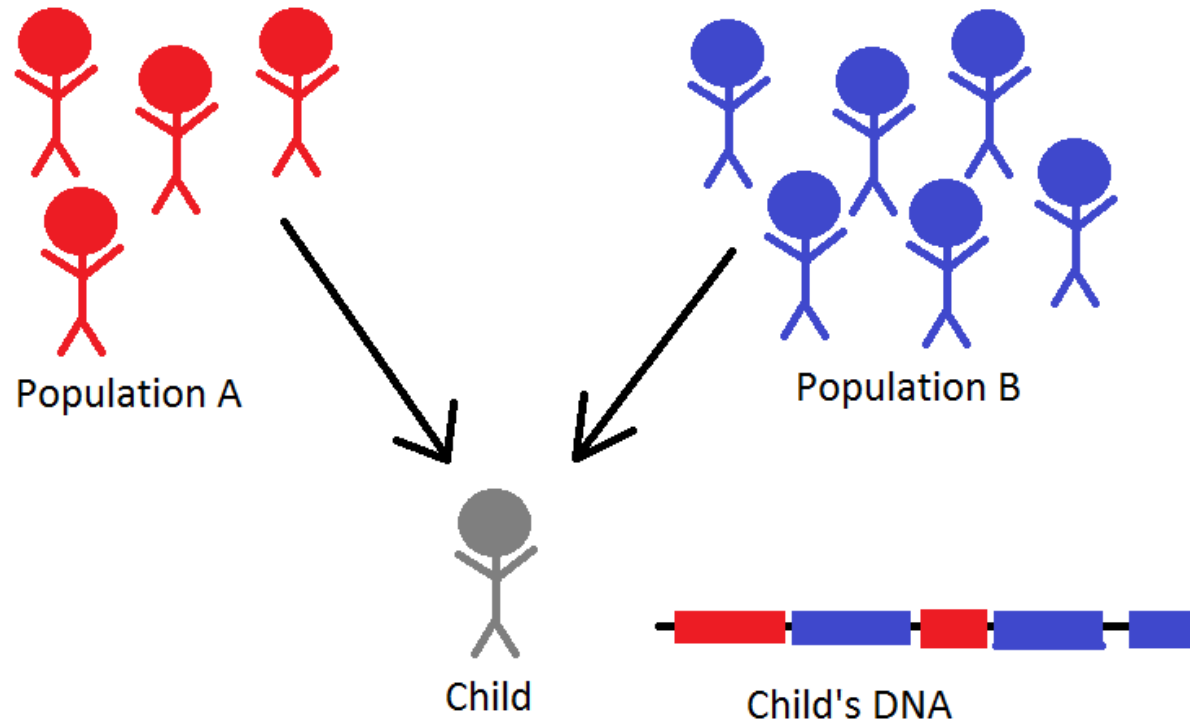


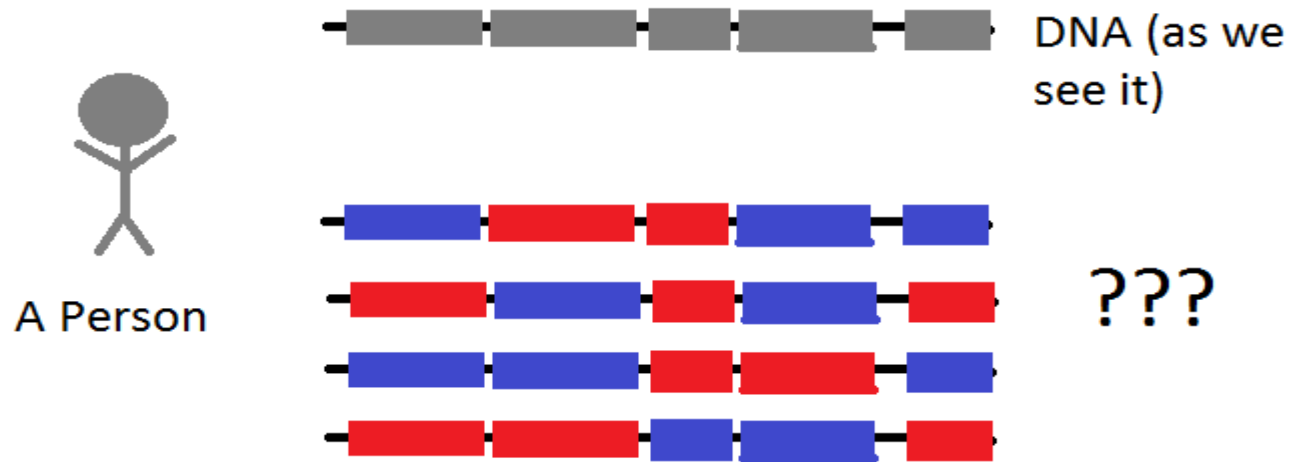
Statistical Algorithms in Population Genetics

Adam Pershan

The General Story



Goal: to infer local ancestry



Which one (if any) is correct?

Why do we care?

- Interesting for studying human populations and evolution
- Interesting for studying other organisms and their evolutionary tracks
- Clinical ramifications for disease tracking: for example, Native American ancestry has been associated with an increased risk of relapse in children with some types of leukemia

What type of program are we looking for?

- Fast
- Accurate
- Flexible

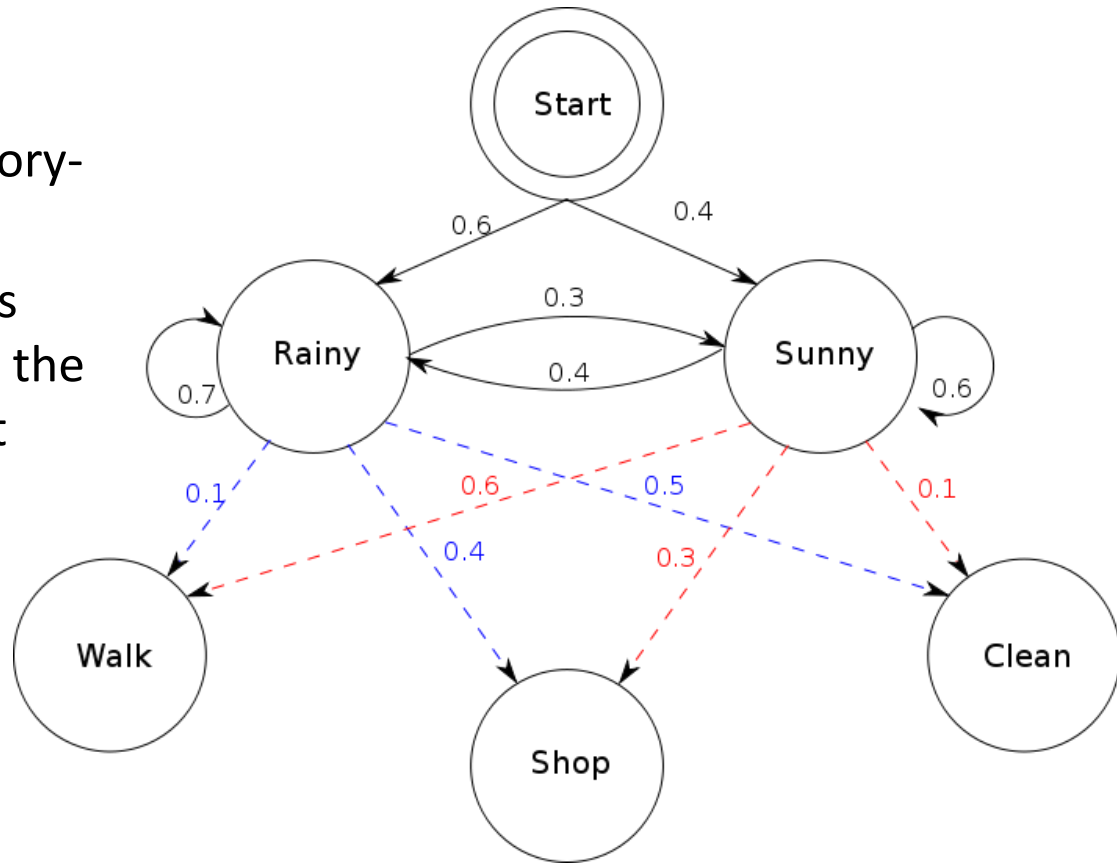
Method	Information from ancestral populations	Running time ($N = 200,1$ CPU)	Handles multi-way admixed populations	Simulations of African Americans		
				r^2	Haploid accuracy (%)	Diploid accuracy (%)
ANCESTRYMAP	Allele frequencies at a set of AIMs	2 min	No	0.83	97.67	95.35
WINPOP	Allele frequencies at GWAS SNPs	3 min	Yes	0.94	99.36	98.75
HAPMIX	Phased haplotypes at GWAS SNPs	7 h	No	0.97	99.66	99.32

CPU, central processing unit.

Solutions

- Many use Hidden Markov Models (HMM)
- A Markov Process is “memory-less”
- A Hidden Markov process is one where you don’t know the states directly, and you just observe the events

Example of HMM

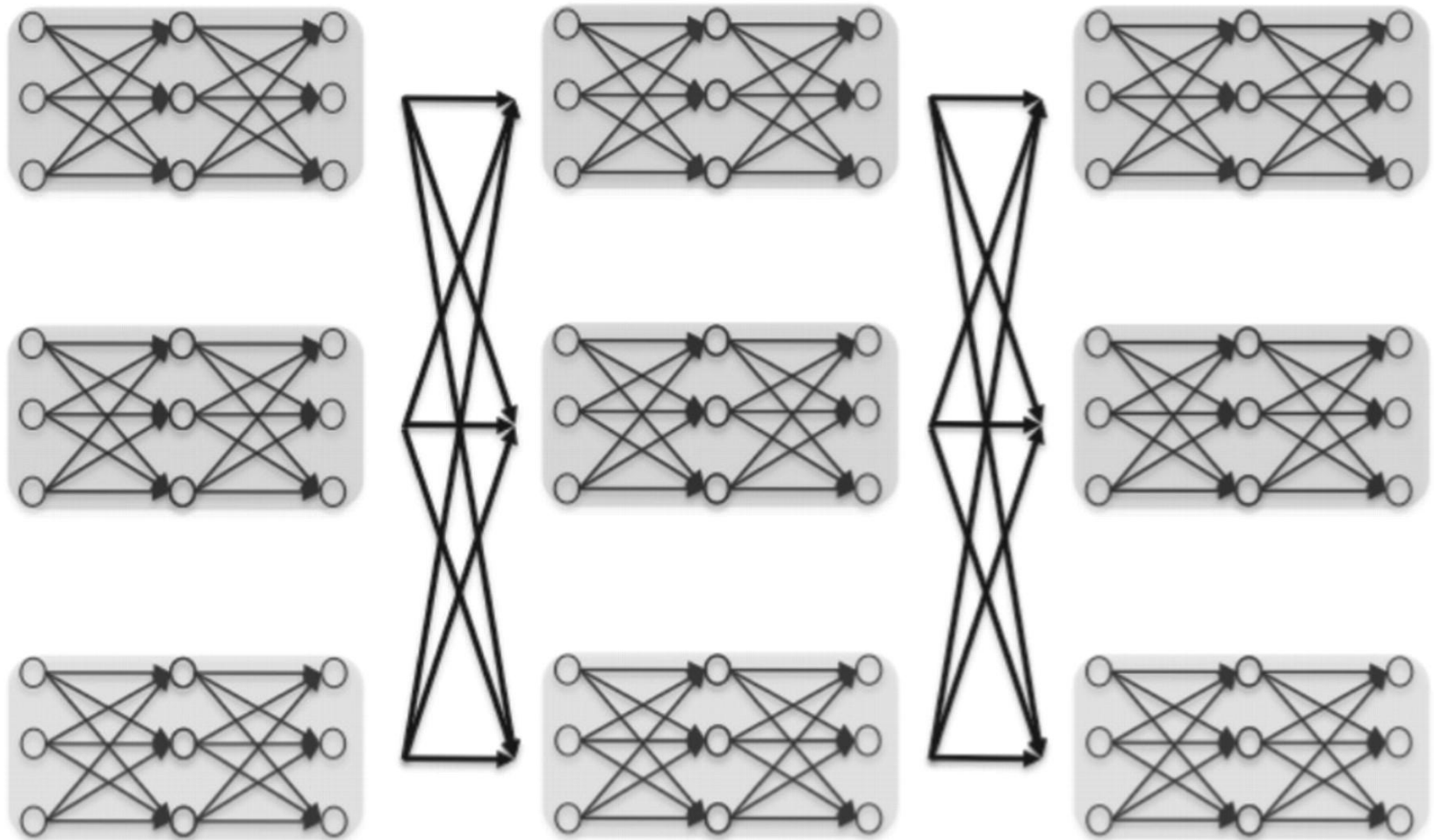


$$P(O|\lambda) = \sum_Q P(O|Q, \lambda) P(Q|\lambda) = \sum_{q_1 \dots q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

HMM in population genetics

- Ancestry is the hidden state
- Work backwards from “child” genome to infer which state most likely led to the observed data at a specific locus
- Requires some information about the ancestor populations, approximated from present-day populations.

Schematic structure of a two leveled HMM-based algorithm



Baran Y et al. *Bioinformatics* 2012;28:1359-1367

My goals for the summer

- Dr Chen's lab works with different species of yeast, and I will be tracing their local ancestry
- I hope to be able to compare the different algorithms that exist and see how they perform with this non-human data

My goals for the summer

- There are several differences between admixture in yeasts and admixture in humans, such as more parent-populations
- If I end up having any ideas for improving any of the existing models for the specifics of this case, that would be pretty cool