Introduction to Coalescent Models – Part II

Biostatistics 666

Key Ingredients

- Coalescent approach
  - Proceeding backwards through time.
  - Genealogy for a sample of sequences.

- Infinite sites model
  - All mutations distinguishable.
  - No reverse mutation.

Important results

- Probability of sampling distinct ancestors for *n* sequences

$$P(n) = \prod_{i=1}^{n-1}\left(1 - \frac{i}{N}\right) \approx 1 - \frac{\binom{n}{2}}{N}$$

- Coalescence time t is approximately exponentially distributed

Tree

- Coalescence Times (in 2N units)

$$E(T_j) = 1 / \binom{j}{2}$$

- Total Length (in 2N units)

$$E(T_{tot}) = \sum_{i=1}^{n-1} \frac{2}{i}$$

- Number of Mutations

$$E(S) = 4N\mu \sum_{i=1}^{n-1} 1/i = \theta \sum_{i=1}^{n-1} 1/i$$

Coalescent approach

- Generate genealogy for a sample of sequences.
  - Introduces computational and analytical convenience.

- Instead of proceeding forward through time, go backwards!

Example II

- What is the probability that two sampled sequences are identical?
  - Proceed until common ancestor…
    - P(CA)=1/2N (diploids)
    - P(CA)=1/N (haploids)
  - … or mutation
    - P(mutation)=2μ
  - Assume only one of these will occur

The answer…

$$P_2(S\text{ is }0) \approx \frac{P_{CA}}{P_{CA}+P_{mut}}$$

$$= \frac{1/2N}{1/2N+2\mu}$$

$$= \frac{1}{1+\theta}$$

Full distribution of S…

- Probability that first *j* events are mutations…

$$P_2(j) = \left(\frac{\theta}{1+\theta}\right)^j \left(\frac{1}{1+\theta}\right)$$

Example…

- 2 sequences
- Population size N = 25,000
- Mutation rate $\mu = 10^{-5}$

- Probability of 0, 1, 2, 3… mutations

And for multiple sequences…

- Proceed back in time, until:
  - One of *n* sequences mutates…
    - Probability approximately $n\mu$
  - A coalescent event occurs…
    - Probability approximately $n(n-1)/4N$

- Using these, define number of mutations during time with *n* lineages

Giving …

$$Q_n(j) = \left(\frac{n\mu}{n\mu + \frac{\binom{n}{2}}{2N}} \right)^j \frac{\frac{\binom{n}{2}}{2N}}{n\mu + \frac{\binom{n}{2}}{2N}} = \left(\frac{\theta}{\theta + n - 1} \right)^j \frac{n-1}{\theta + n - 1}$$

$$P_n(j) = \sum_{i=0}^{j} P_{n-1}(j-i) Q_n(i)$$

Example…

- 3 sequences
- Population size N = 25,000
- Mutation rate $\mu = 10^{-5}$

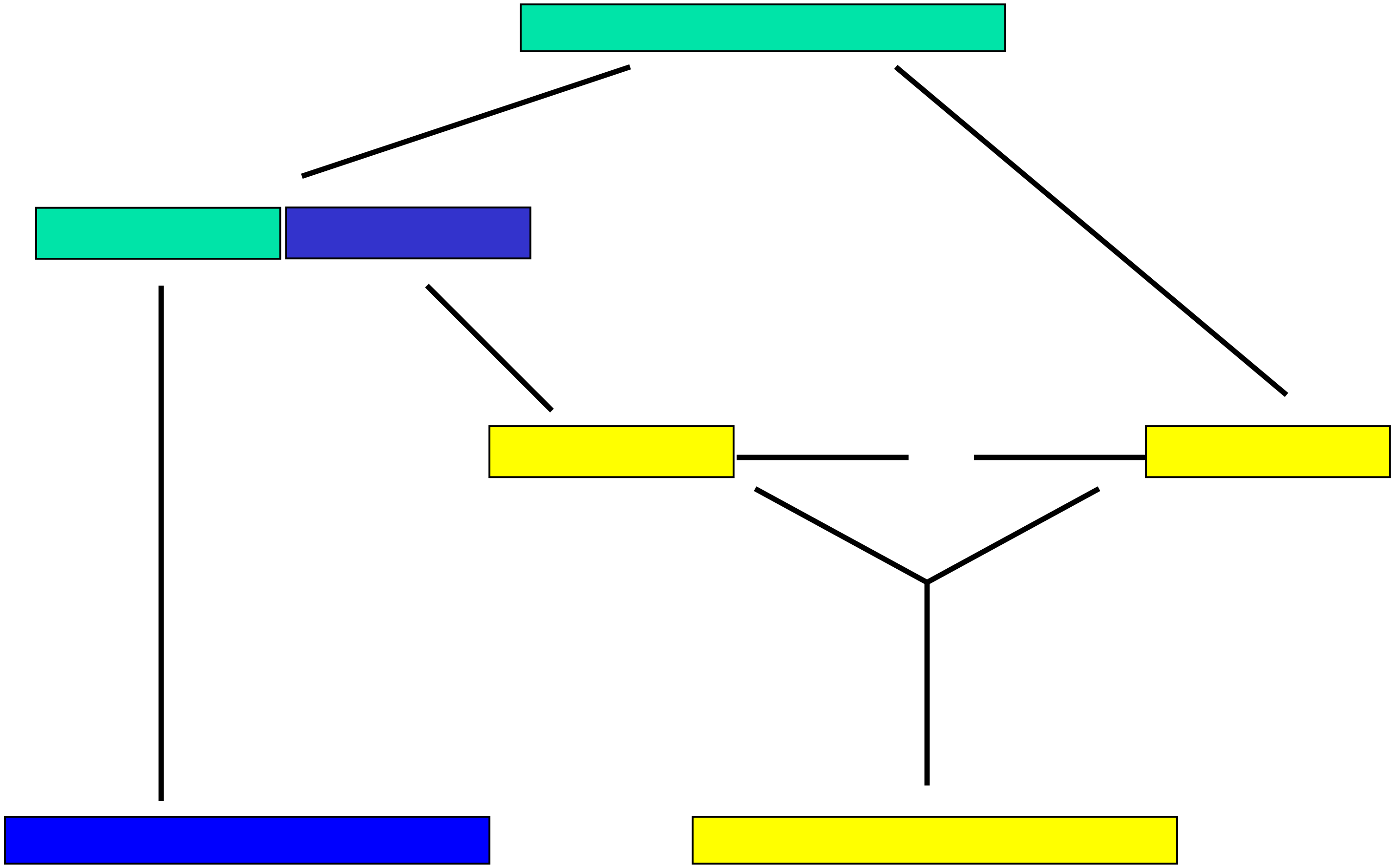
- Probability of 0, 1, 2, 3… mutations

Recombination

- When there is little recombination between two sites, their genealogies should be very similar

- This correlation between genealogies generates linkage disequilibrium

Two Locus Coalescent



Generating Genealogies

- Proceed backwards in time, until…
  - Coalescent event

    $$P_{CA} \approx \binom{n}{2} / 2N$$

    - Reduces number of ancestors by 1
  - Recombination

    $$P_{rec} \approx nr$$

    - May increase number of ancestors by 1

P(First Event is CA)

$$P(\text{no rec}) = \frac{P_{CA}}{P_{CA} + P_{rec}} = \frac{\binom{n}{2}/2N}{\binom{n}{2}/2N + nr}$$

$$= \frac{n-1}{4Nr + n - 1}$$

$$= \frac{n-1}{R + n - 1}$$

Coalescent W/ Recombination

- Analytical results are difficult

- Typical approach is to simulate trees
  - Study sample properties they imply

- We will only discuss these briefly…

Total number of mutations

- Recombination does not change expectation for S…

$$E(S) = 4N\mu\sum_{i=1}^{n-1} 1/i = \theta\sum_{i=1}^{n-1} 1/i$$

- … but it reduces its variance.
  - With large *r*, S is effectively averaged over multiple genealogies

Expectation for Δ^2

- Δ^2 has large variance, and inferences could be inaccurate

$$E(\Delta^2) \approx \frac{1}{1+R}$$

- Rough approximation
  - R>5, θ small, 0.05 < allele frequency p < 0.95
- Hill and Weir (1994) AJHG **54:**705-714

Interesting Questions

- Frequency spectrum of observed mutations
  - Impact of population growth
  - How many mutations are unique?
- Disequilibrium coefficient
  - Joint distribution of (p_A, p_B, D_{AB})
  - Impact of population growth

Recommended Reading

- Richard R. Hudson (1990) “Gene Genealogies and the coalescent process”
    - from Oxford Surveys in Evolutionary Biology, Vol. 7. D. Futuyma and J. Antonovics (Eds). Oxford University Press, New York.