolution to the level of population genetics.

4.8 Detecting potentially interesting sites

We have noticed some sites that are expected to evolve rapidly according to our model, but that actually evolve slowly. This probably indicates that the secondary structure incorporated into our model is an incomplete summary of fitness. It is of interest to determine at which sites our procedure is most inadequate. Sites that are predicted by our model to change but that are actually invariant may be sites that are particularly important with regard to biological function.

4.9 References

Chartrand, P., X. H. Meng, R. H. Singer and R. M. Long. 1999. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. Curr. Biol. **9**:333–336.

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

Dixon, M. T. and D. M. Hillis. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetics analysis. Mol. Biol. Evol. **10**:256–267.

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

Halpern, A. L. and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol. Biol. Evol. **15**:910–917.

Higgs, P. G. 2000. RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **33**:199–253.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**:167–188.

Katz, L. and C. B. Burge. 2003. Widespread selection for local RNA secondary structure in coding regio ns of bacterial genes. Genome Res. **13**:2042–2051.

McCaskill, J. S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**:1105–1119.

Rastegari, B. 2004. Linear time algorithm for parsing RNA secondary structure. Master's thesis, University of British Columbia.

Robinson, D. M., D. Jones, H. Kishino, N. Goldman and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**:1692–1704.

Rocha, E. P., A. Danchin and A. Viari. 1999. Translation in *bacillus subtilis*: Roles and trends of initiation and termination, insights from a genome analysis. Nucl. Acids Res. **27**:3567–3576.

Schmitt, W. R. and M. S. Waterman. 1994. Linear trees and RNA secondary structure. Discrete Appl. Math. **51**:317–323.

Tillier, E. R. M. 1994. Maximum likelihood with multi-parameter models of substitution. J. Mol. Evol. **39**:409–417.

Tillier, E. R. M. and R. Collins. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12**:7–15.

Tillier, E. R. M. and R. A. Collins. 1998. High apparent rate of simultaneous

compensatory basepair substitutions in ribosomal RNA. Genetics **148**:1993–2002.

Workman, C. and A. Krogh. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucl. Acids Res. **27**:4816–4822.

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

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. Science **244**:48–52.