The Coalescent Model

Florian Weber

Karlsruhe Institute of Technology (KIT)

Abstract. The coalescent model is the result of thorough analysis of the Wright-Fisher model. It analyzes the frequency and distribution of coalescence events in the evolutionary history. These events approximate an exponential distribution with the highest rate in the recent past. The number of surviving lineages in a generation also increases the frequency quadratically. Since the number of surviving lineages always decreases as one goes further into the past, this actually strengthens the effect of the exponential distribution.

This model is however incomplete in that it assumes a constant population-size which is often not the case. Further analysis demonstrates that the likelihood of a coalescence event is inversely proportional to the population-size. In order to incorporate this observation into the model, it is possible to scale the time used by the model in a non-linear way to the actual time.

With that extension the coalescent model still has limitations, but it is possible to remove most of those with similar changes. This allows to incorporate things like selection, separated groups and heterozygous organisms.

1 Motivation

When we want to consider how allele-frequencies change in an environment with non-overlapping generations, random mating and no selection, there are a couple of models to choose from: The Hardy-Weinberg model and the more advanced Wright-Fisher model are probably the most well known ones.

Both have serious disadvantages though: While the applicability of Hardy-Weinberg for real-world-populations is questionable at best, the Wright-Fisher model is computationally expensive and cannot perform calculations from the present into the past.

The solution to this is the coalescent model, that solves these problems by analyzing the Wright-Fisher model in detail to allow us focusing on the relevant aspects of it and perform calculations backwards in time while using much less computational power.

The remainder of this paper is structured as follows: In the following section we will look into simpler models that precede the coalescent model. In section 3 we will then develop the basic coalescent model from this and extend it for

non-constant population-sizes in section 4. After that we will mention further possible extensions in section 5 and provide examples for real-world-usage in section 6. In the final section we will then provide a short summary of the key-points presented in this paper.

2 Introduction

2.1 Hardy-Weinberg

The Hardy-Weinberg model assumes an infinite population-size, random mating, a diploid population and no selection.

The basic idea of the model is to look at the initial frequencies of all allele-combinations in the population and to then calculate the frequencies of the next generation if all individuals mate a random partner to create children with a random allele-set (based on the alleles of the parents).

If the requirements for this model were met, the allele-frequencies would stay constant over time. It is however obvious that the infinite population size reduces the applicability of the model: Even if one defines a way to use frequencies with an infinite set in a meaningful way, real-world-populations never have an infinite size.

This drawback is the main-reason for us to look into different models like the Wright-Fisher model.

2.2 Wright-Fisher

The Wright-Fisher model assumes a finite but constant population size, random mating, non-overlapping generations and no selection.

It's approach differs greatly from the Hardy-Weinberg model in that it introduces specific generations and individuals. In its simplest form the model works on haploid populations, though an extension to diploid populations is possible.

The basic idea is to create a list of individuals of the first generation. To calculate the second generation one creates a second list of the same size and fills it by selecting random individuals from the previous generation as parents and simply copying them. It is important to understand that this approach allows to select individuals in the old generation not just once but also either multiple times or never.

One advantage of this approach is that it is easy to implement in software. The following C++-code performs a complete step in four lines of actual code:

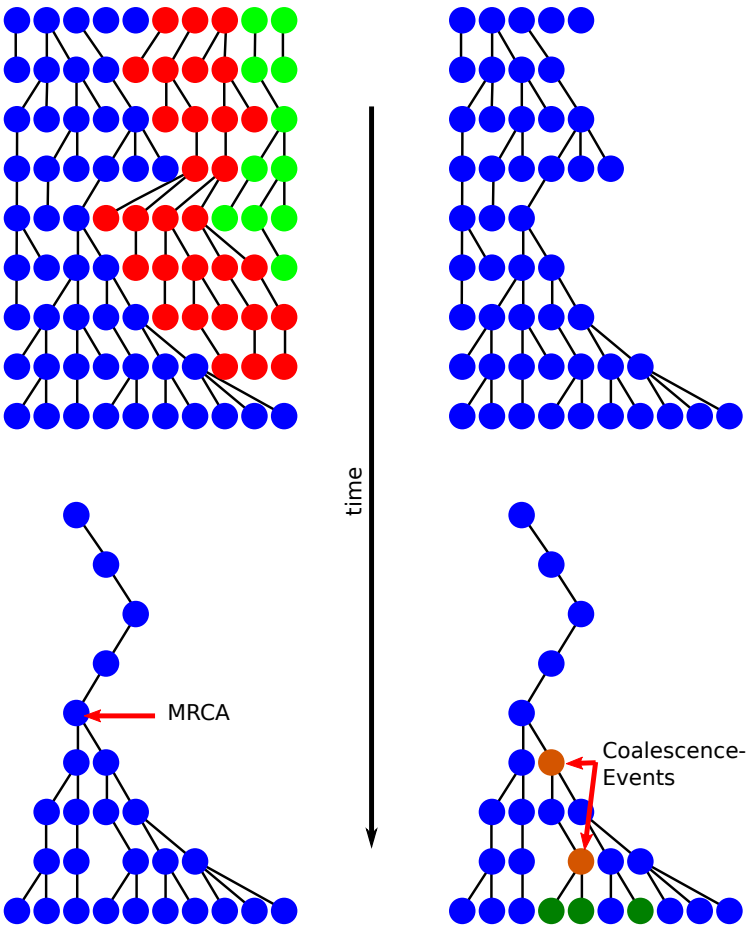

Fig. 1. A possible evolutionary history under the Wright-Fisher model. By reordering the individuals in each generation, it is possible to achieve a simpler visual structure than the direct output of Wright-Fisher. On the top-left we see the entire history of all individuals that lived. On the top-right we see individuals that have the surviving allele. On the bottom-left only the individuals that have descendants in the current generation remain. The individual in the most recent generation with just one individual is the Most Recent Common Ancestor (MRCA). On the bottom-right we see some individuals and their common ancestors marked. Every ancestor (orange) of several later individuals (green) represents a coalescent event in the history.

```
void calc_next_gen(const std::vector<std::uint8_t>& old,
                   std::vector<std::uint8_t>& out) {
    std::random_device rd;
    std::uniform_int_distribution<std::size_t> dist{0, old.size()-1};
    for(auto& individual: out) {
        individual = old.at(dist(rd));
    }
}
```

If we run multiple simulations, we will see that the frequencies do not only not remain constant, but usually end up with a single remaining allele. Once there is only one remaining allele it is naturally impossible for the other alleles to create offspring, which means that they are extinct.

A closer mathematical inspection of the model has shown that, given enough time, it will always converge into that state. The likelihood for an allele to end up as the prevailing one is proportional to its initial frequency. This means that an allele with an initial frequency of 25% will end up as the prevailing one in 25% of all possible simulations. The time until the simulation reaches that state is inversely correlated with the population-size: The larger the population, the longer it will take until all alleles are extinct.

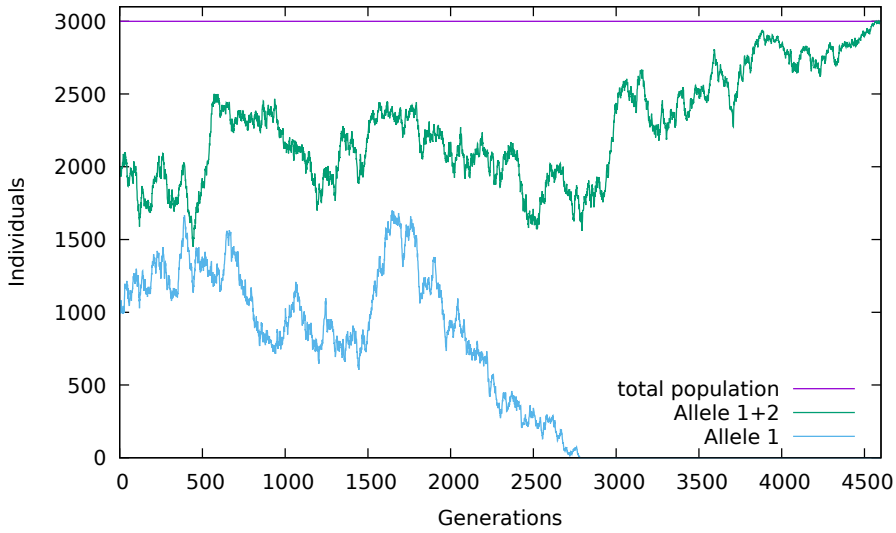

Fig. 2. A simulation of three alleles under the Wright-Fisher model. Initially every allele has 1000 Individuals. The second allele prevails over the other two after 4596 generations.

3 The Coalescent

If we analyze the Wright-Fisher model bottom-up with a focus on coalescence events, we get the coalescent model.

A coalescence event is an event where two roots of sub trees share the same parent. Or, more formally: Given two Generations i and $i+1$, two individuals c_1, c_2 from generation $i+1$ and an individual p from generation i, so that p is the parent of both c_1 and c_2, means that there is a coalescence event between c_1 and c_2.

Since we are currently only considering haploid inheritance of a single gene, the following things are obvious:

- All descendants of c_1 and c_2 are also descendants of p.
- All ancestors of p are also ancestors of c_1, c_2 and all their descendants.
- Together with their descendants both c_1 and c_2 form a tree that is a sub tree of the entire inheritance-tree, where c_1 and c_2 are the respective roots.
- There are no common ancestors between individuals of the different sub trees in a generation later than i.

As a result of this we call p *common ancestor* of all its descendants.

If an individual is the common ancestor of all observed individuals (which may well be the entire population) and there are no later such individuals, we call it the *Most Recent Common Ancestor* or MRCA for short.

A common example for MRCAs are the so called Y-chromosomal Adam and the mitochondrial Eve: Adam is the man from which the Y-Chromosome of all living men descents and Eve is the Woman from which the mitochondrial DNA of all living Humans descents. There are however a couple of problems with it, that are mainly the result of the religious naming:

Some religious people jump to the conclusion that Adam and Eve are the individuals mentioned in Jewish mythology, after whom they are in fact named. This conclusion does however point to a severe misunderstanding of both theory and anthropological evidence for multiple reasons: First and foremost the model does not state that Adam and Eve lived at the same time. This already implies the second problem, namely that they were not the only man or woman when they lived but just the only ones who passed on their genes to the present. Furthermore as MRCAs they are not the only individuals who passed on the genes in question (Common Ancestor), but just the Most Recent ones. Finally there were other previous individuals when Adam and Eve lived, who were MRCAs at that time but are now reduced to common ancestors.

It is worthy to point out that the genes in question are subject to sexual inheritance in a way that more or less resembles haploid inheritance despite belonging to diploid organisms. If this were not the case, analysis would be harder and we would have to resort to less plastic examples.

3.1 Derivation

We consider two generations G_i and G_{i+1} in the Wright-Fisher model with a population size of n.

The likelihood for two individuals for c_1, c_2 from the second generation to pick a common ancestor in the previous generation is $\frac{1}{n}$. This is because c_1 will pick one parent p, leaving $n-1$ individuals that are not parents. When c_2 then picks a parent at random, the change to pick the same parent is trivially $\frac{1}{n}$.

If we ask for the likelihood of picking a same ancestor in the previous two generations, there are two ways to achieve that: The first one is for c_1 and c_2 to pick the same parent. The likelihood for this is $\frac{1}{n}$. The second way is for c_1 and c_2 to pick different parents, but the same grandparents. Since this will scale badly when we ask about even more generations we will invert the question to "How high is the likelihood that there was no coalescence event". By subtracting that value from 1, we can then answer the original questions.

To answer our new question, we simply need to know the likelihood that the individuals and their parents both pick different parents in their generation. That likelihood is $\frac{n-1}{n}$. To get the combined likelihood we then take the product:

$$\left(\frac{n-1}{n}\right) \times \left(\frac{n-1}{n}\right) = \left(\frac{n-1}{n}\right)^2$$

Naturally this scheme works for any number t of generations, so we can generalize this to:

$$\left(\frac{n-1}{n}\right)^t$$

Given that, it is possible to derive that the expected time between coalescence events is n generations. To simplify the model we can rescale that by introducing a variable $\tau = \frac{t}{n}$ and replace i with τn. This results in a representation, in which one unit of coalescent-time corresponds with about one coalescence event:

$$\left(\frac{n-1}{n}\right)^{\lceil \tau n \rceil}$$

As n approaches infinity this term happens to approach $e^{-\tau}$, clearly showing that *the likelihood for two lineages to stay distinct, shrinks exponentially over time.*

Generalizing this to more lineages happens to be easy as well: When we look back for the likelihood of two lineages to stay distinct in the previous generation, we can easily extend this to the likelihood of three alleles staying distinct, by

multiplying it with the likelihood that the third lineage will have a distinct parent from the other two which happens to be $\frac{n-2}{n}$, giving us the following formula:

$$\frac{n-1}{n} \times \frac{n-2}{n}$$

This is because in order to remain distinct, the first individual may choose any parent, the second any parent but the parent of the first and the third any parent but the parents of the first two. Using this scheme, we can extend the formula to an arbitrary number k of individuals:

$$\frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-k+1}{n} = \prod_{i=1}^{k-1} \frac{n-i}{n}$$

It is important to understand that the number of lineages with living descendants (k) may be much smaller than the population-size (n) in any given generation. This is because with every coalescence event, there is also at least one individual (and thereby lineage) in the previous generation that does not procreate (in the constant-population-size model). Therefore any coalescence event that such an individual was part of becomes irrelevant to the simulation and can be ignored. Therefore k will frequently be much smaller than n. If this is the case, it is possible to use the following approximation:

$$\prod_{i=1}^{k-1} \frac{n-i}{n} \approx 1 - \frac{\binom{k}{2}}{n} = 1 - \frac{k(k-1)}{2n}$$

Where $\binom{k}{2}$ is the binomial coefficient, aka the number of ways to pick two lineages from a set of k lineages. This number is about quadratic to k, so we can say that *the likelihood for k lineages to remain distinct in one generation shrinks quadratically with k.*[1]

We can estimate the expected time to the MRCA between k lineages as the sum of the expected coalescence times between the individual coalescence events:

$$E\left[\sum_{k=2}^{n} T(k)\right] = \sum_{k=1}^{n} E\left[T(k)\right] = \sum_{k-2}^{n} \frac{2}{k(k-1)} = 2 - \frac{2}{n}$$

It's easy to see that this time is lower than twice the time for just two lineages, which fits well with the previous claim that the coalescence rate becomes exponentially slower as one goes back in time. It also means that more than half the tree-height is usually spend with just two lineages, whereas most lineages will find their common ancestors relatively fast.

The above is a description for a Markov-process in which each complete set of lineages that exist at some point represents a state and just the number of lineages and individuals define the likelihood for a state-change. The important property here is to understand that the likelihood for a state-change is independent from the time of any previous state-changes, obliterating the need to consider it during computation and analysis.

3.2 Predictions

The result of this is, that the trees created by the coalescence model will usually contain few deep bifurcations with most events happening in the recent past. More specifically the number of relevant lineages shrinks exponentially with the number of generations *and* quadratically with the number of individuals.

Figure 3 provides a visual example of this.

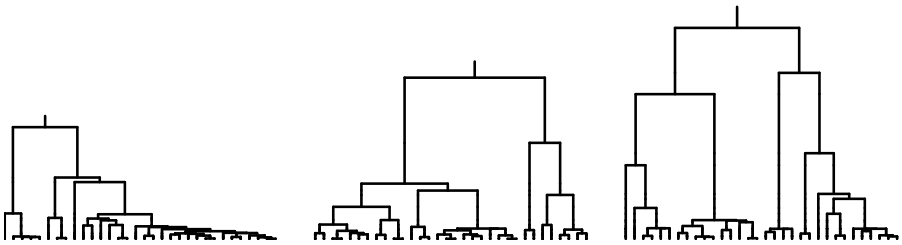

Fig. 3. Exemplary coalescence trees for n=32: Note the small number of deep bifurcations compared to the large number of recent events.[4]

To sum everything so far up:

- The typical history created by the coalescent model has most of its activity in the recent past and is therefore a tree with only a small number of deep bifurcations.
- The coalescent is a likelihood-based model. This means that every thinkable history is possible but potentially extremely unlikely in it.
- Unlike the Wright-Fisher model, the coalescent allows to do calculations backwards in time.
- Thanks to the fact that the model requires no calculations for single individuals and extinct lineages, it is more efficient to compute then the Wright-Fisher model.

In order to use the unmodified coalescent model, we need to assume a finite but constant population size, random mating, non-overlapping generations, at most one coalescent-event per generation and no selection.

4 Non-constant-population-sizes

The coalescent model provides a useful tool to generate plausible histories for individuals in a population of constant size, but a look into reality easily demonstrates that this is often unrealistic. Consider figure 4 that shows the total human population on earth: For 8000 years, it is almost impossible to make out the line, while the number explodes in the last four centuries. As we have seen with the Hardy-Weinberg model, simplifications, even if they appear to be minor, can easily give entirely wrong results. In the following we will see, that this is the case with constant population-sizes as well.

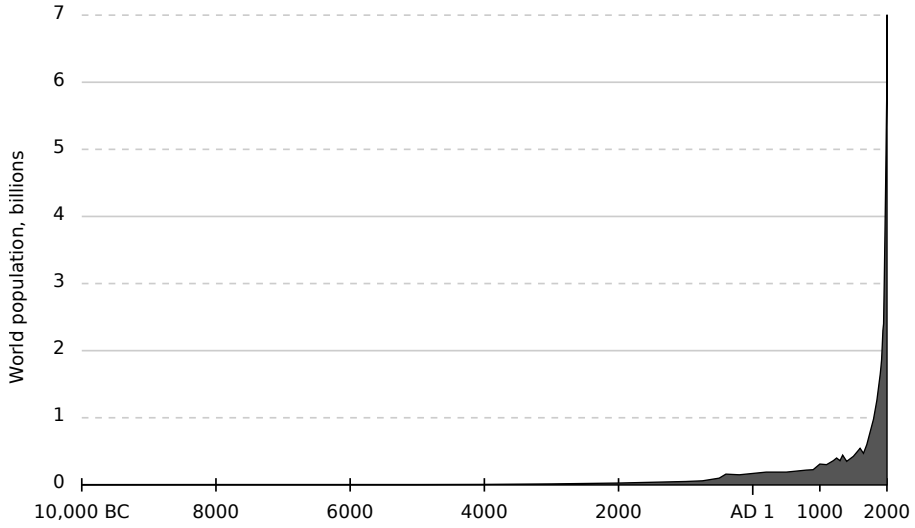

Fig. 4. Human population on earth over time. As it is obviously not even remotely constant, the classical coalescent model is unable to create sensible interpretations.[5]

The target of our remodelling is to be able to make statements about **non-constant, but known** population-sizes.

When considering the previously seen formulas for the likelihood of coalescence events, we can easily see that the likelihood strongly increases with shrinking population-size:

$$\prod_{i=1}^{k-1} \frac{n-i}{n} \approx 1 - \frac{\binom{k}{2}}{n} = 1 - \frac{k(k-1)}{2n}$$

What this means is that the coalescent-rate changes over time. The solution to this is to simply rescale the time.

Recall that we defined our coalescence time as $\tau = \frac{t}{n}$. What this meant was that one unit of τ corresponds to t actual generations.

In order to deal with changing population-sizes we will now redefine this conversion to something that scales with the population-size, namely:

$$\tau = \sum_{i=1}^{t} \frac{1}{n_i}$$

Note that for a constant population-size, the coalescence times are equal, as this equation than simplifies to $\tau = \frac{t}{n}$. This means that special case remains as it is. The results change however if the population-size is not constant. Take for instance a history of five generations with 4, 4, 5, 6 and 6 individuals respectively. With the model for a constant population size we would calculate the average size (5) and thereby conclude that the history covers $\tau = \frac{5}{5} = 1$ units of coalescence time. With the new model we will instead calculate the time like this:

$$\tau = \sum_{i=1}^{t} \frac{1}{n_i} = \frac{1}{4} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6} = \frac{31}{30}$$

We see that even though the average size was the same, the actual evolutionary history uses slightly more coalescence time, due to the greater effect of the smaller generations. If we look at the formula again, we can see that the coalescence time that a generation corresponds to is inversely proportional to its size. Or in simpler terms: *A population with twice the size will get halve the coalescence time.*

4.1 Predictions

This has effects for an exponentially growing population: As the generations in the distant past are much smaller than the more recent ones, rescaling the time in the described way will assign most of the coalescence time to those long past generations. This means that the previous result that most coalescence events are usually recent has to be put into perspective.

Figure 5 provides us with a simple visual representation of the reduced time in larger generations. Furthermore we can see the longterm-effects of different population-growth in figure 6.

4.2 Applicability

The rescaling of the time does in fact converge against the underlying theory for a growing population-size. Practically speaking it is also sufficiently close for some purposes, making it a viable option.

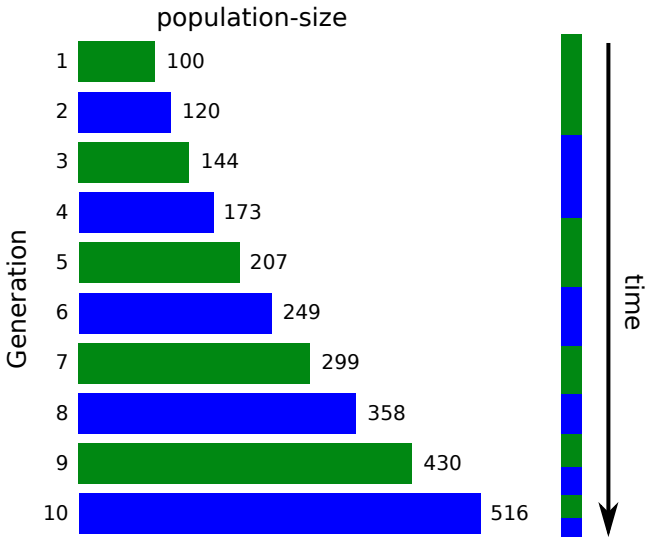

Fig. 5. Population-size vs scaled coalescence time. On the left we see ten generations of an exponentially growing population. On the right we see a to-scale bar-chart that shows how much time the generation corresponds to. Note that the first three generations make up almost halve the time.

One of the plausible things that this approximation tells us, is that for a rapidly growing population, almost no coalescence time passes, which disables the genetic drift during that time.

On the other hand it is important to also understand the limitations of this approach. Marjoram and Donnelly[3] have for instance pointed out, that the common star-like structure that it creates is only realistic if the exponential growth starts from a tiny population-size on, that is to small to use it on human populations.

5 Further Extensions

Aside from non-constant population-sizes the coalescent model also supports many other extensions to reduce its limitations and depend less on questionable assumptions. We will present some of them in the following:

One assumption that often turns out to be unrealistic is for instance random-mating. Realistically speaking, most populations are structured in some way or another, for example into tribes of different sizes. These structures often take the form of clusters with reduced interchange between them. One way to model this is to add parental patches to the Wright-Fisher model with coalescence only

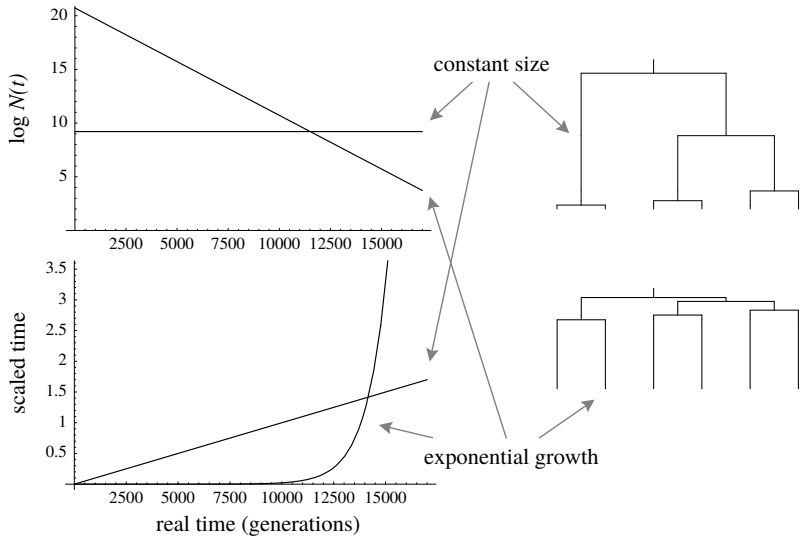

Fig. 6. Exponentially growing and constant populations. Note that the time in the plot flows from right to left! The tree of the constant population has most of its inner nodes far to the bottom with only one very deep bifurcation, as the regular coalescence predicts. The tree for the exponentially growing population on the other has all of its coalescence events in the distant past, due to the non-linear time-scale. [4]

occurring if two individual pick the same patch and the same parent. This can be analyzed in a similar way to the regular coalescent and allows similar usage.

It is easy to imagine that this kind of structure can have a great effect on the overall model, as the effective population-size can be vastly reduced in a certain cluster, implying a vastly increased coalescence-rate. A directly related effect to this is that the individual clusters may have different size and thereby different-coalescence-rates, further complicating the situation.

Another important limitation is that until now we have only considered haploid inheritance. It is however simple to view a diploid population of size n as a haploid population of size $2n$ divided into n patches of two individuals. While a naive approach would allow self-fertilization, a more thorough analysis can deal with that and comes to the result that a diploid organism needs about twice the time until there is a single MRCA, which does exist.

Introducing males and females into this model adds yet another division of the population, that we can deal with in a similar fashion as before. If the number of males and females is about equal, an analysis of autosomal genes will come to the same result for the expected time to the MRCA ($2n$ generations).

Using further analysis, we can even add things like selection, balancing selection or recombination to the model, thereby allowing to lift most of its the initial limitations, except for the non-overlapping generations.

6 Usage

Due to its wide applicability and extendability the coalescent model is used often. One prove of this is the English Wikipedia's long list of programs that make use of it:

BEAST, COAL, CoaSim, DIYABC, DendroPy, GeneRecon, genetree, GENOME, IBDSim, IMa, Lamarc, Migraine, Migrate, MaCS, ms & msHOT, msms, Recodon and NetRecodon, SARG, simcoal2, TreesimJ. [6]

Another number that shows the high significance of this model, is that a google-scholar-search turns up with 59200 results.

One example for such a paper is "mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory"[2] where Atkinson et al. used the model to reconstruct some parts of the human evolutionary history. That paper can also serve as a case-study for the practical use of the model in that it does not only show where the coalescent model has its strength but also demonstrates its weaknesses: While most of the population-size-estimates that they generated appear to more or less match anthropological estimates, we can easily see in figure 7, that they are far off in the case of Oceania. The authors explain this as a result of tribal structures that effectively divided the population into smaller groups, thereby preserving genetic diversity stronger than the total population-size would seem to imply. As previously mentioned, it is in fact possible to model such things as well, but it may not always be easy to do so in a realistic way.

7 Summary

The coalescent is basically the result of analyzing the Wright-Fisher model with statistical methods. It makes certain predictions, that appear to be valid in most of its use-cases, as long as the basic-assumptions of the used variant are met.

The most important aspects that we presented in this article are:

- If the population-size is constant, coalescent-events are, with exponential likelihood, recent.
- The number of historically relevant coalescence events scales quadratically with the number surviving lineages.
- The results of the regular coalescence are not applicable for non-constant population-sizes, but it is possible to bypass this problem by rescaling the time.

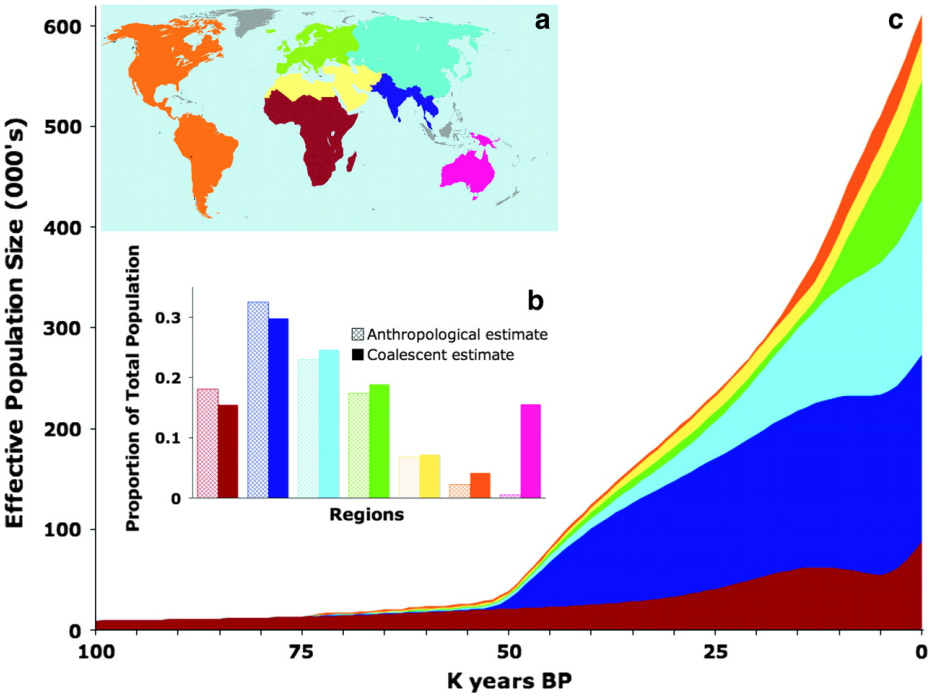

Fig. 7. Predictions of the coalescent model vs. anthropological estimates. [2]

– If one rescales time, the simulated time for a generation is inversely proportional to its size.
– The model does not perform computations for single individuals, single generations or extinct lineages and therefore needs less computational power than the underlying Wright-Fisher model.

Especially as a result of the last point, but also because of its simple extendability the model is widely used in population-genetics.

Sources

Unless otherwise noted, the content of this report is based on the paper "Coalescent Theory"[4] by Magnus Nordborg.

I created all images without citations myself.

References

1. Goncalo Abecasis. Introduction to coalescent models. 2006.
2. Quentin D Atkinson, Russell D Gray, and Alexei J Drummond. mtdna variation predicts population size in humans and reveals a major southern asian chapter in human prehistory. *Molecular biology and evolution*, 25(2):468–474, 2008.
3. Paul Marjoramd and Peter Donnelly. Human demography and the time since mitochondrial eve. *Institute for Mathematics and Its Applications*, 87:107, 1997.
4. Magnus Nordborg. Coalescent theory. *Handbook of statistical genetics*, 2001.
5. El T. Population curve. via Wikimedia Commons [`https://commons.wikimedia.org/wiki/File:Population_curve.svg`].
6. Wikipedia. Coalescent theory — Wikipedia, the free encyclopedia.