Control of the False Discovery Rate Applied to the Detection of Positively Selected Amino Acid Sites

Stéphane Guindon, Mik Black, Allen Rodrigo

Control of the false discovery rate applied to the detection of positively selected amino acid sites

Stéphane Guindon[1,2,*], Mik Black[1,3] & Allen Rodrigo[1,2]

[1] Bioinformatics Institute, University of Auckland, New Zealand.

[2] Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, New Zealand.

[3] Department of Statistics, University of Auckland, New Zealand.

Corresponding author :

Stéphane Guindon.

Bioinformatics Institute and Allan Wilson Centre for Molecular Ecology and Evolution.

University of Auckland.

Auckland, New Zealand.

s.guindon@auckland.ac.nz

Tel. +64-9-3737-599 Ext 83773

Abstract

In this article we consider the probabilistic identification of amino acid positions that evolve under positive selection as a multiple hypothesis testing problem. The null hypothesis "$H_{0,s}$: site s evolves under a negative selection or under a neutral process of evolution" is tested at each codon site of the alignment of homologous coding sequences. Standard hypothesis testing is based on the control of the expected proportion of falsely rejected null hypotheses, or type-I error rate. As the number of tests increases however, the power of an individual test may become unacceptably low. Recent advances in statistics have shown that the false discovery rate – in this case, the expected proportion of sites that do not evolve under positive selection among those that are estimated to evolve under this selection regime – is a quantity that can be controlled. Keeping the proportion of false positives low among the significant results generally leads to an increase in power. In this article, we show that controlling the false detection rate is relevant when searching for positively selected sites. We also compare this new approach to traditional methods using extensive simulations.

Introduction

Most amino acid positions in a protein are likely to evolve under strong functional and structural constraints. These constraints are usually the consequences of a negative (or purifying) selection force that prevents nonsynonymous substitutions from occurring during the course of evolution. Other sites of the sequence are not constrained by natural selection. These positions then evolve under a neutral process and nonsynonymous substitutions occur at the same rate as the synonymous ones. Darwinian selection may also favour nonsynonymous changes at a few positions of the same sequence. These sites evolve under a positive (or diversifying) selection regime. The detection of sites evolving under positive selection is essential to better understand how natural selection acts at the molecular level. For instance, the *env* protein of the HIV-1 virus is usually targeted by the immune system of an infected individual. Several studies (e.g. Ross and Rodrigo, 2002) have shown that positive selection is present at some sites of this protein and is likely to help the virus to escape immune surveillance. Positive selection also appears to be responsible for the excess of replacement substitutions in the antigen recognition sites of the major histocompatibility complex (Hughes and Nei 1988), and in regions involved in species-specific sperm-egg interaction (Swanson et al. 2001).

The heterogeneity of selective patterns across sites has led several authors (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang, Wong, and Nielsen 2005) to propose new methods to detect adaptive evolution at individual positions. These methods focus on the distribution of nonsynonymous and synonymous substitutions across sites in a phylogenetic context. Nielsen and Yang (1998) first described a codon-based model of nucleotide substitutions that allows the nonsynonymous to synonymous rate ratio to vary among sites. Under this model, the substitutions between codons are described as a continuous-time Markov chain, with a state space on the 60 to 63 non-terminating codons, depending on the genetic code under consideration. This model is based on the assumption that a codon site evolves under one of negative,

neutral, or positive selection regime. The nonsynonymous to synonymous rate ratio is expected to be close to zero at amino-acid positions that are evolving under strong purifying selection. Alternatively, a position under positive selection should exhibit a nonsynonymous to synonymous rate ratio greater than one. The parameters of the model explored by Nielsen and Yang (1998) and later by Yang et al. (2000) include the nonsynonymous to synonymous rate ratio for each selection class, frequency parameters that are interpreted as the prior probability that a site falls into any specific selection category, and parameters such as the equilibrium codon frequencies and transition/transversion rate ratio. Most of these parameters are estimated from alignments of protein coding DNA sequences using either a maximum likelihood (Nielsen and Yang 1998) or a fully Bayesian (Huelsenbeck and Dyer 2004) approach.

Nielsen and Yang's model provides a suitable framework for the probabilistic identification of amino-acid sites evolving under positive selection. Indeed, the probability of each selection regime given the data at the site of interest and the parameters of the model (i.e., the posterior probability) can be computed easily using Bayes theorem (Nielsen and Yang 1998). Nielsen and Yang originally suggested the use of a naive Bayes classification method to identify positively selected sites. Under this approach a position is said to evolve under positive selection if the posterior probability of this regime is the highest among the various selection classes to be considered. These authors also indicated that posterior probabilities greater than 0.95 provide strong evidence in favour of the corresponding selection class. This threshold of 0.95 is now commonplace and has been implemented in widely used phylogenetic softwares.

The posterior probability for a site to evolve under positive selection is therefore the statistic that is used to classify each amino acid position in a selection regime. The fixed threshold approach can be seen as a test of the null hypothesis "$H_{0,s}$: position s evolves under a neutral or under a negative selection regime". One way to test this hypothesis would be to estimate the distribution of the statistic under $H_{0,s}$ and reject the null hypothesis if the value of the statistic computed from the data has a probability less than α

(usually $\alpha = 5\%$) of occurring if $H_{0,s}$ was true. In the context of the detection of positively selected sites, the estimation of the null distribution can be accomplished using a parametric bootstrap approach, which has previously been used in phylogenetics to test competing evolutionary hypotheses (Huelsenbeck and Rannala 1997). In practice however, the null distribution is never estimated, as a conservative threshold on the posterior probability of the positive selection regime is thought to correspond to a conservative (but unknown) p-value. This idea is supported by simulation results. For example, Yang et al. (2005) randomly generated sites that did not evolve under positive selection (i.e., $H_{0,s}$ is true for every s). They showed that a threshold of 0.95 on the posterior probability of the positive selection regime leads to a proportion of falsely rejected null hypotheses (type-I error) very close to 0 on average.

When the null hypothesis of a neutral or a negative selection regime is being tested at each amino acid position, the identification of positively selected sites is a multiple hypothesis testing problem. In the absence of positively selected sites, each test has a probability of yielding a significant result, thus the chance of drawing at least one false conclusion increases rapidly with the number of sites to be analysed. The Bonferroni-type correction procedure (Holm 1979) provide protection against any rejection of the null hypothesis across all tests. Using this approach, the probability of making one or more type-I errors is maintained at the desired α level. However, such strict control of type-I error (or familywise error rate, FWER) comes at a high cost : with an increasing number of tests, maintaining a low chance of making even one type-I error decreases the power (i.e., the chance of rejecting the null hypothesis when it is false). Hence, the number of positively selected sites detected is often unacceptably small when a Bonferroni-type correction is used. However, the main drawback with such correction is that it does not provide any information on the expected frequency of errors among all the rejected null hypotheses. The proportion of falsely rejected null hypotheses can therefore vary greatly depending on the frequency of truly false null hypotheses in the data to be analysed.

An alternative approach to error rate control was developed by Benjamini and Hochberg (1995, 2000). Instead of controlling the chance of making one or more type-I error, these authors proposed controlling the expected proportion of false discoveries among all the rejected hypotheses, defined to be the false discovery rate (FDR). For instance, if the FDR is controlled at level $\alpha = 5\%$ and twenty sites are classified as positively selected, then the expected number of false discoveries among these is one. Controlling the FDR offers a more relaxed multiple testing criterion than that of FWER control and therefore leads to an increase in power. Note also that one of the standard procedures providing control of the FDR is valid when the tests are positively correlated (Benjamini and Yekutieli 2001). This mathematical property is well suited to our problem, as physical interactions between amino acid residues in a protein are likely to induce positive correlations between patterns of substitutions at distant sites in the alignment. However, such positive correlation between substitution events does not necessarily imply positive correlation between selection regimes acting at distinct sites of the protein. Negative correlations between selection patterns are also likely to occur as a consequence of certain evolutionary constraints. Controlling the FDR in such conditions may be more difficult.

In this paper, we show that control of the FDR can be applied to the detection of positively selected sites. We first describe the standard codon-based models of nucleotide substitutions and the calculation of the posterior probabilities of each selection class. We then detail two methods for controlling the FDR. The first relies entirely on posterior probabilities of the different selection regimes. The second is based on a parametric bootstrap approach. Simulation results that demonstrate the relevancy of these approaches are discussed next.

Codon-based models and posterior probabilities of selection classes

Substitutions between codons evolving under a given selection regime are modelled as a stationary, homogeneous, and time-reversible continuous-time Markov process. In this article we consider a model with three classes of selection regimes which was designated as 'M2a' by Wong et al. (2004). Under such model, the site-specific nonsynonymous to synonymous rate ratio is a random variable whose discrete and non parametric distribution is estimated from the data. Nonsynonymous mutations are deleterious and eliminated under the first selection pattern, so that $\omega_1 < 1$. Rates of nonsynonymous and synonymous substitutions are equal under the second selection process and $\omega_2 = 1$. Nonsynonymous mutations provide a selective advantage under the third selection class, so that $\omega_3 > 1$. Values of ω_1 and ω_3, as well as p_1 and p_2, the equilibrium frequencies of the different selection classes, can be estimated from the data using a maximum likelihood approach (e.g., Nielsen and Yang, 1998 or Yang et al., 2000) or a Bayesian one (Huelsenbeck and Dyer 2004).

The posterior probability of an event is the probability that this event occurred, assuming that the underlying model is correct. If we assume that the sequences were generated under the substitution model M, the posterior probability that site D_s has been generated under the selection process x is given by Bayes theorem :

$$P(\omega_x | D_s, M) = \frac{p_x P(D_s | \omega_x, M)}{P(D_s | M)}, \tag{1}$$

where $P(D_s|M) = \sum_{x=1}^{3} p_x P(D_s|\omega_x, M)$ is the likelihood of M at site D_s, while $P(D_s|\omega_x, M)$ is the joint likelihood of M and the selection regime x at site D_s. The values of ω_x, p_x and the parameters of M are usually maximum likelihood estimates, and the calculation of the posterior probability of each selection class is referred to as an empirical Bayes (EB) approach. Indeed, the prior probability of each class (p_x for class x) is estimated from the data. This method does not take into account the sampling errors in ω_x and

p_x. Hence, the prior probability estimates, and therefore the posterior probabilities, can be unreliable when analysing small data sets, or very similar sequences (Suzuki and Nei 2004). To address this issue, Yang et al. (2005) recently proposed a Bayes empirical Bayes (BEB) approach. This method assigns a prior to the parameters of the substitution model (including p_x) and integrates over their uncertainties. Using simulations, these authors have shown that the BEB approach generally provides more reliable estimates of the posterior probabilities than those obtained with the standard empirical Bayes method.

Two different approaches to control the false discovery rate (FDR)

Newton et al. (2004) recently described a direct posterior probability approach to control the FDR in the context of microarray experimentation. Our goal here is to construct the largest list of positions that evolved under positive selection under the constraint that the expected proportion of false discoveries among this list is bounded by α. The first step in building this list is to rank the sites according to decreasing values of $\beta_s = 1 - P(\omega_3|D_s, M)$. Let the list, L, contain positions k having values β_k less than some bound δ. The expected number of false discoveries among the sites included in the list is :

$$C(\delta) = \sum_k \beta_k 1\{\beta_k \leq \delta\} \tag{2}$$

where $1\{.\}$ is an indicator function. Equation (2) holds because β_k is the conditional probability that placing site k on the list creates a type-I error. The expected rate of false detections given the data is simply $C(\delta)/|L|$ when $|L| > 0$ and zero otherwise. The goal is then to find $\delta \leq 1$ as large as possible so that $C(\delta)/|L| \leq \alpha$, where α is the desired FDR. It is important to note that β_s is the probability of making a type-I error if we reject the null hypothesis that site D_s does not evolve under positive selection. However, β_s does not correspond to the probability of incorrectly rejecting the null hypothesis as the result of a statistical test, and thus is not equivalent to a frequentist p-value.

8

A second method to control the FDR is based on parametric bootstrap. The parametric bootstrap was originally proposed in phylogenetics to estimate the variances of model parameters, such as the tree topology (Felsenstein, 2003, p.357-358) . This approach also provides an elegant tool for testing complex evolutionary hypotheses. Indeed, parametric bootstrap is used to estimate the distribution of the difference of likelihoods computed under a null and an alternative hypothesis for data generated under the null hypothesis (Huelsenbeck and Crandall 1997).

In this paper we use the parametric bootstrap to identify positively selected sites under the control of the FDR framework. The first step is to generate synthetic data, denoted as D^*, under M, the best model estimated from the original data set. The true selection class at each site is also recorded. Let $c_s^* \in \{1, 2, 3\}$ be the true selection class at position s, with 1, 2 and 3 corresponding to negative, neutral and positive selection respectively. $P(\omega_3|D_s^*, M)$ is the posterior probability of the positive selection regime at site s of the synthetic data set, given the model estimated from the original data set. The next step is to create a list of sites ranked according to decreasing values of $\theta_s^* = P(\omega_3|D_s^*, M)$. This list contains positions k having values θ_k^* greater than some bound δ^*. The observed proportion of false discoveries among the sites included in the list is :

$$\text{FDR}(\delta^*, D^*) = \frac{\sum_k 1\{s_k^* \neq 3, \theta_k^* \geq \delta^*\}}{k}. \tag{3}$$

Let δ_α^* be the value of δ^* so that $FDR(\delta^*, D^*) \leq \alpha$. This procedure of generating synthetic data and recording the value of δ_α^* is repeated several times in order to estimate the distribution of δ_α^*. $\overline{\delta_\alpha^*}$ is the mean of this distribution. Sites with $P(\omega_3|D_x, M) \geq \overline{\delta_\alpha^*}$ define a list of positions that are classified as positively selected.

Simulations

We conducted simulations in order to assess the ability of Newton et al.'s direct posterior probability approach (DP) and our parametric bootstrap (PB) method to control the FDR. We first generated three hundred random phylogenies, each comprising thirty taxa, using the standard speciation process described in Kuhner and Felsenstein (1994). This process makes the trees molecular clock-like, so we created a deviation from this model by multiplying every branch length by $(1 + X)$, where X followed an exponential distribution with expectation μ. The μ value represents the extent of deviation and was identical within each tree, but different from tree to tree and equal to $0.2/(0.001 + U)$, where U was uniformly drawn from $[0.0, 1.0]$. The tree length was eventually rescaled so that the sum of branch lengths was uniformly distributed in the $[0.1, 5.0]$ range. Hence, 0.1 and 5.0 nucleotide substitutions per codon site were expected on the smallest and largest tree respectively.

We then generated sequences along each of these trees using a computer program written by one of the authors (S.G.). These homologous sequences are $N = 500$ codon sites long. They were generated under model M2a with a uniform distribution of equilibrium frequencies of codon states and equal transition and transversion rates. We tested four different sets of parameters to analyse the efficiency of the different approaches to control the FDR with respect to the intensity of positive selection. The frequency of positively selected sites (p_3) was set to 0.10 and four different intensities of positive selection were tested : $\omega_3 = 1.5$, 2.0, 4.0 and 10.0. $\omega_3 = 1.5$ or 2.0 corresponds to weak positive selection pressures while $\omega_3 = 4.0$ or 10.0 corresponds to strong positive selection conditions. We generated three hundred data sets under each of the four models defined by the various positive selection intensities.

We also generated sequences along the same trees under a codon-based model that is probably more realistic than M2a from a biological perspective. This model has two classes of selection regimes. The

first is defined by nonsynonymous to synonymous rate ratios uniformly distributed between 0.0 and 1.0. This class therefore generates sites under a range of selection regimes going from strict neutrality to strong negative selection. The second class is defined by nonsynonymous to synonymous rate ratios uniformly distributed between 1.0 and 10.0. Sites falling in this category therefore evolve under various intensities of positive selection. The equilibrium frequency of the first selection regime is set to 0.9 and the sequences are $N = 500$ codon sites long. One hundred data sets were generated according to this model.

A phylogeny was then estimated from each of the synthetic data set under a general time-reversible model of nucleotide substitution (Lanave et al. 1984) with gamma distributed substitution rates across sites (Yang 1994) and a proportion of invariable sites. Although it would have been more consistent with the rest of the analysis to estimate the tree topology under a codon-based model, in practice, nucleotide-based models are used because they are much more computationally tractable. Moreover, it is likely that slight variations in the tree topology do not drastically affect the estimated values of the codon model parameters.

The tree topology was estimated using the program PHYML (Guindon and Gascuel 2003). The parameters of the M2a codon-based model of substitution were then estimated from each sequence data set and the corresponding estimated phylogenetic tree. The free parameters of the M2a model along with the tree branch lengths were adjusted using standard numerical optimisation methods (Press et al. 1988). For each data set and fitted model in hand, we computed the posterior probability of each selection regime using the standard EB approach (Equation (1)) as well as the BEB method developed by Yang, Wong and Nielsen (2005). We then applied Newton et al.'s method combined with EB or BEB to control the FDR at the $\alpha = 5\%$ level. Our parametric bootstrap approach was also used to control the FDR at the same level. For each original data set, D, of length $N = 500$, we generated one hundred bootstrapped data sets D^* of length N along the tree estimated from D. These sequences were generated under the model M2a with parameters also estimated from D. For each of the hundred synthetic data sets D^*, we estimated one value

of δ_α^* (see the previous section) using the EB approach to calculate the posterior probabilities of the positive selection class at each site. The value of $\overline{\delta_\alpha^*}$ which was used to control the FDR corresponds to the mean taken among the hundred values of δ_α^*. Note that this parametric bootstrap approach does not require additional tree topology estimation or numerical parameter optimisation. The time needed to compute $\overline{\delta_\alpha^*}$ is simply the product of the time to compute the likelihood of the model by the number of bootstrapped data sets. Preliminary tests on simulated data sets suggested that only fifty bootstrap iterations were enough to converge to stable estimates of $\overline{\delta_\alpha^*}$, making this approach computationally tractable.

Newton et al.'s method is straightforward to implement once the posterior probabilities of the positive selection class are known at each site. A program written by one of us (S.G.) was used to compute the EB posterior probabilities. CODEML (from the PAML package version 3.14 (Yang 1997)) was used to compute the BEB posterior probabilities. We then combined the direct posterior probability approach with both EB and BEB. These two combinations of methods are referred to as DP+EB and DP+BEB respectively. We also implemented the parametric bootstrap approach combined with the EB posterior probabilities. This method is referred to as PB+EB. We would have liked to combine the parametric bootstrap approach with the BEB calculation of the posterior probabilities. However, we were unable to formulate a suitable implementation of this combination of methods.

Results

Synthetic data were generated in order to test the ability of the direct posterior probability and the parametric bootstrap to control the FDR at level $\alpha = 5\%$. The first two sets of sequences contained on average 10% of sites evolving under weak positive selection ($\omega_3 = 1.5$ for the first set and $\omega_3 = 2.0$ for the second). Figure 1 presents the distributions of the observed proportions of sites that do not evolve under positive selection among those that are estimated to be positively selected.

Even if the average proportions of false discoveries are above the $\alpha = 5\%$ level, this quantity is maintained below the desired threshold in most data sets under the DP+EB approach (Figure 1a and 1b). Note that the distribution of the observed proportions of false positives is bimodal when positive selection is the weakest (Figure 1a, $\omega_3 = 1.5$). The values centred around PFP$\simeq 0.8$ are the consequence of poor estimation of the substitution model parameters. In such conditions, the model failed to distinguish between truly positively selected sites and neutrally evolving ones. Despite this, the DP+BEB method provides a stringent control of the FDR (Figure 1c and 1d). The average proportions of false discoveries are indeed very close to the desired $\alpha = 5\%$ threshold in both simulation conditions. This proportion is also maintained below the threshold in a great majority of the data sets. Note also that the distribution under very weak positive selection pressure (Figure 1c) is not bimodal, contrasting with the results obtained with DP+EB. This result illustrates the advantage of integrating over the uncertainties in the substitution model parameter estimates when calculating the posterior probabilities of the selection regimes. The results obtained under the PB+EB method are less encouraging (Figure 1e and 1f). While the mode of the distributions are still below the $\alpha = 5\%$ threshold, the averages are well above this limit. The proportion of false discoveries is also greater than $\alpha = 5\%$ in most of the data sets that were analysed.

Two additional sets of synthetic sequences were analysed. Each of these data sets contained on average 10% of sites evolving under strong positive selection ($\omega_3 = 4.0$ or $\omega_3 = 10.0$). Figure 2 presents the distributions of the observed proportion of false discoveries obtained with the three different methods already tested. The average proportions of false discoveries are now systematically close to $\alpha = 5\%$, contrasting with the previous results. The three approaches therefore successfully maintain the average proportions of false discoveries below or close to $\alpha = 5\%$. It is also interesting to note that the fraction of data sets for which the detection rate is below $\alpha = 5\%$ is the highest among the three methods when sites are classified with the PB+EB approach. This suggests that, in such strong positive selection conditions, PB+EB provides a more stringent control of the FDR than both DP+EB or DP+BEB.

13

Table 1 shows the average rates of detection of positively selected sites obtained with methods based on control of the FDR (at the $\alpha = 5\%$ level) as well as the standard fixed threshold approach (the value of the threshold being $P(\omega_3 > 0.95)$). The detection rate is measured for each data set and corresponds to the number of positively selected sites that are correctly identified divided by the total number of positively selected sites. The detection rates are virtually equal to zero under weak selection conditions when the FDR is controlled using DP+BEB. Hence, the stringent control of the FDR under weak selection conditions using this approach comes at the price of very low detection rates. PB+EB seems to outperform the other methods here. However, one must bear in mind that this method often failed to control the FDR if positive selection is weak (see Figure 1e and 1f). The detection rates are much higher in strong positive selection conditions. In such conditions, the parametric bootstrap outperforms the direct posterior probability approach : the three methods show identical detection rates (Table 1, columns 2-4) but PB+EB has the lowest average proportions of false positives (see Figure 2). The methods based on control of the FDR outperform the standard fixed threshold approach (combined with both EB and BEB calculation of the posterior probabilities). Here again, it is important to recall that the fixed threshold approach is very conservative in these conditions. Hence, the FDR obtained with this approach is usually much lower than $\alpha = 5\%$.

We next tested the performance of the different methods when the model used to make the inferences is distinct from the true model of evolution. Sequences were generated under a codon-based model with two classes of uniformly distributed nonsynonymous to synonymous rate ratios. Sites belonging to the first class evolved under a variety of selection regimes going from strong negative selection pressure to a strictly neutral process. Sites falling in the second class evolved under weak to strong positive selection. We then fitted an incorrect model (M2a) to these sequences and searched for positively selected sites using the DP+BEB, DP+EB and PB+EB approaches. Table 2 shows the detection rates and proportions of false positives obtained from the analysis of one hundred data sets. The three methods successfully maintained

the proportion of false positives below the desired $\alpha = 5\%$ threshold. The majority of the proportions of false positives observed among the hundred data are also clustered close to the median (see the values of the 25% and 75% quantiles). These results and the histograms of the proportions of false positives (not shown) indicate that the FDR is successfully controlled in a great majority of data sets. The detection rates are the highest when the FDR is controlled using the parametric bootstrap approach. Note also that the detection rates obtained with the fixed threshold approach combined with BEB and EB are 0.52 and 0.54 respectively. Hence, here again, a significant proportion of positively selected sites are correctly identified under the control of the FDR framework (at the 5% level) and ignored if a fixed threshold method (with $P(\omega_3 > 0.95)$) is used.

Discussion

The purpose of this paper was to introduce recent advances in statistics that are well suited for the probabilistic identification of positively selected sites in coding sequence alignments. We describe two approaches that aim to control the false discovery rate (FDR) by utilising posterior probabilities of selection classes computed at each codon position of the alignment. The first method is the direct posterior probability approach recently proposed by Newton et al. (2004) in the context of microarray data analysis. The second approach is based on a parametric bootstrap procedure. In theory, both methods build the largest list of codon positions that contains, on expectation, no more than a proportion, α, of sites that do not evolve under positive selection. Newton et al.'s method creates this list from the posterior probability of the positive selection regime computed at each site. The parametric bootstrap approach estimates the distribution of the smallest posterior probability of the positive selection regime in the list.

We tested the ability of both methods to maintain the average proportion of false positives below $\alpha = 5\%$ using simulations. The simulation settings were defined in order to best approximate the practical aspects

15

behind the identification of positively selected sites. Hence, while sequences were generated under known codon-based models, the phylogenetic trees and substitution parameters that were used to detect site-specific positive selection were all estimated from the synthetic data. Various positive selection intensities and frequencies were also tested in order to encompass a variety of biologically plausible conditions.

While all the methods tested here successfully control the FDR in strong positive selection conditions, it is generally not the case when positive selection is weak. These poor performances are expected because our substitution models can hardly tell the difference between sites evolving under a weak positive selection regime and those submitted to a strictly neutral process. Hence, no method is likely to achieve satisfying performances in such context. On one hand, approaches that successfully maintain the observed proportion of false discoveries below 5% hardly detect any positively selected sites. The direct posterior probability method combined with the Bayes empirical Bayes calculation of the posterior probabilities belongs to this category. On the other hand, the approaches that show reasonable detection rates are also hampered by unacceptably high proportions of false discoveries, generally above the 5% threshold. The parametric bootstrap approach as well as the Newton et al.'s method combined with the empirical Bayes calculation of the posterior probabilities belong to this category. Similar results have already been reported recently (see Scheme 4, Table 5 in Yang, Wong and Nielsen, 2005)

The control of the FDR is much more satisfactory in strong positive selection intensity conditions. Indeed, the average proportions of false discoveries are close to $\alpha = 5\%$ in a large majority of data sets (Figure 2). In such conditions, positively selected sites and neutrally evolving ones are very distinct in terms of the relative abundance of nonsynonymous and synonymous changes. Hence, any relevant statistical method will easily distinguish between the two patterns. Successful control of the FDR at reasonable levels comes with high detection rates of positively selected sites. These detection rates are actually well above those achieved by the standard fixed threshold approach (Table 1, third and fourth rows), as expected.

To intensities, no method is likely to efficiently control the FDR (nor the type-I error rate) provided that the inferences rely on the current codon-based models. On the other hand, the FDR can be controlled when the intensity of positive selection if strong enough (say $\omega_3 \geq 4.0$). In terms of the number of positively selected sites that are detected in such conditions, controlling the FDR is preferable to the standard fixed threshold approach, which is usually excessively conservative.

Additional simulations have shown that the FDR can be controlled when the model that generated the sequences is different from the one that was used to identify the positively selected sites (see Table 2). Our results indicate that the different methods based on the control of the FDR perform almost equally, even if the detection rates are slightly higher with the parametric bootstrap approach. Controlling the FDR under such simulation settings lead to a controlled increase of the number of false positives as well as a significant increase of the detection rates as compared to the standard fixed threshold approach. These results do not demonstrate that the methods to control the FDR will perform well if applied to real data sets. However, it is clear that the combination of the M2a codon-based model of substitution and the new methods to control the FDR presented in this article can be robust to model misspecification.

Newton et al.'s direct probability approach generally outperforms the parametric bootstrap method when positive selection is weak (Figure 1). In such conditions, the substitution parameters are often very poorly estimated. The estimation of the cut-off values under the parametric bootstrap approach then solely relies on bootstrapped data sets that were generated under a wrong model of substitution. Newton et al.'s method performs better under such conditions probably because it does not rely on such heavy inference procedure applied to poorly informative data. The parametric bootstrap method provides a much more stringent control of the FDR in strong positive selection conditions (Figure 2). The proportions of positively selected sites detected by this method are also generally larger than those obtained under the direct posterior probability method (Table 1). This suggests that in weak positive selection conditions, control of the FDR

17

should be based on the simplest procedures, such as the direct posterior probability. Under strong positive selection conditions, accurate estimations can be achieved using more sophisticated methods, based on parametric bootstrap for instance.

It is also worth mentioning that, unlike Newton et al.'s approach, several different statistics can be used to classify sites using the parametric bootstrap. In this article, the statistic is the posterior probability of the selection regimes computed at each codon position. Hence, the ranking of the original sites according to this statistic is the same under the direct posterior probability approach and the parametric bootstrap. However, sites could also be ranked according to the ratio between the likelihood calculated under the positive selection regime and the likelihoods under the other selection regimes, for instance. This new statistic could replace the posterior probability at the core of the parametric bootstrap approach. Newton et al.'s method is not as flexible because it solely relies on posterior probabilities of the different hypotheses, by definition.

Readers should also be aware that the methods described in this paper are not only restricted to the "site models" introduced by Nielsen and Yang (1998) and Yang and Nielsen (2002). We use these models here because they are very well suited for the identification of positively selected sites, which is relevant from a biological perspective. However, "site-branch models" (Yang and Nielsen 2002; Guindon et al. 2004) (i.e., models that take into account the likely site and lineage heterogeneity of selection processes), are probably more realistic than the simple "site models". Moreover, these models provide an adequate framework to ask the question : "Is this specific lineage evolving under positive (or negative, or neutral) selection at this (or these) site(s) ?". Answering such question requires the use of probabilistic tools that are identical to those involved in the detection of positively selected sites under the "site models". Hence, the control of the FDR should also be considered with interest in this context.

18

Acknowledgements

References

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J. R. Stat. Soc. B., **57**:289-300, 1995.

—. The adaptive control of the false discovery rate in multiple hypothesis testing with independent statistics. J. Educ. Behav. Statist., **25**:60-83, 2000.

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, **29**:1165-1188, 2001.

Felsenstein, J. Inferring phylogenies. Sinauer Associates, Inc., 2003.

Guindon, S. and Gascuel, O. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol., **52**:696-704, 2003.

Guindon, S., Rodrigo, A., Dyer, K. and Huelsenbeck, J. Modeling the site-specific variation of selection patterns along lineages. Proc. Natl. Acad. Sci. USA., **101**:12957-12962, 2004.

Holm, S. A simple sequentially rejective multiple test procedure. Scand. J. Stat., **6**:65-70, 1979.

Huelsenbeck, J. and Dyer, K. Bayesian estimation of positively selected sites. J. Mol. Evol., **58**:661-672, 2004.

Huelsenbeck, J. and Rannala, B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science, **276**:218-219, 1997.

Huelsenbeck, J. P. and Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. Ann. Rev. Ecol. Syst., **28**:437-466, 1997.

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

Kuhner, M. and Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol., **11**:459-468, 1994.

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

Newton, M., Noueiry, A., Sarkar, D. and Ahlquist, P. Detecting differential expression with a semi-parametric hierarchical mixture method. Biostatistics, **5**:155-176, 2004.

Nielsen, R. and Yang, Z. Likelihood models for detecting positively selected amino acid sites and application to the HIV-1 envelope gene. Genetics., **148**:929-936, 1998.

Press, W., Flannery, B., Teukolsky, S. and Vetterling, W. Numerical Recipes in C. Press Syndicate of the University of Cambridge, Cambridge, 1988.

Ross, H. and Rodrigo, A. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J. Virol., **76**:11715-11720, 2002.

Suzuki, Y. and Gojobori, T. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol., **16**:1315-1328, 1999.

Suzuki, Y. and Nei, M. False-positive selection identified by ML-based methods: examples from the *sig1* gene of the diatom Thalassoria weissflogii and the *tax* gene of a human T-cell lymphotropic virus. Mol. Biol. Evol., **21**:914-921, 2004.

Swanson, W., Yang, Z., Wolfner, M. and Aquadro, C. Positive darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA., **98**:2509-2514, 2001.

Wong, W., Yang, Z., Goldman, N. and Nielsen, R. Accuracy and power of statistical methods for

detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics, **168**:1041-1051, 2004.

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

—. PAML : a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci., **13**:555-556, 1997.

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

Yang, Z., Wong, W. and Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol., **22**:1107-1118, 2005.

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

Figure 1. Distributions of the observed proportions of false positives obtained under weak positive selection conditions. PFP stands for proportion of false positives. a and b : FDR is controlled at level $\alpha = 0.05$ using the direct posterior probability approach. The calculation of the posterior probability of positive selection at individual sites is performed using the empirical Bayes method. c and d : FDR is controlled at the same level $\alpha = 0.05$ using the direct posterior probability approach combined with the Bayes empirical Bayes calculation of the posterior probabilities. e and f : FDR is controlled at the same level using the parametric bootstrap approach combined with the empirical Bayes calculation of the posterior probabilities of the selection regimes. a, c and e : the simulated data sets were generated under model M2a with $\omega_0 = 0.0$, $\omega_1 = 1.0$, $\omega_3 = 1.5$ and $p_0 = 0.45$, $p_1 = 0.45$, $p_2 = 0.10$. b, d and f : the sequences were generated under the same conditions except that $\omega_3 = 2.0$. $F(0.05)$ gives the percentage of the three hundred data sets for which the proportion of false positives is less than 0.05. $\widetilde{\text{PFP}}$ is the median of these values and $\overline{\text{PFP}}$ is the mean.

Figure 2. Distributions of the observed proportions of false positives obtained under strong positive selection conditions. a, c and e : the simulated data sets were generated under model M2a with $\omega_0 = 0.0$, $\omega_1 = 1.0$, $\omega_3 = 4.0$ and $p_0 = 0.45$, $p_1 = 0.45$, $p_2 = 0.10$. b, d and f : the sequences were generated under the same conditions except that $\omega_3 = 10.0$. See the caption of Figure 1 for additional informations.

Table 1
Average detection rates of positively selected sites.

	DP		PB	Fixed threshold	
	+BEB	+EB	+EB	+BEB	+EB
$\omega_3 = 1.5$	0.00	0.29	0.31	0.00	0.27
$\omega_3 = 2.0$	0.08	0.15	0.19	0.05	0.11
$\omega_3 = 4.0$	0.61	0.61	0.61	0.45	0.45
$\omega_3 = 10.0$	0.91	0.91	0.91	0.84	0.84

Note.— The detection rate is the number of positively selected sites that are correctly identified divided by the total number of positively selected sites. This ratio is computed for each data set and the table shows the values averaged over the three hundred data sets analysed under each simulation setting. Left to right : the second, third and fourth columns give the detection rates achieved by the methods that control the FDR (i.e. direct posterior probability (DP) and parametric bootstrap (PB)) at the $\alpha = 5\%$ level. The fifth and sixth columns give the detection rates obtained under the standard fixed threshold approach combined with both empirical Bayes and Bayes empirical Bayes calculation of the posterior probabilities. The value of the threshold is fixed to $P(\omega_3 > 0.95)$. The different lines of the table correspond to various selection intensities.

Table 2
Detection rates and control of the FDR when the inferences rely on a wrong model.
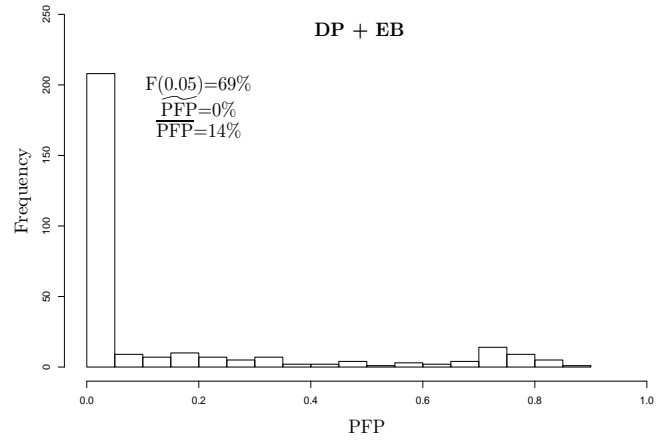
Method	Detection Rates	Proportions of False Positive
DP+BEB	0.64 (0.55-0.79)	0.01 (0.00-0.03)
DP+EB	0.66 (0.57-0.80)	0.02 (0.00-0.03)
PB+EB	0.67 (0.56-0.80)	0.03 (0.00-0.04)

Note.— See the captions of Table 1 and Figure 1 for the definitions of detection rates and proportion of false positives. The FDR was controlled at the $\alpha = 5\%$ level. The numbers in parentheses correspond to the 25% and 75% quantiles. DP stands for direct posterior probability, PB, for parametric bootstrap. EB and BEB stand for empirical Bayes and Bayes empirical Bayes calculation of the posterior probabilities of the selection classes respectively. Sequences were generated under a mixture of two uniform distributions of the values of the nonsynonymous to synonymous rate ratio. The inferences relied instead on the M2a model of substitution.
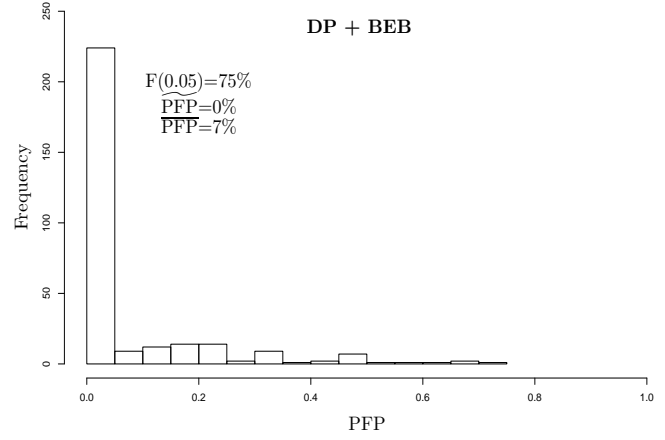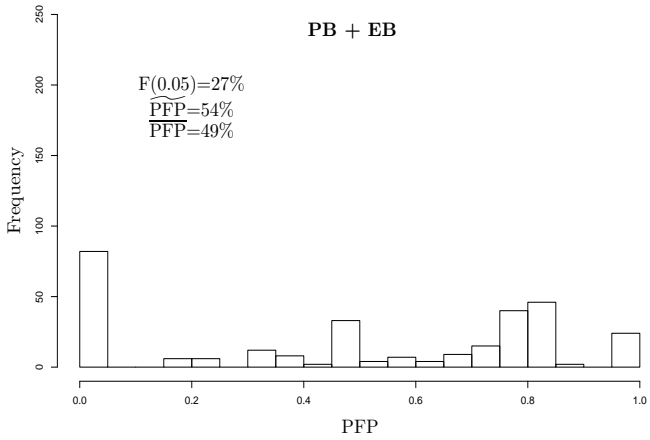
Figure 1.

a)

DP + EB

F(0.05)=47%
$\widetilde{\mathrm{PFP}}$=5%
$\overline{\mathrm{PFP}}$=8%

b)

DP + EB

F(0.05)=63%
$\widetilde{\mathrm{PFP}}$=4%
$\overline{\mathrm{PFP}}$=5%

c)

DP + BEB

F(0.05)=43%
$\widetilde{\mathrm{PFP}}$=6%
$\overline{\mathrm{PFP}}$=7%

d)

DP + BEB

F(0.05)=55%
$\widetilde{\mathrm{PFP}}$=5%
$\overline{\mathrm{PFP}}$=5%

e)

PB + EB

F(0.05)=49%
$\widetilde{\mathrm{PFP}}$=5%
$\overline{\mathrm{PFP}}$=8%

f)

PB + EB

F(0.05)=76%
$\widetilde{\mathrm{PFP}}$=2%
$\overline{\mathrm{PFP}}$=3%

Figure 2.