

Simulating Population Genetics on the XT5

E. A. Duenez-Guzman, A. D. Vose, M. D. Vose, S. Gavrilets

The University of Tennessee, Knoxville

May 1, 2009

NICS

National Institute for Computational Sciences

University of Tennessee & Oak Ridge National Laboratory

Simulate Evolution And Speciation Dynamics

- *individual-based*
- *explicit-genetics*
- *discrete-time*
- *stochastic model*

Computational challenges – *Lessons learned*

- *Integration*
- *Lazy evaluation*
- *Equivalence*
- *Precomputation*
- *Tuning*
- *Time warp*

$$I(\alpha, \beta, \gamma, \delta, \xi)$$

- *time spent in habitat*
- *carrying capacity*
- *mating*
- *viability*

$\alpha, \beta, \gamma, \delta, \xi$: various rational functions of model parameters
and additive phenotypic characters of individuals

Canned Integration Routines?

```
> Digits := 16;  
> I_(0.91381,0.095649,0.57591,4.1584,1.4782);
```

0.00097...

```
> Digits := 32;  
> I_(0.91381,0.095649,0.57591,4.1584,1.4782);
```

0.00687...

```
> Digits := 64;  
> I_(0.91381,0.095649,0.57591,4.1584,1.4782);
```

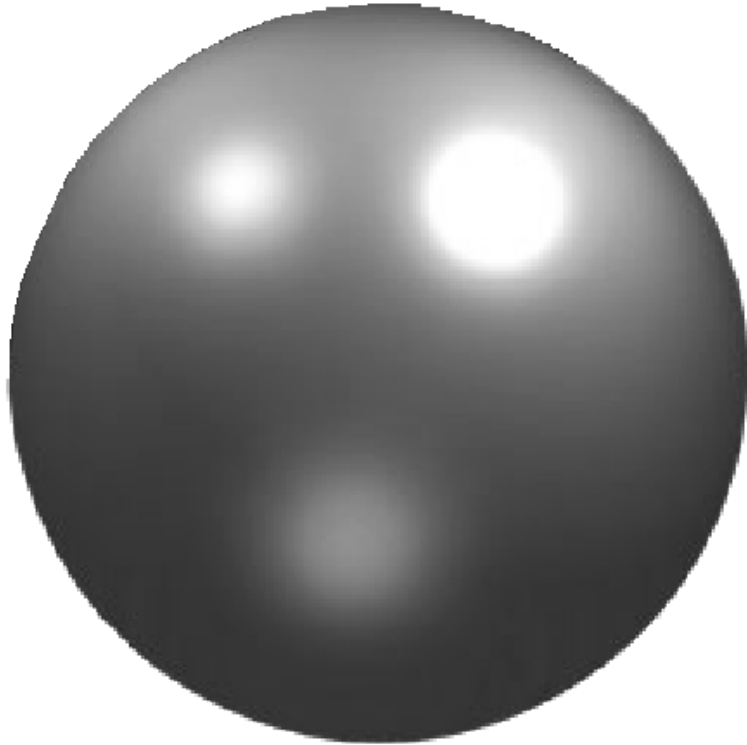
0.0219...

```
> Digits := 128;  
> I_(0.91381,0.095649,0.57591,4.1584,1.4782);
```

0.0322...

method	accuracy	speed
Maple	-	- - -
NAG	- -	- -
GSL	- -	- -
Hand-Coded Quadrature	-	-

$$\int_u^v \exp\left(-\left(\frac{t-a}{b}\right)^2 + (t^{-1} - d)^2\right) \operatorname{erfc}(t^{-1} - d) \frac{dt}{t}$$



Boas and Schoenfeld : *Residues on the Riemann Sphere*

Theorem 1. *Let F be holomorphic in the extended plane except for a finite number of singularities, let F be holomorphic on (a, b) except for simple poles, and let F be holomorphic at a and at b . Then*

$$P.V. \int_a^b F(t) dt = -(R + r)$$

where R is the sum of the residues of

$$F(z) \log\{(z - a)/(z - b)\}$$

for z in the extended plane but not on $[a, b]$, and r is the sum of the residues of

$$F(z) \log\{(z - a)/(b - z)\}$$

for z on (a, b) .

Essential singularities at 0 and ∞

Singularity at ∞ : the residue at $z = 0$ of

$$-z^{-1} \exp\left(-\left(\frac{z^{-1} - a}{b}\right)^2 + (z - d)^2\right) \operatorname{erfc}(z - d) \log \frac{1 - uz}{1 - vz}$$

Singularity at 0 : the residue at $z = 0$ of

$$z^{-1} \exp\left(-\left(\frac{z - a}{b}\right)^2 + (z^{-1} - d)^2\right) \operatorname{erfc}(z^{-1} - d) \log \frac{z - u}{z - v}$$

Tools

Lemma 2. *Let ψ have a simple pole at ζ with residue ξ . If ϕ is holomorphic at ζ , then the residue of $\psi(z)\phi(z)$ at $z = \zeta$ is $\xi\phi(\zeta)$*

Lemma 3. *Let ζ be a point of the Riemann sphere where either $\phi'(\zeta) \neq 0$ or else ϕ has a simple pole. Let $\omega = \phi(\zeta)$ and let ψ either be holomorphic at ω or have an isolated singularity there. If φ is a local inverse of ϕ in a neighborhood of ω , then the residue of $\psi(\phi(z))$ at $z = \zeta$ is equal to the residue of $\psi(z)\varphi'(z)$ at $z = \omega$.*

Rational Approximation

$$\begin{aligned}\exp(-f(t)) &= \exp(s) \exp(-s - f(t)) \\ &\approx \exp(s) \text{rational}_0(f(t) + s)\end{aligned}$$

$$\exp(x^2)\text{erfc}(x) \approx \begin{cases} 2 \exp(w^2) & \text{if } w \leq -3 \\ \text{rational}_1(w) & \text{if } -3 \leq w \leq 0 \\ \text{rational}_2(w) & \text{if } 0 \leq w \leq 10 \\ \text{rational}_3(w) & \text{if } 10 \leq w \end{cases}$$

Relative Error

If A approximates positive integrand F ,

$$\begin{aligned} \left| 1 - \frac{\int A}{\int F} \right| &= \left| \frac{\int F \{1 - A/F\}}{\int F} \right| \\ &\leq \frac{\int F |1 - A/F|}{\int F} \\ &\leq \|1 - A/F\|_{\infty} \end{aligned}$$

$$I(\alpha, \beta, \gamma, \delta, \xi)$$

1% tolerance : 0.000048 seconds (2.5 GHz AMD K10 core)

Lazy Evaluation

Carrying capacity and selective pressure constrain speciation
(limit number of phenotypes)

Number of integrals is quadratic in number of phenotypes

Compute integrals as needed and cache in memory for later use

Caching

Multi-level least-recently-used scheme (threaded splay trees)

First level : integrals associated with phenotype p
(create perfect hash h_p for p)

Second level : keys of the form $\langle h_p, h_{p'} \rangle$
integrals related to interaction of p with p' .

Increasing generations \implies Increasing phenotypes

Thrashing

- Second level cache is *quadratic* in phenotypes
- Can't eliminate cache; Must Reuse Integrals!

Distributed Second-level Cache

- BC : *SLOW* (cpu utilization 90%)
- AD : *faster* (cpu utilization 10%)

Equivalence

$$\text{Ne}(I, I') = \int \xi(I | u, v) \xi(I' | u, v) d\lambda(u, v)$$

$$I \equiv J \iff \xi(I | u, v) = \xi(J | u, v)$$

$$\text{Ne}(c, c') = \text{Ne}(I, I') \quad \text{where } I \in c, I' \in c'$$

$$\begin{aligned}
\sum_{I'} \text{Ne}(I, I') &= \sum_{I'} \text{Ne}(I, I') \sum_{c' \in \mathcal{C}} [I' \in c'] \\
&= \sum_{c' \in \mathcal{C}} \sum_{I'} \text{Ne}(I, I') [I' \in c'] \\
&= \sum_{c' \in \mathcal{C}} \sum_{I'} \text{Ne}(c, c') [I' \in c'] \\
&= \sum_{c' \in \mathcal{C}} \text{Ne}(c, c') \sum_{I'} [I' \in c'] \\
&= \sum_{c' \in \mathcal{C}} \text{Ne}(c, c') |c'|
\end{aligned}$$

Average behavior \iff Many runs

- Reduce cache *size*
- Avoid refilling empty caches
- Aggregate caching memory devoted to processors on a node
- Reduce run-time/run-space *overhead* for caching integrals

Precomputation

- Precompute integral-based functions of equivalence classes
- Memory map the read-only file of results

Naive implementation

- 32×32 patch of demes
- 4,150 children per deme per generation
- 100,000 generation epoch

Approximately 5,344,509,440,000,000 integrations

Kraken @ 90% utilization (66,048 cores) : over 1.8 *months*

Estimate average behavior (10 runs) : over 1.5 *years*

Canned integration routines \implies results could be meaningless

Less naive

Ten runs in parallel by using 10,250 cores : under 1.4 hours

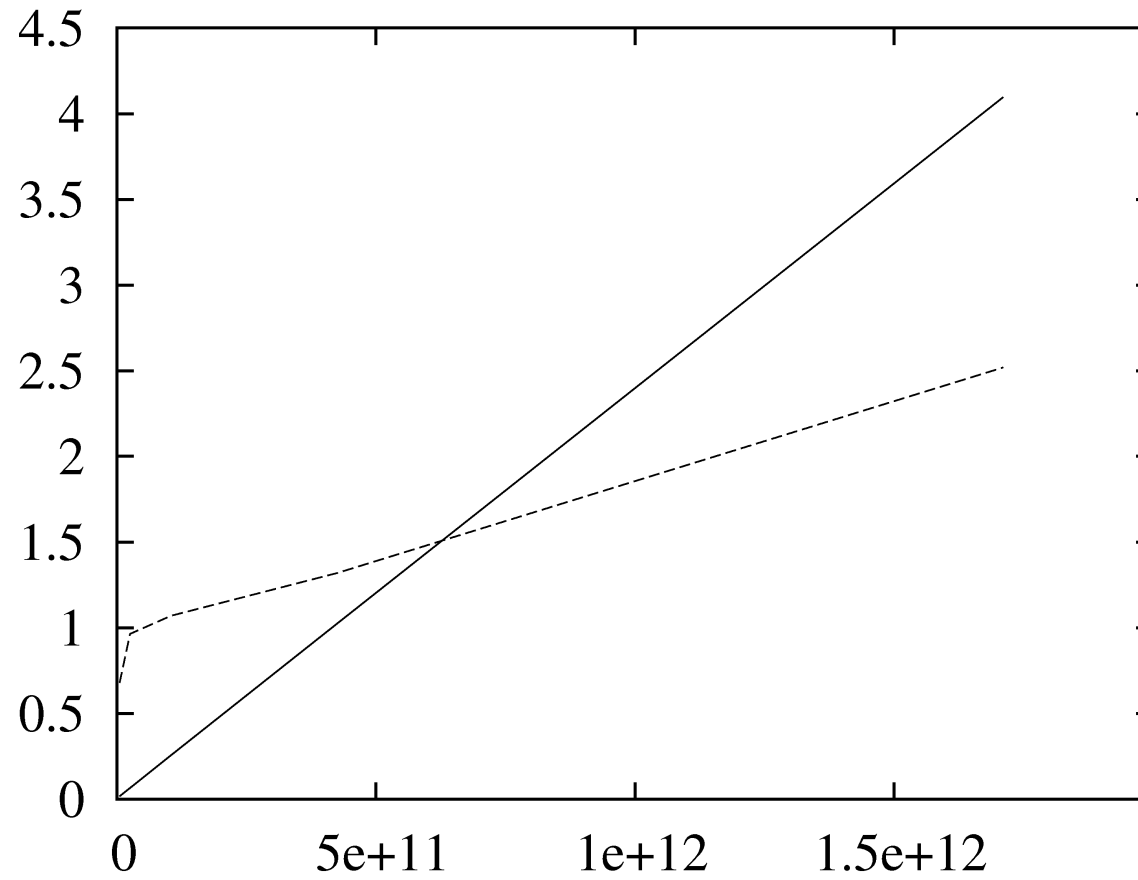
Some degree of confidence in the results

Increasing complexity \implies More equivalence classes

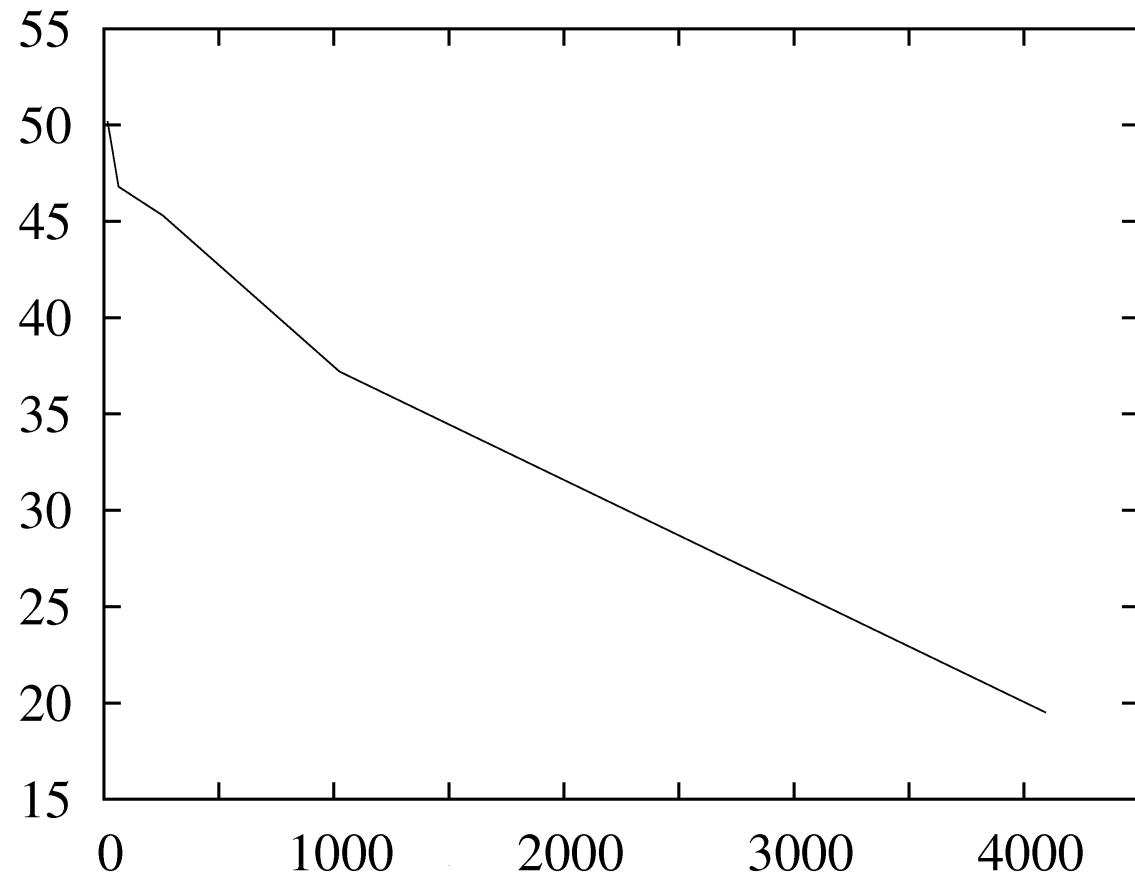
- Limited memory per node (16GB maximum)
- Need memory pages to efficiently map the result file
- Thrashing will set in as genome complexity increases

$$\# \text{ equivalence classes} = (1 + 2 * \text{gene-bit-complexity})^4$$

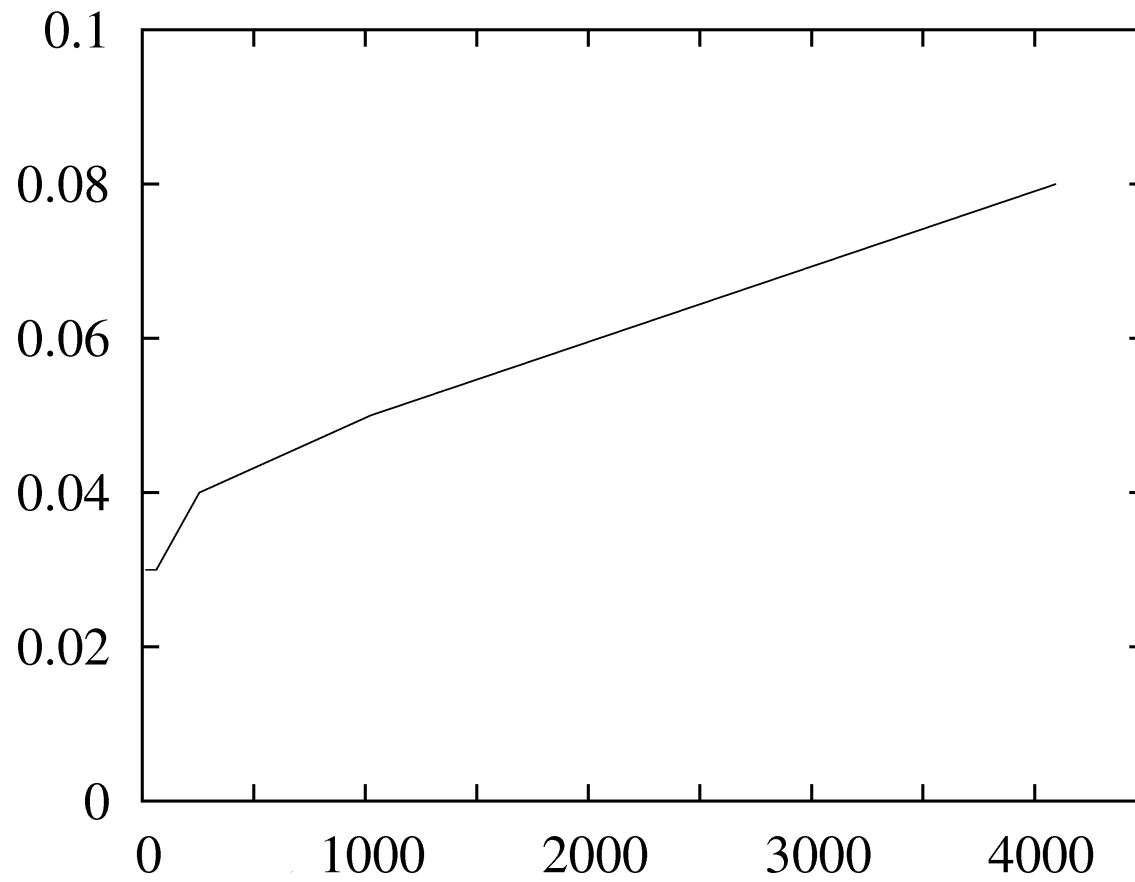
Scaling



CPU utilization



Time Per Generation



Extrapolating...

66,048 cores \implies 66,047 demes

- 1 second per generation
- 26.8 hours per epoch
- 27.6 trillion individuals

Tuning...

Logging 100 times per epoch

- Serialized: deme \longrightarrow output node \longrightarrow disk

Simulation essentially waits for logging to complete

MPI_THREAD_FUNNELED

- Buffer output at deme level
- Thread (at deme level) asynchronously writes buffers to disk

MPI transactions are eliminated

Disk writes are parallel

Potential speedup : 1 ↘ 0.05 seconds per generation

Time warp

Anolis model

- Spatial : demes on two dimensional grid
- Nearest-neighbors exchange genetic material
- Between-deme migration completes before next generation

Asynchronous Migration

Acknowledgements

“New approaches to the modeling of speciation”

National Institutes of Health (GM56693)