CA Simulation of Biological Evolution in Genetic Hyperspace

Michael A. Saum[1] and Sergey Gavrilets[1,2]

[1] Department of Mathematics
University of Tennessee, Knoxville
Knoxville, TN
msaum@math.utk.edu

[2] Department of Ecology and Evolutionary Biology
University of Tennessee, Knoxville
Knoxville, TN
gavrila@tiem.utk.edu

Abstract. Realistic simulation of biological evolution by necessity requires simplification and reduction in the dimensionality of the corresponding dynamic system. Even when this is done, the dynamics remain complex. We utilize a Stochastic Cellular Automata model to gain a better understanding of the evolutionary dynamics involved in the origin of new species, specifically focusing on rapid speciation in an island metapopulation environment. The effects of reproductive isolation, mutation, migration, spatial structure, and extinction on the emergence of new species are all studied numerically within this context.

1 Introduction

From the fossil records and radioactive dating we know that life has existed on earth for more than 3 billion years [1]. Until the Cambrian explosion around 540 million years ago, life was restricted mainly to single-celled organisms. From the Cambrian explosion onward however, there has been a steady increase in biodiversity, punctuated by a number of large extinction events. These extinction events caused sharp but relatively brief dips in biodiversity and the fossil record supports these claims. In our attempt to understand some of the dynamics involved in this process, we decided to look at the speciation process and see if we could model it in a way that would provide insight into some of the factors which determine the dynamic behavior of what is an extremely complex process.

Speciation is the process by which new species are formed via evolutionary dynamics. Speciation can be controlled (or driven) by a number of factors including mutation, recombination and segregation, genetic drift, migration, natural and sexual selection [1–3]. Throughout this paper we say that two populations are of different species if they are reproductively isolated, i.e., no mating producing both viable and fertile offspring between the two populations occurs. That is, we will use the biological species concept [1, 2]. In our model, we can identify reproductively isolated populations by measuring the differences in their genes;

if their genes are sufficiently different, then there is a very low probability that they can mate to produce viable and fertile offspring.

Speciation processes are difficult to verify via experiments or observations. Primary of course is the fact that the time-scales involved in speciation typically are much longer than human life span. In addition, there does not exist a continuous fossil record documenting new species, i.e., there are many gaps in the fossil record. Moreover, existing data on genetic differences between extant species can be interpreted in a number of alternative ways.

We are thus led to different methods of investigating the speciation process by using mathematical models. By necessity, models limit the number of parameters associated with complex behavior. This implies that all factors may not be taken into account in the simulation of complex processes. However, computer models do provide a metaphor for the actual dynamics, assuming of course the model's algorithms accurately reflect in some sense the actual dynamics being modeled, i.e., the model is consistent.

Here, we describe a stochastic cellular automata explicit genetic model of speciation in an island metapopulation. Typically, cellular automata used in biological application are characterized by a rather small number of states: two or, very rarely, three, usually focusing on whether a patch is occupied or not [4–12]. However, even the simplest known biological organisms have hundreds of genes and hundreds of thousands of DNA base pairs [1, 3]. This implies that the number of possible genetic states for an organism is astronomically large. For example, assuming that an organism has only 500 genes each coded by 1000 DNA base pairs, there can be potentially $4^{500000} \approx 9.9 \times 10^{301029}$ different genetic states. This enormous dimensionality requires one to develop new methods of modeling, analyzing, and visualizing the behavior of the corresponding cellular automata. Below we describe some of the approaches that we have developed within the context of studying speciation.

2 The CA Deme-Based Metapopulation Model

A common method for performing numerical studies of biological evolution and speciation is to use an individual-based model in which a finite collection of individuals are tracked through the birth-reproduction-death cycle as well as the migration-mutation-survival cycle. Unfortunately, individual-based models require an enormous amount of computational resources to obtain meaningful results and are currently not practical for studying large-scale biological diversification. Here, instead of an individual-based model we build a deme-based model [3, 13, 14] in which for each local population we explicitly describe only the genetic state of its most common genotype. This simplified approach is justified if mutation and migration are sufficiently rare and the local population size is sufficiently small so that only a negligible amount of genetic variation is maintained within each local population most of the time. We will ignore the dynamics of local population sizes. Following Hubbell [15], we disregard ecological differences between the species. Our main focus will be on genetic incompatibilities (i.e. reproductive isolation) between different populations.

Reproductive isolation will be defined by the threshold model [3, 13] in which two genotypes are not reproductively isolated and, thus, belong to the same species if they differ in less than K_m genes. We will refer to parameter K_m as mating threshold. In some implementations of the model, we allow for multiple populations per patch. A simple heuristic approach for doing this is to introduce another threshold genetic distance, say K_c ($> K_m$), reaching which will allow for coexistence in a patch. We will refer to parameter K_c as coexistence threshold. If the genetic divergence between two populations is below K_c, the competition between them prevents their coexistence.

We consider here a large area divided into smaller connected areas called patches. Each patch can be empty or occupied by one or more populations. We model the habitat patches as nodes on a two dimensional grid. This is a spatially explicit metapopulation model (which is often also called a *lattice model* or *stepping-stone model*), in which migration is restricted to close or neighboring patches.

Our metapopulation model simulates evolution of bit strings in a two dimensional geometry. Each bit string can be considered to represent the DNA of a population. The length L of this binary DNA string is specified as input. Note that the number of possible genetic states is 2^L. We then simulate metapopulation dynamics within and between a given set of habitat niches (or patches).

What we are left with then after a time is a situation in which many genetically different populations exist in different habitat patches. Through a clustering process, we can then determine which populations are *close* to each other genetically by some measure. This process of grouping thus determines clusters of similar populations, or species.

Our model dynamics occur on a time generation basis. For each generation we determine stochastically whether each of the major events occurs in the following order:

1. Patch Extinction.
2. Single Population Extinction.
3. DNA Strand Mutation.
4. Population Migration.

Patch extinction is a situation where all populations in a specific patch go extinct. The exact details are not important, it could be due to depletion of a viable food supply in the habitat patch or due to some catastrophic extinction event which wipes out the populations such as a fatal disease epidemic.

Single population extinction can occur under similar circumstances, however rather than the whole patch (which can include many populations) going extinct, only a single population within the patch goes extinct.

Migration of individuals has two effects. First, migrants can found a new population in a patch previously not occupied by a species. Second, migrants coming into an occupied patch can bring genes that may spread in a local population (see below).

Bit strings change independently at each locus. The probability per generation that an allele at a locus changes to an alternative state is set to be

$$\mu_e = \mu + m\,\mathcal{N}, \tag{1}$$

where μ is the probability of mutation per locus, m is the probability of migration, and $\mathcal{N}$ is the number of neighboring populations of the same species that have the alternative allele fixed at the locus under consideration. Expression (1) utilizes the fact that the probability of fixation of an allele that does not affect fitness is equal to its frequency [16]. With migration, new alleles are brought in the patch both by mutation (at rate μ) and migration (at rate $m\mathcal{N}$). In this approximation, the only role of migration is to bring in new alleles that are quickly fixed or lost by random genetic drift. For example, if initially both the focal population and its four neighbors have allele 0 at the locus under consideration, then the probability that an alternative allele 1 is fixed in the focal population per generation is $\mu_e = \mu$. However, once this has happened, the probability of focal population switching back to allele 0 is $\mu_e = \mu + 4m$. If the migration rate m is much larger than the mutation rate per locus μ, switching back will happen much faster. As time increases, populations accumulate different mutations, diverge genetically and become reproductively isolated species.

3 Model Implementation

There are two main computer programs utilized to implement our model of the speciation process, Evolve and Cluster. As described above, *Evolve* simulates the evolution of bit strings in a two dimensional grid based geometry undergoing evolutionary dynamical processes. *Cluster* then determines which group of bit strings or populations are within a specified Hamming distance of each other. The clustering method is single linkage clustering [17] with an input parameter K. In most cases, we set parameter K to the mating threshold K_m. This procedure produces clusters of mutually compatible populations (i.e. biological species).

Since the clustering process is hierarchical in nature, output from Cluster can also be used to identify and group populations in a taxonomic manner, providing insight into the hierarchical structure of the simulated populations. For example, let us specify an increasing sequence of clustering thresholds $K_1 < K_2 < K_3 < \ldots$. Then, all populations at a genetic distance less than K_1 can be thought of as belonging to the same species, all populations at genetic distances that are larger or equal than K_1 but are smaller than K_2 can be thought of as belonging to different species within the same genus, all populations at genetic distances that are larger or equal than K_2 but are smaller than K_3 can be thought of as belonging to different species and genera within the same family, etc.

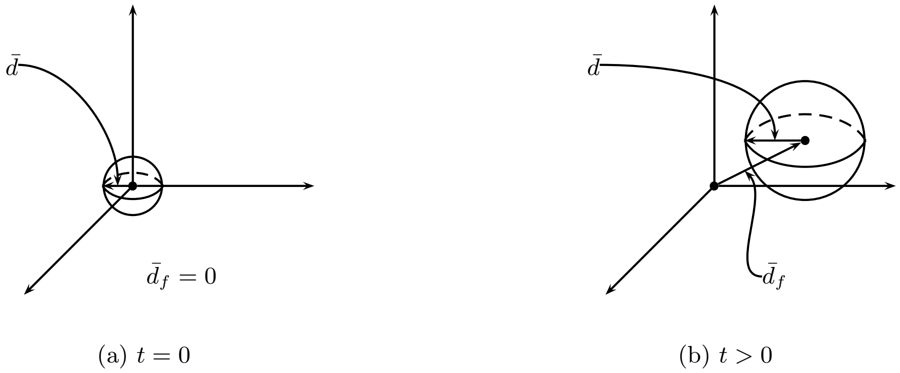
Evolve-Cluster accepts a wide variety of input and produces a wide variety of output. In order to provide focus on identifiable trends, we will concentrate in this paper on the following input to and output from the Evolve-Cluster simulations as shown in Table 1. (Note that there is no correlation between the input and output items, they are just lists).

Table 1. Evolve-Cluster Input/Output Parameters of Interest

Input Parameters	**Output**
Geometry (1D, 2D, size)	Number of Clusters (Species), N_S
Bitwise Mutation Probability, μ	Average Pairwise Distance, $\bar{d}$
Deme Extinction Probability, E_D	Average Distance from Founder, $\bar{d}_f$
Population Extinction Probability,E_p	Time to Speciation, T
DNA strand length, L	Duration of Radiation Event, τ
Population Migration Probability, m	Cluster Diameter
Patch Carrying Capacity	Cluster Range Distribution
Mating, Coexistence and Clustering Thresholds	Cluster Average Pairwise Distance

One can visualize our model as follows: Each population is a point in a genetic hyperspace; the clade (i.e., the whole system of populations) is a cloud of points which changes its size, structure, and location in the genetic hyperspace. The diameter of this cloud can be characterized by the average pairwise distance $\bar{d}$ between members of the clade measuring how diversified the clade is. The average distance to the founder $\bar{d}_f$ characterizes the extent of the overall change (see Figure 1). As time increases, populations get farther and farther away from each other while at the same time moving farther away from the founding population. Of course there is a limit as to how much $\bar{d}_f$ and $\bar{d}$ increase due to the finite number of loci under consideration. In fact it can be shown that $\bar{d}_f \to \frac{L}{2}$ and $\bar{d} \to \frac{L}{2+g(\mu)}$ as $t \to \infty$. [Here, $g(\mu) \to 0$ as $\mu \to \infty$ and $g(\mu) > 0$ for all $\mu > 0$. Essentially $g(\mu) \sim 1/\mu$.]

In addition, we can easily calculate how long it takes for speciation to occur, how many species emerge, and what parameters affect the rate of speciation and species diversity.



(a) $t = 0$

(b) $t > 0$

Fig. 1. The average pairwise distance $\bar{d}$ and the average distance to the founder $\bar{d}_f$ at two different time moments. The clades are represented by the spheres.

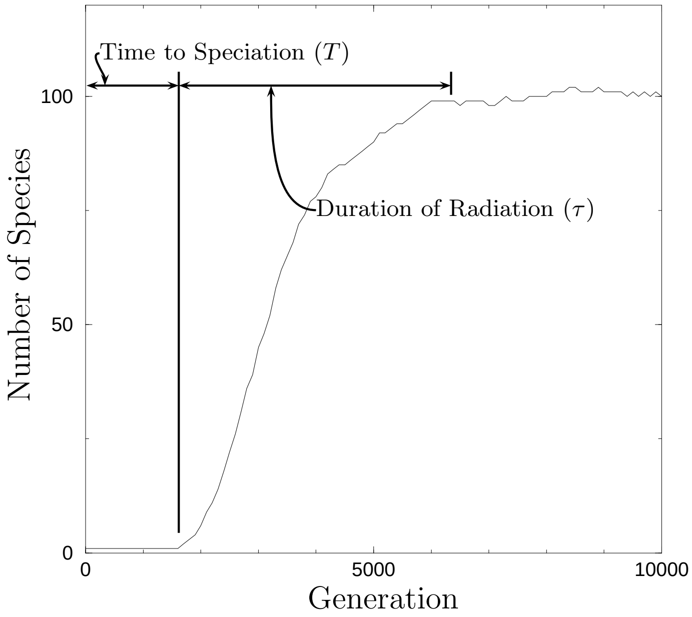


Fig. 2. A typical speciation curve

Figure 2 illustrates a *typical* speciation curve (i.e., the number of species or clusters vs. time). This figure also explains the meaning of two statistics: the time to speciation T and the duration of radiation τ. Note that the number of species stays at 1 for a small amount of time, then rises relatively quickly to reach a stochastic equilibrium level.

All data and results reported in this paper are based on multiple runs of the same set of parameters, usually between 30 and 50 repeats.

Distance from the Founder, $\bar{d}_f$

One quick check that our model is working well is based on analysis of how certain dynamics match the theory. In [18, Eq. 4c], it was shown that the average Hamming distance from a single founding population changes according to equation

$$\bar{d}_f(t) = \frac{L}{2}[1 - \exp(-c(\mu)t)] \tag{2}$$

where $c(\mu)$ a function only of μ, the mutation rate. This is basically a solution to a random walk problem on the binary hypercube. Our model showed that the fit to Equation 2 over hundreds of runs with varying parameter sets truly is a function only of the mutation rate μ and time. This perhaps is the single best indication that our model is performing well with prediction and is internally consistent with the basic mathematical evolutionary theory concepts of mutation, migration, and extinction.

4 Parameter Studies

Since the Evolve-Cluster model seems to be modeling some aspects of the speciation process well when compared with other models, it now remains to identify other characteristics of our model. Specifically we will be analyzing the effect of changing input parameters to first see if the results make qualitative sense and then use our model to uncover *hidden* trends and quantitative results.

Geometry Size, Mating Threshold, and Clustering Threshold

Figure 3 contains summary graphs of nine different parameter sets. The graphs are ordered from top to bottom increasing in 2-D geometry size, 10×10, 14×14, and 20×20. The graphs are ordered from left to right increasing in mating threshold $K_m = 5, 10$, and 15. Each graph is the summary of fifty runs with $L = 256$, $m = 0.02$ and $\mu = 0.00004$. On each graph there are five curves. The three speciation curves are for the different clustering thresholds K, while the other two curves are the average pairwise distance ($\bar{d}$) between all populations and the average distance from the founder ($\bar{d}_f$) as a function of time.

In our model, extensive diversification occurs relatively fast. The graphs in Figure 3 illustrate the fact that $\bar{d}$ dominates initially, while $\bar{d}_f$ eventually becomes larger than $\bar{d}$ and stays that way. In addition, the asymptotics are consistent with those discussed in the previous section. This trend can be understood by considering the metaphor introduced above; the ball changes diameter quicker than moving away from the origin initially, i.e., genetic changes go into producing diversity at a rate quicker than moving the clade as a whole genetically away from the founding population. After a short time, movement away from the founder dominates while at the same time genetic diversity between the populations also increases.

In our model, the probability of a genetic change μ_e (see equation 1) depends on the number of neighboring populations of the same species and, thus, on mating threshold K_m. With a higher K_m, there are more neighbors of the same species which effectively reduces the rate of change and dampens $\bar{d}$ expansion. This is evidenced by the fact that the higher the mating threshold K_m, the closer the curves $\bar{d}$ and $\bar{d}_f$ track each other. Since the number of loci L and mutation probability μ are the same in all of these cases, the $\bar{d}_f$ curve is the same in all graphs as expected. It also appears that the larger the size of the system, the greater the difference between $\bar{d}$ and $\bar{d}_f$, although the asymptotics still remain the same as described above. This can be explained by the fact that with a larger geometry, $\bar{d}$ increases unchecked by physical boundaries until boundary effects coupled with the finite number of loci L effectively dampens $\bar{d}$ expansion and the asymptotics take over.

As the clustering threshold K increases, the number of species decreases. This is as expected, since larger clusters (clusters containing more populations) implies there are less clusters. It is also clear that the number of species increases as geometry size increases. It appears here that boundary effects do play a role in speciation, effectively suppressing the speciation process to some extent.

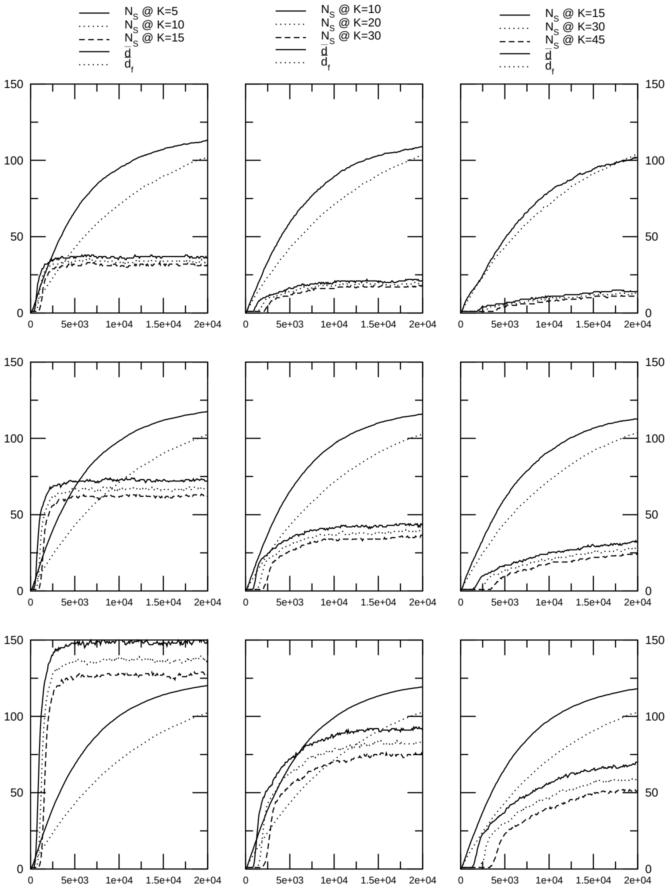


Fig. 3. The effects of geometry size, mating threshold K_m, and clustering threshold K on the number of species N_S, the average pairwise distance $\bar{d}$, and the average distance to the founder $\bar{d}_f$ as functions of time.

The time to speciation T increases as K_m increases. This is due to the fact that it takes longer to accumulate enough genetic differences to separate populations into new species. The duration of radiation τ increases as K_m increases. This is due to the observation that radiation still occurs, but is not as rapid as at lower mating threshold values, more evidence of negative mutation pressure applied by the higher mating threshold.

There are other observations which can be made from the graphs shown in Figure 3, including

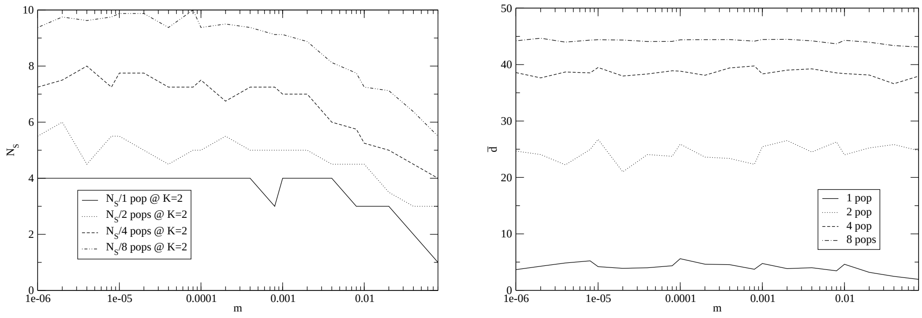
- T increases as geometry size increases,
- τ is approximately constant as geometry size increases,
- The difference between the number of species at different clustering levels remains constant in time,

- The difference between the number of species at different clustering levels is approximately constant as mating threshold increases,
- The difference between the number of species at different clustering levels increases as geometry size increases,
- τ appears to be much less that T in all cases.

Migration and Patch Carrying Capacity

One of the parameter studies undertaken was to increase the carrying capacity of each patch in the geometry so that multiple populations per patch could exist at any time. With multiple populations allowed, the evolutionary dynamics consist of a series of population splits followed by accumulation of additional genetic differences between emerging species which eventually allows for their coexistence in the same patch (when genetic distance is $> K_c$), which in turn leads to range expansions and increase in the number of populations per patch.

Figure 4 illustrates some results for a clustering threshold of $K = 2$ letting migration rate m vary. Part (a) shows the number of species in the system which we normalized by the patch carrying capacity (i.e., the number of populations per patch). Note that increasing the patch carrying capacity increases the number of species N_S in the system disproportionately. N_S is essentially constant with a slight decreasing trend as m increases. Part (b) shows that the average pairwise distance $\bar{d}$ increases with the patch carrying capacity; $\bar{d}$ does not appear to depend on the migration rate. Overall, allowing for multiple populations per patch stimulates population expansion into multiple ecological habitat niches allowing for rapid speciation to occur in parallel resulting in even more diversification, all in approximately the same time frame.



(a)

(b)

Fig. 4. The effects of migration rate m on the normalized number of species and on the average pairwise distance $\bar{d}$ in a model with 1, 2, 4 or 8 populations per patch.

5 Conclusions

Our CA based metapopulation model allows us to investigate the dynamics of genetic diversification in a large dimensional state space. The adaptive radiation regime observed in the model is a rich source of data for helping one to better understand the speciation process.

References

1. Futuyma, D.J.: Evolutionary biology. Sinauer, Sunderlands, MA (1998)
2. Coyne, J.A., Orr, H.A.: Speciation. Sinauer Associates, Sunderland, MA (2004)
3. Gavrilets, S.: Fitness landscapes and the origin of species. Princeton University Press, Princeton, NJ (2004)
4. Wolfram, S.: Statistical mechanics of cellular automata. Reviews of Modern Physics **55** (1983) 601–644
5. Huberman, B.A., Glance, N.S.: Evolutionary games and computer simulations. Proceedings of the National Acedemy of Sciences USA **90** (1993) 7716–7718
6. Keymer, J.E., Marquet, P.A., Johnson, A.R.: Pattern formation in a patch occupancy metapopulation model: A cellular automata approach. Journal of Theoretical Biology **194** (1998) 79–90
7. Keymer, J.E., Marquet, P.A., Velasco-Hernández, J.X., Levin, S.A.: Extinction thresholds and metapopulation persistence in dynamic landscapes. American Naturalist **156** (2000) 478–494
8. Molofsky, J., Durrett, R., Dushoff, J., Griffeath, D., Levin, S.: Local frequency dependence and global coexistence. Theoretical Population Biology **55** (1999) 270–282
9. Molofsky, J., Bever, J.D., Antonovics, J.: Coexistence under positive frequency dependence. Proceedings of the Royal Society London Series B **268** (2001) 273–277
10. Durrett, R., Buttel, L., Harrison, R.: Spatial models for hybrid zones. Heredity **84** (2000) 9–19
11. Carrillo, C., Britton, N.F., Mogie, M.: Coexistence of sexual and asexual conspecifics: a cellular automaton model. Journal of Theoretical Biology **275-285** (2002) 217
12. Ganguly, N., Sikdar, B.K., Deutsch, A., Canright, G., Chaudhuri, P.P.: A survey on cellular automata. Technical Report Centre for High Performance Computing, Dresden University of Technology (December 2003)
13. Gavrilets, S., Acton, R., Gravner, J.: Dynamics of speciation and diversification in a metapopulation. Evolution **54** (2000) 1493–1501
14. Gavrilets, S.: Speciation in metapopulations. In Hanski, I., Gaggiotti, O., eds.: Ecology, genetics and evolution of metapopulations. Elsevier, Amsterdam (2004) 275–303
15. Hubbell, S.P.: The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton (2001)
16. Kimura, M.: The neutral theory of molecular evolution. Cambridge University Press, New York (1983)
17. Everitt, B.S.: Cluster analysis. Arnold, London (1993)
18. Gavrilets, S.: Dynamics of clade diversification on the morphological hypercube. Proc. R. Soc. Lond. B **266** (1999) 817–824