

# URN MODELS in GENETICS: Wright-Fisher Models and the Ewens Sampling Lemma

Sorin Istrail

Department of Computer Science

Brown University, Providence

`sorin@cs.brown.edu`

October 23, 2012

# Outline

- 1 Outline
- 2 Urn Models
- 3 Wright-Fisher Urn Model
- 4 Mutation, Random Genetic Drift and Selection
  - Mutation
    - Mutation: typical values for parameters
    - Equilibrium
  - Random Genetic Drift
    - Probability of fixation
  - Selection
- 5 Wright-Fisher Model and Markov chains
- 6 The Infinite Allele Model and the Ewens Sampling Lemma
  - Ewens Sampling Lemma - first formulas

# Urn Models

- We will use Urn Models that capture the essence of the **Wright-Fisher Model** for a Mendelian theory of evolution. They are **Polya Urn Models**
- In the standard Polya urn schemes there is a fixed set of admissible colors that can appear in the urn. A new species (new color) may appear at some point. In such a model we must allow the number of colors to increase with time, and consider an infinite set of admissible colors (**Infinite Allele Model**).

# Wright-Fisher Urn Model

- The hereditary process can be represented by balls in an urn. The balls undergo a rebirth process and produce balls in subsequent urn.
- Let us start with two colors representing two genes.
- The model was introduced by Fisher in 1922 and by Wright in 1930.
- Suppose that there are  $m$  balls in the urn, of which  $i$  are white (a first gene) and  $m - i$  are blue (a second different gene).
- Balls are sampled with replacement,  $m$  times, to give a chance for each ball to appear once on average.
- If a sampled ball is white, we deposit a white ball in the new urn; if the sampled ball is blue we put a blue ball in the new urn.

# Wright-Fisher Urn Model

- The number of white balls that appear in the size  $m$  sample is  $\text{Bin}(m, \frac{1}{m})$  distributed.
- On average, there are  $m \times \frac{i}{m} = i$  white balls in the new urn, and on average there is no change in the proportion of the white balls.
- After the new urn has been filled with  $m$  balls, a new urn is created next, and so on.
- There is a positive probability that the urn becomes monochromatic. Once this is achieved the urn remains the same forever.
- This can be seen by modeling the process by a **Markov model**.

# Mutation, Random Genetic Drift and Selection

- Mutant alleles that have little effect on the phenotype of the organism may remain in the population until either became fixed or lost due to stochastic forces.
- The simplest mathematical model is **Random Mating Model** in monoecious population. Changes in allele frequencies are caused by mutation, random genetic drift, and selection.

- Consider a locus with two alleles  $A_1$  and  $A_2$  and let  $p(n)$  and  $q(n) = 1 - p(n)$  the frequencies of  $A_1$  and  $A_2$ , respectively in generation  $n$ . Suppose non-overlapping generations. And suppose that  $u$  is the rate of  $A_1$  mutating to  $A_2$  and  $v$  the rate of  $A_2$  mutating to  $A_1$ . Alleles can mutate only once per generation. Assuming no other forces but mutation, we have the following recurrence:



$$p(n+1) = (1-u)p(n) + v(1-p(n))$$

- An  $A_1$  allele in generation  $n + 1$  could have been either an allele  $A_1$  in generation  $n$  that did not mutate, this with probability  $(1 - u)$  or an allele  $A_2$  that mutated into  $A_1$  with probability  $v$ .
- The above recurrence equation can be solved in terms of the initial frequency of  $A_1$ , say,  $p(0)$ :

$$p(n) = \frac{v}{u + v} + (p(0) - \frac{v}{u + v})(1 - u - v)^n$$



# Mutation: typical values for parameters

- The mutation probabilities  $u$  and  $v$  are typically quite small or the order of  $10^{-6}$ .
- For large  $n$  the term  $(1 - u - v)^n$  goes to 0. E.g., for  $n = 10^6$  and  $u = v = 10^{-6}$  the terms  $(1 - u - v)^n = 0.13$  and for  $n = 10^7$  it gets to  $(1 - u - v)^n = 2.06 \times 10^{-9}$ .
- So mutation alone can be slow changing frequencies.  
Although mutations are the source of variation, their role in evolution may be limited.

# Equilibrium

- In the limit, assuming the population is infinite,  $p(n)$  becomes  $\frac{v}{u+v}$ . This value is called **equilibrium**, it happens when  $p(n+1) - p(n) = 0$  which means that the allele frequencies do not change over time.
- If we denote the equilibrium frequency by  $\hat{p} = p(n) = p(n+1)$  and solve for  $\hat{p}$  we get  $\hat{p} = \frac{v}{u+v}$  and  $\hat{q} = 1 - \hat{p} = \frac{u}{u+v}$ .
- If both  $u$  and  $v$  are positive, it follows that **mutation allows for the maintenance of the two alleles in the population.**

# Random Genetic Drift

- The assumption of an infinite population allowed us to use the deterministic formulation when we modeled mutation.
- In **finite populations** the random sampling of gametes alone causes changes in gene frequencies. This process is known as **random genetic drift**.

# Random Genetic Drift

- Let us look at a single locus with two alleles  $A_1$  and  $A_2$ .
- Assume a randomly mating diploid population of size  $N$  (or, which is the same in this case, a haploid population of size  $2N$ ) with non overlapping generations.

# Wright-Fisher Model

- In each generation,  $2N$  gametes are sampled at random from the parent generation. If the  $Y(n)$  denotes the number of gametes of type  $A_1$  at generation  $n$  then in the absence of mutation and selection, the number of  $A_1$  alleles at time  $n + 1$  is given by a binomial distribution. This is the **Wright-Fisher Model**.
- Namely, the probability that there are  $j$  gametes of type  $A_1$  at generation  $n + 1$ , given that there were  $i$  gametes of type  $A_1$  at generation  $n$  is

$$P(Y(n+1) = j \mid Y(n) = i) = \binom{2N}{j} p^j (1-p)^{2N-j}$$

where  $p = \frac{i}{2N}$

# Evolution is quite unpredictable due to stochastic nature of the model

- The discrete time stochastic process is an example of Markov chain
- Since there are no mutations in the model, eventually one of the two alleles will be lost (and the other fixed).
- The larger the population size, the longer this process of fixation takes.

# Probability of fixation is the initial probability

- If the frequency of allele  $A_1$  was  $\pi_1$ , then one can show that the probability of fixation of allele  $A_1$  is  $\pi_1$ , and consequently, the probability of fixation of  $A_2$  is  $1 - \pi_1$ .
- Ewens's argument: After a long enough time, all individuals in the population must have descended from just one of the individuals present at generation 0. The probability that this common ancestor was of type  $A_1$  is equal to the relative frequency of  $A_1$  at generation 0 which is  $p_{i_1}$ .

# Selection

- Selection can act on different parts of life history of an organism: differential fecundity and viability are such examples. The simplest model of viability selection assumes that that selection affects survival between the zygote and adult stage of a diploid organism in a randomly mating population of infinite size, in which generations do not overlap.
- It is assumed that each genotype has a fixed specified **fitness**. Since the population is infinite changes in allele frequencies can be described by deterministic equations.
- For a site with two alleles  $A_1$  and  $A_2$  there are three genotypes:  $A_1A_1$ ,  $A_1A_2$ ,  $A_2A_2$ . We denote their fitnesses by  $w_{11}$ ,  $w_{12}$ ,  $w_{21}$ . In the case of viability selection the fitness  $w_{ij}$  reflects the relative survival of zygotes of genotype  $A_iA_j$ .



# Fitness

- If the population is in Hardy-Weinberg equilibrium and the frequencies of  $A_1$  and  $A_2$  are  $p(n)$  and  $q(n) = 1 - p(n)$  respectively, then ignoring mutations then the next generation allele frequencies are

$$p(n+1) = \frac{p(n)(p(n)w_{11} + q(n)w_{12})}{\bar{w}}$$

and

$$q(n+1) = \frac{q(n)(q(n)w_{22} + p(n)w_{12})}{\bar{w}}$$

and  $\bar{w}$  is defined by

$$p(n+1) + q(n+1) = 1$$

It follows that

$$\bar{w} = p^2(n)w_{11} + 2p(n)q(n)w_{12} + q^2(n)w_{22}$$

# Model Predictions

- The predictions are as follows
- *Directional Selection*:
  - If  $w_{11} > w_{12} > w_{22}$  then  $A_1$  becomes **fixed in the population**.
  - If  $w_{11} < w_{12} < w_{22}$  then  $A_2$  becomes **fixed in the population**.
- *Overdominance*: If  $w_{11}, w_{22} < w_{12}$  then a **stable polymorphism** results.
- *Underdominance*: If  $w_{12} < w_{11}, w_{22}$  then the **polymorphism is unstable** and depending of the initial allele frequencies either  $A_1$  or  $A_2$  become fixed.

# Markov chain

- Suppose we have only two balls in the urn initially. One white and one blue.
- We represent the evolution by the states of a **Markov chain** with three states, where a **state** is described by the number of white balls in it. So we have the states  $S_0, S_1, S_2$ .
- We choose to start in state  $S_1$ . From state  $S_1$  we can move into state  $S_0$ , if no white balls appear in the next urn; from state  $S_1$  it can remain in state  $S_1$  if the next urn has the one white ball; or from state  $S_1$  it can move to state  $S_2$  if the next urn has two white balls. The corresponding probabilities are  $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ . These numbers are  $P(X = k)$  for  $k = 0, 1, 2$  for a Binomial random variable distributed line  $Bin(2, \frac{1}{2})$ . These three numbers correspond to the row 1 of a Markov chain transition matrix  $M$ .

# Markov chain



$$M = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}$$

- In  $n$  steps, we obtain

$$M^n = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2^n-1}{2^{n+1}} & \frac{2}{2^{n+1}} & \frac{2^n-1}{2^{n+1}} \\ 0 & 0 & 1 \end{pmatrix}$$

- Let us start in state  $(S_1, \text{i.e., the state with number of white balls} = 1)$ .
- It follows that after  $n$  steps the probability state vector is:

$$\begin{aligned}\pi_n &= (0, 1, 0)M^n = \left(\frac{2^n - 1}{2^{n+1}}, \frac{2}{2^{n+1}}, \frac{2^n - 1}{2^{n+1}}\right) \\ &= \left(\frac{1}{2}, 0, \frac{1}{2}\right) + \left(-\frac{1}{2^{n+1}}, \frac{1}{2^n}, -\frac{1}{2^{n+1}}\right) \rightarrow \left(\frac{1}{2}, 0, \frac{1}{2}\right)\end{aligned}$$

- this vector converges to  $(\frac{1}{2}, 0, \frac{1}{2})$
- The uniformity of color comes randomly at an exponential rate. Rather quickly, the urn is in either a state of all white or all blue. Ultimately, there is 0 probability of a split. And probability is  $\frac{1}{2}$  of an ultimate all white urn. And same for an all blue urn.

# Mutation - always a new color

- In urns balls of new colors occur by **mutation**. So through the stochastic process, old colors may disappear and new colors may appear. So the concept of stationarity in the Markov chain context is no longer available.
- Alternatively we can think of the **stationarity of partitions** among the existing and ever changing colors.
- A partition of  $m$  species (genes) describes how many colors there are that are presented by  $i$  balls, for various values of  $i$ . For example if the number of balls is 4 and there are one black, two red and two green and four blue balls the partition would be  $(1, 2, 0, 1)$ .  $m = 1 \times 1 + 2 \times 2 + 0 \times 3 + 1 \times 4$ .
- If in time one we observe that there are in the urn, one black, two pink and two yellow, and four red balls the partition is again  $(1, 2, 0, 1)$ .

# Mutation - Balls new colors

- Suppose we are drawing  $m$  balls with replacement from an urn, and we are filling a new urn with balls of colors determined by the sample.
- When we draw a ball of a given color from the current urn we return it in the ball and we placed a ball with the same color in the new urn with probability  $1 - \alpha$  and a new color with probability  $\alpha$ . Mutation always creates a new color.

## stationary partition assumption

- Consider the stationary probability  $Q_2$  of two balls of the same color in the new urn.
- We have two situations to consider:
  - 1 the two balls arrived by drawing the same ball twice in the current urn assumed in stationary state and no mutation occurred;
  - 2 the two balls arrived by drawing two balls of same color in the current urn, assumed stationary and no mutation occurred.



## Computing $Q_2$

- The probability of choosing the same ball (the same parent) is

$$\sum_{k=1}^m \frac{1}{m^2} = \frac{1}{m}$$

- The probability of being children of two different balls of same color chosen with probability  $Q_2$  from the stationary parent urn is computed as follows.
- First, the probability of choosing two different balls is  $1 - \frac{1}{m}$ .
- Two balls of the same color in a stationary urn cannot be mutations from the parent urn, because necessarily all mutations are of different colors.
- The stationary probability satisfies the equation:

$$Q_2 = \frac{1}{m}(1 - \alpha)^2 + (1 - \frac{1}{m})Q_2(1 - \alpha)^2$$

## Computing $Q_2$

- The solution of the above equation is

$$Q_2 = \frac{(1 - \alpha)^2}{(1 - \alpha)^2 + m\alpha(2 - \alpha)}$$

- as  $\alpha = 10^{-5}$  typically we obtain a good approximation;

- 

$$Q_2 = \frac{1}{1 + 2m\alpha} = \frac{1}{1 + \theta}$$

- where

$$\theta = 2m\alpha$$

## Formulas for $Q_3, Q_4, \dots, Q_i$



$$Q_3 = \frac{2}{(\theta + 1)(\theta + 2)}$$



$$Q_4 = \frac{6}{(\theta + 1)(\theta + 2)(\theta + 3)}$$



$$Q_i = \frac{(i - 1)^2}{(\theta + 1)(\theta + 2) \dots (\theta + i - 1)}$$

# The Ewens Sampling Lemma

## Theorem (Ewens 1972)

Let  $\mathbf{A}_n = (A_1, A_2, \dots, A_n)$ .  $A_i$  is the number of colors represented by exactly  $i$  balls in the sample. The size of the sample is  $m = \sum_i^n A_i$ . Then:

$$P(\mathbf{A}_n = \mathbf{a}_n) = \frac{n! \theta^{(\sum_i^n a_i)}}{\prod_i^n i^{a_i} \prod_i^n a_i! \prod_{i=0}^{n-1} \theta + i}$$