Amino Acid Trait Association Model

Model description

We developed a pair of nested evolutionary models, a null and an alternative model, to test for associations between a target amino acid at an alignment site and a target binary trait. The target amino acid being tested, denoted $targetaa$, represents one of the 20 possible amino acids. The target trait, denoted $\mathcal{T}$, represents the trait being tested for an association with the target amino acid. The non-target trait is denoted $\mathcal{N}$. The traits, like the amino acids, are assumed to have evolved in an evolutionary manner.

The null model treats the amino acid evolution at a site and the trait evolution as independent of one another, whereas the alternative model treats the target amino acid and the target binary traits as potentially associated. This potential association is introduced into the alternative model via a dependence parameter λ.

For a given amino acid site and set of traits we used maximum likelihood estimation to estimate the parameters of both models and to obtain the maximum likelihood values. The maximum likelihood values were used to compare both models using a likelihood ratio test (LRT) and to calculate a p-value. If the LRT rejects the null model ($p < 0.05$) in favour of the alternative model, this suggests that the target amino acid and target trait are associated.

This association can be a positive association: the target amino acid and the target trait tend to co-occur together, or a negative association: the target amino acid and the target trait tend to actively avoid co-occurring together. When the maximum likelihood estimate for the dependence parameter is larger than one, $\hat{\lambda} > 1$, this suggests a positive association, and when $\hat{\lambda} < 1$, this suggests a negative association.

This test is somewhat analogous to a chi-squared test of association, except it accounts for phylogenetic correlations. A chi-squared test will treat each observation of amino acid and trait at the tips of a phylogeny as independent events, when in reality they are produced by an evolutionary process where the underlying number of events leading to those observations may be small. This is commonly referred to as a founder effect (Bhattacharya *et al.*, 2007), and can result in chi-squared associations whose significance is erroneously inflated. Our test avoids this by explicitly modelling the potential dependence between the amino acid and trait evolutionary processes and testing its significance relative to a model that treats

them as independent of one another.

Note that the traits, like the amino acids, are assumed to evolve in an evolutionary manner along the tree, and therefore our test is only appropriate where the trait can be described by an evolutionary process. This is the case for the HP and LP traits because they are a direct function of the presence or absence of an insertion, which is generated by an insertion-deletion evolutionary process along the tree. This test would not be appropriate for a trait such as patient survival, which represents a propensity along the tree rather than propagating in a discrete manner. For traits such as these we recommend the test outlined in Bhattacharya *et al.* (2007).

A formal description of the model is given as follows: the joint evolution of amino acids and traits are modelled using a 40×40 substitution model Q, that combines a 20×20 amino substitution model, A, and 2×2 by trait model, T. The trait model is a two-state continuous-time Markov model akin to Felsenstein's 1981 DNA substitution model (Felsenstein, 1981). The joint model is given as follows:

$$Q_{ij,mn} = \begin{cases} \mu A_{ij}\lambda & \text{if } i \neq j \text{ and } j = targetaa \text{ and } m = n = \mathcal{T} \\ \mu A_{ij}\frac{1}{\lambda} & \text{if } i \neq j \text{ and } j \neq targetaa \text{ and } m = n = \mathcal{T} \\ \mu A_{ij} & \text{if } i \neq j \text{ and } m = n = \mathcal{N} \\ \pi_{\mathcal{T}}\tau\lambda & \text{if } i = j = targetaa \text{ and } m = \mathcal{N} \text{ and } n = \mathcal{T} \\ \pi_{\mathcal{T}}\tau\frac{1}{\lambda} & \text{if } i = j \text{ and } j \neq target\text{aa and } m = \mathcal{N} \text{ and } n = \mathcal{T} \\ \pi_{\mathcal{N}}\tau & \text{if } i = j \text{ and } m = \mathcal{T} \text{ and } n = \mathcal{N} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where i and j represent the initial and end amino acid states, respectively, and m and n represent the start and end trait states, respectively. This formulation is motivated by the RNA base-pairing model of Muse (1995).

μ is a site-specific amino acid substitution rate, and τ is the trait substitution rate. A is a rate matrix given by the LG2008 substitution model. π^A is a vector of 20 amino acid frequencies as specified by the LG2008 model, and $\pi_{\mathcal{T}}$ and $\pi_{\mathcal{N}}$ represents the frequencies of the non-target ($\mathcal{T}$) and target traits ($\mathcal{N}$) states, respectively.

The equilibrium frequencies, π, of the alternative model are given by four separate cases corresponding to the two possible values for traits ($\mathcal{T}$ or $\mathcal{N}$), and whether the amino acid (aa) matches the target amino acid ($targetaa$) or not:

$$\pi_{aa=targetaa,trait=\mathcal{T}} = k^{-1}\pi^A_{aa}\pi_{\mathcal{T}}\lambda \quad (2)$$

$$\pi_{aa\neq targetaa,trait=\mathcal{T}} = k^{-1}\pi^A_{aa}\pi_{\mathcal{T}}\frac{1}{\lambda} \quad (3)$$

$$\pi_{aa=targetaa,trait=\mathcal{N}} = k^{-1}\pi^{A}_{aa}\pi_{\mathcal{N}} \tag{4}$$

$$\pi_{aa\neq targetaa,trait=\mathcal{N}} = k^{-1}\pi^{A}_{aa}\pi_{\mathcal{N}} \tag{5}$$

Where $\kappa = (\lambda + \frac{1}{\lambda})\pi_{\mathcal{T}} + 2\pi_{\mathcal{N}}$ is a normalising constant.

These equilibrium frequencies provide a intuitive way of understanding the influence of the association parameter λ. It is possible to get a sense of the expected frequencies of particular amino acid and trait associations for given values of λ. Furthermore, they can be used to predict, for a single sequence, the posterior probability of a trait given the amino aa at the target site:

$$p(trait = \mathcal{T}|aa = targetaa) = \frac{\pi_{\mathcal{T}}\lambda}{\pi_{\mathcal{N}} + \pi_{\mathcal{T}}\lambda} \tag{6}$$

$$p(trait = \mathcal{T}|aa \neq targetaa) = \frac{\pi_{\mathcal{T}}}{\pi_{\mathcal{N}}\lambda + \pi_{\mathcal{T}}} \tag{7}$$

Also note that the model is time-reversible, and therefore an unrooted tree can be used if the equilibrium probabilities are taken to be the initial probabilities at any rooting of the tree (Felsenstein, 1981).

Simulations

Table 1: Summary of benchmarks results

Simulated association strength	**Simulated rate of trait evolution**	**Recall**	**Precision**
Weak (2.0)	2.0	0.23	0.94
Intermediate (4.0)	2.0	0.35	0.96
Strong (8.0)	2.0	0.50	0.93
Weak (2.0)	4.0	0.26	0.92
Intermediate (4.0)	4.0	0.48	0.92
Strong (8.0)	4.0	0.64	0.96
Weak (2.0)	7.5	0.28	0.97
Intermediate (4.0)	7.5	0.62	0.98
Strong (8.0)	7.5	0.73	0.95

The Influenza H7 ML tree and corresponding HP and LP traits were taken and amino acid alignments were simulated. Each alignment consisted of 500 amino acids, with the first twenty sites of each alignment simulated as being associated with the traits, each having a different one of the 20 canonical amino acids as the target amino acid. The remaining

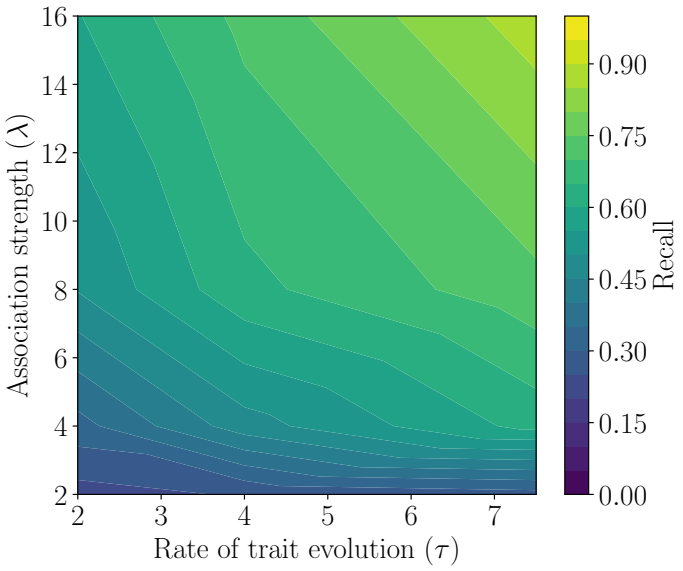


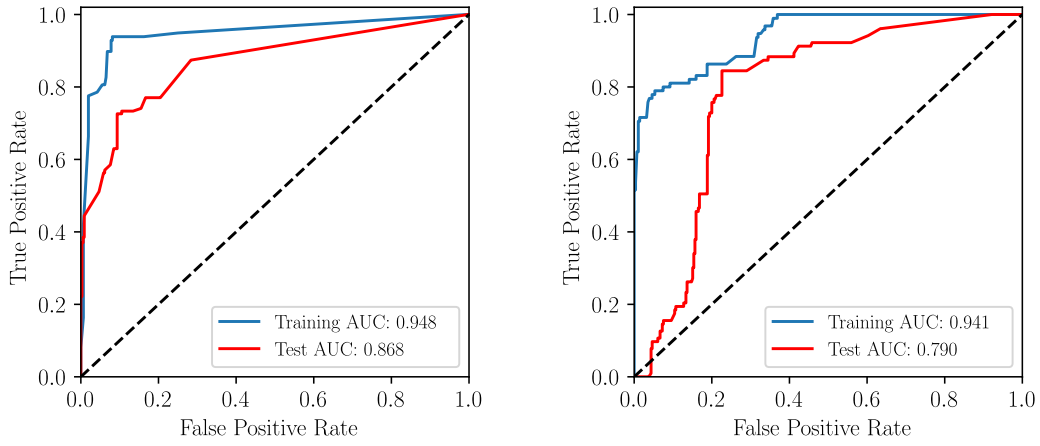
Figure 1: Contour plot of recall (blue-green-yellow colour gradient) as a function of simulated rate of trait evolution (x-axis) and simulated association strength (y-axis).

480 amino acid sites were simulated under the LG2008 model and were therefore treated as being independent of the traits. Three different degrees of association were simulated: weak, intermediate, and high, combined with three different rates of trait evolution (2.0, 4.0, and 7.0) - the inferred rate of trait evolution in the H7 ML tree was ∼4.0 and so this was selected as an intermediate value.

To account for the potential error introduced during tree inference, an ML tree was inferred using FastTree (Price *et al.*, 2010) for each of the simulated alignments. Potential associations were then estimated using our model and the Bhattacharya method (Bhattacharya *et al.*, 2007) on the first 40 sites of each alignment. The first twenty sites were used to measure the number of true positive and false negative detections, whereas the remaining twenty sites were simulated as independent of the traits and were used to measure the number of true negative and false positive detections. The recall and precision were calculated for each simulation using the number of true positives (TN), false-positives (FP), and false-negatives (FN). Recall and precision are defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$



(a) H7 ROC curve

(b) H5 ROC curve

Figure 2: Receiver operating characteristic (ROC) curves for H7 and H5 training and test datasets.

The results in Table 1 indicate that our model has a false-discovery rate (FDR= $100\% \times [1.0 - \text{Precision}]$) of $\sim 5\%$ across all test conditions which is consistent with our p-value significance threshold of 0.05. The recall of the model increases with the simulated association strength as expected, with stronger associations being more easily detected. Likewise, the recall of the model increases with a higher rate of trait evolution (Table 1 and Figure 1), which is also expected given that higher rates of trait evolution imply a greater number of trait events along the tree and therefore more the power in being able to detect an association.

Predicting high pathogenicity

H7 and H5 mutational panels were used to construct a predictive model of high pathogenicity (HP). This model uses the presence or absence of amino acid mutations to predict HP. For the H7 subtype there were 23 mutations selected (all H7 mutations in Table 1 and all H7 mutations with an association p-value less than 0.01), whereas for the H5 subtype there were 33 mutations selected (all H5 mutations in Table 1 and all H5 mutations with an association p-value less than 0.001). Taxa in the test and training datasets with missing segments were imputed. This was necessary because most taxa are not available as complete genomes.

Python's sklearn library was used to fit two random forest classifiers, one for H7 and one for H5 subtypes, respectively. The H7 and H5 subtype alignments were used as training datasets The classifiers each used 1000 decision trees, with a maximum decision tree depth of 4. Area Under the Curve (AUC) values of 0.868 and 0.790 were obtained for the H7 and H5 subtype test datasets, respectively (Figure 2).

Table 2: Mutational panel for H7 and random forest classifier importance weights for the training dataset. Positions are those in the H6 alignment.

Segment	**Mutation**	**Importance weight**
HA	143T	0.203
HA	274I	0.086
HA	438I	0.017
HA	384N	0.104
HA	402K	0.040
HA	335T	0.059
HA	287K	0.014
HA	175G	0.000
HA	152P	0.005
HA	175E	0.000
PB1	154D	0.109
PB1	152L	0.079
PB1	473I	0.044
PB1	709I	0.020
PB2	355K	0.003
PB2	480I	0.094
PB2	356I	0.005
PB2	584I	0.033
PB2	640I	0.009
PB2	655A	0.006
PB2	661A	0.003
PB2	702R	0.003
NS1	56A	0.021
NS1	180T	0.042

Table 3: Mutational panel for H5 and random forest classifier importance weights for the training dataset. Positions are those in the H5 alignment.

Segment	**Mutation**	**Importance weight**
HA	379R	0.219
HA	242I	0.014
HA	145L	0.016
HA	154I	0.016
HA	154L	0.033
HA	127L	0.119
HA	167T	0.073
HA	204I	0.061
HA	157P	0.053
HA	214V	0.052
HA	199N	0.084
HA	172T	0.008
HA	145A	0.000
HA	145V	0.004
HA	145I	0.000
HA	145Q	0.000
HA	145P	0.000
HA	519I	0.018
HA	145D	0.000
HA	145T	0.000
HA	170N	0.007
HA	145N	0.000
HA	489R	0.008
HA	87V	0.000
HA	149A	0.022
HA	145E	0.000
PB1	113I	0.032
PB2	674T	0.040
PB2	451T	0.027
PB2	627K	0.032
PB2	508Q	0.006
M	101K	0.031
M	205I	0.025

Software availability

Julia source code (compatible with Windows and Linux) is available at:
https://github.com/michaelgoldendev/trait-evolution

References

Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., *et al.* 2007. Founder effects in the assessment of hiv polymorphisms and hla allele associations. *Science*, 315(5818): 1583–1586.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6): 368–376.

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics*, 139(3): 1429–1439.

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3): e9490.