Winter 2014

Investigating The Effects of Food Chain on Sympatric Speciation Using ECOSIM

Sara Bandehbahman
University of Windsor

Recommended Citation

Bandehbahman, Sara, "Investigating The Effects of Food Chain on Sympatric Speciation Using ECOSIM" (2014). *Electronic Theses and Dissertations*. 5016.
https://scholar.uwindsor.ca/etd/5016

Investigating The Effects of Food Chain on Sympatric Speciation Using ECOSIM

By

Sara Bandehbahman

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of **Computer Science**
in Partial Fulfillment of the Requirements for
the Degree of **Master of Science**
at the University of Windsor

Windsor, Ontario, Canada

2013

Investigating The Effects of Food Chain on Sympatric Speciation Using ECOSIM

by

Sara Bandehbahman

APPROVED BY:

Dr. Trevor Pitcher
Department of Biological Sciences

Dr. Luis Rueda
School of Computer Science

Dr. Robin Gras, Advisor
School of Computer Science

12/20/2013

DECLARATION OF CO-AUTHORSHIP

I hereby declare that this thesis incorporates material that is result of joint research, as follows:

This thesis also incorporates the outcome of a joint research undertaken in collaboration with Sara Bandehbahman and Andrei Cerghet under the supervision of professor Robin Gras. The collaboration is covered in Chapter 1 of the thesis and whenever a biological discussion is required. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of co-author was primarily through the provision of required background biological information.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from the co-author to include the above materials in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

ABSTRACT

The study of sympatric speciation in evolutionary biology is facing the obstacle of unifying empirical studies with existing theoretical investigations. Disruptive selection due to preferential food resource usage is considered as the main hypothesis to explain the sympatric speciation occurrence in empirical studies.

We extend an individual based evolving predator-prey ecosystem platform called "EcoSim" [Gras et al. 2009a] to model a dual resource system. We investigated whether and in which conditions the selective pressures acting on foraging behaviors drove sympatric speciation. We observed clear results showing some behavioral modifications occurring as a consequence of preferential resource usage. We also observed many cases where the sympatric speciation criteria described in the literature were fulfilled. Using several machine learning techniques, we extracted explicit rules that can predict with a very high accuracy the occurrence of sympatric speciation based on ecological factor observations. Moreover, we confirmed that the existence of a second food resource is determinant for the emergence of sympatric phenomenon. We also proved that our method is able to discover very generic rules which may later be used to structure empirical studies.

DEDICATION

To my dear husband and my beloved son, who patiently endured all the way far from me...

ACKNOWLEDGEMENTS

I would like to gratefully thank Dr. Robin Gras my supervisor, for giving me the opportunity and continued guidance to apply and extend my knowledge in a cross disciplinary field, working at the forefront of computer science and evolutionary ecology. I would like to extend my gratitude towards Andrei Cerghet, my research partner and biology cohort, who meticulously helped me in collecting all the required biological information needed for this research. I would like to give my special thanks to Morteza Mashayekhi and all my fellow students in the Dr. Gras lab, for the humbling experience of allowing me to take part in such an enriching learning environment. Lastly I would like to thank my committee members Dr. Trevor Pitcher from Biology department, and Dr. Luis Rueda from Computer science department for accepting to allocate part of their valuable time to evaluate my research.

This work was made possible by the facilities of Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

Chapter 1

INTRODUCTION[1]

The study of speciation continues to be a fundamental area of interest in biology. Historically, Darwin's approach to studying natural selection[Darwin 1859]was the focus of evolutionary research. Mayr in the mid-20th century shifted the focus towards studying patterns of gene flow through geographical isolation, sparking interest in sympatric and allopatric speciation as modes of evolution [Via 2001].

Sympatric speciation is defined as the splitting of an ancestral species into two or more reproductively isolated groups without geographical isolation of those groups[Coyne 2007]. While allopatric speciation or geographical speciation, is the result of geographical isolation between populations, where "in the absence of gene flow, reproductive isolation arises gradually and incidentally"[Hoskin et al. 2005].

The dominant view about speciation focuses on isolating mechanisms arising from restricted gene flow due to geographical barriers, eventually leading to genetic divergence [Turelli et al. 2001]. Sympatric speciation has received less support in the literature, despite several corroborating mathematical models and more recently empirical evidence [Wilson et al. 2000][Jiang et al. 2008]. Observing active speciation in a natural environment through empirical observation is often hardly possible for the most complex forms of organisms due primarily to generation times, which dictate the span necessary for genetic divergence to accumulate, as well as population tracking to guarantee geographic isolation is maintained. For these reasons empirical evidences to support sympatric speciation are only slowly accumulating and new tools, and methods are being developed and employed to identify the underlying evolutionary mechanisms.

---

[1] This is the outcome of joint research

In recent years computer simulations of ecological systems have been proposed, with the purpose of modeling different evolutionary mechanisms described in natural studies through individual-based behavioral modeling, with the scope of identifying quantitative patterns of genetic drift giving rise to speciation[Gras et al. 2009]. In this research, the aim is investigating the criterion affecting sympatric speciation, and finding their relative importance, using an individual based evolving predator-prey ecosystem platform called “EcoSim” [Gras et al. 2009]. This study aims to answer two broad questions. The first question is: can we find any instances of sympatric speciation as a result of divergent eating behavior? If we can find such instances, the second question would be: can we elucidate the behavioral patterns observed in species, which give rise to sympatric speciation? To answer the first question we simulated a dual food resource version of EcoSim to allow the emergence of divergent eating behavior. The results of the simulation were analyzed to find if there might be any instances of sympatric speciation according to the criteria mentioned in the first chapter. We found some promising result showing the occurrence of sympatric speciation in 5 out of 20 separate runs of the EcoSim model. Using the results of the simulation, machine learning techniques were applied to derive the environmental and behavioral conditions which have the highest influence on sympatric speciation.

The first chapter starts with describing different points of view shared on the topic, beginning with a definition of the concepts of species and speciation, and ending with a discussion about the difficulty to observable natural cases and about the future prospects through simulation studies. In the second chapter, the EcoSim model is described briefly, beside a short description about the other existing computational models. In the third chapter, the modifications which were applied to EcoSim to make it suitable as a platform for investigating sympatric speciation are explained. In this chapter also the approaches for finding the instances of sympatric speciation are explained. The fourth chapter shows the results after applying the approaches proposed in the chapter 3. We show that many cases were observed where the sympatric speciation criteria described in the literature were fulfilled. In this chapter, we also show that some behavioral modifications occurred as a consequence of preferential resource usage to confirm that the existence of a second food resource is determinant for the emergence of sympatric phenomenon. And finally in the fifth chapter, machine learning tools are employed for finding the answer of the second question. Using several machine learning techniques, we extracted explicit rules that can predict with a very high accuracy the occurrence of sympatric speciation based on ecological factor observations, confirming that our method is able to discover very generic rules, which may later be used to structure empirical studies.

Chapter 2

SYMPATRIC SPECIATION

As it was already mentioned, sympatric speciation, from the Greek ‘same place’, involves the splitting of an ancestral species into two or more reproductively isolated groups without geographical isolation of those groups[Coyne 2007]. According to [Coyne 2007] sympatric speciation involves natural selection driving a population in two different directions at once. For example, a population of herbivorous insect may be selected to use two very distinct types of food, while they share a common area. Under strong condition for food selection, the population may be divided into two subpopulations, each specialized on a different resource, where the hybrids suffering a reduced fitness compared with the rest of population as they were not adapted to either resource. Then the speciation can occur under two possible types of isolation, which prevent gene flow between subpopulations. The first scenario is habitat isolation, which is, for example, where the insects mate exclusively on the resource they use, and the second scenario is sexual isolation, which happens when insects choose mates or individuals using the same resource. Both scenarios will result in preventing gene flow between subpopulations while they are living in a single area and as a result sympatric speciation can occur.

2.1 Definitions

2.1.1 Species

In order to initiate a discussion on speciation, it is important to define the focus of the study, a species, and its importance in a biological context. Carolus Linnaeus [Linnaeus 1758] first introduced the binomial naming system for species used to this day, which introduced the concepts of taxonomic rank to order organisms, which at the time were distinguished through morphological features. Fast forward in time, the combination of Darwinism and Mendelian genetics paved the way for the modern evolutionary synthesis, tying in natural selection with the laws on inheritance to explain the species concept [Cerghet 2013]. To this day, the concept attempts to reconcile macro evolutionary changes obtained from studying fossil records with micro evolutionary changes observed, often through bacterial stains. There continue to be distinctions in categorizing certain organisms, examples of which include asexual organisms [Turelli et al. 2001] due to unclear clade formations, morphologically identical species [Mayr 1996] that often differ highly in their genetic content(as it is the case for many protozoa and bacteria), as well as organisms within the same initial species population that have developed reproductive isolation [Turelli et al. 2001]. Reproductive isolation is the driving force behind speciation mechanism and more specifically behind sympatric speciation and is the focus of our research. Mayr [1996], points out an important idea, “The word ‘species’ conveyed the idea of a class of objects, members of which shared a set of defining properties.” The defining root of the word species should never change, rather the tools and methods we use to categorize organisms to uniquely fit a species category must continue to improve. The species concept continues to evolve itself, being most recently explained by a combination of traditional morphology, gene sequencing, and reproductive compatibility [Turelli et al. 2001].

Following a Darwinian train of thought, the question of why species exist as discrete categories of genetic storage, and whether there is any significance to it, is most simply answered by understanding that nature is a product of evolution driven by selection [Mayr 1996]. Coyne and Orr [Coyne and Orr 1998], outline three hypotheses to address the discrete species concept: 1) Species showing discontinuous, stable states of matter, 2) Species adapt to discontinuous ecological niches, 3) independent evolution occurs through gaps created by reproductive isolation. Of these, reproductive isolation can only be distinguished for sexual species, and may be categorized into two main groups, sexual vs. asexual reproduction. Reproductive isolation relies on the ideas that either divergent or disruptive selection works in context on allopatry or sympatry, leading to eventual selection isolation [Turelli et al. 2001]. To test the concept of reproductive isolation through sympatry, it is interesting to consider the degree of distinct speciation occurring when comparing sexual vs. asexual taxa. Asexual organisms are difficult to observe in large groups, although *Bacillus subtillis* bacteria have been found to form discrete sympatric clusters in a natural study performed in the American Dessert [Coyne and Orr 1998]. In reference to sexual reproduction, cases are discussed farther in greater detail. Mayr[1996], makes a compelling argument, that biological species as a defined system with purpose, allows for superior gene combinations adapted adequately to adverse environments through sexual reproduction, and only the abandonment of such reproduction would destroy such a system.

2.1.2 Ecological Speciation

Divergent selection is the driving force behind ecological speciation. This process occurs most often between separate populations and less often between subpopulations in diverse environments of the same initial species. As gene flow is blocked between the populations, natural selection combined with the genetic drift act upon traits leading to genetic isolation [Schluter 2001]. The concept of ecological speciation revolves around the environment, which includes the abiotic features affecting the species habitat, ranging from climate to the spatial range, as well as biotic factors centering on inter-specific and intra-specific interactions: foraging, predation, etc. Such speciation may arise indirectly, as species adapt to a new environment, or through competitive inhibition, they are forced to change their foraging strategy, leading to new morphological and physiological features and distinct behavioral traits.

Ecological speciation can lead to a variety of isolation outcomes, namely pre-mating isolation where individuals mate strictly in their preferred habitats, and as a consequence does not interact with other individuals with different ecologic preferences [Schluter 2001]. Post mating isolation may occur arising from gametic or zygotic incompatibility, and often hybrids' low fitness, leading to sterility or no viable offspring. This is further compounded by the Wallace effect, also known as Reinforcement, where two populations of the same species that were initially separated come back into contact.

If they underwent reproductive isolation, they can no longer interbreed to produce viable offspring. If they underwent partial reproductive isolation, they will produce hybrids, which may or may not be viable [Ollerton 2005]. A separate model for ecological speciation is centered on the by-product mechanism, where reproductive isolation is indirectly favored through selection and other traits driving genetic differentiation. Pre-mating isolation was observed by Dodd [1989] in species of *Drosophila pseduoobscura* and at the time attributed to “unknown mechanisms,” yet has since been described as a case for by-product mechanisms [Schluter 2001]. Allopatric speciation most likely occurs as a result of divergent selection through the by-product mechanism, whereas sympatric speciation is a consequence of disruptive selection, or ultimately as a result of the Wallace effect, where extreme traits are favored to interact with a set environment over intermediate traits. Given strong enough selection pressures and minimal gene flow, the geographic isolation phase may not be a requirement, alluding to a possible mechanistic action for sympatric speciation [Schluter 2001][Turelli et al. 2001].

2.1.3 Differentiating: Allopatric vs. Sympatric

When traditionally described in population biology studies, speciation through allopatric means is the first case scenario prescribed if a new species is to arise. By virtue, if there are no forces acting upon reproductive compatibility due to geographic barriers between separate populations, then reproductive incompatibility is the inevitable outcome and thus allopatric speciation as the pivotal mode of evolution [Turelli et al. 2001]. It is easy to imagine the types of selective pressures, that are placed on populations of species undergoing allopatry, whether through emigration, formation and changes of geographic landscapes, such as, mountain ranges, and lakes or ever-growing forced isolation, from human agricultural and civil undertakings. The process of adaptive radiation, which explains rapid speciation, when a few individuals from a population occupy a new habitat, coupled with novel selection pressures being faced, would be expected to result in different mating strategies, and subsequent changes in morphology. As these heritable variations begin to accumulate within the new population leading to morphological changes, given enough time, reproductive isolation from the parental population will be the outcome.

If a new species is to arise within a set environment, where most likely there are pressures for reproductive compatibility to persist, then there must be rather restrictive conditions present to allow for sympatric speciation to be observed [Via 2001]. To justify the process of speciation has taken place between two sympatric sister species, their environment and habitats must be scrutinized to verify that sufficient evidence is collected for both initiating the speciation process as well as its eventual completion, while fully succumbed to physical isolation [Bush 1994].

At present, there are two main methods for showing empirical evidence used to justify a speciation event as sympatric. When used jointly, the combined efficacy is often enough to distinguish between a scenario where this evolutionary process may have occurred, and instead exclusively justify it as the only plausible scenario [Via 2001]. The first line of empirical evidence pertains to assessing trends observed between the sister species on a phylogenetic scale and identifying relevant patterns that match those reported in literature. Molecular phylogenetics has paved the way for accurate calculations of speciation rates using DNA sequencing, yet inherent problems persist. Barraclough and Nee [2001], outline two prime issues: First, phylogenetic trees rely on forcing the species concept relationship into a clade to identify evolutionary patterns, which biologists are still in heavy debate over. Second, non-speciation events such as extinction of a species, which may lead to phenotypic variation being observed in the surviving sister species for habitat compensation, will affect the construction of phylogenetic trees [Barraclough et al. 1998]. The second line of empirical evidence used to corroborate a sympatric event, revolves around identifying distinct ecological and genetic conditions and behaviors that aid the process of speciation [Via 2001], which I have further outlined in the following section.

2.2 Behavioral and reproductive strategies leading to sympatric speciation

Many factors are believed to influence progression towards sympatric speciation. Sexual selection leading to mate choice, competition for habitat and trophic niche and preferential resource use are among the most discussed in literature [Thibert-Plante and Hendry 2011]. Sexual selection relies predominately on two factors, male display and preference for variants of that display by females [Arnegard and Kondrashov 2004]. [Takimoto et al. 2000] developed a model which showed that female preferences for particular types of males affects the rate of speciation, also female preference is able to override the cost that males incur for developing ornamentation in the context of speciation, and even small costs to female mate choice still maintain a speciation state. Nevertheless, strong disruptive selection for display traits requiring multi polymorphic loci would also require selective pressures on female preference towards these variants, which is highly unlikely, unless they develop in symmetry through a gradual process [Arnegard and Kondrashov 2004]. For these reasons, sympatric speciation is not believed to occur solely through sexual selective pressure. For example, East African Crater Lake Cichlids believed to have undergone sympatric speciation through sexual selection [Schliewen et al. 1994], are now being described as obtaining the pre-zygotic isolation mechanisms as a result of ecological disruptive selection [Arnegard and Kondrashov 2004].

Competition for habitat and niche are best exemplified using predatory animals. For example, the coexistence of two sympatric and morphologically similar bat species, *Rhinolophusaffinis* and *Rhinolophuspearsoni* was studied in a cave habitat [Jiang et al. 2008]. Diet analysis identified that both species of bat foraged on certain prey types exclusively, as well as an overlap in the diet. The overlap in the diet did not lead to competitive exclusion as expected due to preferential foraging in different microenvironments within the cave ecosystem. This exemplifies how coexistence of natural competitive predators can occur even in enclosed environments, through selective pressures to differentially exploit the same trophic niche, which likely lead to sympatric speciation.

The focus of our study centers on preferential resource use causing selective pressures towards sympatric speciation. It is believed that selection for speciation through ecological divergence is centered on empirical evidence showing fitness loss in producing hybrids, which have reduced efficiency in harvesting distinct resources compared to the parental species [Rice and Hostert 1993]. This selective pressure, acts as a driving force leading to reproductive isolation mechanisms, either by reducing the probability of mating between individuals in populations using different foraging strategies, or by promoting mating among those members of the same population [Lu and Bernatchez 1999]. Resource partitioning is most likely to occur when proper conditions are formed: relaxed inter-specific competition, increased intra-specific competition, as well as open niche environments to exploit. These can lead to ecological pressures developing discrete polymorphic organisms with shared phylogenetic histories, driving sympatric speciation. Some examples include Arctic Charr (*Salvelinusalpinus)*, with four identified sympatric morphs, Tiger Salamander (*Ambystomatigrinum*), and African Finches (*Pyrenestesostrinus*) with identifiable differences in beak morphologies, all developed to exploit differential resources [Skulason and Smith 1995].

Rice and Hostert [Rice and Hostert 1993], describe discrete resource polymorphisms as initial steps in the progress towards sympatric speciation. For such speciation to initiate through niche-specific adaptation, two distinct mechanisms are proposed. First, a homogenous environment splits a population in two based on selection for extreme phenotypes, secondly two parapatric populations experience differential selective forces also pulling towards two extremes. If the selection is working on traits essential for resource use, then subsequent reproductive isolation may occur. Note, while I have not discussed parapatric speciation, it is defined as a gradient between sympatry and allopatry, where a common living area exists for genetic transfer to take place between populations [Dingle et al. 2010]. Laboratory experiments indicate that reproductive isolation formed in this way through pleio-trophy, where one gene has heightened control over various phenotypic features, leading to polymorphisms, follows a step-wise fashion: First, there is exploitation of novel resources, this leads to decreased intra-specific competition, which in turn leads to diverse selection pressures adapting organisms preferentially to each niche via mutations, and finally eventual reduction in gene flow between subpopulations [Skulason and Smith 1995]. These experimental findings identify mechanisms of sympatric speciation through resource partitioning that we focused on investigating.

Sympatric speciation occurrence has been empirically studied and categorized most thoroughly in relation to foraging behaviors and specialization in feeding patterns. Under the ecological theory of adaptive radiation, resource-based divergent natural selection is mentioned as the ultimate cause of diversification [Lu and Bernatchez 1999]. Two major processes are inferred under this theory. The first mentions that phenotypical divergence of populations and species is driven by differences in the resource and competitive environment they experience. The second infers that reproductive isolation evolves as a consequence of the same forces causing phenotypical and ecological divergence. Url et al [1999],referencing to [Dobzhansky 1951], mentions that the theory of adaptive radiation predicts that selection will favor the development of mechanisms, either favoring mating among members of a given population or limiting reproduction between populations to reduce the probability of producing hybrids. According to this paper "a major argument for the role of ecologically divergent selection in speciation is the evidence for a fitness cost of producing hybrids of intermediate phenotype with reduced efficiency for resource exploitation relative to parental species, thus acting as a post-mating isolation mechanism."

2.3 Challenges in observing sympatry: Advantage of computational simulations

The main obstacle being faced in current evolutionary research on sympatric speciation is unifying empirical studies with proposed theoretical investigations.

Regardless of the presented theoretical models, biologists continue to debate about the fact that sympatry could appear in nature. These arguments range from placing sympatry in the realm of possibilities under sufficiently strong selective pressures [Kondrashov and Mina 1986], to those arguments which argue sympatry can only arise under restrictive conditions that are biologically improbable [Coyne 2007]. In order to identify sympatry as the mode of speciation for a biological system, a number of difficulties must initially be overcome.

First, the rapid divergence rate of sympatric speciation expected to occur in nature has been linked to the rate of adaptive radiation, falling within the range of 10^4 – 10^6 years [Bolnick 2004]. Continuous empirical observation for such a prolonged time is impossible and speculations are therefore, inherent in the scientific process. Secondly, a consensus must be reached on a set of conditions that encompass sympatric speciation. For the purpose of this research, we have decided to follow the requirements outlined by Bolnick and Fitzpatrick [2007], which pertaining to a set of species are as follows: (1) Largely or completely overlapping geographic ranges, (2) Complete divergence must be observed, partial divergence cannot be considered, implying reproductive isolation, (3) Clades defined by sympatric speciation must be due to sister species or be a part of a monophyletic endemic group, (4) The happenstance of allopatric/parapatric evolution in the groups past, must be considered highly unlikely, in a sense rejecting the alternative hypothesis. It is difficult to completely fulfill the aforementioned requirements through empirical observation in natural studies; species possess dynamic changes that range based on environment, habitat use, life history traits and mating strategies.

By using computational simulations we are able to experimentally control for many discrete factors and develop a model for the complex species interactions that give rise to sympatric speciation. This scientific approach takes advantage of computational resources allowing speciation events to be observed and analyzed on realistic time scales, as well as quantitative analysis of all pertinent information obtained in the process. We use the EcoSim program, an individual-based evolving-behavior model [Gras et al. 2009] discussed in full in the next chapter, associated with a bimodal distribution of resources to answer the following pertinent biological question: Can a speciation event occur in sympatry as a result of behavioral differences towards resource partitioning?

Chapter 3

ECOSIM Model

EcoSim was created to study biological and ecological theories. It can simulate a generic ecosystem with behaviors similar to those found in nature. There are few other such simulations but all of them are on a much smaller scale, compared with EcoSim. In the following sections, first a brief description of the other existing models is presented, and then the EcoSim model specifications are introduced.

3.1 Other existing models

One such previous models is Echo, which is a “genetic ecosystem model in which evolving agents are simulated in a resource-limited environment.”[Hraber et al. 1997]. In this system, each agent replicates itself with possible mutation when it has acquired enough resources to copy its genome. The agents can acquire resources with interaction with other agents (combat, trade or mating) or from the environment. The author claims that this mechanism for endogenous reproduction is much closer to the way fitness is addressed in natural setting than fitness functions in genetic algorithms.

Polyworld is another software developed by Larry Yaeger [1994] to evolve Artificial Intelligence through natural selection and evolutionary algorithms. It displays a graphical environment in which a population of trapezoid agents search for food, mate, have offspring, and prey on each other. The population is typically only in the hundreds, as each individual is rather complex and the environment consumes considerable computer resources. In this model, each individual makes decisions based on a neural network which is derived from each individual's genome. The genome determines the individuals’ size, speed, color, mutation rate and a number of other factors and is randomly mutated at a set probability, which are also changed in descendant organisms.

Tierra[Thearling and Ray 1994] is another computer simulation developed by Thomas S. Ray in the early 1990s in which computer programs compete for central processing unit (CPU) time and access to the main memory. In this context, the computer programs in Tierra are considered to be evolvable and can mutate, self-replicate and recombine. Tierra has been used to experimentally explore the basic processes of evolutionary and ecological dynamics. As there is no explicit, or exogenous fitness function built into the model the authors claim that this may allow for more "open-ended" evolution, in which the dynamics of the feedback between evolutionary and ecological processes can itself change over time.

Avida is another artificial life software platform to study the evolutionary biology of self-replicating and evolving computer programs (digital organisms) [Ofria and Wilke 2004], which was inspired by the Tierra system. Unlike Tierra, Avida assigns every digital organism its own protected region of memory, and executes it with a separate virtual CPU. By default, other digital organisms cannot access this memory space, neither for reading nor for writing, and cannot execute code that is not in their own memory space. A second major difference is that the virtual CPUs of different organisms can run at different speeds, such that one organism executes, for example, twice as many instructions in the same time interval as another organism. The speed at which a virtual CPU runs, is determined by a number of factors, but most importantly, by the tasks that the organism performs: logical computations that the organisms can carry out to reap extra CPU speed as a bonus.

Mark Bedau and Norman Packard developed a statistical method of classifying evolutionary systems and in 1997, [Egri-Nagy and Nehaniv 2003]applied these statistics to Evita, an Artificial life model similar to Tierra and Avida, but with limited organism interaction and no parasitism, and concluded that "Tierra-like systems do not exhibit the open-ended evolutionary signatures of naturally evolving systems.” [Egri-Nagy and Nehaniv 2003]

3.2 EcoSim Model Specifications

The approach for simulating individuals’ behavior, which was used by Gras et al [2009] in EcoSim for the first time in a large scale individual-based evolutionary process, is to use a Fuzzy Cognitive Map (FCM) as the behavioral model for the individuals. The FCM enables individuals to perceive their environment and to choose their action based on perceptions. The FCM of each agent, being coded in its genome, allows the evolution of the agent behavior through the epochs of the simulation [Gras et al. 2009a]. The notion of species is also implemented in such a way that species emerge from the evolving population of agents. To our knowledge, EcoSim is the only system that allows the modeling of links between behavior patterns and speciation without any exogenous fitness function. A notable amount of data can be produced by the model, including the number of individuals, level of energy by individual, choice of action, age of the individuals, and average FCM associated with each species, which allows for numerous investigations on macro-evolutionary processes.

“EcoSim as a virtual ecosystem has shown coherent behaviors of the whole simulation with the emergence of patterns also observed in existing ecosystems providing a general framework for the study of several specific ecological problems ”Several studies have been already done using EcoSim” [Gras et al. 2009a]. Devaurs et al. [2010]have shown that the behavior of this model is realistic by comparing the species abundance patterns observed in the simulation with real communities of species. Furthermore, the complexity has been evaluated [Farahani 2010] and the chaotic behavior [Golestani et al. 2010] with multi-fractal property [Golestani et al. 2011] of the system, have been proven as it has been observed in real ecosystems. In [Golestani and Gras 2012] the effects of small geographic barriers on the speciation in EcoSim are measured. EcoSim can be also used in studying important phenomena in nature such as speciation [Mashayekhi and Gras 2012], extinction [Sedehi 2012], and sexual selection mechanism.

3.2.1 Individuals

There are two types of individuals in EcoSim, predator and prey. Each individual possesses several characteristics such as: age, minimum age for breeding, speed, vision distance, level of energy, and amount of energy transmitted to the offspring. Energy is provided to the individuals by the resources (food) they find in their environment. Prey consumes grass, which is dynamic in quantity and location, whereas predator hunts for prey individuals. Each individual performs one unique action during a time step, based on its perception of the environment. Each agent has its own genome that codes for its FCM and its behaviors are determined by the interaction between the FCM, and the environment.

Individuals gain energy by eating one unit of food (grass for prey, and meat for predator)and for each action, they spend some energy depending on the action (e.g. breeding, eating, running) and on the complexity of their behavioral model (number of existing edges in their FCM).

Individuals live in a world made up from 1000 by 1000 cells, thus in this system local policies can be simply enforced. For example, eating can be done only if the food is in the same cell as the individual. The system goes through discrete time steps, in each time step every individual can percept, act and as a result of acting, change the environment. Each Individual has some properties, mostly physical capabilities, like energy and age. Each individual has an FCM which is used as the behavioral model of the individuals, allowing for observing divergent behaviors among different individuals. The FCM, which is represented by a matrix, is coded in the genotype of the individual.

The system is evolving. Mating occurs if two individuals live in the same cell, have a minimum age, have a minimum level of energy, are genetically close enough and both have chosen the Reproduce action among different actions that they can do. When a new offspring is created, it is given a genome which is a combination of the genomes of its parents with some possible mutations.

The system has a speciation mechanism which makes a species split if the members of the species are not genetically similar enough.

3.2.2FCM

An FCM is a graph, representing relations among some elements. Each node is a concept, and each edge demonstrates the influence of a concept onto another.

A positive weight associated with an edge corresponds to an excitation of the destination concept from the source concept, whereas a negative weight is related to an inhibition (a zero value indicates that there is no influence between the two concepts). The influence of the concepts in an FCM with n concepts can be represented in an n×n matrix.

A number is associated with each concept, called the activation level of the concept. Activation levels are updated at each time step, using the current activation level and the weighted some of the activation levels of other concepts affecting this concept transformed by a non-linear function. Figure 3.1 The FCM (behavioral model) of Prey Individuals in the standard EcoSim model. The width of each edge shows the influence value of that edge. Color of an edge shows inhibitory (red) or excitatory (blue) effects.

The FCM is a matrix, which can be represented as a graph which contains a set of nodes c, each node c_i being a concept, and a set of edges i, each edge c_{ij} representing the influence of the concept c_i on the concept c_j. A positive weight associated with the edge i_{ij} corresponds to an excitation of the concept c_j from the concept c_i, whereas a negative weight is related to an inhibition (a zero value indicates that there is no influence of c_i on c_j).

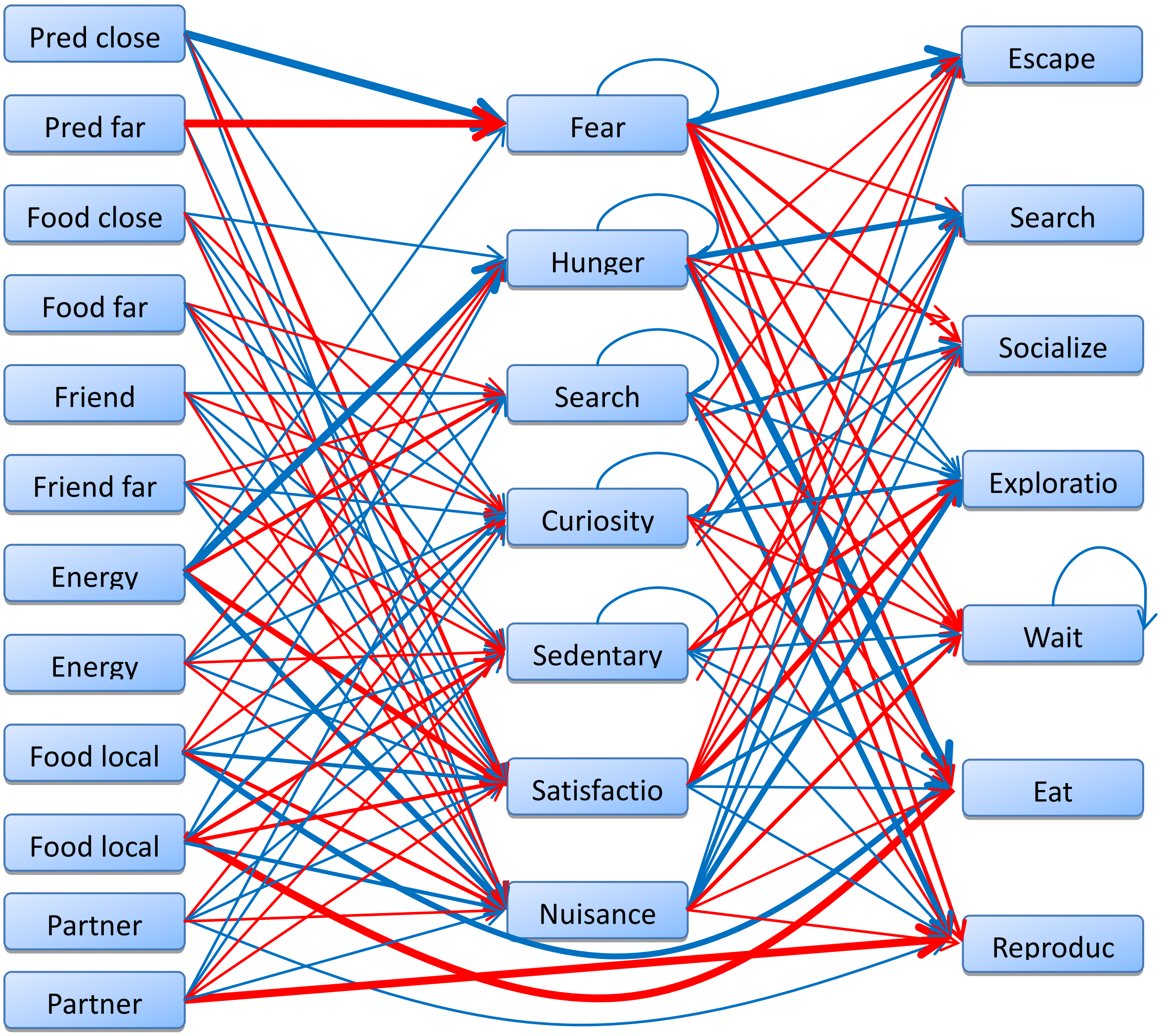
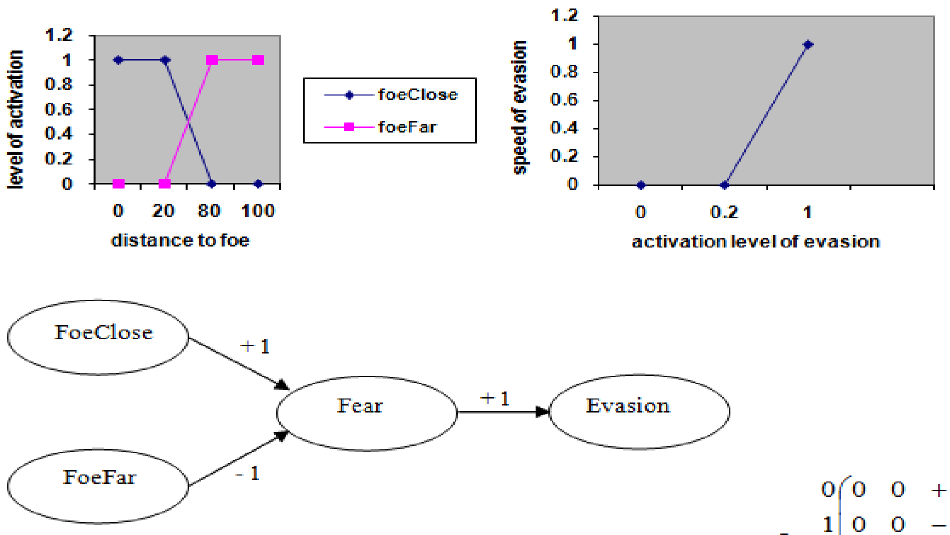


Figure 3.1 The FCM (behavioral model) of Prey Individuals in the standard EcoSim model. The width of each edge shows the influence value of that edge. Color of an edge shows inhibitory (red) or excitatory (blue) effects. The FCM enables individuals to percept their environment and based on their perception choose their next actions.

3.2.3 Concepts

Three different kinds of concepts are defined in the system, Sensitive, Internal and Motor. Sensitive concepts are set by a mapping from observation of the environment to a perception. At initialization, the Sensitive concepts affect Internal concepts and Internal concepts affect Motor concepts respectively but evolution can add edges between any kind of concepts allowing some complex feedback loops to emerge.

The activation level of a Sensitive concept is computed by performing a fuzzification of the information the individual perceives in the environment. For an Internal or Motor concept C, the activation level is computed by applying the de-fuzzification function on the weighted sum of the current activation level of all the concepts having an edge directed toward C. Finally, the action of an individual is selected based on the maximum value of the Motor concepts' activation level. Activation levels of the Motor concepts are used to determine the next action of the individual and their amplitude. As a very simple example, **Error! Reference source not found.** shows the three layers discussed above.



$$L = \begin{array}{c} \begin{array}{c} 0\\ 1\\ 2\\ 3 \end{array} \begin{pmatrix} 0 & 0 & +1 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 0 & +1\\ 0 & 0 & 0 & 0 \end{pmatrix} \\ \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \end{array}$$

Figure 3.2 An FCM for detection of foe (predator) and decision to evade with its corresponding matrix (0 for 'Foe close', 1 for 'Foe far', 2 for 'Fear' and 3 for 'Evasion') and the fuzzification and defuzzification functions. The closer the foe is, the more frightened the agent is. Depending on the foe distance and the fear level the agent will decide to evade or not. The more frightened the agent is, the faster it will evade.

Chapter 4

Modeling Sympatric Speciation using EcoSim Model

As it was mentioned before, EcoSim consists of two different kinds of agents, prey and predator, which along with a source of food (Grass), form the food chain in **Error! Reference source not found.** (left). As discussed in section 1.1.3, resource-based divergent natural selection is mentioned as the ultimate cause of diversification, under the ecological theory of adaptive radiation, therefore, to study sympatric speciation as a result of divergent eating behavior, it is required to have more than one source of food for the prey individuals. For this purpose, a new source of food for the prey has been added to the system, and the food chain is changed as in the **Error! Reference source not found.** (right).

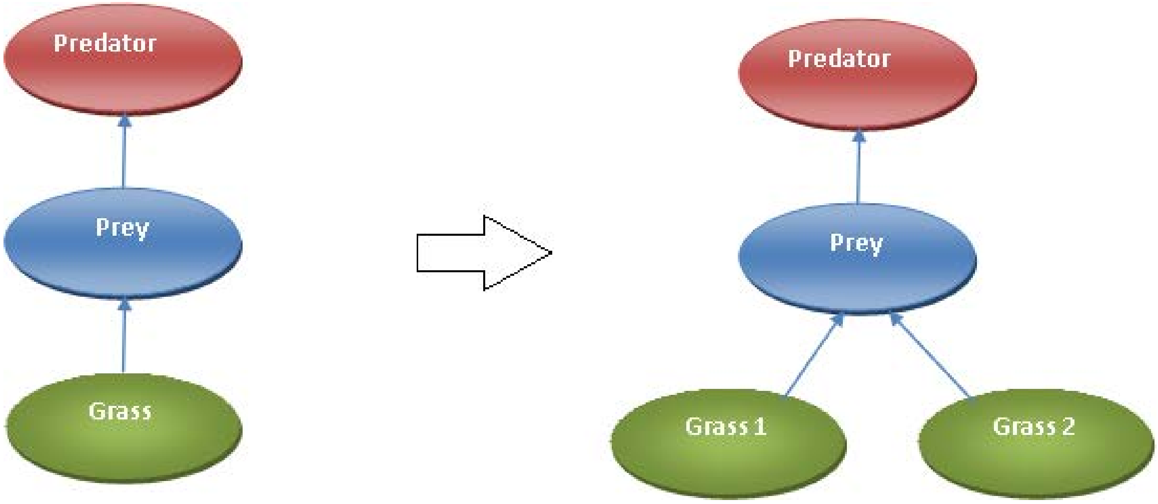


Figure 4.1 The Food Chain in the Standard EcoSim (left), and dual resource EcoSim (right). A second resource is added for the prey to allow the emergence of divergent eating behavior.

In the original FCM of the prey individuals, there are six concepts, which are directly related to the food consumption by the prey, four Sensitive concepts including ‘*FoodClose*’, ‘*FoodFar*’, ‘*FoodlocalHigh*’, and ‘*FoodLocalLow*’ and two Motor concepts, including ‘*SearchForFood*’, and Eat. In the initial FCM, two Sensitive concepts out of the four mentioned (‘*FoodLocalHigh*’ and ‘*FoodLocalLow*’) in addition to having an effect on the Internal concept Hunger, ‘*SearchPartner*’, Curiosity, Sedentary, Satisfaction, and Nuisance, have also direct effect on the Motor concept Eat. The two Motor concepts ‘*SearchForFood*’ and Eat are affected by all the internal concepts on the initial FCM. For avoiding any initial bias for the new food resource, it is required to add four new Sensitive concepts of *‘FoodClose2’*, *‘FoodFar2’*, *‘FoodLocalHigh2’*, and *‘FoodLocalLow2’* to the FCM of the Prey. In addition, the effects these new Sensitive concepts have on the Internal concepts are made identical to those of the Sensitive concepts related to the first food resource by copying the edges going towards them. Two new Motor concepts ‘*SearchForFood2’* and ‘*Eat2*’ are also added to the FCM, and the edges from Internal concepts and Sensitive concepts are also a copy of those that go towards ‘*SearchForFood1’* and *‘Eat1’*. **Error! Reference source not found.** shows a small part of the prey individuals FCM focusing on the eating action. The left-hand side graph is related to the standard EcoSim, and the right-hand side graph shows the same part of the FCM after adding six new concepts (in red).

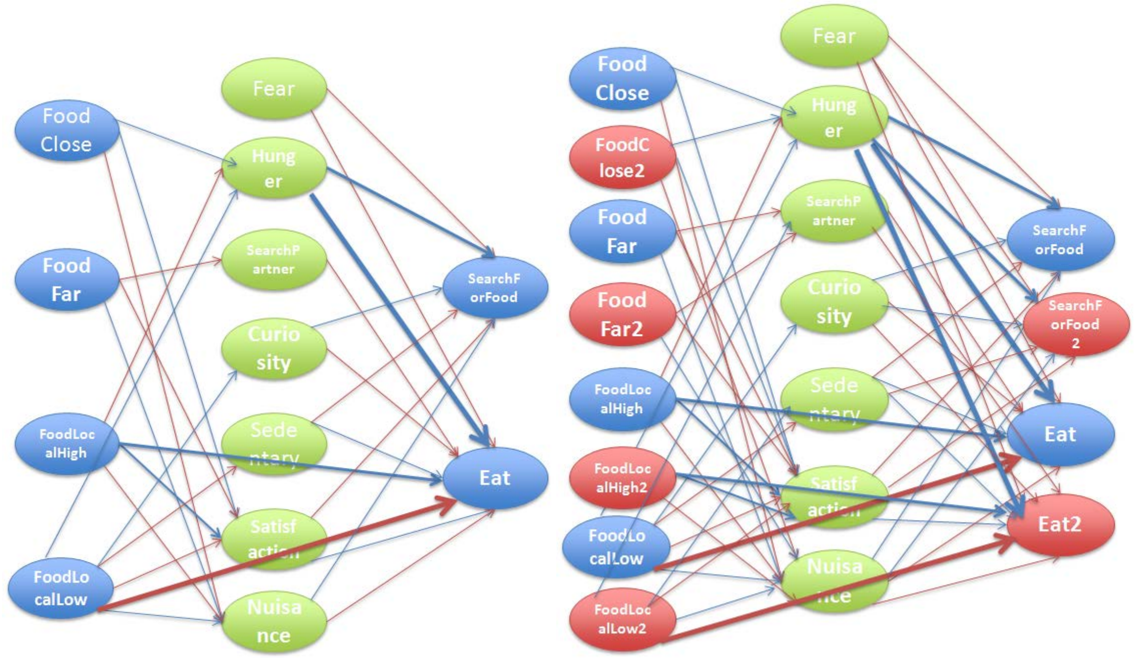


Figure 4.2 left. A small part of the standard EcoSim’s prey individuals’ FCM, which is related to eating actions. Right: Six new concepts are added to the prey individuals’ FCM in dual resource version of the EcoSim to allow the prey to detect and consume the second source of the food.

Error! Reference source not found. represents the complete initial FCM of the prey individuals after adding an extra source of food.

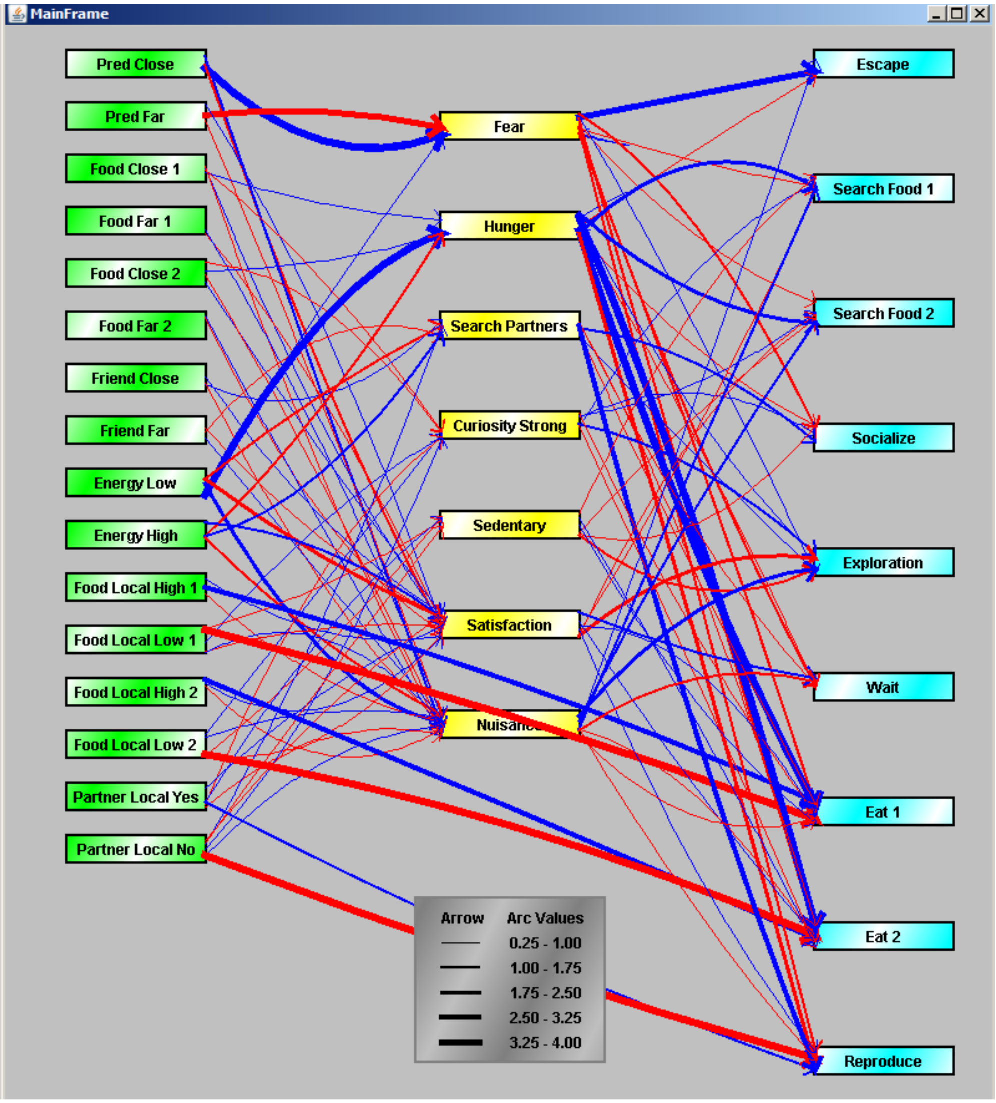


Figure 4.3 The initial Prey's FCM including concepts and edges for the dual resources version of the EcoSim. The width of each edge shows the influence value of that edge. Color of an edge shows inhibitory (red) or excitatory (blue) effects of the source concept on the destination concept.

The new food resource that is added to the ecosystem, has some characteristics that can be customized, which allows to make the two food resources different. These characteristics are described in the Table 4.1.

Parameter	Description
ValueGrass	energy value for a consumed grass
MaxGrass	maximum number of grass in a cell
SpeedGrowGrass	speed of growing grass
ProbaInitialGrass	initial probability of grass per cell
ProbaGrowGrass	probability of diffusion of grass

Table 4.1 The characteristics of the food resource for the prey individuals.

4.1 Approaches

A set of runs was executed to observe the effects of resource partitioning on speciation. As expressed in natural studies, sympatric speciation is a rare observance at best, so a high series of runs will allow us a greater sample size to evaluate. We allow the simulations to run for approximately 25,000 time steps. We begin to evaluate the effects of speciation from approximately time steps 15,000 – 20,000, allowing the simulation and populations of species enough time to stabilize. As discussed at the first chapter, four criteria should be met in order to consider a speciation event as having a sympatric origin. Each criteria and subsequent strategy which is used to verify the criteria are presented in Table 4.2. The simulations were run over a time span of a few months using the joint computational resources available through SHARCNET[2], until the appropriate amounts of time steps were completed, and all necessary data was stored individually for each simulation.

---

[2] This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network SHARCNET:www.sharcnet.ca) and Compute/Calcul Canada.

Criterion	**Strategy**
1. Sister Species	Phylogenetic analysis
2. Complete divergence	Ratio of reproductive events leading to hybrid offspring
3. Overlapping geographic ranges	Calculating average distance of all individuals between sister species
4. Allopatric/Parapatric alternate hypothesis	Reject based on culmination of phylogenetic tracking and biogeographic data

Table 4.2 Sympatric speciation required criteria and chosen strategy for the verification of each criterion.

4.2 The first step: finding the runs in which Sympatric speciation happenstance is more likely

Resource preference among sympatric species which coexist in overlapping habitats has been studied in nature. Examples range from: Terrestrial habitats, including sympatric bat species of Rhinolphusaffinis and Rhinolphuspearsoni [Jiang et al. 2008], and resource host-mediated selection by Hawthorne fly's [Feder and Filchak 1999], aquatic habitats cohabited by arctic charr, Salvelinusalpinus, as well as aerial habitats shared by African Finches, Pyrenestesostrinus[Skulason and Smith 1995].

After completion of the runs, the information about all the individuals and species, such as all their actions, their breeding information, all the information about their FCM or behavioral model, and a complete set of information about their environment, such as the geographical location of the individuals or the food abundance distribution in the environment are available. Having these complete sets of information allows us to analyze them to detect whether the four mentioned required criteria for the happenstance of sympatric speciation hold, and examine the occurrence or non-occurrence of this phenomenon.

A first step for finding the runs in which the occurrences of sympatric speciation due to divergent eating behavior are more likely, would be finding those runs in which there are some species, which are specialized to a source of food, or in other words some species which show more preferences to a source of food than to the other. If we can find such species, showing preferential behavior in two different resources while they are co-existing during a period of time, then we can examine the four required criteria on these species to figure out the occurrence of sympatric speciation.

4.2.1 Species Categorizing algorithm

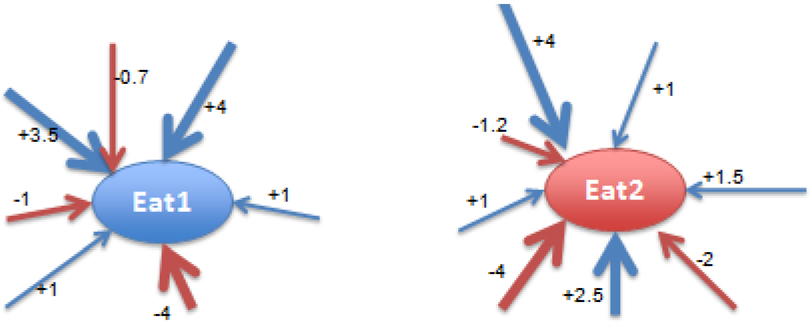
To find out whether some species are showing preferential behavior to a specific food resource than to the other one, we developed two different studies: one study about each species' average behavioral model, and another one about each species' real actions. In the first study, we look at the behavioral model (FCM) of each species' individuals to find out whether they are more likely to consume one special source of food more than the other one, based on their FCM. In the second study, the real actions of individuals in each species and their perceptions of the available resources nearby are taken into account to judge whether they are showing any preferential behavior to one resource or not. The two mentioned approaches are explained in the following subsections. Each of these approaches will help us to categorize the species, into three different groups, based on their resource preferences. Group one, the species which are more likely to choose Grass1 rather than Grass2, group two, which are more interested to consume Grass2, and group three, which are species without any specific preferences on any source of food, who simply choose the closest available food resource.

After applying each of the two following approaches, we would have a set of species, which are classified into three groups. The next step would be trying to find sets of two species, each this one specialized on a different resource, and verifying whether the four mentioned required criteria for sympatric speciation hold for them or not.

4.2.2 Species Categorizing algorithm based on FCM behavioral model (FCM-Clustering)

For the first study, as it was mentioned, we developed an algorithm which categorizes species based on their behavioral model or FCM. To determine whether a species shows any preferential behavior to a special source of food, we calculated the weighted sum of all the edges, which have influence on the Motor concept Eat1 and Eat2 separately, and based on the results, we categorized the species to three groups. We defined a threshold for the difference between the values associated to the incoming edgesforEat1 and Eat2 concepts. If the difference between the weighted sums of Eat1 and Eat2 in the FCM of the species was not significant enough, those species were assigned into the group three, otherwise they were assigned to either group one if the value associated to Eat1 was greater than the value associated to Eat2, or to group two in the opposite case. A difference smaller than about the 10 percent of the highest observed differences between the weighted sums of Eat1 and Eat2 actions among all the prey species in all the dual resource submitted runs was considered as not significant enough, which allows for ignoring the species which do not show to have a significant eating preference, and find the species with the strongest specialization behavior according to their behavioral model.

Error! Reference source not found. represents an example of incoming edges to the concepts Eat1 and Eat2 in an FCM. The algorithm, which is implemented in C programming language, parses all the FCM files to find the category in which each species belongs to, and to find out the distribution of the population belonging to each group. For a Run with around 25,000 time steps, it takes around 10 to 20 minutes for the program to complete the calculation.



Weighted sum of Eat1= (-4)+(+1)+(+4)+(-0.7)+(+3.5)+(-1)+(+1)= 3.8

Weighted sum of Eat2=(-4)+(+2.5)+(-2)+(+1.5)+(+1)+(+4)+(-1.2)+(+1)= 2.8

Difference= 3.8 – 2.8=1 > Threshold=0.5 ----> The species Belongs to Group1

Figure 4.4 The weighted sum of all the edges which have a direct influence on eat1 and eat2 actions were calculated to decide which action is the dominant action at the behavioral model of the species.

After categorizing all species, this algorithm starts from the beginning of the simulation and counts the number of individuals belonging to each group in each time step. This will allows us to find out if there might be some species specialized to Food1 and some other species specialized to Food2, which are co-existing, and perform the required tests on these species to find out whether sympatric speciation occurrence is likely or not.

For example, **Error! Reference source not found.** represents the resource preference distribution of all the population of prey individuals for one of the runs of the simulation, as an example of the output of this algorithm (The final results, after applying this algorithm on all the runs, are presented in the Chapter 4, where the results are discussed). The horizontal axis is representing the time steps, and the vertical axis is representing the percentage of prey population belonging to each group. As it can be observed, starting from around time step 17400, we have a notable amount of population belonging to group one and group two. This will gives us an insight to search through the species around the mentioned time step, and examine the four required criteria on those species.

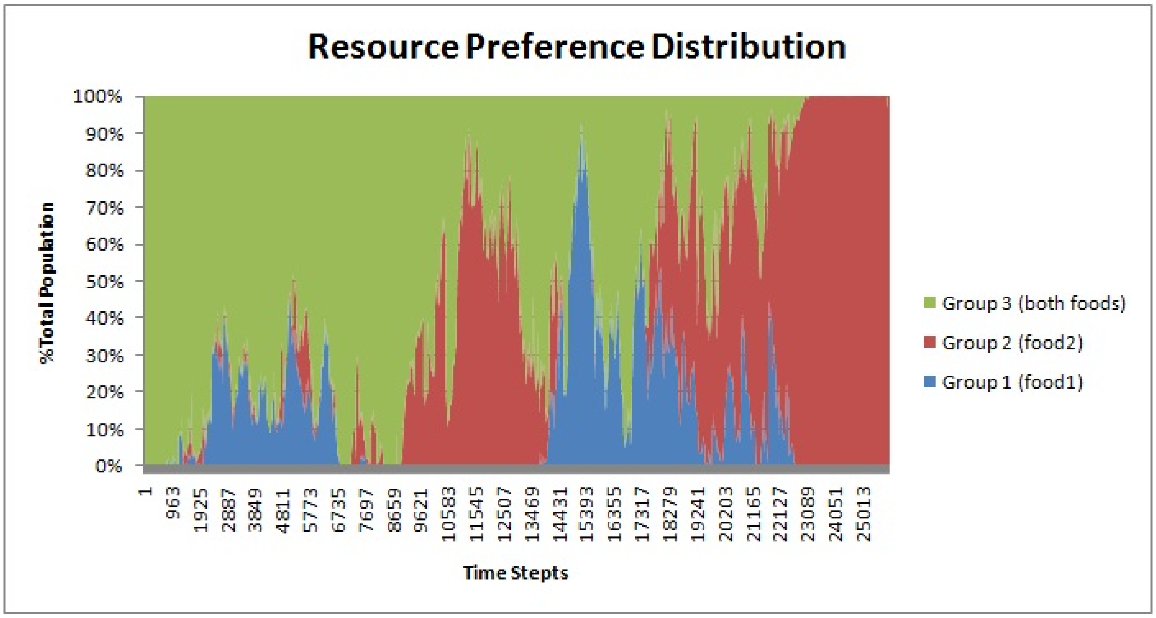


Figure 4.5 Resource preference distribution of the prey population based on the FCM, for Food1 (blue), Food2 (red), and Both foods (green). Each individuals preference is calculated for the duration of the simulation based on their behavioral model. The horizontal axis is representing the time steps, and the vertical axis is representing the percentage of prey population belonging to each group.

Error! Reference source not found. represents the difference between the importance of Eat1 and Eat2 actions, based on the weighted sum of all the edges, which have an influence on these motor concepts for the three groups. These difference calculations were used to categorize species into the three groups and then to find the percentage of population belonging to each group. It can be observed from the figure 9, that the average measured differences for the three groups, are well separated from each other. The average and the standard deviation of the measured differences are calculated and presented in the Table 4.3 confirming that these three groups are well separated from each other.

Difference Of Eat1 minus Eat2	Group1	Group2	Group3
Average	2.411524	-2.15045	-0.03114
Standard Deviation	1.088271	0.943909	0.166314

Table 4.3 The Average and the Standard Deviation of the differences between the weighted sum of Eat1 and Eat2 actions in the FCM of the species belonging to each group.

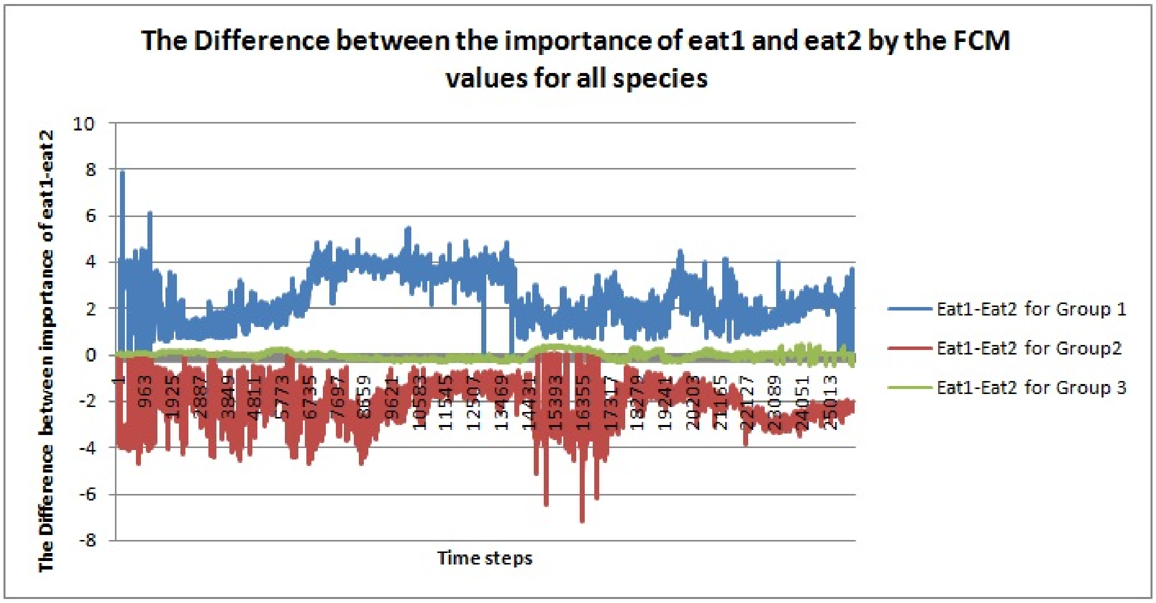


Figure 4.6 The difference between the importance of Eat1 and Eat2 actions in the FCM of prey individuals by calculating the total weighted sum of the edges influencing eat1 and eat2 concepts for the three groups. These difference calculations were used to categorize species into the three groups for finding the percentage of population belonging to each group.

4.2.3 Species Categorizing algorithm based on Individuals' actions and perceptions(Action-Perception Clustering)

In the second study, we assigned each species to one of the three previously mentioned groups, based on their real behavior, rather than their behavioral model. A simple logical model is used for this purpose. For each species, we look at their rate of Eat1 and Eat2 actions and at the same time, we take into account their perception from the environment, regarding the availability of local resources nearby.

Five simple logical rules are applied to each species' rates ofEat1 and Eat2 actions, and the rate of their perceptions about the available Food1 and Food2. The five rules are as follows:

1- if Eat 1 >> Eat 2 and (Foodlocalhight1 ~= Foodlocalhigh2
or Foodlocalhight1 << Foodlocalhigh2) => Group 1
2- if Eat1 << Eat2 and (Foodlocalhight1 ~= Foodlocalhigh2
or Foodlocalhight1 >> Foodlocalhigh2) => Group 2
3- If Eat1 ~= Eat2 and (Foodlocalhight1 >> Foodlocalhigh2) => Group 2
4- If Eat1 ~= Eat2 and (Foodlocalhight1 << Foodlocalhigh2) => Group1
5- Any other remaining species => Group 3

The symbol "<<" is used to indicate "much smaller than", and similarly the symbol ">>" is used to indicate "much greater than". The symbol "~=" is used for showing "approximate equality". At the first rule, we look at the rate of eat actions for the individuals of each species, and if the rate of Eat1 is much greater than the rate of Eat2, while the abundance of food2 is much higher than food1, or they are approximately at the same abundance, it can be concluded that the individuals in this species is more interested to consume Food1 rather than Food2 and belongs to group 1, as despite the high availability of the other source of food, it tends to consume food1 more often. A threshold was used for the minimum required difference between the rate of Eat1 and Eat2 actions to be able to claim that the rate of one of the actions is much greater than the other one. This threshold was selected such that, it insures that the rate of one action should be almost twice higher than the other one to be counted as much greater, or they will be considered approximately equal. Similarly, another threshold was used for the difference between available resources, to find out whether their abundances are approximately equal, or one of them is much more available than the other one.

The other rules also can be explained similarly, as in the second rule, when the rate of Eat2 action is much greater than Eat1 action, while the abundance of Food1 nearby is much greater or almost equal to the abundance of Food2 nearby, it can be concluded that the species belongs to group2, the group of species, which prefer to consume Food2 more than Food1.

The third and fourth rules, are related to the situations where the rate of eat actions are almost equal, while the abundance of one resource is much greater than the other one, which means that the species is more interested in the less available source of food, and that despite the shortage of the desired food, they are still searching for it and consume it at the same rate of the more available resource.

Finally, for the final rule, it can be said that, the species which were not assigned to any group based on the four previous mentioned rules, are not specialized on any specific resource, or are not showing any preferential behavior and therefore they belong to the third group.

For example, Error! Reference source not found. represents the resource preference distribution of all the population of prey individuals based on their real eating behavior and their perception about the available resources at their environment for the run discussed in the Section 3.2.2, as an example of the output of these species categorizing algorithm (more details on the results of applying this algorithm on all the runs are presented in the Chapter 4.) The horizontal axis represents the time steps, and the vertical axis represents the percentage of prey population belonging to each group. As it can be observed, starting from around time step 21000, we have a notable amount of population belonging to group one and group two. This will gives us an insight to search through the species around the mentioned time step, and examine the four required criteria on those species.

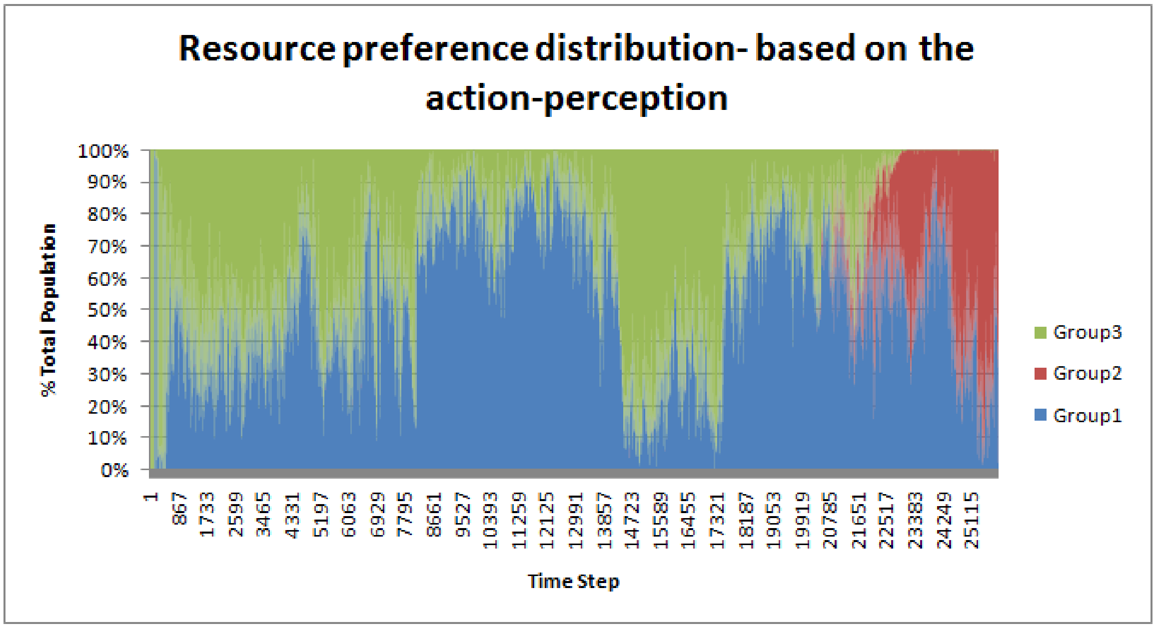


Figure 4.7 Resource preference distribution of the prey population based on the action-perception clustering method, for Food1 (blue), Food2 (red), and both resources (green). Each individuals preference is calculated for the duration of the simulation based on their real eating behavior and their perception about the local food available. The horizontal axis represents the time steps, and the vertical axis represents the percentage of prey population belonging to each group.

4.3 Verifying the first criterion: Sister Species

The first criterion, which needs to be considered for any set of two species, is determining whether they arose as a monophyletic endemic group, sister species, or neither of the two. By the nature of our simulation, where all prey and predators arise from a single lineage that is created as the simulation begins in a fixed geographic area, we chose not to consider monophyletic endemic lineages. For future studies, we can observe population spatial dynamics to experimentally determine regions on the surface of the world for which we can consider that two populations are too far away to consider their lineages endemic.

Two major problems are faced when creating phylogeny trees for natural studies. First, an issue arises with sampling, where all individuals within a species should be theoretically sampled to create a phylogenetic lineage with maximum accuracy, taking into account the exact time of the speciation event. This is in general impossible given the times scale involved and the rarity or difficulty of studying certain species. By leaving out individuals of a species, sampling bias is introduced, and assumptions become an inherent process of generating a tree. Another issue arises once a given individual considered to belong to a phylogeny is studied and expected to reflect the evolutionary traits attributed to its lineage. Subjective bias can be introduced by the researcher depending on the taxonomic approach they use [Barraclough and Nee 2001]. These sampling and taxonomic errors are avoided in our simulations, as every unique individual belonging to each species is tracked over evolutionary time, and their data is stored for future use. This allows us to determine the exact moment a speciation event occurs, in order to design exact trees for phylogenetic analysis.

Focusing on the sister species concept, we constructed phylogeny trees to identify species whose behavioral model expressed a tendency for one of the two grass resources. This allowed us to categorize species on a phylogeny tree differentially based on a tendency for resource preference.

From the previous step, we have a list of species, which are categorized into three groups. To fulfill the first required criteria, we need to find a set of sister species, such that one of them belongs to group1, and the other one belongs to group2. If we can find such set of species, then we can test the other required criteria on this set of species.

To build a phylogeny tree, we developed a code which reads the information about all the prey individuals saved in the files which are called "MinSave Files". During the simulation run, for each time step, a MinSave file is being saved, containing all the information about all the prey and predator species, such as their parent specie ID, and new speciation events. The phylogeny maker program, first reads all the MinSave files to find the parent species ID and the children species ID and the time steps in which speciation happened. It also finds the life span of each species in order to set the length of their branch in the phylogeny tree. The information regarding the tree is being saved in the memory while keeps track of the species ID of parent species, The species ID of the children species, the time step of their birth and their life span. This program also loads the information about the group category of each species, which is already saved in a result file from previously mentioned species categorizing algorithm. For visualizing the phylogeny tree, we were looking for a way to visualize it with a pre-defined color for species belonging to each group, to make it easier to spot sister species belonging to different groups. We used the graphical editor for phylogenetic trees called TreeGraph [Stöver and Müller 2010], which accepts the input in XML schema format, which enables us to include many specifications to each branch such as color, weight, length, etc. The phylogeny maker program, writes the information about the tree in XML format and the result file is opened by TreeGraph software.

As already mentioned, we had two approaches for assigning species into each group, one based on their FCM, and the other one based on their actions. Therefore, we can assign colors to the branches of the phylogeny tree, based on the two different approaches in separate phylogeny trees.

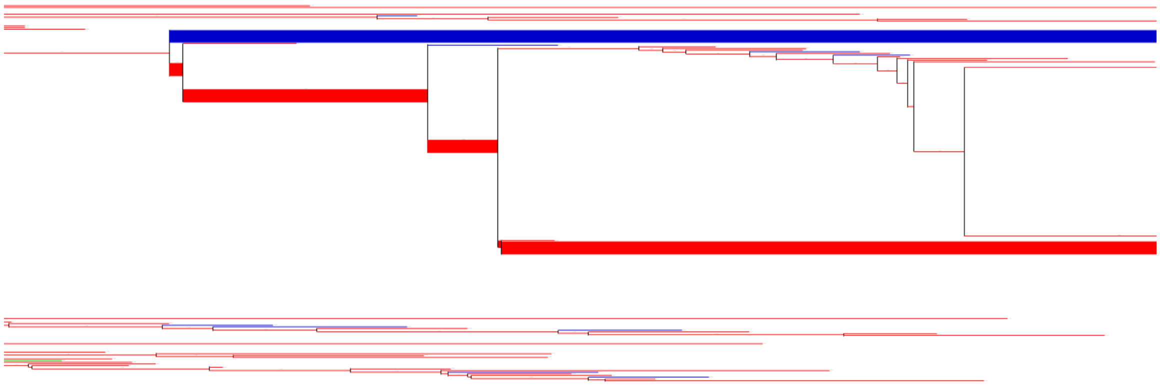


Figure 4.8 A truncated phylogeny tree originating from one of the species, showing preferential resource use, blue for Food1 and red for Food2. The bold lines identify the branches belonging to a set of candidate sister species showing preferential resource behavior.

Error! Reference source not found. shows a truncated phylogeny tree, originating from one of the species existing at time step 17400, categorized based on the real behavior of the species. As it can be observed, a set of sister species were found, exhibiting tendencies toward differential behaviors arising from preferential resource use, bold blue for Food1, and bold red for Food2, each existing for well over 400 time steps before extinction. All other branches shown in light blue and light red, are other species with a short life span.

As the phylogeny tree of each run would be extremely huge, representing thousands of species and speciation events, it would be almost impossible to spot sister species with different resource preferential behavior manually. For this purpose, an algorithm was developed to find all the instances of sister species using the several filtering criteria. One requirement is that one of the species should be a member of group1, and the other one should be a member of group2. Another requirement is that the species should be able to survive for at least 100 time steps, as there are many cases of species with small population in the simulation, which get extinct after less than 50 time steps, and which should be considered as noise. We selected the minimum life span of 100 time steps requirement to select the strongest species, which live well over the average life span of all the species. For example, Figure 4.9 shows the histogram of the life span frequency for the prey species in run s33. The horizontal axis shows the length of the life span, and the vertical axis shows the number of the species having that life span. As it can be observed there are many species with a very small life span (less than 100 time steps), which should be filtered as they are considered as noise in the speciation mechanism. The Average life span for the prey species in our dual resource ecosystem is 84 time steps, with the standard deviation of 195. Selecting the species with the life span of greater than 100 allows us to study the behavior of the species, for which the life span is well over the average life span of the prey species.

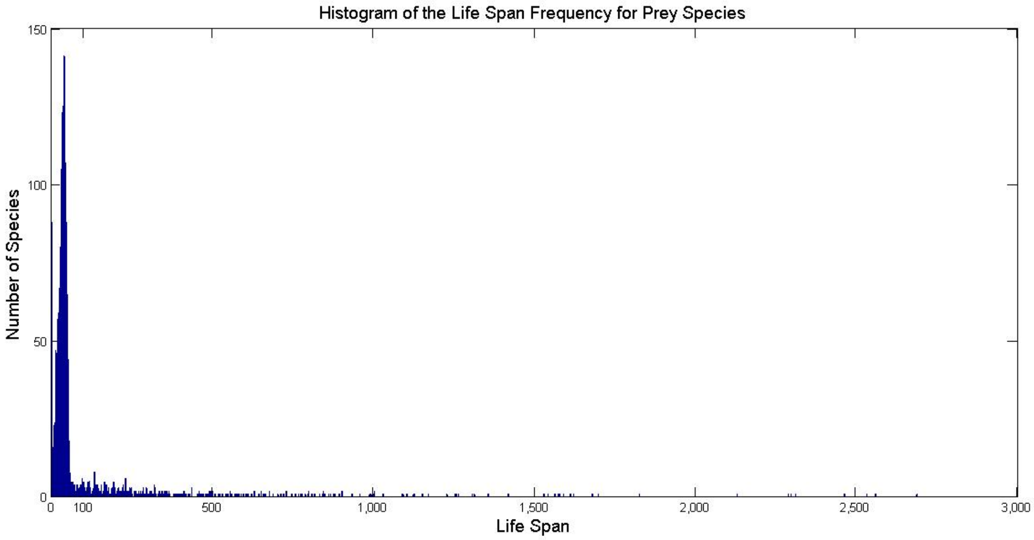


Figure 4.9 Histogram of the life span frequency for the prey species in run s33. The horizontal axis shows the length of the life span, and the vertical axis shows the number of the species having that life span. There are many species with a very short life span (less than 100 time steps), which should be filtered as they are considered as the noise in the speciation mechanism.

The algorithm searches all the phylogeny tree using a depth first search method, and writes all the instances of sister species, which match the first criterion in an output file.

As the speciation event in EcoSim is a two means clustering method, at each speciation event, we would have only one new species emerging from a parent species. This means that in some cases, where a species might have the potential to generate more than one new species, we might observe two consecutive speciation events with a very short time. For these cases, we still take into account those species with such sequential speciation events as sister species, where the difference between their originating time step is less than five time steps.

4.4. Verifying the second criterion: Complete divergence

The second criterion, which requires fulfillment for sympatric speciation, is the observance of complete divergence between the two sister species. A way to identify the extent of divergence between two species is by imposing a limit on the number of hybridization events. Speciation is a continuous process; imposing any finite threshold implies a subjective approach towards the experiment [Bolnick and Fitzpatrick 2007]. In order to measure true reproductive isolation between two species, a genetic basis must be considered [Coyne 2007]. In our simulations, we have measured divergence on the basis of hybridization events, as a ratio of all reproductive events occurring throughout evolutionary time, for all individuals pertaining to either of the two sister species. This can also be considered a measure to differentiate inter-specific reproduction from intra-specific reproduction.

In order to verify whether this criterion holds for a set of sister species, we calculated the ratio of hybridization events occurring between all individuals belonging to either of the two sister species. For this purpose, we used the information saved in part of the MinSave files containing the ID of the parent of each individual. Then for all the individuals in the two sister species, we checked and counted if any intra-specific reproduction is occurred. Every reproduction event between all individuals belonging to either species is taken into account to calculate the ratio of intra-specific reproduction versus inter-specific reproduction occurring at each time step.

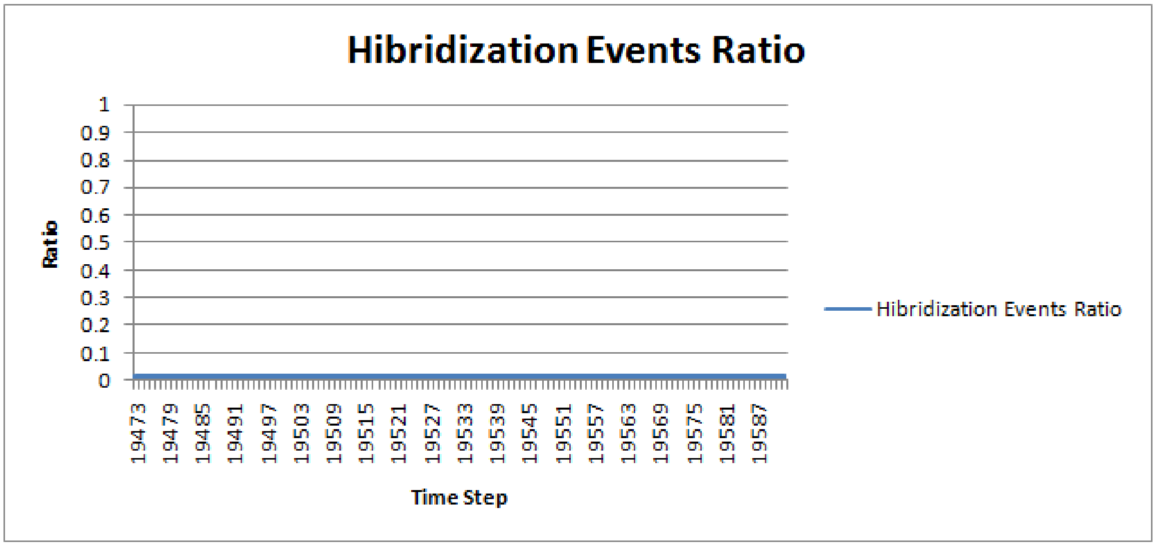


Figure 4.10 The ratio of all intra-specific reproduction versus inter-specific reproduction events occurring at each time step of the simulation between all the individuals belonging to a couple of candidate sister species.

Error! Reference source not found. shows the ratio of hybridization for the two sister species which were found in the previous example phylogeny tree. As it is observed, there were no hybridization events occurring during the persistence of the shorter lived species at any given time steps prior to its extinction.

4.5. Verifying the third required criterion: Overlapping geographic ranges

The third criteria, centered on spatial distribution of a species, requires that the two sister species share an overlapping geographic range. Geographic barriers were traditionally believed to be the main source of restricted gene flow from a spatial perspective, with extensive empirical support showing sister species tendency to be separated at a discontinuous geographic scale, allopatry, or associating through minimal geographic overlap, parapatry[Barraclough et al. 1998]. In order to measure the spatial scale at which to study the habitat of the species, it is important to consider the dispersal ability of all individuals belonging to that species [Bolnick and Fitzpatrick 2007]. When assessing a study based on resource distribution or host-plant mediated interactions, the dispersal ability of each individual, rather than the average of the population as a whole becomes important. Often the resources may be distributed on a continuous overlapping scale throughout the population, but in the case of certain host plant-mediated insects, their dispersal patterns do not overlap, creating pockets of “micro-allopatry,” resulting in internal pockets segregated spatially from the population as a whole [Berlocher and Feder 2002]. Allopatric gene flow barriers would arise through differential dispersal rates for a population that may initially be believed to coexist on a sympatric scale.

To justify that the speciation event occurred between individuals sharing a common geographical range, we calculated the average distance expressed in cell units, for all individuals belonging to either of the two sister species, and then based on this information, we calculated the minimum distance between the two closest individuals, the average distance of the 200 closest individuals, and total average distance between all the individuals in either species. This helps us to find out how close the species are living within the environment.

Error! Reference source not found. shows such calculations for an example of two candidate sister species, which were found in the phylogenetic tree and satisfied the first and the second required criteria, starting from their speciation event. As it can be observed, the distance between the two populations at the time of speciation is very small, as there are at least 200 individuals from either species, which are leaving in the same cell than an individual of the other species, and the total average distance between all the individuals of the two population is about 10 cells. We can conclude that these two species have been leaving at a shared geographical area, and therefore, the third required criterion holds for them.

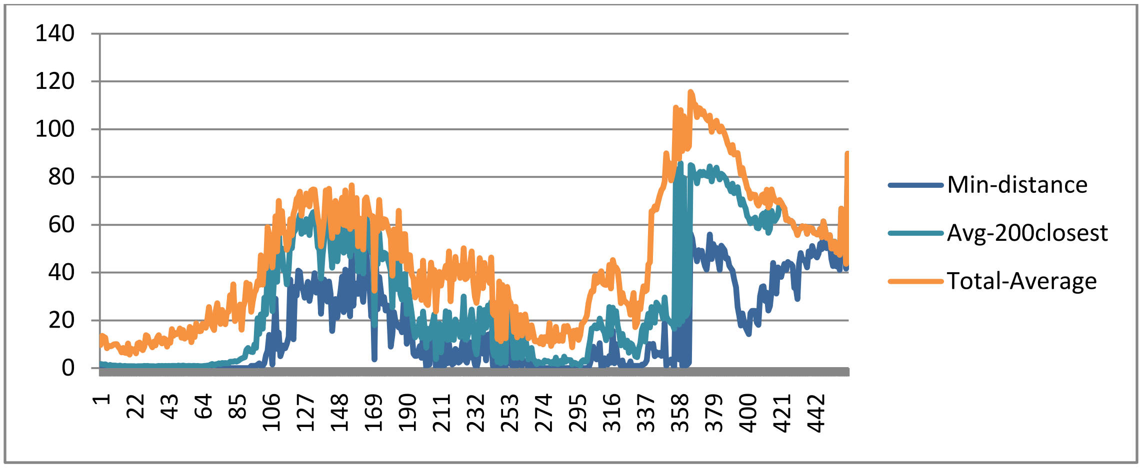


Figure 4.11 The minimum distance, the average distance of the 200 closest individuals, and the total average distance between all the individuals, belonging to a couple of candidate sister species from their speciation event and the subsequent time steps.

In order to be able to compare the distance of the set of candidate sister species, with the average distance of all the sister species in our simulation, we calculated the measures of the minimum distance, the average distance of 200 closest, and the total average distance between all the individuals of every couple of sister species. There were around 10000 couples of sister species, in a run with 25000 time steps, with about 3000 of these couples that had a life span of more than 100 time steps. We calculated distance measures for the 3000 couples of sister species with a life span of more than 100 time steps, then we calculated the average of all the minimum distances, the average of all the average distance of the 200 closest individual distances, and the average of all the total average distances. **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.**, respectively compares the minimum distance, the average distance of the 200 closest individuals, and the average distance for all the individuals belonging to the set of candidate sister species, with the averages calculated from the other 3000 sister species in the same run, during the first 200 time steps after the speciation event. As it can be observed, the average distance between the individuals of the set of candidate sister species, in much less than the average distances of all other sister species. The calculation of the minimum and the average distance of all the sister species, helps us to find an idea about the extent of the living area which sister species share in our simulation, to select a threshold for the highest minimum distance allowed, as well as the maximum allowed average total distance between the individuals of the sets of candidate sister species, to determine the thresholds to be used for testing the third required criteria.

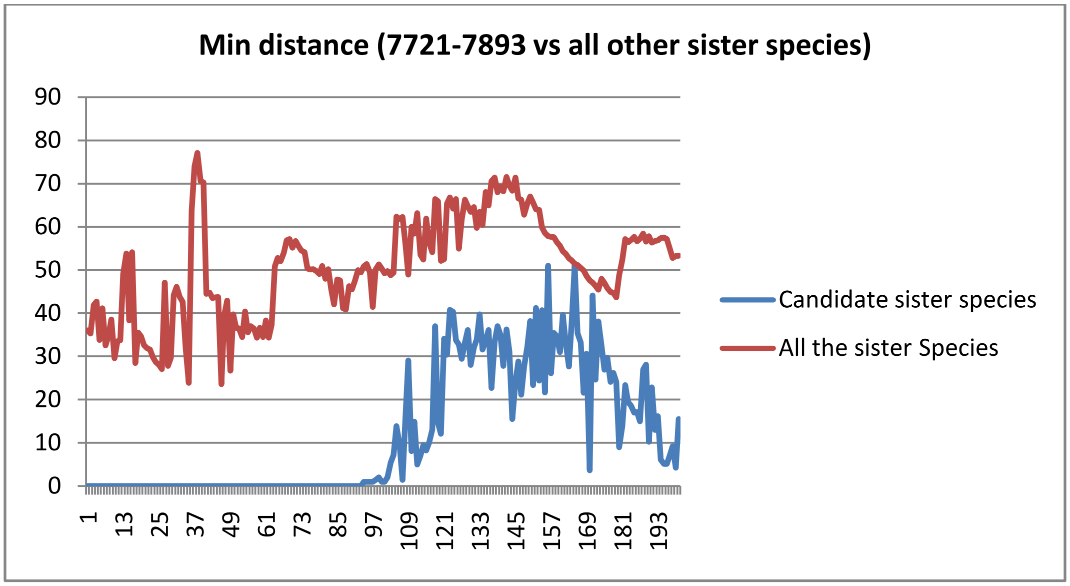


Figure 4.12 The minimum distance between the individuals of sister species. The blue curve shows the minimum distance between an example set of candidate sister species, the red curve shows the average of minimum distances between populations of all the sister species in the simulation during the first 200 time steps after their speciation event.

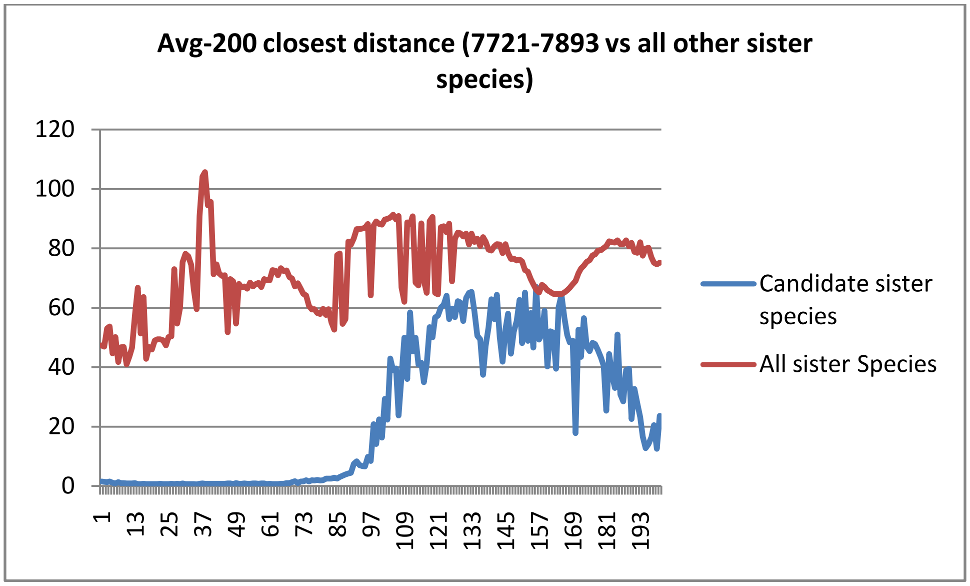


Figure 4.13 The average distance between the 200 closest individuals belonging to sister species. The blue curve shows the average distance between 200 closest individuals belonging to an example set of candidate sister

species, the red curve shows the average of the average distances between 200 closest individuals belonging into the populations of all the sister species in the simulation during the first 200 time steps after the speciation event.

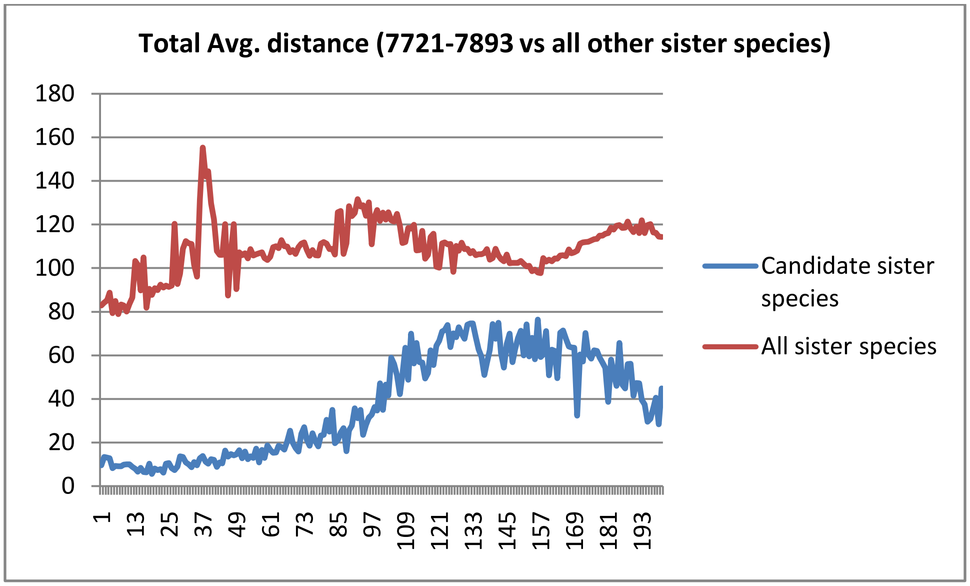


Figure 4.14 The total average distance between all the individuals belonging to sister species. The blue curve shows the average distance between all the individuals belonging to an example set of candidate sister species, the red curve shows the average of average distances between all the individuals belonging to the populations of all the sister species in the simulation during the first 200 time steps after the speciation event.

To strengthen this result, we took into account the minimum distance, the average distance of 200 closest individuals and also the total average distances between all individuals belonging to sister species that did not arise through sympatry, in order to show the statistical significance of the spatial overlap exhibited by sympatric sister species. For this purpose, t-test was applied to the measures of distances for every set of candidate sister species and all other sister species. The results of the t-test between the distance of the individuals of one set of candidate sister species (7721-7893), which are visualized on the **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.**, and all other sister species are shown in the Table 4.4 indicating that the distances between our studied species are significantly different from the distances between all other sister species.

The calculations of the distances between the individuals of the sister species were calculated for all the couples of species, which have already successfully passed two previously required sympatric speciation criterions. Those couples of sister species for which the minimum distance between their individuals and the average distance between their closest 200 individuals was 0 during the first 50 time steps after the speciation event, and the total average distance between their populations was less than 13 during the same time, were considered as the species which successfully passed the third required criteria as well, which is living at a shared geographical location.

p-value	T Test- Sympatric species distance compared with all other sister species – 200 t.s.										
	Min dis.	Avg. 50 closest	Avg.100 closest	Avg.150 closest	Avg.200 closest	Avg.250 closest	Avg.300 closest	Avg.350 closest	Avg.400 closest	Avg.450 closest	Avg.500 closest	Total Avg.
	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001	Less than 0.0001

Table 4.4 The results of applying t-test, on the distances between the individuals of one example set of candidate sister species(7721-7893), and the average distances between the populations of all the sister species in the simulation. The results show that the distances between our studied candidate species are significantly different from the distances between all the other sister species.

4.6. Verifying the fourth required criteria: Reject Allopatric/Parapatric alternate hypothesis

The last required criteria originally postulated by [Coyne and Orr 2004], and modified by [Bolnick and Fitzpatrick 2007], requires that the accumulated evidence for the biogeography and evolutionary history of the group of sister species, restricts the possibility of speciation occurring during a phase of allopatry/parapatry. According to [Bolnick and Fitzpatrick 2007]there is a significant amount of redundancy apparent in this criterion, as the third criterion takes into account the biogeography of the two species in relation to one another, and the second criterion examines their phylogenetic lineage.

In a natural study, it is possible that errors exist through an experimental procedure in determining the evolutionary history of either sister species due to sampling errors, and the major concern for these criteria is ruling out a period in the group's shared lineage where the populations may have split, either on a parapatric or allopatric spatial scale. If such a split occurred in their evolutionary past, and the groups later cohabited a shared geographic range, it is possible that differential selective pressures leading to speciation occurred during their isolated phase. This last criterion is an alternative hypothesis, in essence, and is most accurately fulfilled by justifying that no such period of geographic isolation occurred between the two species believed to have arisen through sympatry.

The benefit of our simulation studies is that all phylogenetic and biogeographic tracking are accessible in relation to every individual within a population. No sampling errors are inherent within the experimental procedure, and no assumptions are made in relation to population tracking. This allows us to justify that sympatric speciation has occurred between the two sets of sister species based on their complete biogeographic and phylogenetic history. Moreover, in our simulation, there are never any physical barriers that restraint the movements or isolate populations. We can then reject the alternative hypothesis, thus fulfilling the final criterion.

Chapter 5

THE RESULTS: DID SYMPATRIC SPECIATION HAPPEN?

In this chapter, results from a set of simulation runs, which were tested with the previously mentioned approaches to verify the probable happenstance of sympatric speciation, are presented.

5.1. Submitting the runs of the EcoSim

More than 50 runs of the two resources version of the EcoSim with different initializations in terms of the foods' specifications, were executed on the Shared Hierarchical Academic Research Computing Network (SHARCNET) servers, each for about three months, allowing us to have the simulation executed for about 25000 time steps for every run, which is long enough for observing the evolutionary behavior of the species through the simulation. Also 10 runs of the classic version of the EcoSim (one source of food) were submitted, to be able to compare the results, whenever required.

Grass Specifications	Food1	Food2	Food(Standard EcoSim)	Description
ValueGrass	250	400	325	energy value for a consumed grass
MaxGrass	4	4	8	maximum number of grass in a cell
SpeedGrowGrass	0 .3	0.2	0.3	speed of growing grass
ProbaInitialGrass	0.187	0.187	0.187	initial probability of grass per cell
ProbaGrowGrass	0.0016	0.0014	0.0016	probability of diffusion of grass

Table 5.1 The specifications of foods in the submitted runs (both the standard and dual resource version of the EcoSim). The Food2 is more valuable in terms of the amount of energy, which transfers to the prey individuals, while Food1 can grow faster and has a higher probability of getting spread through the environment. These specifications will allow the individuals to get specialized to a source of food, either for the matter of gaining more energy or just for having easier access to it.

A total of 20 runs with the attributes of the Food1 and Food2 described in the Table 5.1 were submitted, and shown promising results in terms of finding some evidence of divergent eating behavior on different species. For these experiments, Food1 and Food2 differed in the amount of energy transferred to the prey after eating each resource, and also the probability of the diffusion, and the growth speed of the grasses. These two resources of food are identical in terms of other specifications such as the maximum number of food per cell and the initial probability of food availability per cell. The Food2 provides more energy to the, while Food1 can grow faster and has a higher probability of getting spread through the environment. These specifications allow the individuals to get specialized to a source of food, either for the matter of gaining more energy or just for having easier access to it. Some experiments with different levels of the attributes assigned to Food1 and Food2 were executed for finding the suitable level of the attributes to observe the emergence of divergent eating behavior. In the first experiment, 20 runs were submitted such that Food1 and Food2 were completely identical, and the FCM-Clustering method and Action-Perception Clustering method presented in Section 5.2 were computed. No sign of specialization on a specific source of food was observed in any of these 20 runs. In a second experiment, Food2 was selected to be more valuable in terms of the energy transferred to the prey, while all the other attributes were selected to be identical. 20 more runs were submitted with these new attributes, and the resource preference clustering methods were applied on the results. It was observed that after about 11000 time steps, all the prey species were specialized on Food2, the more valuable source of food. This is why we decided to change some other attributes of the foods to make both resources worthy for the species to compete for. In the last experiment, 20 runs with the attributes mentioned in the Table 5.1, were submitted, such that both foods were dominant from a different aspect. Food1 was growing faster with a higher probability of the diffusion in the world, and Food2 was transferring more energy to the individuals. With this experiment the emergence of species being specialized on different food resources were observed, and therefore the attributes at the Table 5.1 were kept for the Food1 and Food2.

The values of two parameters ValueGrass and MaxGrass are normalized compared to the values in the standard simulation with one source of food, in order to have a similar amount of energy available for the prey to be able to compare the results with the standard simulation in relevant cases. For example, the ValueGrass in the standard simulation is 325, which is equal with the average values for the Food1 and Food2. The Maximum number of food per cell also, which is 8 in the standard simulation is divided by two for each type of food in the two resources version.

After having the results of all the runs for at least 25000 time steps, we started to test each run separately to verify whether we can find a run with some instances of sympatric speciation. The steps described in the previous chapter were applied to each run. Species of each run were categorized into three groups based on their resource preferences, either based on their behavioral model or by their real behavior. Then the phylogeny tree of the species splitting events was constructed for each run, and it was searched for finding the instances of sister species, each specialized in a different source of food. If such instances of sister species could be found in a run, then the hybridization ratio between the individuals belonging to either sister species and the minimum and the total average distances between all the individuals within each species were calculated, and if the hybridization ratio was around 0 (less than 0.01), and the total average distance of individuals was relatively small (less than 13 cells), those sets of sister species were considered as a positive instance of sympatric speciation based on different resource preference.

5.2 Finding the positive results

With the FCM-Clustering method in which the species were categorized into three groups based on their behavioral model, or FCM, we were not able to find many instances of sympatric speciation. The problem of this approach is that there might be some edges in the FCM which have a positive influence on an action, and those genes are coming from a concept which is an important concept at the FCM of the individual such that the concept is frequently being activated. On the other hand, there might be some other genes from another less important concept, which might have a negative influence on the same action.

For calculating the weighted sum of the genes influencing the Eat1 and Eat2 actions, we did not take into account the importance of the concepts which the genes are originating from, therefore, the effect of an important frequently active positive gene, is not compensated by a negative less important gene with the same absolute weight influencing the same action. The consequence would be that, there might be some species, which are specialized on one specific source of food, but we were not able to find them by just looking at their FCM. This is why we decided to categorize species into the three groups based on Action-Perception Clustering approach, which instead of the behavioral model of species takes into account the real behavior of the individuals in the ecosystem.

With the Action-Perception Clustering method, we could find some instances of sympatric speciation in some of the runs. Table 5.2 summarizes the total number of runs, in which we could find some instances of sympatric speciation.

Number of Runs	**Number of the Sympatric instances**
5 runs	Well over 10 pairs
7 runs	1 or 2 pairs
8 runs	No positive instances

Table 5.2 The number of observed sympatric speciation instances in a total of 20 runs. Five runs were found, with more than 10 instances of sympatric speciation, seven runs with only one or two instances of sympatric speciation, and eight runs without any happenstance of this phenomenon.

As it can be observed, from a total of 20 tested runs, we found five runs, with more than 10 instances of sympatric speciation, seven runs with only one or two instances of sympatric speciation, and eight runs without any happenstance of sympatric speciation. We observed that sympatric speciation was a rare type of speciation in our model as its frequency was only between 0.01% and 0.4% of all the speciation events in the runs, where this phenomenon was observed. This low frequency complies with the fact that this kind of speciation is a rare phenomenon in nature, as there are only a few widely accepted examples of sympatric speciation in empirical level [Bolnick and Fitzpatrick 2007].

# of couple of Sister species	Run S10		Run S19		Run S25		Run S33		Run S34	
Initial number	8449		9106		10173		10880		9770	
After applying the first criterion	FCM-Clustering	Action-Perception Clustering	FCM-Clustering	Action-Perception Clustering	FCM-Clustering	Action-Perception Clustering	FCM-Clustering	Action-Perception Clustering	FCM-Clustering	Action-Perception Clustering
	2	12	1	13	1	19	4	53	2	15
After applying the second criterion	1	12	1	13	1	19	4	53	2	15
After applying the third criterion	1	12	1	11	1	17	3	47	2	15

Table 5.3 Initial number of sister species, and the number of sister species which successfully passed the requirements needed by each criterion, for the five runs with over 10 instances of sympatric speciation.

Table 5.3, summarizes how the speciation events were filtered step by step, after applying the proposed verification approaches for each criterion, for the five runs with the highest number of observed sympatric speciation. It can be noticed from the table, that with the Action-Perception Clustering method, we were able to find a higher number of sister species to fulfill the sympatric speciation requirements. It can also be observed that the major parts of the speciation events are filtered after applying the first criterion, which selects the sister species which are specialized on different food resources and have a life span greater than 100 time steps. All the sister species which fulfilled the first criterion, also successfully passed the second required criterion (they found to be reproductively isolated). A small number of sister species which passed the first and the second criterion, were failed to pass the third criterion, as they found to be living far from each other. Table 5.4 shows the number of positive instances, found in each one of the five runs with more than 10 sympatric instances.

Run	S10	S19	S25	S33	S34
# Positive instances	12	11	17	47	15

Table 5.4 The number of observed sympatric speciation instances in each run.

The results of these five runs were used for creating a data set for more studies on the characteristics of the species generated by sympatry, to find if there might be any shared behavioral patterns among them. This is explained in detail at the next chapter.

5.3 More details of one of the runs

Presenting all the results of all the five runs which we observed having promising results would be beyond the scope of our discussion, in this subsection we only focus of run S33, in which we observed the highest number of sympatric speciation happenstance.

Error! Reference source not found., represents the total population of all the prey individuals belonging to all species of prey existing at every time step.

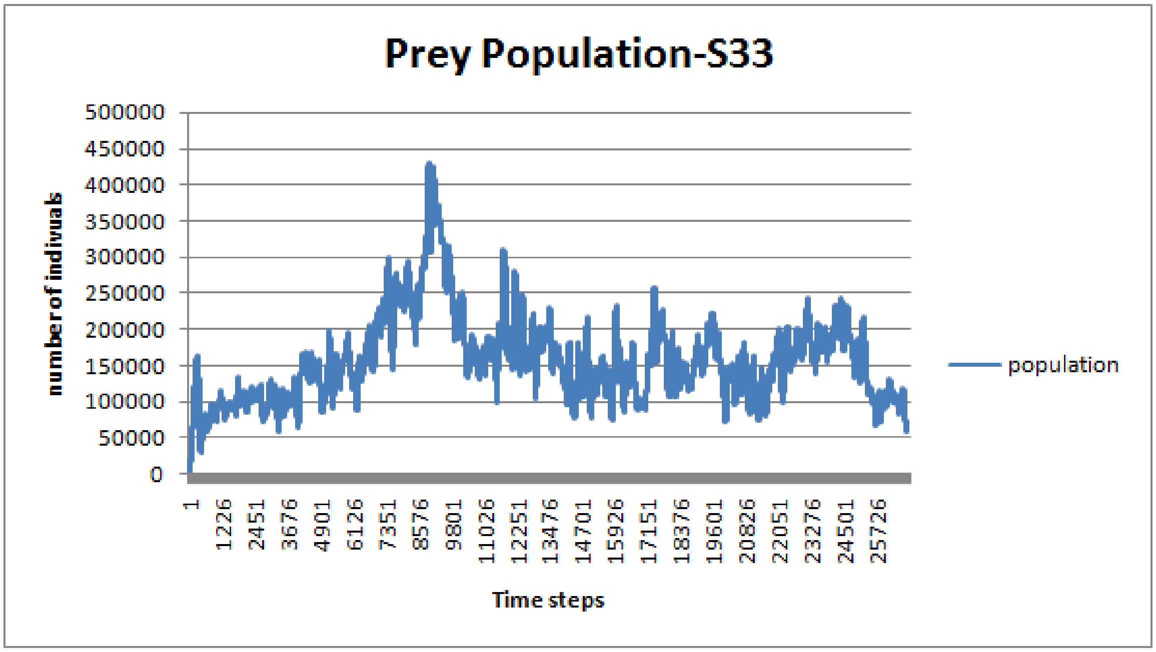


Figure 5.1 The total population of all the prey individuals belonging to all the species of prey existing at every time step (Run S33).

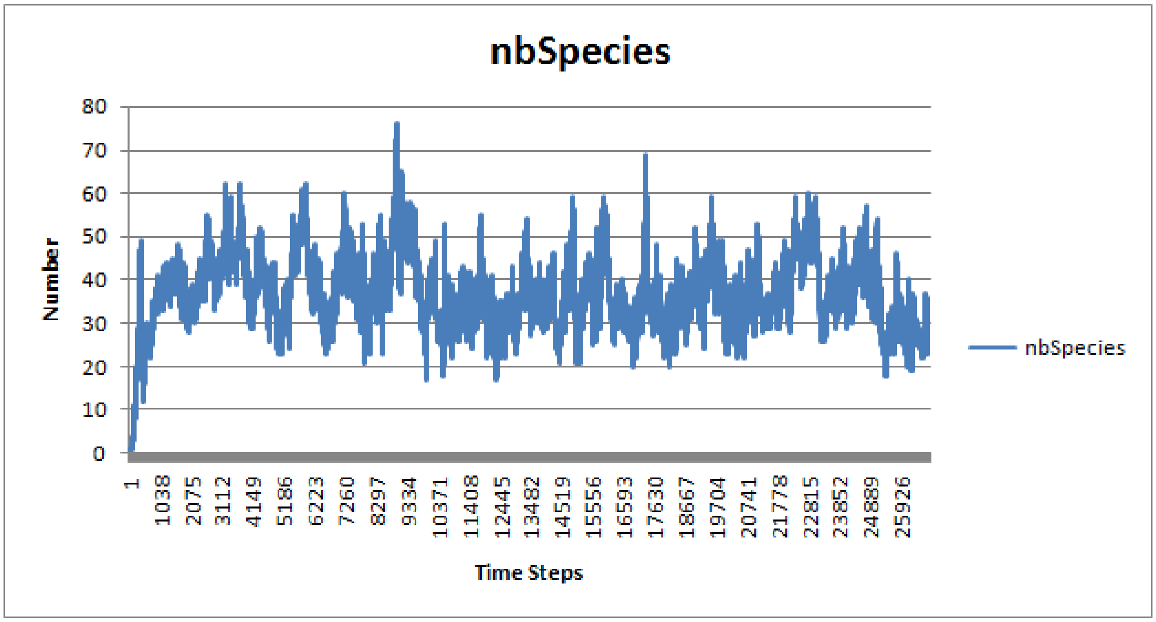


Figure 5.2 The total number of prey species existing at every time step. (Run S33

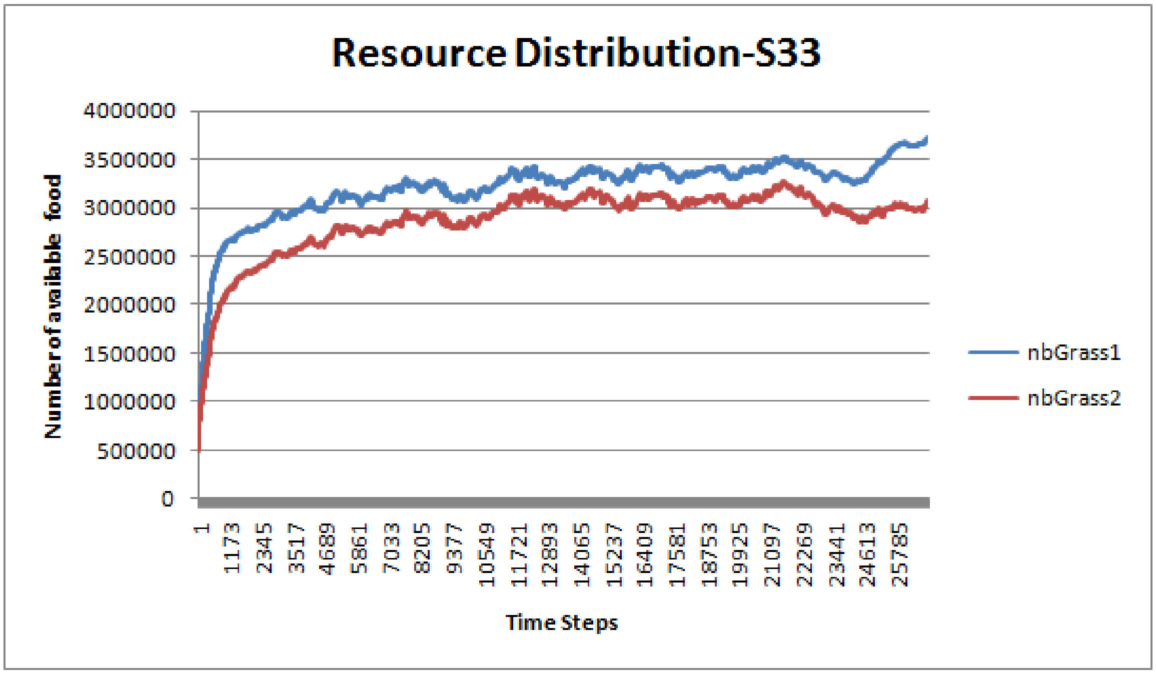


Figure 5.3 The total resource abundance of Food1 (blue) and Food2(red) during the simulation (Run S33). Food2 is the more valuable resource, in terms of the number of the energy, transmitted to the prey individuals, while Food1 can grow faster, with a higher probability of diffusion in the world.

The total number of prey species existing at every time step is give in **Error! Reference source not found.**. The total number of species at every time step is fluctuating between 20 and 60 continuously.

Error! Reference source not found., shows the rate of success or failure of searching for each resource, as a ratio of all actions performed by all prey individuals, at every time step of the simulation. The figure shows that, the two search for food actions, have a quite rare meaning that the prey have no difficulty to find some food.

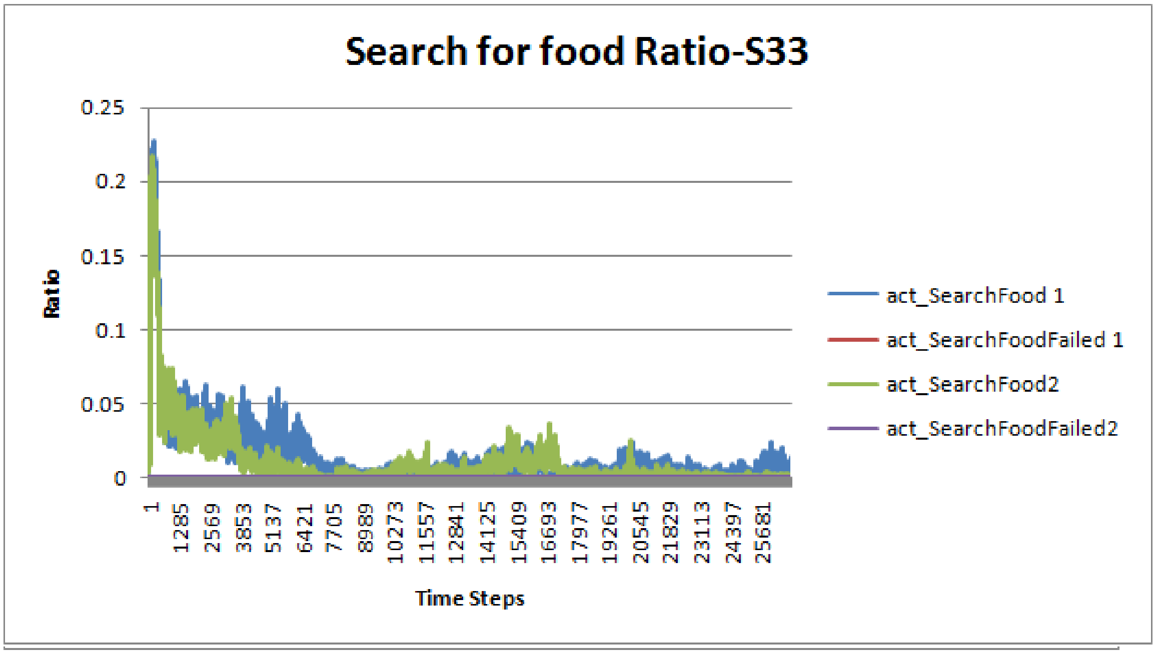


Figure 5.4 The success or failure of searching for each resource as a ratio of all actions performed by all the prey individuals at every time step of the simulation. (Run S33)

Error! Reference source not found., represents the success or failure of eating each food resource as a ratio of all actions performed by all prey individuals at every time step. It can be notice that the rate of Eat1 action is higher than the one of Eat2 from the beginning of the simulation. The reason can be explained easily as, at the beginning of the simulation where the individuals are not specialized on any specific source of food, they just simply choose the one resource which is more available, and according to **Error!**

Reference source not found., Food1 is more available than Food2, from the very beginning of the simulation. It can also be observed that starting from about time step 20000, there is an increasing trend for Eat2 action, and at the same time a decreasing trend for the Eat1 action, such that the ratio of these two actions, cross each other at about time step 22000, and from that time step the rate of Eat2 is notably higher than the one of Eat1. This could be explained by looking at Resource preference distribution of prey population for Food1 (blue), Food2 (red) or both resources (green). Each individual's preference is calculated at every time step for the duration of the simulation The chart shows that, starting from about time step 22000, a large part of the prey population was specialized on Food2, which means that despite the fact that Food1 is still more available in the environment, they prefer to consume Food2 more than Food1.This can explain the increase in the trend of Eat2 action after this time step. Looking at the **Error! Reference source not found.** also shows that, starting from the time step 22000, the difference between the total number of available Food1 and Food2 in the world is getting higher, compared with the steady difference between their level of availability before that time step.

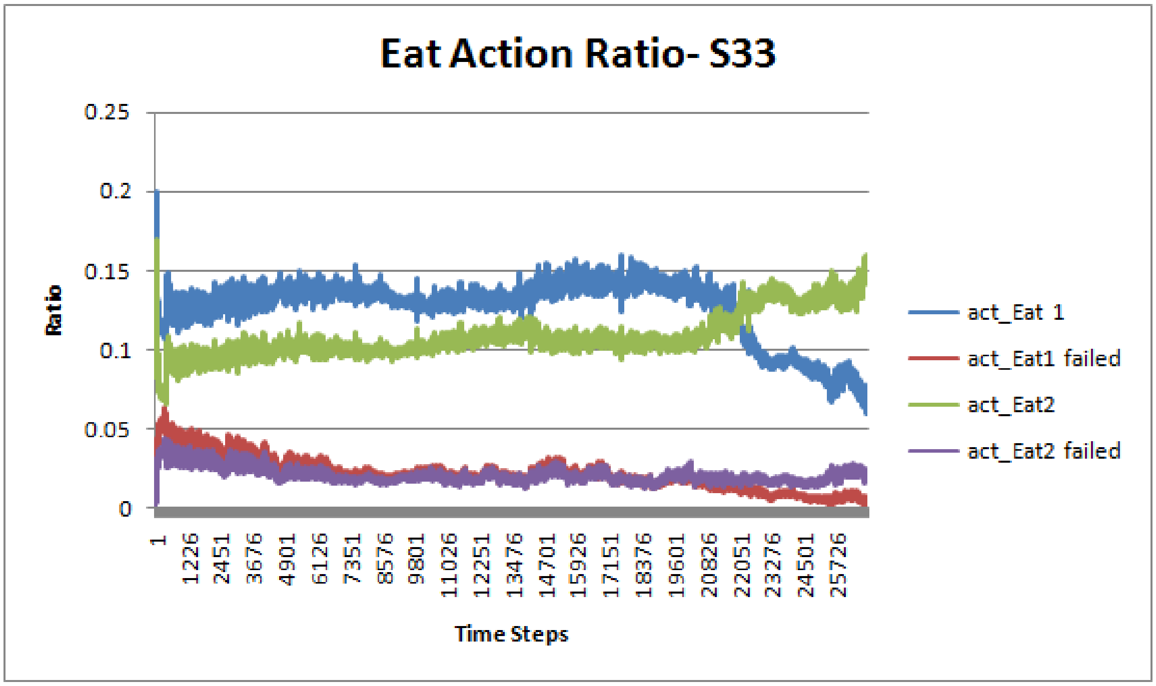


Figure 5.5 The success or failure of eating each resource as a ratio of all the actions performed by all the prey individuals at every time step of the simulation (Run S33). The rate of Eat1 action is higher than the one of Eat2 from the beginning of the simulation. It can be observed that starting from about time step 20000, there is an increasing trend for Eat2 action, and at the same time a decreasing trend for the Eat1 action, such that the ratio of these two actions, cross each other at about time step 22000, and from that time step the rate of Eat2 is notably higher than the one of Eat1.

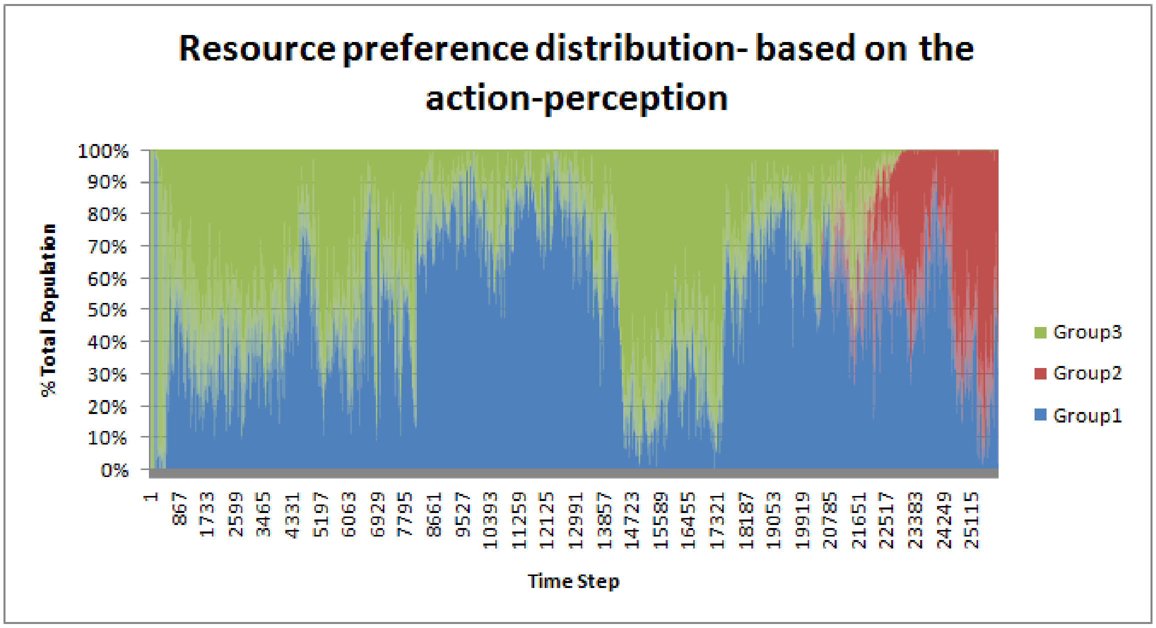


Figure 5.6 Resource preference distribution of prey population for Food1 (blue), Food2 (red) or both resources (green). Each individual's preference is calculated at every time step for the duration of the simulation. The

horizontal axis shows the time step, and the vertical axis shows the percentage of population belonging to each group at every time step.

5.4 Comparing the results of sympatric sister species, with all the other sister species

For all the sets of sister species with the minimum lifespan of 100 time steps, the hybridization ratio between the individuals member of the sister species, and also the average geographical distance between their individuals, was calculated by the same method which was used for the verification of the second and the third required criteria of sympatric speciation. These measures allow us to compare the extent of the geographical overlapping area and also the amount of reproduction isolation of our sympatric sets of sister species, with all the other sets of sister species to find out how significant the required criteria for the sympatric speciation are. **Error! Reference source not found.** shows the scatter plot of hybridization ratio and the average geographical distance between the individuals of all the sister species in the five runs in which we observed more than 10 instances of sympatric sisters. Each red circle corresponds to a couple of sympatric sister species, showing the level of hybridization ratio between the sisters' population and the average geographical distance between their individuals. Each green circle corresponds to a couple of species which were not verified as being sympatric after verification of the required criteria. These couples of sister species are labeled as non-sympatric in the plot. As it can be observed, the sympatric sisters are strongly clustered in the lower right part of the graph while the non-sympatric sister species are distributed along the two axes meaning that the non-sympatric sister species are either not completely reproductively isolated or live in non-overlapping area.

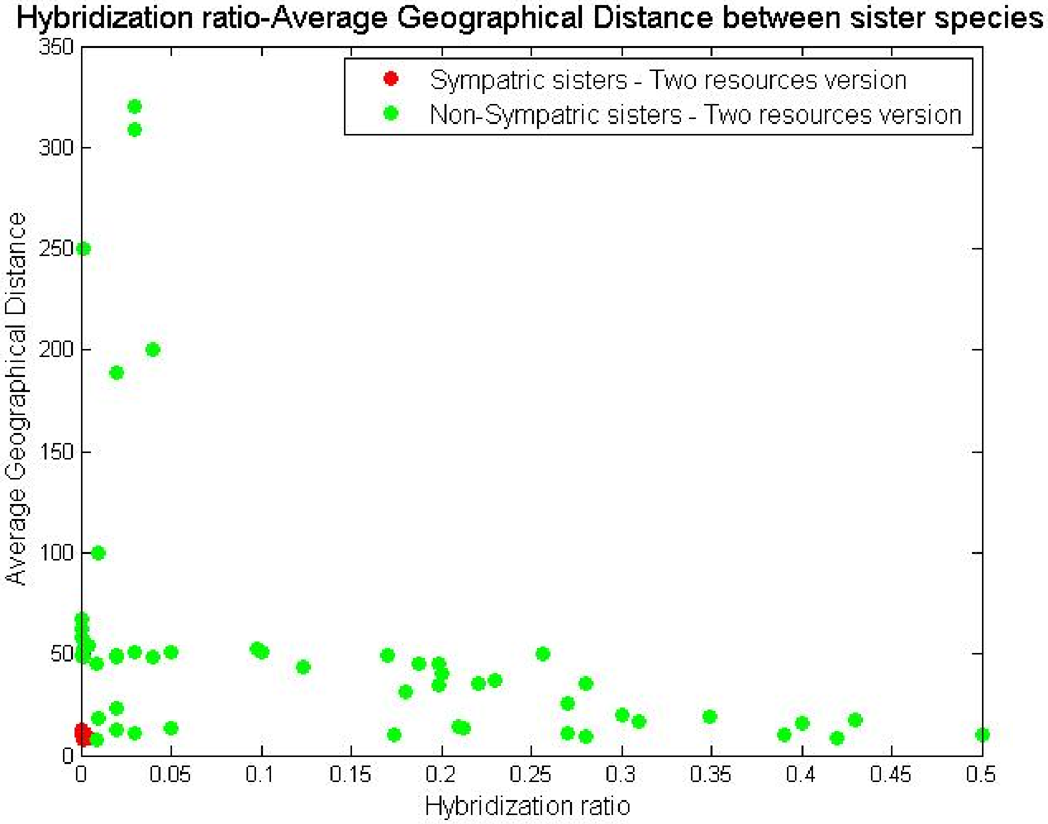


Figure 5.7 Hybridization ratio and the average geographical distance between all the individuals of sister species in dual resource version of the EcoSim. Red circles are representing the sympatric sister species, and green circles are representing the non-sympatric sisters species. The sympatric sisters are strongly clustered in the lower left part of the graph while the non-sympatric sister species are distributed along the two axes meaning that the non-sympatric sister species are either not completely reproductively isolated or they are living in a non-overlapping area.

The logarithmic plot of the hybridization ratio and the average geographical distance for the sister species is presented at **Error! Reference source not found.**. In this plot, the difference between the measures, for the sympatric sisters and non-sympatric sisters species is even more clear.

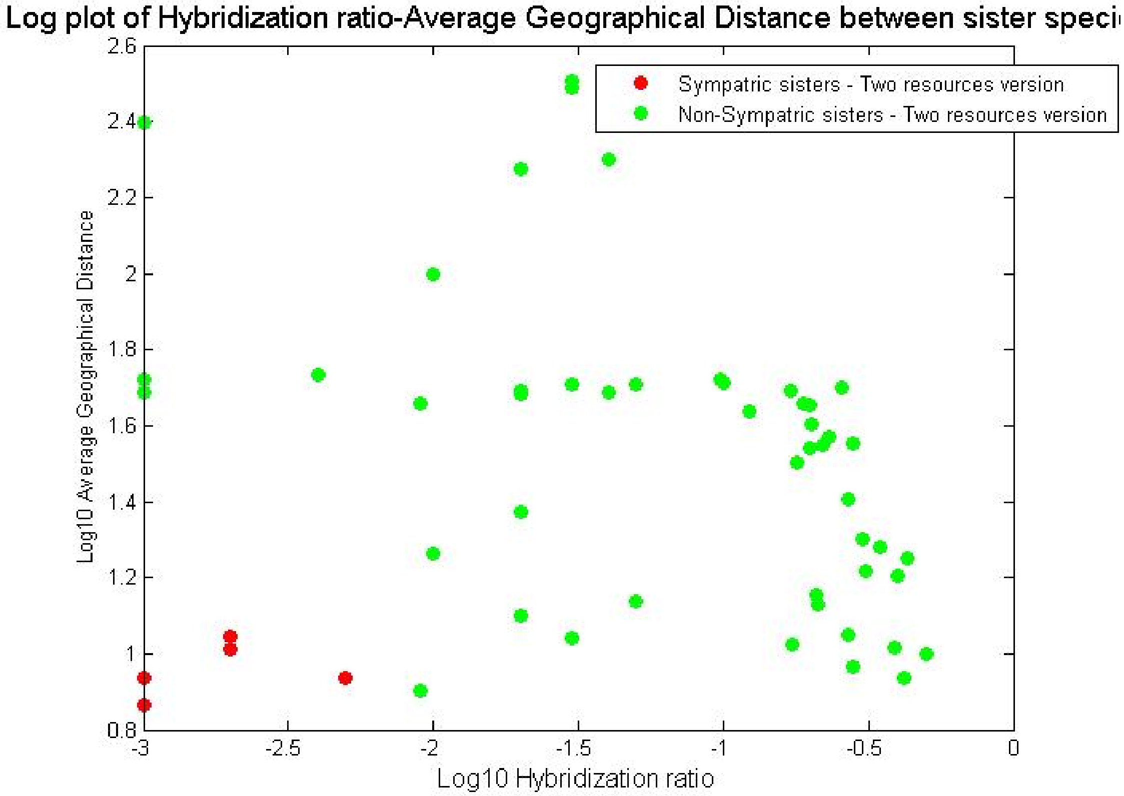


Figure 5.8 The logarithmic plot of the hybridization ratio and the average geographical distance between all the individuals of sister species in dual resource version of the EcoSim. Red circles representing the sympatric sisters species, and green circles representing non-sympatric sisters species. The difference between the measures, for the sympatric sister species and non-sympatric sister species is even more clear.

One of the main hypotheses for emergence of sympatric speciation is that the presence of two different resources could cause a reduction of gene flow leading to speciation. To verify this hypothesis, we applied the same criterion to prove the existence of a sympatric speciation event to all the couple sister species that where generated by five runs of the version of EcoSim in which only one source of food exist. **Error! Reference source not found.** shows the scatter plot of hybridization ratio and the average geographical distance of the sister species for both versions of the EcoSim, the classic version with one resource for the prey individuals, and two resource version of the EcoSim. The blue circles show the hybridization ratio and the average geographical distance for all the sister species in five classic runs of the EcoSim with only one resource available, while the red and green circles, as in the **Error! Reference source not found.**, represent the sympatric sister species and the non-sympatric sister species respectively, in the two resource version of the EcoSim.

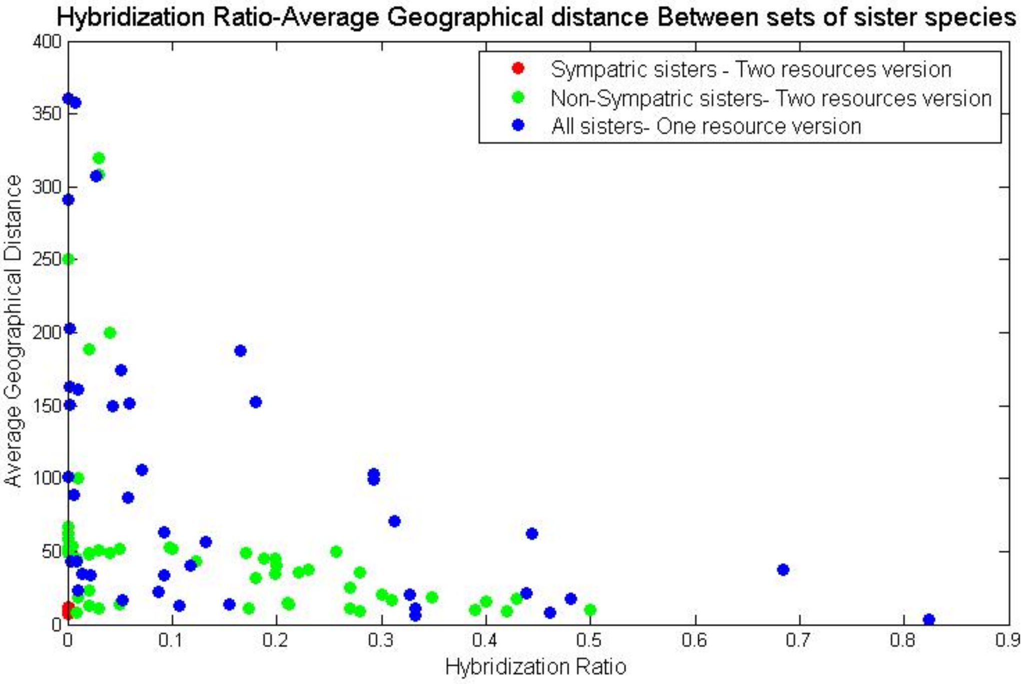


Figure 5.9 The hybridization ratio and the average geographical distance between all the individuals of the sister species in the two resource version of the EcoSim (Red circles representing sympatric sisters species, and green circles representing non-sympatric sisters species), and the standard version of the EcoSim (blue circles representing all the couples of sister species). No instances of sister species, were found to be satisfying the required criteria for the sympatric speciation in the one resource version of the EcoSim. Those species which were satisfying the criteria in the two resource version of the EcoSim, were those who were specialized in two different resources.

Error! Reference source not found. shows the same results using a logarithmic scale. With this scale it appears clearly how differentiate are the sets of sympatric sister species, compared with all the other species in two resource version and the classic version of the EcoSim.

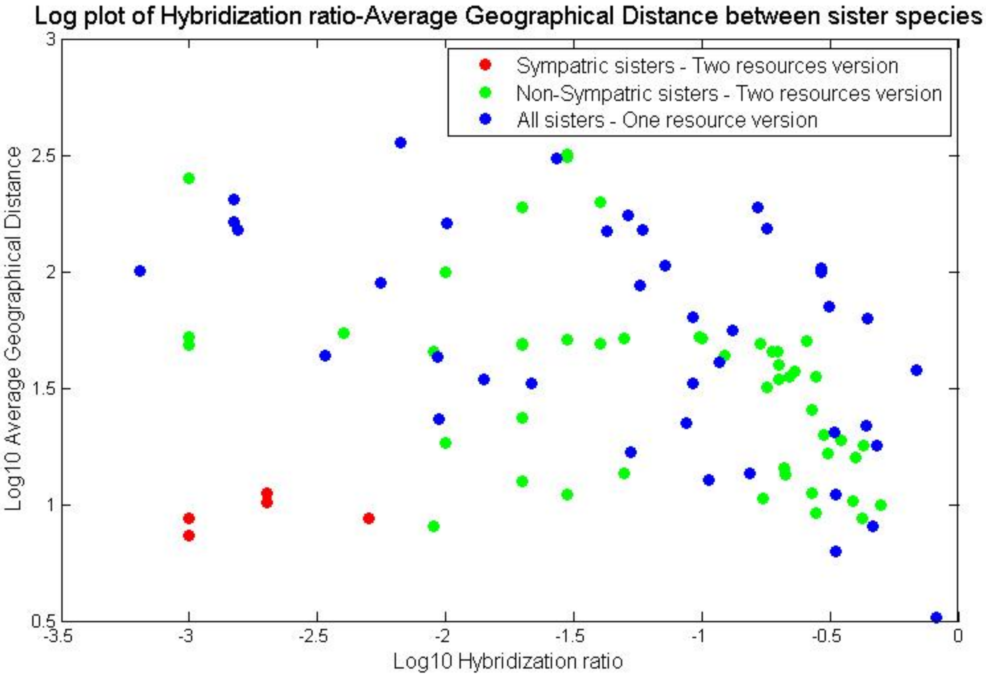


Figure 5.10 The logarithmic plot of the hybridization ratio and the average geographical distance between all the individuals of sister species in the two resource version of the EcoSim (Red circles representing sympatric sister species, and green circles representing non-Sympatric sister species), and one resource version of the EcoSim (blue circles representing all the couples of sister species). In this scale, it appears clearly how differentiate are the couples of sympatric sister species, compared with all the other species in dual resource version and the classic version of the EcoSim.

No instances of sister species, were found to be satisfying the required criteria for the sympatric speciation in the one resource version of the EcoSim. Moreover, those species which were satisfying the criteria in the two resource version of the EcoSim, were those who were specialized in two different resources. Therefore, it can be concluded that the divergent eating behavior has been the reason of reproduction isolation between the sister species, leading to sympatric speciation in our model. This constitutes a strong confirmation of the hypothesis of the importance of multiple resources for sympatric speciation to occur.

Chapter 6

WHAT ARE THE SHARED PATTERNS AMONG SYMPATRIC SPECIES: Applying the machine learning techniques

As it was mentioned in the previous chapter, we found 5 runs from a total of 20, with more than 10 instances of speciation events which have the required evidences of sympatric speciation. We used the results of these runs, for a more detailed study on the specific conditions leading to sympatric speciation.

6.1 Preparing the dataset

To investigate the conditions leading to sympatric speciation, versus those of speciation which cannot be considered as sympatric, we used the results of the 5 runs in which we found a notable amount of sympatric speciation events as the source for our data set. The 'sympatric species' were labeled as positive instances, and the information of other alive species at the same period of time were considered as negative instances, in the dataset.

At the beginning all the attributes of each species were selected for creating the initial dataset. The attributes could cover a wide range of information about each species, from some general information such as population size of each species or their interbreeding ratio, or the amount of their transferred energy to the child, to some behavioral specifications such as the rate of choosing different actions, or their perception of their environment. A complete list of the initial attributes, used for creating the datasets, and a short description about each one of them, can be found in the Appendix A.

6.2 Preprocessing the dataset

Four out of the five datasets, were imbalanced, where the amount of positive samples were only one third of the amount of negative samples. To solve the problem of imbalanced dataset, two main approaches have been mentioned [He and Garcia 2009], the first one is to assign distinct costs to misclassified samples and minimizing the overall cost on the training dataset, and the second one is either over sampling the minority class, or under sampling the majority class. For this dataset we used the smote algorithm [Chawla and Bowyer 2002] to resample the minority class, which is the one corresponding to our sympatric species. One other dataset was imbalanced in a reverse order, where the negative class was the minority class. For this run we oversampled the negative class. After making the datasets balanced, each dataset had around 6000 to 7000 instances, where each instance was the attributes of a species which is either in the positive class or the negative class.

6.3 Attribute selection

As already mentioned we started with 81 attributes each one explaining some characteristics about the species. The next step is to find the best set of attributes which could be used for classifying the dataset to gain the most accurate results. For this purpose, we applied some different attribute selection methods and tried to combine their results using voting to find the best set of attributes.

Table 6.1 represents the list of attributes and the result after applying attribute selection methods. We applied the Info Gain Attribute Evaluator implemented at Weka [Hall et al. 2009] combined with the Ranker search method. We also applied cfs subset Evaluator with three different search method, Best First, Greedy Stepwise and Genetic Search[Hall et al. 2009]. The attributes in the table are sorted by their score returned from Ranker plus Info Gain attribute Evaluator.

The Ranker combined with InfoGain attribute evaluator, assigns a score to each attribute based on their relative importance for the learning process. The lower the rank of an attribute the more important the attribute is. The best first search method combined with

Cfs subset evaluator, only selected 8 attributes, which already had a high score based on the rank returned by Ranker combined with InfoGin attribute evaluator method. The Greedy stepwise method combined with the Cfs subset evaluator also returned a rank for the first 20 important attributes. The genetic search method combined with Cfs attribute evaluator were applied on a 10 fold cross-validation attribute selection basis. If an attribute was selected by evaluation on all the 10 folds, a score of 100% were assigned to that attribute, and similarly if an attribute was not selected by evaluation on any fold, a score of 0% was assigned to that attribute. At the first step, we removed the attributes with the lowest score in all the attribute selection methods. For this purpose, we removed the attributes which had a score of less than 30% in Genetic Search with Cfs subset evaluator method, or their rank was higher than 40 on Ranker with InfoGain attribute evaluator. The removed attributes had already a low score in GreedyStepwise+Csf method and they were not selected by BestFirst+Csf method. The attributes which are highlighted in red, were removed at the first step and the number of the features were reduced to 29.

ID	Attribute	Ranker +InfoGain	bestFirst +Cfs	Greedy Stepwise +Cfs	Genetic Search +Cfs
15	distEvol	1	●	1	100%
21	nbArc	2	●	4	100%
16	stateOFbirth	3		10	80%
76	concept_socialize	4	●	3	100%
31	act_EatRatio	5	●	2	90%
38	reprodFailed_energy	6	●	5	40%
74	concept_searchFood	7	●	7	70%
80	concept_eat2	8		9	60%
69	concept_curiosity	9		12	90%
60	concept_foodLocalHigh	10	●	6	70%
61	concept_foodLocalLow	11		16	70%
33	act_Eat2Ratio	12	●	8	100%
63	concept_foodLocalLow2	13			90%
62	concept_foodLocalHigh2	14		18	10%
1	nbSpecies	15		20	0%

72	concept_nuisance	16			40%
68	concept_searchPartner	17		17	0%
40	parent1_reproductionEnergy	18			40%
71	concept_satisfaction	19			10%
59	concept_energyHigh	20			20%
58	concept_energyLow	21			0%
29	act_ExplorationRatio	22			70%
67	concept_hunger	23			80%
32	act_EatFailedRatio	24		11	50%
81	concept_reproduce	25			10%
18	Energy	26			40%
78	concept_wait	27			20%
70	concept_sedentary	28			10%
12	Entropy	29		13	50%
42	parent2_reproductionEnergy	30			40%
75	concept_searchFood2	31			0%
36	act_ReproduceFailedRatio	32			30%
19	Speed	33			20%
11	deadEnergy	34			60%
10	deadAge	35			80%
25	act_SearchFood2Ratio	36			50%
46	reasonReproduceFailed_PartnerEnerg	37			30%
22	act_EscapeRatio	38			0%
47	reasonReproduceFailed_PartnerActed	39		19	10%
2	nbIndividual	40			20%
79	concept_eat	41			10%
5	interBreedingRatio	42			30%
6	deadRatio	43			10%
50	concept_predClose	44			20%
51	concept_predFar	45			40%
77	concept_exploration	46			20%
17	Age	47			40%
13	diversitySpatial	48			20%
64	concept_partnerLocalYes	49			30%
65	concept_partnerLocalNo	50			0%
14	diversitySpatialRatio	51			80%
35	act_ReproduceRatio	52			20%

4	birthRatio	53			30%
66	concept_fear	54			40%
3	individualRatio	55			30%
20	Compactness	56			10%
27	act_SocializeRatio	57			10%
73	concept_escape	58			0%
34	act_EatFailed2Ratio	59			40%
41	parent2_reproductionAge	60			0%
23	act_SearchFoodRatio	61			0%
48	reasonReproduceFailed_PartnerAction	62			30%
28	act_SocializeFailedRatio	63			30%
30	act_WaitRatio	64			10%
7	deadAgeRatio	65		15	20%
49	reasonReproduceFailed_PartnerDist	66			10%
43	DistMating	67			20%
39	parent1_reproductionAge	68			20%
37	reprodFailed_age	69			10%
44	reasonReproduceFailed_Energy	70		14	40%
56	concept_friendClose	71			70%
57	concept_friendFar	72			0%
9	deadKilledRatio	73			50%
8	deadEnergyRatio	74			10%
45	reasonReproduceFailed_NoPartner	75			10%
54	concept_foodClose2	76			0%
55	concept_foodFar2	77			0%
53	concept_foodFar	78			0%
24	act_SearchFoodFailedRatio	79			0%
26	act_SearchFoodFailed2Ratio	80			0%
52	concept_foodClose	81			0%

Table 6.1 list of the attributes tested by different attribute selection methods. The Info Gain Attribute Evaluator implemented at Weka [Hall et al. 2009]combined with the Ranker search method, and the applied cfs subset Evaluator with three different search methods, Best First, Greedy Stepwise and Genetic Search [Hall et al. 2009] are employed. The attributes are sorted based on their score on ranker-info Gain attribute evaluator method. The attributes highlighted in red were removed at the first step of attribute selection.

6.4 Classification

At this step we tried to find a suitable classifier to classify the dataset with the aim of getting the highest accuracy on classification of the model, with the minimum number of attributes and rules, which will make it more convenient to explain the rules shares by each class.

We employed the J48 classifier in Weka [Hall et al. 2009], the CRF combined rule extraction and feature elimination method in supervised random forest classification[Liu et al. 2012], and the random forest classification combined with feature selection using hill climbing method [Mashayekhi and Gras 2013] to our datasets separately, to choose a suitable method for classification. We first tested each dataset separately to extract the rules on each run. The next step would be combining all the data sets together to find the patterns which are shared by all the runs.

For the first step, we found the J48 classification method suitable for our classification as it returned the lowest number of rules, compared to Random Forest methods. However, Random Forest methods obtained the highest level of accuracy on classification, but in this case the accuracy returned by J48 was reasonably high, therefore we used J48 classifier for this step and left random forest for the next step which we combine four out of five datasets together as the training set and use the fifth dataset as the test set.

S19	J48		Random Forest-CRF method		Random Forest - Hill Climbing			
#Features	Accuracy	#Rules	Accuracy	#rules	Average Accuracy	STD	Average #rules	STD
29	96.26%	17	99.99%	460	98.95%	0.002	40.66	5.70

Table 6.2 Total accuracy and number of rules returned by three different classifier on one of the datasets, for compari and choosing the appropriate classifier.

Table 6.2, shows the total accuracy and the number of rules returned by the three classification methods on one of the data sets. The J48 was selected for classifying each dataset separately as it returns the lowest number of rules with a high accuracy.

6.4.1 Classification of each dataset separately

J48 classifier was employed with different attribute selection methods, in order to find the minimum number of attributes, the minimum number of rules and the highest accuracy. The classification started with 29 attributes which were selected using the attribute selection method discussed in previous sub section. Then we tried to prune the decision tree by increasing the minimum number of instances per leave. This technique will help to reduce the number of rules, which would be very helpful when we try to explain the rules related to each class.

A small part of each data set were put aside, to be used as a validation set. Pruning and removing features were applied step by step to each dataset. Table 6.3, Represents these steps for the dataset S19. The number of selected attributes, the number of returned rules, the total accuracy, the TR Rate, and the ROC Area is presented for the 10 fold cross validation and the validation set. Starting from 29 features and 17 rules, we ended up to 5 features and 11 rules, despite the fact that the total accuracy was decreased from 96.26% to 86.79%. However it would be less complicated to discuss 11 rules with only 5 features rather than 17 long rule with 29 features, and an accuracy of more than 86% means that the main properties have been captured and can provide a first analysis of the condition leading to sympatric speciation.

S19	#Features	#Rules	J48----Train Set – 10 fold C.V.					Test set (Unseen data from the same run)				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	29	17	96.26%	0.96	0.96	0.98	0.98	95.7%	0.97	0.94	0.98	098
	29	12	89.3%	0.89	0.89	0.94	0.94	89.3%	0.89	0.89	0.94	0.94
	20	39	99.21%	0.99	0.99	0.99	0.99	99.4%	0.99	0.99	0.99	0.99
	20	18	96.75%	0.95	0.97	0.99	0.99	97.4%	0.96	0.98	0.99	0.99
	20	15	93.08%	0.92	0.94	0.97	097	94.1%	0.89	0.94	0.97	0.97
	20	12	89.89%	0.90	0.89	0.94	0.94	90%	0.82	0.98	0.96	0.96
	8	35	99.32%	0.99	0.99	0.99	0.99	99.5%	1	0.99	0.99	0.99
	8	19	92.82%	0.91	0.94	0.97	0.97	91.8%	0.86	0.96	0.98	0.98
	8	14	90.51	0.91	0.89	0.96	0.96	89.1%	0.86	0.91	0.97	0.97
	8	12	89.31	0.87	0.90	0.94	0.94	90.12%	0.88	0.91	0.96	0.96
	6	40	99.53	0.99	0.99	0.99	0.99	99.92%	0.99	0.99	1	1
	6	17	94.17	0.93	0.95	0.98	0.98	94%	0.9	0.98	0.99	0.99
	6	15	92.26	0.92	0.91	0.97	0.97	94.1%	0.93	0.94	0.98	0.98
	6	14	90.81	0.91	0.89	0.95	0.95	88.9%	0.83	0.94	0.97	0.97
	6-b	11	91.61%	0.93	0.98	0.95	0.95	87.65%	0.85	0.90	0.92	0.92
	5	53	99.04%	0.98	0.99	0.99	0.99	99.4%	0.99	0.99	0.99	0.99
	5	16	90.49%	0.88	0.92	0.96	0.96	91.6%	0.87	0.95	0.97	0.97
	5	14	90.36%	0.90	0.90	0.95	0.95	90.5%	0.87	0.94	0.96	0.96
	5	11	86.79%	0.85	0.88	0.93	0.93	85.5%	0.75	0.95	0.93	0.93

Table 6.3 The classification results of dataset S19, using J48 decision tree. Pruning techniques and feature removing methods are applied step by step to minimize the number of rules.

The mentioned method is also applied to all the other datasets. Table 6.4, Table 6.5, Table 6.6, and Table 6.7 respectively shows the summarized results of the classification for the datasets S33, S34, S25 and S10.

S33	#Features	#Rules	Train Set – 10 fold C.V.					Test set (Unseen data from the same run)				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	7	14	99.90%	0.99	0.99	0.99	0.99	100%	1	1	1	1
	2	4	99.66%	0.99	0.99	0.99	0.99	98.8%	1	0.97	0.98	0.98

Table 6.4 The classification results of dataset S33, using J48 decision tree. Pruning techniques and feature removing methods are applied step by step to minimize the number of rules.

S34	#Features	#Rules	Train Set – 10 fold C.V.					Test set (Unseen data from the same run)				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10	15	99.79%	0.99	0.99	0.99	0.99	100%	1	1	1	1
	6	7	99.31%	0.99	0.99	0.99	0.99	97.21%	0.95	0.98	0.99	0.99
	4	5	96.74%	0.97	0.96	0.98	0.98	94.35%	0.90	0.97	0.97	0.97

Table 6.5 The classification results of dataset S34, using J48 decision tree. Pruning techniques and feature removing methods are applied step by step to minimize the number of rules.

S25	#Features	#Rules	Train Set – 10 fold C.V.					Test set (Unseen data from the same run)				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	81	30	99.00%	0.98	0.99	0.99	0.99	99.00%	0.98	0.99	0.99	0.99
	9	14	95.86%	0.95	0.96	0.98	0.98	94.14%	0.96	0.92	0.96	0.96
	7-a	11	92.73%	0.9	0.94	0.97	0.97	87.53%	0.86	0.88	0.93	0.93
	7-b	9	88.11%	0.91	0.85	-/94	0.94	82.58%	0.86	0.79	0.88	0.88

Table 6.6 The classification results of dataset S25, using J48 decision tree. Pruning techniques and feature removing methods are applied step by step to minimize the number of rules.

S10	#Feat	#Rule	Train Set – 10 fold C.V.	Test set (Unseen data from the same run)

			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	1-a	2	99.98%	1	1	1	1	99.89%	1	0.99	0.99	0.99
	6	9	99.94%	0.99	1	0.99	0.99	99.84%	0.99	0.99	0.99	0.99
	2	4	99.45%	0.99	0.99	0.99	0.99	98.66%	0.99	0.98	0.98	0.98
	1-b	2	98.48%	0.98	0.98	0.98	0.98	98.66%	0.99	0.98	0.98	0.98

Table 6.7 The classification results of dataset S10, using J48 decision tree. Pruning techniques and feature removing methods are applied step by step to minimize the number of rules.

6.4.1.1 Extracting the rules

Classification using J48, returns a decision tree for each data set, each leaf being a rule assigned to a specific class. These rules can be discussed with biologist to find if there might be any rules in our model which is also observed in nature, or if there might be any other rule in our model which might be interesting for biologist. **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.** show the decision trees related to datasets S19, S33, S34, S25, and S10 respectively.

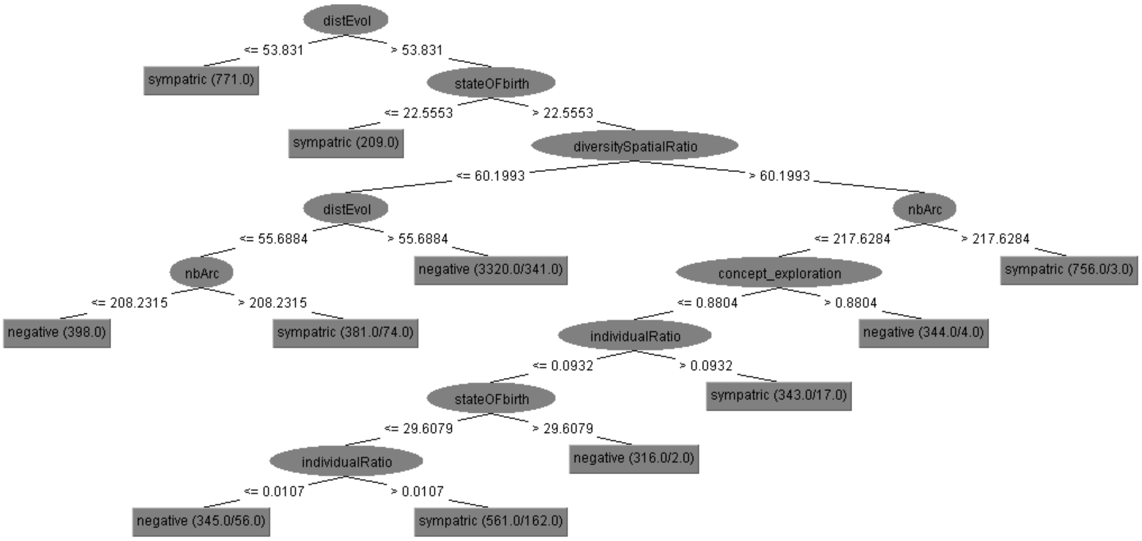


Figure 6.1 Decision tree with 11 rules, related to the dataset S19. Each leaf being a rule assigned to a specific class.

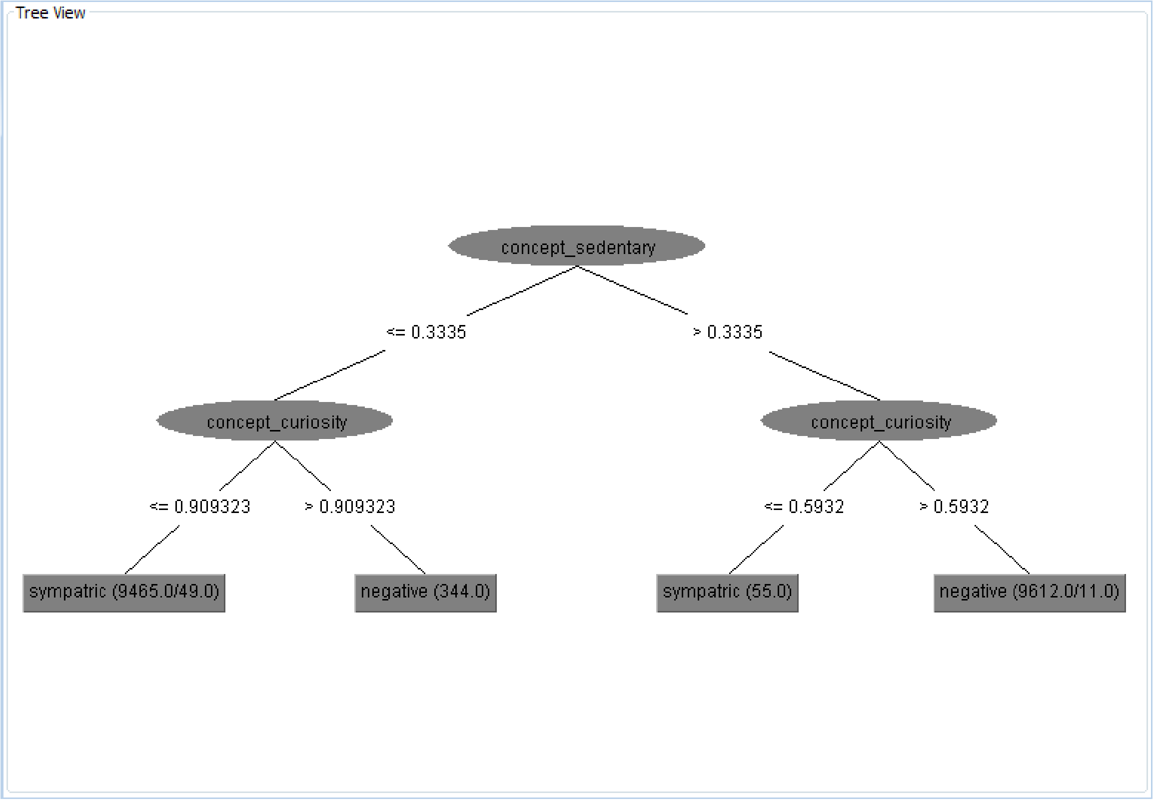


Figure 6.2 Decision tree with 4 rules, related to dataset S33. Each leaf being a rule assigned to a specific class.

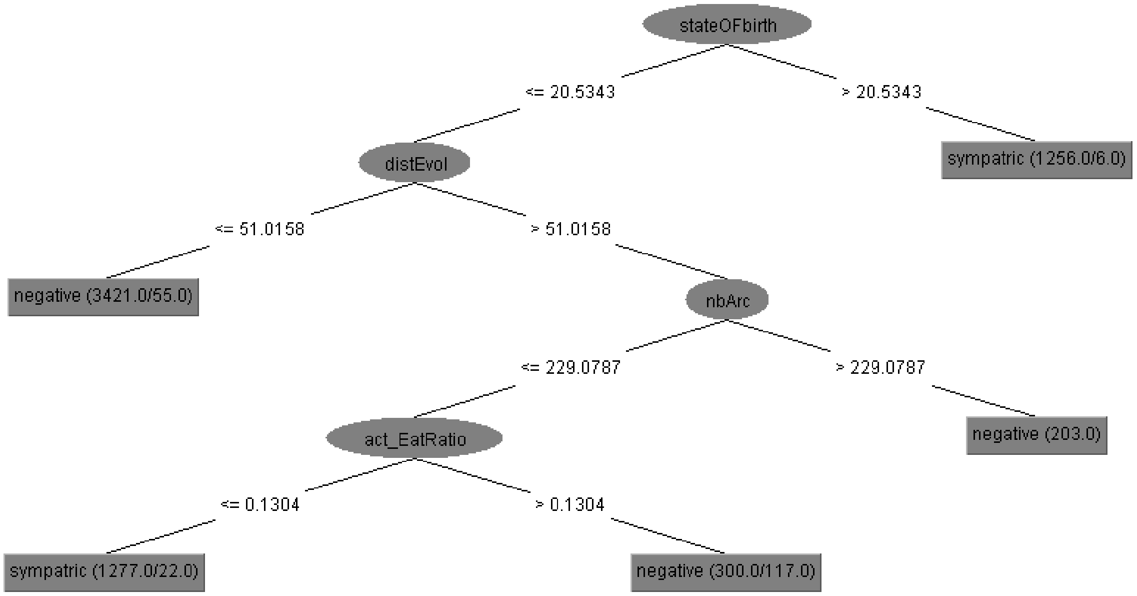


Figure 6.3 Decision tree with 5 rules, related to Dataset S34. Each leaf being a rule assigned to a specific class.

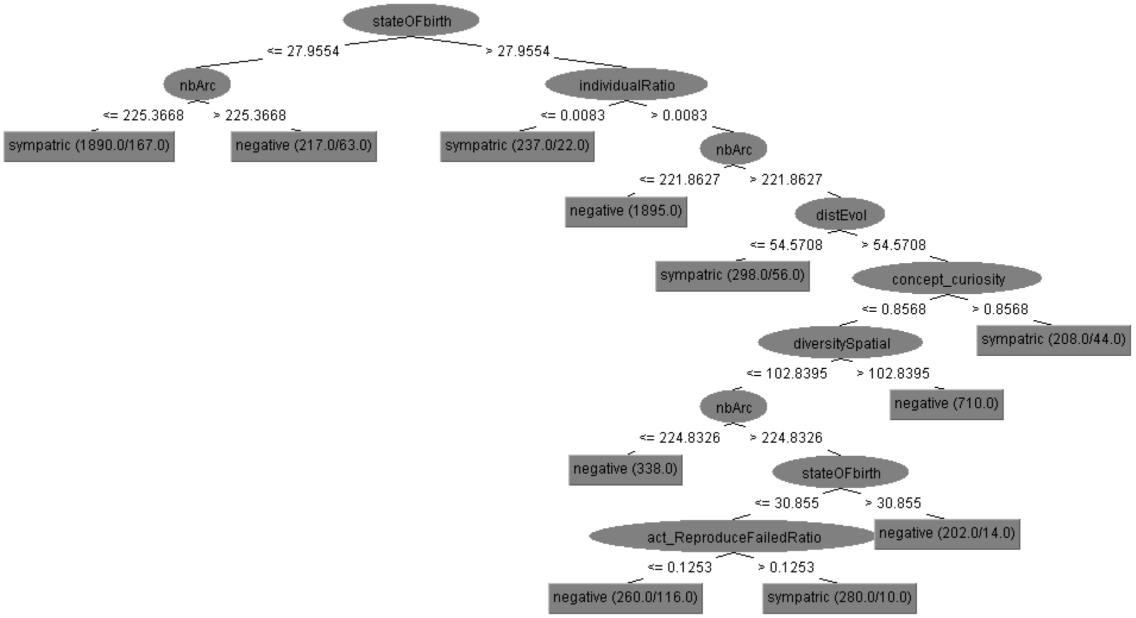


Figure 6.4 Decision tree with 11 rules, related to dataset S25. Each leaf being a rule assigned to a specific class.

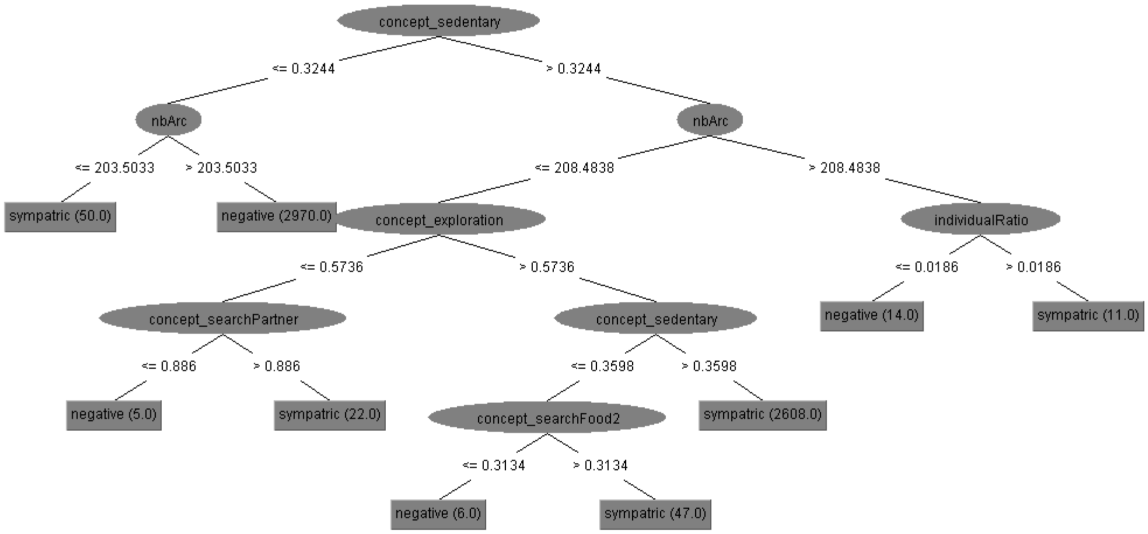


Figure 6.5 Decision tree with 9 rules, related to dataset S10. Each leaf being a rule assigned to a specific class.

6.4.2 Classification of all the datasets together

The results of all the five runs were combined together to create a dataset for finding the shared patterns between the sympatric species in different runs. 30% of the dataset was put aside to be used as the validation set. Two methods of feature selection were applied to our dataset, the Info Gain Attribute Evaluator implemented at Weka [Hall et al. 2009] with the Ranker search method, and the cfs subset Evaluator with Genetic Search method. Initially there were 81 attributes in the dataset. We started by removing the attributes having the lowest score returned by both attribute selection methods step by step. At the first step, we removed 56 attributes whose score was less than 30% in the cfs subset Evaluator with Genetic Search method, or whose rank in the Info Gain Attribute Evaluator with the Ranker search method was higher than 30. Only 25 attributes were kept after the first step of feature selection. The J48 classification method was applied to the dataset with the remaining set of attributes. The pruning method was also applied at the same time by increasing the minimum number of objects per leaf which allows to decrease the number of leaves and consequently to decrease the number of rules per class. For this purpose, there is an option in the input parameters of the J48 Weka classifier named “minNumObj”, where the minimum number of the instances per leaf can be defined. Increasing the minimum number of instances per leaf will lead to a more pruned tree with less rules, but at the same time it will decrease the total accuracy. Therefore, it is very important to choose this number such that, beside decreasing the number of rules, it can be ensured that the total accuracy remains at a reasonable level. This number is highly dependent on the structure of the dataset in term of the number of instances and attributes. For our dataset with around 41000 instances, we found the best results by choosing at least 400 instances per leaf. Despite removing a high number of attributes, the total accuracy only dropped around 1%, from 97.25% with 81 attributes, to 96.34% with only 25 attributes. The number of rules also decreased by 13 rules, from 69 with 81 attributes to 56 with 25 attributes. The feature removing step was repeated by removing 5 other attributes with the lowest score in both attribute selection methods. J48 classification method was applied again to the dataset with remaining 20 attributes, and the decision tree was pruned to lower the number of rules. The total accuracy slightly decreased to 94.95% and the number of rules dropped to 42. These steps were repeated three more times and 13, 11, and 9 attributes were selected respectively after each step. Table 6.8 shows the results of the classification after each step of removing attributes for 10 fold cross validation and the validation set.

All	#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	81	69	97.25%	0.97	0.96	0.99	0.99	96.98%	0.97	0.96	0.97	0.96
	25	56	96.34	0.96	0.95	0.99	0.99	95.91%	0.96	0.95	0.99	0.99
	20	42	94.95%	0.95	0.94	0.98	0.98	93.21%	0.93	0.93	0.98	0.98
	13	30	92.95%	0.92	0.93	0.97	0.97	91.65%	0.91	0.92	0.97	0.97
	11	20	91.03%	0.89	0.92	0.97	0.97	90.86%	0.90	0.91	0.95	0.95
	9	17	89.67%	0.89	0.90	0.95	0.95	89.03%	0.88	0.90	0.95	0.95

Table 6.8 The classification results of all datasets combined together, using J48 decision tree. Pruning techniques and feature removing methods are applied step by step to minimize the number of rules and attributes.

Error! Reference source not found. shows the decision tree returned by the J48 classifier on all the dataset combined together, with 11 attributes and 20 rules. The returned rules still need to be discussed with biologists to explain the patterns shared by all the sets of sympatric species in the different runs of the model.

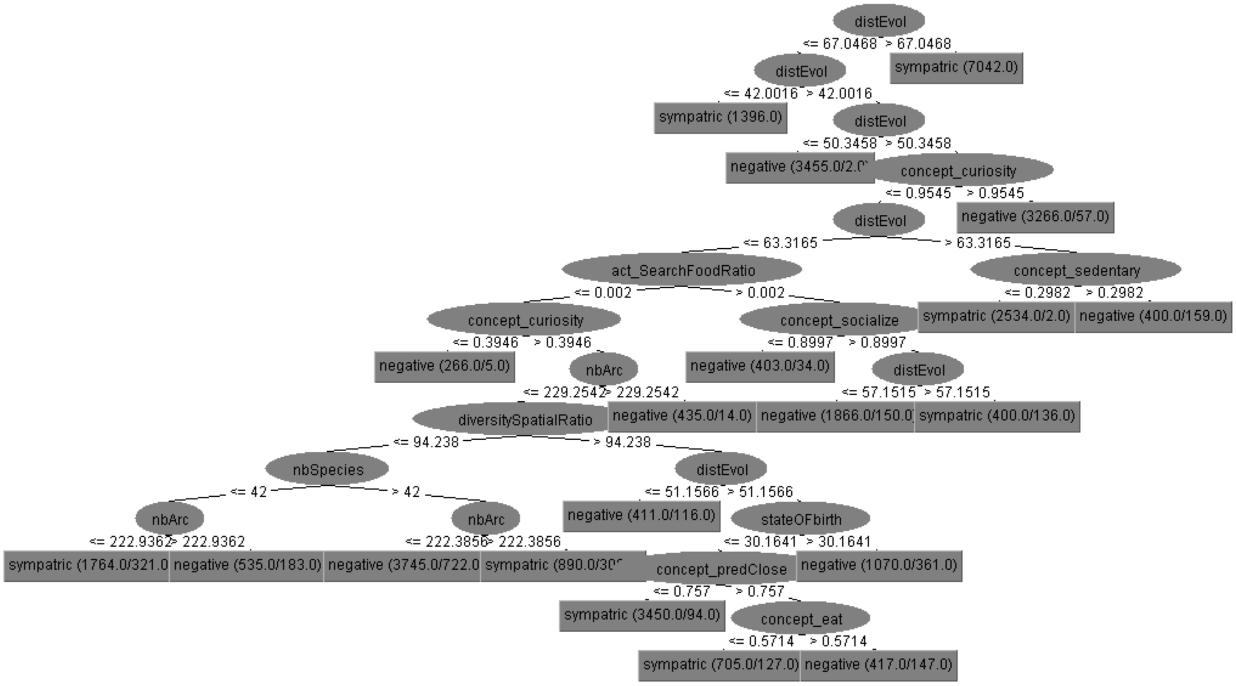


Figure 6.6 The decision tree returned by J48 classifier on all the dataset combined together, with 11 attributes and 20 rules.

6.4.3 Using four runs as train set and the fifth run as the validation set

To evaluate how generic of the rules we discovered are, we repeated our classification process five more times, each time combining the results of four out of the five data sets together and using them as the train set, and using the results of the fifth dataset as the validation set. 10 attributes were selected by applying the Info Gain Attribute Evaluator implemented in Weka [Hall et al. 2009] with the Ranker search method, and the cfs subset Evaluator with Genetic Search method using the step by step removing of the attributes with the lowest score as discussed in the sub-section 6.4.2. J48 decision tree and random forest classification methods were applied at each experiment. As it is expected the total accuracy of the validation set in this experiment is much lower than the total accuracy of 10 fold cross validation on the train set, due to the fact that the validation set is created from the results of a different run. It was observed that the random forest method strongly outperformed the J48 algorithm on the validation set and has always also a higher accuracy on the training set. The results of the five experiments are summarized on the Table 6.9, Table 6.10, Table 6.11, Table 6.12, and Table 6.13 respectively. The averages of the results of the classification on the five experiments are presented at the Table 6.14. It can be observed that with the random forest method, we can predict the occurrence of sympatric speciation on the train set, with the average accuracy of 99.97%, and for an unseen validation set from a different run with the average accuracy of 82.22% which is a quite high accuracy showing that our method is able to discover very generic rules that could be useful for biologists.

All datasets except s10 used as train set

All-s10	#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set –s10				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10	28	91.60%	0.92	0.90	0.97	0.97	41.00%	0.29	0.51	0.38	0.38

All-s10	#Features	#Rules	RandomForest----Train Set – 10 fold C.V.					Validation set –s10				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10		99.96%	1	1	1	1	90.85%	0.81	0.99	0.99	0.99

Table 6.9 The results of the classification using j48 and random forest classification methods. The results of all runs except s10 are used as the train set, and the results of run s10 are used as the validation set.

All datasets except s19used as train set

All-s19	#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set –s19				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10	70	99.85%	0.99	0.99	0.99	0.99	58.94%	0.08	0.94	0.51	0.51

All-s19	#Features	#Rules	RandomForest----Train Set – 10 fold C.V.					Validation set –s19				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10		99.97%	1	1	1	1	61.95%	0.15	0.94	0.74	0.74

Table 6.10 The results of the classification using j48 and random forest classification methods. The results of all runs except s19 are used as the train set, and the results of run s19 are used as the validation set

All datasets except s25usedas train set

All-s25	#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set –s25				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10	27	96.09%	0.95	0.95	0.99	0.99	56.37%	0.45	0.64	0.51	0.51

All-s25	#Features	#Rules	RandomForest----Train Set – 10 fold C.V.			Validation set –s25		
			Total	TP Rate	ROC Area	Total	TP Rate	ROC Area

			Total accuracy					Total accuracy				
	10		99.98%	1	1	1	1	71.33%	0.34	0.98	0.76	0.76

Table 6.11 The results of the classification using j48 and random forest classification methods. The results of all runs except s25 are used as the train set, and the results of run s25 are used as the validation set

All datasets except s33 as train set

All-s33	#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set –s33				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10	24	94.42%	0.92	0.95	0.97	0.97	75.76%	0.70	1	0.85	0.85

All-s33	#Features	#Rules	RandomForest----Train Set – 10 fold C.V.					Validation set –s33				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10		99.96%	0.99	1	1	1	97.84%	0.97	1	1	1

Table 6.12 The results of the classification using j48 and random forest classification methods. The results of all runs except s33 are used as the train set, and the results of run s33 are used as the validation set

All datasets except s34 as train set

All-s34	#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set –s34				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
	10	25	94.91%	0.94	0.95	0.98	0.98	66.27%	0.14	0.95	0.51	0.51

All-s34	#Features	#Rules	RandomForest----Train Set – 10 fold C.V.					Validation set –s34				
			Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	

	10		99.98%	1	1	1	1	89.14%	0.90	0.87	0.97	0.97

Table 6.13 The results of the classification using j48 and random forest classification methods. The results of all runs except s34 are used as the train set, and the results of run s34 are used as the validation set

The average of five experiments

#Features	#Rules	J48----Train Set – 10 fold C.V.					Validation set –				
		Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
10	34.8	95.37%	0.94	0.94	0.98	0.98	59.66%	0.33	0.80	0.67	0.67

#Features	#Rules	RandomForest----Train Set – 10 fold C.V.					Validation set –				
		Total accuracy	TP Rate		ROC Area		Total accuracy	TP Rate		ROC Area	
10		99.97%	0.99	1	1	1	82.22%	0.63	0.95	0.89	0.89

Table 6.14 The average results of five experiments of classification using j48 and random forest classification methods. At each experiment four out of five data sets were used as the train set, and the fifth data set were used as the validation set.

Chapter 7

CONCLUDING REMARKS

7.1 Limitation In Our Study[3]

There are a number of limitations in our study from a biological perspective. Our simulations are a tool to answer pertinent biological questions on a broad scope, and therefore cannot be used to directly model an ecological system with high specificity. For example, we can model a three-tier food chain as described in our study, but we cannot apply our results to directly answer the life histories of a set of species belonging to a three-tier food chain observed in nature. Such a simulation is orders of magnitude more complex in behavioural patterns and interactions, than what we are currently able to model and explore. In relation to preferential resource use by organisms, species living in sympatry have been documented in nature to express differential trophic structures, such as variation in teeth length and mouth size for amphibians, as a result of phenotypic polymorphisms, allowing them to better adapt to their environment [Skulason and Smith 1995]. Resource polymorphisms can also occur in many species living in sympatry, such as certain species of Neotropical fish, where they incorporate different strategies while foraging for the same types of algae [Skulason and Smith 1995]. In our simulations, there is no phenotypic variation at this stage of development that could induce an adaptive benefit for resource partitioning between populations of individuals, so we cannot at this time incorporate into our results the effects of selective pressures giving rise to phenotypic variation. However, the latest version of the EcoSim will allow some physical properties of the organisms to evolve, for example their size, and associated to that their need of energy consumption. This can be a simple model of the 'differential trophic structures' which later can be applied to our model.

---

[3] This part is the outcome of joint research

7.2 Conclusion

We extend an individual based evolving predator-prey ecosystem platform called "EcoSim" [Gras et al. 2009] to model a dual resource system. We investigated whether and in which conditions the selective pressures acting on foraging behaviors drove sympatric speciation. We have shown that the sympatric speciation criteria proposed originally by Coyne and Orr and modified by Bolnick and Fitzpatrick in 2007 were observed in 12 of our runs out of 20. We observed clear results showing some behavioral modifications occurring as a consequence of preferential resource usage. We also observed many cases where the sympatric speciation criteria described in the literature were fulfilled. Our novel individual-based behavioural model of evolution allows us to approach pertinent biological questions through a system with higher complexity than has previously been published in literature. Using several machine learning techniques, we extracted explicit rules that can predict with a very high accuracy the occurrence of sympatric speciation based on ecological factor observations. Moreover, we confirmed that the existence of a second food resource is determinant for the emergence of sympatric phenomenon. We also proved that our method is able to discover very generic rules which may later be used to structure empirical studies.

7.3 Future Direction

As the future work, we will continue on analysis of the results of the runs in which we observed sympatric speciation and will discuss the obtained rules which were found to be shared by sympatric species with biologists to identify broad phylogenetic patterns leading to speciation, potentially for a set of sympatric species in relation to resource partitioning, which can be used to aid or model natural studies and conservational efforts in the future.

We will also look at the runs in which we observed sympatric speciation as a whole, and compare those runs with the runs in which we did not observe any evidence of the happenstance of sympatric speciation to find more information about the probable environmentally or behaviourally differences which were lead to having sympatric speciation in some runs and no sign of this phenomenon in other runs.

Another future step also could be adding morphological ornaments to individuals, and also adding some attributes to the food resources such as size, to express differential trophic structures as observed in nature.

REFERENCES/BIBLIOGRAPHY

ARNEGARD, M.E. AND KONDRASHOV, A.S. 2004. Sympatric speciation by sexual selection alone is unlikely. *Evolution; international journal of organic evolution 58,* 2, 222–37.

BARRACLOUGH, T.G. AND NEE, S. 2001. Phylogenetics and speciation. *Trends in ecology & evolution 16,* 7, 391–399.

BARRACLOUGH, T.G., VOGLER, A.P., HARVEY, P.H., AND BARRACLOUGH, G. 1998. Revealing promote speciation. *353*, 1366, 241–249.

BERLOCHER, S.H. AND FEDER, J.L. 2002. Sympatric speciation in phytophagous insects: moving beyond controversy? *Annual review of entomology 47,* 773–815.

BOLNICK, D. 2004. Waiting for sympatric speciation. *Evolution 58,* 4, 895–899.

BOLNICK, D.I. AND FITZPATRICK, B.M. 2007. Sympatric Speciation: Models and Empirical Evidence. *Annual Review of Ecology, Evolution, and Systematics 38,* 1, 459–487.

BUSH, G.L. 1994. Sympatric speciation in animals: new wine in old bottles. *Trends in ecology & evolution 9,* 8, 285–8.

CERGHET, A. 2013. *Sympatric Speciation – Investigations in an Ecosystem Simulation (EcoSim) Using a Novel Individual-based Behavioural Model*. .

CHAWLA, N. AND BOWYER, K. 2002. SMOTE: synthetic minority over-sampling technique. *Artificial Intelligence 16,* 1, 853–867.

COYNE, J. AND ORR, H. 2004. *Speciation*. .

COYNE, J.A. 2007. Quick guide Sympatric speciation. *Current Biology 17,* 18, 787–788.

COYNE, J.A. AND ORR, H.A. 1998. The evolutionary genetics of speciation. *Philosophical transactions of the Royal Society of London Series B Biological sciences 353,* 1366, 287–305.

DARWIN, C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. *Murray, London*.

DINGLE, C., POELSTRA, J.W., HALFWERK, W., BRINKHUIZEN, D.M., AND SLABBEKOORN, H. 2010. Asymmetric response patterns to subspecies-specific song differences in allopatry and parapatry in the gray-breasted wood-wren. *Evolution; international journal of organic evolution 64*, 12, 3537–48.

DODD, D. 1989. Reproductive isolation as a consequence of adaptive divergence in Drosophila pseudoobscura. *Evolution 43*, 6, 1308–1311.

EGRI-NAGY, A. AND NEHANIV, C. 2003. Evolvability of the genotype-phenotype relation in populations of self-replicating digital organisms in a Tierra-like system. *Advances in Artificial Life*.

FARAHANI, Y.M. 2010. Complexity and Chaos Analysis of a Predator-Prey Ecosystem Simulation. *COGNITIVE 2010, The …*, 52–59.

FEDER, J.L. AND FILCHAK, K.E. 1999. It's about time: the evidence for host plant-mediated selection in the apple maggot fly, Rhagoletis pomonella, and its implications for fitness trade-offs in phytophagous insects. *Entomologia Experimentalis et Applicata 91*, 1, 211–225.

GOLESTANI, A. AND GRAS, R. 2012. Identifying Origin of Self-Similarity in EcoSim, an Individual-Based Ecosystem Simulation, using Wavelet-based Multifractal Analysis. *Proceedings of the World Congress on Engineering and Computer Science*.

GRAS, R., DEVAURS, D., WOZNIAK, A., AND ASPINALL, A. 2009a. An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model. *Artificial life 15*, 4, 423–63.

GRAS, R., DEVAURS, D., WOZNIAK, A., AND ASPINALL, A. 2009b. An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model. *Artificial life 15*, 4, 423–63.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND H., I. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations 11*, 1.

HE, H. AND GARCIA, E.A. 2009. Learning from Imbalanced Data. *Knowledge and Data Engineering, IEEE Transactions on 21*, 9, 1263–1284.

HOSKIN, C.J., HIGGIE, M., MCDONALD, K.R., AND MORITZ, C. 2005. Reinforcement drives rapid allopatric speciation. *Nature 437*, 7063, 1353–6.

HRABER, P.T., JONES, T., AND FORREST, S. 1997. The ecology of echo. *Artificial life 3*, 3, 165–90.

JIANG, T., FENG, J., SUN, K., AND WANG, J. 2008. Coexistence of two sympatric and morphologically similar bat species Rhinolophus affinis and Rhinolophus pearsoni. *Progress in Natural Science 18*, 5, 523–532.

KONDRASHOV, A.S. AND MINA, M. V. 1986. Sympatric speciation: when is it possible? *Biological Journal of the Linnean Society 27*, 3, 201–223.

LINNAEUS, C. 1758. *Systema Naturae*. .

LIU, S., PATEL, R.Y., DAGA, P.R., ET AL. 2012. Combined rule extraction and feature elimination in supervised classification. *IEEE transactions on nanobioscience 11*, 3, 228–36.

LU, G. AND BERNATCHEZ, L. 1999. Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes. *Coregonus clupeaformis 53*, 5, 1491–1505.

MASHAYEKHI, M. AND GRAS, R. 2012. Investigating the Effect of Spatial Distribution and Spatiotemporal Information on Speciation using Individual-Based Ecosystem Simulation. *GSTF International Journal on Computing 2*, 1, 98–103.

MASHAYEKHI, M. AND GRAS, R. 2013. Rule Extraction from random forest: RF+HC method. *to be appeared in the international conference on machine learning, ICML 2014.*

MAYR, E. 1996. What Is a Species , and What Is Not ? *Philosophy of science 63*, 2, 262–277.

OFRIA, C. AND WILKE, C.O. 2004. Avida: a software platform for research in computational evolutionary biology. *Artificial life 10*, 2, 191–229.

OLLERTON, J. 2005. Flowering time and the Wallace Effect. *Heredity 95*, 181–182.

RICE, W. AND HOSTERT, E. 1993. Laboratory experiments on speciation: what have we learned in 40 years? *Evolution 47*, 6, 1637–1653.

SCHLIEWEN, U., TAUTZ, D., AND PÄÄBO, S. 1994. Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature 386*, 629–632.

SCHLUTER, D. 2001. Ecology and the origin of species. *Trends in ecology & evolution 16*, 7, 372–380.

SEDEHI, S. 2012. Studying Mate Choice and Species Extinction using EcoSim. http://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1322&context=etd.

SKULASON, S. AND SMITH, T. 1995. Resource polymorphisms in vertebrates. *Trends in Ecology & Evolution 10*, 9.

STÖVER, B.C. AND MÜLLER, K.F. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics 11*, 7.

TAKIMOTO, G., HIGASHI, M., YAMAMURA, N., AND URL, S. 2000. A Deterministic Genetic Model for Sympatric Speciation by Sexual Selection A DETERMINISTIC GENETIC MODEL FOR SYMPATRIC SPECIATION BY SEXUAL SELECTION. *54*, 6, 1870–1881.

THEARLING, K. AND RAY, T. 1994. Evolving multi-cellular artificial life. *Artificial Life IV IV*, July, 283–288.

THIBERT-PLANTE, X. AND HENDRY, A P. 2011. Factors influencing progress toward sympatric speciation. *Journal of evolutionary biology 24*, 10, 2186–96.

TURELLI, M., BARTON, N.H., AND COYNE, J.A. 2001. Theory and speciation discuss recent examples of progress in each of these areas . *16*, 7, 330–343.

VIA, S. 2001. Sympatric speciation in animals: the ugly duckling grows up. *Trends in ecology & evolution 16*, 7, 381–390.

WILSON, A B., NOACK-KUNNMANN, K., AND MEYER, A. 2000. Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: sexual selection versus ecological diversification. *Proceedings. Biological sciences / The Royal Society 267*, 1458, 2133–41.

APPENDICES

Appendix A

The list of the initial attributes, used for creating the datasets, and a short description about each attribute.

Id	Attribute	Description
1	nbSpecies	Total number of currently alive
2	nbIndividual	The total prey population size
3	individualRatio	Species population size , divided by total population size
4	birthRatio	Total number of new born individuals, divided by species population size
5	interBreedingRatio	Number of interbreeding events (new born individuals with parents from different species), divided by the species population size
6	deadRatio	Number of dead individuals, divided by the total number of individuals in that species
7	deadAgeRatio	Number of dead individuals due to old age, divided by total number of deaths in the species
8	deadEnergyRatio	Number of dead individuals, due to lack of energy, divided by total number of deaths in the species
9	deadKilledRatio	Number of killed individuals, divided by total number of deaths in the species
10	deadAge	Average death age in a species
11	deadEnergy	The average energy of dead individuals in a species
12	Entropy	Diversity of alleles for all loci based on an entropy calculation
13	diversitySpatial	Dispersal level of individuals based on the average distance towards the species center
14	diversitySpatialRatio	the square roots of sum of the square of actual distances of each individual from the species center, divided by the total number of individuals
15	distEvol	Average genetic distance between the reference genome (origin) and the current genomes
16	stateOFbirth	The amount of energy transferred to the child from parent at the birth time
17	Age	The average age of individuals in the species
18	Energy	The average energy of individuals in the species
19	Speed	The average speed of individuals in the species
20	Compactness	The average number of individuals per cell
21	nbArc	Average number of arcs (genes) in the FCM of individuals
22	act_EscapeRatio	Percentage of population that chose Escape action
23	act_SearchFoodRatio	Percentage of population that chose search for food 1action and succeed
24	act_SearchFoodFailedRatio	Percentage of population that chose search for food 1action and failed
25	act_SearchFood2Ratio	Percentage of population that chose search for food 2 action and succeed
26	act_SearchFoodFailed2Ratio	Percentage of population that chose search for food 2 action and failed
27	act_SocializeRatio	Percentage of population that chose socialization action and succeed
28	act_SocializeFailedRatio	Percentage of population that chose socialization action and failed

29	act_ExplorationRatio	Percentage of population that chose exploration action
30	act_WaitRatio	Percentage of population that chose wait action
31	act_EatRatio	Percentage of population that chose eat 1 action and succeed
32	act_EatFailedRatio	Percentage of population that chose eat 1 action and failed
33	act_Eat2Ratio	Percentage of population that chose eat 2 action and succeed
34	act_EatFailed2Ratio	Percentage of population that chose eat 2 action and failed
35	act_ReproduceRatio	Percentage of population that chose reproduction action and succeed
36	act_ReproduceFailedRatio	Percentage of population that chose reproduction action and failed
37	reprodFailed_age	The average age of individuals which failed to complete the reproduction action
38	reprodFailed_energy	The average energy of individuals which failed to complete the reproduction action
39	parent1_reproductionAge	The average age of parents 1 for the reproduction action
40	parent1_reproductionEnergy	The average energy of parents 1 for the reproduction action
41	parent2_reproductionAge	The average age of parents 2 for the reproduction action
42	parent2_reproductionEnergy	The average energy of parents 2 for the reproduction action
43	DistMating	The average genetic distance between mates
44	reasonReproduceFailed_Energy	The amount of unsuccessful reproduction actions due to lack of energy, divided by the total number of unsuccessful reproduction actions
45	reasonReproduceFailed_NoPartner	The amount of unsuccessful reproduction actions due to no available partner, divided by the total number of unsuccessful reproduction actions
46	reasonReproduceFailed_PartnerEnerg	The amount of unsuccessful reproduction actions where the reason is that partner does not have enough energy, divided by the total number of unsuccessful reproduction actions
47	reasonReproduceFailed_PartnerActed	The amount of unsuccessful reproduction actions where the reason is that partner has already acted, divided by the total number of unsuccessful reproduction actions
48	reasonReproduceFailed_PartnerAction	The amount of unsuccessful reproduction actions where the reason is that partner has chosen a different action, divided by the total number of unsuccessful reproduction actions
49	reasonReproduceFailed_PartnerDist	The amount of unsuccessful reproduction actions where the reason is that partner distant is greater than distance mating threshold, divided by the total number of unsuccessful reproduction actions
50	concept_predClose	Average activation level of predator-close concept
51	concept_predFar	Average activation level of predator-far concept
52	concept_foodClose	Average activation level of Food1-close concept
53	concept_foodFar	Average activation level of Food1-far concept
54	concept_foodClose2	Average activation level of Food2-close concept
55	concept_foodFar2	Average activation level of Food2-far concept
56	concept_friendClose	Average activation level of friend-close concept
57	concept_friendFar	Average activation level of friend-far concept
58	concept_energyLow	Average activation level of energy-low concept
59	concept_energyHigh	Average activation level of energy-high concept
60	concept_foodLocalHigh	Average activation level of local food1-highconcept
61	concept_foodLocalLow	Average activation level of local food1-low concept

62	concept_foodLocalHigh2	Average activation level of local food2- high concept
63	concept_foodLocalLow2	Average activation level of local food2- low concept
64	concept_partnerLocalYes	Average activation level of partnerlocal-yes concept
65	concept_partnerLocalNo	Average activation level of partnerlocal-no concept
66	concept_fear	Average activation level of fear concept
67	concept_hunger	Average activation level of hunger concept
68	concept_searchPartner	Average activation level search for partner concept
69	concept_curiosity	Average activation level of curiosity concept
70	concept_sedentary	Average activation level of sedentary concept
71	concept_satisfaction	Average activation level of satisfaction concept
72	concept_nuisance	Average activation level of nuisance concept
73	concept_escape	Average activation level of escape concept
74	concept_searchFood	Average activation level of search for food1 concept
75	concept_searchFood2	Average activation level of search for food2 concept
76	concept_socialize	Average activation level of socialize concept
77	concept_exploration	Average activation level of exploration concept
78	concept_wait	Average activation level of wait concept
79	concept_eat	Average activation level of eat1 concept
80	concept_eat2	Average activation level of eat2 concept
81	concept_reproduce	Average activation level of reproduction concept

VITA AUCTORIS

NAME: Sara Bandehbahman
PLACE OF BIRTH: Esfahan, Iran

YEAR OF BIRTH: 1981

EDUCATION: 22 Bahman High School, ShahinShahr, Iran, 1997

TolooeKhorshid Pre-University School, ShahinShahr, Iran, 1998

University of Najaf Abaad, B.Sc., Najaf Abaad, Iran, 2002

University of Windsor, M.Sc., Windsor, ON, 2013