Models

Will DeWitt, Sarah Hilton, Andy Magee

Goal and Housekeeping

Andy re-implemented the model from Nasrallah and Huelsenbeck (2013), which models pair-wise epistasis due to RNA secondary structure. Below is the model as Andy implemented it in `RevBayes`. [**SH comment:** Comments from Sarah look like this.] [**WD comment:** Comments from Will look like this.] [**AM comment:** Comments from Andy look like this.]

RNA epistasis model

Here is the model from Nasrallah and Huelsenbeck (2013). [**SH comment:** I am just copying over what is found in the paper. We should change this around in an order we like and with notation we like.]

$\boldsymbol{Q}$ is the instantaneous rate matrix describing changes from doublet $\boldsymbol{x}$ to doublet $\boldsymbol{y}$. For $\boldsymbol{x} = (x_1, x_2)$, x_1 is the 5' nucleotide and x_2 is the 3' nucleotide.

$$\boldsymbol{Q} = \xi \times \begin{cases} \pi_{\boldsymbol{y}} S_{x_1,y_1} & \text{if single substitution at the 5' site,} \\ \pi_{\boldsymbol{y}} S_{x_2,y_2} & \text{if single substitution at the 3' site,} \\ \pi_{\boldsymbol{y}} S_{x_1,y_1} S_{x_2,y_2} d & \text{if double substitution where } \boldsymbol{x} \text{ and } \boldsymbol{y} \in \boldsymbol{W}, \\ 0 & \text{if any other double substitution,} \\ -\sum_{\boldsymbol{x} \neq \boldsymbol{y}} Q_{\boldsymbol{x},\boldsymbol{y}} & \text{if } \boldsymbol{x} = \boldsymbol{y} \end{cases} \quad (1)$$

$\boldsymbol{S}$ is the GTR exchangeability matrix (Tavaré, 1986) and S_{x_i,y_i} is understood to be the element in S governing the rate of exchangeability between nucleotide x_i and y_i (by definition, $S_{x_i,y_i} = S_{y_i,x_i}$), $\boldsymbol{W} = AT, CG, GC, TA$ is the set of Watson-Crick pairs, $\boldsymbol{\pi} = (\pi_{AA}, \pi_{AC}, ..., \pi_{TT})$ are the stationary state frequencies of the 16 possible doublet states, d controls rate of double to single mutations between doublets, and ξ is the rate-scaling factor.

Points to clarify

How do we interpret d?

d is the relative rate of double to single mutations between doublets or the "strength" of epistatic interactions.

[**AM comment:** Nasrallah says relative proportion, we've been saying relative rate. Are these the same?]

Here is a copy of table 1 from Nasrallah and Huelsenbeck (2013):

	$\pi_{\boldsymbol{y}} = \pi_{y_1}\pi_{y_2}$	$\pi_{\boldsymbol{y}} \neq \pi_{y_1}\pi_{y_2}$
$d = 0$	Independent and nonepistatic	Model inadequacy
$d > 0$	Dependent but nonepistatic	Dependent and epistatic

[**SH comment:** We talked about this last time but I want to make sure I totally understand. When we are simulating with $d = 0$, we are actually in the "model inadequacy" quadrant?] [**AM comment:** Yes!]

How do we normalize Q (ξ)?

[**SH comment:** From slack: 0.5 * Pr(single) + Pr(double)]

[**AM comment:** It is not completely clear how ξ is defined in the paper. Naïvely, I first assumed it was defined in the usual way (which RevBayes can do for us), simply taking a weighted sum of the off-diagonal elements $\xi^{-1} = \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y} \neq \boldsymbol{x}} \pi_{\boldsymbol{x}} Q_{\boldsymbol{x},\boldsymbol{y}}$. However, this normalizes the rate matrix on paired sites, which would count both doublet substitutions and single-substitutions equally. This formulation does not guarantee that the number of expected substitutions per unpaired epistatic site is the same as per non-epistatic site. Thus it seems that the appropriate normalization should be defined by,

$$\xi^{-1} = \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y} \neq \boldsymbol{x}} \pi_{\boldsymbol{x}} \times \begin{cases} \frac{1}{2} Q_{\boldsymbol{x},\boldsymbol{y}} & \text{if single substitution,} \\ Q_{\boldsymbol{x},\boldsymbol{y}} & \text{if double substitution of the allowed type,} \end{cases}$$

This *should* recognize the fact that single substitutions change only one site in the pair. We have to do this normalization ourselves, and can't just make RevBayes do it for us, but that's more me whining than anything important.]

[**WD comment:** Looking at the NH paper, the ξ "scaling factor" is a unified rate parameter for both Q and Q^*, but for Q the worry is that it's controlling expected numbers of events on pairs of sites, not on individual sites like for Q^*. To sort out what's up with ξ it I was thinking we'd want to compute the expected number of single subs per site in

a pair as

$$\mathbb{E}_{\mathbf{Q}}[\text{number of pair events}]\left(\frac{1}{2}\mathbb{P}(\text{single sub})+\mathbb{P}(\text{double sub})\right),$$

then demand that this equals the expected number of subs per site for the null model

$$\mathbb{E}_{\mathbf{Q}^*}[\text{number of events}].$$

Oh that's exactly what SKH has above, but I needed to get there myself ;).]

Given the fit parameter values for a GTR model, how do we simulate under this model?

[**SH comment:** From github, " This is the tree inferred under GTR+GAMMA using RAxML version 8.2.12. Inferred model parameters

alpha shape parameter = 0.440894

relative exchange rates (ac ag at cg ct gt) = 1.882161 7.009179 0.914813 0.495852 7.666181 1.000000

base frequencies = 0.340152 0.190828 0.225045 0.243974"]

[**AM comment:** Under the parameterization in the paper (standard MrBayes/RevBayes parameterization), we first take the RER and simplex them, yielding $\boldsymbol{r} = (0.0992, 0.3695, 0.0482, 0.0261, 0.404$ Then we make the symmetric GTR rate matrix $\boldsymbol{S}$ from the relative rates (for entirely too much detail, we put the elements of $\boldsymbol{r}$ into the upper diagonal of $\boldsymbol{S}$ row-wise, and the lower-diagonal column-wise, to make the symmetric exchangeability matrix). Then we draw the doublet stationary frequencies $\boldsymbol{\pi}$ from a Dirichlet(2,...,2) distribution. Then we assemble an unscaled version of 16 x 16 matrix $\boldsymbol{Q}$ as described 1, while calculating ξ. Then we normalize $\boldsymbol{Q}$, tell RevBayes that's our rate matrix on the flu tree, and draw from the induced phylogenetic CTMC distribution. Rev simulates these as characters 0,1,2,...,D,E,F. We also simulate the non-epistatic sites, and with an R script we turn the 16-state characters into pairs of sites, and add them to the nonepistatic sites.]

References

Nasrallah CA, Huelsenbeck JP. 2013. A phylogenetic model for the detection of epistatic interactions. *Molecular Biology and Evolution*. 30:2197–2208.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*. 17:57–86.