Modelling Canalisation of a Genetic Network

Author: Matthew Campos

Submitted August 2020

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science/Research at Imperial College London

Formatted in the journal style of Evolution & Development

Submitted for the MSc in Computational Methods in Ecology and Evolution

1 Declaration

All raw data collected were from the simulation I created for my project. The simulation requires the input of genetic model systems, and mathematical equations used to derive output. The model system and equations used in the simulation were sourced from the work of /textitOmholt et al, and *Gjuvsland et al.* I was reponsible for data processing, cleaning and analysis. All analyses presented in the paper are from the simulations, with the help of my supervisor.

2 Acknowledgements

I would like to firstly thank Dr. Scott Rifkin for being a wonderful supervisor and guiding me throughout the project. This includes understanding background knowledge, results and overall research purposes. Secondly, thank you to Dr. Thomas Bell for agreeing to be my internal supervisor, making sure I am aware of the process of the project and ensuring my safety during such difficult times.
Finally I would like to thank the laboratory of Dr. Rifkin- Antonia Darragh, Jessica Bloom, Alexis Cugini, Yang Bing and Rachel Goodridge for being very welcoming and having wonderful and insightful weekly meetings. Good luck with everything!

3 Abstract

4 Introduction

Species migration can result in the following: (i) it allows individuals and species to colonise new areas and create new subpopulations. Over time, ecological events cause species to become reproductively isolated. This is known as allopatric speciation and it leads to the splitting of lineages, differentiating both related populations long-term. This differentiation also increases with geographic distance. Without gene flow, both sets of species are able to rapidly evolve in their local optimums /citegarcia1997genetic. Overall, species that are reproductively isolated have more pronounced modifications which can be observed phenotypically and genotypically (Pongratz, Gerace & Michiels 2002, Sato, Isagi, Sakio, Osumi & Goto 2006). (ii) Migration can allow isolated species to attempt to colonise each other's habitats. If species are capable of interbreeding, parapatric or peripatric speciation may occur, depending on distance. This introduces new sets of alleles into an environment and as species interact and pass on its heritable genes, it changes the developing genetic makeup of local species. The latter case includes gene flow which helps to maintain the genetic diversity in an area but has been shown to homogenize populations over long periods of time, through the recombination of genes (Sato et al. 2006). Advancements in genotypic techniques now enable us to study the genotypic effects and further our understanding of phenotype-genotype relationship. As organisms evolve, phenotypic evolution is assisted with genotypic evolution. Collective expression of certain genes through pathways assist in the morphology and behaviour we observe in species. Hereditary genome alterations through random changes in molecular mechanisms change varying aspects of the species (Chandrasekaran & Betrán 2008) These molecular changes induced by mutation and recombination lead to the variation of descending species (Chandrasekaran & Betrán 2008, Ohno 1999, Brown 2002). Over time, evolutionary forces involving drift and selection acts on these polymorphisms and those most fit passes their variant genome, and phenotype as a result to future generations. This is the foundation of Darwin's theory natural selection.

To better understand species evolution, we can focus on the development of their genome network. Orr showed that there is variation with respect to genetic differences or gene influence on phenotype. The effects of adaptive and non-adaptive processes vary among species where there is no common set of genes involved, nor is the effects and interactions of the genes similar for species (Orr 1998) Although generalizations cannot be made of genetic function and interactions, what can be considered is the pattern at which these genetic processes develop over time. Genetic network simulations can be used to understand these patterns of evolution and the effect on phenotype-genotype relationships. Long temporal periods allow genetic interactions within a network to robustly develop, canalising the network (Orr 1998, Lynch 2007). Lynch highlighted the significance of non-adaptive processes as well in shaping genetic networks. His study showed that networks can still evolve its architecture and become redundant even without the influence of natural selection (Lynch 2007). Robustness can evolve from the effects of epistasis, additivity and dominance, all of which are connected (Omholt, Plahte, Øyehaug & Xiang 2000).

Species evolution is non-linear, descending with modification and constant splitting from lineages of a common ancestor. This continuous process over long temporal periods results in the accumulation of optimal genetic adaptations that results in a robust network structure. This can be quantified through fitness or reproductive success. There is a balancing act as selection aids to

propagate fitter variants in a population, while mutation and environmental change limits such propagation (Burt 1995). What this study focuses on is how these forces affect the development of genes and the genetic network. Specifically, with the effects of gene flow on a genetic network that has evolved in isolation. The patterns of change that a genetic network undergoes with these adaptive and non-adaptive evolution processes. Even once a robust structure is reached, how does the structure resist change and maintain its network despite perturbations and evolutionary processes. As species evolve, studies have shown that pathways have a safety margin, that make them resistant to change such as mutations (Bourguet 1999). Species best suited to their environment will evolve to their local optima, which we can represent as a quantitative value. The further apart these values are, what I label as environmental distance, the greater the variance of the two species. The concern is on how a network responds when these evolutionary forces come into play and seeing the evolving genetic interactions. Investigating the effects of changing migration rates, two variant genetic networks, environmental distance and patterns of migration.

Ecological events eliminate barriers and allow species to migrate into new environments, introducing new sets of genes in an environment. The presence of variant genes and network structures from gene flow hinders local adaptation and fixation of adaptive genes (Burt 1995). Using quantitative trait loci (QTL) we are able to numerically interpret and visualise the patterns of change. Previous research looked at the effects of gene flow, selection and mutation at generating local adaptation at the phenotypic level, showing how maintenance of alleles and linkage is important in adaptation (Yeaman & Whitlock 2011). It was shown that with random perturbations and aid of genetic modifiers, there are bounds for which selection for canalization can act on, leading to evolution of robustness. They also showed that under migration selection balance, selection for robustness increases with the migration rates (Proulx & Phillips 2005). This research will be looking at the changes in genetic architecture dynamics and the interactions of the varying systems. As a genetic network evolves, there exists a threshold which is actively regulating these homeostatic genes (Gjuvsland, Plahte & Omholt 2007). As selection for robustness occurs within the local population, it can give insight into the change in architecture and statistically significant interaction (Gjuvsland, Hayes, Omholt & Carlborg 2007).

Using a multi-locus system, I will construct a genetic network and simulate the effects over many generations and see how the output of the network changes, specifically looking at allelic interactions and tracking the fitness over time. Variance in fitness should decrease as a genetic network becomes robust, making it resistant to perturbations. Fitness can be quantified as reproductive success and is represented by passing on quantitative values generated from the alleles. These values are used to derive the trait values of individuals of which phenotypic values are then calculated and used as probabilities for fitness. The expectation is that after migration, a more robust network is formed when compared to before migration. At the start allowing new alleles to enter the population will result in a less robust network and more susceptible to perturbations (García-Ramos & Kirkpatrick 1997). Especially when the migrant network is a different structure, gene flow will allow maladaptive alleles to enter and should those be passed on, will impose a fitness cost to individuals (Tigano & Friesen 2016) However, over time, the network should adapt and become resistant to such perturbations.

5 Methods

To understand the effects from the evolutionary forces, I wrote a R script that constructs genetic networks and simulates their evolutions, allowing migration to occur between two populations. All functions to perform adaptive and non-adaptive processes were written from scratch and implemented in the simulation. The following functions are:

- Population: initialises the starting populations of specified size where each individual (row) contains 12 allele sites (4 per gene). Since it is a di-allelic model, it is a 2-dimensional array.
- Fitness: determines the fitness value of each individual based on their trait values and used as a probability for offspring contribution. A heavy tailed Cauchy distribution is used to determine fitness value from trait values. Each individual has a probability of passing on their genotype to the next generation and function randomly samples from the distribution to select parents, representative of genetic drift.
- Mutation: produces an array same dimensions as the population and random uniformly distribution of values to determine which sites undergo mutation based on inputted mutation rate. Generates a new value using a normal function with current value as the mean and a standard deviation of 0.001.
- Recombination: randomly chooses which site, and if any consecutive sites downstream, to switch allele values for each individual.
- Migration: using a uniform distribution, randomly generates values for each individual in the migrant population to determine which individuals will migrate and replace those in the main population. Population is kept constant in both populations.

5.1 The model

A di-allelic interlocus model from the research of *Omholt et al.* In this case, all the genes are hereditary, representing only the regulatory and coding region which determine protein expression and rate of expression. Studies has shown that mutations along the coding region are known to cause morphological variation within species (Stern & Orgogozo 2009). Similar to the work of Omholt et al, this model structure evolves dominance through epistatic interactions and regulatory effects. Using a system of equilibrium solutions and solved ordinary differential equations (ODE), simulated protein concentrations corresponding to phenotype are measured over time (Omholt et al. 2000). Here I consider the sites as quantitative factors of protein function, and trait value is determined by protein concentrations. The greater the amount of protein expressed, the larger the trait value. The model consists of three genes, X_1, X_2 and X_3. Let j represent the genes where $j = 1, 2, 3$, each gene X_j consists of two alleles, X_{j1} and X_{j2}. This leads to the following formula:

$$y_j = X_{j1} + X_{j2} \tag{1}$$

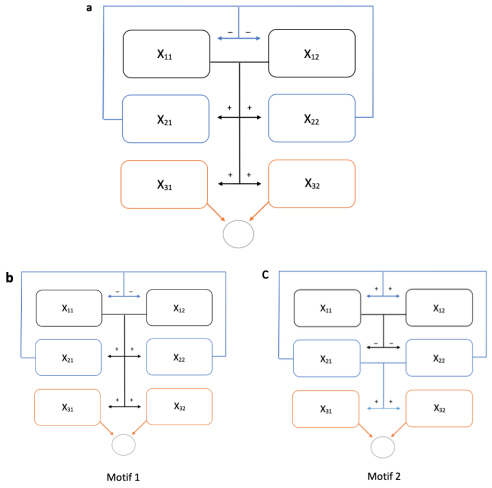

Figure 1: Diagram showing the genetic model and two variants used to represent the migrant population. (a) Interlocus model of the population in focus. Lines labelled with mathematical symbols showing the interactions between genes. Gene X_1 interacts with both gene X_2 and X_3, positively regulating both of them. To limit site values below infinity, gene X_2 is reponsible for negatively autoregulating X_1. There is an output for each gene where $j = 1, 2$ and $y_j = x_{j1} + x_{j2}$. Gene X_3 contains the trait values for each individual, which is the output. Circle represents phenotype which is determined from trait values using a Cauchy Distribution. (b) and (c) represent models for the migrant population. (b) is the same pathway and regulation as (a) however (c) is switched where gene X_1 negatively autoregulates X_2, and gene X_2 postively regulates X_1 and X_3. Again, output of gene X_3 are the trait values used to derive fitness.

Where y_j is the total protein concentration at each gene. There are four sites which represent the different factors affecting protein production. These are a, γ, ϑ and P, where a is the protein production rate while γ is the degradation rate (Omholt et al. 2000). For both sets of populations, a single gene, X_3 determines the trait value for individuals and quantifiably differentiates the populations in terms of morphology (Orr 2001). For the population in focus, gene X_1 positively regulates gene X_2 and gene X_3, and gene X_1 is negatively regulated by gene X_2. This is to regulate trait value and prevent the value from exceeding to infinity. As gene X_2 increases in expression, it decreases X_1 expression, negatively autoregulating the system and limiting its value. Let $j = 1, 2, 3$ and $i = 1, 2$, from the separate researches of *Omholt et al*, and *Gjuvsland et al*, R_j is a regulatory Hill Function representing a Michaelis-Menten mechanism, where $S(y_j, \theta, P) = \frac{y_j^P}{y_j^P + \theta^P}$. The Hill Function explains the relationship between regulator and producer, where $\vartheta$is the amount of regulator needed for 50% production rate and P affects the steepness of the curve (Gjuvsland, Hayes, Omholt & Carlborg 2007, **?**). Should the network be negatively regulated, it leads to the following equation:

$$R_j(y) = 1 - S(y, \theta_j, P_j), j = 1, 2 \tag{2}$$

And if positively regulated:

$$R_j(y) = S(y, \theta_j, P_j), j = 1, 2 \tag{3}$$

Again, letting $j = 1, 2, 3$, as gene X_1 positively autoregulates gene X_2 and gene X_3, and gene X_2 negatively autoregulates gene X_1, this results in the following equations:

$$R_{1j}(y_2) = 1\text{–}S(y_2, \theta_{2j}, P_{2j}), \tag{4.1}$$

$$R_{2j}(y_1) = 1\text{–}S(y_1, \theta_{1j}, P_{1j}), \tag{4.2}$$

$$R_{2j}(y_1) = 1\text{–}S(y_1, \theta_{3j}, P_{3j}) \tag{4.3}$$

μ is the ratio of $\alpha$and γper locus. Using the equilibrium solutions, total protein concentration is calculated by the following equations:

$$y_1 = \mu_{11}(1–S(y_2, \theta_{21}, P_{21})) + \mu_{12}(1–S(y_2, \theta_{22}, P_{22})) \tag{5.1}$$

$$y_2 = \mu_{21}(S(y_1, \theta_{11}, P_{11})) + \mu_{22}(S(y_1, \theta_{12}, P_{12})) \tag{5.2}$$

$$y_3 = \mu_{31}(S(y_1, \theta_{31}, P_{31})) + \mu_{32}(S(y_1, \theta_{32}, P_{32})) \tag{5.3}$$

5.2 Migrant network

For the first motif, the genetic network will be the same as the main population, just evolving to a different local optimum value. For the second motif however, the difference is that gene X_1 negatively regulates gene X_2, while gene X_3 and gene X_1 are positively regulated by gene X_2. The formulas used to derive y_1, y_2 and y_3 values for the migrant population are as follows:

$$y_1 = \mu_{11}(S(y_2, \theta_{21}, P_{21})) + \mu_{12}(S(y_2, \theta_{22}, P_{22})) \tag{6.1}$$

$$y_2 = \mu_{21}(1–S(y_1, \theta_{11}, P_{11})) + \mu_{22}(1–S(y_1, \theta_{12}, P_{22})) \tag{6.2}$$

$$y_3 = \mu_{31}(S(y_2, \theta_{31}, P_{31})) + \mu_{32}(S(y_2, \theta_{32}, P_{32})) \tag{6.3}$$

This is to represent the concept of speciation but can still integrate in the other population and interbreed.

5.3 The simulation

A total of 44 different permutations of environmental distance, genetic network structure, migration rates and migration patterns were simulated for 1,200 generations each run. For the effect of genetic drift and to account for the large deviations of values, a Cauchy distribution is used to generate fitness probabilities per generation. Since the Cauchy distribution is characterized for its heavy tails, it is able to account for values that greatly deviate from desired trait value. The values entered in the Cauchy distribution are the desired trait values. It is important to note that environment is kept constant both spatially and temporally. The main population was kept at a constant trait value 50 with a standard deviation 8, while the migrant population alternated between 65 and 80 with standard deviation 10. The large standard deviations characterise the varying forms of morphology that can be noticed in species. In addition to this, these trait values also represent the environments of both populations and the local optimums they evolve to. The probabilities extracted from the Cauchy distribution are for parental contribution to offspring for the next generation. As the network evolves to a stable state, the feedback loops should maintain the homogeneity of the system. For the simulation we assume that both populations have the same size and stay constant, with migrants replacing individuals. There is no spatial structure and all individuals have an equal chance of being replaced. Both populations undergo stabilising selection towards different specified trait values, thus homogenizing the population over time (Sato et al. 2006). These trait values represent the different environments. Alleles for each individual can either be homogenous, using a uniform distribution to determine starting value between 0.1 and 0.3 for both populations, or heterogenous, using a uniform to randomly

generate the starting allele values, again between 0.1 and 0.3. To get a value from equation 2, there must be a value for y, so uniformly distributed starting values were generated for the three genes. If the population is homogenous, each individual in the population started with the same value at each gene, otherwise have differing values if heterogenous. Recombination is equal chance at any locus and interchanges the alleles and everything downstream. Mutation can occur at each locus by randomly deviating from the current value. The probability is the same constant for both populations where each locus has an equal chance of mutating. Mutation value is kept constant at 0.0011 per site. A cis-mutation in the second gene will affect the corresponding value in the first gene as in equation 3, gene y_2 negatively autoregulates gene y_1, while a trans-mutation in gene y_1 will affect the expression in gene y_3. As mentioned before fitness is determined by phenotypic value as the offspring contribution per generation. Each individual per generation has no limit as to how many times they can be a parent, however the standard deviation of 8 and 10 in the Cauchy distribution ensures varying combination of parents are produced. Migration rates varied between 1%, 3% and 5%. As migrant individuals enter the population, they randomly replace individuals in the population. With constant population size, this represents immigration and emigration. Furthermore, low migration rates were used to prevent migration population from completely replacing the original population and allowing the network to be able to adapt to the new values. Both populations have a burn-in period of 80 generations to evolve in their own environments before migration can happen. Also, migration can only occur until the 700th generation. The remaining 500 generations were to assess how the network responds to the migration. Pattern of migration was also considered, varying between each generation, every 10 generations, every 5 generations and random (between 1% and 5% each occurence) after the 80th generation.

5.4 Analysis

Analysis was done on the recorded fitness, trait values and population arrays. At the end of each simulation, fitness is normalised by dividing fitness probabilities with the medians of Cauchy distributions, with 1.0 being the highest possible fitness value. Firstly, control conditions of no migration were simulated to see how rapidly isolated networks evolve. Without migration, I expect rapid evolution of allele values (García-Ramos & Kirkpatrick 1997). However, which of the two starting genetic network makeup: homogenous or heterogenous evolve faster to their optimums. With migration, I wanted to see its effect on a robust network and how long it takes for the network to recover after migration. As the migration rate and patterns were set, can track allele values in-between periods of migration and analysing recovery. By the 700th generation, the network should be more robust that it recovers faster compared to when migration starts in the 80th generation. In addition to this, the length of time foreign alleles persists in the environment as how migration patterns affect this (W. Morris, E. Diffendorfer & Lundberg 2004). Finally comparing robustness and variance in before and after migration. Comparing them, variance after migration should be less as the network has had more time for evolutionary processes to act on and evolve.

6 Data Code and Availability

All code can be found in the GitHub Repository:
https://github.com/matthewcampos/CMEECourseWork.git

References

Bourguet, D. (1999), 'The evolution of dominance', *Heredity* **83**(1), 1–4.

Brown, T. A. (2002), How genomes evolve, *in* 'Genomes. 2nd edition', Wiley-Liss.

Burt, A. (1995), 'The evolution of fitness', *Evolution* **49**(1), 1–8.

Chandrasekaran, C. & Betrán, E. (2008), 'Origins of new genes and pseudogenes', *Nature Education* **1**(1), 181.

García-Ramos, G. & Kirkpatrick, M. (1997), 'Genetic models of adaptation and gene flow in peripheral populations', *Evolution* **51**(1), 21–28.

Gjuvsland, A. B., Hayes, B. J., Omholt, S. W. & Carlborg, Ö. (2007), 'Statistical epistasis is a generic feature of gene regulatory networks', *Genetics* **175**(1), 411–420.

Gjuvsland, A. B., Plahte, E. & Omholt, S. W. (2007), 'Threshold-dominated regulation hides genetic variation in gene expression networks', *BMC Systems Biology* **1**(1), 57.

Lynch, M. (2007), 'The evolution of genetic networks by non-adaptive processes', *Nature Reviews Genetics* **8**(10), 803–813.

Ohno, S. (1999), Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999, *in* 'Seminars in cell & developmental biology', Vol. 10, Elsevier, pp. 517–522.

Omholt, S. W., Plahte, E., Øyehaug, L. & Xiang, K. (2000), 'Gene regulatory networks generating the phenomena of additivity, dominance and epistasis', *Genetics* **155**(2), 969–980.

Orr, H. A. (1998), 'The population genetics of adaptation: the distribution of factors fixed during adaptive evolution', *Evolution* **52**(4), 935–949.

Orr, H. A. (2001), 'The genetics of species differences', *Trends in ecology & evolution* **16**(7), 343–350.

Pongratz, N., Gerace, L. & Michiels, N. K. (2002), 'Genetic differentiation within and between populations of a hermaphroditic freshwater planarian', *Heredity* **89**(1), 64–69.

Proulx, S. R. & Phillips, P. C. (2005), 'The opportunity for canalization and the evolution of genetic networks', *The American Naturalist* **165**(2), 147–162.

Sato, T., Isagi, Y., Sakio, H., Osumi, K. & Goto, S. (2006), 'Effect of gene flow on spatial genetic structure in the riparian canopy tree cercidiphyllum japonicum revealed by microsatellite analysis', *Heredity* **96**(1), 79–84.

Stern, D. L. & Orgogozo, V. (2009), 'Is genetic evolution predictable?', *Science* **323**(5915), 746–751.

Tigano, A. & Friesen, V. L. (2016), 'Genomics of local adaptation with gene flow', *Molecular ecology* **25**(10), 2144–2164.

W. Morris, D., E. Diffendorfer, J. & Lundberg, P. (2004), 'Dispersal among habitats varying in fitness: reciprocating migration through ideal habitat selection', *Oikos* **107**(3), 559–575.

Yeaman, S. & Whitlock, M. C. (2011), 'The genetic architecture of adaptation under migration–selection balance', *Evolution: International Journal of Organic Evolution* **65**(7), 1897–1911.