第五讲
遗传算法
Lecture 5
Genetic Algorithm

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

  The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

二进制编码
Binary Coding

- 我们已经讲述了十进制到二进制的转换，二进制由于和物理元器件，主要是半导体元器件的稳定的二值状态相对应，对于计算机对数据的处理，包括计算、存储、传输等非常便捷。

  We have introduced the conversion from decimal numbers to binary numbers. Due to the correspondence between binary numerals and the stable binary states among physical especially semiconductor devices, the processing of data by computers, such as computing, storing and transmitting are relatively convenient in the way of binary numbers.

- 在某些算法应用中，如果能恰当地运用二进制编码，也能起到便捷算法实现的目的。下面，我们以遗传算法为例，讲解二进制编码的运用。

  In some applications, the solution to the problem can be simplified especially from the implementation perspective if binary coding is adopted. We demonstrate this by introducing the genetic algorithm.

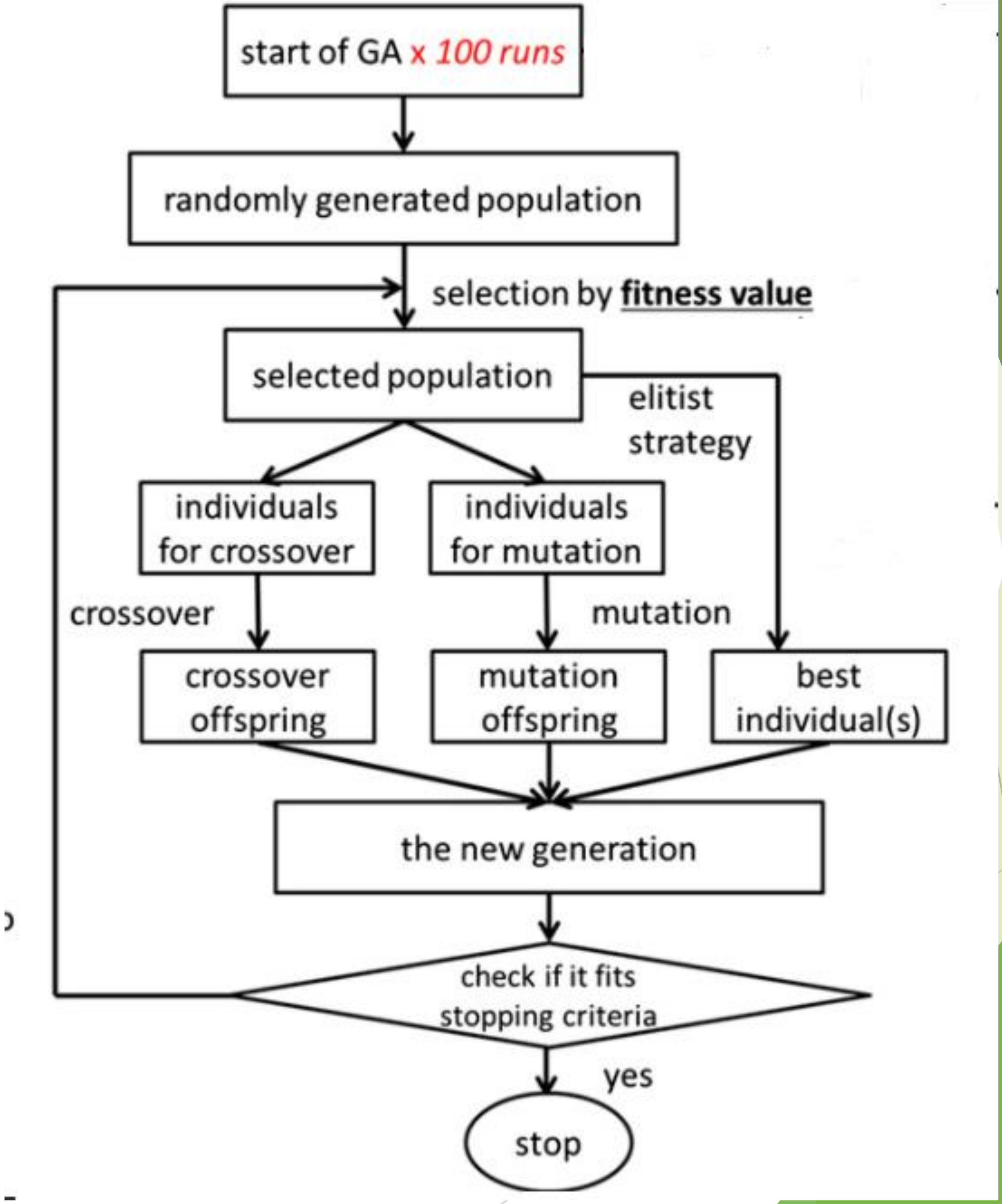
遗传算法
Genetic Algorithm

- 遗传算法是一种以自然选择和遗传理论为基础，将生物进化过程中，种群中个体的适者生存，个体间染色体的交叉，以及个体内基因变异等机制相结合的随机优化算法。其由美国密歇根大学的J. Holland教授于1975年提出。

  Genetic algorithm is based on natural selection and genetic theories. It simulates the mechanisms during the biological evolution, such as the survival of the fittest, the crossover of chromosomes among individuals, and the mutation of genes in individuals, etc. It was proposed by the American Prof. J. Holland in Michigan University in 1975.

遗传算法
Genetic Algorithm

- 遗传算法总体流程（The overall flow-chart of genetic algorithm）。



遗传算法
Genetic Algorithm

- 个体：指带有染色体特征的实体；

  Individual: an entity expressing characteristics of chromosomes

- 种群：个体的集合，该集合内个体数称为种群大小

  Population: A collection of individuals, the number of individuals is called the population size

- 进化：种群逐渐适应生存环境的过程；生物的进化是以种群的形式进行的。

  Evolution: The process of which the populations are gradually adapted to the living environment. The evolutions usually take place in the unit of populations.

- 适应度：度量某个物种对于生存环境的适应程度。

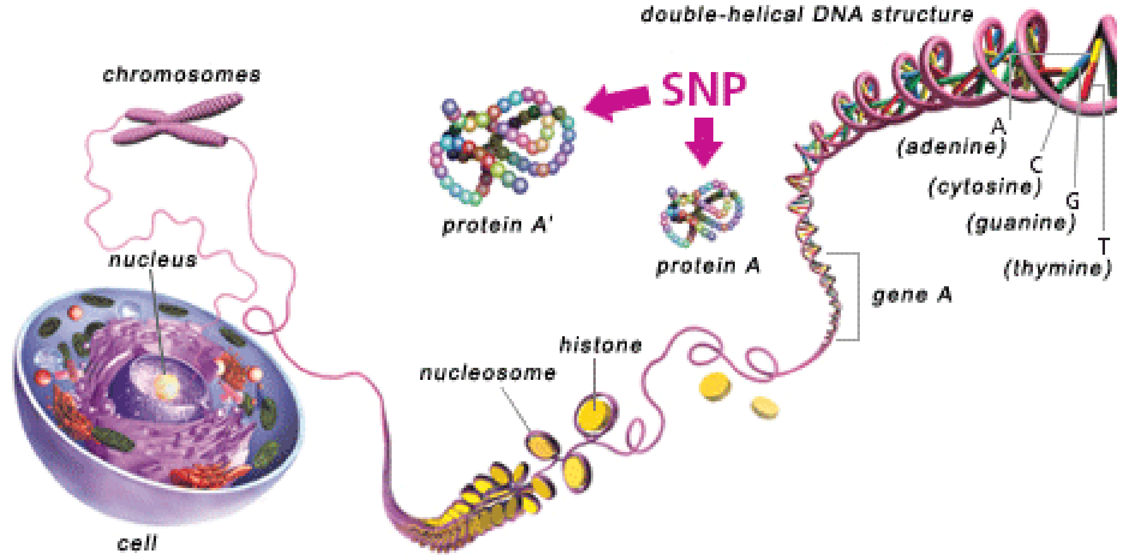
  Fitness: Measurement of the species' adaptation to its living environment.

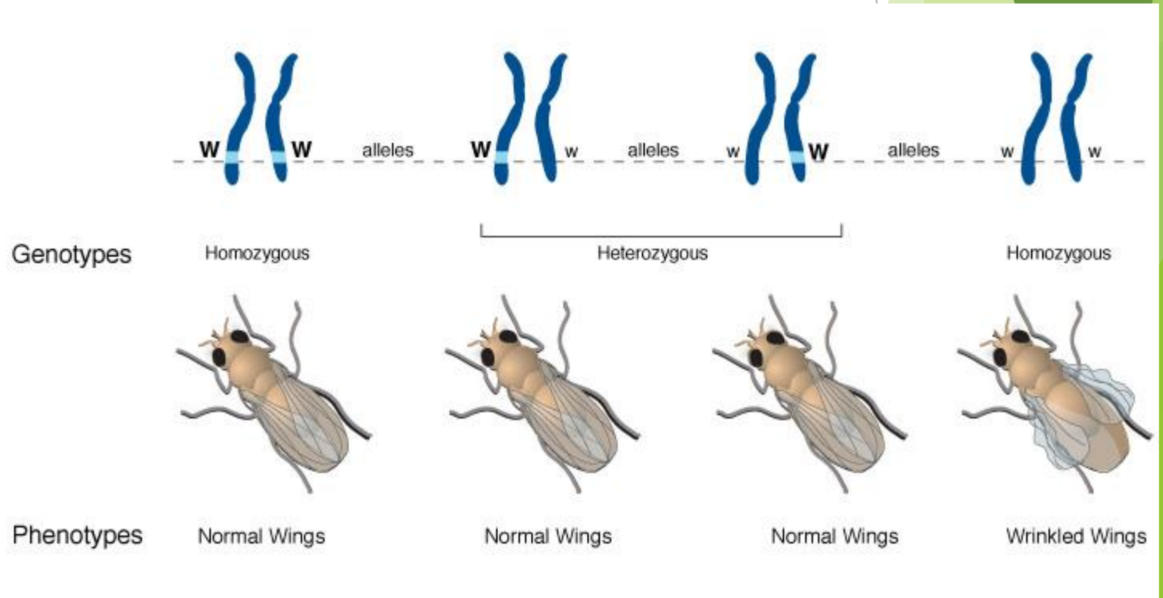
遗传算法
Genetic Algorithm

- 基因型：性状染色体的内部表现

  Genotype: The internal manifestation of the trait chromosome;
- 表现型：染色体决定的性状的外部表现

  Phenotype: External manifestation of the traits determined by the chromosome

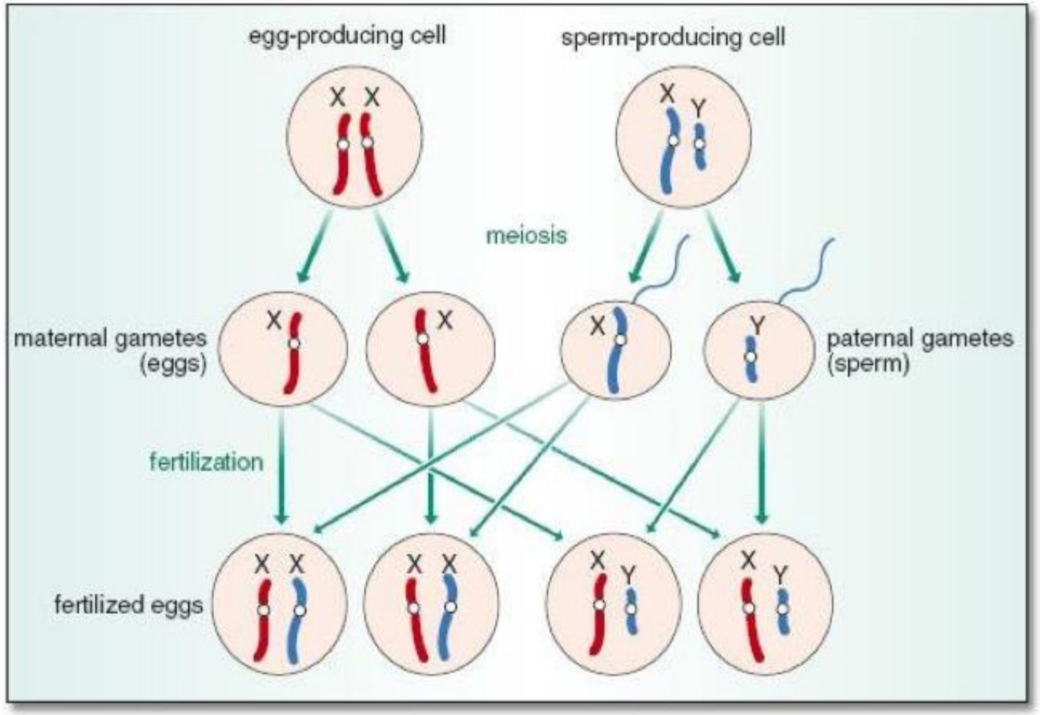




遗传算法
Genetic Algorithm

- 复制：单个亲本的遗传物质通过克隆全数或半数传递给子代的过程。

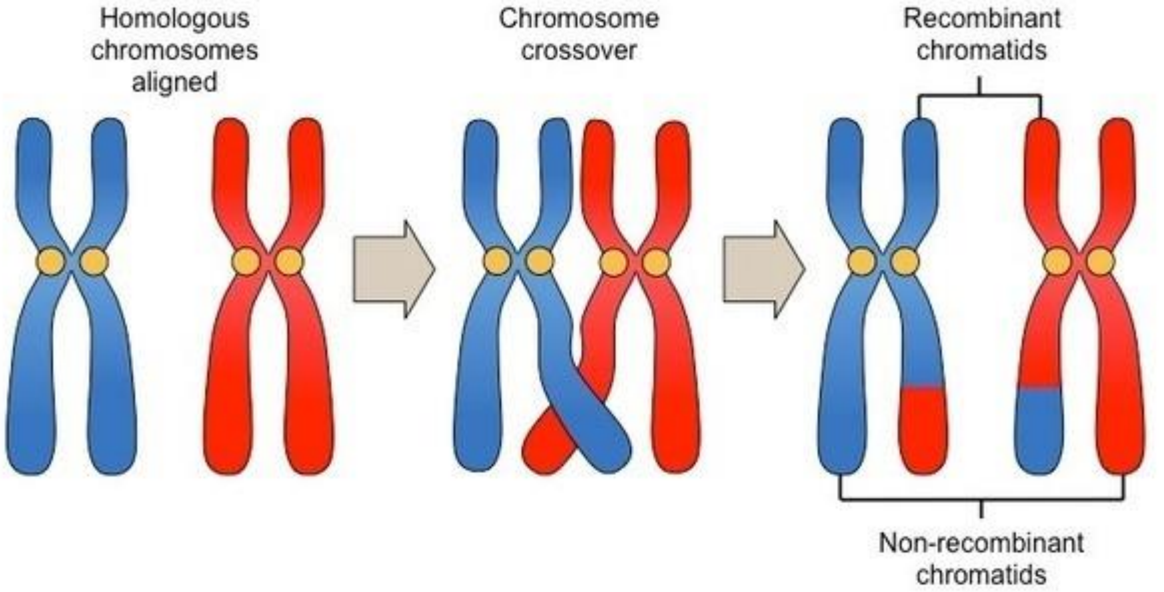
  Reproduction: The process of passing all or half of the genetic materials from the parent to the offspring via cloning of the chromosomes.



遗传算法
Genetic Algorithm

- 交叉：两个染色体在同一位置处的DNA被切断，前后两串分别交叉组合，形成两个新的染色体的过程，也称基因重组或杂交；

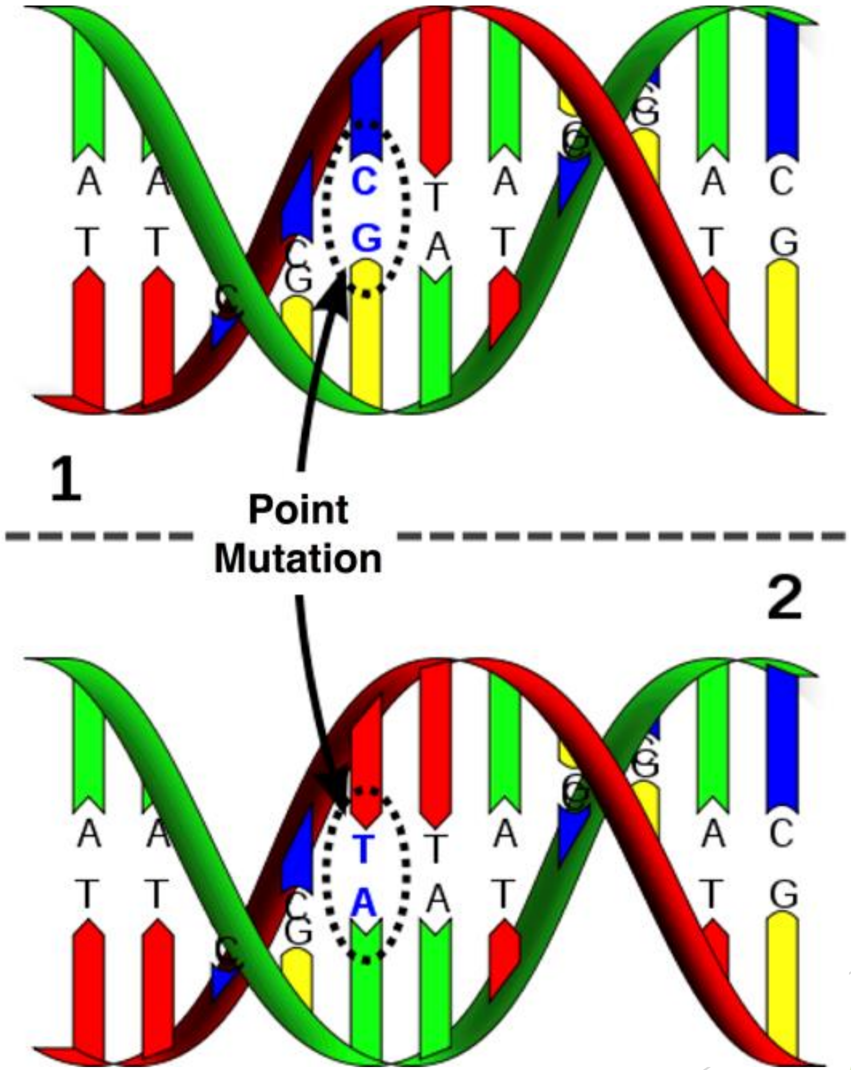
  Crossover: DNA is cut off at one of the same locations of the two chromosomes, and the front and back strings are cross-combined to form two new chromosomes. Also known as gene recombination or hybridization;



遗传算法
Genetic Algorithm

- 变异（或突变）：反映为DNA序列的变化，主要由细胞分裂过程中的DNA复制错误引起，其它原因包括辐射，化学接触等。

  A mutation is a change in a DNA sequence. Mutations are mainly caused by DNA copying mistakes made during cell division. Other factors include radiation, chemical contact, etc.



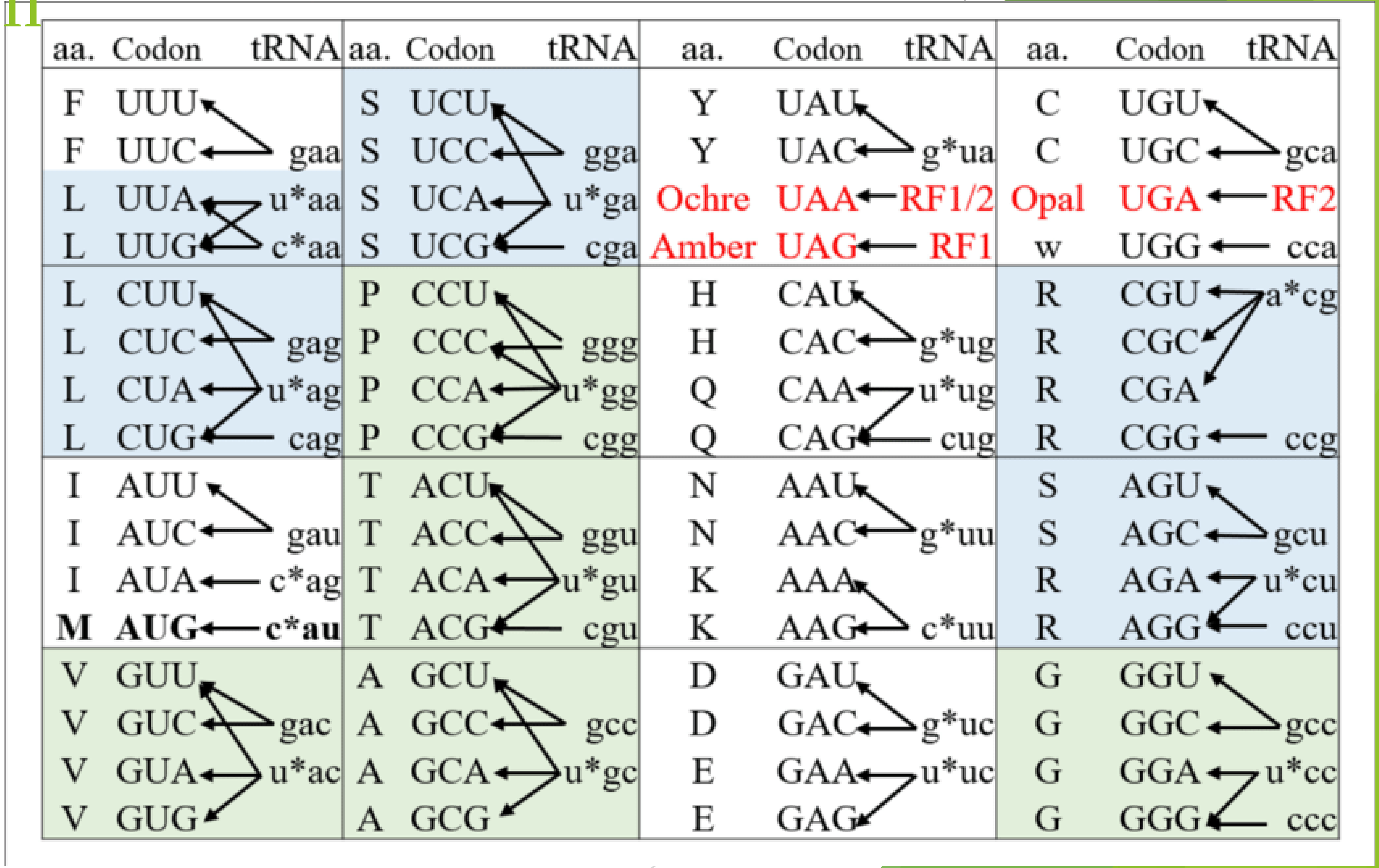
遗传算法
Genetic Algorithm

- 编码：从表现型到基因型的映射。

  Coding: the mapping from the phenotype to genotype.

- 解码：基因型到表现型的映射。

  Decoding: the mapping from the genotype to phenotype.



aa.	Codon	tRNA	aa.	Codon	tRNA	aa.	Codon	tRNA	aa.	Codon	tRNA
F	UUU		S	UCU		Y	UAU		C	UGU	
F	UUC	gaa	S	UCC	gga	Y	UAC	g*ua	C	UGC	gca
L	UUA	u*aa	S	UCA	u*ga	Ochre	UAA	RF1/2	Opal	UGA	RF2
L	UUG	c*aa	S	UCG	cga	Amber	UAG	RF1	w	UGG	cca
L	CUU		P	CCU		H	CAU		R	CGU	a*cg
L	CUC	gag	P	CCC	ggg	H	CAC	g*ug	R	CGC	
L	CUA	u*ag	P	CCA	u*gg	Q	CAA	u*ug	R	CGA	
L	CUG	cag	P	CCG	cgg	Q	CAG	cug	R	CGG	ccg
I	AUU		T	ACU		N	AAU		S	AGU	
I	AUC	gau	T	ACC	ggu	N	AAC	g*uu	S	AGC	gcu
I	AUA	c*ag	T	ACA	u*gu	K	AAA		R	AGA	u*cu
M	**AUG**	**c*au**	T	ACG	cgu	K	AAG	c*uu	R	AGG	ccu
V	GUU		A	GCU		D	GAU		G	GGU	
V	GUC	gac	A	GCC	gcc	D	GAC	g*uc	G	GGC	gcc
V	GUA	u*ac	A	GCA	u*gc	E	GAA	u*uc	G	GGA	u*cc
V	GUG		A	GCG		E	GAG		G	GGG	ccc

遗传算法
Genetic Algorithm

- 我们以求如下函数在给定区间内的最大值问题来说明利用遗传算法解决问题的过程。

  We exemplify the genetic algorithm by solving the problem of finding the maximal value in specific range for the following function:

$$f(x) = x \cdot \sin(10\pi x) + 2\,, x \in [-1, 2]$$

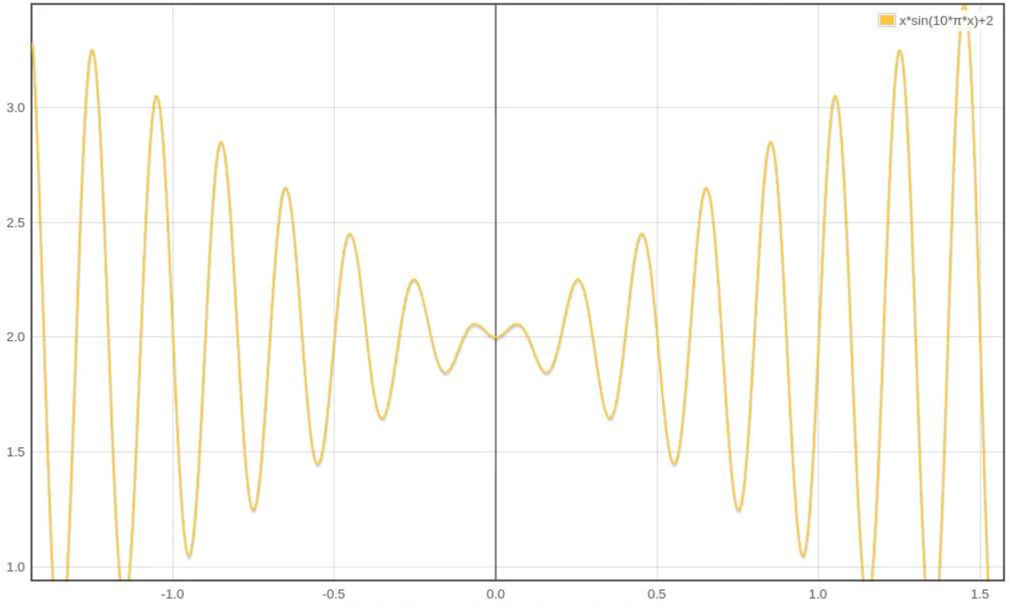
- 我们尽管可以通过求函数的一阶导数$f'(x)$的值为0的解来求极值，但由于$f'(x) = 0$的高度非线性性，这个方程并不容易求解，因此我们考虑用遗传算法直接求解。

  Although it is possible to directly find the maximal value of $f(x)$ by solving the equation $f'(x) = 0$, where $f'(x)$ is the derivative of $f(x)$. However, since $f'(x) = 0$ is highly non-linear, we can apply genetic algorithm to ease the difficulty in solving the equation.

遗传算法
Genetic Algorithm

- 注意到$f(x)$在求解区间内跳变很多，用传统的数值方法如牛顿法，不恰当的初值选取可能将求得局部极大值。

  Notice that $f(x)$ bumps up and down many times during the range of solution, so the traditional methods such as Newton method might be ineffective due to the sensitivity to initial values, which might lead to local optimum.



遗传算法
Genetic Algorithm

- 正如算法流程图所示，遗传算法的第一步是要初始化一个种群，其中每个个体代表问题的一个可行解（未必是最优解）。我们首先要解决的问题便是如何表示这些可行解。

  As indicated in the algorithm flowchart, the first step of genetic algorithm is to initialize a population, in which the individuals represent candidate solutions (might not be optimal solutions at the beginning). The first thing we need to deal with is how to represent these candidate solutions.

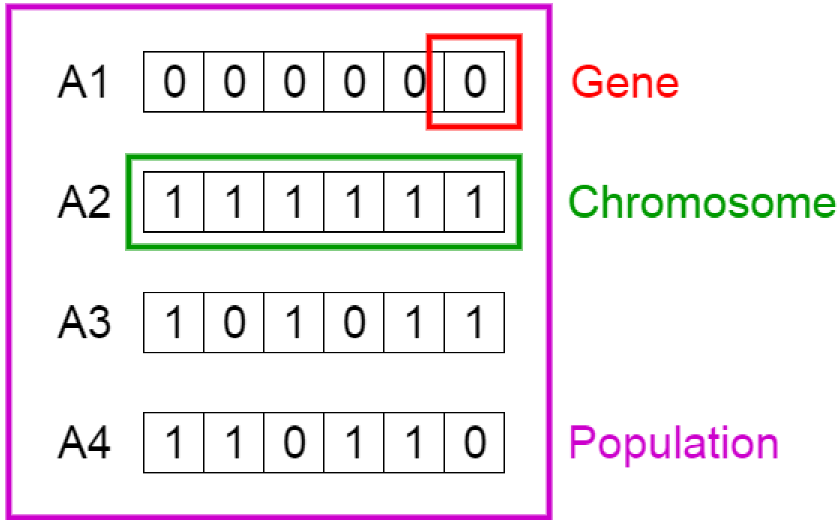
- 从生物学上我们可知，生物的性状由染色体确定。对此问题而言，个体的性状就是一个表示数值的大小的标量。因此，一条染色体足够。对于基因，我们作一个简化，认为单个的位就可以表示基因。

  From biology we know that the traits of individuals are decided by the chromosome. For the exemplified problem, the trait is just a scalar, so one chromosome is enough. For gene, we do simplification and treat each bit as one gene.

遗传算法
Genetic Algorithm

- 在之前的数制转化中，我们着重讲了整数的转化。虽然对一些要求解是整数的问题，可以用二进制直接编码，但这里不行，我们需要借助于前面所讲的基因编码与解码的概念。

  The number system conversions established before just focus on integers. Binary numbers can directly be used for the problems that require integers to be solutions. However, it is not plausible here, the concepts of genetic coding and decoding just discussed need to be leveraged for solution representation.



A1	0	0	0	0	0	0
A2	1	1	1	1	1	1
A3	1	0	1	0	1	1
A4	1	1	0	1	1	0

Gene

Chromosome

Population

遗传算法
Genetic Algorithm

- 首先我们要意识到用计算机的有限位长表示浮点数，绝大多数情况下都是有限精度。这里面有两个概念，即离散与量化。如要用计算机存储一天的温度变化，首先，温度随时间的变化是一个连续值，如果不离散化，时间段可以呈无穷细，计算机可能不能存储，因此，一般1秒记录一次足够。如果不量化，则无法用有限的位数表示实数。

  First, it is needed to realize that to represent float point numbers with limited bits in computer only results in limited precision. There are two concepts in this regard, namely, digitization and quantization. For example, to log the temperature change for one day, since temperature is analogous, it has uncountable values in any tiny time range. Without digitization, it cannot process the case such as acquiring the temperature in one second. Also, without quantization, the computer might not be able to store even one temperature value.

遗传算法
Genetic Algorithm

- 针对上面问题，因为解的区间为$[-1,2]$，长度为3，假设期望精度为10^{-6}，则需要将区间分为3×10^6份。由于$\lceil\log_2 3\times10^6\rceil=22$，因此至少需要22个二进制位来表示一个可行解。因此，编码过程是将一个可行解对应为一个22位的二进制串。而解码过程与之相反，但由于需要将二进制串对应为可行解，因此还需要一个映射到可行解区间的步骤。

  For the problem above, a precision of 10^{-6} demands dividing the range into 3×10^6 equal breadths due to the range $[-1,2]$ of the solution. Since $\lceil\log_2 3\times10^6\rceil=22$, a bit string at least of length 22 is needed to represent a candidate solution. So, the coding process is to mapping a candidate solution to a bit string, the decoding process is to delineate a bit string into a candidate solution.

遗传算法
Genetic Algorithm

- 如上所述，解码过程分为两步：

  As aforementioned, the decoding is a two-step process:

  - 将二进制串代表的数转化为十进制数

    Convert the binary number to decimal number

$$[b_{21} \quad b_{20} \quad \cdots \quad b_0]_2^t = \left(\sum_{i=0}^{21} b_i \cdot 2^i\right)_{10}^t = n_{10}^t$$

  - 将十进制数转化为可行解

    Convert the decimal number to a candidate solution

$$x^t = -1 + n_{10}^t \frac{2-(-1)}{2^{22}-1}$$

遗传算法
Genetic Algorithm

- 适应度函数衡量了种群中的个体对生存环境的适应程度，其以评分的形式呈现，间接地决定了在迭代过程中，特定个体是否产生新个体，或产生新个体的概率。在大多数情况下，评分即目标函数的取值。

  The fitness function determines that how fit an individual adapts to the environment for surviving. It gives each individual a fitness score, which indirectly determine that during the iterations, a given individual is allowed or not for reproduction, or the probability that an individual will be selected for reproduction.

- 在此例中，计算如下函数的值作为适应度即可：

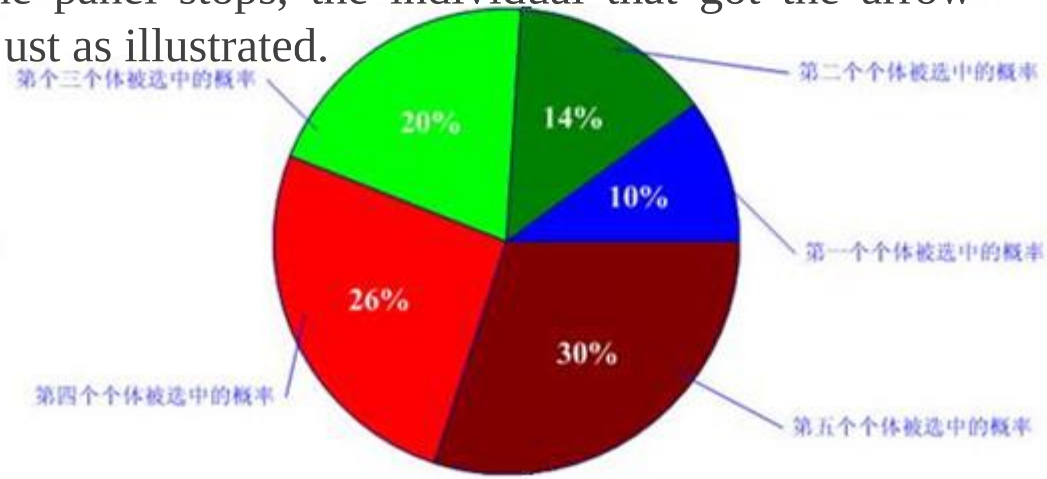
  In this example, we can simplify it as computing the value of the following function:

$$f(x) = x \cdot \sin(10\pi x) + 2$$

遗传算法
Genetic Algorithm

- 当有适应度值之后，可能根据适应度值设计不同的选择算法，如下面所示的轮盘算法。其预处理步骤可以是根据各个适应度值求softmax，然后根据比例画在圆盘上，随机轮转，当轮盘停下来时，指针指向的个体保留，如下图所示：

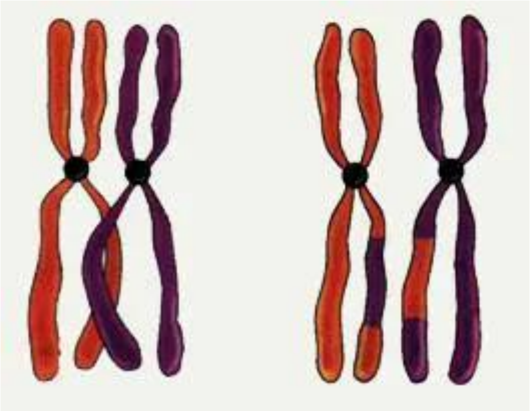
  To select the appropriate individuals for reproduction, different selection algorithms can be designed based on the fitness values, for example, a variant of round-robin algorithm. The softmax can be applied to fitness values as pre-processing step to find the probabilities of each individual’s chance to reproduce. Then these probabilities are drawn on the round panel. When the panel stops, the individual that got the arrow points wins the chance to reproduce, just as illustrated.

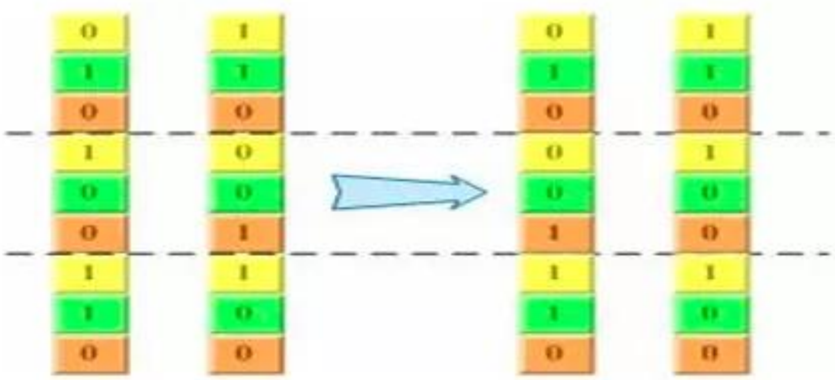


遗传算法
Genetic Algorithm

- 基因重组或交叉的过程，特别是当采用二进制编码时，与高中所学染色体联会过程类似，如下图所示：

  The crossing over process of chromosomes, especially when binary coding is adopted, is analogous to what we’ve learned about the synapsis process of homologous chromosome pairs, as illustrated below:



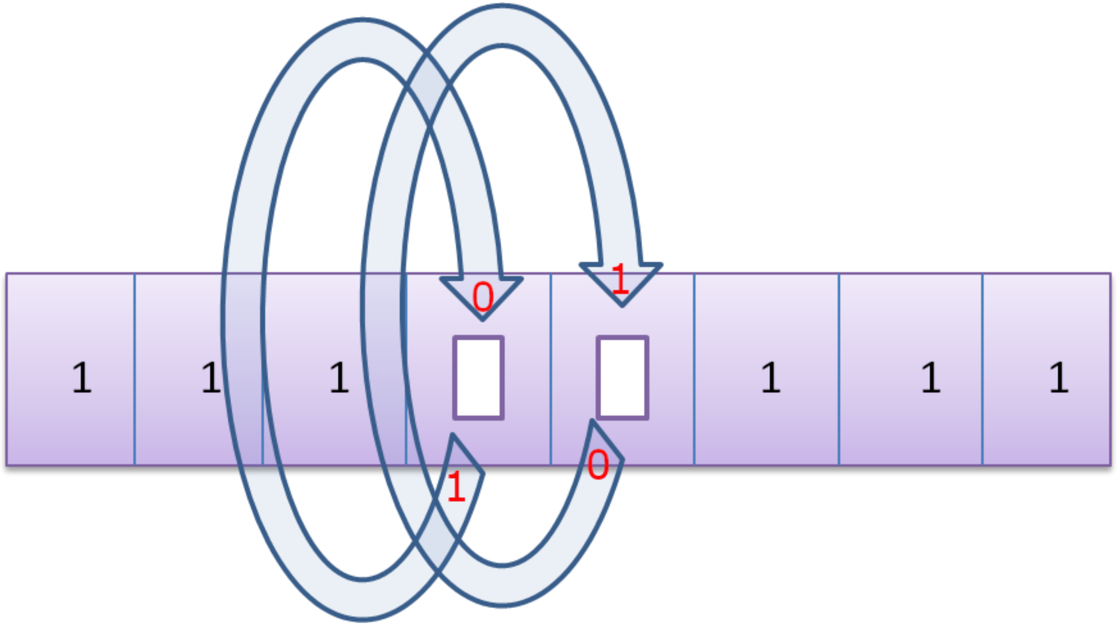


0	1	0	1
1	1	1	1
0	0	0	0
1	0	0	1
0	0	0	0
0	1	1	0
1	1	1	1
1	0	1	0
0	0	0	0

遗传算法
Genetic Algorithm

- 对于变异操作，特别对于二进制编码来说，即作一个位的翻转即可，如下图所示：

  For mutation operation, especially when binary coding is adopted, it is as simple as flipping the bit in specific locations in the bit strings, as illustrated below:



遗传算法
Genetic Algorithm

- 通过遗传算法，我们将阐述以下思想：算法的外延没有特别清晰的界定，符合算法定义特征的计算过程，均可视为算法。在计算过程中，某些子过程，也可以与某些算法相关联。

  By introducing genetic algorithm, we want to show that it is hard to impose a clear boundary for the extension of algorithm. Any process that is in accordance with the characteristics of the definition of algorithms can be treated as an algorithm, though for a complicated process, some sub-process can also be linked to a certain algorithm.

习题
Problems

- 理解如下源代码中的遗传算法的实现并根据自己的理解进行改进：

  To understand the implementation of genetic algorithm in the following source file and to improve it according to your understanding:

  https://github.com/mingyr/intro2algo/blob/master/intro2algo/genetic_algo.cpp