Identifying Species by Genetic Clustering

Jaimie Murdock and Larry S. Yaeger

School of Informatics & Computing, Indiana University, Bloomington, IN 47408
jammurdo@indiana.edu, larryy@indiana.edu

Abstract

Complex artificial life simulations can yield substantially distinct populations of agents corresponding to different adaptations to a common environment or specialized adaptations to different environments. Here we show how a standard clustering algorithm applied to the artificial genomes of such agents can be used to discover and characterize these subpopulations. As gene changes propagate throughout the population, 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, and provide evidence of both sympatric and allopatric speciation in the Polyworld artificial life system. Analyzing intra- and inter-cluster fecundity differences and offspring production rates suggests that speciation is being promoted by a combination of post-zygotic selection (lower fitness of hybrid offspring) and pre-zygotic selection (assortative mating), which may be fostered by reinforcement (the Wallace effect).

Introduction

Artificial life simulations exhibit complex agent-based behaviors, which persist and evolve through genetic recombination and mutation. Unless explicit speciation is built into the simulation, identifying emergent species in these simulations is difficult, both theoretically and practically. Here we demonstrate a technique for identifying subpopulations of agents using a clustering algorithm to identify groups of agents with shared genetic attributes. The resulting clusters might reasonably be considered distinct species, and allow us to identify some of the different adaptation mechanisms adopted in the simulation. Examining the temporal distribution of these clusters allows us to better understand the evolutionary course of speciation and adaptation in our simulations, and may offer some insights into speciation in biological ecosystems.

Understanding speciation is one of the key problems in biology. Much debate centers around the role of *allopatric* (geographically isolated) vs *sympatric* (shared environment) species divergence. The significance and driving forces of sympatric speciation have been controversial since the ideas were introduced by Wallace (1899) and championed by Dobzhansky (1937). *Disruptive selection* (adaptation to distinct fitness peaks) in combination with *reinforcement* (the selection pressure that results from reduced fitness of hybrids; aka the *Wallace effect*) leads to *assortative mating* (a preference for related partners) thus providing a basis for sympatric speciation. Despite the simplicity and attractiveness of these ideas, the so-called Modern Synthesis largely discarded the idea of selective speciation, instead attributing divergence to more readily observable geographic isolation (Mayr and Provine, 1998), and a variety of models (reviewed in (Kirkpatrick and Ravigné, 2001)) have led many to conclude that sympatric speciation, while possible, will only be found under very limited circumstances (Felsenstein, 1981). However, though the jury is still out, empirical evidence for reinforcement driving sympatric speciation does exist ((Sætre et al., 1997; Ortiz-Barrientos et al., 2004; Silverton et al., 2005) and others) and recent theoretical and modeling work have suggested potential mechanisms (such as competition overwhelming selection towards a single method of resource utilization) for overcoming the perceived limitations on sympatry (Dieckmann and Doebeli, 1999; Kondrashov and Kondrashov, 1999; Van Doorn and Weissing, 2001). For high-level reviews see (Butlin and Tregenza, 1997; Tregenza and Butlin, 1999; Weissing et al., 2011). In this work, both *pre-zygotic* (pre-mating) and *post-zygotic* (post-mating) selection are observed, suggesting reinforcement may be playing a role in our speciation events—both sympatric and allopatric followed by population mixing.

In the life sciences clustering algorithms are applied in many areas, including the analysis of clinical information, phylogeny, genomics, and proteomics (Zhao and Karypis, 2005). Mallet (1995) proposed gene clustering as a preferred method for the rigorous identification of biological species (as opposed to taxonomic features). We seek to import these concepts and tools from the realm of biology into our artificial life work to help us better understand the evolutionary dynamics of our model ecosystem, though we believe there may be some general principles that apply to both artificial

and natural ecosystems.

The use of gene clustering for speciation has been explored in genetic algorithms by Hocaoglu and Sanderson (1995) and in computational ecosystems by Aspinall and Gras (2010). The Aspinall and Gras predator-prey simulator has some traits in common with ours, but defines two distinct agent classes which do not interbreed, and the clustering analysis is performed during the simulation and allowed to control reproductive success, thus allowing it to drive the speciation process. By contrast, there is no impact of cluster membership or genetic distance on reproductive success in the work reported here, and all gene clustering analysis is performed *post hoc*, after a simulation has run its course.

Clustering algorithms (reviewed in Hartigan (1975); Kaufman and Rousseeuw (2005)) rely upon two key elements: the distance function used to measure object similarity and the algorithm used to partition the data. The distance function must account for the "curse of dimensionality" (Bellman, 1957) intrinsic to high dimensional spaces in general and evolutionary algorithms employing large, high-dimensional genomes in particular. Clustering algorithms with a pre-specified number of clusters—such as k-means clustering (MacQueen, 1967)—though widely used, suffer from the simple fact that the number of clusters may not be known *a priori*.

Information theory (Shannon, 1948) allows us to partially alleviate the curse of dimensionality. Through the process of variation and selection those genetic dimensions which most affect an agent's fitness will be selected for and conserved, thus exhibiting low entropy across the population of agents, while those which are inconsequential will descend into a random distribution. By weighting genetic dimensions with *certainty* (i.e., 1 - entropy) those genetic features most significant to the agents' survival and reproduction will be emphasized during the partitioning into clusters, while spurious proximity in the inconsequential dimensions is ignored.

Algorithmically, we have chosen to use the QT (Quality Threshold) Clustering algorithm (Heyer et al., 1999; Scharl and Leisch, 2006), which clusters based on a maximum intra-cluster distance (diameter), rather than a set number of clusters.

The Artificial Life Software

This research was carried out using Polyworld (Yaeger, 1994), a computational ecology with a long history, in which populations of haploid agents evolve, each possessing a suite of primitive behaviors (move, turn, eat, mate, attack, light, focus) under continuous control of an Artificial Neural Network (ANN) employing (in this case) discrete-time, firing-rate neurons with synapses that adapt via Hebbian learning. The wiring diagram of the ANN is encoded in the organism's genome, via a statistical description of the number of neural groups of excitatory and inhibitory neurons, synaptic connection densities, regularity of connections, and learning rates. The only epistatic interaction between genes derives from the role played by the genes expressing the number of neural groups and the number of neurons in each group in controlling whether the corresponding inter-group and inter-neuron connections are expressed. For a detailed discussion of Polyworld's genetic encoding scheme, see (Yaeger, 1994).

Input to the ANN consists of pixels from a rendering of the scene from each agent's point of view. Output from the ANN consists of the aforementioned primitive behaviors. For the simulation discussed here, there are 2,486 genes devoted to specifying the neural topologies (but not synaptic weights) of ANNs with up to 217 neurons and 45,854 synaptic connections. The actual neuron count ranged from 14 to 163, with a mean of 48, and the synapse count ranged from 46 to 9,034, with a mean of 656. A small number of genes (8) characterize the agents' simple morphologies, metabolisms, and meta-genetics, in terms of size, strength, maximum speed, fraction of energy contributed to offspring, ID (green color component), mutation rate, number of crossover points, and lifespan. Thus there are 2,494 genes in all used in the clustering process.

All actions of the agents consume energy, so they must replenish their energy levels by seeking out and consuming food or by killing and eating other agents. Normally there are also per-neuron and per-synapse energy costs, but for consistency with some evolution-of-complexity experiments these were disabled for the results reported here. Reproduction occurs when two collocated agents simultaneously express their mating behaviors.

The simulation is initially seeded with a uniform population of agents that have the minimum number of neural groups and a nearly minimal number of neurons and synapses. While predisposed to some potentially beneficial behaviors, such as running towards food (green) and away from aggression (red; see (Yaeger, 1994) for details on color use in Polyworld), these seed organisms are not a viable species. I.e., without evolution they cannot sustain their numbers through their reproductive behaviors and will inevitably die out.

For these analyses the world was configured as in (Yaeger et al., 2008), with two barriers running 90% of the depth of the world, but left open for the remaining 10% of the world, so populations are able to mix relatively easily, but not with complete freedom. 80% of the food is grown in a patch occupying 40% of world depth at the open end of the barriers, 20% in a patch occcupying 10% of world depth at the closed end of the barriers. This layout may be seen in Figure 1.

As simulations progress both the structural architecture of the ANNs and the activation of every neuron at every time step for every agent may be recorded, thus permitting investigation into evolutionary trends in network structure and function (Yaeger et al., 2010). Agent genomes may also be recorded, and these recorded genomes serve as the basis for

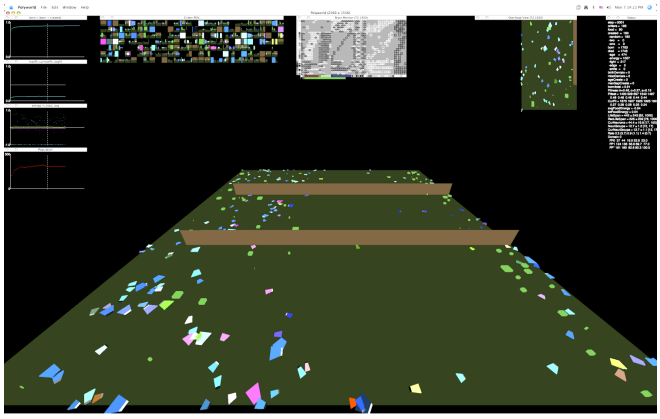

Figure 1: Polyworld simulation environment

the clustering analysis described here. Some genes exhibit smooth, general trends over the course of the simulation, but others demonstrate short, sharp changes that correspond to temporal cluster boundaries, as will be discussed later.

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. For the distance function, we used entropy-weighted Euclidean distance over each agent's genome. For the clustering algorithm, we used a variation of the QT-Clust algorithm (Heyer et al., 1999; Scharl and Leisch, 2006), with the addition of a new algorithmic improvement to allow for multiple cluster selection on each pass and a precalculation of point-wise distances for greater efficiency.

The Distance Function

Genomic data in artificial life simulations are afflicted by the curse of dimensionality (Bellman, 1957), and the current Polyworld genome consists of nearly 2,500 genes! Fortunately, the process of selection in evolutionary algorithms gives a way to identify genes which are likely to differentiate subpopulations. 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. By taking the information certainty (1 - Shannon Entropy) of each gene, the relative importance of each dimension may be used to weight the many dimensions:

$$H(g) = -\sum_{i=0}^{N_s} p(g_i)\ log_2(p(g_i))$$

$$certainty(g) = 1 - H(g)$$

where g is a specific gene, the g_i are the gene values (states), and N_s is the number of possible gene states. Probabilities were calculated for 16 bins of 16 gene values, capturing the full range of these 8-bit genes (0-255), over the entire population of 29,564 agents extent during the full evolutionary simulation.

While each gene of the Polyworld genome is specified by an 8-bit value, the full range of genetic values may not be expressed over the course of a simulation. In comparing genomic data, the difference along this distribution is more important than the raw score. To address this issue when calculating genetic distances between agents, we have normalized the measure of each gene dimension, by calculating the genes' z-scores:

$$z(x) = \frac{x - \mu}{\sigma}$$

where x is the raw gene value, μ is the mean value of that gene, and σ is the standard deviation of that gene's values).

After normalizing gene values to produce gene z-scores, distances are calculated between z-scores, weighting the relative importance of each gene by its certainty. Our distance metric is therefore the certainty-weighted squared-Euclidean distance of 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 and their genomes, N_g is the total number of genes in the genome, w_i is the certainty calculated for each specific gene i, and $z(x_i)$ and $z(y_i)$ are the z-scores of gene i in the genomes of agents x and y.

The QT-Clust Algorithm

Clustering algorithms rely upon the fixation of one or more variables: number of clusters, similarity of elements in the cluster, or number of elements in each cluster. Effective clusters should maximize inter-cluster distances, while minimizing intra-cluster distances (cluster diameter). Traditional k-nearest-neighbor approaches (MacQueen, 1967) require the number of clusters to be specified *a priori*. Additionally, these algorithms encounter the hubness phenomenon in which a centroid may be a common nearest-neighbor in Euclidean space, building large diameter clusters. This phenomenon is exacerbated by high-dimensionality (Beyer et al., 1999; Radovanović et al., 2010).

To avoid these issues, we have opted to use the QT-Clust algorithm (Heyer et al., 1999; Scharl and Leisch, 2006), which is a nearest-neighbor clustering approach fixing cluster diameter (ϵ), rather than the number of clusters. This algorithm is perticularly well suited for data discovery problems, such as gene analysis (the original use case). Adjustment of the cluster diameter parameter provides a means of controlling cluster fit that is both more intuitive and practical than algorithms requiring explicit specification of the number of clusters. (E.g., we are unlikely to have chosen

Algorithm 1: QT-Clust

Input: G, ϵ
Output: Clusters

if $|G| \leq 1$ **then**
  output G
else
  `// Cluster building`
  foreach $i \in G$ **do**
    $flag := TRUE$; $C_i := i$;
    while $flag$ *and* $C_i \neq G$ **do**
      find $j \in G - C_i : diameter(C_i \cup j)$ is min;
      if $diameter(C_i \cup j) > \epsilon$ **then**
        $flag := FALSE$
      else
        $C_i = C_i \cup j$
  `// Cluster selection`
  $C := C_0 ... C_{|G|}$;
  while $|C| > 0$ **do**
    identify set $P \in C$ with max cardinality;
    $G := G - P$;
    $C := X \in C : |X \cap P| = 0$;
    output P;
  $QT\_Clust(G, \epsilon)$

values of 8, 29, and 108 for the number of clusters we ended up focusing our attention on, but specifying cluster diameter in terms of standard deviations that produced these clusterings seemed reasonably natural.) The iterative approach used by QT-Clust also avoids issues of hubness common to nearest-neighbor clustering algorithms by creating an ϵ-neighborhood graph around each agent. The largest of these groupings is then selected and removed from the population to be re-clustered, thus eliminating the effect of outliers and hubs (Radovanović et al., 2010).

The algorithm has two stages. First, a cluster is built starting with each agent within the population (G). The cluster is built by adding the next closest agent to the cluster, until a threshold (ϵ) of maximum distance is reached. Cluster construction may be done in parallel for a significant speed increase. Then, each of these clusters is passed through a filtering step, which selects the largest candidate that does not overlap with a previously selected cluster, until no viable candidates remain. This multiple selection amortizes the time complexity of the original QT-Clust algorithm, while maintaining its quality control advantages. After filtering, unclustered elements are then reclustered within the remaining population until all elements are classified.

Results

We ran this algorithm on Polyworld simulation data containing 29,564 agents (distributed over 30,000 time steps), contained in 1.9GB of genomic data. Simulation parameters are identical to those presented in previous work on the evolution of neural complexity (Yaeger et al., 2008). While previous work has focused on general trends, combining the results of multiple runs and applying standard tests of statistical significance, here we wish to tease apart the dynamics of a particular simulation, and we are interested in the degree to which cluster analysis and a species/sub-population perspective can inform the understanding of those dynamics. We would expect the details of cluster/species formation to vary from run to run, even when nothing changes but the pseudo-random number generator's seed, and have seen hints of such variation in previous work on complexity trends.

ϵ	1.5	1.75	2	2.125	2.25	2.5	2.75
# clust	2063	750	108	29	8	3	3

Table 1: Resulting cluster counts for different ϵ thresholds

For the discussion below, we define ϵ as a factor of the sum of all certainty weightings:

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

This sum is equivalent to the weighted distance between two genomes which differ by 1 standard deviation on each dimension, due to z-score normalization. Thresholds were set between 1.5 and 3 times the sum of the certainty values, at increments of .25.

Behavior Across Different Thresholds

Table 1 shows the number of clusters identified for varying levels of ϵ. Figure 2a-c show the population of each cluster over time for $\epsilon = 2.0, 2.125, 2.25$. The progression from a large diameter to a smaller diameter shows each cluster splintering. Whether these show heirarchical clustering is a question for further empirical study.

Temporal Trends in Clusters

Figure 2b shows that while the larger clusters tend to be replaced serially over time, other, smaller clusters emerge, coexist with one or more of the larger clusters for extended periods of time, and are ultimately extinguished, suggesting the emergence, persistence, and decline of subordinate species. This also suggests we may be seeing reproductive isolation of sub-populations, despite the fact that Polyworld does not in any way inhibit cross-cluster reproduction. This could be due to pre-zygotic, assortative mating preferences (unpublished work suggests agents attend to both genetically and behaviorally determined color expressions) or to post-zygotic disruptive selection effects in a Dobzhansky-Muller manner—hybrid offpsring expressing neural architectures that are sub-optimal in themselves, or in combination with "physiological" characteristics that affect energy requirements. We look at both possibilities below.

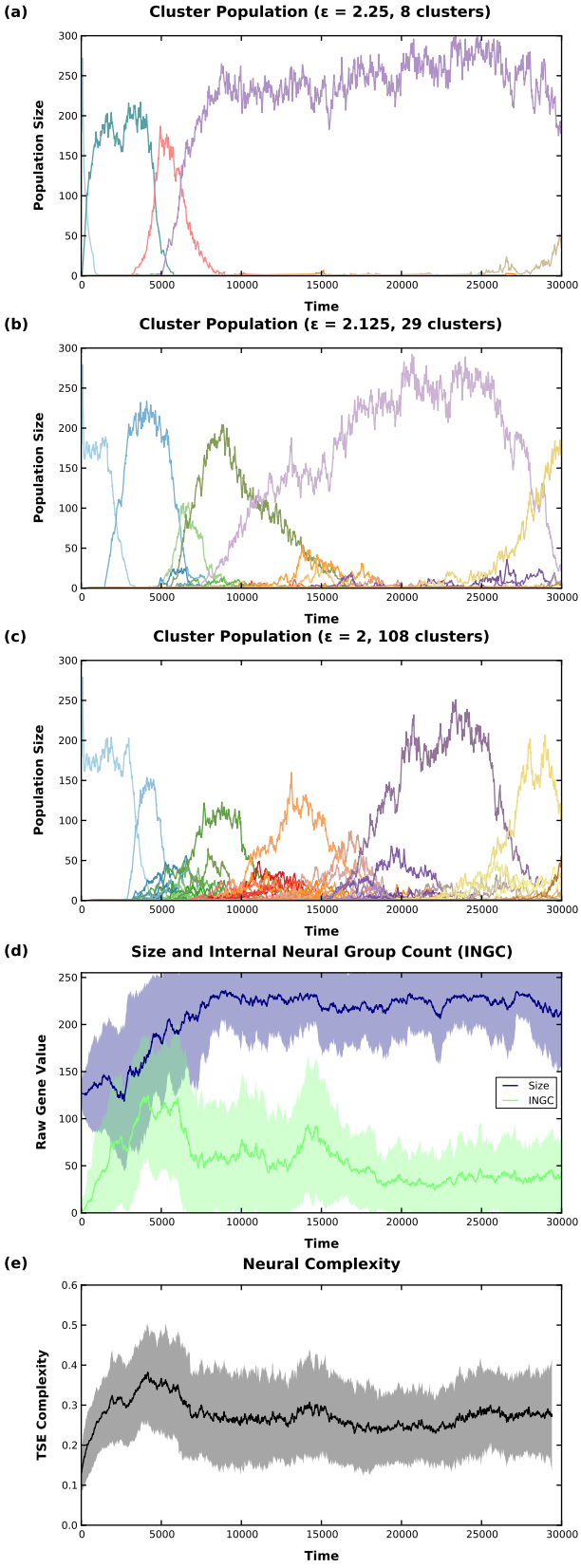

Figure 2: Temporal trends in cluster populations for $\epsilon = \{2.0, 2.125, 2.25\}$, two high-certainty genes (size and internal-neural-group count) exhibiting different selection behaviors, and TSE complexity. Genes and complexity shown as population means with standard deviation bands.

		Temporal			Neural	Genetic	
	Size	**Start**	**Peak**	**End**	**Complexity**	**Size**	**INGC**
0	1062	0	78	3749	0.2445	133.6	35.1
1	2278	1311	4087	8441	0.3626	172.4	101.0
5	769	4408	6694	11585	0.3657	202.5	98.3
9	3983	5119	8772	17509	0.3058	224.7	62.3
16	205	9168	14722	20192	0.3563	221.3	106.3
17	767	8795	13813	27611	0.3257	215.8	77.1
21	16732	6394	20672	30000	0.2876	225.4	41.9
23	397	11487	28572	30000	0.2861	200.5	38.8
24	273	13207	26594	30000	0.2619	185.3	29.1
27	2202	15126	29565	30000	0.3114	222.0	45.7

Table 2: Raw data from QT-Clust with $\epsilon = 2.125$. Shown are the origin, peak, and extinction of each major cluster, the TSE complexity, and mean values of the size and internal-neural-group-count (INGC) genes. Gene values are in the raw 0–255 range. Clusters with < 700 members appear in light gray. Clusters with < 200 members are not shown.

For the larger clusters, from cluster populations alone we cannot distinguish between roughly monotonic, anagenetic (within lineage) changes and true cladogenetic (divergent) speciation. However, long periods of temporal overlap during transitions suggest we may be seeing true speciation in large clusters as well, as distinct, new clusters emerge and are simply more successful than either the short-lived small clusters or the previous large cluster.

Temporal Trends in Genes

The use of clusters allows us to identify genetic differences between different subpopulations, including temporal trends in specific genes known to distinguish different subpopulations. Figure 2d shows different selection patterns for two high-certainty genes positioned below the cluster population graphs to allow comparison of their temporal trends. Table 2 shows the corresponding raw data for all clusters with a population size greater than 200.

The size gene ($certainty = 0.3515$) 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 to the third major cluster, size has reached the level at which it will plateau—around 220. By contrast, the internal-neural-group-count gene ($certainty = 0.2058$) shows a more variable selection pattern, which corresponds to trends in neural complexity as discussed below. These changes also correspond to cluster emergence and decline, as discussed in Cluster Characterization.

Neural Complexity

Tononi-Sporns-Edelman neural complexity (TSE complexity) (Tononi et al., 1994) gives an indication of the neural structure and function for each agent. Figure 2e shows the mean TSE complexity over time for the simulation being analyzed. In a past study, complexity was shown to be highly selected for only during periods of behavioral adaptation of the agents to their environment (Yaeger, 2009), in keeping with the tautology of evolutionary selection applying only when the subject of selection confers an evolutionary ad-

Clusters	children	grandchildren	child-rate	grandchild-rate
Same	2.04 (0.02)	4.04 (0.05)	6.54 (0.06)	10.6 (0.2)
Diff	1.89 (0.00)	3.57 (0.03)	5.11 (0.12)	7.76 (0.3)

Table 3: Reproductive success—numbers of offspring from parents of the same or different clusters and child-production rates per 1,000 contacts with agents from same or different clusters (stderr in parens), using $\epsilon = 2.125$.

vantage. The current results are in general agreement with previous simulations, showing strong selection for complexity in early populations during the period in which they are evolving to adopt an Ideal Free Distribution (Fretwell and Lucas, 1970; Fretwell, 1972) of agents to the heterogeneous resources of the simulated environment, plateauing around step 7500, and followed by a long stretch of relative stability lasting for the rest of the simulation. However, we see here a bump in complexity around t=15,000, unique to this particular simulation, that our clustering analysis reveals to be the result of a corresponding bump in internal-neural-group count deriving from the emergence and decline of a pair of specific sub-populations (clusters 16 and 17).

Discussion and Conclusions

Whether discussing the larger clusters, that replaced each other somewhat serially, or smaller clusters that represented sub-populations coexistent with the larger populations, we think it may be reasonable to conceive of these clusters as *species* within our artificial simulation. Since the simulation does not explicitly prevent interbreeding between clusters or base reproductive success on genetic distance, perhaps they should be considered *proto-species*, but the fall and rise of sub-populations, with significantly different genetic makeup from the dominant population, suggests a degree of specificity and persistence of species identity. Even the dominant populations may demonstrate speciation and competition between species, given the degree to which they overlap in time; e.g., note in Figure 2b that the cluster rising to dominance at the end of the run (light orange – cluster 27) first appeared barely over half way through the simulation (t=15,126) well before the previous dominating population (light purple – cluster 21) had reached its peak population (t=20,672). This occurs despite a relatively simple environment in which agents are free to mix and in which there is only one kind of energy resource (two if you distinguish between food that is grown and food derived from the carcasses of agents that are killed).

As Mallet (1995) 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." Lacking geographic isolation, sympatric speciation is typically thought to require disruptive selection to elicit distinct phenotypes and genotypes, coupled with selection for assortative mating to elicit reproductive isolation.

If disruptive selection and poor hybrid fitness are playing a role in balancing gene flow, we should see differences in the fitness, as measured by fecundity, of offspring from parents belonging to the same or to different clusters. To investigate this hypothesis we examined the number of children and the number of grandchildren produced by agents born to parents from the same or from different clusters. The left-hand columns of Table 3 summarize the results. Though the differences are modest, the offspring of parents from the same cluster produce more offspring than do the offspring of parents from different clusters, and those offspring are themselves more fecund. The magnitude of the differences are about 10x the standard error rates observed in the population (shown in parentheses), thus there is at least a modest post-zygotic selection pressure at work. Amplified across multiple generations it is easy to see how intra-cluster breeders will outperform inter-cluster breeders and produce ever more distinct sub-populations—species—even sympatrically. This is basically the first half of Wallace and Dobzhansky's proposed route to sympatric speciation.

If reinforcement is producing pre-zygotic selection and assortative mating, we should see differences in the rate at which agents produce offspring when they come in contact with agents from the same or different clusters. To investigate this possibility we examined the number of children and grandchildren produced *per contact* with other agents from the same or different clusters. For this analysis it is important to normalize birth rates by contact counts, since any kind of temporal, behavioral, or geographical isolation can and does significantly skew the number of potential reproductive encounters between same and different clusters for a given agent. The right-hand columns of Table 3 summarize these results. Both the child- and grandchild-production rates (per 1,000 contacts) are greater for encounters with agents from the same cluster than for agents from a different cluster. Here again, though the magnitude of the differences is small, they are roughly 10x the standard error rates observed in the population. Thus there is at least a weak pre-zygotic selection pressure at work.

Certain characteristics of the current simulated environment—especially the partial barriers, that are a holdover from previous experiments looking at the evolution of complexity—make it difficult to entirely tease apart sympatric vs allopatric speciation. In movies showing cluster membership over time we see clusters emerge and persist alongside existing clusters in a fully 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 we currently believe both forms of speciation are to be found in these simulations. A sample movie can be found at: http://informatics.indiana.edu/larryy/cluster_movie.zip

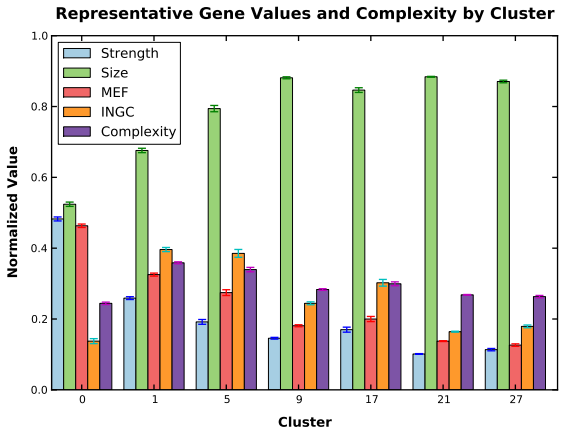

Figure 3: Means and standard error bars for the strength, size, mate-energy-fraction (MEF) and internal-neural-group-count (INGC) genes, along with neural complexity, for clusters with more than 700 agents, using $\epsilon = 2.125$.

Cluster Characterization

Clustered sub-populations can be characterized by their respective cluster centroid. Figure 3 provides a set of cluster fingerprints, summarizing the raw data in Table 2 for clusters of $\epsilon = 2.125$ (showing only the major clusters, with populations > 700). Specific evolutionary trends can be correlated to the rise and fall of particular species. The increase in size is readily apparent, along with a general decline in mate energy fraction and strength, and a variation in the internal-neural-group count. The earlier clusters 1 and 5 explore larger neural structures, achieving higher complexity. All dominant clusters exhibit a trend towards reduced energy consumption (low mate energy fraction and strength) and increased energy capacity (large size). Cluster 17 shows an exploratory population with slightly higher internal-neural-group count and neural complexity, coupled with a reduced emphasis on energy conservation, as evidenced by an increased strength and mate energy fraction, and slightly smaller size. This exploratory population is present in the middle third of the simulation, emerging out of the dominant cluster 21, but having only limited success, and, together with cluster 16, is responsible for the bump in internal-neural-group count and complexity as previously discussed.

Future Directions

One direction is to apply these analysis methods to simulations with simpler environments, in order to eliminate the possibility of allopatric speciation. We are also investigating methods from the evolutionary biology literature, such as "heat maps" of genetic diversity versus geological origins of parents, that might help us quantify degrees of sympatric vs allopatric speciation. An analysis of the temporal history of the fecundity and child-production rates discussed here might help distinguish pre-zygotic and post-zygotic selection and clarify the role of reinforcement in producing assortative mating.

Alternative clustering algorithms are also of interest. Information theory-based algorithms, such as that of Gokcay and Principe (2002), which maximizes cross-entropy between clusters, look particularly attractive. Alternatively, adopting a rival-penalization method, such as the k*-means algorithm (Cheung, 2002), may provide a better metric for cluster selection than cluster diameter. It might also be interesting to adapt the heirarchical clustering scheme of Aspinall and Gras (2010), regardless of whether we adopt their practice of allowing clusters to modulate reproductive success. Such a comparison would provide insight into whether or not varying the thresholds of QT-Clust is suggesting heirarchies of sub-populations, as hinted by Figure 2.

Any of these clustering methods, including the current one, would allow us to evaluate the effectiveness of a "miscegenation function", which establishes a probability of reproductive success that is inversely proportional to genetic distance between potential mates, that was long ago built into Polyworld, but which has never been explored to any substantial degree.

With the existing data, a study of the geographic locality of the origin and spread of each species may yield information about environmental effects on selection and degrees of sympatric vs allopatric speciation. This may provide theoretical insights into a common real-world speciation scenario in which initial allopatric (regional) divergence is followed by sympatric divergence, as seen in Darwin's Finches and other taxa (Huber et al., 2007). We would also like to apply these methods to simulations with clearly differentiated niches that are geographically either overlapping or isolated, to distinguish and quantify the relative effects of niche specialization vs geographic isolation.

References

Aspinall, A. and Gras, R. (2010). K-means clustering as a speciation mechanism within an individual-based evolving predator-prey ecosystem simulation. In An, A., Lingras, P., Petty, S., and Huang, R., editors, *Active Media Technology*, volume 6335 of *Lecture Notes in Computer Science*, pages 318–329. Springer Berlin / Heidelberg. 10.1007/978-3-642-15470-6_33.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In Beeri, C. and Buneman, P., editors, *Database Theory – ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg. 10.1007/3-540-49257-7_15.

Butlin, R. K. and Tregenza, T. (1997). Is speciation no accident? *Nature*, 387:551–553.

Cheung, Y.-m. (2002). k*-means – a generalized k-means clustering algorithm with unknown cluster number. In Yin, H.,

Allinson, N., Freeman, R., Keane, J., and Hubbard, S., editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2002*, volume 2412 of *Lecture Notes in Computer Science*, pages 147–154. Springer Berlin / Heidelberg. 10.1007/3-540-45675-9_48.

Dieckmann, U. and Doebeli, M. (1999). On the origin of species by sympatric speciation. *Nature*, 400:354–357.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia Univ. Press, New York, NY.

Felsenstein, J. (1981). Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, 35:124–138.

Fretwell, S. D. (1972). *Populations in a seasonal environment*. Princeton Univ. Press, Princeton, NJ.

Fretwell, S. D. and Lucas, H. L. (1970). On territorial behavior and other factors influencing habitat distribution in birds. *Acta Biotheoretica*, 19:16–36.

Gokcay, E. and Principe, J. C. (2002). Information theoretic clustering. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 24:158–171.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA.

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

Hocaoglu, C. and Sanderson, A. C. (1995). Evolutionary speciation using minimal representation size clustering. In *Evolutionary Programming'95*, pages 187–203.

Huber, S. K., León, L. F. D., Hendry, A. P., Bermingham, E., and Podos, J. (2007). Reproductive isolation of sympatric morphs in a population of Darwin's finches. *Proceedings of the Royal Society B: Biological Sciences*, 274(1619):1709–1714.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Blackwell.

Kirkpatrick, M. and Ravigné, V. (2001). Speciation by natural and sexual selection: models and experiments. *Am Nat*, 158:S22–S35.

Kondrashov, A. S. and Kondrashov, F. A. (1999). Interactions among quantitative traits in the course of sympatric speciation. *Nature*, 400:351–354.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

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

Mayr, E. and Provine, W. (1998). *The evolutionary synthesis: perspectives on the unification of biology*. Harvard University Press, Cambridge.

Ortiz-Barrientos, D., Counterman, B. A., and Noor, M. A. F. (2004). The Genetics of Speciation by Reinforcement. *PLoS Biol*, 2(12):e416.

Radovanović, M., Nanopoulos, A., and MirjanaIvanović (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning*, 11:2487–2531.

Sætre, G.-P., Moum, T., Bureš, S., Miroslav Král, M. A., and Moreno, J. (1997). A sexually selected character displacement in flycatchers reinforces premating isolation. *Nature*, 387:589–592.

Scharl, T. and Leisch, F. (2006). The stochastic qt–clust algorithm: evaluation of stability and variance on time–course microarray data. In Rizzi, A. and Vichi, M., editors, *Compstat 2006—Proceedings in Computational Statistics*, pages 1015–1022. Physica Verlag, Heidelberg, Germany.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Silverton, J., Servaes, C., Bliss, P., and Macleod, D. (2005). Reinforcement of reproductive isolation between adjacent populations in the park grass experiment. *Heredity*, 95:198–205.

Tononi, G., Sporns, O., and Edelman, G. (1994). A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Nat. Acad. Sci.*, 91:5033–5037.

Tregenza, T. and Butlin, R. K. (1999). Speciation without isolation. *Nature*, 400:311–312.

Van Doorn, G. S. and Weissing, F. J. (2001). Ecological versus sexual selection models of sympatric speciation: a synthesis. *Selection*, 2:17–40.

Wallace, A. R. (1899). *Darwinism*. Macmillan, London.

Weissing, F. J., Edelaar, P., and van Doorn, G. S. (2011). Adaptive speciation theory: a conceptual review. *Behav Ecol Sociobiol*, 65:461–480.

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

Yaeger, L. S. (2009). How evolution guides complexity. *HFSP*, 3(5):328–339.

Yaeger, L. S., Griffith, V., and Sporns, O. (2008). Passive and Driven Trends in the Evolution of Complexity. In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., 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.

Yaeger, L. S., Sporns, O., Williams, S., Shuai, X., and Dougherty, S. (2010). Evolutionary Selection of Network Structure and Function. In Fellerman, H., Dörr, M., Hanczyc, M. M., Laursen, L. L., Maurer, S., Merkle, D., Monnard, P.-A., Støy, K., and Rasmussen, S., editors, *Artificial Life XII: Proceedings of the Twelfth International Conference on the Simulation and Synthesis of Living Systems*, pages 313–320. MIT Press, Cambridge, MA.

Zhao, Y. and Karypis, G. (2005). Data clustering in life sciences. *Molecular Biotechnology*, 31:55–80.