TEACHING THE GENOME GENERATION

## *Sequence Comparison with CYP2C19*

The Jackson Laboratory
*Leading the search for tomorrow's cures*

NIH National Institutes of Health
*Turning Discovery Into Health*
SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

## Sequence Comparison with *CYP2C19*

Did you know that humans are, on average, 99.9% genetically identical? Only 0.1% of all our DNA bases are different, but those differences are what influence our traits and help make us each who we are.

The Cytochrome P450, family 2, subfamily C, polypeptide 19 (*CYP2C19*) gene codes for the Cytochrome P450, family 2, subfamily C, polypeptide 19 (CYP2C19) protein, which is an enzyme that catalyzes many reactions involved in drug metabolism, synthesis of cholesterol, steroids, and other lipids. Variants in the human *CYP2C19* gene are associated with differences in drug metabolism. In this activity, you'll compare *CYP2C19* DNA sequences from different individuals and different organisms.

When comparing DNA sequences, **percent identity** provides a measure of how similar two sequences are. The formula for percent identity uses the total number of nucleotide **positions** in the sequence comparison and the number of nucleotide positions that are different, or **divergent**, between the sequences:

$$Percent\ Identity = \frac{\#\ positions - \#\ divergent\ positions}{\#\ positions}\ x\ 100\%$$

### *Part 1. Compare to a Reference Sequence*

One common type of sequence comparison is comparing an individual's DNA sequence to a reference sequence. A **reference sequence** is a DNA sequence that is assumed by scientists to be a representative example of the genetic material of a specific species. Reference sequences are typically created by combining the DNA sequences of multiple individuals from the same species.

Comparing an individual's DNA to a reference sequence allows us to identify variants, or differences, between the individual's DNA sequence and the reference.

In this example, we will compare an individual person's DNA sequencing data for the *CYP2C19* gene to the human reference sequence for *CYP2C19* to identify which *CYP2C19* gene variant that person carries. We'll call this person **Jean**.

The sequence comparison below looks at a portion of the *CYP2C19* gene for two sequences:

- The **Query** sequence is **Jean**'s *CYP2C19* sequence data.
- The **Subject** ("Sbjct") sequence is the human **reference** sequence for *CYP2C19.*

When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from **Jean** and which is the **reference** sequence.

```
_______ Query  1    ATTTTCCCACTATCATTGATTATTTCCCAGGAACCCATAACAAATTACTTAAAAACCTTG  60

_______ Sbjct       ............................G.............................
```

1. How many nucleotides are different between the two sequences within this portion of the *CYP2C19* gene? What is/are the variant nucleotide(s) in the individual's DNA sequence and the reference DNA sequence?

2. In this comparison, there are 60 positions total. What is the percent identity for this comparison?

3. The entire human *CYP2C19* gene sequence is 90,209 nucleotides. What is the percent identity for the entire *CYP2C19* sequence? You can assume that all the other nucleotides are the same between the two sequences. Round your answer to two decimal places.

Sometimes researchers know which gene a DNA sequence is from, but other times they don't. Comparing sequences and calculating percent identity can help them figure out which gene their sequence is from. Since humans are 99.9% genetically identical to each other on average, a very high percent identity provides more confidence that the sequence is from a given gene.

4. If you didn't know which gene **Jean**'s DNA sequence was from, would you feel confident that this sequence was from *CYP2C19* given the percent identity you calculated in question 3? Why or why not?

5. What if the percent identity was 99%? 95%? 75%? Justify your answer.

*Part 2. Compare Across Species*

Imagine that you are a researcher studying drug metabolism, and you want to learn more about the *CYP2C19* gene. You plan to use a model organism to conduct some experiments to better understand the impact of the *CYP2C19* gene on drug metabolism. Model organisms, such as mice and fruit flies, are often used as a representation of human biology because they are easier to study in controlled environments and share much of the same physiology as humans.

When you try to find the sequence for *CYP2C19* in mice, you are surprised to find that mice do not have a *CYP2C19* gene! However, mice do have several genes that share sequence similarity with human *CYP2C19*, like cytochrome P450, family 2, subfamily C, polypeptide 65 (*CYP2C65*). These two genes are part of the same gene family, or group of genes with similar functions.

You decide to compare a small section of the mouse reference sequence for *CYP2C65* (**Query**) to the human reference sequence for *CYP2C19* (**Sbjct**). When comparing across species, it can be helpful to record which sequence corresponds to which organism. In the figure below, write in which sequence is **mouse** and which is **human**.

```
_____ Query  1     ATTTTCCTGCTGTCATTGATTATCTACCAGGAAGACACAGAAAATTACATAAAAATTTTG  60

_____ Sbjct        .......CA..A...........T.C..G....CC..T.AC.......T......CC...
```

1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Now, look at a different section of mouse *CYP2C65* (**Query**) and human *CYP2C19* (**Sbjct**).

```
_____ Query  1     ACAATCCTCGGGACTTTATTGATTGTTTCCTGATCAAAATGGAACAGGAAAAGCACAACC  60

_____ Sbjct        ....C.....................C.................GA..........A....
```

3. How many nucleotides are different between the two sequences? What types of differences are they?

4. What is the percent identity for this comparison?

While the start of the sequences for human *CYP2C19* and mouse *CYP2C65* are similar, the mouse sequence (1941 nucleotides) is much shorter than the human sequence (4131 nucleotides). The percent identity for the comparison of the whole protein coding sequences of human *CYP2C19* and mouse *CYP2C65* is about 30%. For genes shared between mice and humans, percent identity is 85% on average.

5. Given a 30% identity for human *CYP2C19* and mouse *CYP2C65*, would you feel confident using mice as a model for your study on drug metabolism? Justify your answer.

6. What is one other system or method you might be able to use to study human *CYP2C19* and drug metabolism?

Comparing DNA sequences across species can also provide information about how species evolved over time. In general, the greater the percent identity of DNA sequences between two species, the more recently they have shared a common ancestor. It takes time for genetic differences to accumulate, so organisms with fewer genetic differences are typically more closely related.

7. Which species would you expect to have a more similar DNA sequence to human *CYP2C19*: mice or chimpanzees? Justify your answer.

Unlike mice, chimpanzees have a *CYP2C19* gene. Check your hypothesis by comparing a portion of the chimpanzee *CYP2C19* reference sequence (**Query**) and the human *CYP2C19* reference sequence (**Sbjct**).

```
_________ Query  1      ATTTTCCCACTATCATTGATTATTTCCCGGGAACCCATAACAAATTACTTAAAAACCTTG  60

_________ Sbjct         ............................................................
```

8. How many nucleotides are different between the two sequences? What types of differences are they?

9. What is the percent identity of this section of the *CYP2C19* gene?

10. Compare your percent identity calculations for questions 2 and 4 (mouse) with your calculation for question 9 (chimpanzee). Do they support your hypothesis about whether the mouse *CYP2C65* or chimpanzee *CYP2C19* DNA sequence is more similar to the human *CYP2C19* sequence? Explain your reasoning.

11. If not, why do you think that might be?

### *Part 3. Compare Within Species*

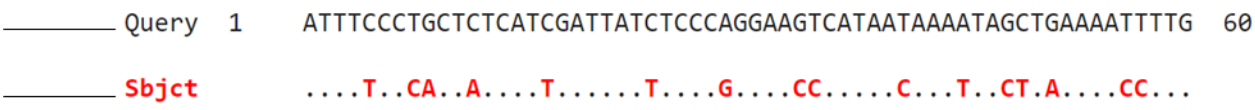
Another type of sequence comparison is comparing the DNA sequences of different genes within the same species. Comparing different genes within the same species can help scientists identify gene families.

Gene families are groups of genes with similar functions. Comparing sequences helps identify gene families because a gene's sequence determines its associated protein's structure, which determines protein function. This is especially useful in organisms where a full genome sequence is not known. By comparing new gene sequences to known genes, scientists can determine if the new gene serves a similar function to a known gene.

Comparing genes within species can also provide us information about evolution. Sometimes, as species evolve, genes get duplicated. Over time, these gene duplicates accumulate changes and become different enough that they serve different, yet related, functions.

Let's look at an example. *CYP2C19* is in the same gene family as cytochrome P450 family 2 subfamily C member 8 (*CYP2C18*). How similar are these two sequences?

Start by comparing a small section of the human *CYP2C18* reference sequence (**Query**) to the human *CYP2C19* reference sequence (**Sbjct**). When comparing sequences, it can be helpful to record the source of each sequence. In the figure below, write in which sequence is from ***CYP2C18*** and which is from ***CYP2C19***.




1. How many nucleotides are different between the two sequences? What types of differences are they?

2. What is the percent identity for this comparison?

Overall, human *CYP2C18* has an 84% identity with human *CYP2C19*.

3. Given this percent identity and what you know about the function of CYP2C19, what might you predict the function of CYP2C18 in the body to be?