Verbindende Übung Biowissenschaften

Teil des Moduls "Einführung in die Biowissenschaften"

Sonja J. Prohaska

Computational EvoDevo Group
Universität Leipzig

Leipzig, WS 2025/26

Why is evolution is central to bioinformatics?

- **Explains sequence similarity**

  homology arises from common ancestry; allows meaningful sequence comparison and function prediction
- **Provides statistical models**

  substitution models, phylogenetic inference, and dN/dS ratios quantify evolutionary change
- **Reveals functional importance**

  conserved sequences indicate essential genes, protein domains, and regulatory elements
- **Structures biological knowledge**

  evolutionary concepts like orthology, paralogy, and gene family evolution structure how we compare and interpret sequences
- **Inspires computational methods**

  genetic and evolutionary algorithms mimic variation, selection, and recombination for optimization and modeling

Genotype-to-Phenotype Problem

When it was first proposed that (1) the **genotype** is **inherited** while the phenotype is not and (2) the genotype must hold the **information necessary to build an organism**, this gave rise to the **genotype-to-phenotype problem**.
The challenge is to ...

- understand **how** the genotype gives rise to the phenotype
- disentangle which genes are causally related to which characters/traits
- predict phenotypic changes from genotypic changes and *vice versa*
- intervene with the development of a trait in a controllable manner
- engineer organisms with specific traits

Genotype to Phenotype

Let's assume...

- the **genotype** is the set of heritable genetic units termed "genes"
- a **gene** is an **abstract** entity
- its *instances* are termed "alleles"
- the **phenotype** is the set/sum of observable characteristics
- a **character** (or **phene**) is an **abstract** entity
- its *instances* are termed "traits"

gene	alleles	character	traits
g^1	$\{g_1^1, g_2^2\}$	p^1	$\{p_1^1, p_2^1, p_3^1, p_4^1\}$
g^2	$\{g_1^2, g_2^2\}$	p^2	$\{p_1^2, p_2^2\}$
g^3	$\{g_1^3, g_2^3, g_3^3\}$	p^3	$\{p_1^3\}$
$\vdots$		$\vdots$	
g^n		p^m	

Marked in red and blue are, the **particular** genotype and the **particular** phenotype of a (haploid) organism, respectively.

Caution! The notation does not implied relations between g^x and p^x!

Genotype and Phenotype dissected into Units

- we seem to know how to split up the **genotype** into meaningful **units**,

  i.e. the "genes" g^1 to g^n
- to derive character **units** p^1 to p^m from the **phenotype** is challenging
- which characters turn out to be "observable" is ambiguous

  and depends on the "measurement" method
- example: human "eye color" is a character, its instances are "blue", "brown", "green" a.o.
- frequently, the "gene expression profile", i.e. the amount of mRNA expressed per gene is used as a "molecular phenotype" which is commonly assumed to be predictive of the amount of proteins present or active
- reasons for using the "molecular phenotype":

  – mRNAs are convenient units

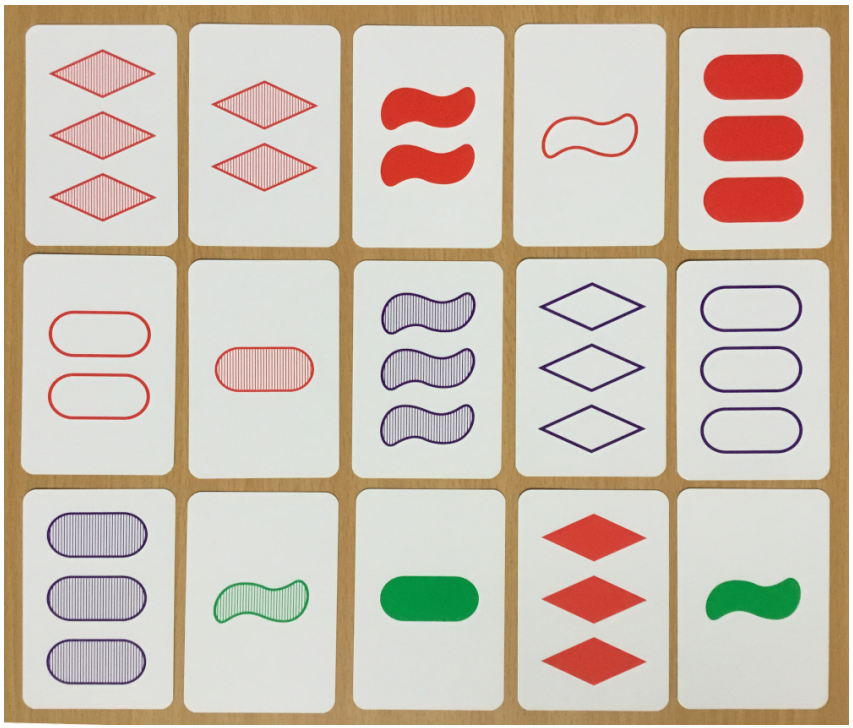
  – the amount of mRNA present is easy to measure

  – the relations between "gene" and "phene", in the sense of "level of gene expression",

  is relatively well known

Exercise: Phenotypic characters

Given is a population with 15 individuals.
Indetify the characters and their respective traits.



Question: Are some traits linked with others?
Do the exercise NOW! Hints to the solution are on the next slide!

Exercise – Hint: Phenotypic characters

Hint: there are four characters, with three traits each.

Question: Are some traits linked with others?

Exercise – Solution: Phenotypic characters

Solution

character	traits		
color	red	green	purple
shape	bar	wave	diamond
fill style	solid	hatched	open
number	one	two	three

Question: Are some traits linked to others?

Answer: Yes (in the given set which is very small).

- All green individual have one symbol. (strong)
- All purple individual have three symbols. (strong)
- Individuals with two symbols are always red. (strong)
- Individuals with diamonds are never green. (weak)
- Individuals with a solid fill style are never purple. (weak)
- Individuals with an open fill style are never green. (weak)
- Individuals with one symbol are never purple. (weak)

Evolution

Evolution give rise to the "unity and diversity" of life.

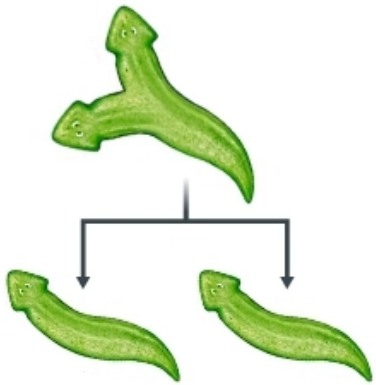
"What are the premises for evolution to happen?"
or *"How does life evolve?"*

- Multiplication
- Variation
- Fitness
- Selection
- Heredity
- (Environment, e.g. limited resources)

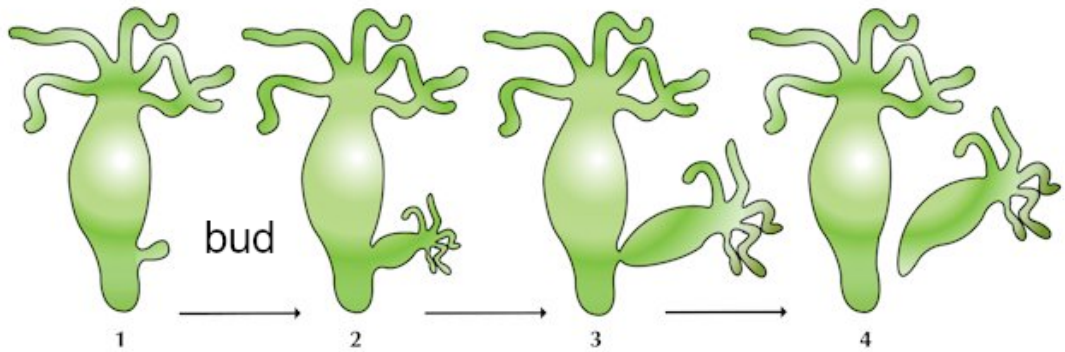
Multiplication

reproduction (in biological terms) is the process by which new individual organisms are produced.

- **asexual reproduction**
  by **binary fission**
  – one "parent" cell produces two "doughter" cells of equal size
  – classic example: bacteria, fission yeast (fungi)
  – example: Planaria (multicellular animals) – *bottom left*
  by **budding**
  – one "parent" cell produces one, smaller "doughter" cells or spore
  –classic example: budding yeast (fungi)
  – z.B. Hydra (multicellular animal) – *bottom right*
- **sexual reproduction**
  by **fusion** of gametes of opposite sex
  only Eukarya have **Mitosis and Miosis** enabling sexual reproduction



planaria



"parental" hydra

"daughter" hydra

Heredity

- **heredity ("Vererbung")** is the passing on of traits from parents to their offspring

  Caution! the passing on is indirect, meaning

  not the traits (phenotype) themselves are passed on

  but rather the genetic material (genotype) causing the appearance of the traits
- inheritance ("das Erben") of traits by other sources than the genetic material,

  e.g. epigenetic modifications, is know
- heritable ("erbbares") materials need to be replicated (at least once) and properly segregated
- the object of heredity is mainly the genotype

Understanding the importance of heredity:
What if offspring did not inherit the genetic material from their parents?

Variation

- in the context of evolution, **variation** refers to **genetic variation**
- genetic variation is the difference in genetic material among individuals (or populations)
- **mutation** is the ultimate source of variation

  it includes, point mutation (i.e. single nucleotide change),

  (nucleotide/sequence) insertion, deletion, translocation, and inversion
- **genetic variation can cause phenotypic variation**
- variation may have deleterious or beneficial phenotypic effects
- however, most mutations are neutral (and show no phenotypic effect)
- **genetic/meiotic recombination**

  is the exchange of genetic material between homologous chromosomes during Meiosis

  and involves **crossovers**
- variation gives rise to genetic variability and diversity (which is usually good)
- (random) heritable/genetic variation occurs at the genotypic level

Understanding the importance of variation:
What if the genetic material would be passed from generation to generation without any variation?

Fitness

- **fitness** is linked to **survival** and

  the survival of an organism's/individual's **genes in the offspring**

  relative to other (directly competing) organisms/individuals
- fitness is **"an individual's ability to propagate its genes"** (Wikipedia)
- evolutionary fitness is not/less about **physical fitness**

  unless, it has a positive effects on the **reproductive fitness**

  example: due to higher physical fitness, a parent can gather more food for the offspring, causing more offspring to reach reproductive age
- fitness is linked to **competition**

  intra-species competition, e.g. for food, maiting partners, a.o.

  inter-species competition, e.g. between predator and prey
- fitness is linked to **adaptation**

  biotic and abiotic environmental changes will

  reduce the reproductive fitness of individuals less well adapted to the situation
- **in modelling**, fitness is often defined as a probability of an individual to contribute its genes to the next generation

Understanding the importance of fitness:
What if all individuals of a population had the exact same fitness?

Selection

- **natural selection** is the differential survival and reproduction of individuals

  due to differences in phenotype (not necessarily genotype)
- in contrast to **artificial selection** (e.g. breeding), it is not intentional
- while selection acts on the phenotypic differences that result in differences in fitness, it is obligatory that the corresponding change is a heritable (genetic) change
- **sexual selection** is a mode of natural selection

  members of one biological sex choose mates of the opposite sex to mate
- sexual selection can lead (mostly) males to extreme efforts to demonstrate their fitness
- this may lead to the evolution of features whose function is actually deleterious to the individual's survival
- selection occurs on the phenotypic level in an environmental context





Evolutionary Theory

The (data) structure created by the following assumption (made by Darwin in md 19th century)

- common descent
- descent with modification
- changes accumulate gradually
- new species arise from pre-existing ones
- a single species can split into two or more new, distinct species
- species can go extinct

gave rise to the proposal of an **hierarchical tree** of life.

Evolution: an Optimization Process?

In a constant environment and over time, mutation and natural selection drive the population towards its **fitness optimum**.

Evolutionary algorithms (EAs) are a type of **optimization algorithm** that mimics biological evolution to find optimal solutions to complex problems.

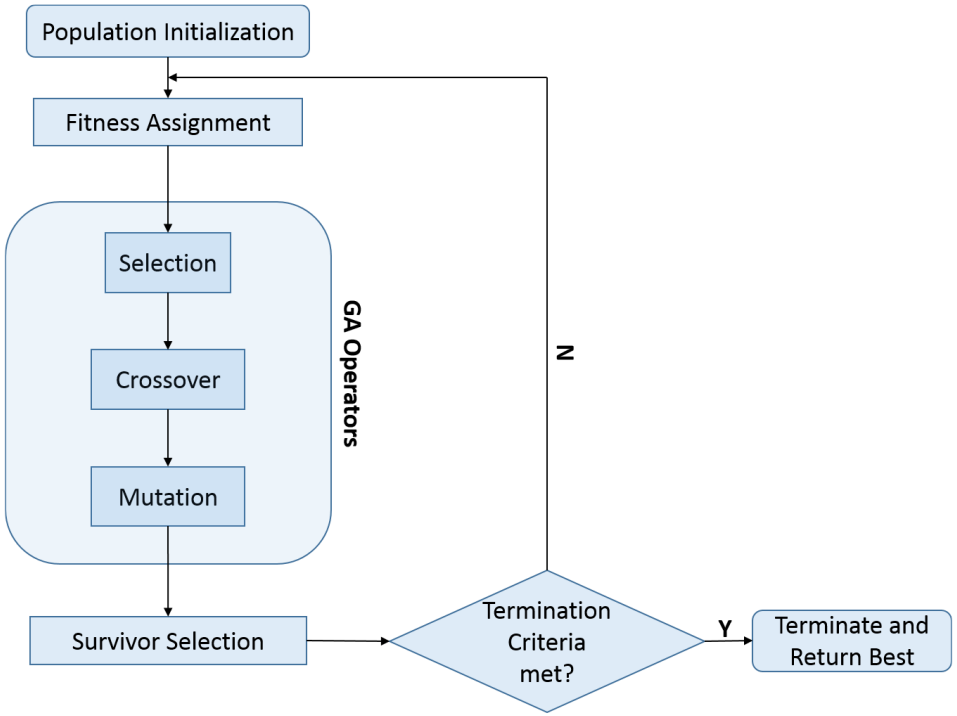
To understand the analogy between evolution and evolutionary algorithms, it is relevant to understand the following processes:

- point mutation
- meiotic recombination and crossover
- random mutation
- natural selection

Evolutionary Algorithms (EA)

Most evolutionary algorithms (EAs) are **metaheuristics** and **nondeterministic**.

- Genetic Algorithms (GAs)
- Evolution Strategies (ESs)
- Differential evolution
- Evolutionary Programming (EP)



Genetic Algorithms

What we need is ...

- **population**: a population A of size N
- **individuals**: individuals of the population, i.e. $a_1 ... a_N$
- **genomes**: each individual a has a genome S of size L

  that represents the solution attributes coded as a **bit string**,

  i.e. $s_1, ... s_L$, where $s_x \in \{0, 1\}$

  Here, for simplicity, attributes are represented by **one** bit*

  such that the contribution of an attribute is either *zero* or *one*
- **fitness** function: each individual has a fitness $f(a)$,

  $F(a) = \sum_{i=1}^{L} s_i w_i$

  where w_i is the **value** of the attribute s_i
- **selection** method: to select from the fittest individuals

  – e.g. preserve the x fittest individuals (i.e. the elite)

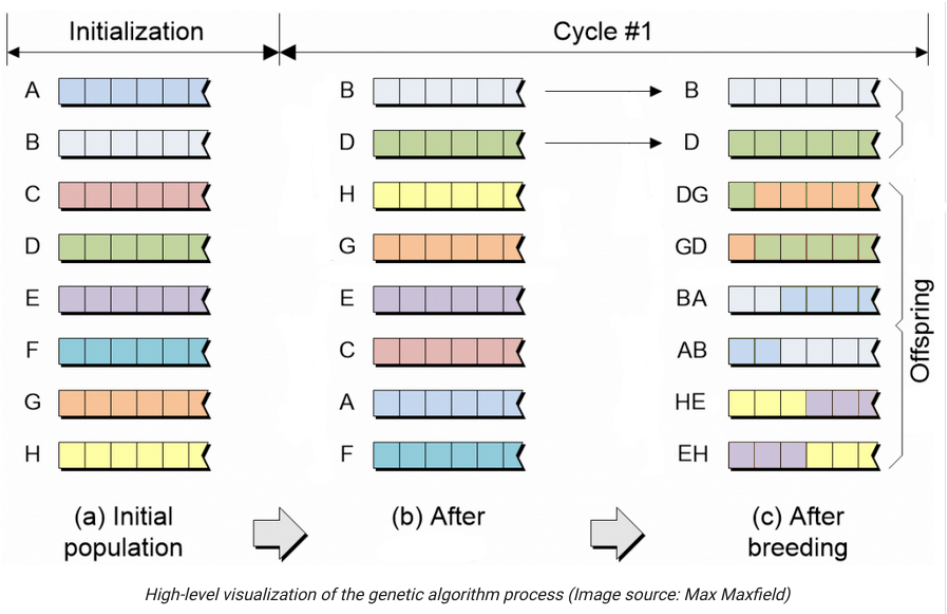
  – e.g. roulette wheel selection: pick individuals with probability p proportional to their fitness
- **mutation**: to introduce variation

  – **crossover** partial strand exchange between two individuals

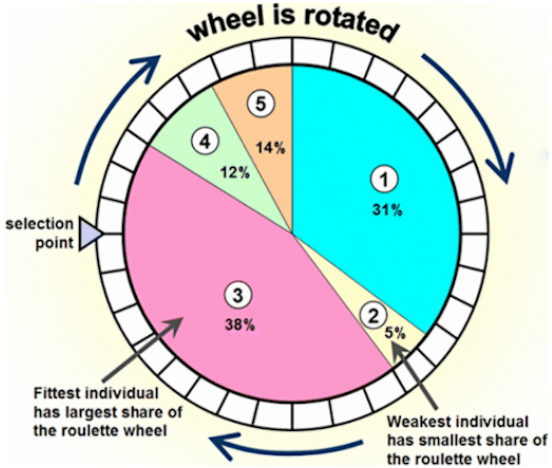
  – **(point) mutation** bit flip
- **termination**: max number of generations, fitness threshold, or fitness saturation

* usually, a decoded **word** of length l is the solution attribute x

Genetic Algorithms - crossover



High-level visualization of the genetic algorithm process (Image source: Max Maxfield)



a) A bis H represent different random genomes of zeros and ones (not shown).

b) After sorting by fitness genomes B and D are top ranking.

c) The top ranking genomes B and D are copied to the next generation.

   Mate pairs are chosen using roulette wheel selection.

   A single crossover is introduced at a random position resulting in chimeric offspring.

- **Roulette wheel** for five genomes (1-5) and their relative fitness.

Notice, the roulette wheel on the right is from a different example than the population(s) on the left.

The Knapsack Problem

The elves are loading Santa's sled. The last bag is allowed to weight only 10 kg. Five gift items are still left. Their total weight is to high to put all of them into the bag.

	1	2	3	4	5	total
weight	5000g	4000g	1200g	700g	600g	11,5kg
value	300	200	150	125	105	880



Santa's request is to maximize the total value without exceeding the weight limit.

How would you solve the problem (without knowing about GA)?

The Knapsack Problem

Show that the following greedy approaches fail to find the optimal solution.

greedy approaches

- approach **A**: sort the item **descending** by weight
- approach **B**: sort the item **ascending** by weight
- go through this list and check
- will the weight limit be exceeded by adding this item?
  - if the anwser is NO, add the item
  - if the answer is YES, pick the next item
- **A** returns the item combination 1, 2, and 4 (weight 9.7kg, value 625)
- **B** returns the item combination 5, 4, 3, and 2 (weight 6.5kg, value 580)
- complexity $O(n)$ (n is the number of items)

optimal solution

- it is easy to show that
  by putting all gifts together and removing either item 1, 2, 3, 4, or 5
  we quickly find the optimal solution:
- the item combination 1, 3, 4, and 5 (weight 7.5kg, value 680)

Applying Genetic Algorithms to the Knapsack Problem

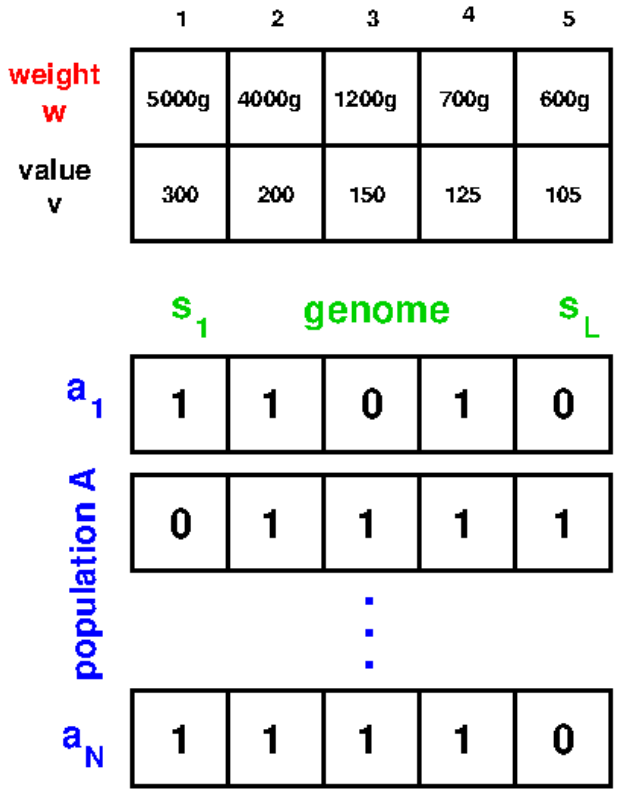
- generate N individuals
- with a random **genome** $\{s_1, ...s_L\}$

  $s_i = 1$, item i is in the bag

  $s_i = 0$, item i is *not* in the bag
- assign **fitness** according to $F(a)$

  takes the weight limit into account by setting

  $F(a) = 0$ if the weight limit is exceeded
- **selection**
- **crossover**
- **mutation**
- loop to "assign **fitness**"
- **termination**



You will need to fiddle with the parameters but it's worth it.
Actually, the problem is a bit too small to value its advantages. Try a larger set of items!

More approaches to the Knapsack Problem

genetic algorithm

- samples the solution space around optima
- non-deterministic
- pro: likely to return the optimum/optima
- pro: complexity $O(gmn)$ with g the number of generations, m the population size, and n the number of items
- con: requires fiddling with encoding and parameters

brute force

- enumerates and evaluates all possible solutions
- deterministic
- pro: will return the optimum/optima
- con: complexity $O(2^n)$ with n being the number of items

dynamic programming

- based on a recurrent formula (e.g. sequence alignment, RNA folding)
- pro: elegant
- pro: complexity $O(Wn)$ with W the weight limit, and n the number of items
- con: takes a lot of memory