University of Glasgow | School of Computing Science

The effects of natural selection on the shape of population graph

Adam Kurkiewicz

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 24, 2016

Abstract

Placeholder for the abstract, which will be written at the end.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: ________________ Signature: ________________

Contents

Chapter 1

Introduction

Introduction. This should place the work in context and will be developed from the introduction produced for assessment.

1.1 Context of the work

The publication of ”On the Origin of Species” (Darwin 1859) is generally accepted to be a founding stone of still vigorously researched field of evolutionary biology. Over the years, as the understanding of the field deepened, it became clear that certain scientific questions within the field can only be answered by obtaining progress in related fields, like establishing laws governing inheritance (genetics) (Corcos and Monaghan 1993), understanding the role of organic molecules in the process of inheritance (molecular biology) (Watson and Crick 1953), understanding population dynamics of predator-prey system using both continuous and discrete methods (mathematical biology), understanding the character of population variability at molecular level (neutral theory of molecular evolution, Markov processes) (Kimura 1968) (Dayhoff 1973), studying of historical populations (archaeology, anthropology, linguistics) (Theunissen and Perlin-West 1989), proposing and verifying hypotheses about populations based on quantitative evidence (statistics) (Allentoft et al. 2015), modeling and simulating biological systems *in silico* (systems biology), and many others.

Next paragraph does not have to be included in the final version of the dissertation.

While the research presented in this work falls within the realm of computational biology, one should understand the difficulty of both undertaking the research in the field and presenting the results, namely, the broadness of topics and tools involved, which need to be understood and taken into account if the research is to be of any significance. A part of the struggle in communication is one of vocabulary, as each of the related fields uses specialized jargon, and researches with expertise in one field might not understand the other fields’ jargon. To ease this difficulty in the beginning of each chapter we shall include a glossary of terminology in use.

1.2 Research question

The scientific question that we will try to answer in this piece of work is directly connected with evolution by natural selection and perhaps might have even occurred to Darwin himself:

Research Question (First Attempt)**.** *How much time does it take for an advantageous trait in an individual to spread across entire population?*

As we will approach the question on the grounds of computational biology, it is perhaps useful to rephrase it in a language more familiar to computer science, which will also serve us the purpose of disambiguating the question's meaning. Firstly, let us introduce the notion of the **population graph**, which we will understand to be a directed acyclic graph, with vertices representing all individuals ever alive and directed edges representing the parent-child relationship between individuals. Additionally, each vertex will be labelled with complete information on individual's traits. For any point in time all individuals alive at that point will be referred to as "population at time t", or simply "population", if the time can be inferred from the context. Representing individuals as a graph has an advantage of allowing us to use a rich set of tools for analysing the shape of the population graph, such as graph centrality measures, or even simply most recent common ancestry of sets of vertices.

Finally, this allows us to reword the research question:

Research Question (Reworded)**.** *How does natural selection affect the shape of the population graph?*

Even after this linguistic juggling the question remains pretty vague. We never specify any parameters of population at any time t, such as the mating scheme, migration, competition for resource, survivability, number of offspring, etc. These parameters will be discussed in chapter 3, along with the appropriate methodology, to which we will prepare ground in chapter 2. We will discuss the architecture of the computer program which constitutes the solution, asdf and is on its own a valuable contribution to the field in chapter 4. The results will be presented in chapter 4 and the conclusions will be drawn in chapter 5.

Chapter 2

Review of the field

This should be a critical survey of the relevant literature, adhering to normal academic conventions in citing references, etc. Here I need to discuss coalescent method vs. forward-in time method vs. sample-based methods.

explain trait, explain mendelian genetics, explain homozygous, heterozygous, dominant, sweep, polymorphism example of polymorphism: impression of bitterness while eating brussel sprouts, linkage equilibrium, panmictic population, allele, epigenetics, probability density function.

2.1 Basic inheritance

Presented work is by no means the first attempt at quantitative study of population genetics. What follows is a review of the most relevant achievements in the field.

2.1.1 Early days

Ever since three European scientist simultaneously and independently rediscovered & experimentally confirmed Gregor Mendel's theory of inheritance [find the original works of DeVries, Correns and Tscherma, Roberts, H. F. Plant Hybridization before Mendel. Princeton: Princeton University Press, 1929.], the interplay between Mendelian Inheritance and Darwinian Evolution has been hotly debated. One claim popular at the time has fallaciously asserted that Mendelian Genetics cannot be a correct model of trait inheritance [cite this guy whose name starts with U]. The reasoning at the core of the claim was that dominant alleles over time sweep the entire population, thus eliminating variation within the population, rendering evolution by natural selection impossible. This misconception was debunked in a very short letter to the editor of "Science" by famous English number theorist Hardy 1908. The letter is a truly master example of a presentation of non-trivial mathematical thought in a manner both rigorous and understandable, even to non-experts. We shall replicate Hardy's argument, for it will become an important tool through the rest of the work. Before we do, let us recall Mendel's laws of inheritance.

2.1.2 Mendelian Inheritance

Let us start by briefly sketching a relevant subset of the laws of Mendelian Inheritance. A more comprehensive reference can be found in an annotated English translation of the original work of Mendel by Corcos and Monaghan 1993.

Firstly, we will be acting under the assumption that each animal characteristic (trait) is either present, or not present in a given organism because of:

Assumption 1. *possession by this organism of one, two or none copies of a gene encoding this characteristic.*

Assumption 2. *the mode of expression of the characteristic, which can be either dominant or recessive.*

We have to remark that these assumptions may not be obviously true for certain seemingly continuous, partially environment-dependent characteristics such as height, weight or individual's position in the autism spectrum. However experiments using various methods have consistently asserted that 80% of human height is inheritable (Silventoinen et al. 2003; Visscher et al. 2006). Although the challenge of identifying all genes affecting height is still open, recent progress using a cross-discipline technique called Genome Wide Association Study (GWAS) identified over 40 genes, which contribute 5 percent points towards total height heritability. To play devil's advocate further, it has been recently established that important traits can be acquired and inherited by non-genetic means [cite palm oil study, cite the mouse study, cite Auchwitz study], giving rise to the field of Epigenetics. We shall consider both environmental and epigenetic factors to be outside of the scope of this research. A curious reader is referred to an accessible, popular science discussion of the topic of GWAS, height inheritance and epigenetics by Maher (2008). A concerned reader should find assurance in that a great many traits do obey Mendelian's Inheritance straightforwardly, a comprehensive list given by [find some comprehensive list].

Before we proceed let us recall the following, standard notation: if an allele encoding a trait is determined to be dominant we will represent it with capital letter, e.g. A. In such case a lack of trait will be considered recessive and be denoted as a. Under the assumption 1, we will consider all three possible genotypes: (A, A) – homozygous dominant, (A, a) – heterozygous and (a, a) – homozygous recessive. Of these the first two the genotypes (A, A) and (A, a) will result in an individual possessing the characteristic. An individual with the last genotype (a, a) will not exhibit the trait. Analogously, one can consider a trait inherited recessively, and in such case the above description is true when "trait" is mutually exchanged with "lack of trait", "not exhibit the trait" with "exhibit the trait", etc.

Further, we will assume that

Assumption 3. *in offspring each of the two alleles present comes from the parents, precisely one from the mother and one from the father. Each of the parent's two alleles has a 50% chance of being inherited.*

We can use our assumptions 1, 2, 3 to work out that in a specific case of breeding two heterozygous individuals, an (A, a) mother and an (A, a) father, each individual offspring will have either of 3 possible genotypes with probabilities $\frac{1}{4}$ for (A, A), $\frac{1}{2}$ for (A, a) and $\frac{1}{4}$ for (a, a). Table 1 gives all possible outcomes of breeding all possible genotypes involving one characteristic along with corresponding probabilities. If the desired mode of inheritance is dominant, the reader can interpret (A, a) and (A, A) as trait-positive and (a, a) as trait-negative. For the recessive case, (a, a) is trait-positive and both (A, A) and (A, a) are trait-negative.

Finally, we have to remark that the assumption 3 may not be correct in a rare cases of sex-linked characteristics (e.g. colour-blindess). These will not be focused on in this research.

2.1.3 Hardy-W principle

Let us define a distribution of a genotype in the population as the ratio of dominant homozygotes to heterozygotes to recessive homozygotes, symbolically $p : 2q : r$. 2q appears in the ratio to ease the computation. Ingenious insight of Hardy is in observing that there are $p : 2q : r$, which can be considered stable, whereas others, although unstable will eventually converge to a stable configuration. To define the exact meaning of stable asdasd let us write in full generality a mass density function for any infinite, randomly breeding population

[here we need to give a proof that if the ratios are p:2q:r, then eventually]

2.1.4 Work of Fisher

2.1.5 Branching Processes

2.2 Coalescent method

all but the most simple models of the coalescent method , which can make it prohibitively

2.3 Forward-Time Modelling

for this let us echo concerns raised by **peng09**

2.4 Population graph

Write about Rohde and Olsen's work. In particular, write about migration around various islands. Discuss the use of MRCA and IA.

Chapter 3

Report of research work

This should bring out the insights and new ideas produced by the review and may continue to suggest further practical work, which would test these ideas. Most importantly, the report should clearly state the overall value of the work which has been done - this should not be lost in a welter of detail.

Here I need to discuss the

Chapter 4

Report of practical work

This section should summarize the practical work done and its results, concentrating on any unusual or original features. Enough detail should be included to assure the reader that the work has been professionally and competently done, and that the results are trustworthy. Any design features, algorithms and data structures of special interest should be described. Documentation to full software engineering standards is not required.

glossary locus, loci, polyploid, chromosome, DAG, extension modules

4.1 Natural Selection Framework

An important contribution of this research work is an implementation of a forward-time population modeling framework. The framework is then used to investigate the research question that we're interested in by focusing on particular population models of our interest. However, the framework does not force its users into any particular choice of the model. A wide range of real-life populations can be modeled by implementing appropriate mutation, breeding, mating, natural selection, migration and inheritance schemes. The schemes are implemented as python functions, and can be glued together using the framework's API into a full population model. This approach is not dissimilar from [cite simuPOP], a particularly successful forward-time simulation framework, which has been used extensively to model [give a list of things it's been used for]. Let us discuss the design of the framework focusing on similarities and differences with simuPOP.

4.1.1 High-Level overview

The central unit of any simulation in the framework is a ”sim”. Sims are computational abstractions of biological individuals. Sims are assigned partners in accordance with a mating scheme, can sexually reproduce with other sims of opposite sex, and produce offspring in accordance with a breeding scheme. Both mating and breeding schemes can be chosen from any provided by the framework or be user-defined. Some mating and breeding schemes are capable of taking into account the individual's relative fitness and therefore can be used to get insights into the process of natural selection. Sims live on the surface of a unit sphere. Each sim after birth is assigned the position of its female parent. After birth and before mating and breeding each sim migrates in accordance with a specified migration scheme. Each sim possess a genome, which is an encoding of that sim's genes. The set of all sims in the simulation forms the set of vertices V in the population graph $G_p = (V, E)$. A pair of sims (u, v) belongs to the set of edges E of G_p if the sim u is a parent of the sim v. It is easily noticed that G_p is a Directed Acyclic Graph (DAG), as no sim can be its own ancestor. Finally sims are serialised into json format and output to hard drive for further analysis. The analysis stage consists of computing MRCA of each generation of sims in the population graph.

simuPOP provides most of the same functionality, however there are significant differences between the two programs. Exact comparison is difficult, because simuPOP is mature software in constant development for over a decade, whereas the proposed framework is an effect of 7 months of part-time development.

Most notably, simuPOP does not allow for graph-theoretic analysis of simulated populations, and does not support outputting population graphs into permanent storage for later analysis. Aside from the educational aspect of creating my own framework this was the main reason why the effort was undertaken.

4.1.2 Performance

SimuPOP achieves low memory and CPU footprint by using python's *extension modules*, which allows the implementation a much more direct access to underlying hardware than pure python implementation. However, there is a significant price that simuPOP pays for this strength.

Firstly, development is slowed down by higher cognitive burden put on the programmer, as C++ requires manual memory management and python's C++ API is arguably less friendly than pure python's. A reader critical of the previous statement can judge for themselves by looking at this snippet of simuPOP's source code:

```
PyObject * genoObj = PyTuple_New(alleles.size());
// set value
for (size_t j = 0; j < alleles.size(); ++j)
    PyTuple_SET_ITEM(genoObj, j, PyInt_FromLong(alleles[j]));
return genoObj;
```

which could be equivalently expressed in pure python

```
genoObj = tuple(alleles)
return genoObj
```

Secondly, a study correlating error rates with programming languages has shown C++ programmers make more errors than python programmers: [cite error rates: macbeth.cs.ucdavis.edu/lang_study.pdf]. These results must be treated with a pinch of salt – correlation does not imply causation and arguably more depends on the quality of the programmer than on the quality of programming language, additionally the methodology of parsing commit messages to detect bug fixes employed by this study is rather experimental.

Thirdly, the use of extension modules limits execution environment of simuPOP to Cython, whereas the framework proposed has been successfully tested both under python 2, 3 (the entire framework) pypi and jython (with exception of the graph analysis module).

Let us end this note on performance by saying that the performance of the framework is satisfactory for our purpose. Timing one of the simulation scripts (with MRCA analysis) reveals that the simulation of an initial population of 1000 sims for 70 generations accomplishes in under 2 minutes on consumer grade hardware.

4.1.3 Genome

SimuPOP allows for extensive modeling of genomes, such as modeling multiple loci situated on multiple chromosomes even in polyploid organisms [cite source code: sourceforge.net/p/simupop/code/HEAD/tree/trunk/src/individual.cp

Proposed framework takes much simpler approach consistent with our requirement of investigating a single feature, in its present form allowing for only one locus capable of containing only 2 alleles.

4.1.4 Migration

Mine is more accurate because of the unit sphere.

4.1.5 Mutations and Inheritance

4.1.6 Natural Selection

4.1.7 Generations

discrete generations as in simuPOP. This is unlike Rohde and Olhsen.

4.1.8 Mating and Breeding

Depending on capture difficulty in random mating -¿ random fluctuations in population

Chapter 5

Conclusion

This section should summarize the work and its significance, comparing practical results with theoretical expectations where appropriate. Suggestions for possible extensions, or research projects arising from the work, may also be included here.

Chapter 6

References

A complete ordered list of any references which have been cited or consulted. Appendices. Full listings of software written, with selected test results, should be included in the accompanying CD. A summary log should also be appended, recording significant events in the progress of the project.

Bibliography

Allentoft, Morten E. et al. (2015). “Population genomics of Bronze Age Eurasia”. In: *Nature* 522.7555, pp. 167–172. ISSN: 0028-0836. DOI: `10.1038/nature14507`. URL: `http://www.nature.com/nature/journal/v522/n7555/full/nature14507.html`.

Corcos, Alain F. and Floyd V. Monaghan (1993). *Gregor Mendel’s Experiments on plant hybrids: a guided study*. Rutgers University Press. ISBN: 0813519217.

Darwin, Charles (1859). *On the origin of species. Or the preservation of favoured races in the struggle for life.* London, Albemarle Street: John Murray.

Dayhoff, Margaret (1973). *Atlas of protein sequence and structure. Supplement*. Washington, D.C: National Biomedical Research Foundation. ISBN: 0912466073.

Hardy, G. H. (1908). “Mendelian Proportions in a Mixed Population”. In: *Science* 28.706, pp. 49–50.

Kimura, M. (1968). “Evolutionary Rate at the Molecular Level”. In: *Nature* 217.5129, pp. 624–626. DOI: `10.1038/217624a0`.

Maher, Brendan (2008). “Personal genomes: The case of the missing heritability”. In: *Nature* 456, pp. 18–21. DOI: `10.1038/456018a`.

Silventoinen, Karri et al. (2003). “Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries”. In: *Twin Research and Human Genetics* 6 (05), pp. 399–408. ISSN: 1839-2628. DOI: `10.1375/twin.6.5.399`. URL: `http://journals.cambridge.org/article_S1369052300004001`.

Theunissen, L. T. and Enid Perlin-West (1989). *Eugène Dubois and the Ape-Man from Java*. Kluwer Academic Publishers. ISBN: 1556080816.

Visscher, Peter M et al. (2006). “Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings”. In: *PLoS Genet* 2.3, e41. DOI: `10.1371/journal.pgen.0020041`. URL: `http://dx.plos.org/10.1371/journal.pgen.0020041`.

Watson, J. D. and F. H. C. Crick (1953). “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171.4356, pp. 737–738.