Genotype–phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate

Timothy D. O’Connor* and Nicholas I. Mundy
Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK

ABSTRACT

Motivation: Mapping between genotype and phenotype is one of the primary goals of evolutionary genetics but one that has received little attention at the interspecies level. Recent developments in phylogenetics and statistical modelling have typically been used to examine molecular and phenotypic evolution separately. We have used this background to develop phylogenetic substitution models to test for associations between evolutionary rate of genotype and phenotype. We do this by creating hybrid rate matrices between genotype and phenotype.

Results: Simulation results show our models to be accurate in detecting genotype-phenotype associations and robust for various factors that typically affect maximum likelihood methods, such as number of taxa, level of relevant signal, proportion of sites affected and length of evolutionary divergence. Further, simulations show that our method is robust to homogeneity assumptions. We apply the models to datasets of male reproductive system genes in relation to mating systems of primates. We show that evolution of semenogelin II is significantly associated with mating systems whereas two negative control genes (cytochrome b and peptidase inhibitor 3) show no significant association. This provides the first hybrid substitution model of which we are aware to directly test the association between genotype and phenotype using a phylogenetic framework.

Availability: Perl and PHYBR scripts are available upon request from the authors.

Contact: to252@cam.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

One of the major issues in evolutionary genetics research is the relationship between genotype and phenotype. Natural selection acts on phenotypes and indirectly leaves a signal at the molecular level. The connection between the two levels is important because it ties together the effects of natural selection. Thus, selection for a phenotype can change the genetic variation for specific genes or genomic regions.

Within the field of molecular evolution, the study of adaptation has focused on methods for detecting selection in coding sequences, with any inferences about phenotypic evolution being indirect. At the forefront of this enquiry, Yang, Nei, Goldman and others (Goldman and Yang, 1994; Nei and Gojobori, 1986; Yang, 2007) developed computational models of molecular evolution to distinguish between neutral mutation and selection. These codon models focus on the ratio (dN/dS) of the rate of non-synonymous or protein altering changes to the rate of synonymous or silent changes assumed to estimate the neutral rate of evolution (Goldman and Yang, 1994; Muse and Gaut, 1994).

At intraspecies level, and occasionally at the closely related interspecies level, quantitative trait locus (QTL) analyses have been designed to detect specific regions of the genome associated with a given trait (Slate, 2005). These methods typically use pedigree information or known population structure to make specific crosses for particular phenotypes (Lynch and Walsh, 1998). The crosses are then genotyped using SNP or other markers across the whole genome and statistical associations of the linkage disequilibrium between genotype and phenotype are identified. Other studies use association mapping to identify genomic regions involved in phenotypic differences, or perform candidate gene associations, e.g. MC1R in relation to colouration differences (Nachman et al., 2003; Theron et al., 2001).

A few studies have looked for associations at the interspecies level using phylogenetics. The two main approaches used are regression analysis between evolutionary rate and phenotypic variation and codon branch-site models with phenotypes assigned to branches.

In the regression analyses published to date, dN/dS ratios are calculated for each branch in the tree using the free-ratios model (Yang, 1998) and a regression is performed by (i) pairing the dN/dS ratio for each terminal branch with the phenotype value for its terminal node or (ii) pairing the dN/dS ratio for every branch with the reconstructed phenotype on that branch. Using the first approach in primates, Dorus et al. (2004) found a positive correlation between levels of sperm competition (mean number of partners in a periovulatory period) and the dN/dS ratio of semenogelin II (SEMG2), a gene encoding a protein involved in primate semen. Later, Hurle et al. (2007) added additional taxa and performed a similar analysis but found no significant trend.

In a similar approach, Herlyn and Zischler (2007) found a negative correlation between the dN/dS in sperm ligand zonaadhesin (ZAN) and primate body weight dimorphism. In birds, Nadeau et al. (2007) employed this method to study correlations between pigmentation genes and sexual dimorphic colour variation in galliforms. Also, they used the second method and correlated dN/dS ratios for internal and terminal branches and ancestral reconstructions of sexual dimorphism in colouration over the phylogenetic tree. Both methods showed a correlation between MC1R, but not other pigmentation genes, and dimorphic colouration (Nadeau et al., 2007).

The second method employed the use of branch-site codon tests which test for changes in selection pressure on particular branches with phenotypes of interest. This method tests for positive selection

*To whom correspondence should be addressed.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/25/12/i94/192975
By guest
on 28 July 2019

[10:14 15/5/2009 Bioinformatics-btp231.tex]
by comparing a null model of neutral evolution to a model of positive selection on those branches (Zhang et al., 2005). Ramm et al. (2008) reanalysed SEMG2 as well as SEMG1 in primates using the codon models. They found that branches leading to species with high levels of sperm competition (multimale mating systems) show significant evidence of positive selection in SEMG2 but not SEMG1. Branches leading to species with low levels of sperm competition show no evidence for positive selection at either locus. In addition, they tested seven rodent semen proteins and found that Ssv2, the rodent orthologue to SEMG2, showed significant evidence for positive selection on branches leading to taxa with high relative testis size.

All of these tests can be criticized on theoretical grounds. For tests using phenotypic states derived from terminal taxa, the phenotypic state is applied to a whole branch without regard to its evolution. This creates a problem because some portion of the branch being associated with a phenotype is potentially misapplied, by ignoring the timing of the evolutionary loss or gain of the phenotype. For tests relying on phenotypic character reconstruction for internal assignment, error in reconstruction is not taken into account in downstream analyses.

One way around these difficulties is the maximum likelihood approach, which assigns characters to terminal nodes and probability distributions for those characters to internal nodes (Felsenstein, 1981). Thus, it estimates the ancestral state in terms of a probability distribution and integrates over the whole distribution. The probability distribution is calculated by accounting for all combinations of character state and numbers of changes (Felsenstein et al., 2004).

The maximum likelihood framework allows us to pull from a large body of statistical research. One applicable area includes methods designed to detect coevolution both at the phenotype–phenotype level (Pagel, 1994) and the genotype–genotype level (Pollock et al., 1994; Williams et al., 2006) and phenotype evolution across the phylogeny. The Independent model $Q$ matrix or substitution rate matrix is given as:

$$Q_i(\mathbf{g}_i, \mathbf{p}_i) = \begin{cases} Q_{gi}(\mathbf{g}_i, \mathbf{p}_i) & \text{if } \mathbf{g}_i = \mathbf{g}_j \text{ and } p_i \neq p_j \\ Q_{pj}(\mathbf{g}_i, \mathbf{p}_i) & \text{if } p_i = p_j \text{ and } \mathbf{g}_i \neq \mathbf{g}_j \\ 0 & \text{if } \mathbf{g}_i \neq \mathbf{g}_j \text{ and } p_i \neq p_j \end{cases}$$

where $g_i$ is the genotype state and $p_i$ is the phenotype state at point $i$. $Q_{gi}$ is the genotype rate matrix and $Q_{pj}$ the phenotype rate matrix. Double mutations, where both the genotype and phenotype are changing at the same moment are fixed to zero to allow the methodology to distinguish between actual associations and those that occurred by chance on the same branch. This follows the philosophy of the coevolution models (Pagel, 1994). When there is a single change, the rate is calculated based on its respective rate matrix.

The Dependent model $D$ uses scaling or weighting parameters to modify the rate of evolution for the genotype given the state of the phenotype, thus testing for an evolutionary association of the gene to various states of the phenotype. The Dependent model $Q$ matrix is defined similarly to the Independent model as:

$$Q_D(\mathbf{g}_i, \mathbf{p}_i) = \begin{cases} Q_{gi}(\mathbf{g}_i, \mathbf{p}_i) & \text{if } \mathbf{g}_i = \mathbf{g}_j \text{ and } p_i \neq p_j \\ Q_{pj}(\mathbf{g}_i, \mathbf{p}_i) & \text{if } p_i = p_j \text{ and } \mathbf{g}_i \neq \mathbf{g}_j \\ 0 & \text{if } \mathbf{g}_i \neq \mathbf{g}_j \text{ and } p_i \neq p_j \end{cases}$$

The scale or weight parameter is then $W_{ip}$, with a different value for the given phenotype. The Independent model is a subset of the Dependent model by setting all of the weight parameters to one.

Since the time and rate are mathematically confounded in Markov models, they are simultaneously calculated as a product (Yang, 2006), we use a mixture model approach to separate the weight parameters from the basic rate parameters and branch lengths (Pagel and Meade, 2004). In a likelihood ratio test (LRT), the Independent model is compared with a model containing a proportion of sites evolving under the Independent model and a proportion of sites evolving under the Dependent model with the same branch lengths and rate parameters, the only difference being the scaling parameters and the proportion of sites. In addition, the branch lengths for the phenotype are estimated using the molecular data under the assumption that they estimate divergence distances because estimating branch lengths and rate parameters from a single phenotype character can overparameterize the data, thus violating maximum likelihood assumptions (see Yang (2006), pp. 124–126). In other words, a single binary data point cannot be used to estimate rate parameters and branch lengths (when $N = k$, the number of parameters is 13 branch lengths and one rate parameter). After the branch lengths were calculated from the genotype data, the phenotype rate parameter was estimated on its own because when combined with the genotype data the likelihood surface of the phenotype rate parameter was overshadowed by those of the genotype, creating optimization difficulties (Fig. 1).
The models and likelihood tests were implemented using the phylogenetic software package HYPHY (Pond et al., 2005) (see Supplementary File 2 for an example HYPHY script). This program is flexible in creating likelihood functions and optimizing them with a conjugated gradient ascent algorithm (Hestenes and Stiefel, 1952) with bracketing. We set the number of iterations per variable to 1026 as recommended by the HYPHY authors’ web site to help with flat likelihood surfaces. The phenotype tended to create flat likelihood surfaces, see Figure 1, due to their low level of information content (a single data point across all species). Each model was run a minimum of five times from random starting positions in both the simulated and real datasets. A typical run with 32 taxa and five random search starts takes about 30 min on a 2.8 GHz Intel Xeon processor with 512 MB of RAM running Bio-linux 4 with a few of the runs taking up to a day.

2.2 Model interpretation

The parameters estimated can be used to understand the evolutionary relationship between genotype and phenotype. As is standard procedure the Qp parameters are measured in expected substitutions per site per unit time. The Qg parameters are measured as expected substitution changes per unit time as there is only one site or data point. The weight parameters (Wp), with their association with the Qp can be interpreted as a rate multiplier. This means that a weight equal to one is the same rate as the background substitution rate, and a weight equal to 10 has a 10-fold higher expected substitution per site per unit time than the background.

This scaling effect in the Dependent model is caused by a change in evolutionary pressure associated with a particular phenotype. In principle, a major reason for a change in rate associated with a particular phenotype is an altered selective regime occurring under that phenotype, such as positive selection or reduced constraint. For example, species under high sperm competition are predicted to have a higher rate of change in coding regions involved in sperm competition because of a higher dN due to directional selection. However, it is important to note that other formal causes of an association between phenotype and evolutionary rate are possible, including effects involving neutral processes. Examples of these are an effect of the phenotype on mutation rate and an effect of the phenotype on rate of fixation of mildly deleterious substitutions. One way to discriminate between neutral and selective effects would be that the former would have genome-wide effects whereas the latter would be gene specific.

2.3 Model implementation

The models and likelihood tests were implemented using the phylogenetic software package HYPHY (Pond et al., 2005) (see Supplementary File 2 for an example HYPHY script). This program is flexible in creating likelihood functions and optimizing them with a conjugated gradient ascent algorithm.
alternative model (in the null case $W_0(1) = W_1(1) = \text{length} + 2$). We then ranged the tree length from 0.5 to 5 with 50 datasets generated for every 0.5 increment in length. In contrast to previous simulations, the tree length was scaled to be exactly the length specified rather than the average length of a random tree. This was done to examine tree length in a more specific manner. All three scenarios previously described were tested where the results of the null case gave us the FP rate and the mild and extreme scenarios gave us two measures of sensitivity. Sensitivity was measured as the number of true positives divided by the number of actual positives ($N = 50$).

2.5 Primate data

As a test case generated from real data, we analysed the semenogelin I (SEMG1), semenogelin II (SEMG2), and peptidase inhibitor 3 (PPI) data sets that have previously been tested for an association with mating system and sperm competition. In addition, we analysed the mitochondrially encoded cytochrome b (CTFB) and portions of the zonadhesin ligand (ZAN) (Herlyn and Zischler, 2007). $dN/dS$ ratio in ZAN has been shown in primates to be negatively associated with body dimorphism, another measure of sexual selection. Sequences submitted by previous studies were downloaded from Genbank (Doros et al., 2004, Herlyn and Zischler, 2007; Herle et al., 2007; Jensen-Seaman and Li, 2003) (for GI numbers see Supplementary Table S1). Sequences were aligned using the linsi settings of MAFFT (Katoh et al., 2002) and manually checked for codon position. Premature stop codons are common in these datasets (Herle et al., 2007; Jensen-Seaman and Li, 2003) and sequence information after those positions was excluded for those taxa. We used the phylogenetic trees as previously published (Herlyn and Zischler, 2007; Herle et al., 2007) and estimated the branch lengths as part of the maximum likelihood tests. Previous results were verified for SEMG1 and SEMG2 by following the codon-based method of Ramm et al. (2008) but with more taxa included. This method assigns terminal branches for a given phenotype as fore branches and tests for selection by comparing model A and model A with $\omega_2 = 1$ (Wong et al., 2004; Zhang et al., 2005) from the PAML package (Yang, 2007).

Phenotypic information was assigned based on a binary classification of multimale-multifemale or not, similar to high and low sperm competition consistent with Herle et al. (2007). The one exception was the classification of dispersed breeding system (Pongo abelii and Microcebus murinus) being grouped as under low sperm competition because the sexual selection will not be as strong as with the multimale-multifemale case.

To test for heterogeneity, we calculated the likelihood under the GTR (Tavaré, 1986; Yang, 1994; Zharkikh, 1994) model, GTR + $\Gamma$ (Yang, 1996), and separate GTR matrices, with each repeated a minimum of five times from random starting positions to mitigate problems with optimization. Further, we calculated the likelihood of each dataset under the Independent model, Dependent model and Dependent model with the weight parameters fixed to each other.

3 RESULTS

3.1 Binary phenotype simulations

To examine the robustness of our methods to confounding factors of maximum likelihood in a phylogenetic framework, we simulated under four key variables: tree length (sum of all branch lengths), number of taxa, proportion of sites affected and strength of the association. We ran each permutation of these variables 100 times to create a distribution of LRT values that could then be compared with different significance thresholds ($\chi^2_{0.05}$, $\chi^2_{0.01}$). For results see Supplementary Table S2. We also ran simulations to evaluate the FP rate for the different variables.

The FP rate for the tests were within the acceptable range as expected by chance. The average number of significant tests across the other variables (proportion of sites, number of taxa and tree length), at the $\chi^2_{0.05}$ (df = 1) level was 4.94 with the greatest number being 9. Similarly for the 0.01 significance level the average number of significant tests at the $\chi^2_{0.01}$ (df = 1) level was 1.39 with the maximum being 4.

A critical feature of the method is the strength of association that it is able to detect. The two scenarios used here are described in Section 2, with the extreme case being a 1000-fold difference in rate between the two different phenotypes and the mild case being a 10-fold difference in rate. The average number of significant tests for the extreme case was 78.1 (max 100) at the $\chi^2_{0.05}$ level. The mild case averaged 65.0 (max 96).

Tree length is a measure of evolutionary divergence time with the greater amount of time conferring a higher probability of observing the underlying signal. The average number of significant results with a tree length of 1 under the more extreme scenario was 69.1 (max 93) and under the more mild scenario was 48.0 (max 72) at the $\chi^2_{0.05}$ level. In contrast, when the tree length was 3 the average under the extreme scenario was 87.1 (max 100) and 82 (max 96) for the more mild case, again under the $\chi^2_{0.05}$ level.

When this variable is examined more in depth, by a series of 0.5 incremental steps, the FP rate stays consistently low and the sensitivity is in 75–100% range after a tree length of 1 (Fig. 3). With a tree length of one the expected number of substitutions per site across the whole tree is one.

The number of taxa provide the data with which to measure the signal, i.e. the more taxa the greater number of instances to estimate your parameters and detect the signal you are searching for. Here, the average significant result with eight taxa, the fewest tested, was 58.5 (max 74) for the extreme case and 45.0 (max 70) under the mild case. With 32 taxa this number rose to 76.7 (max 96) for the mild case and 89.0 (max 95) for the extreme case. The 16 taxa case produced a result similar to the 32 taxa case: 86.8 (max 100) for the extreme and 73.3 (max 93) for the mild case.

The proportion of sites had a less drastic effect on the success of the method. The equal proportion of 0.5 had the best results with an average of 81.8 for the extreme case and 69.8 for the mild case. The proportions 0.25 and 0.75 did only slightly worse, with 57.8 and 67.3 respectfully for the mild case and 65.3 and 78.5 for the extreme case.
This system of LRTs provides the first models of which we are aware that are specifically designed to answer questions of genotype–phenotype integrating across the whole phylogeny. Previous methods had difficulties with the comparison of genotypic evolutionary rate parameters such as $dN/dS$ on branches and related phenotypes of extant taxa (Dorus et al., 2004; Herlyn and Zischler, 2007; Hurle et al., 2007) or ignoring error in ancestral phenotypic reconstructions (Nadeau et al., 2007). Our method overcomes these issues by estimating both phenotypic and genotypic evolution in an integrated framework over the entire tree.

### 3.2 Primate data

We obtained similar results to Ramm et al. (2008) for SEMG1 and SEMG2 using a similar procedure of the branch-site models (see Section 2). SEMG2 was significant for the model A versus model A (fixed $a = 1$) with a $P$-value of 0.009 ($df = 1$), the fore branches being set to terminal branches with taxa under high sperm competition and in this case including orangutan. When orangutan is excluded the $P$-value is still significant at 0.017 even after correction for multiple testing ($N = 2$). Again, parallelling their results, SEMG1 for both high and low sperm competition branches and SEMG2 for low, were not significant. SEMG1 was also not significant for either set of branches.

Next we tested for an association using our models. Tests for the Dependent versus the Independent model were highly significant for PI3, SEMG1 and SEMG2 (our unpublished data). From this we tested for violations of the rate heterogeneity assumption and all five datasets were highly significant ($P < 0.001$ for GTR versus 2x GTR). This held for individual codon positions as well, except that some codon positions in PI3 and ZAN were not significant (our unpublished data) presumably because of low power from the small number of nucleotides. But even PI3 and ZAN had some codon positions with significant heterogeneity.

When the second test was used (all weight parameters equal to one), PI3 and CYTB, our two negative controls, were found to be insignificant (Table 1). In contrast, our positive control, SEMG2, had a had significant $P$-value of 9.95e-4 (Table 1). SEMG2 retains significance at $P<0.005$ after Bonferroni correction ($N = 5$).

Saimiri boliviensis has a duplicated SEMG1 with no SEMG2 (Hurle et al., 2007) and both a and b copies of SEMG1 were included in the previous analysis. When either parologue was included alone in the analysis, the $P$-values were still insignificant.

### 4 DISCUSSION

This system of LRTs provides the first models of which we are aware that are specifically designed to answer questions of genotype–phenotype integrating across the whole phylogeny. Previous methods had difficulties with the comparison of genotypic evolutionary rate parameters such as $dN/dS$ on branches and related phenotypes of extant taxa (Dorus et al., 2004; Herlyn and Zischler, 2007; Hurle et al., 2007) or ignoring error in ancestral phenotypic reconstructions (Nadeau et al., 2007). Our method overcomes these issues by estimating both phenotypic and genotypic evolution in an integrated framework over the entire tree.

#### 4.1 Performance on simulated data

The method performed well on the various simulated scenarios and should be applicable to many enquires at various evolutionary time scales. We have only shown the use of the method in the binary phenotype case and hope to extend the models to accommodate a greater number of phenotype categories.

The Independent model versus Dependent model LRT is very susceptible to violations of rate homogeneity assumptions and we do not recommend its use. But the Dependent model with weight parameters fixed to each other versus Dependent model LRT is accurate in spite of rate heterogeneity.

Both scenarios investigated had a strong effect, 10-fold and 1000-fold changes in rate. Other simulation studies have shown that low levels of signal can make it difficult for likelihood methods to detect true positives (Wong et al., 2004). For example, both Adaptosite (Suzuki et al., 2001) and the site models implemented in codeml have difficulties detecting sites evolving with a $dN/dS$ of 1.5, from those evolving with $dN/dS$ of 1. Similarly, when we test our method with a weak scenario, 2-fold, our method has low power (our unpublished data). Results obtained from the method are conservative in nature and further investigations into sensitivity are needed.

In all our simulations, 1000 nt were used. We found that when this number was varied from 250 to 3000 the method performed well (our unpublished data). With less information it did not perform as well but was consistently conservative with a FP rate within acceptable limits and sensitivity increasing rapidly with the length of the alignment.

#### 4.2 Primate mating system and evolutionary rate of key proteins

We tested the method in a system where high rates of amino acid change have been associated with a behavioural/life history phenotype in primates at more than one locus, and where an association with high rates of overall nucleotide substitution is plausible. This signal is different from previous analyses because its focus is overall evolutionary rates associated with phenotype rather than adaptive positive selection identified by estimating $dN/dS$. We found that, as hypothesized, SEMG2 shows a significant associations...
between genotype and phenotype. This is unsurprising given its known functions in male reproduction but is reassuring in terms of the use of our method. SEMG1 and ZAN, even though involved in the same function are not associated. This has been observed in SEMG1 before (Ramm et al. 2008). However, there is evidence that on at least some lineages (human-chimpanzee) there is positive selection or elevated rates (Jensen-Seaman and Li, 2003; Kingan et al., 2003). In the case of ZAN, it was previously associated with dimorphic body size (Herlyn and Zischler, 2007), but not directly with mating systems as performed here.

We believe the association of SEMG2 to be functionally related because of the data previously presented on the molecular and cellular function of the proteins in question (Dixon and Anderson, 2002; Hurle et al., 2007; Lea et al., 2001), but as was stated before, this is not direct evidence of selection. We have not identified the specific sites that make this association but in future work we hope to provide such methods. Reasons for the large proportion of sites associated is not yet clear and further work will be needed to determine whether they are primarily evolving neutrally, under selection, or with a gene-specific explanation.

One caveat that should be taken into consideration is that FPs can arise when a limited amount of data is analysed or assumptions are violated. For example, when CYTB is examined with the same taxon sampling as SEMG2 it comes out as significant, whereas with more data (N = 27) it is not. One possible explanation for this is that CYTB is known to violate molecular clock assumptions (Nabholz et al., 2008) and we make this assumption in calculating phenotype parameter values and branch lengths.

4.3 Particulars of the models

Since the method is currently nucleotide based, it is not constrained to just protein evolution but can be applied to non-coding regions as well. The method can, theoretically, be expanded to use any number of genotypic rate models but its use in those scenarios has not been attempted here. Preliminary work with codon models has proven computationally difficult as the rate matrix is extremely large and difficult to evaluate [matrix exponentiation used in calculating probabilities of transitions is cubic at best with respect to the number of dimensions (Stoer et al., 2002) using eigen decomposition]. This method is not a search for selection but a first step in evaluating whether genes are involved in a particular function or phenotype. As mentioned previously, in addition to positive selection on a locus involved in the phenotype, other causal relationships are possible. For example, relaxation of constraint at a particular locus may also be associated with a phenotype, which could be a consequence of adaptive mutations upstream in an interacting pathway. This method could be the first step in localizing such a signal.

Hughes (2007) in his critique of maximum likelihood positive selection techniques mentions that functional associations are rarely investigated further as follow up to the detection of selection. Like these previous methods, our method is just the first step to identify candidate genes or interaction pathways, enabling the search for causal mutation(s) for a phenotype whether SNP, indel or major mutation, to be narrowed. Taken with methods to detect selection both at the genotype and phenotype level, system-level questions of evolutionary relationships between genotype and phenotype can be addressed using our method.

From a molecular evolution perspective, this method can be interpreted as an attempt to characterize rate heterogeneity or variations in constraint. Typically, rate heterogeneity is viewed as a confounding factor in phylogenetics (Pagel and Meade, 2004, 2008; Yang, 1996; Zhou et al., 2007), which is true in the search for associations between species. But it can also be viewed as a non-random signal of biological processes. Specifically, this method has the potential to relate heterogeneous signal to a meaningful biological relationship, even if not a causal relationship.

As previously mentioned, we hope to extend these models to detect specific sites that have associations with phenotypes. In addition, we hope to develop the models further to search more directly for causative sites, mostly by examining the rate of change of the phenotype compared with the state of an individual nucleotide or the reverse of this. Eventually, we hope that these methods can be used at the genomic level to detect functional associations between many genes and genomic regions and the phenotypic selection that has shaped their evolution.

5 CONCLUSION

We have successfully developed a hybrid substitution model, under a maximum likelihood phylogenetic framework, to test associations between the rate of evolution of genes and phenotypes. This method is successful under a variety of simulated situations and robust to site rate heterogeneity. In addition, we have applied our method to data sets of primate semen proteins and mating system and have shown that SEMG2 is significantly associated, while the control genes PUL and CYTB and two other candidate genes (ZAN and SEMG1) are not. This method can generate hypotheses based on molecular evolution which can then be verified using more direct functional assays and gives researchers an additional computational tool in their search for evolutionary relationships between genotype and phenotype.

ACKNOWLEDGEMENTS

We would like to thank Ziheng Yang for useful discussion and advice. We would also like to thank three anonymous reviewers for their insights and discussion. Most of the computation for the simulation and analyses were done on the CamGrid cluster via the mole server (http://mole.bio.cam.ac.uk) at the University of Cambridge and we thank the Cambridge eScience Center for its support of the system. This research could not have been completed in a timely manner without these services.

Funding: The Gates Cambridge Trust (to T.D.O), Leverhulme Trust (to N.J.M).

Conflict of interest: none declared.

REFERENCES

Dixon,A.L. and Anderson,M.J. (2002) Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatol., 73, 63–69.

Dona,S. et al. (2004) Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat. Genet., 36, 126–1329.

Dummond,A. and Strimmer,K. (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics, 17, 662–663.

Fawcett,T. (2006) An introduction to ROC analysis. Pattern Recogn. Lett., 27, 861–874.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17, 360–376.

Felsenstein,J. et al. (2004) Inferring Phylogenies. Sinauer Associates Sunderland, MA.
Nadeau,N.J. et al
Lynch,M. and Walsh,B. (1998)
Nachman,M.W. et al
Muse,S.V. and Gaut,B.S. (1994) A likelihood approach for comparing synonymous and
correlates negatively with body weight dimorphism in primates. Evolution, 48, 289–298.

Hestenes,M.R. and Stiefel,E. (1952) Methods of conjugate gradients for solving linear
Hughes,A.L. (2007) Looking for Darwin in all the wrong places: the misguided quest

Pagel,M. (1994) Detecting correlated evolution on phylogenies: a general method for

Sunderland, MA.

Slate,J. (2005) Quantitative trait locus mapping in natural populations: progress, caveats
and future directions. Mol. Evol., 14, 363–379.

Theron,E. et al (2001) The molecular basis of an avian plume polymorphism in the wild A melanocorax-l1 receptor protein mutation is perfectly associated with the melanin plume morph of the humanapi, Corvus frugilegus. Curr. Biol., 11, 550–557.

Williams,S.E. et al (2006) SLPI and elafin: one glove, many fingers. Clin. Sci., 110, 21.

Wong,W.S.W. et al (2004) Accuracy and power of statistical methods for detecting
adaptive evolution in protein coding sequences and for identifying positively
selected sites. Genetics, 168, 1041–1051.

Yang,Z. (1996) Among-site rate variation and its impact on phylogenetic analyses.
Trends Ecol. Evol., 11, 367–372.

Yang,Z. (1998) Likelihood ratio tests for detecting positive selection and application to
primates lsozyme evolution. Mol. Biol. Evol., 15, 568–573.

Yang,Z. (2006) Computational Molecular Evolution. Oxford University Press, New York, USA.

Yang,Z. (2007) PAML: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.,
24, 1586–1591.

Yanag,C.-I. et al (2007) Detecting coevolution in and among protein domains. PLoS Comput. Biol., 3, e21.

Yanag,C.-I. et al (2008) Identifying coevolving partners from paralogous gene families. Evol. Bioinform., 4, 97–107.

Zhang,J. et al (2005) Evaluation of an improved branch-Site likelihood method for
detecting positive selection at the molecular level. Mol. Biol. Evol., 22, 2472–2479.

Zharikhin,K. (1994) Estimation of evolutionary distances between nucleotide sequences.
J. Mol. Evol., 39, 315–320.

Zhou,Y. et al (2007) Evaluation of the models handling heterocytos in phylogenetic inference. BMC Evol. Biol., 7, 1471–2148.