Mathematical Modelling of reproduction and evolution

Saurabh & Burhan

Boltzmann Lectures
Horizon Physics and Astronomy Club
Indian Institute of Technology
Madras

11-10-2020

Overview of the talk

- A simplified model of fitness function.
- Evolution of the above fitness function with time in cases of asexual and sexual reproduction, finding timescales, steady states and rates.
- Maximum tolerable mutation rate and largest genome size given a fixed mutation rate, cost explorability and cost redundancy tradeoffs
- Rate of Information acquisition and its connection to fitness, comparison of information acquisition for sexual and asexual reproduction.
- Maxwell's demon, Landauer's exorcism, and second law of thermodynamics. Finding why evolution happens from an information perspective.
- Overview and summary of the model. Validity(and generality) of assumptions, implications and examples in real life. Conclusion and further directions.

Reproduction and Variation: An essential feature of all Organisms

- An essential feature of all the living organisms.

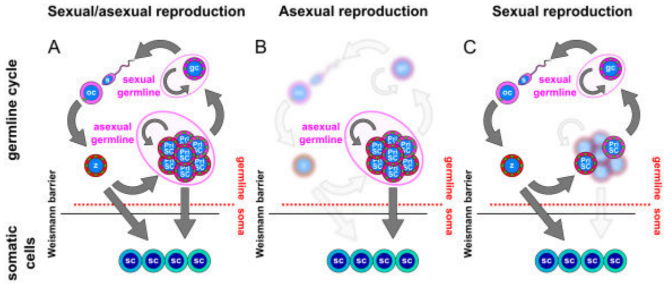
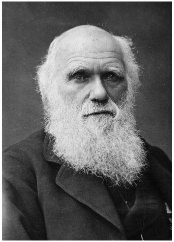


Figure 1: Modes of reproduction
Adopted from (Solana, 2013). [1]

[1] Jordi Solana, 'Closing the circle of germline and stem cells: The Primordial Stem Cell Hypothesis', January 2013, EvoDevo 4(1):2

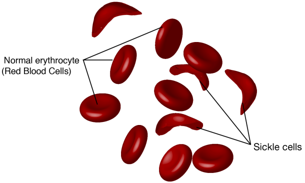
Evolution and Natural Selection

- Slight variations among the individuals.
- Competition for finite resources
- Individuals most suited to their environment survive and reproduce. These species will gradually evolve.

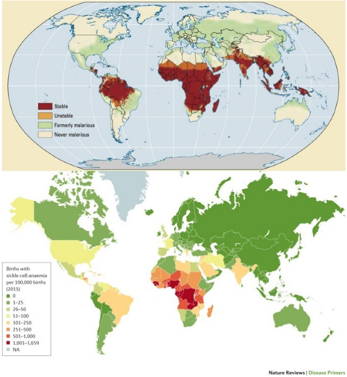


Survival of the Fittest?

- A case of Sickle Cell Disease



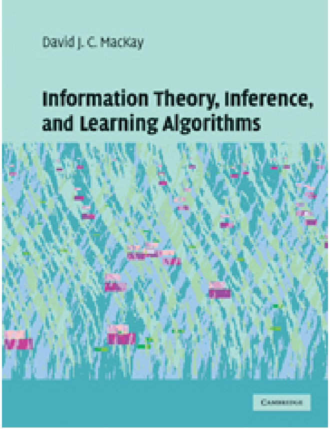
(a) Sickle cell disease
Image Courtesy: NIH



(b) Top: Malaria prevalence (From Wikipedia)
Bottom: Sickle cell prevalence From Ochocinski et al [a]

[a] - Ochocinski et al, "Life-Threatening Infectious Complications in Sickle Cell Disease: A Concise Narrative Review", Front. Pediatr., February 2020

The Model



(a) Information Theory, Inference, and Learning Algorithms

(b) Prof. David J. C. Mackay

The Model

- Genotype of each individual is a vector **x** of G bits. $x_g = 1$ for a good or fit state and $x_g = 0$ for a bad or unfit state.

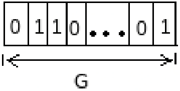


Figure 4: Example of **x**

- Fitness of an individual is given by

$$F(x) = \Sigma_{g=1}^{G} x_g \tag{1}$$

- The bits in the genome could be considered to be either genes with fit and unfit alleles, or nucleotides of the genome.

Theory of mutations

- We assume that each individual gives rise to two progeny, the parent dies and each bit in the genome of the progeny flips with probability of mutations equal to m. Then the top 50 percentile are chosen as the new parents and the process continues.
- We work with normalised fitness f $= \frac{F(x)}{G}$, with f $\in (0,1]$. Since one can achieve f=0.5 by simple random choice, we work with excess fitness $\delta f = f - 0.5$.
- If the parent generation has a mean excess fitness of δf, the mean excess fitness and variance of progeny is given by:

$$\bar{\delta} f = (1-2m)\delta f$$

$$\sigma^2 = \frac{m(1-m)}{G} \approx \frac{m}{G}$$

- If parents have an excess fitness with mean $\delta f(t)$ and variance $\beta \frac{m}{G}$, the selected progeny will have:

$$\delta f(t+1) = (1-2m)\delta f(t) + \alpha\sqrt{1+\beta}\sqrt{\frac{m}{G}}$$

$$\sigma^2(t+1) = \gamma(1+\beta)\frac{m}{G}$$

Theory of mutation....2

- We assume variance is in dynamics equilibrium with $\sigma^2(t+1) = \sigma^2(t)$.
- Subsituting α, β, γ for gaussian distribution, we get the following differential equation for mean excess fitness evolution:

$$\frac{d\delta f}{dt} \approx -2m\delta f + \frac{\sqrt{m}}{G}.$$

- The solution for the derivative is an exponential. We have:

$$\delta f(t) = \frac{1}{2\sqrt{mG}}(1 + ce^{-2mt}) \tag{2}$$

- Steady state excess fitness is given by:

$$\delta f = \frac{1}{2}(1 + \frac{1}{\sqrt{mG}})$$

Theory of combination

- We choose a large population size. Divide into 2 pairs randomly. Each pair gives rise to four offsprings, and the couple dies. Now top 50 percent of the progeny are selected so that population again returns to original value and the process continues.
- We assume homogeneity, ie fraction of bits that are in the good state are same.
- The expected value of the child's fitness will be the same as their parents fitness.
- Standard deviation can be modelled after sum of Bernoulli variables which is $\sqrt{\frac{f(1-f)}{G}}$.
- After selection the mean moves to the right by a distnce proportional to stdev. For a standard gaussian, the proportionality constant is $\sqrt{\frac{2}{\pi+2}}$.
- In differential equation form, it can be written as:

$$\dot{f} = \eta\sqrt{\frac{f(1-f)}{G}}$$

Solution for this equation is:

$$f(t) = \frac{1}{2}(1 + sin(\frac{\eta(t+c)}{\sqrt{G}}), t + c \in (\frac{-\pi\sqrt{G}}{2\eta}, \frac{\pi\sqrt{G}}{2\eta}, c = sin^{-1}(2f(0) - 1). \quad (3)$$

- Thus the system reaches optimal fitness in roughly $\frac{\pi\sqrt{G}}{\eta}$ generations.

Maximal tolerable mutation rate

- When both the models of variation are combined, what is the maximum mutation rate tolerated by a species that undergoes sexual reproduction?
- $$\frac{df}{dt} \approx -2m\delta f + \eta\sqrt{2}\sqrt{\frac{m + (1-f)f/2}{G}}$$
- for positive slope of fitness, $m < \eta\sqrt{\frac{f(1-f)}{G}}$ for sexual reproduction and $m < \frac{1}{G}\frac{1}{(2\delta f)^2}$ for purely asexual reproduction.
- The maximum tolerable mutation rate in sexual reproduction is of order $\sqrt{G}$ times greater.
- If mutation flips on an average mG bits, the probability that no bits are flipped in one genome is roughly e^{-mG}.
- The size of asexually reproducing species has to be of the order of e^{mG}, for a species to persist.
- Max. tolerable mutation rate for asexual species is close to $\frac{1}{G}$
  Max. tolerable mutation rate for sexually reproducing species is close to $\frac{1}{\sqrt{G}}$

Basics of Information Theory

- Here we introduce the concepts of information theory needed for this session
- Shannon entropy for a rV x with N probabilistic outcomes, with the i^{th} event having a probability p_i is given by:

$$S = -\sum_{i=1}^{N} p_i log_2 p_i = E[log_2 \frac{1}{p(x)}]$$

- If all outcomes are equiprobable, entropy is simply $log_2 N$. This is the number of bits of information required to specify the state. One can also interpret this as the complexity corresponding to binary search, since we have knowledge of what outcomes occur and can arrange them and search.
- If some outcomes are more probable, naturally we will use a modified binary search. The expression can be derived by the following: Make blocks for all i in the set, with length Mp_i where M is a large number.
- The number of questions one would have to ask to determine an outcome should be $logM$. But wait! We are doing extra counting, since there are multiple points in each block(corresponds to an outcome) and we do not need to ask more questions once inside the block(of length Mp_i. We fall inside a block i with probability p_i.

Basics of Information Theory....2

- Hence the expected number of questions one would ask would be given by:

$$logM - \sum p_i logMp_i = \sum -p_i logp_i$$

- Entropy is a measure of uncertainty of a system. Information is gained when uncertainty decreases.
- After an outcome(which has a probability of p) occurs, we gain $log(\frac{1}{p})$ bits of information. A fair coin toss resulting in heads means we gain 1 bit. If an event is highly likely(unlikely) to happen, we gain little(huge) information after its occurence
- Entropy which is the weighted sum of these is the information **required** to specify the outcome of a random variable.

Rate of Information Acquisition

- If the normalised fitness is 1, that would mean that the organism has figured which of the two states x_g is better suited, and we have acquired G bits of information. If it is half, which is the most random case, it has not acquired anything more than the case of random case.
- Ergodicity assumption: We assume that given a normalised fitness f, each $x_g, g \in G$ takes the value of the better suited bit with probability f. Obviously there can be some bits more necessary for survival and these could take the better value with more probability and others with less, but this is a case of lesser uncertainty. Our assumption involves lesser information.
- We define the information acquired wrt to fully random case to be:

$$I = \sum_g log_2 \frac{f_g}{\frac{1}{2}} = \sum_g log_2 2 - log_2 \frac{1}{f_g}$$

Plugging in ergodicity, we have $f_g = f, \forall g \in G$

$$I = Glog_2 \frac{2F}{G}$$

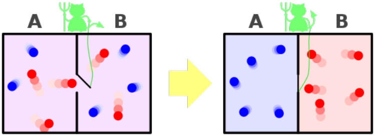
Differentiating with time, we get

$$\boxed{\dot{I} = \frac{G\dot{F}}{F}} \qquad (4)$$

Rate of Information Acquisition

- From the previous equation we see that, increase in fitness with time is accompanied by an acquisition of information with time.
- Expanding around f = 0.5, rate of information acquisition is twice that of fitness.
- Plugging in $\frac{dF}{dt}$ from the previous sections we see that organisms participating in sexual reproduction have a rate of fitness change of O($\sqrt{G}$) while asexually reproducing species have O(1). In terms of information, asexually reproducing species acquire around 1 bit per generation, while sexually reproducing species acquire $\sqrt{G}$ bits per generation. This falls in line with our intuition of more interaction with the population pool produces fitter organisms
- But now I want you to focus on the equation for information acquisition rate. Notice that we have made very weak assumptions here, and the result is model independent to a large degree.
- This observation that fitness increase and information acquisition accompany each other is **very important**

Maxwell's demon and Landauer's exorcism



- Maxwell formulated a thought experiment. Assume we start with a container of ideal gas separated by a partition. A demon opens a small door to let fast moving molecules pass to one side and slow moving to another side.
- Hence one chamber warms up and other cools down, and entropy is decreased.(Perpetual machine?)
- Landauer debunked this argument by the following argument: After getting a result of measurement, the demon has to erase the information in his scratchpad, increasing entropy by an amount at least equal to the decrease in entropy by separating. Even if the demon doesn't erase it straight away, Bennett showed that however well prepared the demon might be, he will run out of storage and has to erase information eventually.

SLOT and Information perspective of reproduction

- Erasing randomness is hard. Resetting an N bit register dissipates N kT ln 2 amount of heat(reversibly). Connection to computers: Modern transistor based computers dissipate around 500 kT heat. A reversible computer(e.g. quantum computer) dissipates kT ln 2 heat, but dissipating more heat is the cost you pay for reliable computers.
- The second law of thermodynamics, cast in information theoretic form says that any isolated system cannot gain information with time.
- Now coming back to our model. Since fitness increase and information acquisition go hand in hand, what has happened here that allowed fitness to increase. What is the other system that has lost information?
- We assumed that only the fittest 50 percentile survive to take part in the next stage. The bottom half died out and their bits became more random(loss of information).
- Death is not necessary. One could think of the environment as a Maxwell's demon, which selects the fast moving particles(fitter population) and separates them.
- If we had not separated the unfit population, these would have bred with the fit population and over time, we would not have gotten any increase in average fitness. Removal of these from the breeding process by some method is necessary for evolution, simply because we have to lose information of some subset to get an information increase, and hence fitness increase by the previous equation.

The Cost of Males

- Why have male offsprings if a female can produce offsprings parthenogentically?

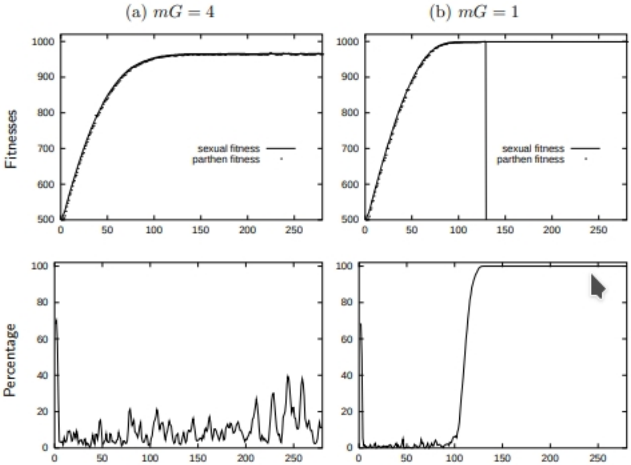


Figure 5: Horizontal axis is time. Vertical axis show fitnesses of the two subpopulations and the percentage of population that is parthenogenetic.
Results when there is a gene for parthenogenesis, and no interbreeding, and single mothers produce as many children as sexual couples

The Cost of Males

- During the 'learning' phase of evolution, the fitness increases rapidly, pockets of parthenogens appear briefly, but then disappear within a couple of generations as their sexual counterparts overtake them in fitness.
- Once the population reaches its top fitness, the parthenogens can take over if the mutation rate is low.
- However, in an unstable environment or in species with high mutation rates, parthenogens always lag behind the sexually reproducing communities. This is consistent with the argument of Haldane and Hamilton of **Kin selection**

Additive fitness function

- Is it reasonable to model fitness as a sum of independent terms? What are the assumptions?
- According to Maynard Smith, the more good genes you have, the higher you come in the pecking order.
- Real fitness function might involve multiple interactions of the individual genes (bits). In this case, a crossover might reduce the average fitness.
- From our model we can predict that, evolution will have favoured species that used a representation of the genome that corresponds to a fitness funtion that has only weak interactions.

Thank you