Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Detecting Positively Selected Sites From Amino Acid Sequences: An Implicit Codon Model

Zheng Ouyang and Jie Liang

Abstract— **Fixation of advantageous mutations is an important evolutionary force driving the accelerated protein diversification. However, the standard phylogenetic approach to infer positive selection is based on relative rate of nonsynonymous to synonymous substitutions, and requires the knowledge of DNA sequences, hence precludes its application to family of remotely related sequences where saturated substitutions occur. In this study, we develop a new method to detect positive selection directly from amino acid sequences by treating codon usage as hidden parameters.**

For a given amino acid sequence set and a phylogenetic tree, we use a reversible continuous time Markov process as our evolutionary model. This model has fewer parameters than normal amino acid evolutionary model, with only transition/transversion rate ratio, nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$), and codon usage. Similar to earlier work, we assume that ω is a random variable with different probabilities to take a set of discrete values. Those with $\omega > 1$ model sites under positive selection. We use the Bayesian Monte Carlo method to estimate model parameters, as it allows implementation of complex model of sequence evolution. Here unobserved DNA sequences are sampled from protein sequences based on distributions parametrized by codon usages, based on the fact that both protein sequences and the native protein-encoding DNA sequences have the same phylogenetic tree. The object is that sampled DNA sequences should fit the same phylogenetic tree as well as the native DNA sequences. Data set of β-globin sequences from vertebrates is used to verify our model. We are able to detect all eight positive selection sites, which were originally reported using native nucleotide sequences. Our work shows that although nonsynonymous/synonymous rate ratio is defined at codon level, it can be used to detect selective pressures of amino acid sequences by our implicit codon-based model.

I. INTRODUCTION

Inferring selection pressure at individual amino acid sites provides an important approach for studying the mechanisms of protein evolution and function [23]. Several methods have been developed that can detect functional important sites based on evolutionary conservation [2,7,12,21]. However, high levels of variability also signify functional importance [1,9–11,14,17,25]. The presence of such residues experiencing positive selection can be inferred from the observation that the rate of nonsynonymous nucleotide substitution d_N is higher than that of synonymous substitution d_S in protein-coding genes [6,24]. One expects $\omega = d_N/d_S = 1$ if no Darwinian selection is acting on the DNA sequences; $\omega < 1$ (selection against new mutations) if there is negative selection; and $\omega > 1$ (selection for new mutations) if there is positive selection. Since d_N/d_S ratio is a proxy for the strength of selection, it can be used to search for regions of functional importance. For example, Sawyer *et al.* identified a small segment of the primate TRIM5α protein that experiences positive selection [18], including those involved in mutagenesis, confirming the importance of the segment in species-specific retroviral inhibition. In fact, many proteins have been detected to be under positive selection, which may be involved in immunity against viral attacks, reproduction, and acquirement of new functions after gene duplications [20].

Although several methods have been developed for detection of the presence of positive selection [11,19], they are based on an explicit codon substitution model, and all requires the knowledge of the DNA sequences. This precludes the application of these methods to remotely related protein families, where saturated substitutions occur. Pupko *et al.* developed a method for detecting positive selection from amino acid sequences [15]. However, this counting method is based on the calculation of chemical-distances between residues, and relies on definitions of conservative and radical substitution of residues solely on the physicochemical properties of residues. This approach therefore is subjective, and may lead to ambiguous conclusions.

In this study, we develop a new model to estimate selection pressure and to infer adaptive evolution using amino acid sequences alone as input. Taking a Markovian process as the model of codon substitution, our method can estimate $\omega = d_N/d_S$ ratio at individual amino acid sites through an implicit codon model by translating the amino acid sequences back into likely codon sequences. We use a Bayesian approach to estimate our model parameters, including the ω ratios [5] and the probabilities of the usage of individual codons at an amino acid site.

II. MATERIALS AND METHODS

A. Markov model of codon substitution

We use a reversible continuous time Markov process as our evolutionary model [3]. We assume mutations occur at the three codon positions independently, and therefore only single-nucleotide substitutions to occur instantaneously. Mutations involving more than one position will have very small probabilities of occurrence and will be ignored. At codon level, the states of the Markov process are the set $\mathcal{C}$ of 6l sense codons. The three nonsense (stop) codons are not considered in the model, as mutations to or from stop codons can be assumed to affect drastically the structure and function

This work is supported by grants from NSF (CAREER DBI0133856), NIH (GM68958 and GM079804), and ONR (N000140310329). We thank Dr. Chung-I Wu for helpful discussion.

Department of Bioengineering, SEO, MC-063 University of Illinois at Chicago 851 S. Morgan Street, Room 218 Chicago, IL 60607–7052, U.S.A. jliang@uic.edu

1-4244-0788-5/07/$20.00 ©2007 IEEE

of the protein and therefore will rarely survive. We use a 61 × 61 rate matrix $\boldsymbol{Q}$, whose entries q_{ij} are substitution rates of codons at an infinitesimally small time interval. Specifically, we have: $\boldsymbol{Q} = \{q_{ij}\}$, where the diagonal element is $q_{i,i} = -\sum_{i,j \neq i} q_{i,j}$ so that row sums of $\boldsymbol{Q}$ equal zero. The codon substitution model specifies the relative instantaneous substitution rate from codon i to j at one site is:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 positions,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition.} \end{cases}$$

Parameter κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio, and π_j is the stationary frequency of codon j. In this model, κ and π are common for all sites, and $\omega = d_N/d_S$ ratio vary among sites,

The transition probability matrix of size 61×61 after time t is [8]:

$$\boldsymbol{P}(t) = \{p_{ij}(t)\} = \boldsymbol{P}(0)\exp(\boldsymbol{Q} \cdot t),$$

where $\boldsymbol{P}(0) = \boldsymbol{I}$. Here $p_{ij}(t)$ represents the probability that codon i will mutate into codon j in time interval t. To ensure that the nonsymmetric rate matrix Q is diagonalizable for easy computation of $\boldsymbol{P}(t)$, we follow the reference [22] and insist that $\boldsymbol{Q}$ takes the form of $\boldsymbol{Q} = \boldsymbol{S} \cdot \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix who entries are the composition of codons, and $\boldsymbol{S}$ is a symmetric matrix whose entries need to be estimated.

B. Codon usage and DNA sequence sampling

Because of the degeneracy of the genetic code, the problem of generating a reliable nucleotide sequence from an amino acid sequence of a protein is complex [13]. Since the amino acid sequence has the same phylogenetic tree as the native nucleotide sequence, our aim is to find probable DNA sequences compatible to the amino acid sequences which can fit the phylogenetic tree well. The probability of each compatible DNA sequence will be estimated. This is a more realistic goal than finding the exact native DNA sequence. For amino acid residue type k, let the number of synonymous codon for residue k be s_k. The unequal usage of the s_k codons can be modeled by assigning s_k weights $w_1, \cdots, w_{s_k}$. Since they sum to 1, there are $s_k - 1$ free parameters to be estimated. According to the number of synonymous codon listed in Table I, there are $3 \times (6-1) + 5 \times (4-1) + 1 \times (3-1) + 9 \times (2-1) + 2 \times (1-1) = 41$ free codon usage parameters. We use this set of parameters $\mathcal{W} = (w_1, \cdots, w_{41})$ to sample putative DNA sequences that are compatible to the given amino acid residue sequence.

Amino acid	L,S,R	VAG P,T	I	F,C,Y,Q,N H,D,E,K	M,W
#synonymous codon	6	4	3	2	1

TABLE I: Amino acids and the number of synonymous codon.

C. Likelihood function of a fixed phylogeny.

We assume that a reasonably accurate phylogenetic tree $\boldsymbol{T} = (\mathcal{V}, \mathcal{E})$ is given. Here $\mathcal{V}$ is the set of nodes, namely, the union of the set of observed s sequences $\mathcal{L}$ (leaf nodes), and the set of $s-1$ ancestral sequences $\mathcal{I}$ (internal nodes). $\mathcal{E}$ is the set of edges of the tree. For an alignment of s codon sequences of length n, let the vector $\boldsymbol{x}_h = (x_1, \cdots, x_s)^T$ represent the observed codons at position h for the s sequences, h ranges from 1 to n. Without loss of generality, we assume that the root of the phylogenetic tree is an internal node k. For node k and node l separated by divergence time t_{kl}, the time reversible probability of observing residue x_k in a position h at node k and residue x_l of the same position at node l is:

$$\pi_{x_k} p_{x_k x_l}(t_{kl}) = \pi_{x_l} p_{x_l x_k}(t_{kl}).$$

Given a set $\mathcal{S}$ of s DNA sequences $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ translated from multiple-aligned amino acid sequences of length n based on proposed codon usage $\mathcal{W}$, the specified topology of the phylogenetic tree $\boldsymbol{T}$ and the set of edges, the probability of observing the s number of codons $\boldsymbol{x}_h$ at position h is a sum over all possible codon assignments to the interior nodes of the phylogenetic tree :

$$f(\boldsymbol{x}_h|\mathcal{W}, \boldsymbol{T}, \kappa, \omega_h, \boldsymbol{\pi}) = \pi_{x_k} \sum_{\substack{i \in \mathcal{I} \\ x_i \in \mathcal{C}}} \prod_{(i,j) \in \mathcal{E}} p_{x_i x_j}(t_{ij}).$$

Let $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_n)$. Assuming independence of the substitutions among residues, after summing over the set $\mathcal{C}$ of all possible codon types for the internal nodes $\mathcal{I}$, the probability of observing multiple-aligned amino acid sequences is:

$$\begin{aligned} f(\mathcal{S}|\mathcal{W}, \boldsymbol{T}, \kappa, \boldsymbol{\omega}, \boldsymbol{\pi}) &= f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n|\mathcal{W}, \boldsymbol{T}, \kappa, \boldsymbol{\omega}, \boldsymbol{\pi}) \\ &= \prod_{h=1}^{n} f(\boldsymbol{x}_h|\mathcal{W}, \boldsymbol{T}, \kappa, \boldsymbol{\omega}, \boldsymbol{\pi}). \end{aligned}$$

To detect positive selection, we follow the "M3" model of Yang *et al.* (2000) [24]. There are three categories of the nonsynonymous/synonymous rate ratio ω at each site. Let the proportions of codon sites in the different categories at a site be p_1, p_2, and p_3, where $p_1 + p_2 + p_3 = 1$, and let the corresponding ω be $\omega_1 < \omega_2 < \omega_3$. Categories with $\omega > 1$ model sites under positive selection. Let $\boldsymbol{p}_1 = (p_{1,1}, \cdots, p_{1,n})$, $\boldsymbol{p}_2 = (p_{2,1}, \cdots, p_{2,n})$, and $\boldsymbol{p}_3 = (p_{3,1}, \cdots, p_{3,n})$, and denote $\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3)$. The probability of observing data $\mathcal{S}$ is:

$$\begin{aligned} & f(\mathcal{S}\,|\mathcal{W}, \boldsymbol{T}, \kappa, \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{p}) = \\ & \prod_{h=1}^{n} \left(\sum_{K_i=1}^{3} f(\mathcal{S}_h\,|\mathcal{W}, \boldsymbol{T}, \kappa, \boldsymbol{\pi}, \omega_{K_i}, K_i)\, p_{K_i} \right), \end{aligned} \tag{1}$$

where K_i is the selection category at position h and $\boldsymbol{\pi}$ is the stationary frequency of codon. This can be used to calculate the log-likelihood function $\ell = \log f(\mathcal{S}|\mathcal{W}, \boldsymbol{T}, \kappa, \boldsymbol{\pi}, \boldsymbol{\omega}, \boldsymbol{p})$.

When the tree is given and $\boldsymbol{\pi}$ is fixed to the empirical frequencies in the sampled DNA sequences, this model only has 47 parameters. Among these, $\omega_1, \omega_2, \omega_3, p_1, p_2$ are the 5 site-specific parameters to be estimated, κ is a site-independent parameter, and $\mathcal{W}$ contains 41 parameters. We use $\boldsymbol{\theta}$ to denote all these parameters that need to be estimated.

D. Bayesian estimation

Our goal is to estimate the values of the free parameters. Here we adopt a Bayesian approach, where parameter estimates are obtained from samples drawn from the posterior probability distribution of the parameter.

We use a prior distribution $\pi(\boldsymbol{\theta})$ to encode our past knowledge of the free parameters. We then describe
$btheta$ by a posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{S})$, which summarizes prior information available on
$btheta$ and the information contained in the observations $\mathcal{S}$ of multiple sequence alignment.

After integrating the prior information $\pi(\boldsymbol{\theta})$ and the likelihood function $f(\mathcal{S}|\boldsymbol{\theta}, \boldsymbol{T})$ (Eqn 1), the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{S}, \boldsymbol{T})$ can be estimated up to a constant as:

$$\pi(\boldsymbol{\theta}|\mathcal{S}, \boldsymbol{T}) \propto \int f(\mathcal{S}|\boldsymbol{\theta}, \boldsymbol{T}) \cdot \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Our goal is to estimate the posterior means of parameters in $\boldsymbol{\theta}$:

$$\mathbb{E}_\pi(\boldsymbol{\theta}) = \int \boldsymbol{\theta} \cdot \pi(\boldsymbol{\theta}|\mathcal{S}, \boldsymbol{T}) d\boldsymbol{\theta}.$$

E. Prior distribution

We choose "noninformative" priors that have the smallest effect on the results of the analysis. For ratio parameters κ, the prior is taken as a beta distribution $B(\kappa; \alpha = 1.0, \beta = 1.0)$. The other ratio parameters ω_1, ω_2 and ω_3 are also drawn from the beta distribution $B(\kappa; \alpha = 1.0, \beta = 1.0)$.

The family of Dirichlet distributions is the obvious choice for specifying priors of codon usage $\mathcal{W}$. Dirichlet priors assign densities to groups of parameters that measure proportions (*i.e.*, parameters that must sum to 1). For a specific amino acid type k, the Dirichlet prior of codon usage $\pi(w_1, \cdots, w_{s_k}) = Dir(w_1, \cdots, w_{s_k}|\alpha_1, \cdots, \alpha_{s_k})$ has s_k parameters, each corresponds to the relative frequencies of one synonymous codon. We use uniform Dirichlet priors by setting all α_j to 1.0, which means that every combination of the parameters is assigned the same prior density. The prior distribution $\pi(p_1, p_2, p_3)$ for the selection category of ω frequencies is similarly taken as $Dir(p_1, p_2, p_3|\alpha_1, \alpha_2, alpha_3)$, with the assignment of $\alpha_1 = \alpha_2 = \alpha_3 = 1.0$.

F. Positively selected sites

In this work, we follow an approach similar to Huelsenbeck and Dyer 2004 [5] and take the advantage of a full Bayesian approach to determine the probability that each amino acid site is under positive selection. Assume that $\omega_{K_i=3} > 1.0$. The probability that the i-th codon site is in positive class $K_i = 3$ is obtained by integrating over all possible combinations of transition/transversion rate ratios, nonsynonymous/synonymous rate ratios and codon usage. The likelihood of each site is calculated under several three ω values and then the values are summed to give the site likelihood. The posterior probability of the site being positively selected is the proportion of this sum originating from categories that are positively selected (ω> 1).

$$f\left(K_i = 3|\mathcal{S}\right) = \frac{f(\mathcal{S}|K_i = 3)p_3}{\sum_{j=1}^{3} f(\mathcal{S}|K_i = j)p_j}$$

Sites at which this probability is larger than a threshold value (say, 90, 95, or 99%) are identified as potentially under positive selection.

G. Markov Chain Monte Carlo

Since it is impossible to directly integrate the marginal distributions for posterior probability calculation, we run a Markov chain to generate samples drawn from the target distribution $\pi(\boldsymbol{\theta}|\mathcal{S}, \boldsymbol{T})$. Starting from $\boldsymbol{\theta}_t$ at time t, we generate a new $\boldsymbol{\theta}_{t+1}$ using the proposal function: $T(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$. The proposed new matrix $\boldsymbol{\theta}_{t+1}$ will be either accepted or rejected, depending on the outcome of an acceptance rule $r(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$. Equivalently, we have:

$$\boldsymbol{\theta}_{t+1} = A(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) = T(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) \cdot r(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}).$$

To ensure that the Markov chain will reach stationary state, we need to satisfy the requirement of detailed balance, *i.e.*,

$$\pi(\boldsymbol{\theta}_t|\mathcal{S}, \boldsymbol{T}) \cdot A(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) = \pi(\boldsymbol{\theta}_{t+1}|\mathcal{S}, \boldsymbol{T}) \cdot A(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t).$$

This is achieved by using the Metropolis-Hastings acceptance ratio $r(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$ to either accept or reject $\boldsymbol{\theta}_{t+1}$, depending on whether the following inequality holds:

$$u \leq r(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) = \min\{1, \frac{\pi(\boldsymbol{\theta}_{t+1}|\mathcal{S}, \boldsymbol{T}) \cdot T(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t)}{\pi(\boldsymbol{\theta}_t|\mathcal{S}, \boldsymbol{T}) \cdot T(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})}\},$$

where u is a random number drawn from the uniform distribution $\mathcal{U}[0, 1]$. With the assumption that the underlying Markov process is ergodic, irreducible, and aperiodic, a Markov chain generated following these rules will reach the stationary state [16].

We collect m correlated samples of the $\boldsymbol{\theta}$ matrix after the Markov chain has reached its stationary state. The posterior means of the rate matrix are then estimated as:

$$\mathbb{E}_\pi(\boldsymbol{\theta}) \approx \sum_{i=1}^{m} \boldsymbol{\theta}_i \cdot \pi(\boldsymbol{\theta}_i|\mathcal{S}, \boldsymbol{T}).$$

In this paper, the state space of this Markov chain includes transition/transversion rate ratios κ, codon usage parameters $\mathcal{W}$, probabilities of being in three selection classes and ω values for those selection classes. The chain is constructed by randomly selecting a parameter, proposing a new state for the parameter, and deciding whether the new state is accepted or rejected.

H. Move set

The move set determines the proposal function, which is critical for the rapid convergency of a Markov chain. For parameters ω_1, ω_2 and ω_3, they were initially set to 0.1, 1.0, and 3.0, respectively. Their current values were changed by adding or subtracting with equal probability a random value drawn uniformly from intervals of widths, 0.1, 0.5 and 1.0, respectively. If the proposed new value ω_i^{t+1} for current ω_i^t is outside of the range of allowed values, $|\omega_i^{t+1} - \omega_i^t|$ is added/substracted to ω_i^t so the new value is within the valid range ($10^{-5} < \omega_1 < 0.1$, $0.1<\omega_2<3$, $1<\omega_3< 10^3$). The κ parameter was initialized to 1.0, then took new value as a ratio of two random variables, which sum up to one and are drawn from beta distribution. The valid range is between 0.001 and 1000. For the probabilities of belonging to one of the three classes (p_1, p_2, p_3), the new values $(p_1^{t+1}, p_2^{t+1}, p_3^{t+1})$ are also sampled from a Dirichlet distribution $Dir((p_1, p_2, p_3|\alpha_1, \alpha_2, \alpha_3)$, where α_i is set to $p_i^t / \sum_i p_i^t$. Similarly, the parameters $\mathcal{W}^{t+1} = (w_1^{t+1}, \cdots, w_{s_k}^{t+1})$ for codon usage are drawn from $Dir(w_1, \cdots, w_{s_k}|\alpha_1, \cdots, \alpha_{s_k})$, where α_i is set to $w_i^t / \sum_i w_i^t$. Since the sampling space of codon usage is much larger than that of other model parameters, we assigned 65% of the moves for changes in $\mathcal{W}$, and 35% of the moves for all other model parameters. We assume residues in sequences belonging to the same species are likely to adopt similar codon usage, we constrain that amino acids in one sequence has the same codon usage. This improves the mixing of the Markov process. Further improvement can be obtained by assigning initial values of codon usage parameters $\mathcal{W}$ based on statistics of codon frequency in genomic sequences.

I. Dataset

We use a dataset of 17 β-globin sequences from vertebrate species. Each sequence contains 144 amino acids. Protein sequences were translated from the dataset which was originally collected by Yang et al. [24, 26] from the EMBL and GenBank databases. Phylogenetic tree was built by ProML [4], which implements a maximum likelihood estimator for protein amino acid sequences. To reduce the computational complexity, we fix the phylogenetic tree during entire Markov process. The amino acid sequences used here is available at (`http://gila.bioengr.uic.edu/lab/dataset/beta-globin.fasta`).

III. Results

This method was implemented in C, and part of the data structure, functions for matrix operation and statistical distributions were adapted from the program MrBayes v3.1 [5]. Multiple amino acid sequences alignment of β-globin sequences and the phylogenetic tree were used as input, and the algorithm was run with different number of Markov moves ranging form 10,000 to 400,000. Samples were taken for every 100th moves. The first 200,000 samples were discarded as the chain is still in the "burning-in" period. In this study, we only collect samples if there are positively selected sites with posterior probability > 0. Figure 1 shows the log probability of observing the data for each sampled state at different time steps. The chain was started with a fixed tree and branch lengths. The likelihood of the initial state is poor, and the chain quickly found parameters that could explain the data better. After about 225,000 steps, a plateau in the log likelihood is reached. The probabilities that each site was under the three different types of selection pressure were calculated for each sample. Figure 2 shows the average posterior probability of individual site under positive selection. When the threshold for the posterior probability is set to 99.9%, the detected positively selected sites are completely in agreement with results of Yang *et al*, which were derived from native nucleotide sequences (Table II). We also test protein sequences of HIV-1 env V3 region. The DNA sequences were analyzed by Yang et al. (2000) and sites 28, 66, and 87 are identified as under positive selection pressure with high posterior probability. Among these, our method detected site 28 to be under significant positive selection pressure (p>98%), and site 66 under weak positive selection pressure.

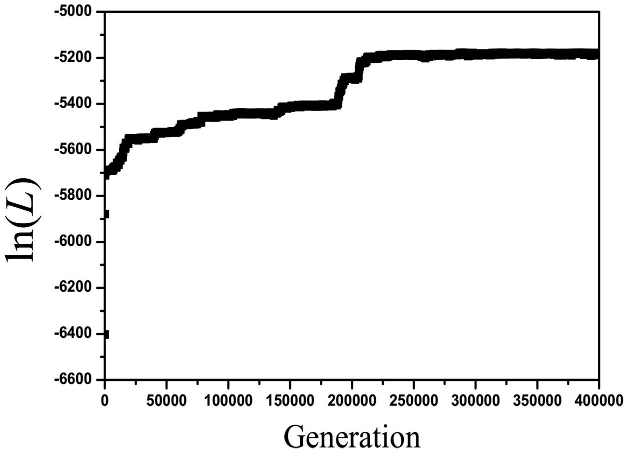

Fig. 1: The log likelihood of the current state over the course of the MCMC analysis.

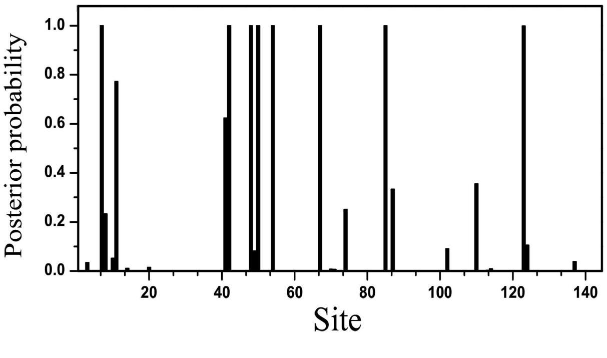

Fig. 2: The average posterior probability of each site being in the positively selected class

Site	Average Posterior Probability	
	MCMC(100,000 Samples)	MCMC(400,000 Samples)
*7	0.996870	0.999973
*42	0.999477	0.999979
*48	0.999562	0.999830
*50	0.999998	0.999990
*54	0.973350	0.999897
*67	0.999987	0.999521
70	0.977635	0.006991
74	0.897890	0.251067
*85	0.993659	0.999806
87	0.999981	0.333273
110	0.597621	0.355364
*123	0.999434	0.999060

TABLE II: Positively selected sites with high average posterior probability.

IV. CONCLUSION

Studying the evolutionary history of proteins is an essential task. A powerful method for detecting selection pressure is considering the ratio of synonymous and non-synonymous substitutions. However, the applicability of this method is limited, as it cannot be applied to remotely related proteins when saturated substitutions occur. In addition, it requires the availability of DNA sequences, which are often difficulty to obtain in practice. Currently, no method is available that detects selection pressure based on this ratio using amino acid sequences alone. In the current work, we have developed an implicit codon model to infer positive selections directly form amino acid sequences at relative high accuracy level. Our method generates possible underlying DNA sequences from known protein sequences. For a given amino acid sequence set and a phylogenetic tree, a reversible continuous time Markov process was used as the evolutionary model at the amino acid level. Data set of β-globin sequences from vertebrates was used to verify our method. We are able to detect all of the eight residue sites known to experience positive selection, which were originally derived from native nucleotide sequences by Yang *et al*. For this test, our model is as effective as the traditional DNA sequence based method, but with the promise of much wider applicability. This work opens a new way to detect selection pressure by examining directly protein sequences, which are far easier to obtain than DNA sequences. Although the rate ratio of nonsynonymous and synonymous substitution is defined at codon level, our work showed that it can be generalized for detecting selective pressures using amino acid residues sequences by applying our Bayesian Monte Carlo method on an implicit codon-based model.

REFERENCES

[1] J. Bielawski, K. Dunn, G. Sabehi, and O. Beja, "Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, pp. 14 824–9, 2004.

[2] G. Casari, C. Sander, and A. Valencia, "A method to predict functional residues in proteins." *Nat. Struct. Biol.*, vol. 2, pp. 171–8, 1995.

[3] Felsenstein, "Evolutionary trees from dna sequences: a maximum likelihood approach." *J. Mol. Evol.*, vol. 17, pp. 368–76, 1981.

[4] J. Felsenstein, "Maximum-likelihood estimation of evolutionary trees from continuous characters." *Am. J. Hum. Genet.*, vol. 25, pp. 471–92, 1973.

[5] J. Huelsenbeck and K. Dyer, "Bayesian estimation of positively selected sites." *J. Mol. Evol.*, vol. 58, pp. 661–72, 2004.

[6] M. Kimura, "The neutral theory of molecular evolution." *Cambridge University Press, Cambridge*, 1983.

[7] L. Li, E. Shakhnovich, and L. Mirny, "Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, pp. 4463–8, 2003.

[8] P. Lio and N. Goldman, "Models of molecular evolution and phylogeny." *Genome. Res.*, vol. 8, pp. 1233–44, 1998.

[9] T. Massingham and N. Goldman, "Detecting amino acid sites under positive selection and purifying selection." *Genetics*, vol. 169, pp. 1753–62, 2005.

[10] R. Nielsen and J. Huelsenbeck, "Detecting positively selected amino acid sites using posterior predictive P-values." *Pac. Symp. Biocomput.*, pp. 576–88, 2002.

[11] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene." *Genetics*, vol. 148, pp. 929–36, 1998.

[12] A. Panchenko, F. Kondrashov, and S. Bryant, "Prediction of functional sites by analysis of sequence and structure conservation." *Protein. Sci.*, vol. 13, pp. 884–92, 2004.

[13] G. Pesole, M. Attimonelli, and S. Liuni, "A backtranslation method based on codon usage strategy." *Nucleic. Acids. Res.*, vol. 16, pp. 1715–28, 1988.

[14] O. Podlaha and J. Zhang, "Positive selection on protein-length in the evolution of a primate sperm ion channel." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, pp. 12 241–6, 2003.

[15] T. Pupko, R. Sharan, M. Hasegawa, R. Shamir, and D. Graur, "A chemical-distance-based test for positive darwinian selection," *WABI '01: Proceedings of the First International Workshop on Algorithms in Bioinformatics*, pp. 142–155, 2001.

[16] C. P. Robert and G. Casella., "Monte carlo statistical methods," *Springer-Verlag Inc., New York.*, 2004.

[17] R. Sainudiin, W. Wong, K. Yogeeswaran, J. Nasrallah, Z. Yang, and R. Nielsen, "Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system." *J. Mol. Evol.*, vol. 60, pp. 315–26, 2005.

[18] S. Sawyer, L. Wu, M. Emerman, and H. Malik, "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, pp. 2832–7, 2005.

[19] Y. Suzuki and T. Gojobori, "A method for detecting positive selection at single amino acid sites." *Mol. Biol. Evol.*, vol. 16, pp. 1315–28, 1999.

[20] R. Tarrio, F. Rodriguez-Trelles, and F. J. Ayala, "A new drosophila spliceosomal intron position is common in plants," *Proc. Natl. Acad. Sci. USA*, vol. 100(11), pp. 6580–6583, 2003.

[21] Y. Tseng and J. Liang, "Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach." *Mol. Biol. Evol.*, vol. 23, pp. 421–36, 2006.

[22] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." *Mol. Biol. Evol.*, vol. 18, pp. 691–9, 2001.

[23] Z. Yang, "The power of phylogenetic comparison in revealing protein function." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, pp. 3179–80, 2005.

[24] Z. Yang, R. Nielsen, N. Goldman, and A. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites." *Genetics*, vol. 155, pp. 431–49, 2000.

[25] Z. Yang, R. Nielsen, and M. Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution." *Mol. Biol. Evol.*, vol. 15, pp. 1600–11, 1998.

[26] Z. Yang and W. Swanson, "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes." *Mol. Biol. Evol.*, vol. 19, pp. 49–57, 2002.