# Making Sex Tractable: an evolutionary model that satisfies detailed balance

Jüri Lember : University of Tartu

Jenny Poulton, Chris Watkins : Royal Holloway, University of London

# Evolutionary models and genetic algorithms

Nature's no. 1 algorithm : evolution with asexual reproduction

**Nature's no. 2 algorithm : evolution with recombination**

Many population genetic models (Moran Process, Wright-Fisher process)

Hundreds of 'genetic algorithms'

These models yield Markov chains of populations – but these Markov chains are typically hard to analyse.  Stationary distributions hard to obtain.

# Aim: **tractable** and **implementable** model of evolution with mutation, recombination, and selection, for complex genomes and arbitrary fitness landscapes.

We consider evolution as an aimless and endless process.

Under constant conditions (and constant population size) it yields a Markov chain of populations.

We assume **finite set of possible genomes.**

Markov chain is irreducible, and therefore has a stationary distribution.

We would like to derive what the stationary distribution is, for general fitness landscapes.

# Dirichlet-categorical process is a Moran process with mutation

$$DC(x_1, \ldots, x_N; \alpha_0, \alpha_1) = \frac{\alpha_0(\alpha_0+1)\cdots(\alpha_0+n_0-1)\alpha_1(\alpha_1+1)\cdots(\alpha_1+n_1-1)}{\alpha(\alpha+1)\cdots(\alpha+N-1)}$$

where $\alpha_0, \alpha_1 > 0$ are concentration parameters, $\alpha = \alpha_0 + \alpha_1$, and $x_1, \ldots, x_N$ are exchangeable $\{0, 1\}$ variates.

This distribution is exchangeable (by inspection), and can be thought of as an urn-sampling distribution, with a base distribution defined by $\alpha_0, \alpha_1$.

When sampling from the urn, we either copy a randomly selected existing element (from $x_1, \ldots, x_N$, with probability $\frac{N}{\alpha+N}$, or else sample a 'mutation' from the base distribution with probability $\frac{\alpha}{\alpha+N}$

# Parametrising: 3 parameters for a ‘genetic algorithm’

3 parameters: $N$, $\alpha$, and $\beta$

$\alpha = \alpha_0 + \alpha_1$ determines ‘mutation rate’ $u$:

$$u = \frac{\alpha}{N + \alpha} \qquad \alpha = N \frac{u}{1 - u}$$

Note that for constant $u$, $\alpha \propto N$

Define $f(x) \equiv \exp(-\beta\phi(x))$

$\beta$ is a *fitness scaling* parameter: in a genetic algorithm, we choose fitness, unlike in statistics where likelihood is given.

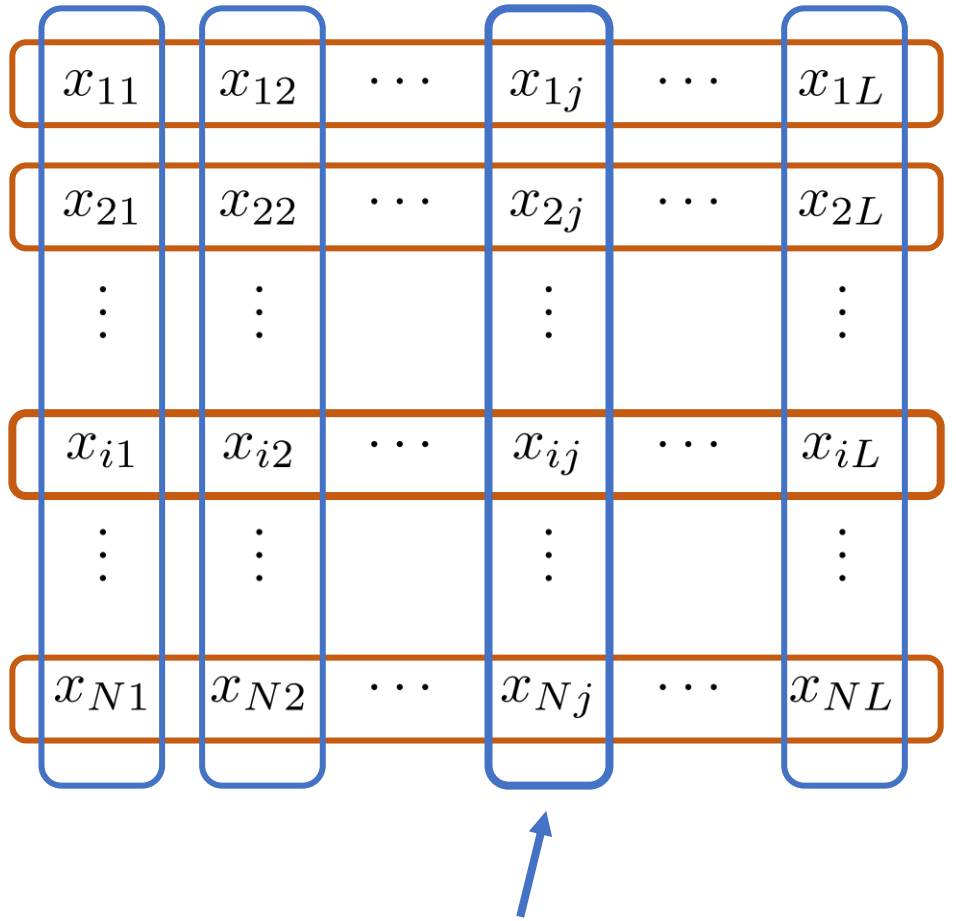
Loose physics analogy: $\phi$ corresponds to energy; $\beta$ is inverse temperature; and $P_\xi$ determines the numerosity of states

# Stationary distribution as ‘crossbar’ factor graph: product of Dirichlet-Categorical processes, with selection

Mutation-selection equilibrium (stationary)
joint distribution of population is a factor graph.

Each **row** is a **genome**. The i’th row factor is

$$f(x_{i1},\ldots,x_{iL}) \equiv \exp(-\beta\phi(x_{i1},\ldots,x_{iL}))$$




Column factors specify breeding.
Row factors specify selection.

Columns are made mutually dependent by fitness;
rows are made mutually dependent by breeding.

Each **column** is a **locus** (or **urn**). The j’th column factor is $\mathrm{DP}(x_{1j},\ldots,x_{Nj};\alpha_0,\alpha_1)$

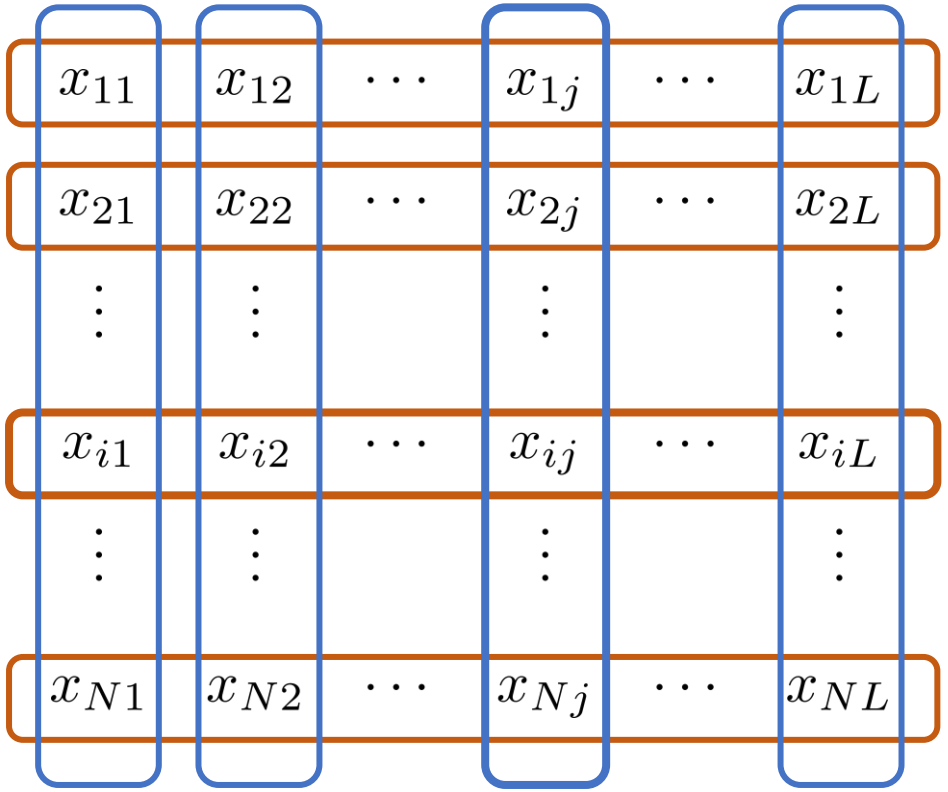
# Stationary distribution looks Bayesian

Let $g_i$ be $i$'th row (genome).
Let $c_j$ be $j$'the column (locus or urn).

$$P_s(\mathbf{X}) = \frac{1}{Z} \prod_{j=1}^{L} P_u(c_j) \prod_{i=1}^{N} f(g_i)$$
$$= \frac{1}{Z} P_u(\mathbf{X}) f(\mathbf{X})$$

where

$$Z = \sum_{\mathbf{X} \in \{0,1\}^{N \times L}} P_u(\mathbf{X}) f(\mathbf{X})$$



# ‘Factorial clustering’ models in statistics

- $N$ items of data $y_1, \ldots, y_N$
- $N \times L$ latent variables $x_{ij}$
- likelihood function $l(y, \mathbf{x})$
- concentration parameters $\alpha_{kj}$ for Dirichlet priors of latent variable

Joint posterior of latent variables and data is:

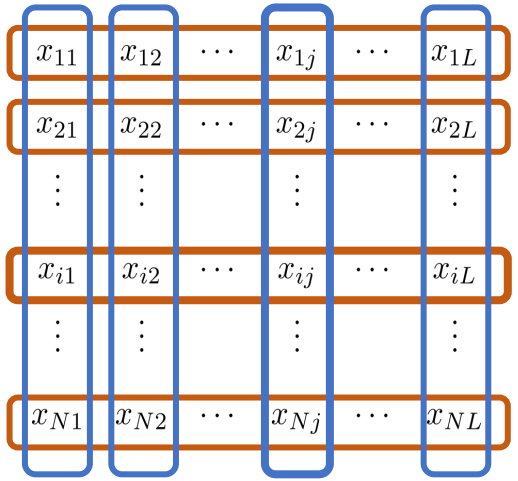
$$P(Y, X) \propto P_u(X; \alpha) \exp(\textstyle\sum_i l(y_i, X_{i:}))$$

Aim is to fit model to data

Fitness is likelihood

Fitness different for each item *i*

Role of Dirichlet process prior is to force latent parameters to be sufficiently similar (few clusters)



# General case: breeding as exchangeable sampling; continuous-time Markov chain where fitness is expected lifetime

Let $\xi$ be an exchangeable discrete-valued process

Genomes $g_1, g_2, \ldots \sim \xi$

Expected lifetime of $g_j$ is $f(g_j)$

While a genome is alive, it contributes to breeding.

$\nu_j$ is death-time of $g_j$

When a genome is 'born', its fitness and lifetime are computed, and its death is scheduled in a priority queue.

**Algorithm 1**

1. Given: genomes $g_1, \ldots, g_N$
2. For $1 \leq i \leq N$, sample $\nu_i \sim \text{Exponential}(\frac{1}{f(g_i)})$; that is, $Pr(\nu_i = t) = \frac{1}{f(g_i)} \exp(\frac{-t}{f(g_i)})$
3. For $m = 1$ to $\infty$ do:
   (a) Let $j = \text{argmin}(\nu_1, \ldots, \nu_N)$ and let $\tau_m := \nu_j$
   (b) Reject $g_j$. ($g_j$ dies...)
   (c) Resample $g_j \sim P_\xi(\cdot \mid g_1, \ldots, g_{j-1}, g_{j+1}, \ldots, g_N)$
   (d) Sample $lifetime_j \sim \text{Exponential}(\frac{1}{f(g_j)})$
   (e) Set $\nu_j := \tau_m + lifetime_j$

# Stationary distribution of algorithm 1

The continuous time Markov chain of populations induced by Algorithm 1 satisfies detaile balance and has stationary distribution:

$$P_s(g_1, \dots g_N) = \frac{1}{Z_{N,\beta}} P_\xi(g_1, \dots, g_N) f(g_1) \cdots f(g_N)$$
$$= \frac{1}{Z_{N,\beta}} P_\xi(g_1, \dots, g_N) \exp\left(-\beta \sum_{i=1}^{N} \phi(g_i)\right)$$

where

$$Z_N = \sum_{g_1, \dots, g_N} P_\xi(g_1, \dots, g_N) \exp\left(-\beta \sum_{i=1}^{N} \phi(g_i)\right)$$

# How well does row-resampling in a ‘crossbar’ factor graph model evolution?

Simplifications in the model:

1. Selection by **viability** not fecundity. Fitness proportional to expected lifetime.
2. **N-way recombination**: each ‘child’ is created with genetic material sampled uniformly from N ‘parents’ instead of from 2 parents
3. Linkage equilibrium in breeding (but not in stationary distribution)
4. **Mutation as sampling from base distribution** ( in DP, mutant does not depend on existing alleles, but in hierarchical DP there can be dependency between mutants)
5. **Haploid** genomes (but can exchangeably sample pairs of haploid genomes...)

Many other details of biological reproduction are of course omitted...

# The modelling dichotomy

Natural recombination and selection is complex
Propose a simplified model

If it succeeds in explaining some phenomenon....we have an abstract insight

If it fails in explaining some phenomenon....then we know that some additional mechanism – that we ignored in the model – is important

# Questions to address?

‘Crossbar’ models are inapplicable to (most) population-genetic questions ...

... but could be used to investigate effects of different fitness landscapes?

Explicit form of stationary distribution allows investigation of stationary distributions for different fitness landscapes.

# Infinite population limits via de Finetti representation

Since $\xi$ exchangeable, by de Finetti's thm, for some measure $\pi$, parameter $\theta$,

$$P_\xi(X) = \int P(X \mid \theta)\pi(\theta)d\theta$$

For the product of Dirichlet categorical processes, we have $\theta = (\theta_1, \ldots, \theta_L)$, where each $\theta_i$ is a Bernoulli parameter.

In the 'biological' limit, where $N \to \infty$ with $u$ and $\beta$ held constant, $\alpha_{lk} \propto N$, so that the prior $\pi$ becomes highly concentrated.

The larger population supplies more 'data'.

In [1] it is shown that under general conditions, the limiting posterior distribution of $\theta$ becomes a delta function.

# Infinite population limit via de Finetti representation

In [1] it is shown that the limiting optimal $\theta^*$ can be found by optimising the expected fitness of genomes generated by a product of binomial distributions, together with a penalty term:

$$\theta^* = \text{argmax}_\theta \; E[\exp(-\beta\phi(\mathbf{x}))] \prod_{l=1}^{L} \prod_{k=1}^{2} \theta_{lk}^{\alpha_{lk}}$$

where:

$\mathbf{x} = (x_1, \ldots, x_L)$, a vector of Bernoulli variates

$\theta_{l0}$ is the probability that $x_l = 0$

and $\alpha_{l0} = \frac{u_0}{1-u}$ is the Dirichlet parameter for the 0 value of the $l$th locus.

# Finite populations : case study of 3 fitness functions

Perfect fitness: only the perfect genome $11\ldots1$ is fit

Fragile: a single bad 0 allele cancels the benefit of all the other 1s.

$$\phi(g) = \begin{cases} 0 & \text{if } g_1 = \cdots = g_L = 1 \\ 1 & \text{otherwise} \end{cases} \qquad \text{Examples : } \phi(11111) = 0 \qquad \phi(111010) = 1$$

Prefix fitness: fitness improves with length of prefix of 1s

Some robustness added:
a single 'path to perfection'

$$\phi(g) = 1 - \frac{1}{L}\max\{i \mid 1 = g_1 = \cdots = g_i\} \quad \text{Examples : } \phi(01111) = 1 \qquad \phi(11101) = \frac{2}{5}$$

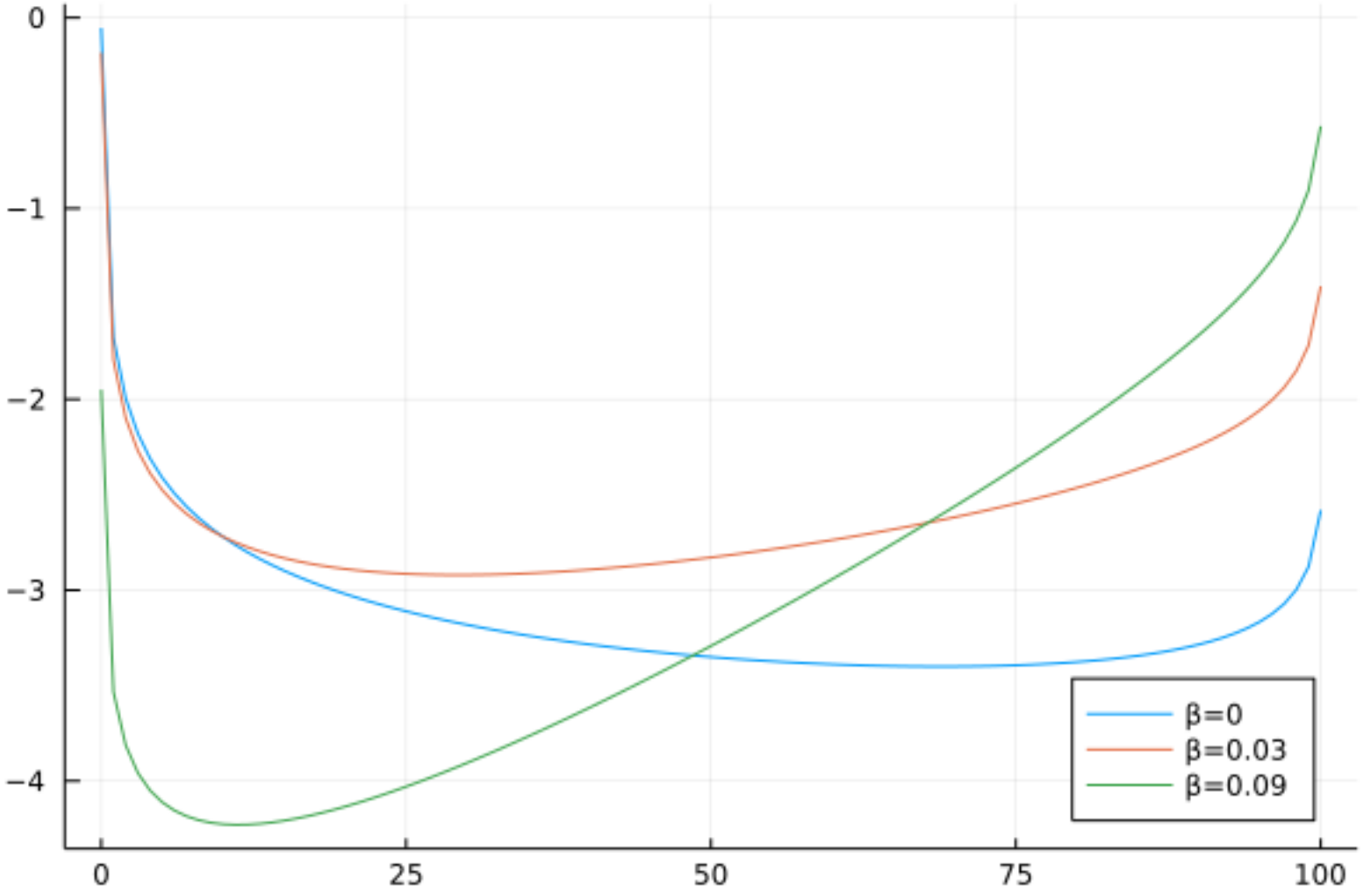
Independent fitness: all good alleles (1) contribute independently to fitness

Robust: each locus contributes independently to fitness

$$\phi(g) = \frac{1}{L}\sum_i (1 - g_i) \qquad \text{Examples : } \phi(01101) = \frac{2}{5} \qquad \phi(11101) = \frac{1}{5}$$

All 3 fitness functions are scaled so that fittest genome has $\phi = 0$, so that $f = 1$, and unfittest genome has $\phi = 1$, so that $f = e^{-\beta}$

# Perfect fitness: distribution of fraction of perfect genomes




Log10 probability

Number of perfect genomes in population of 100

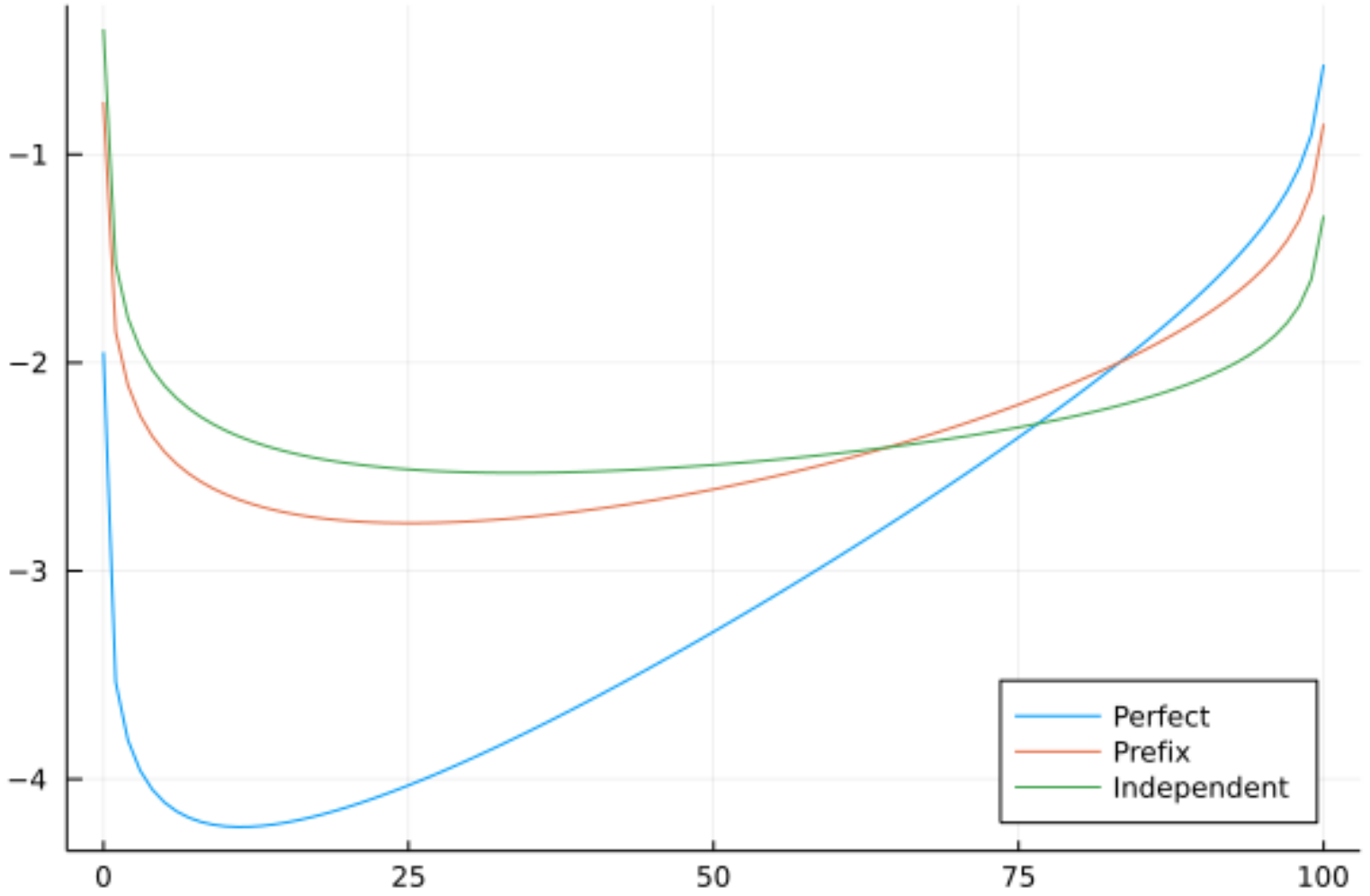
L = 5

N = 100
u = 0.002

Fitness of perfect genome = 1

Fitness of imperfect = exp( -beta )

# Distribution of number of perfect genomes for three different fitness functions, comparably scaled




L = 5

N = 100
u = 0.002

Fitness of perfect genome = 1

Fitness of imperfect = exp( -0.09 )

# Summary

- Suggest a class of evolutionary models that satisfy detailed balance,
  and for which stationary distribution is a ‘crossbar’ factor graph for arbitrary fitness landscapes

- Calculation of stationary distribution for finite and infinite populations, by elementary methods

- Case study of exact finite population stationary distributions for 3 fitness landscapes

- This ‘genetic algorithm’ is actually Metropolis-Hastings with proposals via Gibbs sampling. So.....how do we explain the creativity of natural evolution?

# References

[1] “An evolutionary model that satisfied detailed balance”, J. Lember and C. Watkins,

Methodology and Computing in Applied Probability, 2020

https://arxiv.org/pdf/1902.10834.pdf

[2] “Sex as Gibbs Sampling: a probability model of evolution”, C. Watkins and

Y. Buttkewitz, https://arxiv.org/pdf/1402.2704