Computational Systems Biology: Biology X

Bud Mishra

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

Human Population Genomics

Outline

1. Recapitulation: Coalescence
2. Recombination

“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

2. Recombination

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:

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

Hurdles

- However, there is a serious problem: The integration is over two components: (1) Over all coalescent topologies and (2) over the waiting times corresponding to epochs with different number of ancestors.
- Number of coalescent topologies (closely related to the Catalan number C_n) grows exponentially, and the integral is high dimensional, this computation cannot be done exactly... **Needs many approximation schemes.**

$$C_n = \sum_{i=0}^{n-1} C_i C_{n-i-1} = \left(\frac{1}{n+1}\right)\binom{2n}{n} = \frac{2n!}{(n+1)!n!}$$

Recall: Wright-Fisher Model

- Discrete and non-overlapping generations
- Haploid individuals vs. two subpopulation
- The population size is constant
- All individuals are equally fit
- The population has no geographical or social structure
- The genes do not recombine within the population

Coalescence of a Sample of n Genes

- Recall that

$$Pr(T_k = j) \approx \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{j-1}\binom{k}{2}\frac{1}{2N}.$$

- Thus T_k has approximately a geometric distribution with parameter $\binom{k}{2}/(2N)$. Note that the times $T_2, \cdots, T_n$ are independent.

Continuous Time Approximation

- One unit of time corresponds to the average time for two genes to find a common ancestor: $E(T_2) = 2N$ generations. Time is scaled by a factor of $2N$ (or N or in some cases, $4N$).
- Coalescent becomes independent of the population size. **The structure of the coalescent process is the same for any population as long as the sample size is small relative to population size** $2N$.

$$n \ll 2N.$$

Only the time scale differs between populations when $2N$ varies.

Rescaling Time

- Let $t = j/(2N)$, where j is time measured in generations. $j = 2Nt$. The waiting time, T_k^c, in the continuous representation (for k genes to have $k - 1$ ancestors) is exponentially distributed $T_k^c \sim Exp(\binom{k}{2})$.

$$Pr(T_k^c \leq t) = 1 - e^{-\binom{k}{2}t}.$$

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 $T_k^c \sim Exp(\binom{k}{2})$.
     2. Choose a random pair (i, j) with $1 \leq i < j \leq k$ uniformly from the $\binom{k}{2}$ possible pairs.
     3. Merge i and j into one gene and decrease the sample size by one: $k \mapsto k - 1$.

The Wright-Fisher Model with Mutation

- Impose a process of mutation on top of the process of reproduction.
- Each gene chosen to be passed on is subject to a mutation with probability u.
- If we follow a lineage from the present time to the past, then with probability u the parental gene in generation t differs from the offspring gene at time $t+1$.
- The probability that a lineage experiences the first mutation j generations back is

$$Pr(T_M = j) = u(1-u)^{j-1} \approx \frac{u}{u-1} e^{-uj}.$$

Continuous Approximation

- If time is measured in units of $2N$ generations (like in coalescence) then

  $$Pr(T_M \leq j) = 1 - (1-u)^j \approx 1 - e^{-\theta t/2} = Pr(T_M^c \leq t),$$

  where $t = j/(2N)$, $\theta = 4Nu$ and T_M^c is the time in $2N$ (assumed large) generations units.
- The parameter θ is called the *population mutation rate* or the *scaled mutation rate*.

$n > 2$ Lineages

- Consider n disjoint lineages. The time until the first mutation event in any of the n lineages is exponentially distributed with parameter $n\theta/2$.
- If we wait for mutation events of coalescence events then the parameter of the exponentially distributed waiting time is the sum of the two parameters, which is

$$\binom{n}{2} + \frac{n\theta}{2} = \frac{n(n-1+\theta)}{2}.$$

- Whether the first event is a coalescence or a mutation is determined by a Bernoulli trial:
- With probability

$$\frac{\binom{n}{2}}{\binom{n}{2}+\frac{n\theta}{2}} = \frac{n-1}{n-1+\theta},$$

  the event is a coalescence; and
- With probability

$$\frac{\theta}{n-1+\theta},$$

  it is a mutation.

Stochastic Algorithm to Sample Genealogies with Mutations

- **Algorithm**
  1. Start with $k = n$ genes (sample size). Repeat until $k = 1$:
     1. Simulate the waiting time T_k^c to the next event $T_k^c \sim Exp(k(k-1+\theta)/2)$.
     2. With probability $(k-1)/(k-1+\theta)$ the event is coalescence, and with probability $\theta/(k-1+\theta)$ the event is mutation.
     3. **Case Coalescence**: 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: $k \mapsto k-1$.
     4. **Case Mutation**: Choose a lineage at random to leave. The sample size k remains unchanged.

Outline

1 Recapitulation: Coalescence

2 Recombination

Role of Recombination: An Example

- ApoE Locus: Apolipoprotein E exhibits variation associated with increased risk of Alzheimer's disease. Its gene shows several segregating sites.
- Some haplotypes are much more common than others; thus disagreeing with our basic coalescent model. By just using mutations, we cannot explain why there are many clusters of similar haplotypes, which are different from each other.
- The coalescent tree will need to create deep splits to explain these. But these trees will have topologies that will be very different for different parts of the genome, and will come with very negligible likelihood values.

Role of Recombination

- Genetic recombination creates incompatibilities since it allows different parts of the genome sequence to have topologically different trees.
- LD (Linkage Disequilibrium): LD measure nonrandom association of alleles at different sites. Thus it also measures the correlation among genealogical trees for different segregating sites.
- Recombination can explain observed LD patterns: A weak tendency that highly significant LD is found for sites close to each other. Also LD is smaller further apart the segregating sites are. LD-mapping exploits this effect to find disease causing variant site shares more history with other neutral sites nearby.

Recombination

- Recombination is common in most organisms, but uses different mechanisms for different types of organisms: viruses, bacteria and eucaryotes.
- Biological processes involved: sexual reproduction, gene transfer and template switching.
- However, introducing recombination events into coalescence process poses some difficulties: (a) No single tree can describe a sample of recombining sequences; (b) A directed acyclic graph (**ARG or Ancestral Recombination Graph**) is needed to describe the genealogies; (c) The concepts of a GMRCA (Grand Most Recent Common Ancestor) and other ancestors (not necessarily contributing ancestral material to the extant individuals) must have to be considered.

Recombination in Viruses

- Recombination occurs by **template switching** during the replication process.
- The genomes of a virus is replicated by either a DNA- or and RNA-polymerase depending on the type of virus. The polymerase may jump from one template to a different one.
- For this to occur, the two templates have to be in the same host cell and in close proximity within the cell.
- The recombination rates in virus are determined by the rate of co-infection of the same cell (say in swine) by different variants of the same virus (e.g., human, avian and swine flu viruses).

- The recombination occurs at several points throughout the genome when different templates get close...
- This phenomenon is called **negative inference**: *The probability of recombination occurring at a specific point increases if recombination is occurring near by*.

Flu Virus

- *Evolution and ecology of influenza A viruses.*, by R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers and Y. Kawaoka, MMBR 1992.
- Hypothesis: Aquatic birds are the primordial source of all influenza viruses in other species
- One can do a phylogenetic analysis of the nucleotide sequence of influenza A virus RNA segments. The RNA segments code for the spike proteins (HA, NA, and M2) and the internal proteins (PB2, PB1, PA, NP, M, and NS).
- They come from a wide range of hosts, geographical regions, and influenza A virus subtypes.

- Two partly overlapping reservoirs of influenza A viruses exist in migrating waterfowl and shorebirds throughout the world. These species harbor influenza viruses of all the known HA and NA subtypes.
- Influenza viruses have evolved into a number of host-specific lineages that are exemplified by the NP gene and include equine Prague/56, recent equine strains, classical swine and human strains, H13 gull strains, and all other avian strains. Other genes show similar patterns, but with extensive evidence of genetic reassortment. Geographical as well as host-specific lineages are evident.
- All of the influenza A viruses of mammalian sources originated from the avian gene pool, and it is possible that influenza B viruses also arose from the same source.

- The different virus lineages are predominantly host specific, but there are periodic exchanges of influenza virus genes or whole viruses between species, giving rise to pandemics of disease in humans, lower animals, and birds.
- The influenza viruses circulating in humans and pigs in North America originated by transmission of all genes from the avian reservoir prior to the 1918 Spanish influenza pandemic; some of the genes have subsequently been replaced by others from the influenza gene pool in birds.
- The influenza virus gene pool in aquatic birds of the world is probably perpetuated by low-level transmission within that species throughout the year.

There is evidence that most new human pandemic strains and variants arise as a result of host-switching; often pigs serve as the intermediate host in genetic exchange between influenza viruses in avian and humans.

Recombination in Bacteria

- Recombination occurs by transfer of part of the haploid genome from one cell to a different cell by invasion of a naked DNA molecule: **transfection** or **transduction**.
- If a ssDNA enters the cell it can intrude into the homologous part of the dsDNA of the recipient. Heteroduplex DNA is formed and extends the length of the intruding DNA. Note even if heteroduplex DNA contains mismatches (corresponding to the differences between the donor and the recipient strings), the errors are corrected by the mismatch repair mechanism (involving RecA proteins).
- If the donor string is used as the template, the affected part of donor DNA is effectively recombined.

Recombination in Eukaryotes

- In eucaryotes, accidental breakage of a single strand in one of the two homologous chromosomes leads to invasion of the broken strand into the other chromosome (dsDNA). This leads to a region of a heteroduplex DNA.
- The heteroduplex regions are assumed to be many kilobases. The mismatches in the heteroduplex are recognized and fixed according to the dsDNA template later by the DNA repair system.
- True recombination occurs when both strands break simultaneously or the flanking chromosomal parts are exchanged after strand invasion.

- **Recombination Module**: Needed for the formation and resolution of Holliday junction.
- What daughter molecules get created depend on many factors: (1) the formation of Holliday structure, (2) the creation of heteroduplex, (3) the potential mismatch repair mechanism, (4) Random-walk and absorption of the heterduplex, (5) the timing of the subsequent DNA replication, etc.
- The result is either a **recombination event** (gene conversion event with crossing over) or a **gene conversion event** (gene conversion event without crossing over).

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.

- Assuming k sequences:
  - the time to coalescent even is distributed exponentially with parameter $k(k-1)/2$,
  - the time to recombination event exponentially distributed with parameter $k\rho/2$ and
  - the two events are independent.
- The algorithm thus chooses the waiting time for an event (recombination or coalescent) from an exponential distribution with parameter

$$\binom{k}{2} + \frac{\rho k}{2} = \frac{k}{2}(k-1+\rho)$$

Stochastic Algorithm to Sample Genealogies for *n* Genes

- Next, it resolves the event identity by choosing coalescent with probability
$$\frac{k-1}{k-1+\rho}$$
and by choosing recombination with probability
$$\frac{\rho}{k-1+\rho}.$$
- In case of recombination, the number of ancestral sequences increase by one (from k to $k+1$); in case of coalescent, the number of ancestral sequences decreases by one (from k to $k-1$).

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 $1 \le i < j \le k$ uniformly from the $\binom{k}{2}$ possible pairs. Merge i and j into one gene and decrease the sample size by one: $k \mapsto k-1$.

[End of Lecture #11]

THE END