Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

Human Population Genomics

Outline

1 Recapitulation: Coalescence

"Damn the Human Genomes. Small initial populations; genes too distant; pestered with transposons; feeble contrivance; could make a better one myself."

–Lord Jefferey (badly paraphrased)

Outline

1 Recapitulation: Coalescence

Wright-Fisher Model

- Recombination and coalescent are treated as competing processes that determine the graph structure of the genealogy of n ($n \ll N$) samples from a population of fixed size of N male individuals, and N female individuals.
- A recombination event (ignoring gene conversion events for the time being) takes place with probability r. The recombination point is chosen uniformly along the paternal and maternal sequences, and the sequences recombine.

- **Wright-Fisher Model**: N diploid females $+$ N diploid males.
- Each individual creates a haploid genome by recombination each time they contribute genetic materials to the next generation.
- An individual in the next generation is then made by choosing one haploid genome from the males and the other from the females.
- The $2N$ resulting individuals in the next generation are then divided equally into males and females.

- We wish to view this process backward in time.
- The effect of recombination is that the ancestral material to a specific sequence is found on two DNA sequences in the parent, etc.
- If we focus on a single base on the sequence, it should follow the classical Wright-Fisher model with coalescent and mutation. A very close-by base (without having a recombination event in the intervening region) will also follow the same Wright-Fisher model.
- The tree at the single base position is called a *local tree*. As we move along the genome, base-by-base, the local tree topologies will change only when one encounters a recombination event. The local-trees at different bases are the correlated instances of Wright-Fisher process without recombination.

Algorithm

- We can create a graph structure describing the genealogy process, as follows: As we move backward it time, from one generation to the previous generation, recombination will cause splitting; coalescence will cause merging.
- The scaled mutation rate is $\theta = 4Nu$ and the scaled recombination rate is $\rho = 4Nr$. (ρ is sometimes called the population recombination rate.)
- There are two algorithms to model coalescence with recombination: (i) Hudson's back-in-time algorithm & (ii) Wiuf and Hein's spatial algorithm. We will cover the first only.

Ancestral Recombination Graph (ARG)

- Suppose we start with n extant genes. Suppose that the first (earliest) event encountered is a recombination event. Just after the recombination event, there are n sequences carrying the ancestral materials for the n sequences in the sample.
- Before that recombination event, there are two ancestors who recombined their genetic material to create one of the ancestors (a brand new one) of the n extant sequences in the sample. One ancestor contributed genetic material to the left of the recombination break point, the other to the right of it.

- The waiting time (measured backward, but in terms of generations) for this recombination event to have occurred is geometrically distributed with parameter $r = \rho/4N$.

$$Pr(T_R = j) = r(1-r)^{j-1}.$$

- Thus we have the following (with time now rescaled):

$$Pr(T_R^c \leq t) = 1 - (1-r)^{\lfloor 2Nt \rfloor} \approx 1 - e^{-2Nrt} = 1 - e^{-\rho t/2}.$$

- If there are currently k sequences ancestral to the sample, the time to the next recombination event is distributed exponentially with parameter $k\rho/2$. The recombination event is equally likely to occur in any of the k ancestors, and the position of the recombination breakpoint in the chosen sequence is picked uniformly over he sequence length.

Stochastic Algorithm to Sample Genealogies for n Genes

- **Algorithm**
  1. Start with $k = n$ genes. Repeat until $k = 1$:
     1. Simulate the waiting time T_k^c to the next event $T_k^c \sim Exp(k(k-1+\rho)/2)$.
     2. With probability $(k-1)/(k-1+\rho)$ the event is a coalescent event; otherwise, it is a recombination event.
     3. If it is recombination: choose a random sequence and a random point on the sequence. Create an ancestor sequence with the ancestral material to the left of the chosen point and second ancestor sequence with the ancestral material to the right of the recombination point.Increase the sample size by one: $k \mapsto k+1$
     4. If it is coalescent: choose a random pair (i,j) with $1 \leq i < j \leq k$ uniformly from the $\binom{k}{2}$ possible pairs. Merge i and j into one gene and decrease the sample size by one: $k \mapsto k-1$.

Computational Relevance of the Coalescent Models

- Given data D,
- Parameter(s) Θ,
- Model M.

We wish to make inference re. $f(\Theta|D)$.

$$f(\Theta|D) = f(D|\Theta)\pi(\Theta)/P(D).$$

where $\pi(\Theta) =$ Prior & $P(D) =$ Normalizing constant

The problem

- Given data D,
- Parameter(s) Θ,
- Model M.

Validate the model; interpret the data; ...

- Data sets are growing much larger.
- Larger implies more complex.
- Traditional analysis methods may fail or become computationally intractable. ... $[f(D|\Theta)]$

Possible response

- Construct better theory
- Use simpler (less realistic) models;
- ‘Approximate’ methods.

Ancestral methods with no recom (haploid data)

- **A stochastic (Markov) process.**
- Time between events is exponentially distributed
- As we look back in time two events may occur:
  1. Two lines of ancestry will coalesce to form a single line of ancestry, with prob. $(k-1)/(k-1+\theta)$ where there are currently k lines and $\theta/2$ represents the mutation rate. (Pick a random pair of lines)
  2. A mutation will occur to a line of ancestry, changing the type of a gene, with prob. $\theta/(k-1+\theta)$. (Pick a random line)
- The process continues until there is a single line of ancestry: the most recent common ancestor (MRCA) of the sample.

Coalescent with recombination (diploid data)

- As we look back in time three events may occur:
  1. Two lines of ancestry will coalesce to form a single line of ancestry, with prob. $(k-1)/(k-1+\theta+\rho)$ where there are currently k lines and $\theta/2$ represents the mutation rate. (Pick a random pair of lines)
  2. A mutation will occur to a line of ancestry, changing the type of a gene, with prob. $\theta/(k-1+\theta+\rho)$. (Pick a random line)
  3. A recombination will occur to a line, splitting it into two, with prob. $\rho/(k-1+\theta+\rho)$. (Pick a random line)
- The process continues until there is a single line of ancestry: the most recent common ancestor (MRCA) of the sample.

Points of interest

- Not all mutations on the recombination graph impact the sample.
- Not all recombinations impact the sample.
- The space of possible graph topologies is (very!) infinite (c.f. the finite space of possible coalescent tree topologies).

Ancestral Processes with Recombination

- **Key observation:** Each locus still follows a coalescent
- Explicitly allows for the non-independence of multiple loci and use all data simultaneously.
- Recombination makes life much more difficult.
- Can wait a long time for the MRCA.

Can the coalescent produce human data?

"Calibrating a coalescent simulation of human genome sequence variation," Schaffner, et al. *Genome Research*, **15**:1576-1583, 2005.

Approximating the model: Fast "Coalescent" Simulation

- **Goal**
- *A faster way of producing coalescent data for chromosomal-length regions (cf. existing methods such as Hudson's ms (mksamples software))*

Fast “Coalescent” Simulation

- Why? Growth of genome-wide data (e.g. SNP-chips, next-generation sequencing, etc.)
- New analysis methodologies being developed. Need to test them somehow.
- **Usual strategy: simulate test data**
- **Problem**: traditional (coalescent) models are too slow.
- Simulation-based analysis methods (Rejection algorithms, Importance Sampling, ‘no likelihoods’ MCMC -we will cover this letter)

Fast “Coalescent” Simulation

- **Generating test data**:
  - (a) Real data + perturbation (e.g. bootstrap resampling);
  - (b) Model + simulation (e.g. coalescent)

- *Real data + perturbation*:

  Advantage ‘Model’ is correct; we do not know how the data got there, but it used the correct model;

  Disadvantage Subsequent perturbation adds noise.

- *Model + simulation*:

  Advantage Know what you are getting;

  Disadvantage May take a long while to get it; not clear how accurate the model is...

Finding a faster way

- Use an approximation to the coalescent

  Advantage It will be faster

  Disadvantage It is an approximation (to an approximation)
- Example: Wiuf and Hein's "along the chromosome" algorithm
- Remarks: Builds subset of ARG; Slower than Hudson's ms (larger subset), as it includes many recombinations in non-ancestral material; this suggests a simplification and complexity reduction.

'Vanilla' Rejection method

1. Generate θ from prior π.
2. Accept θ with probability $P(D|\theta)$. [Acceptance rate]
3. Return to step 1.

- Set of accepted θ's forms empirical estimate of $f(\theta|D)$
- If upper bound, c, for $P(D|\theta)$ is known replace step 2 with step 2'. Accept θ with probability $P(D|\theta)/c$.
- In general, $P(D|\theta)$ cannot be computed, so...

Alternate rejection method

1. Generate θ from π.
2. Simulate D' using θ.
3. Accept θ if $D' = D$.
4. Return to step 1.

(Likelihood estimation - Diggle and Gratton, J.R.S.S. B, 46:193-227, 1984.)

Note: Probability may be very small; Method is then very inefficient

Rejection method - (approximate Bayesian computation)

- Suppose we have a good summary statistic S.
  1. Generate θ from π.
  2. Simulate D' using θ, and calculate S'.
  3. Accept θ if $S' \approx S$.
  4. Return to step 1.

(1) Result: $f(\theta|S)$ [rather than $f(\theta|D)$].
(2) Best case scenario: S is sufficient

- We know that if the method works, we are getting a good estimate of $f(\theta|S)$
- This makes the assumption that we know how to find a sufficient statistic(s) S that captures the genome structures well:

Issues

1. How to choose S?
2. How close is $f(\theta|S)$ to $f(\theta|D)$?
3. *Lack of theoretical groundwork/guidance*

MCMC - Metropolis-Hastings

- MCMC — Monte Carlo Markov Chain Techniques
  1. If at θ, propose move to θ' according to "transition kernel"

$$q(\theta \mapsto \theta')$$

  2. Calculate

$$h = min\left\{1, \frac{P(D|\theta')\pi(\theta')q(\theta' \mapsto \theta)}{P(D|\theta)\pi(\theta)q(\theta \mapsto \theta')}\right\}$$

  3. Move to θ' with prob. h, else remain at θ
  4. Return to step 1.
- **Result:** $f(\theta|D)$ ((Metropolis et al. 1953, Hastings 1970)

MCMC 'without likelihoods'

- Algorithm
  1. If at θ, propose move to θ' according to "transition kernel"

$$q(\theta \mapsto \theta')$$

  2. Generate data D' using θ'
  3. If $D' = D$ go to step 4; else stay at θ and go to step 1
  4. Calculate

$$h = min\left\{1, \frac{\pi(\theta')q(\theta' \mapsto \theta)}{\pi(\theta)q(\theta \mapsto \theta')}\right\}$$

  5. Move to θ' with prob. h, else remain at θ
  6. Return to step 1.
- **Result:** $f(\theta|D)$

MCMC 'without likelihoods'

- Algorithm
  1. If at θ, propose move to θ' according to "transition kernel"

$$q(\theta \mapsto \theta')$$

  2. Generate data D' using θ', calculate S'
  3. If $S' \approx S$ go to step 4; else stay at θ and go to step 1
  4. Calculate

$$h = min\left\{1, \frac{\pi(\theta')q(\theta' \mapsto \theta)}{\pi(\theta)q(\theta \mapsto \theta')}\right\}$$

  5. Move to θ' with prob. h, else remain at θ
  6. Return to step 1.
- **Result:** $f(\theta|S)$

How to choose “the right” statistics

- Not possible to just include ‘any and all’ statistics; efficiency will degrade
- A new idea motivated by the concept of sufficient statistics.
- If S_1 is sufficient for θ, then:

$$\begin{aligned} P(\theta|S_1) &= P(\theta|D); \\ P(\theta|S_1, S_2) &= P(\theta|S_1), \quad \forall S_2 \end{aligned}$$

With more statistics, the algorithm will be less efficient — lower acceptance rate

Definition

A set of statistics $S_1, S_2, \ldots, S_k$ are ϵ-sufficient statistics relative to a statistic X if

$$\sup_{\theta} \ln P(X|S_1, \ldots, S_{k-1}, \theta) - \sup_{\theta} \ln P(X|S_1, \ldots, S_{k-1}, \theta) \leq \epsilon.$$

Definition

A score of a statistic S_k relative to a set of statistics $S_1, S_2, \ldots, S_{k-1}$ is defined as follows:

$$\delta_k = \sup_{\theta} \ln P(S_k|S_1, \ldots, S_{k-1}, \theta) - \sup_{\theta} \ln P(S_k|S_1, \ldots, S_{k-1}, \theta).$$

Procedure

- Suppose a data-set D and a set of possible statistics S_1, ..., S_M
  1. For $i = 1, \ldots, N$ (N, very large):
     1. Sample θ_i from prior $\pi()$
     2. Simulate data D_i
     3. Calculate $S_{1,i}$, $S_{2,i}$, ..., $S_{M,i}$
  2. Start with no statistics in the rejection; proceed as follows

Algorithm (applied to rejection method)

- Existing posterior, F_{k-1}, using S_1, S_2, ..., S_{k-1}
- Calculate posterior, F_k, after addition of randomly chosen currently unused stat S_k
- If $\|F_k - F_{k-1}\|$ "sufficiently large" add S_k
- Else do not include S_K
- If S_K added, try to remove S_1, ..., S_{k-1}
- Repeat until no statistic can be added

Coalescence

- Coalescent tree provides a method for stochastic simulation (time moving backward).
- Imagine the parameters Θ governing the evolutionary processes are known (e.g., population size [which determines the coalescent-times], mutation rates, etc.; one may add other parameters to the model: recombination and gene conversion rates, population splitting, migration and outbreeding rates, etc)

- Using these parameters Θ one can imagine calculating all possible event waiting times $\mathcal{W}$ from their (exponential) distributions; use the earliest time to chose a particular event and select the participants in that event (e.g., these two individuals *a* and *b* will coalesce at time T_c). Thus one also creates a possible topology: $\mathcal{T}$.
- One can evaluate the fidelity of such a generated tree (although **it sounds rather infeasible**)...

Probability of a Sample Configuration

- Given a set of sequences: Seq_1, Seq_2, ..., Seq_n we want to understand how they have evolved.
- In particular, at a location, we want to know if one can postulate some effect of selection; linkage disequilibrium, etc.
- One idea is to organize the sequences, in terms of a phylogeny with a topology $\mathcal{T}$. But to put them in a phylogenetic tree, we need to know the parameters Θ and the branch lengths (e.g. the waiting times)... But to know the parameters, we need to understand the population structures (e.g., effective population sizes, bottlenecks, migration, etc.)
- **Hopeless Circularity**. We need to proceed in many bay-steps.

Probability of a Sample Configuration

- For any $\langle \mathcal{T}, \mathcal{W} \rangle$, we can compute the following

$$Pr(Seq_1, \ldots, Seq_n | \mathcal{T}, \mathcal{W}, \Theta) = \prod_{j=1}^{L} Pr(Nuc_j | \mathcal{T}, \mathcal{W}, \Theta),$$

  where Seq_i is the i th sequence, all of same length L nts, Nuc_j is the j th column of nucleotides, $\mathcal{T}$ is the topology relating the sequences and $\mathcal{W}$ the set of branch-lengths.
- If the tree is described by the coalescent process the probability of the sample (not conditioned on any particular tree, but just the parameters) is $Pr(Seq_1, \ldots, Seq_n | \mathcal{T}, \mathcal{W}, \Theta)$ integrated over all branch lengths and possible topologies.

- That is, we have

$$Pr(Seq_1, \ldots, Seq_n|\Theta) = \int_{\mathcal{T},\mathcal{W}} Pr(Seq_1, \ldots, Seq_n|\mathcal{T}, \mathcal{W}, \Theta) f(\mathcal{T}, \mathcal{W}|\Theta) d(\mathcal{T}, \mathcal{W}),$$

  where $f(\mathcal{T}, \mathcal{W}|\Theta)$ is the probability density of $\langle \mathcal{T}, \mathcal{W} \rangle$.
- Using the Bayes' formula, we can then compute:

$$\max\{-\ln Pr(\Theta|Seq_1, \ldots, Seq_n)\}$$

[End of Lecture #12]

THE END