Evolutionary Farming: EvCl and NetSolve

Michael Saum
Department of Mathematics
&
Department of Ecology and Evolutionary Biology
May 10, 2002

Overview

- Research Goals
- Basic Theory
- Data Flow & Processing
- The Problem
- The Solution

Research Goals

- Develop a computer based model simulating evolution and diversification of metapopulations in a spatial setting.
- Explore relationships between various parameters affecting speciation dynamics.

Why?

- Rapid genetic diversification early in a clade's history, at relatively low taxonomic diversity, with an apparent slowdown afterwards, represents a commonly observed pattern of radiation of animal life.

- Best example is *Cambrian Explosion*.

- Life has occupied planet for nearly 4 billion of its 4.5 billion years.

- Until 600 million years ago, there were no organisms more complex than bacteria, multicelled algae, single celled plankton.

- Then, 543 million years ago, in the early Cambrian, within the span of no more than 10 million years, creatures with teeth and tentacles and claws and jaws suddenly appeared.

Basic Theory

- Evolve-Cluster *EvCl* consists of three programs.
  - evolve - main evolution simulator.
  - cluster - species determination.
  - graphics - display.
- *evolve* produces output files processed by *cluster* resulting in data files for analysis and display by *graphics*.

Basic Theory

- EvCl is written in `C` and makes use of `gtk+` to add flexibility in use of data structures and to make available various `X-windows` graphics primitives.
- Evolve simulates evolution of fixed length bit strings in a one or two dimensional grid based geometry.
- Each bit string can be considered to represent the *DNA* of a population, each grid point a habitat niche or *deme*.
- Cluster determines groups of DNA that are within a specified hamming distance of each other; clusters of similar populations are called *species*.

If two populations differ genetically by a lot, they probably won't mate with each other, i.e., different species.

Evolve

- **geometry**: 2-D grid of demes, each deme can hold fixed number of populations.
- **populations**: Each population represented by a binary string of fixed length (DNA).
- Random single population and catastrophic deme extinction.
- Population invasion attempts.
- Colonization of genetically *different* demes.
- DNA mutations.

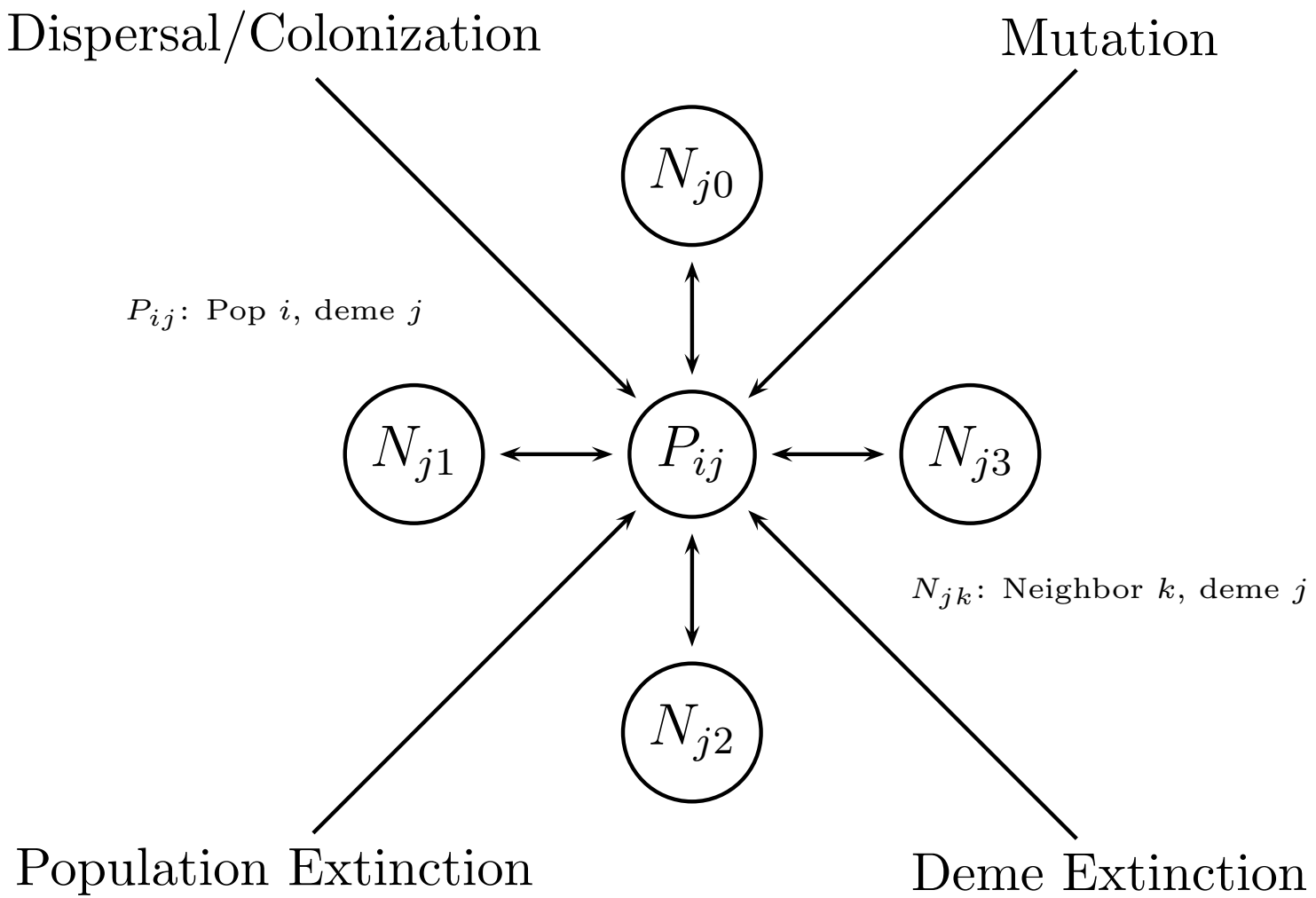
Evolve

- For each generation, simulation treats in order:
  - Deme extinction.
  - Single population extinction.
  - DNA strand mutation.
  - Single population dispersal.
- Demes and populations processed in random order for each generation (time step).

Can be considered in some respects a series of random walks on morphological (genetic) hypercube.

Note that simulation is great simplification of real life; Estimates range from about 30,000 to over 100,000 genes in the human genome.

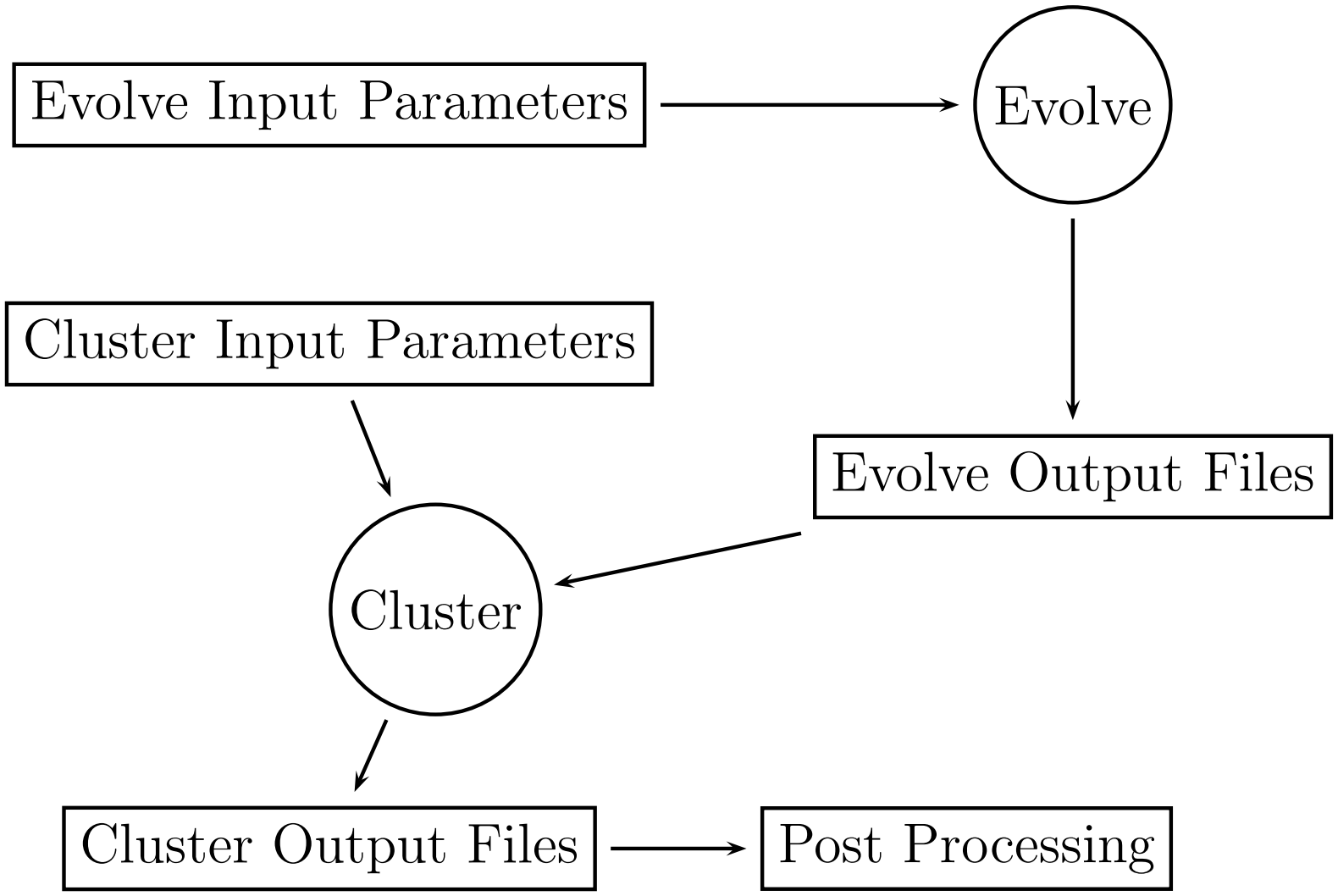
The Big Picture



Cluster

- Populations are clustered together (single linkage clustering) if they are within a specified hamming distance threshold.
- This threshold we call the *mating* threshold.
- Each cluster is termed a *species*.
- Produces a DISPLAY file which can be displayed by the graphics program.
- Produces detailed information on clusters for later processing and analysis.

Data Flow & Processing



The Problem

- Parameters passed to EvCl on command line.
- Must run many separate runs of same parameter set due to stochastic nature of simulation.
- In December, ran 4K runs, each run taking 45-60 minutes on UltraSparc 10.
- Took around 2 1/2 weeks to complete on approximately 8 machines.
- Generated around 4 GB data to be analyzed
- Logistical nightmare managing resources.

The Solution

- Began working with Jack Dongarra and Michelle Miller to move EvCl into NetSolve framework.
- Converted command line driven programs to be functions in a shared library.
- Created EVCL PDF.
- Created client program to access NetSolve via `netsl_farm()` API.
- Installed and Configured NetSolve in $HOME for use on machines in tiem.utk.edu.
- Running with minimal resource management effort.

Remarks

- Programming work not yet finished on EvCl. Current efforts directed at speeding up clustering and benchmarking evolve.
- NetSolve installation was easy and straightforward.
- Current solution dictates that machines in NetSolve domain must share same NFS file space for retention of output files.
- Investigating use of IBP to alleviate this.
- At this point, SInRG resources not required since still tuning EvCl PDF.

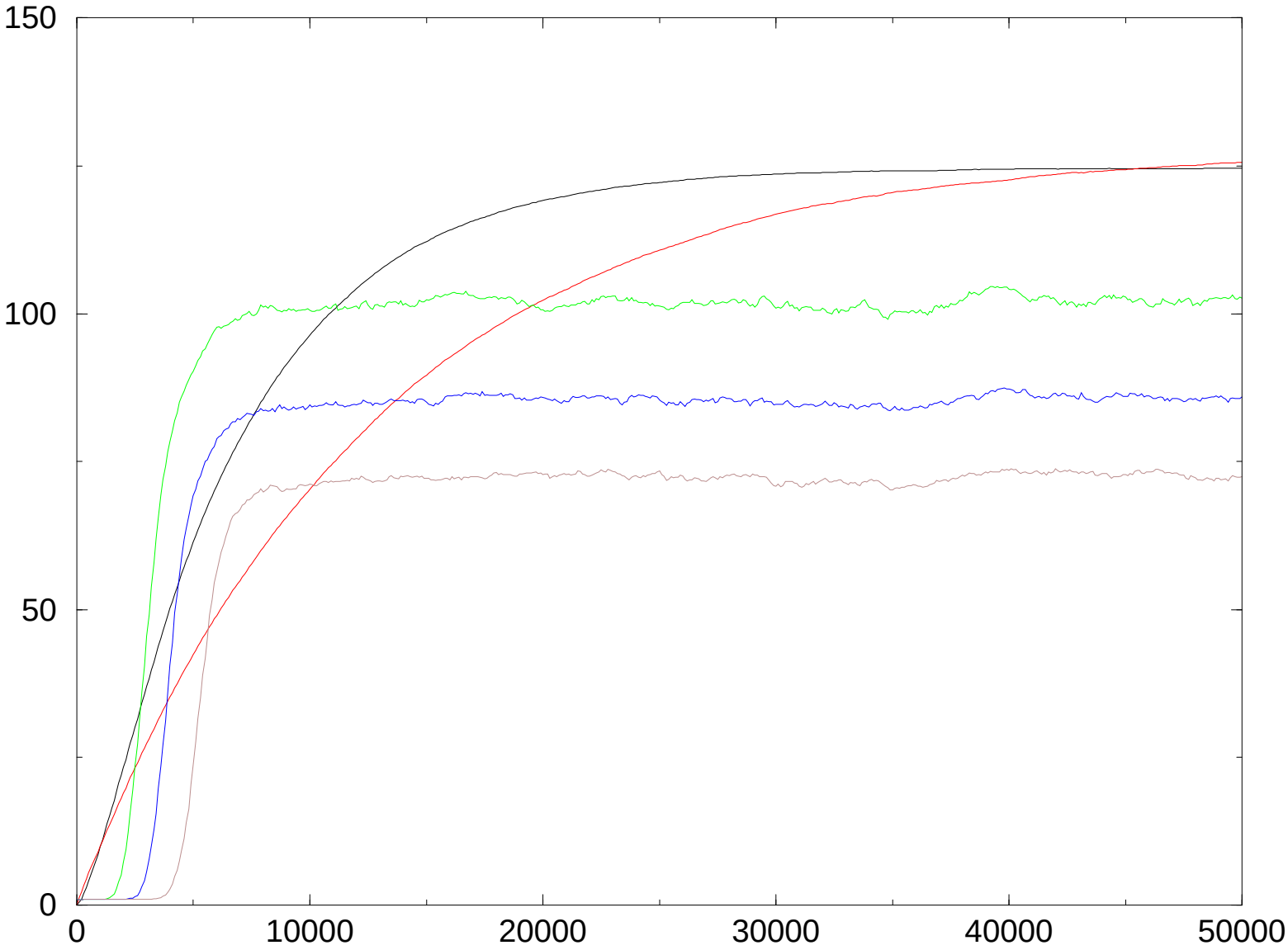
Distance from Founder

Inserting a single population into the clade at time $t = 0$, the average distance d of the clade from it's ancestral state at time t can be approximated as:

$$d = \frac{L}{2}\left(1 - exp\left(-\int_0^t 2\mu \, ds\right)\right)$$

C25 −− 50 Runs

G20x20.y15.U00004.M005



Thu Dec 6 08:54:54 EST 2001