The Path of the Blind Watchmaker

Andy Poggio

SRI International

UC Berkeley

2011

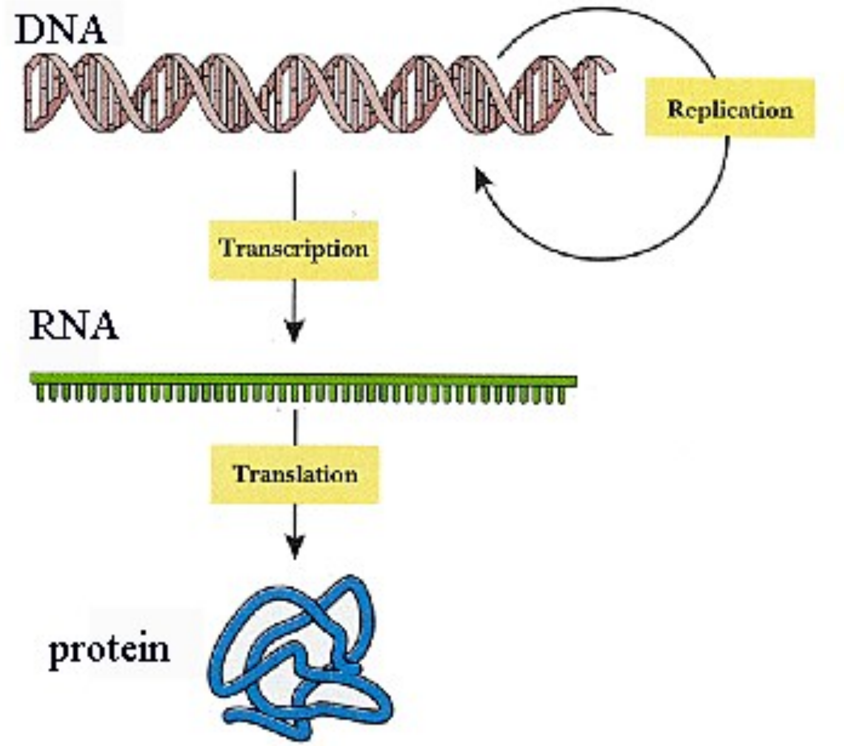
Outline

- Biology background
- The Last Universal Common Ancestor, LUCA
- Simple evolution model
- Reference species and their genomes
- Sequence evolution
- Population evolution
- Applications and future work

Evolution: the blind watchmaker

Central dogma of molecular biology

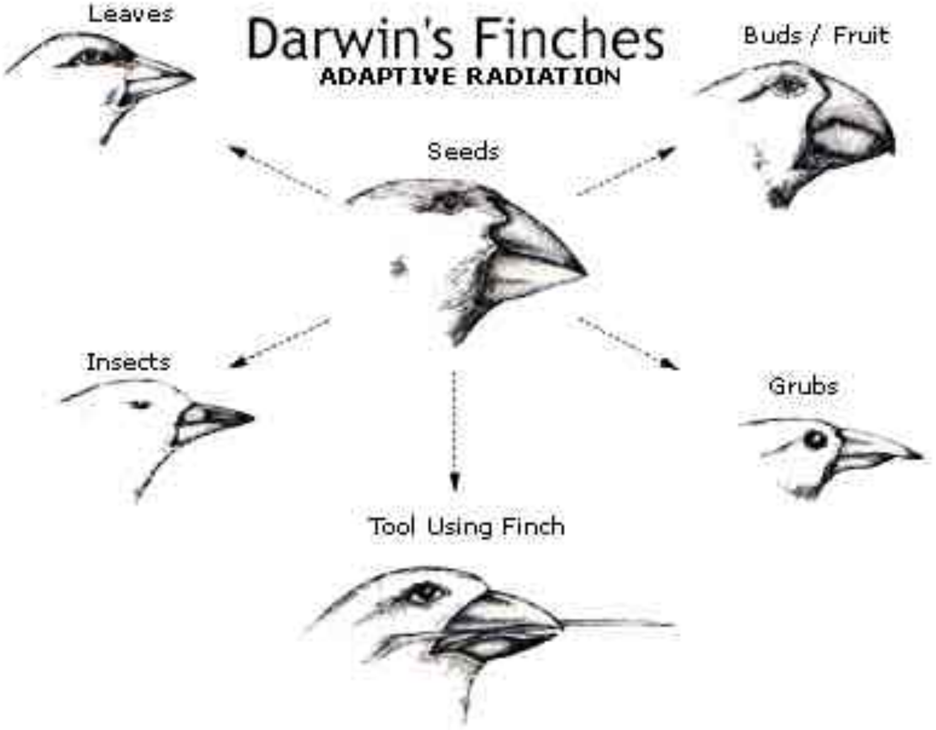
- DNA made up of 4 bases: a, t, c, g
- When replicated, occasional errors (mutations)
- Some DNA in genome is genes that code for proteins and regions that regulate them
  - Homologs are genes that evolved from a common ancestor gene
- Coding DNA transcribed to RNA
- RNA translated to protein by ribosome
- Proteins do work of cell



Evolution process

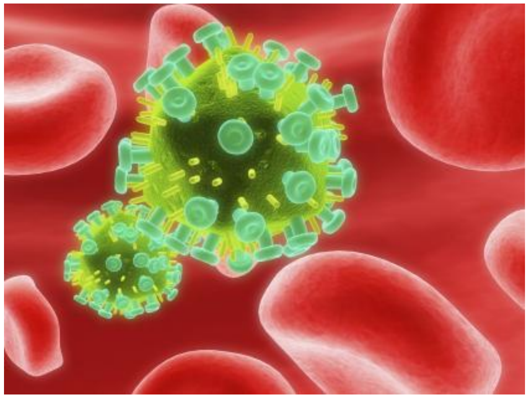
1) Variation of characteristics (genetic mutation)
2) Propagation of variation: reproduction and inheritance (duplicate of parent's genome in offspring)
3) Environment has selective effects on variations (fitness affects longevity and/or fecundity)

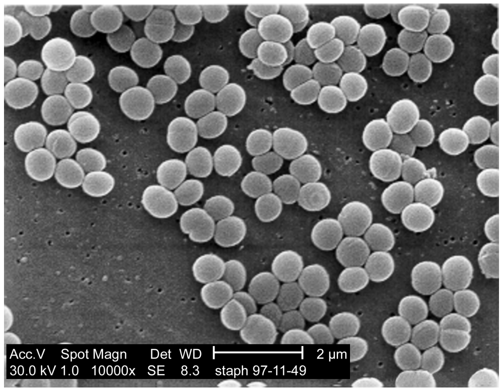
- **With these three components, evolution must occur**



Pathogen evolution

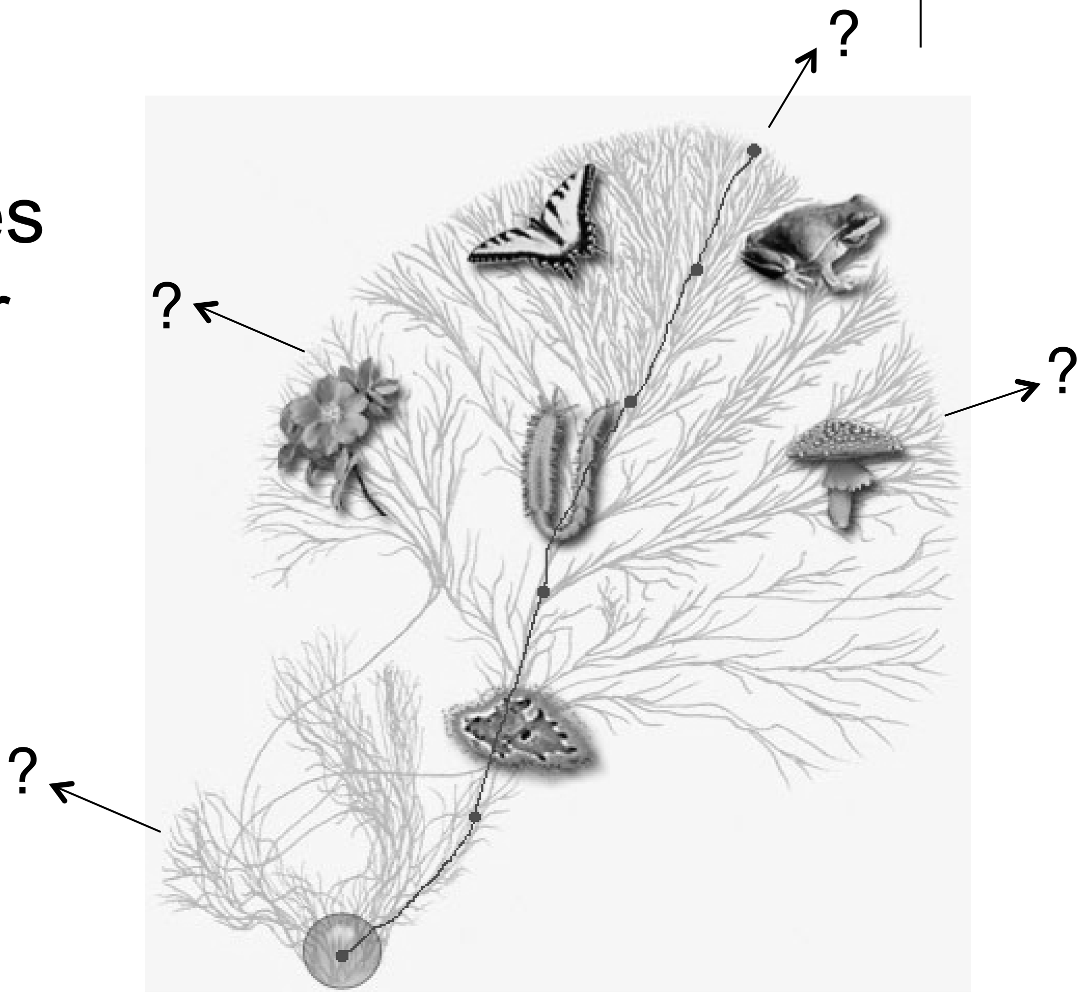
- 3.1 million deaths in 2005 due to HIV virus
- Antibiotic vancomycin “drug of last resort” for bacterial infections
- 20-fold increase in vancomycin-resistant bacteria from 1987-1993
- Pathogens evolve treatment resistance
- We need to understand, predict




Acc.V Spot Magn Det WD 2 µm
30.0 kV 1.0 10000x SE 8.3 staph 97-11-49

Future of life

- Evolution shapes us (and all other life)
  - Stochastically
  - Consciously determined

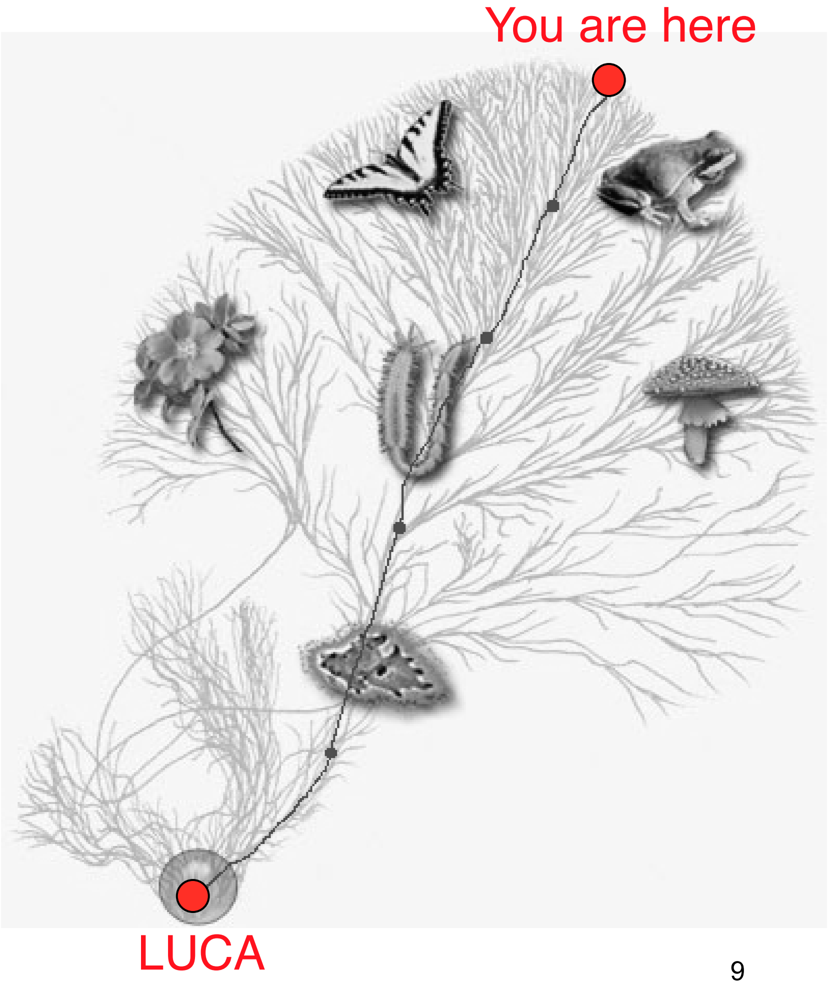


LUCA

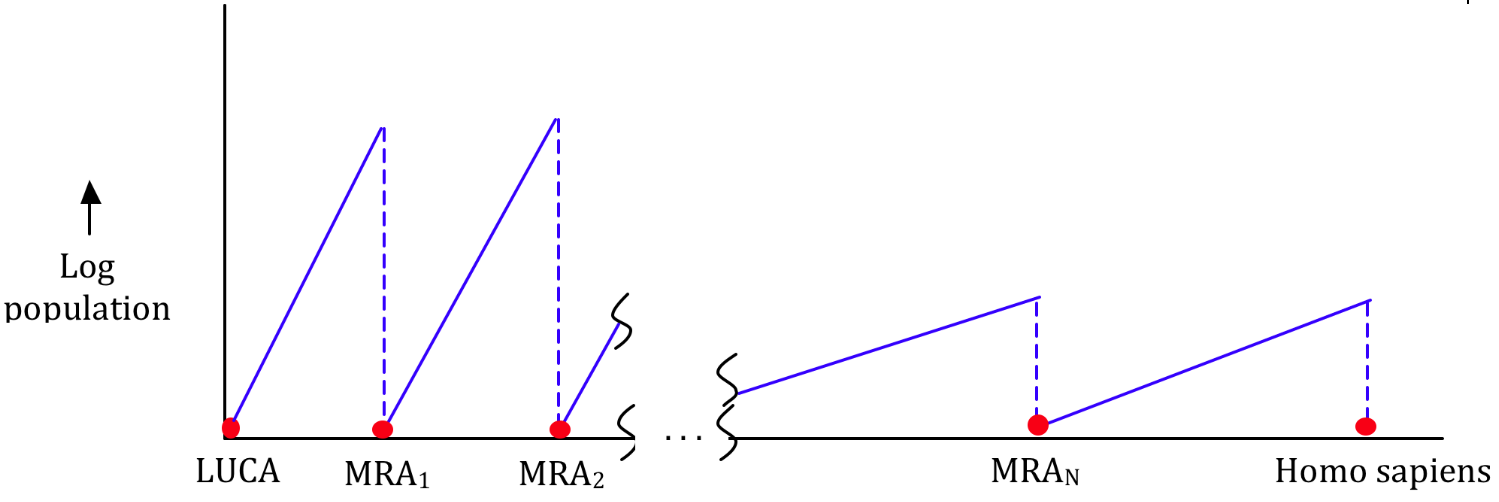
- LUCA is branching point; life exists prior to LUCA
- Consensus:
  - single-celled organism with 500-1000 genes
- Controversy:
  - Simple prokaryote or complex, single-cell protoeukaryote – exons/ introns “piece” together proteins
  - DNA or RNA genome – RNA has high mutation rate, rapid evolution
  - If protoeukaryote, then reductive evolution produced prokaryotes (e.g. bacteria) – prokaryotes “more efficient”

How did we get here from LUCA?

- A simple evolution model:
  - One mutation at a time makes a More Recent Ancestor (MRA)
  - Each MRA proliferates until a next MRA emerges
- Generation ≤ MRA ≤ Speciation



Simple model structure



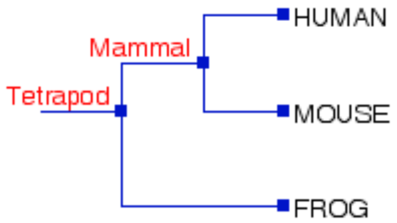
LUCA – Last Universal,
Common Ancestor

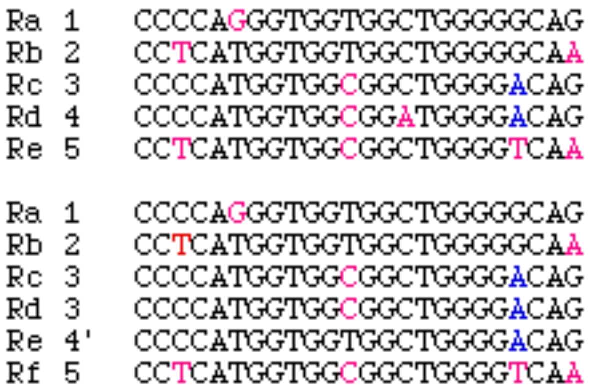
MRA – More Recent Ancestor

- Using mutation rate, growth rate, and sequence length from the literature, calculated $1.1*10^9$ years compared to $3.5*10^9$ years accepted time
- Relevant to actual process but significantly incomplete

Comprehensive model

- Input data: reference species (including LUCA) and their genomes





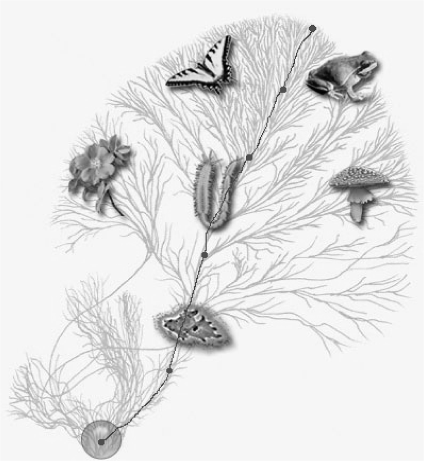
```
Ra  1    CCCCAGGGTGGTGGCTGGGGGCAG
Rb  2    CCTCATGGTGGTGGCTGGGGGCAA
Rc  3    CCCCATGGTGGCGGCTGGGGACAG
Rd  4    CCCCATGGTGGCGGATGGGGACAG
Re  5    CCTCATGGTGGCGGCTGGGGTCAA

Ra  1    CCCCAGGGTGGTGGCTGGGGGCAG
Rb  2    CCTCATGGTGGTGGCTGGGGGCAA
Rc  3    CCCCATGGTGGCGGCTGGGGACAG
Rd  3    CCCCATGGTGGCGGCTGGGGACAG
Re  4'   CCCCATGGTGGTGGCTGGGGACAG
Rf  5    CCTCATGGTGGCGGCTGGGGTCAA
```

- What happened? Sequence evolution model

- How did it happen and how long did it take? Population evolution model

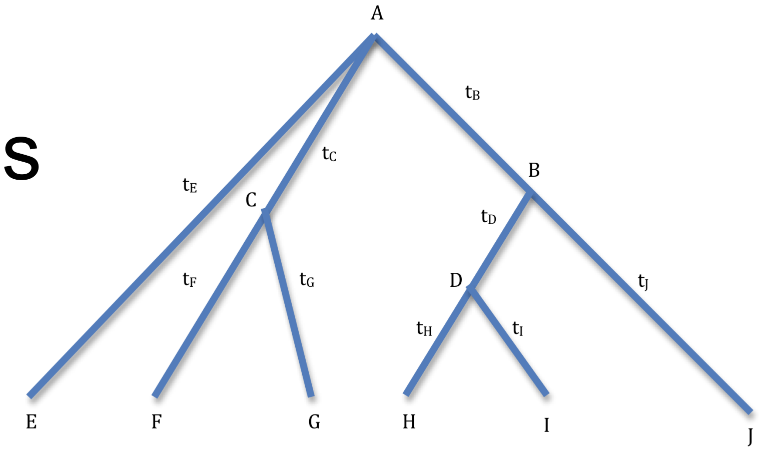
Reference species

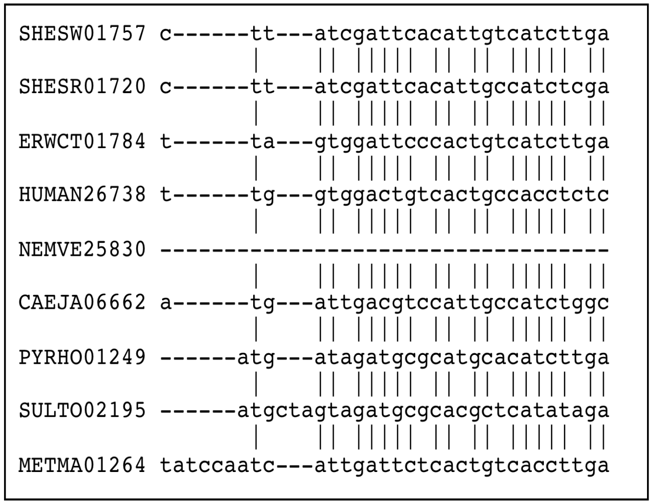


- Chosen for distinctions, not equal time intervals
- LUCA
- LUCAEukaryota -- organelles (e.g. nucleus, mitochondria, chloroplast), multicellular, sexual reproduction, exons/introns
- LUCAMetazoa -- heterotrophic (engulf food), motion, developmental stage due to gene regulation
- LUCAMammalia -- warm-blooded, vertebrate, mothers nourish young, neocortex
- Homo sapiens

Reference species genome reconstruction

- Need actual sequences
- Infer from existing species sequence data:
  - Phylogenetic tree creation
  - Multiple sequence alignment to determine corresponding bases
- Used existing tools together with new tool for reconstruction





```
SHESW01757 c------tt---atcgattcacattgtcatcttga
                  |     || ||||| || || ||||| ||
SHESR01720 c------tt---atcgattcacattgccatctcga
                  |     || ||||| || || ||||| ||
ERWCT01784 t------ta---gtggattcccactgtcatcttga
                  |     || ||||| || || ||||| ||
HUMAN26738 t------tg---gtggactgtcactgccacctctc
                  |     || ||||| || || ||||| ||
NEMVE25830 -----------------------------------
                  |     || ||||| || || ||||| ||
CAEJA06662 a------tg---attgacgtccattgccatctggc
                  |     || ||||| || || ||||| ||
PYRHO01249 ------atg---atagatgcgcatgcacatcttga
                  |     || ||||| || || ||||| ||
SULTO02195 ------atgctagtagatgcgcacgctcatataga
                  |     || ||||| || || ||||| ||
METMA01264 tatccaatc---attgattctcactgtcaccttga
```

Reference species genes

	Nonhomologous	Homologous	Total
LUCA	33		33
LUCAEukaryota	43	33	76
LUCAMetazoa	43	76	119
LUCAMammalia	44	119	163
Homo sapiens	39	163	202

- Nearly 600 genes total
- LUCA deoxyribonuclease, involved in DNA manipulation and repair

```
atggaatacaaacccatgccttatccaatgattgattctcactgtcatcttgatattccagaatttgatc
atgacagagatgaagccattcagaaagccaaaaaaacaggtgttgtcgtaatggtggcaattccggaatt
tgccttgaaagaaattgaaaaagtcttgaaaattttcgaggaaaattacgagaatgttctttcagcactg
ggttttcatcccgatatcggtgaaaaagatatcaactaaaatgaattggataaaagttaagcaatagctg
gaaaggcggtagctatcggagaagtcggcctagattattattactgcaaaacagacgaggaaaggaaaaa
acagagagctttatttgaaaagctgatcgagcttgccaaagaactggaaatgcctgtggttgtgcatgcc
agaatggctgaaagagaagccattaatattctccaagagcttgacggggacatagtcaccgtaatttttc
actcctataccggctctgttgaaaccgcaaaggaaatagtggaagcaggctactttatctcaatggctgg
aattgtgaccttctgtcattccgaacattagcaaaaagttgcagaaaaagtgcccctcgaaaacctgctg
ctcgaaacagattctccttttctggcccctataagacaccggggtcagaaatgagccatggattgttaat
attatccctgaagagattgccagaattaaggaaatggcacttgaagaagttgctgaaataacaactgaaa
acgcacgcaaattttttcctaagctggctcggttgctcaagatataa
```

Mutations

- 14 mutation types
- Essential mutations for model:
  - substitutions

    ```
    atcg
    | ||
    aacg
    ```

  - Insertions/deletions (indels)

    ```
    a-cg        atcg
    | ||        | ||
    atcg        a-cg
    ```

  - Inversions (reverse+complement) `atcg` → `gcta` → `cgat`
- Others common bulk adds or subtracts
- Made survey of empirical mutation rates; arithmetic means of relevant species used

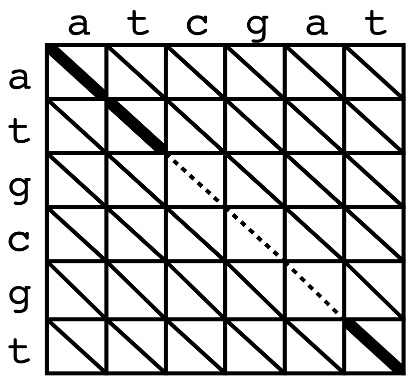
Sequence evolution model

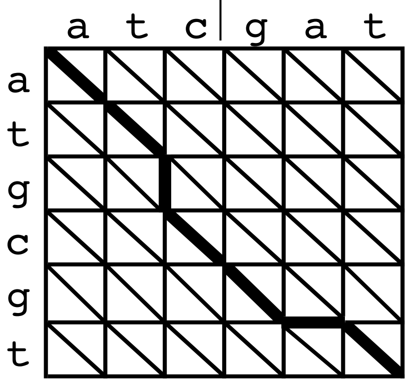
```
cgaaagcggcgttccgaccttcagcggggccatggatggactgt
 ||||| || |||||||||||||| || ||    | |      |
agaaagtggtgttccgaccttcagaggagctggaggt---tatt
```

- Sequence evolution is set of mutations that occurred as one sequence evolved to another
- Determined through pairwise sequence alignment of each reference species gene with predecessor reference species homolog or other gene
- Homologs aligned with homolog in previous reference species
- Nonhomologs aligned with unrelated genes in previous reference species and with random sequences

Sequence alignment

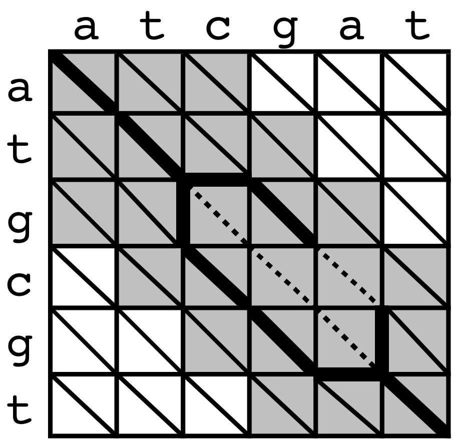
- Global, end-to-end alignment
- Alignment scores based on mutation rates
  - indel and inversion scores are a function of length
- Multiple paths/alignment
  - more paths for longer sequences
- Most probable paths near diagonal
- Nearly 50,000 alignment paths produced





```
a t g c g t
| |       |
a t c c a t
```

```
a t g c g - t
| |   | |   |
a t - c g a t
```



Finding inversions

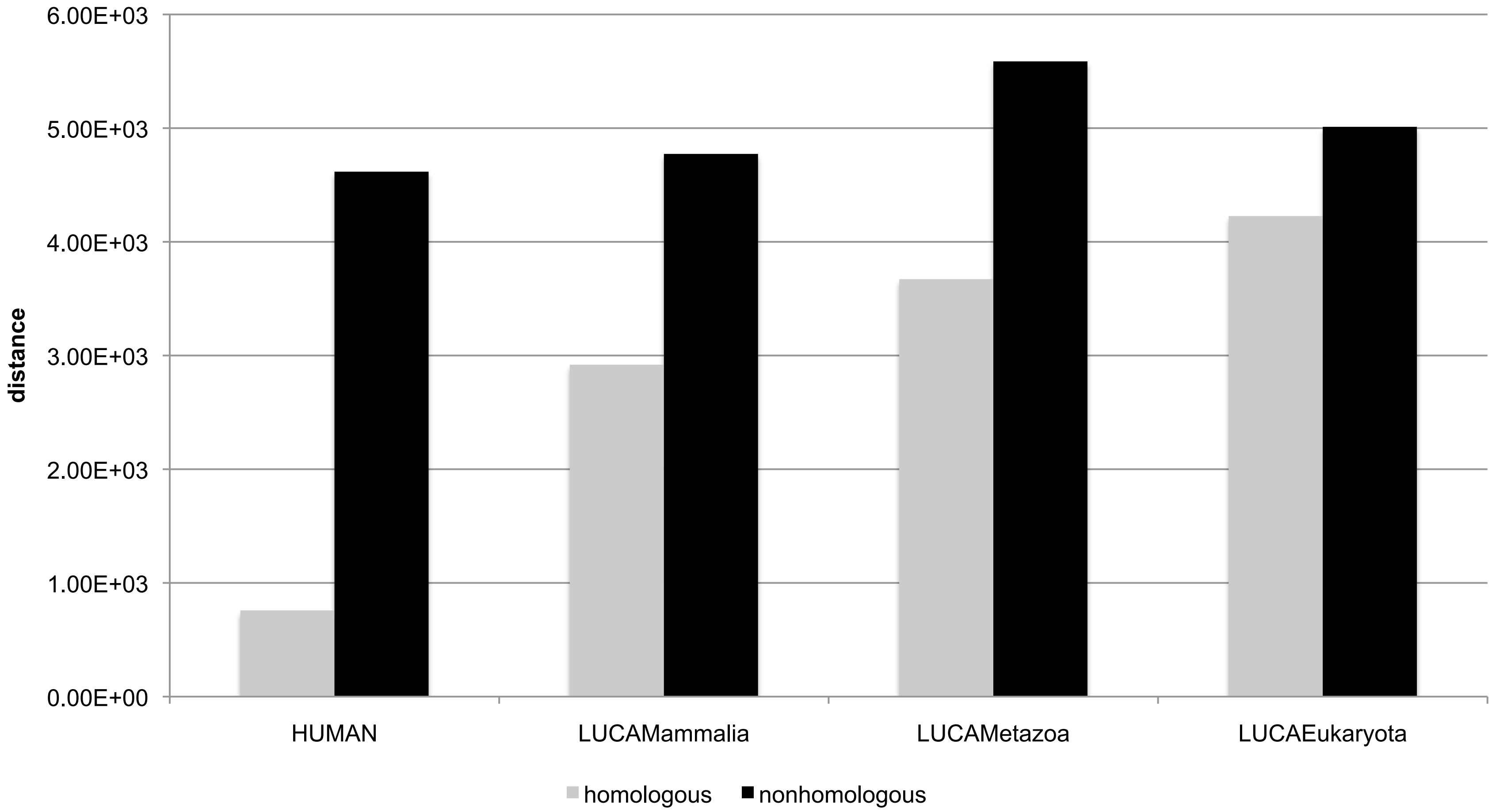
- Distinct from global alignment algorithm – inversions can start/end anywhere; want probable ones
- Inversion must end when no longer probable
- Inversions must be aligned as may contain mutations

	c	c	c	g	a	a
c						
g						
g						
g						
a						
a						

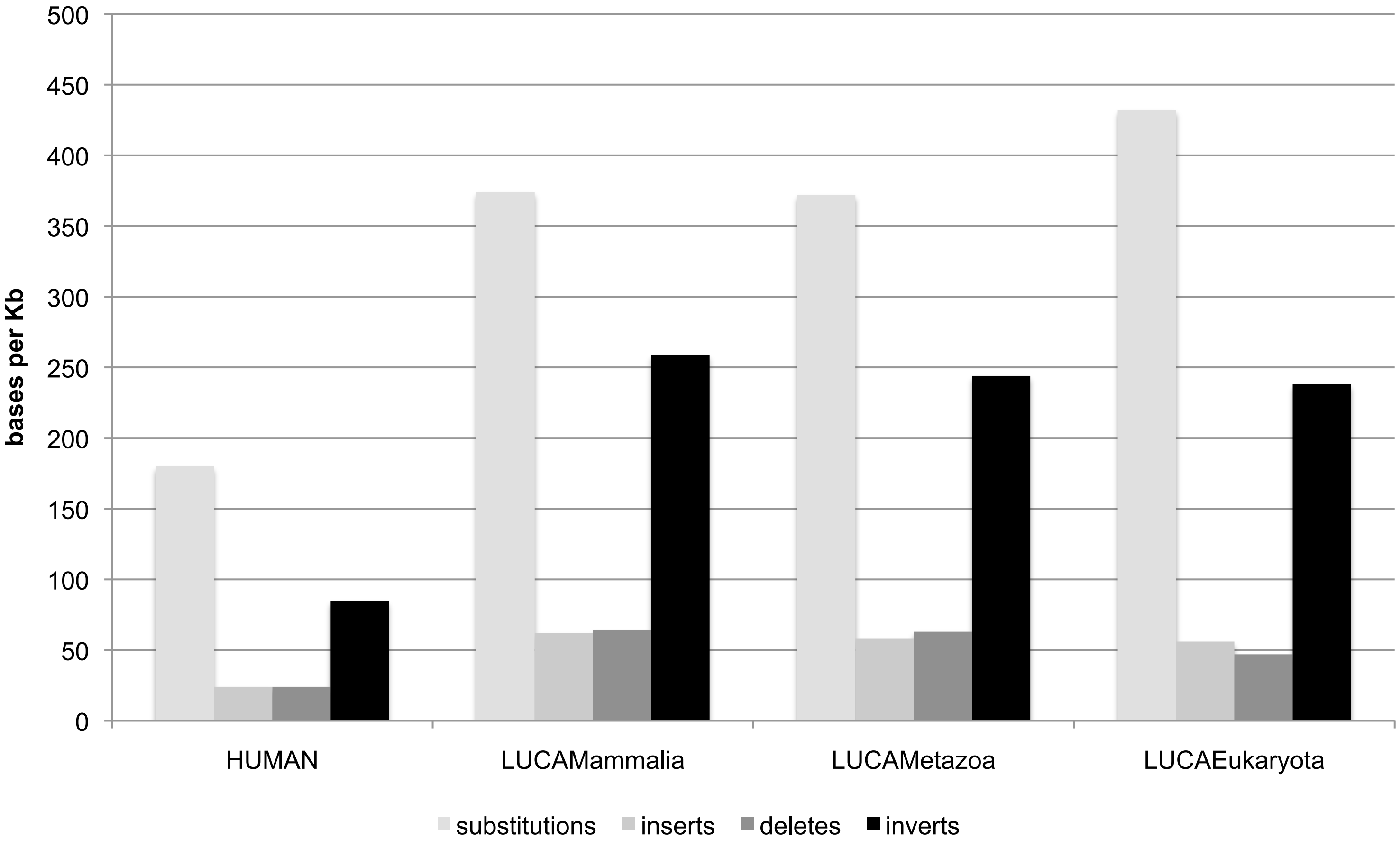
⇨

	c	c	c	g	a	a
T			▨	▨	▨	▨
T		▨	▨	▨	▨	▨
C	■	▨	▨	▨	▨	▨
C	▨	■	▨	▨	▨	▨
C	▨	▨	■	▨	▨	
G	▨	▨	▨	■		

Homologous/nonhomologous distance comparison



Reference species mutation comparison

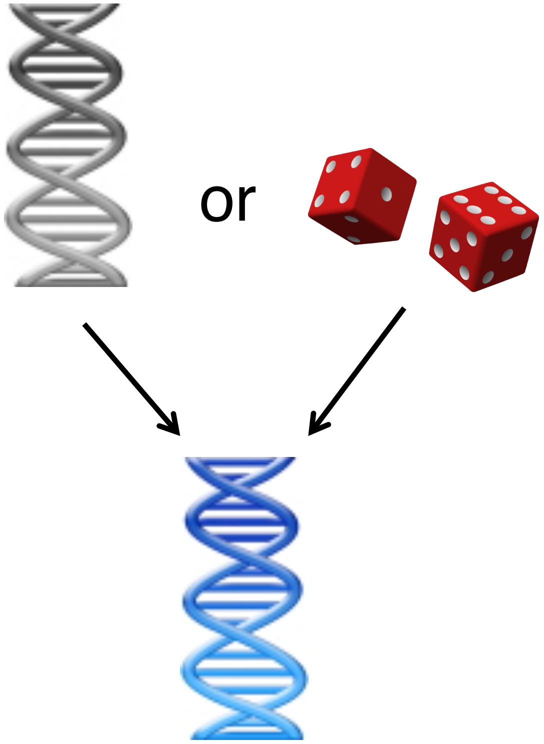


Inversions

- Microinversions length 4 detected under special circumstances
- Minimum length 12
- All alignments performed with and without inversions
- Conclusion: Inversions reduce alignment distance (increase alignment probability), confidence >99%

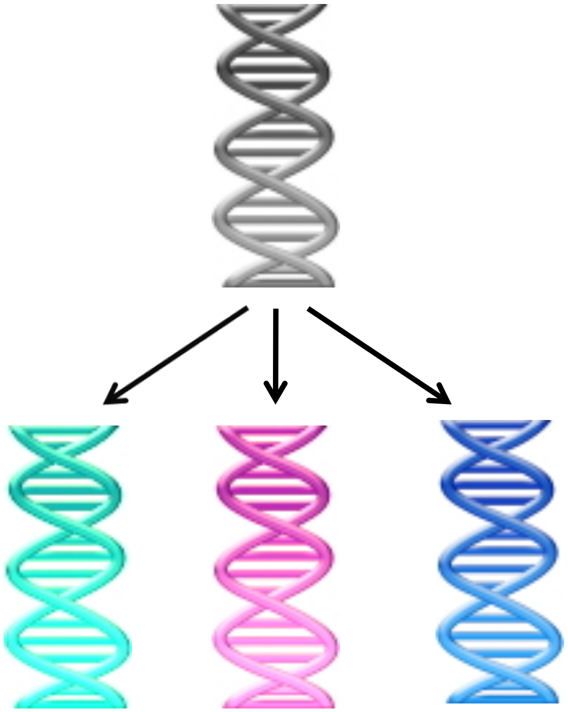
Nonhomologous gene evolution

- Must come from unrelated gene sequence or random sequence
- Modest confidence (>80%) coding sequence more likely for most reference species
- Likely due to protein secondary and tertiary structures that are functional in many contexts

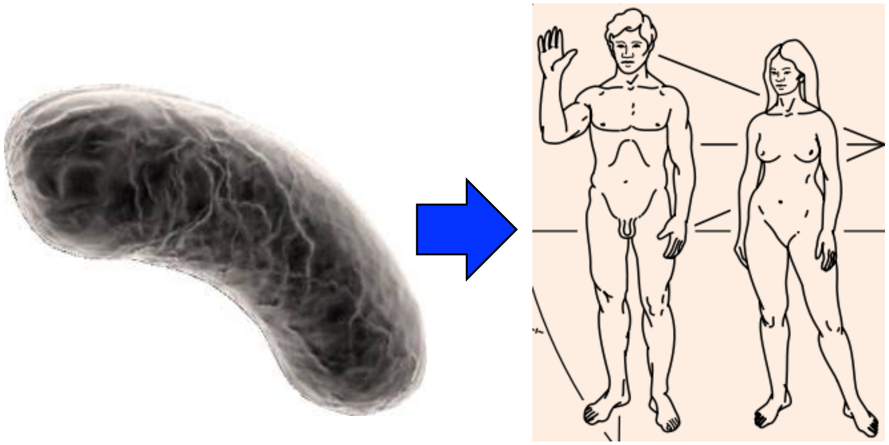


Universal source sequence

- Gene sequence better than random sequence for creating nonhomologous genes – some genes better than others?
- 4 LUCAMammalia genes aligned with 39 nonhomologous Homo sapiens genes
- Small sample size provided modest evidence for universal source sequence
- Best source gene was 21530LUCAMammalia
  - Homologs back to LUCA
  - No consensus function in LUCA
  - Speculation: function is to act as universal source sequence



How to make a Homo sapiens



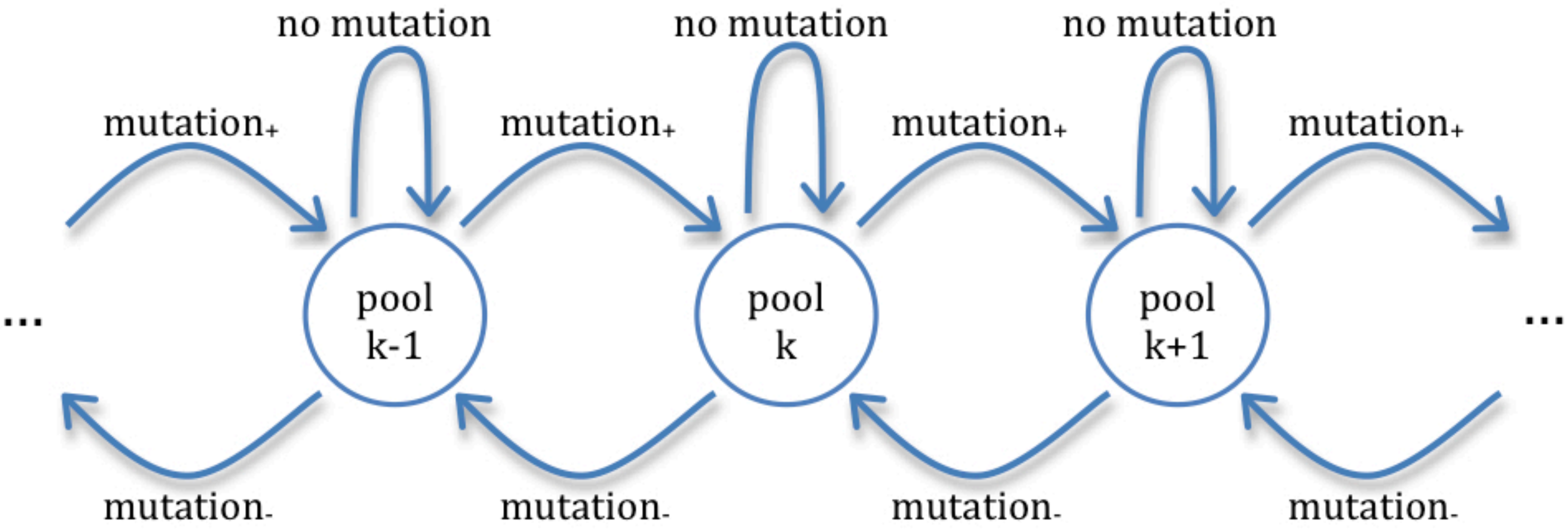
- Start with a LUCA genome
- Insert 26,000,000 bases
- Delete 25,700,000 bases
- Substitute 177,000,000 bases
- Invert 107,000,000 bases
- Add bulk DNA; use any of several available mechanisms
- Enjoy your new species with its consciousness, intelligence, creativity, and empathy
- Key question: how long did it take? Need population model for answer

Population model

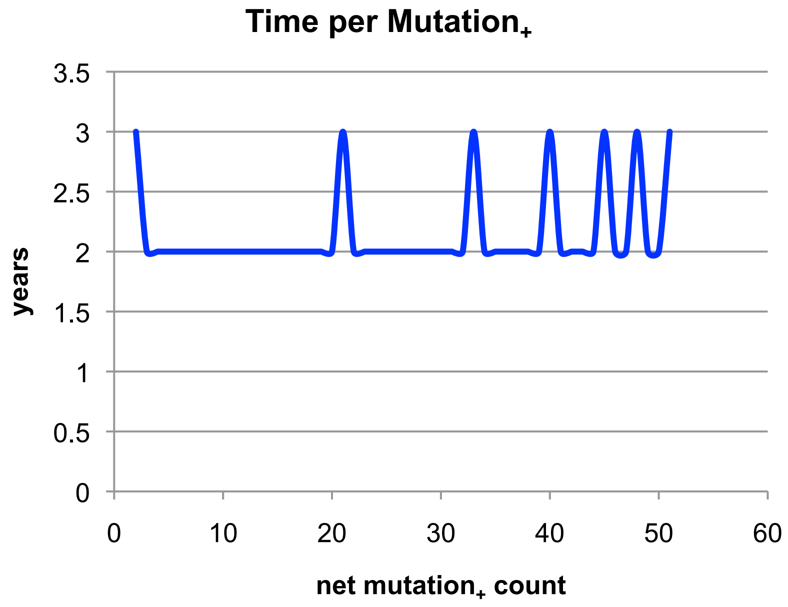
- Population evolution simulation
- Two types of mutations:
  - mutation$_+$ makes an MRA
  - mutation$_-$ nullifies a mutation$_+$
  - probabilities defined by mutation rates survey and sequence evolution model results
  - P(mutation$_+$) < P(mutation$_-$) – many ways to nullify a mutation$_+$
- Confined to LUCAMammalia to Homo sapiens evolution because good estimates for earlier species model parameters not available
- Model sequence length < Homo sapiens effective sequence length
- Standard model length 200, scaled up where needed; other lengths also investigated

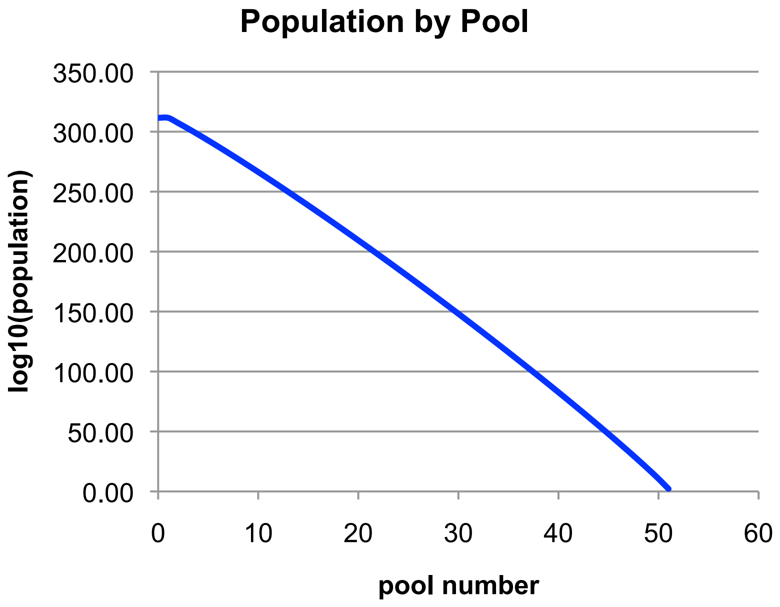
Population pools

- Pools numbered from 0 to *n*
- Pool$_k$ contains individuals with *k* net mutation$_+$s
- Newborns have mutations based on empirical probabilities
- When pool *n* population ≥ 1, model run complete
- Pools whose numbers are close are said to be similar



Population evolution model 0.1



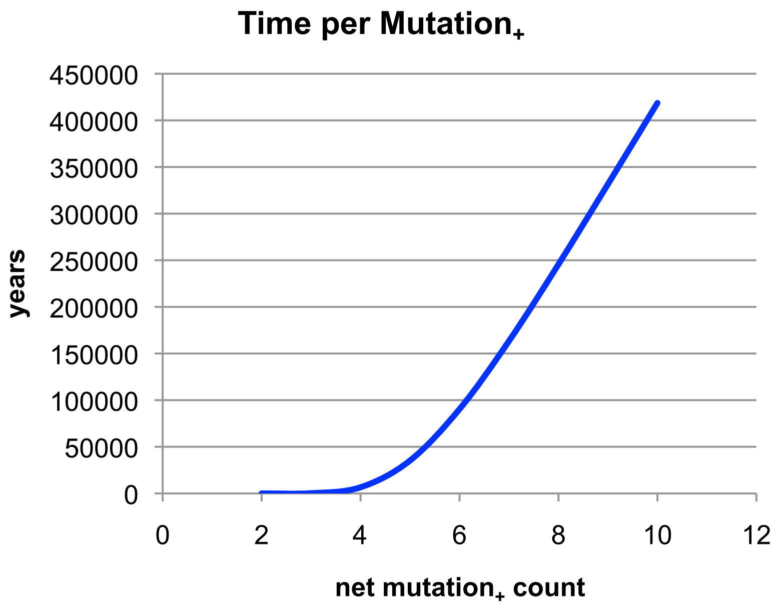


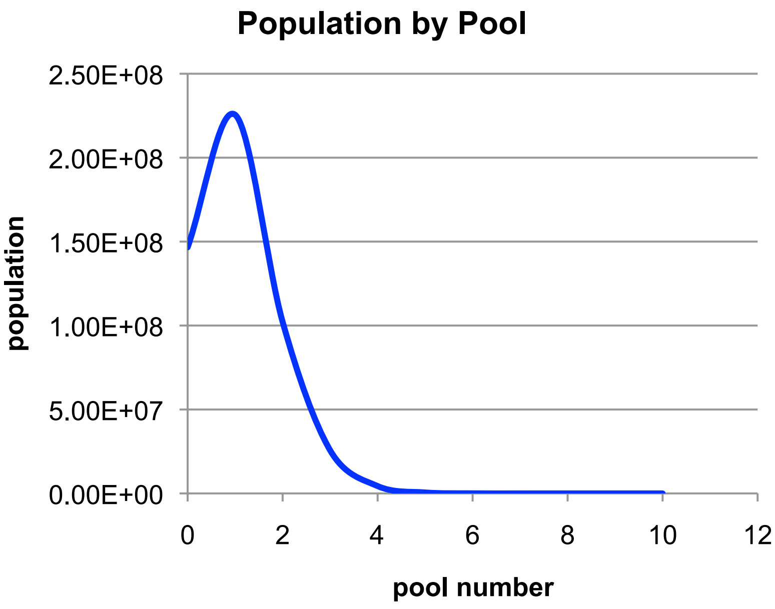
- Reasonable time/mutation$_+$
- Populations problematic

Carrying capacity

- Resources, competition, predation limit species population in an environment
- *g* = *birthRate-deathRate*
- d*pop*/dt = *g*pop**(1-(*pop*/*K*)), *K* carrying capacity
- *pop* approaches K, *g* approaches 0 and *birthRate, deathRate* approach each other
- *birthRate* ≠ 0
- Used mean of mouse and human estimates

Population evolution model 0.2

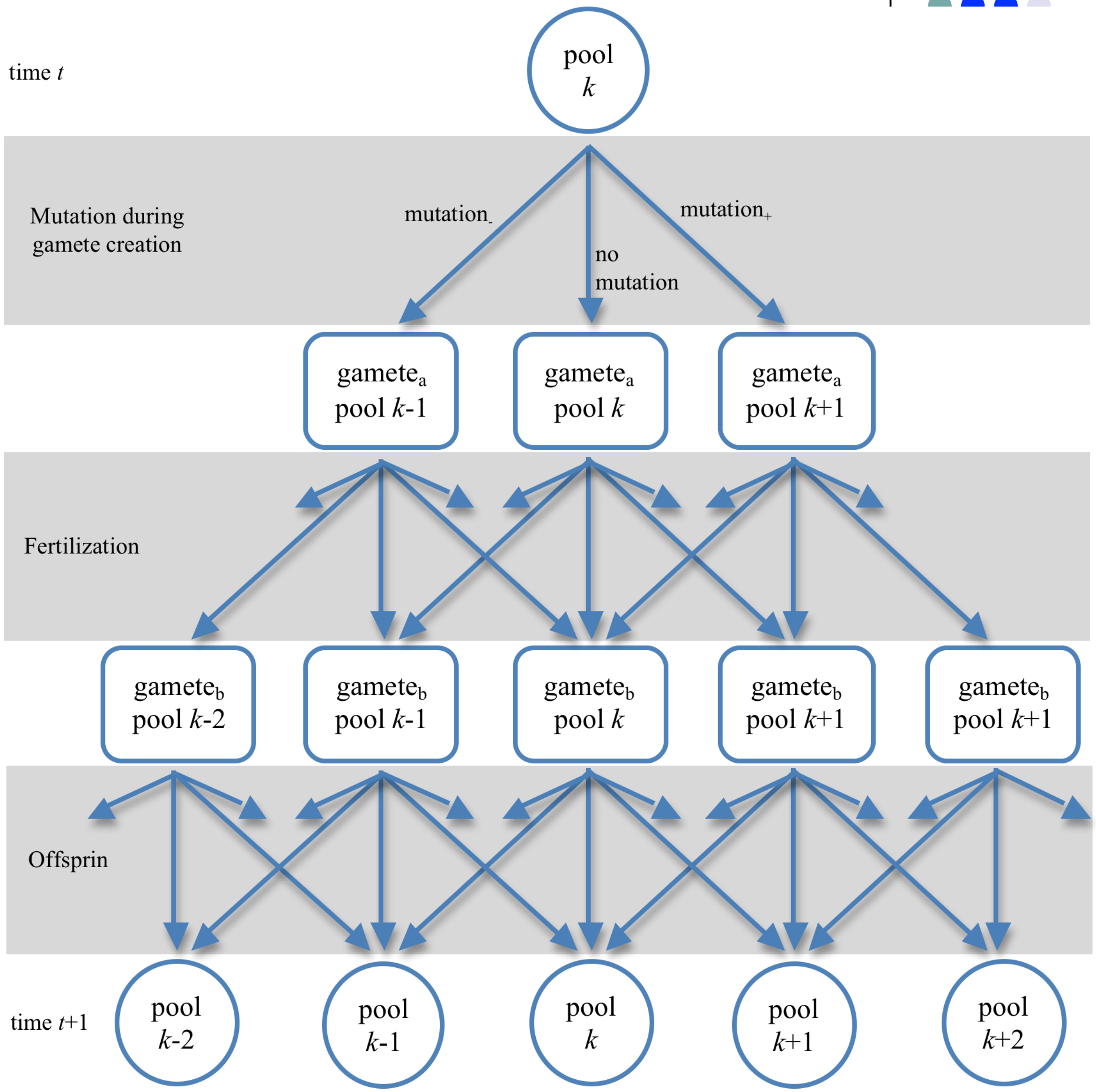




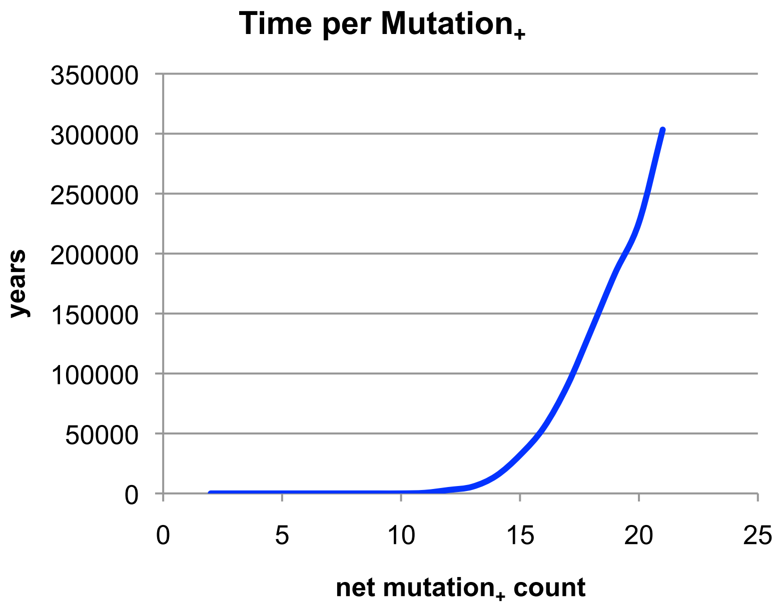
- Time/mutation$_+$ too long (model run terminated early)
- Populations reasonable

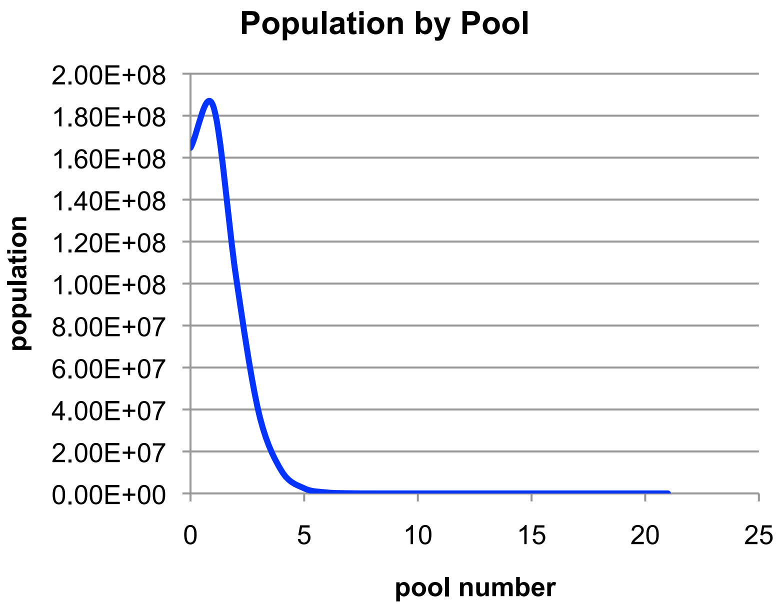
Sexual reproduction

- Two individuals from $pool_k$, $pool_l$ have $(k*l)/n$ $mutation_+$s in common
- They have $(k+l)-(2*(k*l)/n)$ distinct $mutation_+$s
- Offspring inherit all common $mutation_+$s and a binomial distribution of distinct $mutation_+$s
- Zygotes placed in broader pool range than parents
  - parents $pool_8$, $pool_9$
  - zygotes $pool_7$ to $pool_{10}$ inclusive



Population evolution model 0.3



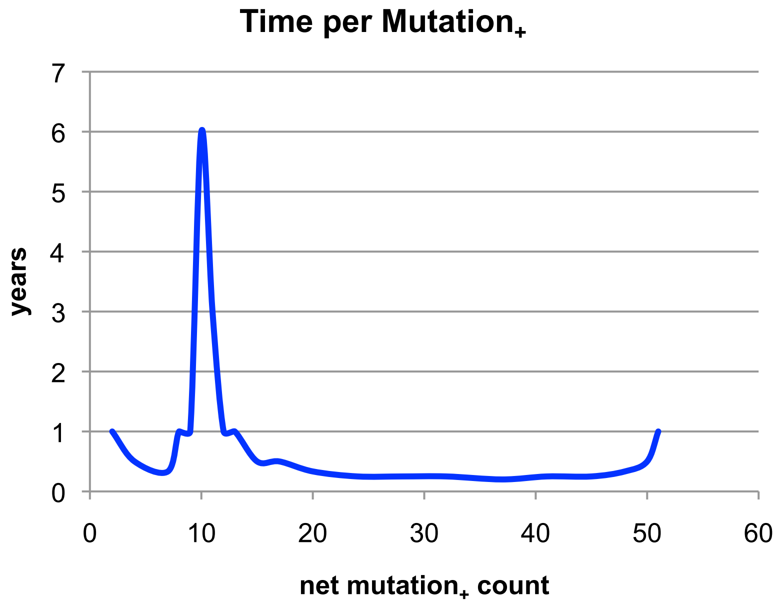


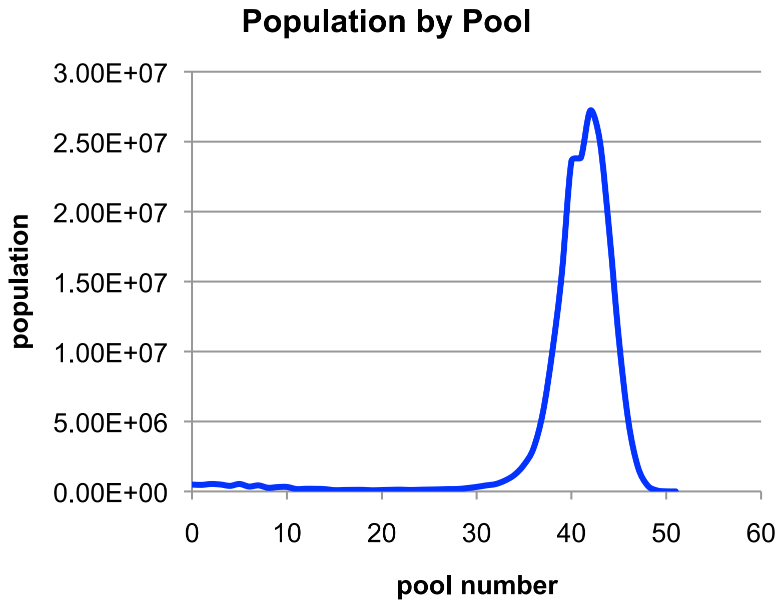
- Time/mutation$_+$ better but still too long

Fitness

- Mutation$_+$s may confer some fitness advantage
- Most fit (highest pool) has fitness 1.0
- Less fit genotype *i* has relative fitness 1-s_i where s_i is the selection coefficient against genotype *i* compared to fittest
- Pool$_i$ with less mutation$_+$s than $pool_{fittest}$ has birth rate reduced by $1-((\mathit{fittest}\text{-i})*s)$ where *s* is selection coefficient for model

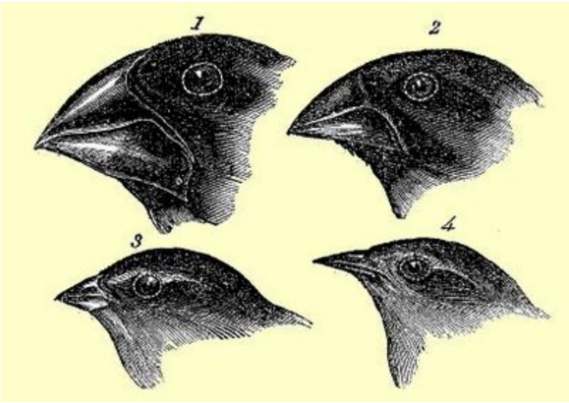
Population evolution model 0.4





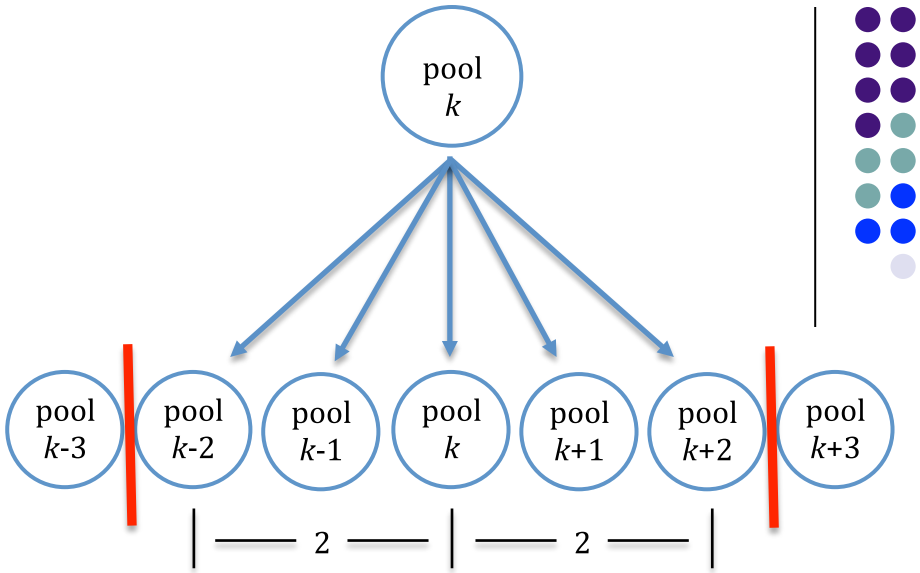
- Fitness selection coefficient 10%
- Time/mutation$_+$ good
- Selection coefficient unrealistically high
- Modest value of 1% more appropriate

Nonrandom mating



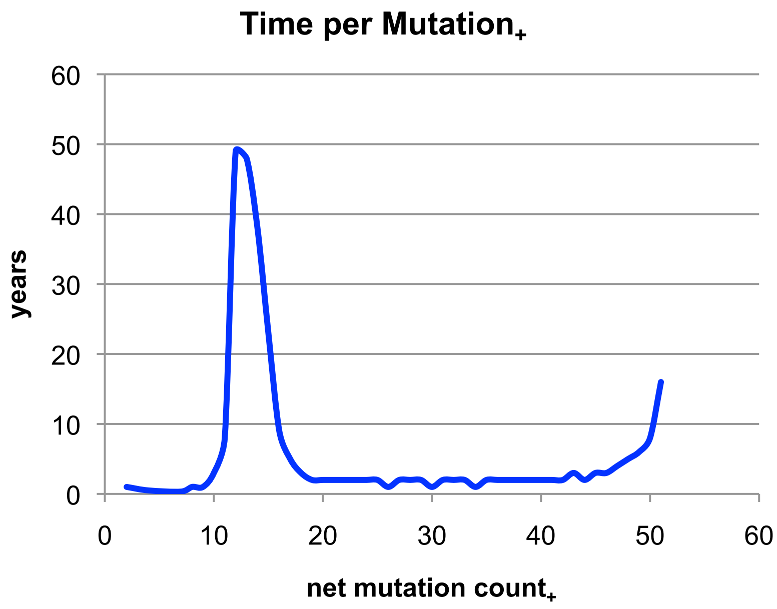
- Classic population models, e.g. Hardy-Weinberg, assume random mating – frequently inaccurate
- Speciation
  - many speciation events between LUCAMammalia and Homo sapiens
  - can’t mate outside of species
  - model sequence length less than Homo sapiens sequence length – speciation implied at boundaries of model sequence length

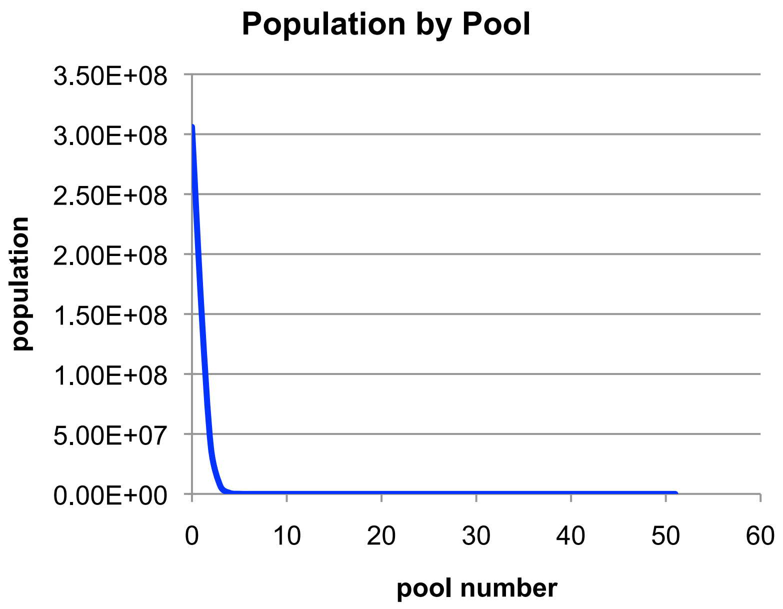
Mating radius



- Maximum difference in pool numbers that two mates can have
- With mating radius 2, pool$_k$ members can mate with pool$_{k-2}$ to pool$_{k+2}$
- Speciation limits mating radius
- Consider mates from pool$_k$ and pool$_l$
  - Offspring go into pools with binomial distribution having peak at $(k+l)/2$; offspring go into pools similar to pool$_k$ and pool$_l$
  - Mammals have small natal dispersal, so mate with individuals from similar pools, hence limited mating radius

Population evolution model 1.0


Time per Mutation$_+$

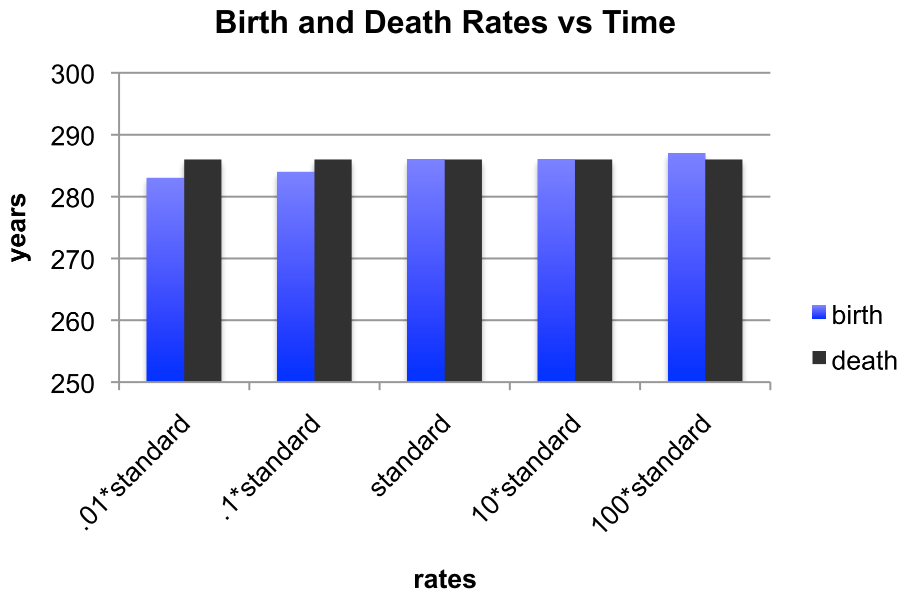

Population by Pool

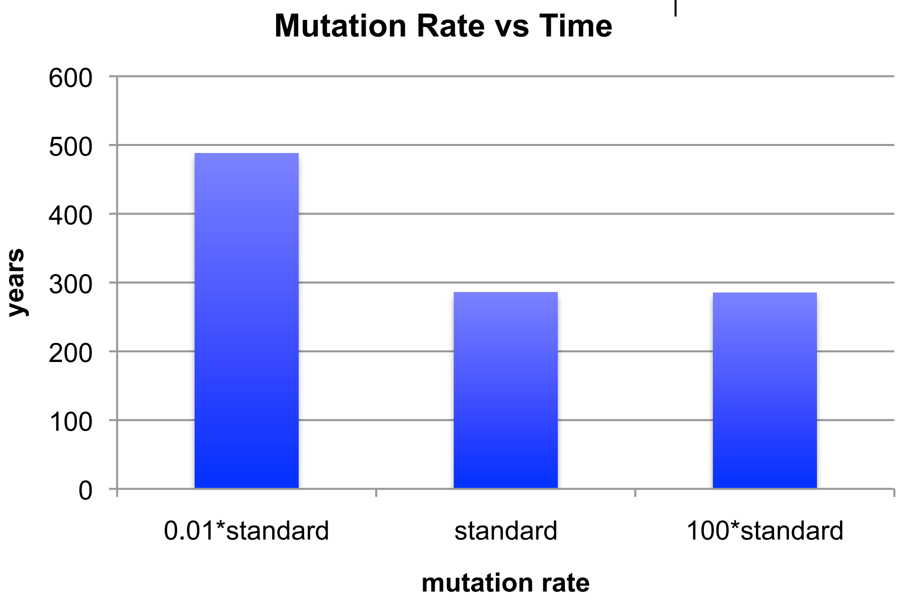
- Standard model has carrying capacity, sexual reproduction, selection coefficient 1%, mating radius 5
- Time/mutation$_+$, population both good

Evolution duration estimate

- Estimate for LUCAMammalia to Homo sapiens
- Using standard model with parameter values obtained from literature or otherwise estimated
- Model duration of 186 million years compares well with broadly accepted estimate of just over 200 million years
- Key question: was there enough time? Model demonstrates that there was
- Using other reasonable estimates for parameters, can obtain values from 0.5 million years to greater than age of universe

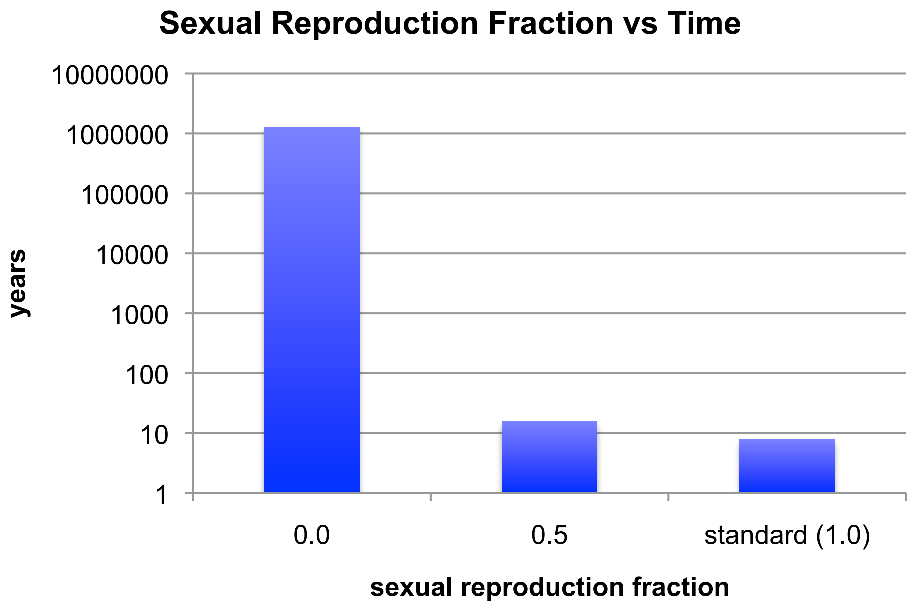
Insensitive population evolution parameters

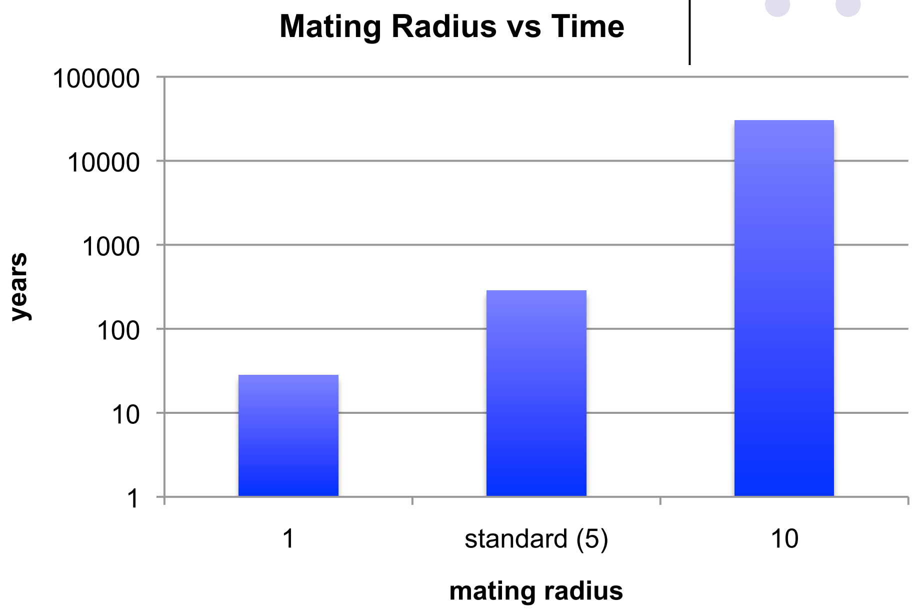




- Birth rate or death rate – very small change over 4 orders of magnitude
- Mutation rate – small change over 4 orders of magnitude

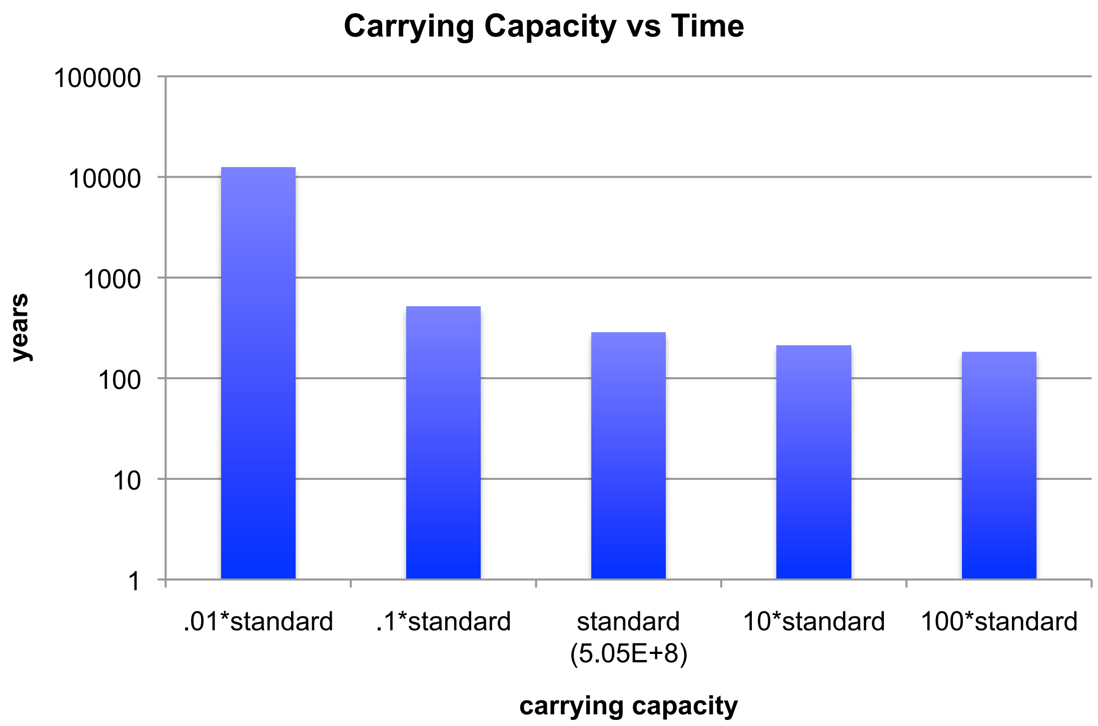
Top 4 population evolution parameters





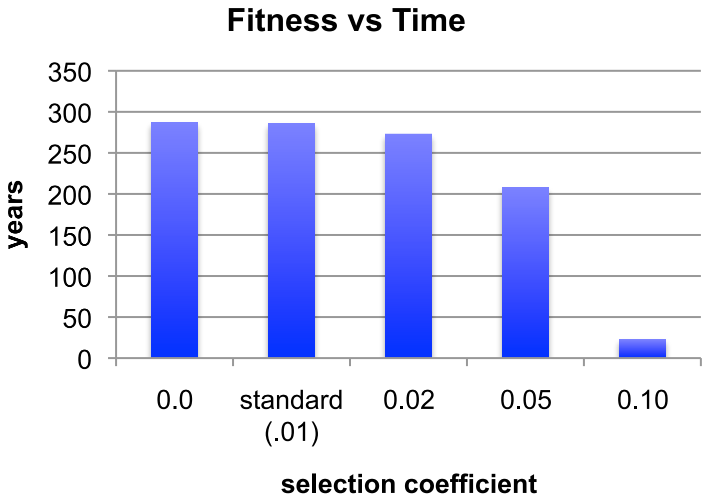
- Sexual reproduction and mating radius both have exponential effects with small changes in parameter values
  - sexual reproduction used model sequence length smaller than standard
- Prokaryote Horizontal Gene Transfer (HGT, absorbing DNA from environment) served same purpose as sexual reproduction
  - model consistent with recent results showing HGT common
- High mating radius sensitivity

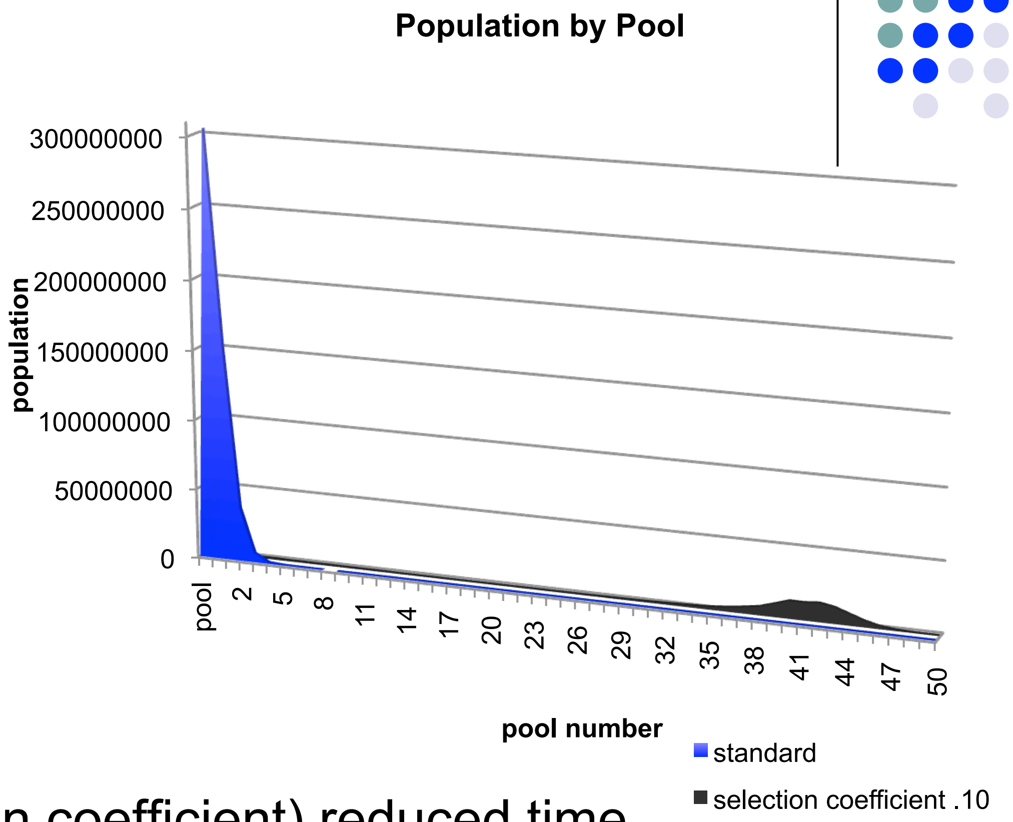
Top 4 population evolution parameters



- Large reductions in carrying capacity increased time by a similar magnitude
- Large increases had modest effect

Top 4 population evolution parameters





- A very high fitness (selection coefficient) reduced time substantially
- It reduces the population of early pools, increasing that of later pools (show model runs)
- Fitness is the only one of the four parameters that asymmetrically favors progress

Fundamental population evolution

- Mutation$_+$s and mutation$_-$s occurred resulting in offspring in higher or lower pools, respectively
- Sexual reproduction produces zygotes in broader pool range than parents; mating radius limited lower-number pool offspring despite higher population
- Increased fitness (selection coefficient) slowed growth of, and ultimately reduced population of, lower-numbered pools; this resulted in increased population of higher-numbered pools
- By limiting how rapidly population pools could grow, carrying capacity slowed evolution to rates we observe in nature

Small population property

- When population << carrying capacity, any sequence produced in time linear to length, independent of other parameters
- This is the case when an individual microbe mutates to have antibiotic resistance
- While conferring advantage, resistance also carries fitness cost, mitigated by subsequent evolution; speculate this is due to small population property

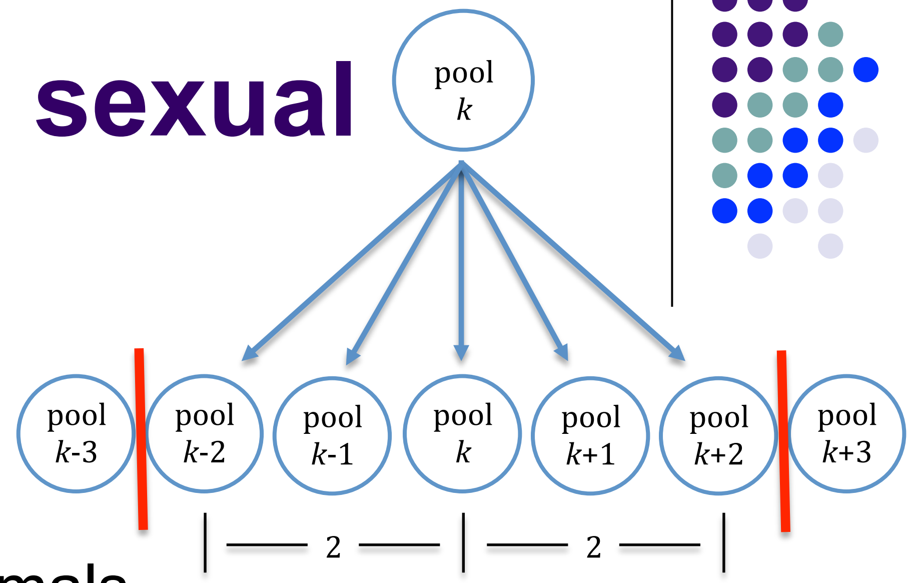
Fitness

- Fitness is only parameter that is not symmetric
  - selection coefficient > 0 benefits higher-numbered pools
- Fitness effect not required for expected evolution duration
  - mean selection coefficient = 0 is sufficient
- Large fitness effect substantially reduces evolution time

Speciation ratchet

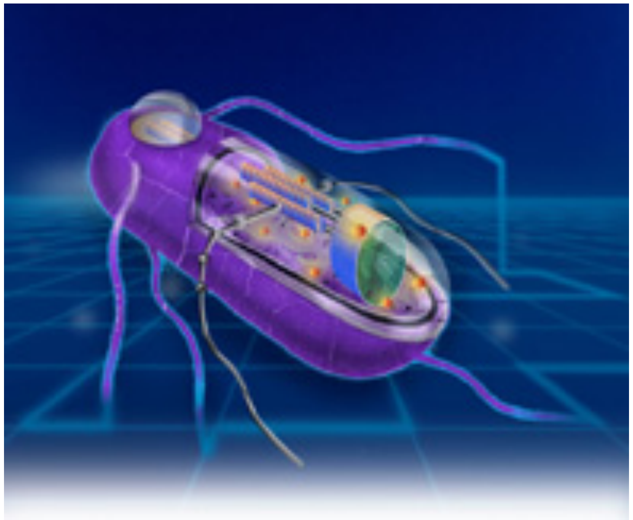
- Speciation prevents mutation$_+$s from regression due to sexual reproduction
- Individuals in new species can’t mate with lower-numbered pools as they are different species
- Does not prevent regression due to mutation$_-$s

Mating radius and sexual attraction



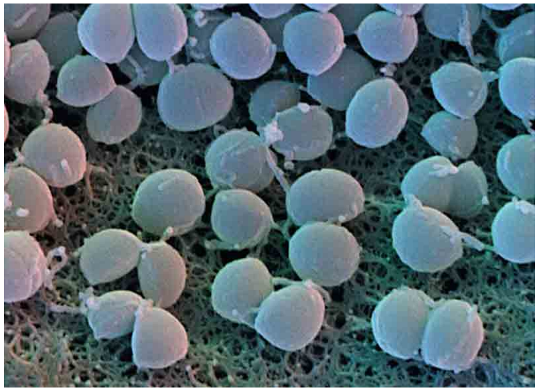
- Radius limited by:
  - must be same species
  - low natal dispersion for mammals
- Sexual attraction may serve to limit mating radius
  - not too different (must be same species)
  - not too similar (otherwise subject to inbreeding issues)
  - Mating with an individual from similar pool provides these characteristics
- Speculation: advantage of limited mating radius partial cause of some human biases such as xenophobia

Application: Synthetic Biology



- Create synthetic organisms with valuable properties, e.g. produce biofuel
- Stability requirement
- Can predict time to loss of property using sequence and population model
- Initial recommendations for high stability:
  - make valuable property resistant to SNPs
  - preclude horizontal gene transfer

Application: pathogen evolution



- Pathogens evolve resistance to drugs (or vaccines)
- Using protein structural prediction or empirical data, determine what pathogen mutation(s) confer resistance to a drug
- Using sequence and population models, predict expected time to resistance emergence
- Use models to determine means to postpone resistance

Future work

- In vivo: determine carrying capacity, fitness, and mating radius values in nature
- In vitro: measure more mutation values, especially inversion rates and lengths
- In silico:
  - complete LUCA and other reference species genome reconstructions
  - apply sequence evolution model to entire reference species genomes
  - confirm or refute universal source sequence hypothesis
  - implement fully multithreaded population model and run it on long model sequence lengths, simulating long periods between speciation events
  - model complete LUCA to Homo sapiens evolution
  - determine heterozygosity effects during population evolution

Thanks for the support

- Thesis committee:
- Prof. Dave Patterson, advisor
- Prof. Adam Arkin
- Prof. Brent Mishler
- Prof. Christos Papadimitriou

- SRI people:
- Steven Eker
- Ashish Gehani
- Merrill Knapp
- Keith Laderoute
- Pat Lincoln
- Ken Nitz
- Carolyn Talcott
- Al Valdes
- And many others who provided advice or processor cycles

- Organizations:
- SRI International
- UC Berkeley
- ONR
- DARPA
- NSF
- NIH
- Sun Microsystems