A Bayesian Framework for SNP Identification

B.M. Webb-Robertson, S.L. Havre, and D.A. Payne

Pacific Symposium on Biocomputing 10:421-432(2005)

A BAYESIAN FRAMEWORK FOR SNP IDENTIFICATION

B.M. WEBB-ROBERTSON, S.L. HAVRE

Computational Biology & Bioinformatics, Pacific Northwest National Laboratory
Richland, WA 99352, USA

D.A. PAYNE

Information Analytics, Pacific Northwest National Laboratory
Richland, WA 99352, USA

As evolutionary models for single-nucleotide polymorphisms (SNPs) become available, methods for using them in the context of evolutionary information and expert prior information is a necessity. We formulate a probability model for SNPs as a Bayesian inference problem. Using this framework we compare the individual and combined predictive ability of four evolutionary models of varying levels of specificity on three SNP databases (two specifically targeted at functional SNPs) by calculating posterior probabilities and generating Receiver Operating Characteristic (ROC) curves. We discover that none of the models do exceptionally well, in some cases no better than a random-guess model. However, we demonstrate that several properties of the Bayesian formulation improve the predictability of SNPs in the three databases, specifically the ability to utilize mixtures of evolutionary models and a prior based on the genetic code.

1. Introduction

Interest in single-nucleotide polymorphisms (SNPs) has exploded in recent years. This interest is evident from the large number of SNP databases publicly available on the web: NCBI SNP database (http://www.ncbi.nlm.nih.gov/SNP), Human SNP database (http://www.broad.mit.edu/snp/human/), hemoglobin database (http://globin.cse.psu.edu), SNP Consortium Ltd. (http://snp.cshl.org/), and many others. The compilation of SNPs is vital to studying important biological problems such as the identification of biomarkers for disease and evolution at the molecular level. High-throughput technologies, such as proteomics via mass spectrometry (MS) hold promise to identify SNPs rapidly at a global scale, but identification by these approaches using brute force is computationally unattractive. Thus, accurate methods to assign probabilities to potentially polymorphic sites are necessary.

Evolutionary information is generally captured by estimating model parameters associated with a set of biosequences, DNA [1-5] or proteins [6-7], from one or multiple organisms. These models are then used in the context of some specific framework, for example phylogenetics [8-9] or sequence alignment [10-12]. Limited evolutionary models have been developed for

SNPs, capturing information at the amino acid [13] and codon [14] levels. No framework for assigning probabilities to individual polymorphic events exists and thus little comparison of both SNP specific and general evolutionary models has been performed.

This paper presents the general framework for SNP identification in the context of Bayesian inference through the use of evolutionary models to assign probabilities to all possible SNPs (confined by an application to MS). The Bayesian framework allows both individual and mixtures of evolutionary models to be used. Additionally, it allows for the injection of additional information in the form of a prior. We demonstrate the Bayesian framework by comparing four specific evolutionary models (one SNP specific [13], two nucleotide evolutionary rate matrices [1,4], and one inter-species amino acid model [7]) on three SNP databases. The first two are databases of disease causing or enhancing SNPs for the human proteins hemoglobin [15-17] and p53 [18-19]. The third is a set of genes that characterize SNPs in eight inbred mouse strains [20]). Lastly, we explore the benefits observed from the Bayesian formulation related to the use of mixtures of evolutionary models and the use of the genetic based prior in comparison to a neutral based prior.

2. Methods

Bayesian statistics is an attractive approach for making probabilistic inferences from biological data because it supports the injection of information related to the data, for example, expert opinion or evolution constraints based on sequence composition or length [21-22]. Due to the uncertainty associated with biological data and the frequent availability of expert opinion, many biological problems can be more easily modeled by Bayesian methods than by other approaches.

The Bayesian framework for SNPs attempts to quantify the belief that a nucleotide at a given position in a genetic sequence underwent a polymorphism. We model the problem at the codon level to observe both individual nucleotides and amino acids. In a Bayesian formulation both the observed and unobserved data are treated as random variables. The general formulation defines the annotated genomic data (G) as the observed data. The unobserved data are the codons (S) and two types of background information – an evolutionary model (Ψ) and a mutational descriptor (M).

2.1. *Background Information*

Evolutionary information is depicted by matrices at either the amino acid or nucleotide level describing the likelihood of one residue being substituted by

another. In this study four evolutionary models (Ψ) are evaluated for comparative value. We first describe each of these four models; two amino acid and two nucleotide. Subsequently, we describe the SNP mutation variable (*M*).

Amino Acid Matrices. The first model is from the *BLOSUM* [7] series of scoring matrices commonly used in sequence alignment. This series is generated from a large set of sequences from multiple species at various levels of sequence identity and thus represents a complex ancient history over speciation. It is believed to be inappropriate for intra-species evaluation so a less divergent matrix, *BLOSUM80* (referred to as *BL80*), is included for comparative value. The second model is a newly developed substitution matrix by Majewski and Ott [13] (referred to as *M-O*). It is based on identified SNPs in the human genome and thus captures recent evolutionary changes.

Nucleotide Matrices. The last two models are based on continuous-time Markov chain models that describe the evolutionary rate of substitution between two nucleotides [9]. A fully parameterized model, a 4x4 rate matrix, requires the estimation of 12 parameters (16 possible substitutions minus the four changes to the same nucleotide). To reduce the parameterization several nested evolutionary models have been developed based on possible transitions between purines and pyrimidines at various levels [1-4]. The number of parameters estimated and their estimation values are dependent upon two factors – the data selected and the model. We use the parameter values estimated by Suchard et al. [9] for two models reflecting different levels of evolution. The first of these is the Tamura and Nei model (*TN93*) [4], parameterized into three rates based on data from the "Tree of Life", representing organisms across all living kingdoms. The last model is the Hasegawa et al. model [1], which calculates the rate matrix for each codon position using two parameters based on primate data, resulting in a less general model.

Mutation Variable. The background information (*M*) is a binary variable that describes a detectable SNP event; defined as a mass changing substitution. Undetectable SNPs include silent mutations (SNPs resulting in no change at the amino acid level) and mutations between leucine (*L*) and isoleucine (*I*) (whose mass is indistinguishable by MS). Additionally, we define mutations to and from *STOP* codons as invalid, assuming that such mutations are typically detrimental to the protein. Additionally, we assume that an observed amino acid change is the result of one SNP per codon and not from double mutations. Thus, given two codons, c_i and c_j, where $a_{(i)}$ and $a_{(j)}$ define their respective amino acids, a valid SNP is expressed explicitly as,

$$M(c_i, c_j) = \begin{cases} 1 & \text{if } c_i \text{ and } c_j \text{ differ by only one nucleotide and} \\ & m(a_{(i)}) \neq m(a_{(j)}) \text{ and } a_{(i)} \neq a_{(j)} \neq STOP \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $m(a_{(i)})$ and $m(a_{(j)})$ are the masses associated with $a_{(i)}$ and $a_{(j)}$, respectively. There are several benefits of defining the mutation variable in this manner. For example, in lieu of a binary definition, probabilities could be defined for SNP events based on mass difference. Also, multiple types of peptide variants could be defined, for example M_1 for SNPs, M_2 for frameshifts, and M_3 for multiple nucleotide polymorphisms.

2.2. *Bayesian Formulation*

In our basic Bayesian formulation (1) the genomic data (G) consists of I codons, (2) the codon substitution (S) represents a mutation to a codon j, j=1,…,64 (s_j), (3) the evolutionary model (Ψ) describes the probability ratio or rate of mutation between two residues, and (4) the mutation variable (M) describes the probability of a substitution between two codons. The Bayesian formulation is described as the joint distribution of the observed and unobserved data: $P(G,S,M,\Psi)$. This is the product of the likelihood (L) and the prior (P):

$$P(G, S, M, \Psi) = L(S, M, \Psi; G)P(S, M, \Psi).$$

The likelihood is the probability of the observed data given the unobserved data:

$$P(G, S, M, \Psi) = P(G \mid S, M, \Psi)P(S, M, \Psi). \quad (2)$$

The prior $P(S,M,\Psi)$ can be decomposed into easily calculable probabilities. The evolutionary model Ψ does not change based on the genetic code or the type of mutation being observed. Thus, we assume independence from S and M: $P(S,M,\Psi)=P(S,M)P(\Psi)$. Lastly, returning to the genetic code, we observe that the probability of observing a given codon is dependent on M, and given that there is only one type of mutation event in this case: $P(S,M)P(\Psi)=P(S|M)P(\Psi)$. Hence the Bayesian formulation observed in Eq. 2 can be expressed as:

$$P(G, S, M, \Psi) = P(G \mid S, M, \Psi)P(S \mid M)P(\Psi). \quad (3)$$

The calculations for implementation occur at the individual codon level. Thus, the above general representation can be given in terms of the individual elements of the observed and unobserved data. The joint distribution of a specific mutation in the genome, observing codon j at the i^{th} position in the genome given a specified evolutionary model, ψ_k, is defined as:

$$P(g_i, s_j, M, \psi_k) = P(g_i \mid s_j, M, \psi_k) P(s_j \mid M) P(\psi_k). \quad (4)$$

The joint distribution in Eq. 4 allows easy calculation of posterior probabilities of interest; for example, specific SNPs describing the probability of observing a SNP in the form of codon s_j at the i^{th} position in the genome. Given a specific evolutionary model, ψ_k, this is formulated in terms of Bayes theorem:

$$P(s_j \mid g_i, \psi_k) = \frac{P(g_i \mid s_j, M, \psi_k) P(s_j \mid M) P(\psi_k)}{\sum_j P(g_i \mid s_j, M, \psi_k) P(s_j \mid M) P(\psi_k)}. \quad (5)$$

Additionally, the Bayesian formulation allows the probability of observing a specific SNP independent of the evolutionary model to be calculated:

$$P(s_j \mid g_i) = \frac{\sum_k P(g_i \mid s_j, M, \psi_k) P(s_j \mid M) P(\psi_k)}{\sum_k \sum_j P(g_i \mid s_j, M, \psi_k) P(s_j \mid M) P(\psi_k)}. \quad (6)$$

To obtain these values of interest, the likelihood and priors much each be calculated.

The Likelihood. There is no loss in information by transforming evolutionary models in the form of symmetric 4x4 or 20x20 matrices into 64x64 codon matrices. We perform this conversion for consistency to make the likelihood calculation straight forward. Accordingly, ψ_k describes the probability ratio or rate of substitution between two codons g_i and s_j. The likelihood also includes the mutation event variable, Eq. 1, which allows only valid SNPs to have non-zero probabilities. Thus, the likelihood can be described as:

$$P(g_i \mid s_j, M, \psi_k) = \psi_k(g_i, s_j) * M(g_i, s_j).$$

The Prior. There are two prior in Eqs 3-6, where $P(\psi_k)$ is the prior belief that the codon mutation model ψ_k fits the data and $P(s_j|M)$ is the probability of observing a given codon s_j (independent of the genomic data) given a mutation of type M. The prior on the evolutionary model can be either defined by the user or *a priori*. We assume *a priori* – all models are equally likely. By assuming that mutations at all positions in the genome are equally likely, the prior $P(s_j|M)$ can be defined directly from the genetic code. Because there are 64 codons, there are 576 possible SNPs between all codons, 385 of these are valid as described by M, Eq. 1. The prior probabilities are calculated for each codon, observing that a given codon, $P(s_j|M)$, can only result from a SNP to nine other possible codons. As illustrated in Figure 1 [14], of the nine possible

SNPs that could result in the codon *AGA*, two are products of silent mutations and one is a *STOP* codon. Thus, in this example, the probability of observing *AGA* given *M* (all possible valid SNPs) is $P(AGA|M)={}^{6}/_{385}$. Alternatively Bayes theorem can be used to arrive at this answer in perhaps a more intuitive manner:

$$P(s_j \mid M) = \frac{P(M \mid s_j)P(s_j)}{P(M)} = \frac{P(M \mid s_j)P(s_j)}{\sum_j P(M \mid s_j)P(s_j)}.$$

The benefit of this approach is that it incorporates prior information on the probability of observing a given codon, $P(s_j)$. For instance, codon frequency information on a specific species could be incorporated.

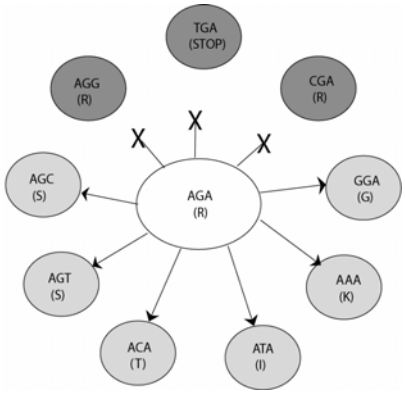

Figure 1. A schematic of the nine possible codons and corresponding amino acids from a SNP to codon *AGA*. In this case there are 6 nonsynonymous SNPs (resulting in an amino acid change), 2 synonymous SNPs, and 1 SNP from a *STOP* codon.

3. Results and Discussion

The human ability to treat disease has developed man beyond a simple 'survival of the fittest' species. Accordingly, due to the potential for identifying biomarkers to treat disease, the majority of SNP data is human focused. We focus on two of these compilations for the proteins hemoglobin [15-17] and p53 [18-19]. These SNPs are functionally important SNPs as they are disease causing or enhancing. Additionally, we observe a third database that characterizes observable SNPs in eight inbred mouse strains [20]. We first discuss the specifics of each SNP databases. Secondly, we evaluate the predictive ability of each individual evolutionary model using the Bayesian framework. Lastly, we assess the benefits of the Bayesian formulation,

specifically the inclusion of the genetic code based prior and the posterior based on mixtures of evolutionary models.

3.1. *The SNP databases*

Hemoglobin is a protein that transports oxygen from the lungs to the peripheral tissues to maintain the viability of cells. It has been largely studied because diseases such as sickle cell anemia have been linked to variants of the protein. The genomic sequence (including the α and β chains) consists of 287 codons, resulting in 1871 valid SNPs. The "*Syllabus of Human Hemoglobin Variants*" [15] is a comprehensive listing of all known human hemoglobin variants. Approximately 541 (29%) of the possible 1871 SNPs are represented in this databases (http://globin.cse.psu.edu).

The p53 protein is the result of a tumor suppressor gene located on human chromosome 17. The p53 gene has been largely studied because mutations of this gene are often accompanied by cancer. This protein is 393 codons in length and has 2559 valid SNPs. From the database, 776, or approximately 30%, of the 2559 SNPs are represented (http://p53.curie.fr/).

The Mouse database contains SNPs identified in eight inbred strains of mice. Although its generation is quite different than that of freely mating populations, it is included for comparative value to determine if any of the defined evolutionary models hold predictive power despite the forced inbreeding. This database covers a much larger genomic space than the human protein-specific databases. The database contains 1307 sequences with 71,798 codons yielding 439,202 possible valid SNPs. Only 1822 SNPs, or 0.4%, of the valid SNPs are represented (http://www.broad.mit.edu/snp/moue).

3.2. *Predictive Ability of Individual Evolutionary Models*

Each database has one or more genes associated with it; from these genes all valid SNPs can be calculated. Each database consists of a list of observed SNPs (a subset of all valid SNPs), which are presumed to be true positives. All the remaining SNPs not represented in a database are assumed to be true negatives. The hemoglobin, p53, and mouse databases have 541, 776, and 1881 true positives and 1330, 1783, and 437,380 true negatives, respectively. To observe the predictive capability of each evolutionary model with the Bayesian framework we use it as a classifier. Given the probabilities assigned to each SNP from the Bayesian model, the true positives and negatives can be used to generate Receiver Operating Characteristic (ROC) curves [23] for each database.

The ROC curve gives a graphical representation of the trade-off between sensitivity and specificity. The plot displays the false positive rate (ratio of false positive to total negatives) versus the true positive rate (ratio of true positives to total positives) at all possible cut-off values. A completely random predictor would give a straight line at a 45° angle – *TP rate* equal to *FP rate*. This is the *Baseline* model. Figure 2 shows ROC curves generated from the individual SNP posterior probabilities (Eq. 5) obtained for each of the four described evolutionary models. Since there are many SNPs with the same posterior probability we generate points on the ROC curve by randomly shuffling the order of the SNPs within any given probability 100 times and display the average.

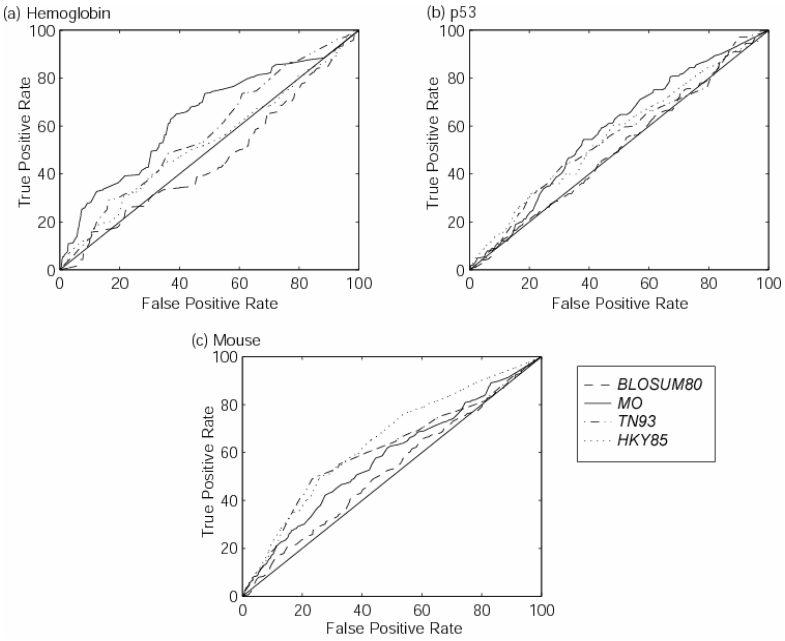

Figure 2. The ROC curves for the (a) hemoglobin, (b) p53, and (c) mouse databases to assess the evolutionary models (1) *BL80*, (2) *M-O*, (3) *TN93*, and (4) *HKY85*.

At first glance, Figure 2 appears to show that the *M-O* model performs the best for both of the disease databases and the *HKY85* model performs best for the mouse database. None of the evolutionary models are exceptionally good classifiers, however it is of interest to compare and contrast these models. The

area under the ROC curve is the most common measure used to compare the discriminative effectiveness of classifiers. DeLong et al. [24] is a standard statistical test for comparing correlated ROC curves that have a large sample size. Tables 1-3 display the area under the ROC curve for each of the evolutionary models and the associated *p*-values. The largest of each database is highlighted in bold.

Table 1. The area under the ROC curves (in parentheses) and *p*-values for primary model comparison for the hemoglobin SNP database.

	BL80 (0.461)	***M-O* (0.639)**	*TN93* (0.571)	*HKY85* (0.528)
Baseline (0.500)	0.0111	<0.0001	<0.0001	0.0586
BL80 (0.461)		<0.0001	<0.0001	0.0008
***M-O* (0.639)**			0.0001	<0.0001
TN93 (0.571)				0.0074

Table 2. The area under the ROC curves (in parentheses) and *p*-values for primary model comparison for the p53 SNP database.

	BL80 (0.507)	***M-O* (0.580)**	*TN93* (0.550)	*HKY85* (0.561)
Baseline (0.500)	0.5563	<0.0001	<0.0001	<0.0001
BL80 (0.507)		0.0001	0.0045	0.0009
***M-O* (0.580)**			0.0518	0.2177
TN93 (0.550)				0.4230

Table 3. The area under the ROC curves (in parentheses) and *p*-values for primary model comparison for the mouse SNP database.

	BL80 (0.527)	*M-O* (0.577)	*TN93* (0.609)	***HKY85* (0.646)**
Baseline (0.500)	0.0005	<0.0001	<0.0001	<0.0001
BL80 (0.524)		<0.0001	<0.0001	<0.0001
M-O (0.577)			0.0006	<0.0001
TN93 (0.609)				<0.0001

Applying a standard threshold of 0.05, all models are significantly different than the *Baseline* model, except *HKY85* for hemoglobin and *BL80* for p53. Additionally, the low *p*-value observed for *BL80* on hemoglobin is a result of this model performing significantly worse than *Baseline* (Figure 2a). The most surprising result is that the human amino acid model does not outperform either of the nucleotide models on mouse despite many mouse genes being orthologous to human. In fact the global *TN93* model has a consistent performance on all three databases. It has the second largest area under the ROC curve for both hemoglobin and mouse. This is significant since in practice evolutionary models specific to the organism under study may not be available. The parameters for *TN93* were generated from species drawn from eukaryotes, eubacteria, halobacteria, and eocytes.

3.3. *Assessing the Benefits of the Bayesian Formulation*

The Bayesian formulation has several benefits, specifically the ability to obtain a posterior probability using multiple evolutionary background information parameters and the ability to define priors. It has been shown in sequence alignment that sensitivity can be improved by summing over multiple substitution matrices [12]. We assess the improvement in sensitivity observed from mixtures of evolutionary models, as well as when the codon defined prior is used in lieu of a neutral defined prior.

Inclusion of Multiple Evolutionary Models. ROC curves were constructed using posterior probabilities calculated for the eleven possible model combinations (Eq. 6). The areas under the ROC curves were compared to the best individual model using DeLong et al. [24]. Table 4 gives these results for p53 and mouse – no model combinations gave an area under the ROC curve greater than *M-O* for hemoglobin. Not surprisingly, the predictability of the *M-O* model, tuned specifically to human SNP data, is not improved for hemoglobin or p53 by including additional models. For mouse, two mixture models perform significantly better than the best individual model; (1) *M-O* and *HKY85*, and (2) *BL80*, *M-O*, and *HKY85*. It appears that the decision to utilize multiple models may be subjective to the evolutionary distance between the organism being studied and the organisms used to generate the models.

Table 4. Comparison of the area under the ROC curve for each model combination that had a larger area than the best individual model and the corresponding *p*-values.

	p53 (*M-O* = 0.580)		Mouse (*HKY85* = 0.646)	
	ROC area	*p*-value	ROC area	*p*-value
M-O/TN93	0.586	0.6241	0.634	
M-O/HKY85	0.588	0.5284	**0.655**	**0.0001**
BL80/M-O/HKY85	0.586	0.6244	**0.654**	**0.0008**
M-O/TN93/HKY85	0.594	0.2166	0.650	0.3222
All	0.592	0.3139	0.647	0.7798

The Prior. The Bayesian SNP model includes a prior on the probability of a codon given a valid SNP, $P(s_j|M)$. This prior is based on the genetic code (Figure 1). Table 5 gives the area under the ROC curve for the genetic code defined prior and for a neutral prior (all codons are equally likely to undergo mutation), as well as the *p*-value comparing the ROC area difference. The results are startling. In all cases the area under the ROC curve for the neutral prior is less than or equal to the genetic code prior, in most cases returning significant *p*-values. Of the twelve generated *p*-values, nine have a *p*-value of less than 0.1. Thus, the inclusion of this prior is beneficial to the overall predictability of the model.

Table 5. The area under the ROC curve generated from the genetic code (C prior) and the *neutral* (N prior) and associated *p*-values.

	Hemoglobin			p53			Mouse		
	C Prior	N Prior	*p*-value	C Prior	N Prior	*p*-value	C Prior	N Prior	*p*-value
BL80	0.461	0.437	<0.01	0.507	0.486	<0.01	0.524	0.525	0.72
M-O	0.639	0.627	<0.01	0.580	0.563	<0.01	0.577	0.577	0.34
TN93	0.571	0.540	<0.01	0.550	0.514	<0.01	0.609	0.597	<0.01
HK85	0.528	0.525	0.21	0.561	0.544	<0.01	0.646	0.645	0.40

4. Conclusions

We propose a Bayesian methodology for assigning posterior probabilities to individual SNPs. We evaluate the model using posterior probabilities associated with one or more evolutionary models on three databases of functional SNPs. These probabilities were used to classify the SNPs represented in each database and observe the sensitivity versus specificity (ROC curves). We observe that none of the models hold strong predictive power (Figure 2). Not surprisingly, the best single model for hemoglobin and p53 is *M-O*, which was generated from human SNP data (Tables 1 and 2). Surprisingly, both nucleotide models outperform the amino acid *M-O* model for mouse; the largest area under all ROC curves for the individual models was 0.646 on *HKY85* for mouse (Table 3).

We also demonstrate that two properties unique to the Bayesian framework improve SNP identification. First, the Bayesian formulation allows inference to be made over mixtures of evolutionary models. Given the specialty of *M-O* to hemoglobin and p53 no improvement in specificity was observed, but for the mouse database two mixture models found an area under the ROC curve that was significantly better than the best individual model *HKY85* (Table 4). Finally, we focus on the prior, a special feature of the Bayesian model. We define our prior as the probability of observing a specific codon given the SNP model from the genetic code. We compare the results of applying our prior to the results using the neutral prior. We observe that the area under the ROC curve for the neutral prior is always smaller than or equal to the genetic code based prior. Furthermore, 75% of the time the area under the curve is significantly smaller for the neutral prior at a *p*-value of 0.05 (Table 5). Although none of the evolutionary models were highly accurate predictors, the Bayesian formulation gives a framework under which prior knowledge or more advanced evolutionary models can be incorporated to assign probabilities to individual polymorphic sites.

Acknowledgments

This work was supported by the U.S. Department of Energy (DOE) through the Computational Sciences and Engineering Initiative Laboratory Directed Research and Development program at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated by Battelle Memorial Institute for the U.S. DOE under contract DE-AC06-76RLO 1830.

References

1. M. H. Hasegawa et al., *J. Mol. Evol.* **22**, 160-74 (1985).
2. M. Kimura, *J. Mol. Evol.* **16**, 111-20 (1980).
3. M. Kimura, *PNAS.* **78**, 454-58 (1981).
4. K. Tamura and M. Nei, *Mol. Biol. Evol.* **10**, 512-26 (1993).
5. Z. Yang and R. Nielsen, *Mol. Biol. Evol.* **17**, 32-43 (2000).
6. M. O. Dayhoff, R. M. Schwartz and B. C. Orcutt, *Atlas of Pro. Seq. Str.* 345-52 (1978).
7. S. Henikoff and J. G. Henikoff, *PNAS.* **11**, 725-36 (1994).
8. T. R. Buckley and C. W. Cunningham, *Mol. Biol. Evol.* **19**, 394-405 (2002).
9. M. A. Suchard et al., *Mol. Biol. Evol.* **18**, 1001-13 (2001).
10. S. F. Altschul et al., *J. Mol. Biol.* **215**, 403-10 (1990).
11. T. F. Smith and M. S. Waterman, *J. Mol. Biol.* **147**, 195-97 (1981).
12. B. M. Webb et al., *Nucleic Acids Res.* **30**, 1268-77 (2002).
13. J. Majewski and J. Ott, *Gene.* **305**, 167-73 (2003).
14. N. Goldman and Z. Yang, *Mol. Biol. Evol.* **11**, 725-36 (1994).
15. T. H. J. Huisman et al., *A Syllabus of Human Hemoglobin Variants* (1996).
16. R. Hardison et al., *Genomics* **47**, 429-37 (1998).
17. R. Hardison, et al., *Hemoglobin* **22**, 113-27 (1998).
18. C. Beroud, et al., *Hum. Mutat.* **15**, 86-94 (2000).
19. T. Soussi, et al., *Hum. Mutat.* **15**, 105-13 (2000).
20. K. Linblad-Toh et al., *Nat. Genet.* **24**, 381-86 (2000).
21. R. D. Knight et al., *Genome Biol.* **2**, (2001).
22. D. J. Lipman et al., *BMC Evol. Biol.* **2**, (2002)
23. J. P. Eagen, (1975). Signal Detection Theory and ROC Analysis (New York: Academic Press.
24. E. R. DeLong et al., *Biometrics* **44**, 387-45, (1988).