Toy models in Population Genetics: some mathematical aspects of evolution

David Aldous

April 6, 2016

Some probability models of real-world phenomena are “quantitative” in the sense that we believe the numerical values output by the model will be approximately correct. At the other extreme, a **toy model** is a consciously over-simplified model of some real-world phenomenon that typically attempts to study the effect of only one or two of the factors involved while ignoring many complicating real-world factors. It is thus “qualitative” in the sense that we do not believe that numerical outputs will be accurate.

As our first examples will show, providing a toy model to support a scientific theory shows that the theory is at least **possibly** correct; whereas if you are unable to provide a supporting toy model then the theory looks dubious.

Today's topics

- Is "evolution by natural selection" mathematically plausible?
- What explains the shape of evolutionary trees?
- What maintains genetic diversity within species?
- Are you related to your ancestors?

What was Charles Darwin's contribution to science? In everyday language we say "the theory of evolution" but this isn't quite right. By the middle of the nineteenth century, once dinosaur and other fossils were being discovered, the proposition that life on Earth has been in existence for a very long time, that earlier species had become extinct and that other species had originated – this proposed **fact** wasn't particularly controversial.

Consider an analogy between

- Empires in human history.
- Species, in the history of life on Earth.

Wikipedia has a list of almost 200 empires, almost all of which no longer exist; the **fact** that empires have risen and fallen was never controversial.

At one level, everything happened for some specific reason – why the Inca Empire or *Tyrannosaurus Rex* are no longer here. But is there any underlying general principle?

Ever since the first historian wrote, many general explanations for the rise and fall of empires have been proposed – divine favor, racial superiority, class struggle, technological superiority, societal ethics, ecological collapse – but none is widely accepted, and indeed are generally taken to reflect prejudices of the era when they were formulated.

In contrast, Darwin's idea of "evolution **by natural selection**" was that there is one underlying *explanation* of this process – natural selection. Darwin and his nineteenth century followers did not have our current notion of genetics and did not seek a mathematical formulation of their theory. And indeed they were aware that there was a difficulty with the whole idea, if approached from a certain common sense view of heredity ("paint mixing", below). Let me first describe the difficulty, and then show how it is resolved in the correct theory of genetics.

If heredity were like paint mixing. Observation of animal breeding might suggest offspring are a mixture of parents, like a mixture of blue and yellow paint makes green paint. Of course this couldn't be the whole story, or every individual in a population would be identical by heredity, but (unaware of genetics) we might imagine heredity working as "mixture of parents, plus individual randomness". And indeed this kind of "additive" model does correctly predict the behavior of some real-world quantitative characteristics, for instance height in humans.

However, let us consider a model for how natural selection might work on a novel hereditable trait, **if** heredity were like paint mixing. We'll give a model that ignores randomness (both in number of offspring and assumed "individual randomness"), but incorporating randomness doesn't change the conclusions.

A paint mixing model. One individual (in generation 0, say) has a new characteristic giving selective advantage α, meaning that the mean number of offspring reaching maturity is $2(1+\alpha)$ instead of 2. Each offspring (generation 1) has only half of the characteristic (this is the "like paint mixing" assumption), so has selective advantage $\alpha/2$, so each generation 1 offspring has mean number $2(1+\frac{1}{2}\alpha)$ offspring in generation 2, and these generation 2 individuals have a quarter of the characteristic. So the "penetration" (sum over individuals of their proportion of the characteristic) of the characteristic in successive generations is

generation	0	1	2
mean number individuals	1	$2(1+\alpha)$	$4(1+\alpha)(1+\frac{1}{2}\alpha)$
proportion of characteristic	1	$\frac{1}{2}$	$\frac{1}{4}$
penetration	1	$1+\alpha$	$(1+\alpha)(1+\frac{1}{2}\alpha)$

As time passes the mean penetration increases, not indefinitely but only to a finite limit

$$\beta(\alpha) = \prod_{i=0}^{\infty}(1 + 2^{-i}\alpha)$$

which for small α is approximately $1 + 2\alpha$. This value doesn't depend on the population size (N, say). So the key conclusion is that the effect of a single appearance of a new characteristic would be, after many generations, that each individual in the population gets a proportion around $(1 + 2\alpha)/N$ of the characteristic.

This conclusion is bad news for a theory of natural selection, because it implies that to become "fixed" in a population, a new characteristic would have to reappear many times – order N times – even when it provides a selective advantage.

The genetic model. How does genetics really work? Here is a (very) toy model. We consider genes (physically, a small segment of a chromosome) rather than individuals, so there are $2N$ genes in each generation. On average, a gene has 1 copy in the next generation, with some s.d. ($= \sigma$, say). For a new allele (the *alleles* are the possible forms of a given gene) which confers a small selective advantage, we suppose the average number of copies becomes $\mu = 1 + \alpha$ for some small $\alpha > 0$. Note this can only be true while the number of copies is small relative to the population, and during that time the number of copies in successive generations behaves as a just supercritical Galton-Watson process described in previous lecture.

In particular, either the new allele disappears from the population quite quickly (extinction, in the Galton-Watson terminology) or the number of copies starts to grow exponentially; then (as in the epidemic model in previous lecture) the proportion of this new allele in the population grows as an S-shaped curve and eventually the allele becomes *fixed* in the population – every gene is this allele.

The mathematical point is that the earlier formula for survival probability of just supercritical Galton-Watson processes can be applied in the present model.

For a single mutation giving an allele with small selective advantage α,

$$\text{the chance that the allele becomes } \textbf{fixed} \text{ is about } \frac{2\alpha}{\sigma^2}. \tag{1}$$

This conclusion is much better news for a theory of natural selection, because now the population size doesn't matter. If the chance above were $1/10$, say, then an advantageous mutation needs to reappear only 10 or 20 times to be likely to become "fixed" in the population, regardless of how large the population size N is.

The whole process of an allele becoming fixed in this way is called a **selective sweep**. Once a sweep is under way, the number of copies grows at rate α per generation, and so

$$\text{duration of a selective sweep} \approx \frac{\log(2N)}{\alpha} \text{ generations .} \qquad (2)$$

So our "toy model" of heredity (which Mendel guessed and was confirmed round 1900) shows that "evolution by natural selection" is at least possible, mathematically. The key point is that genes are discrete entities – metaphorically, heredity is digital not analog.

(Continuous phenotypes like height are affected by many genes – envisage something like a CLT).

A conceptual point is to distinguish between what we have just discussed

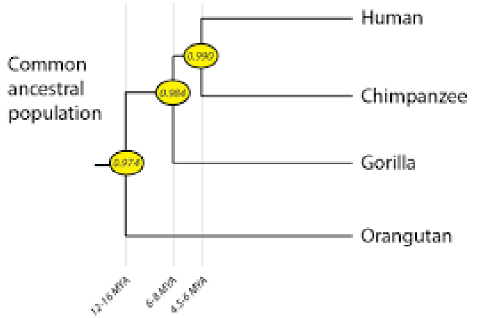
• **microevolution** at the level of **genes** (allele frequencies) within a given species

and what "evolution" means in popular language

• **macroevolution** at the level of **species**.

The evolutionary relationships between species are described graphically via **phylogenetic trees**, and these provide interesting examples of statistical data.

Historically, biologists first did **classification** of species based on **morphology** – the physical structure – of living species. Next they sought to fit extinct species (based on fossils) into the evolutionary tree. Finally DNA provides a quantitative measure of similarity between living species.



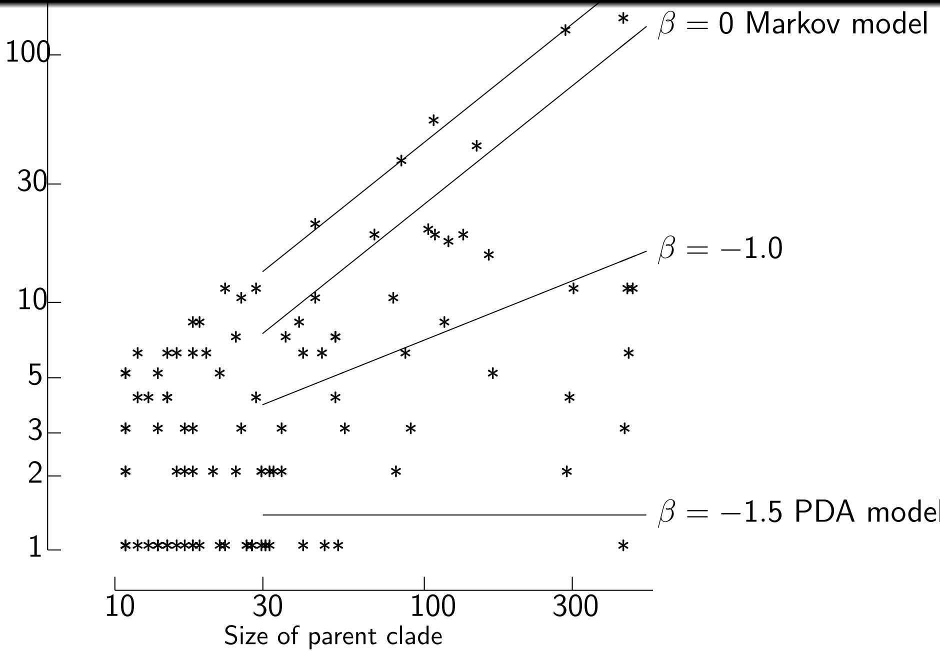
[show hominids]

[show parrots]

[show horse and dinosaur trees]

One can make probability models for macroevolution – see paper *Toy models for macroevolutionary patterns and trends* – but these are "made up" without reference to actual biology. The simplest models just assume there is some chance a species will go extinct and some chance it will produce a daughter species, giving a continuous-time analog of the Galton-Watson process.

A puzzle concerns the "shape" of phylogenetic trees. The data does not match the simple models! The next figure is from my paper *Stochastic Models and Descriptive Statistics for Phylogenetic Trees*, showing the scatter diagram for the splits in a phylogenetic tree of 475 species of seed plants.
[board]


100
30
10
5
3
2
1
10
30
100
300
Size of parent clade
$\beta = 0$ Markov model
$\beta = -1.0$
$\beta = -1.5$ PDA model

Today's topics

- Is "evolution by natural selection" mathematically plausible? YES
- What explains the shape of evolutionary trees? MYSTERIOUS
- What maintains genetic diversity within species?
- Are you related to your ancestors?

Probability models of **micro**evolution refer to more concrete entities – alleles and mutation and "fitness" as defined by number of offspring – even though they then make the unrealistic assumption that one can separate the effect of one gene from all the other factors affecting fitness.

The "gene-centered view of evolution" – that it makes sense to think in terms of genes without regard to species – was popularized by Richard Dawkins in *The Selfish Gene*.

Why test human medicine on mice?

What maintains genetic diversity? We have an everyday notion of species – rabbits and robins and roses – because we can recognize different individuals as similar. After learning about genetics and evolution by natural selection, a rather subtle question arises: why is there any genetic difference at all between different individuals in the same species? That is, if "evolution by natural selection" worked according the the simple "selective sweeps by more fit alleles" story, then why hasn't it already happened, so that all the less fit alleles have been replaced by the most fit allele, leading to individuals genetically identical except for sex-related traits?

Many different answers have been proposed, and undoubtedly many are valid in different contexts. A textbook answer is *heterozygote advantage*i, illustrated by sickle cell anemia in humans. (Given that most phenotype variations presumably arise from complicated interactions between genes, there is much scope for this kind of effect). Another answer is *frequency dependent selection*, where it is advantageous to be different from others, in contexts of predation or competition. Another answer is that we may just be seeing a selective sweep in progress, though the toy model prediction (2) for sweep duration suggests this is unlikely.

[board: time scale of evolution]

We will consider the **neutral theory** which asserts that much of the variation we **see** in a species at a particular time is “non-selective”; different alleles have arisen by chance mutations some time in the past but have almost zero difference in fitness, implying that the frequencies of alleles in successive generations change only in some “random” way rather than being pushed in one direction by selection.

Rather obviously this appeals to mathematical probabilists, so let me show some predictions within this theory.

> *Consider a gene with several alleles $A, B, C \ldots$. In diploid populations consisting of N individuals in each generation there are 2N copies of each gene. An individual can have two copies of the same allele or two different alleles. Assume generations do not overlap. For example, annual plants have exactly one generation per year. In the model, each copy of the gene found in the new generation is drawn independently at random from all copies of the gene in the old generation.*

This is "the **Wright-Fisher model** without mutation or selection", edited from Wikipedia *Genetic Drift*.

The model looks strange as biology, but turns out to be mathematically tractable, and behaves similarly to more plausible models in which parents have offspring independently while some external mechanism keeps the population size roughly stable.

If we never had mutations then eventually the random fluctuations of allele frequency from generation to generation would make all but one allele die out.

We now introduce *mutation* by supposing that, each time a gene is copied, there is a small chance p of a mutation, and that each such mutation produces a brand new allele. The process of “numbers of alleles of different types” is abstractly a certain complicated finite-state Markov chain.

From the theory of Markov chains there must be a stationary distribution for the proportions $X_1 \geq X_2 \geq X_3 \ldots$ of different alleles, listed in decreasing order for definiteness. The remarkable *Ewens's sampling formula* gives the exact distribution of the (X_i), but instead let me derive a simpler statistical measure of diversity.

Consider $S := \sum_i X_i^2$ and note that $\mathbb{E}S$ is the chance that two randomly-picked genes are the same allele; then we can view

$$n_{eff} := 1/\mathbb{E}S$$

as "effective number of different co-existing alleles of the gene" in the population.

[This is an idea we saw in a previous lecture: next slide]

What do these particular statistics $\sum_s p_s^2$ and $-\sum_s p_s \log p_s$ measure?

[board]: spectrum from uniform distribution to deterministic.
Interpret as "amount of randomness" or "amount of non-uniformity".

First statistic has no standard name.
Second statistic: everyone calls it the *entropy* of the probability distribution $\mathbf{p} = (p_s, s \in S)$.

For either statistic, a good way to interpret the numerical value is as an "effective number" N_{eff} – the number such that the uniform distribution on N_{eff} categories has the same statistic.

For many purposes the first statistic is most natural – e.g. the chance two random babies born in 2013 are given the same name.

We shall derive the formula (in the **neutral theory**)

$$n_{eff} \approx 1 + 4Np. \tag{3}$$

The approximation holds in the (realistic) case where N is large and p is small, and we think of $4Np$ as a number – maybe 0.2, maybe 10 – that is neither very large nor very small.

Discussion of formula (3). This confirms and quantifies the idea that pure randomness (mutations without selective advantage) can maintain a fixed level of diversity as time goes by; so the neutral theory is at least a possible explanation of diversity. But how realistic is the model?

What you don't see in the model description or the concluding formula, but is buried in the derivation, is the requirement that the model must have been realistic over the last (order) N generations. Thinking of total species population N as in the millions, this is hardly plausible, since the time involved would become larger than species lifetime. Also, the model implicitly ignores geographic location of individuals – any pair can breed – and it is often argued that what is relevant is a much smaller "effective population size" of interbreeding subpopulations.

Mathematical derivation of formula (3). The key feature that makes the model mathematically tractable is that we can easily study genealogy. Ignoring mutations for the moment, a gene in the present generation is a copy of a gene in the previous generation, which is a copy of a gene in the previous generation, and so on: there is a "line of descent". Now consider two randomly-picked genes in the present generation, and trace back the two lines of descent until they meet, some random number G of generations back, at their "most recent common ancestor". From the definition of the model, at each stage there is chance $1/(2N)$ that the lines merge, and so G has the Geometric$(1/(2N))$ distribution.
[board]
Introducing mutations (without selection) doesn't change the behavior of lines of descent. Given $G = g$, the two sampled genes have the same allelic type if and only if none of the $2g$ copies since the most recent common ancestor caused a mutation, so

$$\mathbb{P}(\text{Two sample genes are the same type}|G = g) = (1 - p)^{2g}.$$

$$\mathbb{P}(\text{Two sample genes are the same type}|G = g) = (1-p)^{2g}. \tag{4}$$

Recalling the interpretation

$$\mathbb{E}S = \mathbb{P}(\text{Two sample genes are the same type})$$

we see that we just want the unconditional probability associated with (4). A textbook calculation is that for X with Geometric(q) distribution

$$\mathbb{E}z^X = \sum_{i\geq 1} q(1-q)^{i-1}z^i = \frac{qz}{1-z(1-q)} \approx \frac{q}{q+1-z}$$

the final approximation holding when $1-z$ is small. Now (4) can be rewritten as $\mathbb{E}S = \mathbb{E}(1-p)^{2G}$, so setting $z = (1-p)^2 \approx 1-2p$ and $q = 1/(2N)$ we get

$$\mathbb{E}S \approx \frac{1/(2N)}{1/(2N)+2p} = \frac{1}{1+4Np}.$$

I will show two more genetics calculations.

MRCA of entire population. A related question that has a nice answer within the Wright-Fisher model concerns $T =$ number of generations back to the most recent common ancestor (MRCA) of the entire population. Note we are referring to a particular gene site, not the entire genome. Write $M = 2N$ for the number of genes in a generation. We saw above that the time (number of generations back) G to the MRCA of *two* individual genes has $\mathbb{E}G \approx M$, and one might expect $\mathbb{E}T$ to be considerably larger. Surprisingly, it isn't. In fact

In the Wright-Fisher model without selection, the mean number of generations back to the MRCA of the entire population is $\approx 2M$. (5)

In the Wright-Fisher model without selection, the mean number of generations back to the MRCA of the entire population is $\approx 2M$.

Imagine starting with a single neutral mutation. If there were no mutations in future, the chance this allele becomes fixed in the population would be $1/M$, by a martingale argument (cf. Lecture 9: the number of copies of the allele fluctuates as a fair game). And the time taken for this "neutral sweep" would be order M, by (5). The chance is much less, and the time much larger, than for a selective sweep (1, 2).

Mathematical derivation of formula (5). Look backwards from the present; in each generation there is some number of "lines of descent" leading to present-generation genes.

[board]

As we proceed backwards these sometimes merge. Where we see $i \geq 2$ different lines of descent, the chance that some two lines meet in the next (previous in time) generation is $\approx \binom{i}{2}/M$. So the mean number of generations for the number of lines to decrease from i to $i-1$ is $\approx M/\binom{i}{2}$. The number of generations T back to the MRCA is the number required for the number of lines to decrease from M to 1, so

$$\mathbb{E}T \approx \sum_{i=2}^{M} M/\binom{i}{2} = M \sum_{i=2}^{M} \frac{2}{i(i-1)} = 2M(1-M^{-1}) \approx 2M.$$

How many of your ancestors are you related to? Genetically, that is. This is fun to say in a popular talk.

You have 2 parents, 4 grandparents, 8 great-grandparents, and back 10 generations you have somewhat less than 1,024 ancestors (some of the lines of descent will have merged). How many of these ancestors have you actually inherited DNA from? You have 46 chromosomes, so if you inherited whole chromosomes you could only be **genetically** related to 46 of the ancestors.

In fact the situation is more complicated because there is *chromosomal crossover* and under a simplified model, one can calculate that you inherited DNA from about 370 of your 10'th generation ancestors. Less than half of them. So even if you can prove that one of your ancestors was King George III or the Qianlong Emperor, this doesn't mean you have royal blood.

So "10" is the critical number of generations, in the sense of the number at which the proportion (of ancestors to whom you are genetically related) drops below half. This is analogous to the criterion by which "7" was deemed to be the critical number of shuffles required to mix a deck of cards.