VISUALISING SPECIATION IN MODELS OF CICHLID FISH

ROSS CLEMENT

Department of Artificial Intelligence and Interactive Multimedia,
Harrow School of Computer Science
University of Westminster
Email: clemenr@wmin.ac.uk

Abstract: An Agent-Based model of speciation in cichlid fish has been implemented. When run, this generates large amounts of trace data in which speciation is an implicit, near unobservable, processes. Fuzzy C-Means Clustering is used to identify species extant at the end of simulation, and the power set of these species is the potential set of ancestral species. Membership values for all fish in each of these theoretical ancestral species are calculated, and total set membership for each of these species is plotted against time. The resulting graph is to be a clear visualisation of the process of speciation, and the appearance and disappearance of intermediate species. Our approach allows the visualisation of speciation resulting in larger numbers of final species than was possible using previous techniques based on measuring correlations between explicit properties of modeled organisms, and is also unaffected by changes to the properties used to model fish.

Keywords: Evolution, Speciation, Data Visualisation, Fuzzy Sets, Cichlids

1. INTRODUCTION

The Cichlid fish are one of the great mysteries of evolution [Barlow, 2000]. In Lake Victoria, hundreds of species of Cichlid have evolved from one or two ancestors in approximately 14,000 years [Seehausen, 2002]. In the tiny crater lake Barombi-Mbo, Cichlid fish have speciated despite any observable physical separation between populations [Schliewen, Tautz, & Pääbo, 1994].

Computer models investigating plausible hypotheses for speciation in Cichlids [Turner & Burrows, 995; Lande, Seehausen, & van Alpen, 2001], and sympatric (without physical barriers) speciation [e.g. Kondrashov & Kondrashov, 1999] have concentrated on the division of one species into two. In such models speciation can be observed by manual viewing of individual (modelled) fish, or by measuring emergent correlation between two fish properties (such as colour and female preference) using a simple measure such as Pearson's correlation coefficient [e.g. Kondrashov & Kondrashov, 1999] or by graphing explicit numerical characteristics of individuals [van Doorn & Weissing, 1999]. In such small-scale simulations, simple measures lead to an easily understood visualisation showing that speciation has occurred, and a simple trace of the process of speciation. These methods are not suitable for simulation of the evolution of large species flocks from single (or a few) ancestors as detailed information about the number, and timing, of speciation events is not revealed. This is unfortunate as the true mystery of Cichlid evolution is exactly how these species flocks arise.

A large scale agent based simulation has been built for the investigation of simulations of Cichlid speciation [Clement, *in prep*]. This systems includes environments that are less abstract than previous simulations, and which are designed to model the environment and characteristics of Lake Victoria rock-living Cichlids. The system allows very flexible creation of models, with agents used to model fish, and properties of the environment such as food sources. The shoreline of Lake Victoria, where rocky regions are separated by sandy, or muddy, regions where rock Cichlids are not found [Seehausen, 1996]. In the simulation system, this is modelled by multiple agent arenas (representing individual rocky reefs), with the (parameterised) possibility of fish migrating between these reefs. With a sufficient number of sufficiently different types of food sources, large numbers of species arise in the simulation. However, as the number of species rises, it quickly becomes impossible to manually extract objective traces of the number and histories of species. This is made more difficult by the generality of the simulation system itself. Fish can be designed using a number of different modelling methods, including the choice of directly modelling numerical phenotypes, or using genetic based models with loci, alleles, and genetic linkage. Hence any general method for tracing species must be independent of fish properties.

Our agent based simulation is most similar to agent based simulations used in Ecology [e.g. Ginot, Le Page, & Souissi, 2002] except that our aim is to understand Cichlid fish speciation, rather than the ecology of the lakes. However, it is extremely unlikely that speciation can be understood (or even

 ISBN 3-936150-25-7

has much meaning) without a detailed understanding, and modelling, of the underlying ecology of the fish.

There is no general agreement in Biology as to what a species is. Popular definitions of species as being groups capable of interbreeding, but not being capable of interbreeding with organisms outside the group are not applicable to natural systems such as Cichlid flocks. Hence a large number of species concepts have been developed [e.g. Paterson, 1993; Futayama, 1998]. Different from observations of natural systems, simulations allow the reproductive history of a fish to be traced both forwards and backwards in time, for as many generations as the simulation is run. Hence, in this work we adopt a new species concept. Two fish of the same species are likely to have common descendants many generations on, and common ancestors many generations back. The research reported in this paper describes the use of Fuzzy methods for the tracking of the emergence and disappearance of species during simulations according to this species concept.

2. METHODS

The data visualisation [e.g. Fayyad *et al,* 2002] method described in this paper is run independently of the simulation program. A simulation is performed by building a model, which includes a definition of the number of agent arenas and then the assignment of agents (both fish, and environmental agents such as food sources) to arenas. Both arenas and agents typically have large numbers of parameters, including frequently building agents by selection of building blocks. The simulation is then run for a predefined time period, and a trace of the simulation is saved on disk. As well as other information, this simulation includes records of all fish born, and their parents. Hence the exact ancestral history of any fish is known.

In order to visualise speciation, we first need to establish the number, and membership of species in the end-state of the simulation. There is no explicit species marker, and hence species groups need to be discovered from implicit patterns in the final population. In biology, species concepts (descriptions of what is and is not a species) is a highly contentious issue, and there is no general agreement on how a species should be defined. In this work, we use two separate species concepts. First, fish from the same species are likely to share common ancestors over (relatively) recent times, while those in different species will only share ancestors at much earlier times. Secondly, if we observe fish breeding, then the two fish breeding are likely to be of the same species.

To obtain the number and memberships of final species groups a critical time period is chosen (by the user, typically about 2000 time steps – each step being roughly equivalent to a week) back from the end of the simulation. All fish from this critical time period which have surviving descendants are found. A binary vector is created for each surviving (at the end of simulation) fish, with a bit for each potential ancestor at the critical time period. This bit is set to 1 for potential ancestors which are actually an ancestor for some fish F, and 0 if the potential ancestor is not actually an ancestor E.g. if we had eight potential ancestors at the critical time period, and a particular fish *F* was descended from potential ancestors 0, 2, and 7, then *F*'s vector is:

$$\text{vector(F)} = 10100001$$

These vectors are then clustered by Fuzzy C-Means Clustering (Bezdek, 1981). As the exact number of species is not known, clusterings are attempted from a (parameterised) minimum species number, up to a maximum species number. The first clustering where the sum of Euclidean distances between each vector and the set centres falls to less than 1.0 is taken as the correct speciation for this set of fish. The set of these (fuzzy set) final species is referred to as *Final={A, B, C, ...}*.

The next step is to discover the speciation history that lead to these species (e.g. *A, B,* and *C*) being present in the final steady state (which occurs in simulations, though is unlikely to happen in real life). To do this we first propose a set of potential ancestral species, which may have existed during speciation. If the set of initial species is *S*, then the set of potential species that may have occurred during speciation is the power set S^*. E.g. if *Final={A,B,C},* then S^**={ {}, {A}, {B},{A,B},{C},{A,C},{B,C},{A,B,C}}*. A set such as *{A,B}* represents a species that was the ancestor of final species *A* and *B*, but was not an ancestor of *C*. *{}* represents a species that was not an ancestor of any final species (i.e. "any other" species), and *{B}* represents the final species B itself.

It is impossible to apply crisp species labels to fish in the process of speciation. At some point in the simulation, the species *{A,B}* will exist, and at some later point, this species may be absent, and the species *{A},* and *{B}* will be present, but this is not an instantaneous event. Speciation is a process which takes time, and the aim of this research is to visualise this process. To track the history of species, we then define a method of calculating the (fuzzy) membership of each fish in each species, and then track the total sizes of these sets over time. This allows us to plot the history of species (represented by fuzzy sets) without having to assign crisp species labels to individual fish.

After fuzzy clustering, only fish alive at the end of the simulation have defined species membership. And, these are only memberships in the final species (e.g. *{A}, {B},* and *{C}*), not the power set of potential ancestral species. Fish from earlier times are labelled by summing the membership weights of all their descendants, and then normalising these weights so that the largest such weight is 1.0). E.g. a fish *F* that has 27.7 *A* descendants, 35.7 *B* descendants, and 0.3 *C* descendants is given weights $w_A(F)=0.776$, $w_B(F)=1.0$, $w_C(F)=0.008$.

To find set memberships of all fish in all potential ancestral species, the following calculation is performed. This calculates the membership of one fish in one potential ancestral species.

$$\mu_{S^*_i}(F_j) = \prod_{Sp \in S^*_i} w_{Sp}(F_j) \times \prod_{Sp \notin S^*_i} w_{Sp}(F_j) \quad (1)$$

This calculation is based on the assumptions that the set of potential ancestors is an exhaustive set, and that all species are mutually exclusive.

Speciation is then observed by tracking two properties over time. First the total membership of each potential ancestral species is tracked over time. Secondly for each breeding where at least one descendant survived, a 'characteristic' species footprint of this breeding is generated by averaging the weights of the mother and father, and mapping this onto membership of the potential ancestral species using (1).

3. RESULTS

Results have been good in the sense that a reliable and repeatable (across different trials) method for tracking speciation has been created. The initial clustering into species is particularly reliable, usually resulting in clusters of fish which have identical ancestor sets at the critical time period, without ancestors being shared by any fish assigned to different species. The following two graphs show the visualisation of speciation in two different cases. Figure 1 shows speciation in a system with two different food sources sufficiently different to motivate speciation into two distinct species. Figure 2 shows speciation in the case of three sufficiently different food sources, and the emergence of three species.

In both cases, speciation appears more or less static for quite some time, before ancestral species disappear fairly suddenly. In the three species case, speciation was much faster than in the two species case. However, small changes to parameters of the model caused large differences in the speed, and result of speciation. Hence, no conclusions can currently be made from this until far more is known about the factors that lead to speciation.

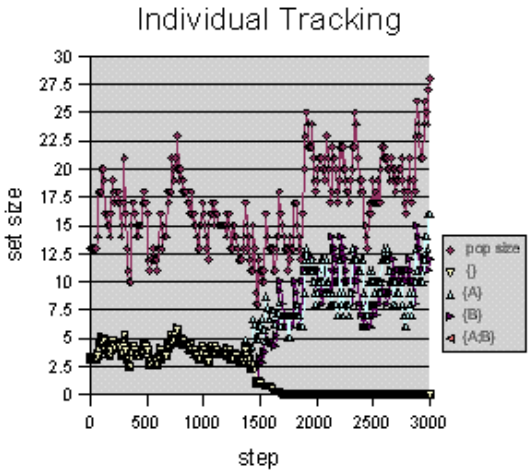

Figure 1a: Two-Species Individuals

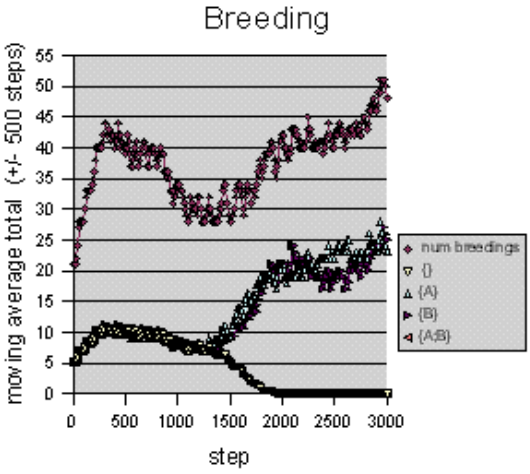

Figure 1b: Two-Species Breedings

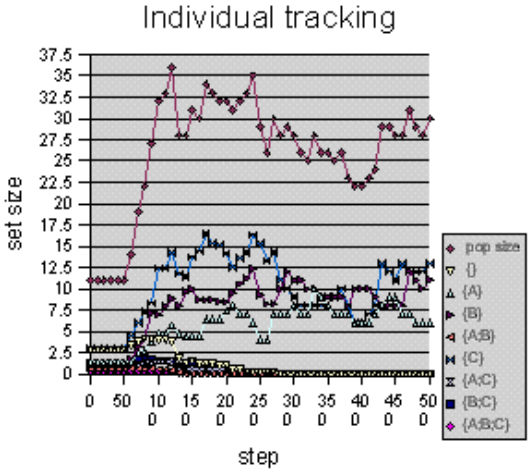

Figure 2a: Three-Species Individuals

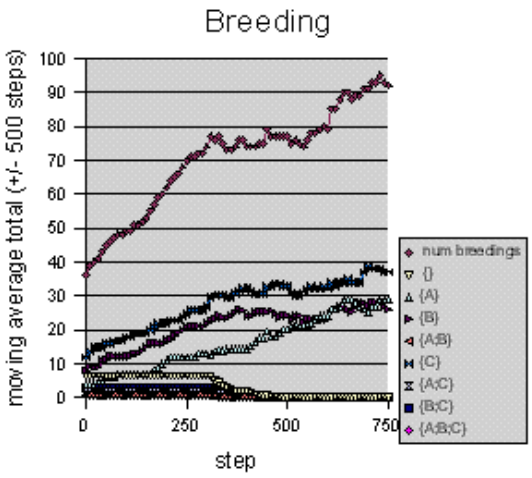

Figure 2b: Three-Species Breeding

4. CONCLUSIONS AND FUTURE WORK

The results show both clarity, and are reproduced across multiple independent trials. The initial fuzzy clustering of fish from the final surviving set performs far better than expected (and far better than several previous attempts to discover species). Typically, all fish from a single species share an identical set of ancestors (from the critical period), and no fish from different species share ancestors from this period. This is partially a result of designing systems that result in stable multi species populations, and the careful (and often experimental) choice of time spans for the simulation such that hybridisation between species had effectively ceased well before the critical period. However, we feel the results clearly indicate that the correct set, and number, of species is being found.

It is more difficult to evaluate the quality of the tracking of speciation over time, without an exact definition of species. However, tracking species both in terms of individuals, and mating events, give broadly comparable results. This supports the claim that the patterns being graphed are true representations of speciation, rather than aspects of speciation only applicable for a single species concept. Also, examination of the exact numerical traces of species (fuzzy set) membership, the exact point when an ancestor species finally disappears (total membership falls to zero) can be detected, allowing an objective measure of the time where speciation is complete.

It is planned to stop future work on this visualisation method and concentrate on using it to learn as much as possible about speciation in the circumstances where it can be used. Future developments in visualisation of speciation will be designed when there is a much better understanding of the modelling of speciation, and exactly what experiments need to be performed to learn more about the theoretical properties of various theories of speciation.

5. ACKNOWLEDGEMENTS

This work benefited greatly from discussions with a large number other researchers in both Computer Science, and Biology, including George Turner, Robert John, Peter Innocent, and Michael Walters.

6. REFERENCES

Barlow, G. 2000. *The Cichlid Fishes: Nature's Grand Experiment in Evolution.* Perseus.

Bezdek, J. 1981. *Pattern Recognition with Fuzzy Object Function Algorithms.* Plenum Press.

van Doorn, G. & F. Weissing F. 2001. "Ecological versus sexual selection models of sympatric speciation". *Selection* **2**: 17-40

Fayyad, U, Grinstein, G., & Wierse, A. 2002. *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann.

Futayama, D. 1998. *Evolutionary Biology*. Sinaeur.

Ginot V., Le Page C. & Souissi, S. 2002. "A multi-agents architecture to enhance end-user individual-based modelling". *Ecological Modelling* 157: 23-41.

Kondrashov, A. & Kondrashov, S. 1999. "Interactions among quantitative traits in the course of sympatric speciation". *Nature* **400**: 351-354.

Lande, R., Seehausen, O., & van Alphen, J. 2001. "Rapid sympatric speciation by sex reversal and sexual selection in Cichlid fish". *Genetica* **112/113**: 435-443

Paterson, H. 1993. *Evolution and the Recognition Concept of Species.* John Hopkins University Press.

Seehausen, O. 2002. "Patterns in fish radiation are compatible with Pleistocene desiccation of Lake Victoria and 14,6000 year history for its Cichlid species flock". *Proc R. Soc. Lond. B. Biol. Sci.* **269:** 491-7.

Seehausen, O. 1996. *Lake Victoria Rock Cichlids.* Verduijin Cichlids.

Schliewen, U, Tautz, d, Pääbo, S. 1994. "Sympatric speciation suggested by monophyly of crater lake Cichlids". *Nature* **368:**

Turner, G. & Burrows, M. T. 1995. "A model of sympatric speciation by sexual selection". *Proceedings of the Royal Society of London Series B Biological Sciences* **260:** 287-292.

7. BIOGRAPHY

Ross Clement received the degree of BSc in Cell Biology (1985), and MSc in Computer Science (1987) from the University of Auckland, New Zealand. He received the degree of Doctor of Engineering (1991) from The Toyohashi University of Technology, Japan. Since 1993, he has been first a Lecturer, then a Senior Lecturer in the Department of Artificial Intelligence and Interactive Multimedia of the Harrow School of Computer Science of the University of Westminster. His research interests include the Simulation of Cichlid Speciation and Evolution, and the application of Artificial Intelligence methods in Education.