## Evolution through programming

This is the 1st time we are doing this course. In this course we will try to explore the relationship between evolution and programming.

Types of programming tasks:

1. Simulations
2. Data related

The regular structure of the course is going to be on all lectures except the 1st and the last – today will be different.

**Nothing in biology makes sense – unless in the light of the theory of evolution.**

We have several topics in this course – they appear in the syllabus. Each week we will do programming homework upcoming to the class, we will discuss it in the 1st half of the class and then there will be a new theoretical topic.

We are expected to submit the programming task, in each class only 2 teams present (2-3 people in each team). We must decide about the order of presentations.

**Topics in the course:**

1. Today – talk about evolution, source&rate of mutations.
2. Fitness landscape and fitness scape.
3. Evolution of cooperation.
4. Neutral selection
5. The genetic code (codons) – why was this code selected
6. Regulation of gene expression – data analysis and programming
7. The notion of phylogeny – phylogenetic trees etc. try to reconstruct models of common ancestors. Over the course of 100 million years that mammals exist, we still have 80-90% similarity between mammals.
8. The evolutionary processes of cancer, B cells, T cells – evolution within the body
9. Genetic algorithms – a subbranch in computer science that tries to solve everyday problems by representing possible solutions as genes and applying the laws of evolution on them.
10. Not everything that evolves is biological – talking about evolution of language, ideas etc. talk about evolution of culture.

Today we will start introducing evolution.

## Lecture 1 – 27Mar2025

Although Darwin did not use the terms genotype and phenotype, he had these notions in mind. Note that the environment cares only about the phenotype and not about the genotype – only what is presented happens. Genotypes instruct the phenotype, they can be introduced with mistakes and the selected phenotypes have by chance the instructions to the phenotype that suits the current conditions best.

Prior to Darwin there was another scientist – LaMarck, but he believed that the mechanism of evolution is inheritance of changes that happen by phenotypes only (if you broke the arm, your child's arm would be broken. All giraffs were born with same length of neck, but some were able to extend their necks to get reward – food and passed it on to their progeny). The model of Lamarck is called inheritance of acquired traits.

This is not entirely wrong – epigenetics can be considered soft Lamarckian evolution, and some things such as language are hard- Lamarckian because the changes in language propagate into dictionaries. CRISPR is an example of Lamarckian evolution – the events that happen to the bacterium (infection by a phage) are acquired into its genome as spacers and is inherited to offsprings.

If we move a bit away from Darwinian evolution we get to a problem – according to Darwin mutations happen purely at random, but there is another option – that mutations occur when and where they are needed the most.

Luria & Delbruck experiment – transfer liquid bacteria into petri plate to get colonies. The bacteria grew in good conditions, but the plate has antibiotics, so we expect no bacteria on the plate but more bacteria on a non-antibiotic plate. In real life we do get antibiotic-resistant colonies. We assume that the resistance is a genetic property (although the Lamarckian possibility is interesting). They could have acquired the mutation in the liquid, or only in the presence of the bacteria – these mutations often come with a trade off so are selected against in non-antibiotic conditions. Luria and Delbruck set to answer the question when do mutations occur (randomly/on demand), and this will be the center of our home assignment this week. We should try to solve the problem in the terms that were available at their experimental time, but we can try to implement modern solutions as well.

Note – some resistance mechanisms are drug-specific, some are more broad; for example there is a phenomenon of multi drug resistant bacteria (such as pumps that pump the drugs back out of the cells).

A solution that was used by microbiologists Ester & Joshua Lederberg and started from a plate that has no drugs. They then took 2 or more plates that do have the drug (same drug) and blotted the bacteria to these plates – if the resistance developed on the 1st dish (not on demand) the same colonies should always be resistant (number & location) and if the mutations happen on demand the number & position would be different. This experiment was conducted after the original experiments of Luria & Delbruck and the results are that the mutations indeed happen randomly and not on demand (pure Darwinian evolution). This mechanism is called "replica plating".

Evolvability – the capacity to evolve (the center of research at the Pilpel lab).

The Luria & Delbruck solution – they decided to solve the problem mathematically and then perform an experiment. They used a single drug, single concentration and single timepoint but they repeated the experiment many times. In each of the replications they grew cells in a liquid culture (for ~10 generations), took one plate and plated the cells on it, and counted the number of resistant colonies they got. Therefore, they can check the properties of the distribution of the number of colonies achieved – mean, variance etc. they tried to characterize the distribution expected in both options (mutation in random or on demand). Rise of mutations upon demand will get the Poisson distribution – the probability to get a "saving" mutation at the timeframe before you die, assuming a constant rate of mutations. K is the number of colonies

K is the number of colonies (saving events in the timeframe) by the chances to get an event in such timeframe. We do not know lambda but we know the probability of k=0 (because in some experiments we did not get any colonies) and therefore we can calculate lambda from the equation.

In a Poisson distribution, the mean and the variance is equal to lambda. If the mean is small the variance would be small too, so the distribution would not be wide and will get a relatively strong peak. But they checked and the mean and variance of the actual distribution were not the same, so they came to conclusion that mutations happen randomly and not on demand. The option of random mutations would generate a different distribution, because we choose a number of cells out of the entire population to plate and X% of the population is already resistant, so we get only a percentage of our population resistant. The percentage depends on the time in which we got the mutation, so the earlier the mutation occurs the more resistant cells we would get. The events are dependent on each other, so we get a distribution that allows outliers, the possibility of 1/2 of the cells being resistant is essentially 0. This generates a probability distribution with a very long tail. Luria & Delbruck made a mathematical description of their non-Poisson distribution and when they made the experiment their equation fit the data.

In the assignment we will try to simulate the Luria & Delbruck experiment. The homework would try to deal also with assigning a cost to the mutation (so the growth rate is lower). Assume:

* P = probability of mutation (same or different for getting mutation and losing mutation?)
* H = depth (number of generations)
* R = repeats of experiment (how many times to we run the simulation)
* Simulate what happens in either scenario (mutations can occur at any timepoint or only on the last generation)

We need to convince ourselves that we get a difference between the two modes?

2nd part of the assignment – can we use this to simulate real life processes? Get processes that are both Darwinian and non-Darwinian? Can we run a simulation that has the two options combined and get a distribution that will indicate if we get the 1st option, the 2nd option or the two combined? We have to show that the distributions are distinct from one another.

Note – we are prompted to use our assignments to further explore other questions if they are more interesting to us. (e.g. epigenetic effect, horizontal gene transfer etc.)

Epigenetics – a level of regulation on the expression of genes, that sits on the genome itself and controls the expression. Mutations that are epigenetic in nature are with a memory time sorter than h (the height of the tree), so epigenetic mutations that appear too early in the tree are forgotten before we exam. If K is the epigenetic memory time, only mutations that occur up to K generations ago will be present in the last generation.

## Lecture 2 – 03Apr2025

**Part 1: presentations about the last week's homework**

It looks complicated to distinguish between the random model (pure Darwinian) and the combined model (both random and induced mutations), because the distributions are similar. Today we would sequence everything but back then they did not know what DNA is, let alone sequence. Can we solve this with sequencing? Hypothetically you could sequence the bacterial population right before plating the cells on antibiotics. One problem is that we do not necessarily know the nature of the beneficial mutation. Another problem is that the error rate is higher than the rate of natural mutation, so our measurement is not adequate to answer such question.

**Part 2: basic models for mathematical and computer analysis of evolution.**

In England even prior to Darwin people stated to think of evolution. Martus thought about the mathematical process of an exponential process (replicating bacteria), with a differential equation (see notebook for further notes).

## Lecture 3 – 10Apr2025

**Part 1: presentation of last week's homework**

No one actually signed up to present to we are discussing what we may have done. Interesting choice of 1st presenter (someone volunteered) – instead of sampling from n who dies and who replicates, he sampled from a vector of two (0 or 1) with given probabilities. Shortest number of generations (until fixation) – the number of individuals from the smaller subset.

Principle of evolution – founder effect may lead to survival of fittest out of the existing possibilities, and not the fittest out of the theoretical possibilities (if only blind people are founders of the population we will probably not get survival of those who can see, because there is none of them to begin with. The founders have a higher chance to fixate, not because they are the best but because they are the founders.

**Part 2: evolution of cooperation**

Evolution can be survival of the fittest, the luckiest, those who happened to be first, etc. sometimes, those who survive are those who are the flattest in the original fitness landscape – where all mutations would not create a disadvantage so it cannot die by a random mutation. Today we will talk about the survival of those who know how to cooperate.

We see in nature that once organisms live next to one another, they are willing to share nutrients (microbial cooperation), use a cooperative life style and in humans, even sacrifice for one another – all of these are inherited traits.

This brings us to a more fundamental notion – frequency dependent fitness effect. Most traits in evolution don't show a frequency dependent effect, but rather have an advantage irrespective of how many others have that trait. Example: dark skin can provide protection from the sun even if all other individuals are white. But other traits have a benefit that is frequency dependent – for example, language is only beneficial if you can talk to other individuals, so it is only beneficial if there are many individuals that share this trait. Other example – if all population gets a vaccine, then not-getting a vaccine has an advantage because you do not have to suffer the side effects, but if the entire population is not vaccinated than the trait of not being vaccinated becomes disadvantageous because you would most likely get sick – so the same trait can be beneficial or harmful, depending on the population level frequency of it.

If we plot benefit against frequency, most traits would be a flat line. If we see a decreasing line (less beneficial if it is more abundant and vice versa) we call it negative frequency-dependent fitness effect (for example – cheating). Other traits have benefits that increases with frequency is positive frequency-dependent fitness effect.

There is a question in evolution – how do we acquire a beneficial trait that requires two mutations to happen, when the 1st of the two has a harmful effect? This can be also the case in positive frequency-dependent fitness effect, because if the trait is non beneficial in lower frequencies it cannot reach the high frequencies required to be beneficial. Offer about language – it has a benefit in large numbers but can also be beneficial in self-thoughts (describing the world to myself) therefore has some benefit in lower frequencies as well.

Fitness landscape: fitness as a function of all genetic traits. We can theoretically move from lower fitness to higher fitness by mutations. What happens when the landscape changes as a function of the process of fitness optimization? The problem – as the population shifts towards the higher fitness, it lowers that fitness. In this case there is migration between two possible stable equilibriums (steady states) – either all have the trait or none, in each of these cases the fitness fixates on one of the edges. On the other hand, there is a third equilibrium when exactly half of the population have the trait – in that case, and slight change will lead to reaching one of the steady states (all or none), so this is an unstable equilibrium.

Evolutionary game theory: an application of game theory that assigns fitness to choices in game theory. Examples:

* Prisoner's dilemma – if both cooperate there is evolutionary benefit, if both defect there is evolutionary cost for both, if one defects and one cooperates only the cooperating one suffers from a loss. In this game the most rational decision is to defect (Nash equilibrium – no player could gain from changing their own strategy, holding that all other players' strategies are fixed: if everyone defects, changing to cooperation would always cause loss). Points: coop+coop = (1,1), coop+defect=(-5,0), defect+defect=(-3,-3), defect+coop=(0,-5). The rational decision is to defect, because in any scenario of the opponent's choice, defection would give higher payoff than cooperation.
* Repeated prisoner's dilemma: the way to escape the Nash equilibrium is to repeat the dilemma many times with the same opponent – that way, some strategies allow cooperation to be favorable. Getting to know your opponent can cause better strategies that would help you both better. In the single iteration, the rational decision is to defect; but in the repeated version choosing to cooperate might allow payoff improvement.
* Snowdrift game: if two players are driving cars and there is snow that needs to be cleared, and the options are to clear it together (both cooperate), only one clears (one defect and one cooperate) or it is not cleared at all (both defect), the rational thing to do would be to do the opposite of the opponent, because the loss from both defecting is greater than the loss from cooperating alone, and the loss from cooperating together is greater than the loss of defecting alone.
* Stag hunt game: hunt either stug (deer) or bunny. A stug is bigger (more payoff), but you cannot hunt it alone. You can go after the bunny by yourself. If both cooperate you can go after the stug, if both defect they go after the bunny. Best strategy depends on other player (if he cooperates it is best to cooperate, else it is best to defect).
* Harmony game: cooperation wins.

Prisoner's dilemma in a population – in any mixed population, defectors have a higher payoff than cooperators.

There are some ways in which cooperation has a positive selection coefficient (has an evolutionary advantage and can be selected for):

* Direct reciprocity – if I help you, you help me. If we play on a prisoner's dilemma many times, we can develop trust and achieve better payoff. The winning strategy in this case would be tit-for-tat. Requires memory to evolve, along with the ability to remember faces.
* Kin selection – I am willing to help someone who is more like me than someone who is not similar. סלקציה על בסיס קרבה. Favors cooperation if r>C/B, when r=genetic relatedness (for example 50% for parent-offspring, average of 50% for siblings), C=cost, B=benefit.
* Indirect reciprocity – everyone remembers what happened because they are watching the game, so you have gain in fitness and gain in reputation (because people will defect or cooperate based on your history with someone else, which has an impact on your fitness). Requires the evolution of a gossip network. I help you – someone helps me.

For direct reciprocity you need a face, but for indirect reciprocity you need to have a name.

## Lecture 4 – 23Apr2025

Started with the tournament – our code started out with problems, then Itamar fixed it and submitted to the tournament. We came out next to last.

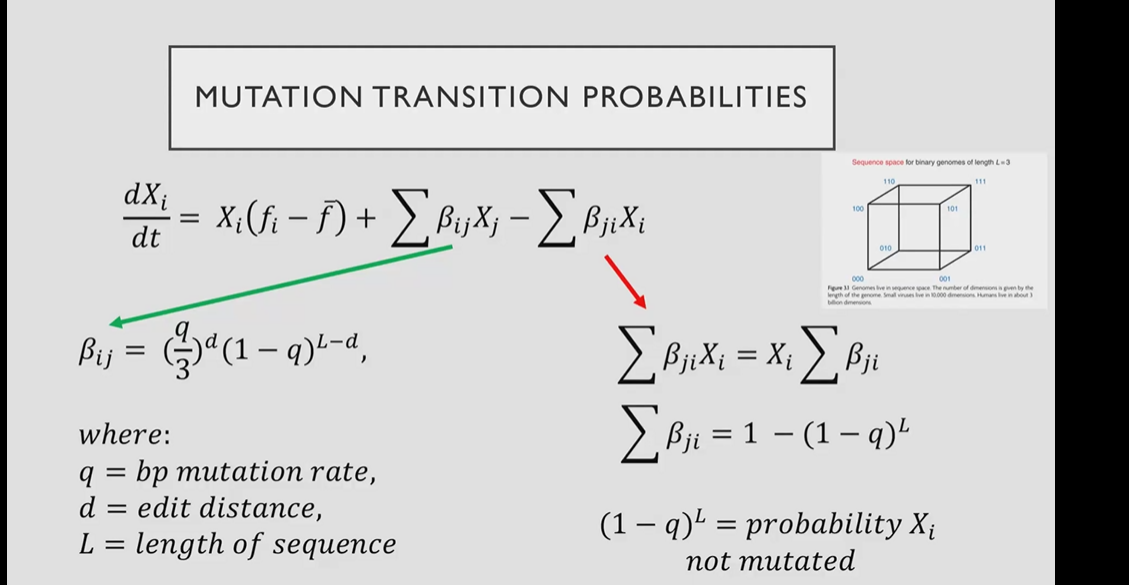
**Fitness landscape in evolution**

The fitness landscape – a metaphor of a landscape: altitude represents fitness, and the spatial location represents the sum of all traits that can contribute to the determination of that fitness.

if we look at a given sequence at length L, the number of sequences that differ from it in only one nucleotide is 3L.

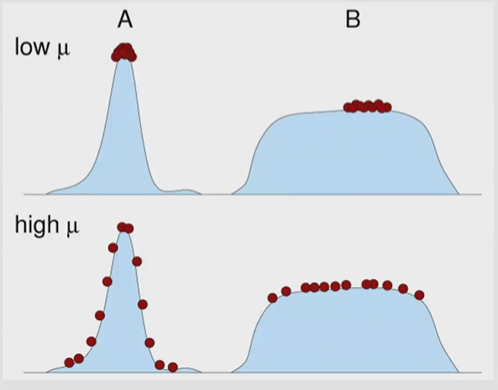
Information content per position – in an alphabet of 2 characters each position only has one piece of information (0 or 1). Therefore, in a binary sequence the length of the sequence is the length of the content of it. In a 4-letter alphabet, each character has 2 bits of information therefore it is more compact. Information theory states that each character in the alphabet has log2(n) when n is the number of characters in the alphabet.

Fitness is a function of the probability of the sequence (combination of traits) to reproduce. You can look at evolution as a random walk in this domain – each mutation can change your fitness, and the goal is to find the highest peak.



Achieving the highest peak is not necessarily possible since to get to it, a population must migrate from a lower peak to the higher peak through an even lower landscape. If mutation rate is very low the population may center around the peak, but if the mutation rate is higher, different variants would be introduced, and the population structure would be more fluid. Usually, mutation rate is considered "safe" if the mutation rate per base is lower than 1/(length of genome). That would keep the average number of mutations per genome replication lower than 1.

Sometimes, having a "flatter" landscape is an advantage, even if the total fitness is lower:



This is called a "quasispecies", because it is a collection of species with not-identical genotypes but a similar phenotype.

Tzahi is not sure that he finished to teach all information needed for the homework, and he will update us.

No lecture next week (יום העצמאות).

## Lecture 5 – 08May2025

There was no homework last week because we did not finish learning the relevant material. Today we will do a recap of last time and also add some more data, and next week will be ordinary back on track (with students presenting their work).

**Fitness landscape** – the relationship between genotype (your traits) and fitness (reproductive rate – can change based on the traits you have). We usually do not get to see the entire number of possibilities in the sequence, but we get some variation that usually originated from the "founder sequence" that arrived by chance.

Evolution is the survival of the quasispecies on the fitness landscape. In this context, flatter regions in the landscape have a higher chance of survival (because mutations would not necessarily cause fitness loss), but this effect depends on mutation rate – higher mutation rate would lead to survival of the flattest, while a lower mutation rate would lead to survival of the fittest.

Actual mutation rate depends on genome size – we want some mutations but in a low rate (compared to the overall genome length), so the longer the genome, the lower the mutation rate should be (to minimize the total number of mutations per genome replication).

Crossing a local minimum (fitness valley) to get to a higher local maximum is a problem in evolution.

Note – fitness landscape varies between different environments – and it is environment-dependent.

**Epistasis** – a phenomenon in genetics in which the effect of a gene mutation is dependent on the presence or absence of mutations in one (or more) other gene, respectively termed "modifier genes". Epistatic mutations have different effects when they occur together or alone.

Epistasis manifests in a non-trivial decline in fitness.

Mutation rates can be inherited epigenetically.

## Lecture 6 – 15May2025

I am making the lecture up from recording. Missed the presentation part because there was no sound in the recording.

We are beginning a chapter of this course that deals with gene expression. Suppose that an organism lives in an environment that presents over time different environmental variables (for example different stresses, we can present them as S1 and S2 to represent different stresses. Different environments may alternate differently between the stresses, but in one environment one stress may be a predicting event for the appearance of the 2nd one. Some organisms may use that phenomenon to use this prediction for their benefit.

Introducing notations:

* S1 and S2 are 2 environmental conditions (for example one is clouds and the 2nd is rain)
* R1 and R2 are cellular responses to these conditions.

Different relationships between S and R:

* The most direct regulation is a direct arrow between Si and Ri.
* Stochastic switching may create a minor effect of R2 in response to S1 (while the main response still being R1) and vice versa. This happens stochasticly by chance.
* Associative learning – while S1 is designated to create R1 it also creates R2 on a lower level as a deterministic choice (not by chance).
* Conditioned regulation – if S1 predicts S2 but S2 does not predict S1, than S1 may activate R1 and R2 but S2 would activate R2 only.

43:00

## Lecture 7: 22May2025

**Evolution of the genetic code and codon usage bias**

Some amino acids are coded by more than one DNA codon, and they are not equally distributed across different genomes. Some of them are more frequent than others, when the identity of the more frequent codon may differ between genomes.

Open questions about the genetic code: why is it that the same aa is coded by a few very similar codons (wobble of the 3rd position)? Why is it in triplets? Why 20 amino acids?

Note – almost all amino acids where the 2nd base is U or C are non-polar amino acids. Codons with the 2nd base being A or G tend to code for polar, acidic or basic amino acids. Looks like also the structure properties of the amino acids seem to follow this order where similar structure is coded by similar codons.

There are a few competing hypotheses for the formation of this pattern:

* Frozen accident
  + Whatever it is now, it happened once and just got stuck and that's it. It happened randomly and got fixed by chance.
  + The code fixed early so any change now is catastrophic (too many mechanisms rely on it)
  + Explains near-university across life
  + Horizontal gene transfer reinforces constraints
  + Predicts limited variation
* Stereochemical affinity
  + For some codons, the amino acids have affinity to the codons that represent them. At that point it can explain the in the beginning we did not need a mediator (RNA as mediator between DNA and protein), and the mediator started only after.
  + Coevolution with biogenesis – new amino acids adopt related codons, so we started with a smaller pool of amino acids and as it expanded we got the new amino acids assigned to codons with similar chemical properties. Predicts orderly expansion of the code.
* Error minimization and robustness
  + The current genetic code minimizes mistakes (mutations), because it allows silent mutations. A single mutation would (in most cases) yield the same amino acid, or a different amino acid that has the same chemical properties (for example, one nonpolar amino acid can change into another nonpolar amino acid).
  + Implies selection but also historical limits
  + Not perfectly optimal but very close to it.

The current genetic code is selected to minimize the impact of mutations on the nutrient budget (oxygen, carbon, nitrogen etc.).

## Lecture 8: 29May2025

Note – the cost of changing codon from nonstop to stop is far greater than changing from stop to nonstop. Many genes have a stop codon coded right after the 1st one (in frame) to make sure that the ribosome indeed stops on time.

**Conditioning and extinction**

sljkl