Molecular Genetics - Mutation

While up to this point we have used the word "mutation" many times, we have not clearly defined mutations, specified how they occur, nor described what causes them. In this unit, we will define mutation, describe different types of mutation, review some causes of mutations, list some of the mutation repair mechanisms in organisms, and examine several ways to detect mutation.

What are mutations?

A mutation is simply a change in DNA sequence. This change can be as large as a chromosomal rearrangement that can be seen under a microscope, or as small as a single nucleotide substitution. Mutations in sex cells (the germ line) can be passed on to future offspring, whereas mutations in non-sex cells (somatic cells) cannot be passed on to future offspring. Mutations can have drastic outcomes (e.g., when a mutation occurs in a locus that encodes or regulates a gene product), or no detectable outcome at all (e.g., when a mutation occurs in a region that does not encode a product). When a detectable effect does occur, more often than not it has a negative impact on organism health/fitness. However, it is possible (although less probable) for a mutation can have a positive outcome.

1 Types of mutation

Large-scale mutations, sometimes referred to as structural mutations, include inversions, translocations, and deletions/duplications that are detectable on a macro scale (e.g., using chromosome staining). Small-scale mutations that can be localized to a specific locus (e.g., base pair) in a gene/genome are known as **point mutations** and **indels**, and these can be detected using cloning/amplification and sequencing technologies (which will be addressed later in this unit).

1.1 Point Mutations

Point mutations are changes where a single nucleotide is replaced by a different nucleotide. These changes can occur anywhere in the genome- including in the coding regions of genes, introns of genes, regulatory regions of genes (e.g., promoter sequences), or non-coding regions. When point mutations occur within the coding sequence (e.g., within the part of the gene that gets interpreted by the ribosome for translation), they can either alter the resulting amino acid sequence (which occurs in nonsynonymous mutations such as missense [change to new amino acid] and nonsense [change to premature stop codon] mutations) or not alter the amino acid sequence (which occurs in synonymous [a.k.a. silent] mutations). Point mutations that occur outside of the coding sense are never classified as nonsynonymous (missense or nonsense) or synonymous - because there is no resultant amino acid that can be changed.

Point mutations can either be **transitions** or **transversions**. A **transition** is when a nitrogenous base is replaced by a different nitrogenous base with the same carbon ring (i.e., a purine is replaced by another purine, or a pyrmidine is replaced by another pyrimidine). For example, an 'A' being replaced by a 'G' is a transition. A **transversion** is when a nitrogenous base is replaced by a different nitrogenous base with a different carbon ring (i.e., a purine is replaced by pyrimidine, or a pyrmidine is replaced by purine). For example, a 'T' being replaced by a 'G' is a transversion. Even though there are more options for transversions, transitions are

more common (because it is more likely for a replication error to occur with a similar-shaped nucleotide).

A **single nucleotide polymorphism** (SNP; pronounced 'snip') is a location in a sequence where there is variation in a biological group of organisms (e.g., population, species, genus, etc.). A SNP is the result of a point mutation, and the different SNPs present in a population are recognized as **alleles** (remember, an allele is simply a unique version of a gene).

1.2 Indels

Indels are insertions or deletions where a nucleotide (or multiple nucleotides) are either removed or added. When an indel of a size that is not a factor of 3 occurs within the coding sequence of a gene, the reading frame for translation is altered. This is known as a **frameshift** mutation.

Summary of mutations in coding sequence

nonsynonymous mutation	point mutation that changes the amino acid sequence
- missense mutation	amino acid changes to a different amino acid (not a stop codon)
- nonsense mutation	amino acid changes to a stop codon (not a different amino acid)
synonymous mutation	point mutation that doesn't change the amino acid sequence
frameshift mutation	indel that changes the reading frame

2 Sources of mutation

Mutations can be caused by **intrinsic factors** (spontaneous, self-caused agents [e.g., replication error caused by DNA polymerase) or **extrinsic factors** (foreign agents [e.g., mutagenic chemicals/radiation]).

2.1 Intrinsic factors

2.1.1 Replication error

Replication error can cause abnormal base pairing and repeat expansion. Abnormal base pairing occurs via a "wobble" effect, wherein incorrect bases are matched due to similar hydrogen bonding conditions (resulting in a transition mutation). Repeat expansion mutations are when repetitive regions of DNA increase in frequency due to the formation of a DNA hairpin (where regions of a nascent strand of DNA bind to nucleotides upstream of their actual base pair). Repeat expansion mutations are the source of variation in **short tandem repeats** (STRs), also known as **microsatellites**. Microsatellites are highly variable loci, and they can be used for paternity tests and forensics. Repeat expansions are also the source of mutation responsible for Huntington's Disease and Fragile X syndrome.

2.1.2 Rare bases

Some rare forms of the nitrogenous bases can cause abnormal base pairing. These are created by proton shifts, deamination, and/or depurination within the nitrogenous bases that then promote different opportunities for hydrogen bonding between bases that usually don't pair.

2.2 Extrinsic factors

2.2.1 Mutagens

Mutagens are foreign, mutation-causing chemicals that either (a) cause alterations to the bases, (b) resemble the form of the bases, or (c) alter the structure of the DNA double helix. The base alterations caused by mutagens are different than the naturally-occuring rare bases described in the "Intrinsic factors" section above. The compounds that resemble the form of bases, known as base analogs, can have a "mimicking" effect where they for a hydrogen bond with the base that normally bonds to the analog. Mutagens that alter the structure of DNA include DNA chelating agents that can wedge themselves inside of a double helix and increase the likelihood of incorrect base pairing during DNA synthesis.

2.2.2 Radiation

Ionizing radiation and **UV radiation** can directly alter DNA structure. Ionizing radiation, which includes alpha particles, beta particles, gamma rays, and X-rays, Can induce a break along the backbone of one (single-strand break) or both (double strand break) of the DNA double helix. UV radiation, which is light with wavelength shorter than that of the visible spectrum, can create a covalent bond between adjacent Thymine bases on the same DNA strand (known as a "kink") that results in the inability of these bases to for hydrogen bonds with the complementary strand.

2.2.3 Transposable Elements

Transposable elements (TEs) can "jump" or be copied from one position of the genome to another, thus altering the sequence of DNA. However, is it more appropriate to consider transposable elements as "extrinsic factors" or "intrinsic factors"? Given that it is hypothesized that TEs originated from the horizontal gene transfer of other organisms into the genomes of diverse organisms, perhaps it is appropriate to consider them as "extrinsic". However, their stability within the genome of many organisms signifies their complete genomic incorporation, and perhaps it is appropriate to consider them as "intrinsic". The categorization is less important than the recognition that these shifty loci are another source of mutation.

3 Detecting Mutation

Originally, detecting mutation was a painstaking process that involved the continual examination of progeny phenotypes after exposure to a mutagen (and even then, this approach could only detect mutations in loci that affected observable phenotypes). Biotechnology breakthroughs in the late 20^{th} century included more sophisticated and objective techniques for detecting mutation. The biotechnology called **polymerase chain reaction (PCR)** and **gel electrophoresis** can be used to detect differences in microsatellite size. PCR and a different biotechnology known as **sequencing** can be used to detect specific changes to a DNA sequence (such as those created by an abnormal base pairing). We will highlight these foundational laboratory techniques in this section, and build upon them in future units to discuss biotech breakthroughs in the 21^{st} century.

3.1 PCR

PCR is basically the isolation and **amplifying** (copying) of a targeted locus. To perform PCR, you must include a few components and reagents in your reaction mixture. This includes:

- **DNA template**: This is the DNA of the individual you want to investigate.
- **deoxyribonucleotide triphosphate (dNTPs)**: These are the monomers that will be used to build the copies of your target locus (A, C, G, and T dNTPs are all needed).
- **Custom-built DNA primers**: Two single-stranded primers flank the target locus.
- **heat-tolerant DNA polymerase**: This enzyme will start at the DNA primers and create the copies of the target locus.
- **Water and buffer solution**: This creates a suitable environment for the DNA polymerase to function.

PCR: Several great videos that will help you understand PCR include `https://www.youtube.com/watch?v=x5yPkxCLads`, `https://www.youtube.com/watch?v=wBrNbbAIAFo`, and `https://www.youtube.com/watch?v=2KoLnIwoZKU`.

Practice Problem: What DNA primers (5 nucleotides in length) would you use to amplify the following DNA sequence?

5'- AACGATAGAAACAGATAGAACACAAGAT -3'

Solution: The first step for a problem like this is to identify the complement of the target sequence. If the sequence is small enough, you can simply write out the whole thing:

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'

After writing it out, design all of the possible primers for the sequence. You can even write all of these out (shown below in red):

3'- TTGCT -5' 3'- TTCTA -5'
5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGA -3' 5'- AAGAT -3'

After designing all of the possible primers, consider which primers will synthesize DNA in the correct direction (over the target locus). Keep in mind that DNA is created from 5' –> 3'.

<– 3'- TTGCT -5' <– 3'- TTCTA -5'
5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGA -3' –> 5'- AAGAT -3' –>

Now select the two primers that will amplify across your target region.

<– 3'- TTCTA -5'
5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGA -3' –>

You have your two primers! Write them out in the conventional 5'–>3' direction:
Forward primer: 5'- AACGA -3'
Reverse primer: 5'- ATCTT -3'

3.2 Gel Electrophoresis

You can visualize molecules and quantify their size using gel eletrophoresis, a molecular laboratory technique that utilizes an electrical current to separate molecules within a gel matrix. The electrical current creates oppositely charged electrodes, and because DNA is negatively charged in travels through the gel matrix towards the positive electrode (the cathode). The gel, at a microscopic level, is a matrix of polymers that makes traveling through it more difficult for larger molecules compared to smaller molecules. Therefore, smaller DNA fragments can move through at a faster rate. In order to visualize the DNA, you can use a chemical that binds to DNA and fluoresces under UV light. Using this technique, you can (1) confirm that your PCR worked (by looking for a band at your expected size), and (2) compare loci that vary in size across individuals (e.g., microsatellites).

3.3 Sanger Sequencing

Sequencing technologies determine the exact nucleotide sequence for a locus or loci. Sanger sequencing is one of the earliest methods of sequencing that was developed, and it is still considered the "gold standard" for determining the sequence of a locus. Other techniques that sequence many loci at a time (i.e., "high-throughput" genomic sequencing) are more cost-effective and time-efficient, but they are still prone to errors (we will discuss these genomic sequencing technologies in a later unit). Sanger sequencing uses the same components as PCR, with a small (but critical) difference. All of the components for Sanger sequencing include:

- DNA template: This is the DNA of the individual you want to investigate.
- deoxyribonucleotide triphosphate (dNTPs): These are the monomers that will be used to build the copies of your target locus (A, C, G, and T dNTPs are all needed).
- **Custom-built DNA primer**: A single-stranded primer adjacent to the target locus.
- heat-tolerant DNA polymerase: This enzyme will start at the DNA primers and create the copies of the target locus.
- Water and buffer solution: This creates a suitable environment for the DNA polymerase to function.
- dideoxyribonucleotide triphosphate (ddNTPs): These are the monomers that will be used to terminate the synthesis of DNA molecules (A, C, G, and T ddNTPs are all needed- but in the old-school way of Sanger sequencing each is used in a separate reaction).

  Did you catch the differences between Sanger sequencing and PCR? If you noticed the primer (singular rather than plural) and ddNTPs, you are right! The single primer is needed because you are only going to look at the sequence in a single direction (you don't want to confuse the forward and reverse strands). However, in most cases you will use both primers- you will just have two separate reactions (one reaction for each primer). The ddNTPs terminate the elongation step, creating sequences of different sizes for a single locus. The "old-school" way of Sanger sequencing described above requires four separate reactions for each of the four ddNTPs. Each of these reactions can be ran on a gel, and together they can be used to infer the nucleotide sequence of a locus. This approach is no longer used- instead, a single reaction with all of the ddNTPs is implemented where each ddNTP has a unique fluorescent die that can be detected by a sequencing machine.

Sanger Sequencing: Several great videos that will help you understand Sanger sequencing include `https://www.youtube.com/watch?v=dVRB4CaLizc`, `https://www.youtube.com/watch?v=FvHRio1yyhQ`, and `https://www.youtube.com/watch?v=e2G5zx-OJIw`.

Practice Problem: If you sequenced the following locus using the "old-school" approach of Sanger sequencing, what would your resulting sequences look like for each reaction? Assume you use the forward primer from the previous practice problem (5'- AACGA -3').

5'- AACGATAGAAACAGATAGAACACAAGAT -3'

Solution: The first step for a problem like this is to identify the complement of the target sequence. If the sequence is small enough, you can simply write out the whole thing:

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'

After writing it out, write out where the primer will bind:

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGA -3'

Then write out all of the sequences and their respective lengths that will occur for the reaction with the 'A' ddNTP (where the ddNTP is shown in bold):

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGAT**A**-3' (7 nucleotides)
5'- AACGATAG**A**-3' (9)
5'- AACGATAGA**A**-3' (10)
5'- AACGATAGAA**A**-3' (11)
5'- AACGATAGAAAC**A**-3' (13)
5'- AACGATAGAAACAG**A**-3' (15)
5'- AACGATAGAAACAGAT**A**-3' (17)
5'- AACGATAGAAACAGATAG**A**-3' (19)
5'- AACGATAGAAACAGATAGA**A**-3' (20)
5'- AACGATAGAAACAGATAGAAC**A**-3' (22)
5'- AACGATAGAAACAGATAGAACAC**A**-3' (24)
5'- AACGATAGAAACAGATAGAACACA**A**-3' (25)
5'- AACGATAGAAACAGATAGAACACAAG**A**-3' (27)
5'- AACGATAGAAACAGATAGAACACAAGAT-3' (28)

Then write out all of the sequences and their respective lengths that will occur for the reaction with the 'G' ddNTP (where the ddNTP is shown in bold):

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGATA**G**-3' (8)
5'- AACGATAGAAACA**G**-3' (13)
5'- AACGATAGAAACAGATA**G**-3' (18)
5'- AACGATAGAAACAGATAGAACACAA**G**-3' (26)
5'- AACGATAGAAACAGATAGAACACAAGAT-3' (28)

Then write out all of the sequences and their respective lengths that will occur for the reaction with the 'C' ddNTP (where the ddNTP is shown in bold):

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGATAGAAA**C**-3' (13)
5'- AACGATAGAAACAGATAGAA**C**-3' (21)
5'- AACGATAGAAACAGATAGAACA**C**-3' (23)
5'- AACGATAGAAACAGATAGAACACAAGAT-3'(28)

Then write out all of the sequences and their respective lengths that will occur for the reaction with the 'C' ddNTP (where the ddNTP is shown in bold):

5'- AACGATAGAAACAGATAGAACACAAGAT -3'
3'- TTGCTATCTTTGTCTATCTTGTGTTCTA -5'
5'- AACGA**T**-3' (6)
5'- AACGATAGAAACAGA**T**-3' (16)
5'- AACGATAGAAACAGATAGAACACAAGA**T**-3'(28)