Genetic Clustering for the Identification of Species

Jaimie Murdock
School of Informatics & Computing
919 E. 10th St.
Bloomington, IN 47408
jammurdo@indiana.edu

Larry S. Yaeger
School of Informatics & Computing
919 E. 10th St.
Bloomington, IN 47408
larryy@indiana.edu

ABSTRACT

Artificial life simulations can yield distinct populations of agents representing different adaptations to a common environment or specialized adaptations to different environments. Here we apply a standard clustering algorithm to the genomes of such agents to discover and characterize these subpopulations. As evolution proceeds new subpopulations are produced, which show up as new clusters. Cluster centroids allow us to characterize these different subpopulations and identify their distinct adaptation mechanisms. We suggest these subpopulations may reasonably be thought of as *species*, even if the simulation software allows interbreeding between members of the different subpopulations. Our results indicate both sympatric and allopatric speciation are present in the Polyworld artificial life system. Our analysis suggests that intra- and inter-cluster fecundity differences may be sufficient to foster sympatric speciation in artificial and biological ecosystems.

Categories and Subject Descriptors

I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems, intelligent agents*; H.1.1 [**Information Systems**]: Models and Principles—*Systems and Information Theory*; I.5.3 [**Pattern Recognition**]: Clustering—*algorithms, similarity measures*

General Terms

Algorithms, Experimentation, Theory

1. THE SIMULATION SOFTWARE

This research was carried out using Polyworld [4], a computational ecosystem evolving populations of agents controlled by artificial neural networks, the topology of which are encoded in the agents' genomes, along with a small number of "physiology" genes. Simulation parameters are identical to those presented in previous work on the evolution of neural complexity [5]. There are a total of 2,494 of these 8-bit genes in each of 29,564 agents, distributed over 30,000 time steps.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

2. THE CLUSTERING ALGORITHM

The clustering task can be divided into two subproblems: the distance function used to measure object similarity and the clustering algorithm used to partition objects.

To address the "curse of dimensionality" and force QT-Clust to focus on the most relevant genes, we take advantage of the fact that genes with a high impact on agent fitness will be selected for and conserved, while those which are inconsequential will trend towards a random distribution. We therefore use the information *certainty* (1 - entropy) of each gene, to weight each of the dimensions:

$$certainty(g) = 1 + \sum_{i=0}^{N_s} p(g_i)\ log_2(p(g_i))$$

where g is the gene, g_i are gene values (states), and N_s is the number of gene states. Probabilities are computed for 16 bins of 16 gene values, over the set of all agents extent during the evolutionary simulation.

Our distance metric is the certainty-weighted squared-Euclidean distance of normalized gene values (z-scores):

$$dist(x, y) = \sum_{i=0}^{N_g} (w_i(z(x_i) - z(y_i)))^2$$

where x and y correspond to two agents' genomes, N_g is the total number of genes in the genome, w_i is the certainty of each gene i, and $z(x_i)$ and $z(y_i)$ are the z-scores of gene i for each agent. Using z-scores addresses the fact that genes may only be expressed over a fraction of their possible range.

We use a version of the QT-Clust algorithm [1] with the addition of an algorithmic enhancement to allow for multiple cluster selection on each pass. QT-Clust bases its clustering analysis on cluster radius, ϵ, which we normalize in terms of cluster standard deviations, using the certainty weightings:

$$\epsilon(x) = x \sum_{i=0}^{N_g} w_i$$

3. RESULTS

We examined ϵ thresholds from 1.5 to 3, at increments of .25, but only discuss $\epsilon = 2.125$ here.

Figure 1 shows that clusters tend to be replaced serially over time. This supports a previous conjecture about the rapid spreading of "good enough" solutions throughout the population [5]. It also suggests that significantly different subpopulations exhibit reproductive isolation, even though the simulation does not explicitly forbid such pairings.

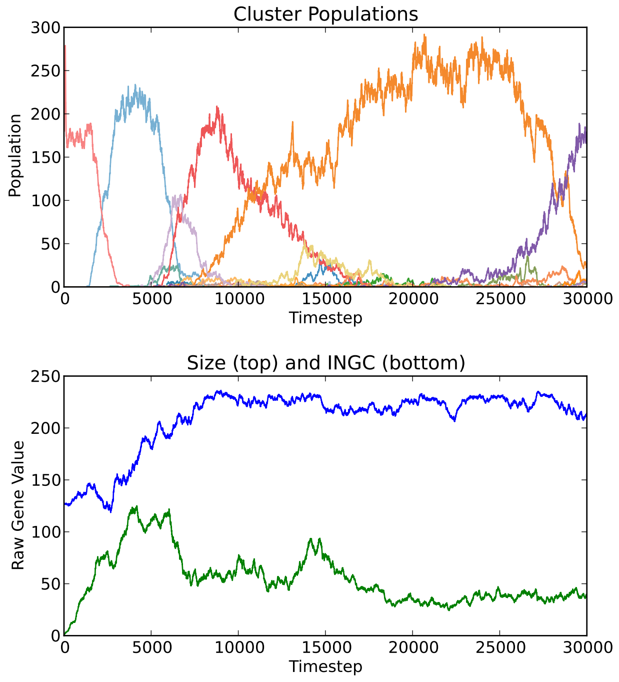

Figure 1: Cluster population over time for $\epsilon = 2.125$ and two high-certainty genes exhibiting different selection behaviors: size ($certainty = 0.3515$) and internal neural group count ($certainty = 0.2058$).

Figure 1 also shows two different gene selection patterns. The size gene shows a nearly monotonic selection pattern. Only the initial seed population has a relatively small size. By the time of the transition from the second major cluster to the third major cluster, size has plateaued.

Other genes are not so uniformly selected. The internal neural group count gene shows a pattern suggestive of punctuated equilibrium, that appears to be the result of species emergence and decline. At timestep 15,000 a sudden spike in the gene value corresponds to the rise and decline of two competing groups (light orange and blue) that emerge at the tail of the third dominant group. Without observation of cluster behavior, the source of this anomaly would be elusive.

4. DISCUSSION AND CONCLUSIONS

Since the simulation allows breeding between agents from different clusters perhaps they are best thought of as *proto-species*, however the temporal nature of cluster replacement and the fall and rise of subpopulations point to reproductive isolation between *species*, even when they overlap in space and time. This suggests the presence of sympatric speciation, despite the lack of any explicitly modeled mate preferences [3].

In proposing the use of clustering algorithms to identify biological species, Mallet [2] notes, “Clusters can remain distinct under relatively high levels of gene flow provided there is strong selection against intermediates; species will be maintained when selection balances gene flow.” In Polyworld, selection operates on variations in body plans and behaviors, derived from differing neural topologies. The mixing, due to genetic crossover, of neural topologies and body plans is liable to produce less viable neural dynamics and resulting behaviors in offspring. Thus the network topologies of Polyworld agents are likely to play a large factor in balancing gene flow, as evidenced by differing average complexity across clusters (not shown).

Parents	total	# children	# grandchildren
Diff Cluster	7335	1.87	3.58
Same Cluster	21600	2.03	4.10

Table 1: Reproductive success for parents from the same or different clusters using $\epsilon = 2.125$

To investigate the effects of cluster membership on reproductive success we examined the number of children and the number of grandchildren produced by pairs of agents from the same or from different clusters. Table 1 summarizes the results. Though the differences are not large, parents from the same cluster are clearly more fecund than parents from different clusters. Amplified across multiple generations it is easy to see how intra-cluster breeders will outperform inter-cluster breeders and produce ever more distinct subpopulations—species—even sympatrically. It seems likely that similar differences in fecundity, due to beneficial or detrimental genetic recombinations, could lead to sympatric speciation in biological organisms as well.

In movies showing cluster membership over time we see clusters emerge and persist alongside existing clusters world wide, in a sympatric fashion. But we also see evidence of allopatric speciation, with new clusters emerging in and coming to dominate one food patch before spreading to the other—in fact, having difficulty invading the second food patch. So evidence suggests both forms of speciation are present in these simulations. A sample movie can be found at: http://informatics.indiana.edu/larryy/cluster_movie.zip

5. REFERENCES

[1] L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research*, 9(11):1106–1115, 1999.

[2] J. Mallet. A species definition for the modern synthesis. *Trends in Ecology & Evolution*, 10(7):294 – 299, 1995.

[3] P. Todd and G. Miller. On the sympatric origin of species: Mercurial mating in the quicksilver model. In R. K. Belew and L. B. Booker, editors, *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 547–554. Morgan Kaufmann, San Mateo, CA, 1991.

[4] L. S. Yaeger. Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or Polyworld: Life in a New Context. In C. G. Langton, editor, *Proceedings of the Artificial Life III Conference*, pages 263–298. Addison-Wesley, Reading, MA, 1994.

[5] L. S. Yaeger, V. Griffith, and O. Sporns. Passive and Driven Trends in the Evolution of Complexity. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 725–732. MIT Press, Cambridge, MA, 2008.