

Mendelian genetics

Heredity is how genetic information passes from one generation to the next; subsequent generations have characteristics of the previous ones.

Gregor Mendel (1822-1884) studied characteristics of pea plants (seed/flower color, height, etc.) and how they are passed to subsequent generations. He proposed two laws of genetics:

Mendel's First Law of Genetics (the law of segregation) basically states that an organism inherits a copy of each gene from each parent. Thus, each organism has two copies of a gene (one on each of the two chromosome copies), and will contribute only one of those copies to offspring. Sometimes these two copies are identical (e.g., green pods, green pods), but sometimes they aren't (e.g., green pods, yellow pods); variations of the same gene are called **alleles**. More technically, a parent will contribute during gamete (sex cell) formation each member of an allelic pair separate from the other member to form the genetic constitution of the gamete.

Mendel's Second Law of Genetics (the law of independent assortment) states that genes are inherited individually during gamete formation; that is, the segregation of the alleles of one allelic pair is independent of the segregation of the alleles of another allelic pair. For example, in pea plants, this would mean that flower color and seed color would be independently inherited.

Mendel did several crosses of plants that were **phenotypically** different (different outward appearance) and observed the characteristics of the subsequent generations. Results from Mendel's experiments are shown in the following table:

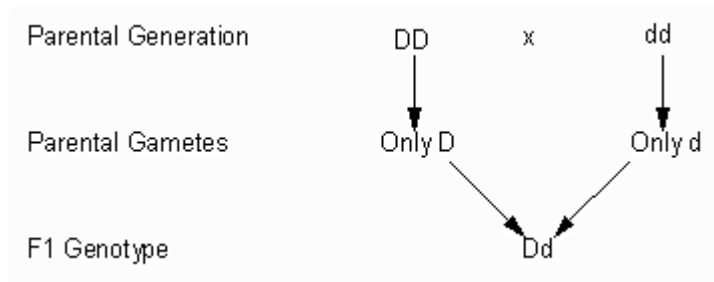
Parental Cross	F ₁ Phenotype	F ₂ Phenotypic Ratio	F ₂ Ratio
Round x Wrinkled Seed	Round	5474 Round:1850 Wrinkled	2.96:1
Yellow x Green Seeds	Yellow	6022 Yellow:2001 Green	3.01:1
Red x White Flowers	Red	705 Red:224 White	3.15:1
Tall x Dwarf Plants	Tall	1787 Tall:227 Dwarf	2.84:1

<http://www.ndsu.nodak.edu/instruct/mcclean/plsc431/mendel/mendel1.htm>

In the F₁ generation, which refers to the offspring of the initial parental crossing, Mendel observed that all the plants exhibited the exact same phenotype for a given characteristic that matched one or the other parent. However, in the F₂ generation, which results from crossing members of the F₁ generation, there is a mix of phenotypes in an approximately 3:1 ratio. The genetic explanation for this phenomenon is that one type of allele for a given gene is **dominant**,

while the other is **recessive**; when both a dominant and recessive allele are present in an individual, only the dominant one is expressed phenotypically. Consider the following example:

Let's say we're crossing tall and short pea plants. Let D be the allele that encodes being tall, and d be the allele that encodes being short. D is the dominant allele, d is the recessive one. In the parental generation, each parent is **homozygous** with respect to height – that is, both copies of the gene are the same allele.



Crossing genotypically different parents produces all **heterozygous** individuals, such that they all have a copy of the D allele and a copy of the d allele. Since D is dominant, all individuals will be phenotypically tall.

In the F2 generation, the random union of gametes produces an equal number of each of the four two-allele combinations; this corresponds phenotypically to a 3:1 ratio of tall to short plants:

	D	d
D	DD (Tall)	Dd (Tall)
d	Dd (Tall)	dd (Short)

Modern genetics

Genes were identified as the unit of inheritance in the 1900s; genes are comprised of DNA and are organized into **chromosomes**. Each individual has two copies of the same chromosome (with the possible exception of the sex chromosomes).

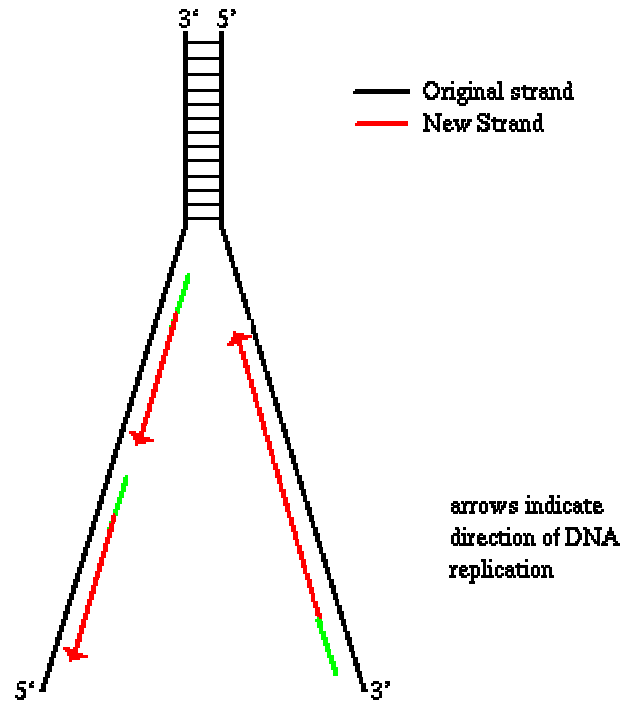
DNA is made of units called **nucleotides**, which are each composed of a sugar, a phosphate and a base. There are four types of bases occurring in DNA, thus four different types of nucleotides: Cytosine (**C**), Guanine (**G**), Adenine (**A**), Thymine (**T**).

In 1953, James Watson and Francis Crick discovered that the structure of DNA is a **double helix**, with complementary base pairs. A is chemically attracted to T, and C with G. For example, the complementary fragment to AGGTAC would be TCCATG

The double-stranded nature of DNA means that it can be replicated by matching complementary nucleotides off a single strand.

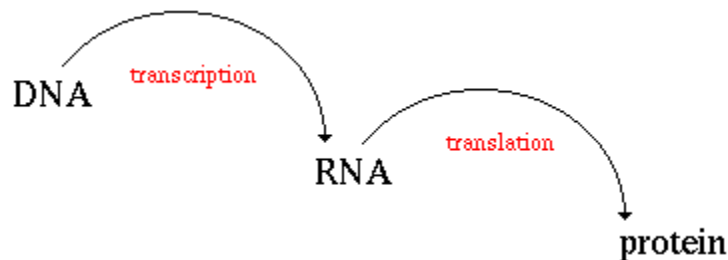


DNA double helix



DNA replication

The **Central Dogma of Molecular Biology** describes the process through which DNA is expressed



In a process called **transcription**, a gene sequence is copied into **mRNA**, which is similar to DNA, except that T nucleotides are replaced by U (Uracil). mRNA serves as a single-strand chemical messenger that encodes the gene that should be expressed.

The next step is **translation**. The cellular machinery starts reading the mRNA at the **start codon**, which is always an AUG nucleotide triplet, and continues until a **stop codon** is hit (UAG, UGA, or UAA). Every three nucleotides together identify one of 20 different **amino acids**, the building blocks of proteins. There are 4^3 possible base pair triplets, and 21 unique end results (20 amino acids + stop signal; the start signal actually encodes a particular amino acid, methionine), so there is a lot of redundancy in the translation table, usually in the 3rd amino acid in the codon.

As the amino acid string is being created, it starts to fold into a 3D structure according to the sequence. The resulting protein is the functional unit of most biological processes.

Course goals

This course will primarily study biology from the point of view of sequences. In particular, the central course ideas are (1) the relationship between sequence and function, and (2) the similarities and differences between different organisms, in terms of genetic sequences.

Summary of computational genomics

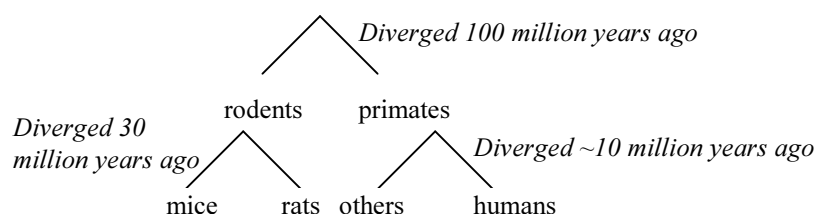
The human genome contains approximately 3 billion letters; traditionally, sequencing techniques existed to read only short sequences of DNA, mainly through gel electrophoresis. Today, we have the complete sequence of the human genome, along with the genomes of other organisms (bacteria, yeast, mouse, drosophila, arabidopsis); these are very long, so this is where we need computer algorithms.

There are basically two tasks: **sequencing DNA**, putting together millions of pieces of short sequences with lots of errors; and **analysis of DNA** sequences – finding the individual genes and other biological features (there are ~30,000 human genes, which comprise only 2% of the genome); comparison of different organisms at the sequence level; and the discovery of networks of gene interactions (microarray technology, clustering).

Computer solutions to these problems include alignment algorithms, probabilistic techniques for sequence analysis, and large systems built from these basic algorithms.

Sequence alignment

The following is an example phylogenetic tree:



Divergence happens through changes in genetic sequences that are propagated to subsequent generations. The mechanism of these changes is called **mutation**, and occurs in several flavors:

Sequence edits are basic kinds of mutations, and include **base substitution**, e.g., AGGTAC becomes AGTTAC; **insertion**, e.g., AGTAC becomes AGGTAC; and **deletion**, e.g., AGGTAC becomes AGTAC.

There are also longer-range mutations, which includes **translocation** (two pieces get exchanged), **reversion** (one piece flips around), and **long-range deletions/insertions**.

Mutations are random processes caused by radiation, errors in replication machinery, etc.. They can do damage if the mutation happens in a critical gene, so mutations that tend to accumulate over generations are those that do not cause changes that affect vital sections of the genome. For example, genes controlling DNA replication are common throughout most organisms. Also, most genes are common among mammals.

Alignment algorithm

Problem definition: best way to match 2 sequences

Example: A G G T A C G A A T _ T A C A
 A G G G* A C _ A A T T T A G* A

“Best” is measured using a scoring function:

- Each **match** gets a reward of m
- Each **mismatch** gets a penalty of $-s$ (denoted by * above)
- Each **gap** gets a penalty of $-d$ (denoted by _ above)

Score = m (#matches) + $(-s)$ (#mismatches) + $(-d)$ (# gaps)

The revised problem definition: Given two sequences of length n and m ,

$$x = x_1 \dots x_n$$

$$y = y_1 \dots y_m$$

Find the optimal alignment of x and y subject to the scoring function.

We can't look at all possible alignments because there are too many: $O(2^{|x|+|y|})$. But because the scoring function is additive, **dynamic programming** can be applied:

Define $x = x_1 \dots x_{i-1} \mid x_i \dots x_n$
 $y = y_1 \dots y_{j-1} \mid y_j \dots y_m$

Score of entire alignment is score of prior portion + score of later portion. In practice, we assume that $x_1 \dots x_{i-1}$ and $y_1 \dots y_{j-1}$ optimally align in some way, and concentrate on trying to align x_i and y_j .

Let $F(i,j)$ = optimal score of aligning $x_1...x_i$ to $y_1...y_j$

Then there are three cases:

- x_i and y_j align; then $F(i,j) = F(i-1, j-1) + \{m \text{ if } x_i = y_j, -s \text{ otherwise}\}$
- x_i aligns with a gap; then $F(i,j) = F(i-1, j) - d$
- y_j aligns with a gap; then $F(i,j) = F(i, j-1) - d$

This is a recursive definition;

- Base case: $F(0,0) = 0$
- Inductive hypothesis: we know $F(i-1, j-1)$, $F(i-1, j)$, $F(i, j-1)$
- Inductive step: calculate $F(i,j)$ as $\max \{ F(i-1, j-1) + m \text{ or } -s$
 $F(i-1, j) - d$
 $F(i, j-1) - d \}$
- $F(n, m)$ is the optimal alignment score

Here, the total number of operations is $O(nm)$, which is a big savings.

Example using a tableau:

$x = \text{AGTA}, y = \text{ATA}, m = 1, -s = -1, -d = -1$

		A	G	T	A
	0	-1	-2	-3	-4
A	-1	1	0	-1	-2
T	-2	0	0	1	0
A	-3	-1	-1	0	2

$F(0,0)$ is an initial gap-gap alignment, set to 0.

$$F(1,0) = F(0,0) - 1 = -1$$

$$F(2,0) = F(1,0) - 1 = -2$$

$$F(1,1) = \max \{ F(1,0) - 1 ; F(0,1) - 1 ; F(0,0) + 1 \} = 1$$

a backpointer is set to point to $F(0,0)$, which yielded the max score at $F(1,1)$

etc.

The optimal score, found in the lower right corner of the tableau, is 2; we can follow the backpointers to get the final alignment:

A G T A
A _ T A

Every alignment has a representation as a path through the matrix (**Needleman-Wunsch / Smith-Waterman matrix**); each subsequent cell in a path is non-decreasing in the x and y directions; the matrix can be divided into sub-matrices corresponding to alignment subproblems. The optimal path passing through (i,j) can be found by solving two smaller alignment problems $(1, 1) \rightarrow (i, j)$ and $(i,j) \rightarrow (n, m)$. These subproblems can in turn be divided into smaller subproblems.

