Statistical Genetics, Spring 2024

Class notes 2

Types of models for mutational processes

1. A nucleotide substitution model describes the process of moving between the 4 states: A,C,G,T. Usually described by a 4×4 matrix with substitution probabilities or rates.

$$P = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix} \end{array} \qquad M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} * & \lambda_{AC} & \lambda_{AG} & \lambda_{AT} \\ \lambda_{CA} & * & \lambda_{CG} & \lambda_{CT} \\ \lambda_{GA} & \lambda_{GC} & * & \lambda_{GT} \\ \lambda_{TA} & \lambda_{TC} & \lambda_{TG} & * \end{pmatrix} \end{array}$$

2. Within coding regions (exons in genes) triplets of nucleotides define codons which encode amino acids - the basic unit of a protein. A codon substitution model is a 64×64 table. Such a table has a huge number of free parameters, so hard to work with.
   Key concept: *synonymous* vs *non-synonymous* mutations. 61 of the 64 triplets encode 20 possible amino acids (≈ 3 per amino acid), and the other 3 encode *stop*. Mutations that do not change amino acid (and hence the protein) are synonymous. In coding regions we do indeed see that they are a lot more common because of *selection*.
   For example: A gene where the synonymous and non-synonymous rate are very different is likely to be important and/or sensitive to structure changes.

3. Tandem repeats, in particular STRs, where the mutation process is a change in the number of repeats (get longer or get shorter).

Nucleotide substitution models

Reminder: A,G are *purines*, while C,T are *pyrimidines*. Mutation within group ($A \leftrightarrow G$, $C \leftrightarrow T$) are called *transitions*, between groups *transversions*.

We can think of the process in discrete time (e.g., generation) with the matrix P above describing transition probabilities, e.g.:

$$p_{AG} = \mathbb{P}(X_{T+1} = G | X_T = A) := P_{AG}.$$

This is a stochastic process with transition matrix P. Hence, in t generations we will get:

$$\mathbb{P}(X_{T+t} = G | X_T = A) = (P^t)_{AG}.$$

Assuming the state space is fully connected, we will have a stationary probability Π, which will solve: $\Pi = \Pi P$.

A more realistic model, and easier to analyze is the continuous time version M above. The parameters $\lambda_{..}$ are exponential "mutation rates" from the first state to the second. It is more common to write it with the "exit rate" on the diagonal:

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\lambda_A & \lambda_{AC} & \lambda_{AG} & \lambda_{AT} \\ \lambda_{CA} & -\lambda_C & \lambda_{CG} & \lambda_{CT} \\ \lambda_{GA} & \lambda_{GC} & -\lambda_G & \lambda_{GT} \\ \lambda_{TA} & \lambda_{TC} & \lambda_{TG} & -\lambda_T \end{pmatrix} \end{array},$$

where $\lambda_A = \lambda_{AC} + \lambda_{AG} + \lambda_{AT}$. Assume we have a distribution over states $q(t) = q(0)M(t)$ at time t, we can now write for infinitesimal $\triangle t$:

$$q(0)M(T + \triangle t) = q(t + \triangle t) \approx q(t) + q(t)M\triangle t,$$

with error of order $O((\triangle t)^2)$, and therefore:

$$\frac{\partial M(t)}{\partial t} = M \cdot M(t),$$

From this simple differential equation we can conclude the general formula for a transition matrix:

$$M(t) = \sum_{i=0}^{\infty} \frac{(Mt)^i}{i!} := e^{Mt}.$$

Now take the diagonalizing transformation $M = R^{-1}DR$, with $D = (d_j)$ diagonal. We get:

$$M(t) = \sum_{i=0}^{\infty} \frac{(R^{-1}DRt)^i}{i!} = R^{-1}\left(\sum_{i=0}^{\infty} \frac{(Dt)^i}{i!}\right) R = R^{-1}e^{Dt}R,$$

where e^{Dt} is a diagonal matrix with $d_j t$ on the diagonal.

A vector Π is the stationary distribution if $\Pi M = 0 \Rightarrow \Pi M(t) = \Pi$.

Time reversibility

A model is time reversible if $\Pi_j M_{ji}(t) = \Pi_i M_{ij}(t)\ \forall i, j$, meaning $M_{ij}(t)$ is the same whether time moves forward or backward. An equivalent formulation:

$$\Pi_j M_{ji} = \Pi_i M_{ij}.$$

When considering substitution models with unequal stationary probabilities ($\Pi_i \neq \Pi_j$,) it is common to express the limitation to time reversibility by making the "relative" transition rates equal $\lambda_{ij} = \lambda_{ji}$ but multiplying them by the complementary stationary probabilities:

$$M(GTR) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} * & \Pi_C\lambda_{AC} & \Pi_G\lambda_{AG} & \Pi_T\lambda_{AT} \\ \Pi_A\lambda_{AC} & * & \Pi_G\lambda_{CG} & \Pi_T\lambda_{CT} \\ \Pi_A\lambda_{AG} & \Pi_C\lambda_{GC} & * & \Pi_T\lambda_{GT} \\ \Pi_A\lambda_{AT} & \Pi_C\lambda_{TC} & \Pi_G\lambda_{TG} & * \end{pmatrix} \end{array},$$

It is common to consider only time-reversible processes for mathematical convenience, although the intuition is not really clear.

Common models

Jukes-Cantor (JC69): Simplest model: $\lambda_{ij} = \begin{cases} \lambda & i \neq j \\ -3\lambda & i = j \end{cases}$ .. In this model easy to see:

1. $\Pi_i = 1/4 \; \forall i$.
2. Number of mutations in given time t distributed $Pois(3\lambda t)$.
3. $$M_{ij}(t) = \begin{cases} 1/4 - 1/4 \cdot \exp\{-4\lambda t\} & i \neq j \\ 1/4 + 3/4 \cdot \exp\{-4\lambda t\} & i = j \end{cases} .$$
   Proof: HW. Can be proven with similar arguments as we did for Poisson parity, or with differential equations.

   This model has just one parameter to estimate from data.

Problem: it is not realistic, ignoring transitions vs transversions, and requiring same marginals.

Kimura 2-rate (K80):

$$\lambda_{ij} = \begin{cases} \alpha & transition \\ \beta & transversion \\ -\alpha - 2\beta & i = j \end{cases} .$$

Typically $\beta << \alpha$.

Model with two parameters. What about Π and Poisson number of events?

HKY85: Combining non-uniform stationary $(\pi_A, \pi_C, \pi_G, \pi_T)$ with difference in transitions and transversions:

$$\lambda_{ij} = \begin{cases} \pi_j \alpha & transition \\ \pi_j \beta & transversion \end{cases} .$$

This model has 5 parameters (why not 6?). Is the number of mutations in time t still Poisson distributed?

Given a sequence, for each one of these models, we can assume the model is the same for all sites, or different. One common assumption, as we did in the binary case, is that there is an overall rate parameter with Gamma distribution: $\lambda \sim \Gamma(\theta, 1)$, so for example for $K80 + \Gamma$ we would get:

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} * & \lambda\beta & \lambda\alpha & \lambda\beta \\ \lambda\beta & * & \lambda\beta & \lambda\alpha \\ \lambda\alpha & \lambda\beta & * & \lambda\beta \\ \lambda\beta & \lambda\alpha & \lambda\beta & * \end{pmatrix} \end{array} ,$$

with three parameters (what is the stationary distribution? why don't we need to estimate the scale parameter of the Gamma distribution?)
Is the number of mutations in this model in time t Poisson distributed? What about $HKY85 + \Gamma$?

STR evolution models

Reminder: Short tandem repeats are short DNA sequences (up to 5 bases) that repeat several times like *ATGATGATGATGATGATG*. The aspect that mutates is the number of copies, and the mutation rate is high (many such copying mistakes), making STR very informative for understanding evolutionary history and relationships.

The simplest standard model (often called **Stepwise model**) is one of symmetric random walk (in discrete or continuous time). Denote $l(t)$ the number of repeats in time t then we can describe the process as:

- Moving from $l(t)$ to $l(t)+1$ at rate λ
- Moving from $l(t)$ to $l(t)-1$ at rate λ

It is easy to see that the number of mutations in this process in time δt has $\text{Pois}(2\lambda \cdot \delta t)$ distribution.

Now assume we are given STR counts in two species (initially we can assume just a single STR, so each species is represented by a number). Since the direction of time does not matter we will denote them by X_0, X_t where the time length is t. How can we estimate t? First we can try the same approach as in SNPs:

$$\mathbb{E}(X_t - X_0) = 0 \ , \ \mathbb{P}(X_t = X_0) = \sum_{i=0}^{\infty} \exp(-2\lambda t) \left(\frac{(\lambda t)^i}{i!} \right)^2 ,$$

we can try to use this, but it does not use the information that we observe the actual difference $X_t - X_0$. Due to symmetry we see that we cannot use its first moment to estimate time, but can we use its second moment? The key is the simple relationship:

$$\mathbb{E}(X_t - X_0)^2 = 2\lambda t.$$

Before proving it, we can easily see how to use it in our canonical problems of calibration and time estimation:

- Given t, $\hat{\lambda} = \frac{(X_t - X_0)^2}{2t}$.
- Given λ, $\hat{t} = \frac{(X_t - X_0)^2}{2\lambda}$.

This is the basis of the famous $(\delta\mu)^2$ method for estimating the "genetic distance" between two species observed at K different STRs:

$$(\delta\mu)^2 = \frac{\sum_{k=1}^{K} (X_{tk} - X_{0k})^2}{K} \ , \ \ \mathbb{E}\left((\delta\mu)^2\right) = 2\lambda t.$$

Proof of result:
For our continuous time case, we can divide the interval into t/δ tiny intervals of size δ. In each interval we have:

$$|X_u - X_{u-\delta}| = \begin{cases} 0 & \text{w.p} \ \ \exp(-2\lambda\delta) \\ 1 & \text{w.p} \ \ 1 - \exp(-2\lambda\delta) - o(\delta) \\ > 1 & \text{w.p} \ \ o(\delta) \end{cases}$$

and therefore:

$$\mathbb{E}(X_u - X_{u-\delta}) \approx 1 - \exp(-2\lambda\delta) \approx 2\lambda\delta.$$

Noting the increments are independent in this model (lack of memory!), with infinitesimal delta (or moving from sums to integrals) it yields the desired result:

$$\mathbb{E}(X_t - X_0)^2 = 2\lambda t.$$

Issues with the stepwise model

Why is the stepwise model not possibly a realistic model, especially over long time spans (far away species)? The basic fact about this model is $\mathbb{E}(X_t - X_0)^2 \to \infty$, meaning:

1. There is no stationary distribution
2. With probability one the length will exit any finite interval, given long enough time
3. In particular, by definition $\mathbb{P}(\exists t : X_t < 0) = 1$, which makes less than zero sense

The model also does not allow changes of size more than 1, which we know occur in practice.

So if we want a more realistic model we need a more complex one, which fixes these issues, and hopefully also follows our understanding of the biology / chemistry underlying these mutation models. For example, to have a stationary distribution we must have that $\mathbb{E}(X_t - X_0)^2 \to C < \infty$ as $t \to \infty$.

The model proposed by Whittaker et al. considers possible relevant effects:

- Allowing for jumps of different sizes
- Dependence on the current count X_t: Do longer STRs have more mutations?
- Allowing different rate for increase and decrease in count
- Allowing interaction: the rates of increase and decrease depend differently on the length

Their most general model for rate or probability of moving from i to j with $i < j$:

$$\mathbb{P}(i \to j) = \gamma_u \exp(\alpha_u i) \exp(-\lambda_u (j - i)).$$

For $i > j$:

$$\mathbb{P}(i \to j) = \gamma_d \exp(\alpha_d i) \exp(-\lambda_d (i - j)),$$

for a total of 6 parameters.