1 Challis-Schmidler Extension Sequence and Indel Models

In our new parameterization, there are no changes to the indel or sequence models, but they are presented here for reference. Let X and Y represent two proteins, where X is the ancestor of Y. Let S^X and S^Y be the sequence of amino acid characters for protein X and Y. The joint likelihood of S^X and S^Y, and an alignment M is:

$$\begin{aligned} P(S^X, S^Y, M|\lambda, \mu, t, Q) &= P(S^X, S^Y|M, t, Q)P(M|\lambda, \mu, t) \\ &= P(S^Y_M|S^X_M, t, Q)P(S^Y_{\bar{M}}|\pi) \\ &\times P(S^X|\pi)P(M|\lambda, \mu, t) \end{aligned} \tag{1}$$

Where λ is the birth rate and μ is the death rate in the indel model. t is the time interval. S^X_M and S^Y_M denote the matched (aligned) positions of S^X and S^Y, $S^Y_{\bar{M}}$ the unmatched positions of S^Y, Q the substitution rate matrix, and π the equilibrium distribution of characters.

2 Challis-Schmidler Extension Structural Model

In the structural component of the model, bonds between C_α's of consecutive amino acids drift in three-dimensional space according to an Ornstein–Uhlenbeck (OU) process. For a protein of length L there will be $L-1$ consecutive bonds. The bonds are represented as the vector between the C_α coordinates of consecutive amino acids.

Let $C_{ij}(t) - C_{i-1j}(t)$ be the jth coordinate of the bond vector between the ith C_α and the $i-1$ C_α. Then the equilbrium distribution of the structural diffusion process would be:

$$C_{ij}(t) - C_{i-1j}(t) \sim N(0, \tau) \tag{2}$$

$\tau = \frac{\sigma^2}{2\theta}$, where σ is the structural diffusivity coefficient and θ is the strength of the reversion to the mean in the OU process. The conditional distribution at time t, given time s:

$$C_{ij}(t) - C_{i-1j}(t)|C_{ij}(s) - C_{i-1j}(s) \sim N((C_{ij}(s) - C_{i-1j}(s))e^{-\theta(t-s)}, \tau(1 - e^{-2\theta(t-s)}) \tag{3}$$

Let B^X and B^Y be the sequence of bond vectors between consective amino acids, of proteins X and Y. Bond vectors in the ancestor B^X and bond vectors in B^Y that are not homologous to bond vectors in B^X are drawn from the equilibrium distribution. Bond vectors are only homologous if they connect amino acids in the

same alignment column (see pair-HMM section). Thus, the joint likelihood of B^X, B^Y and a sequence alignment:

$$\begin{aligned} P(B^X, B^Y, M|\lambda, \mu, t, \tau, \theta, R) &= P(B^X, B^Y|M, t, \tau, \theta, R,)P(M|\lambda, \mu, t) \\ &= P(B^Y_M|B^X_M, t, \tau, \theta, R)P(B^Y_{\bar{M}}|\tau, \theta) \\ &\times P(B^X|\tau, \theta)P(M|\lambda, \mu, t) \end{aligned} \tag{4}$$

Where R is the rotation of B^Y. $P(B^Y_M|B^X_M, t, \tau, \theta, R)$ is calculated according to equation 3. $P(B^Y_{\bar{M}}|\tau, \theta)$ and $P(B^X|\tau, \theta)$ are calculated according to equation 2.

3 Challis-Schmidler Extension Joint Sequence and Structure Model

Independence is assumed between the sequence substitution and structural diffusion processes, conditional on the indel process. The marginal likelihood of the sequence and structure data is obtained by mariginalizing over all possible alignments:

$$\begin{aligned} p(X, Y|\Theta) = &\sum_M P(B^X, B^Y, M|\lambda, \mu, t, \tau, \theta, R) \\ &P(S^X, S^Y, M|\lambda, \mu, t, Q)P(M|\lambda, \mu, t) \end{aligned} \tag{5}$$

with Θ representing the entire parameter set.

4 Challis-Schmidler Extension Pair HMM

The joint sequence and structure model can be written as a pair-HMM. To introduce the pair HMM, the amino acid survival probabilities in the TKF91 model need to be presented:

$$\alpha(t) = e^{-ut} \tag{6}$$

$$\beta(t) = \frac{\lambda(1 - e^{(\lambda-\mu)t})}{\mu - \lambda e^{(\lambda-\mu)t}} \tag{7}$$

$$\beta(t) = \frac{\mu(1 - e^{(\lambda-\mu)t})}{(1 - e^{-ut})(\mu - \lambda e^{(\lambda-\mu)t)}} \tag{8}$$

$\alpha(t)$ is the probability of ancestral survival, $\beta(t)$ is the probability of insertions given at least one surviving descendant, and $\gamma(t)$ is the probability of insertions given ancestral death. Below is the pair-HMM for the joint sequence and structure

model in the Challis-Schmidler model, transitions occur from the state indicated by row labels to the state indicated by column labels.

$$\begin{pmatrix} & \text{Start} & \text{Match} & \text{Delete} & \text{Insert} & \text{End} \\ \text{Start} & 0 & \frac{\lambda}{\mu}(1-\beta(t))\alpha(t) & \frac{\lambda}{\mu}(1-\beta(t))(1-\alpha(t)) & \beta(t) & (1-\frac{\lambda}{\mu})(1-\beta(t)) \\ \text{Match} & 0 & \frac{\lambda}{\mu}(1-\beta(t))\alpha(t) & \frac{\lambda}{\mu}(1-\beta(t))(1-\alpha(t)) & \beta(t) & (1-\frac{\lambda}{\mu})(1-\beta(t)) \\ \text{Delete} & 0 & \frac{\lambda}{\mu}(1-\beta(t))\alpha(t) & \frac{\lambda}{\mu}(1-\beta(t))(1-\alpha(t)) & \gamma(t) & (1-\frac{\lambda}{\mu})(1-\beta(t)) \\ \text{Insert} & 0 & \frac{\lambda}{\mu}(1-\beta(t))\alpha(t) & \frac{\lambda}{\mu}(1-\beta(t))(1-\alpha(t)) & \beta(t) & (1-\frac{\lambda}{\mu})(1-\beta(t)) \\ \text{End} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The pair-HMM for the joint sequence and structure model in the new model follows.

Pair-HMM 1:

$$\begin{pmatrix} & \text{Start} & \text{Match} & \text{Delete} & \text{Insert} & \text{End} \\ \text{Start} & 0 & p_{SM}P(B_0^X)P(B_0^Y) & p_{SD}P(B_0^Y) & p_{SI}P(B_0^X) & p_{SE} \\ \text{Match} & 0 & p_{MM}P(B_i^X)P(B_i^Y|B_i^X) & p_{MD}P(B_i^Y) & p_{MI}P(B_i^X) & p_{ME} \\ \text{Delete} & 0 & p_{DM}P(B_i^X)P(B_i^Y) & p_{DD}P(B_i^Y) & p_{DI}P(B_i^X) & p_{DE} \\ \text{Insert} & 0 & p_{IM}P(B_i^X)P(B_i^Y) & p_{ID}P(B_i^Y) & p_{II}P(B_i^X) & p_{IE} \\ \text{End} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Where p_{jk} is the transition probability from state j to state k in the Challis-Schmidler model. B_i^X is the ith bond vector for protein X. $P(B_i^Y|B_i^X)$ is calculated according to equation 3 and $P(B_i^X)$ and $P(B_i^Y)$ are calculated according equation 2. The major difference between this pair HMM and the one in the Challis-Schmidler model, is that the probabilities in the structure model are no longer emission probabilities, but instead have been multiplied by the transition probabilities. This is because, while a match state will always emit a homologous pair of amino acids according to the sequence model, it might not emit a homologous pair of bonds according to the structure model. For the transition MM the pair of bonds in X and Y will be homologous. This corresponds to the case when there is a "bond length" of 1 for both X and Y. We define bond length to be n+1, where n is the number of gaps between the current amino acid and the preceding amino acid. Bonds are viewed as homologous when they have the same length. In pairwise alignment, both X and Y cannot have a bond length longer than 1. If this were the case, then there would be at least one column with all gaps, and this is not possible. Therefore, it is only possible for the bonds in X and Y to be homologous when there is a MM transition. In all other cases, if a bond is emitted, the bond will be drawn from the stationary distribution. Therefore, to account for how entering a match state from a prior match state is different from

entering the match state from any other state, the transition probabilities have been multiplied by the probabilities in the structure model.

$P(B_0^X)$ and $P(B_0^Y)$ are set to 1 to account for the fact that a bond with the preceding amino acid does not exist for the first amino acid in a protein.

I have added a second possibility for the pair-HMM because it may be easier to think of the new structure model as adding an extra state to the Challis-Schmidler pair-HMM. In this pair-HMM, amino acids are emitted according to the sequence model and bond vectors are emitted according to the structure model. An extra state has been added to account for how a pair of amino acids emitted from a match state could either have homologous are non-homologous bonds.

Pair-HMM 2:

	Start	Hom Bond Match	Non-Hom Bond Match	Delete	Insert	End
Start	0	0	p_{SM}	p_{SD}	p_{SI}	p_{SE}
Hom Bond Match	0	p_{MM}	0	p_{MD}	p_{MI}	p_{ME}
Non-Hom Bond Match	0	p_{MM}	0	p_{MD}	p_{MI}	p_{ME}
Delete	0	0	p_{DM}	p_{DD}	p_{DI}	p_{DE}
Insert	0	0	p_{IM}	p_{ID}	p_{II}	p_{IE}
End	0	0	0	0	0	1

Since no new probabilities have been added to the rows, the transition probabilities for any state should still sum to 1. The emission probability of B_0^X and B_0^Y will be set to 1, to again account for the fact that a bond with the preceding amino acid does not exist for the first amino acid in a protein.

For either of these pair-HMMs, the forward algorithm could be used to marginalize over all possible alignments and calculate the marginal likelihood of sequence and structure data.

5 Herman et al Extension Indel and Sequence Models

With the Herman et al model, marginalizing over the internal node alignments results in exponential complexity. Herman et al avoided this problem by sampling an alignment, augmented with ancestral sequences, $\tilde{M}$. They then calculated the probability of the sampled alignment, $P(\tilde{M}|\Lambda, \Upsilon)$, where Λ is the indel parameters, and Υ is the tree.

The way I am doing things now, $P(\tilde{M}|\Lambda, \Upsilon)$ is calculated by using a pair-HMM similar to Pair-HMM 2, except that the transition probabilities have been modified so that TKF92 is used, instead of TKF91. Statalign starts at the tips of the tree,

and then iteratively calculates the probability of the alignment of child to parent, until it reaches the root. The indel pattern of the alignment of any child to parent specifies a path through Pair-HMM 2. Only the transition probabilities in the Pair-HMM are used to calculate the probability of the alignment.

The probability of the sequence model is, $P(S|\Phi, \tilde{M}, \Upsilon)$, where S is the sequence data and Φ is the sequence model parameters. $P(S|\Phi, \tilde{M}, \Upsilon)$ is calculated using the pruning algorithm.

6 Herman et al Extension Structure Model

I have already stated how we have changed the parameterization of the structural diffusion process so that bond vectors are diffusing instead of amino acids (see section "New Structural Model").

For an OU process on a tree, the joint distribution of the data at the leaves is a multivariate Gaussian with covariance matrix, $\Sigma_{kl}[\tau, \theta, \Upsilon] = \tau e^{-\theta d_{kl}(\Upsilon)}$, where $d_{kl}(\Upsilon)$ is the distance between leaves k and l along the branches of Υ.

Let A be the set of all unique bond lengths for amino acids in column i. Let $B_j^{(M_i)}$ be the length - $|M_i|$ vector obtained by taking the jth coordinate of each bond vector for each observed (leaf) protein containing an amino acid at the ith column with a bond of length $n+1$. The marginal likelihood of the observed bond vectors is then given by a product of $|A|$ bond lengths, L columns of the alignment and the three spatial dimensions:

$$P(B|\tilde{M}, \Theta, \Upsilon) = \prod_{(n+1)\in A} \prod_{i=1}^{L} \prod_{j=1}^{3} N_{|M_i|}(B_j^{(M_i)}|0, \Sigma_{M_i}[\tau, \theta, \Upsilon]) \tag{9}$$

Where Σ_{M_i} is a submatrix of Σ of dimension $|M_i|$ formed by selecting the columns and rows corresponding to the homologous bonds for the alignment column M_i.

7 Herman et al Extension Joint Sequence and Structure Model

The joint sequence and structure model would then be:

$$P(S, B|\tilde{M}, \Phi, \Theta, \Lambda, \Upsilon) = P(S|\Phi, \tilde{M}, \Upsilon)P(B|\tilde{M}, \Theta, \Upsilon)P(\tilde{M}|\Lambda, \Upsilon) \tag{10}$$