# BioMath

## Genetic Inversion: Relationships Among Species

### Student Edition



COMAP


Funded by the National Science Foundation,
Proposal No. ESI-06-28091

This material was prepared with the support of the National Science Foundation. However, any opinions, findings, conclusions, and/or recommendations herein are those of the authors and do not necessarily reflect the views of the NSF.

At the time of publishing, all included URLs were checked and active. We make every effort to make sure all links stay active, but we cannot make any guaranties that they will remain so. If you find a URL that is inactive, please inform us at info@comap.com.


Published by COMAP, Inc. in conjunction with DIMACS, Rutgers University.


COMAP, Inc.
175 Middlesex Turnpike, Suite 3B
Bedford, MA 01730
www.comap.com

ISBN: 1 933223 62 6

Front Cover Photograph: EPA GULF BREEZE LABORATORY, PATHO-BIOLOGY LAB. LINDA SHARP ASSISTANT

# Genetic Inversion: Relationships Among Species

You are a Biomath graduate student conducting research in the Amazon Rainforest. One day you are lucky enough to stumble upon a previously undiscovered creature. To name this creature (perhaps after yourself!) it is important to determine its place on the evolutionary tree of life. After sequencing a section of chromosome 6 you use a computer to look for matches to any known species. The computer finds that the genes from chromosome 6 correspond to the genes in the frillneck lizard (a currently living species) but they are not in the same order. Your task now is to create a phylogenetic tree to show the relationship between the newly discovered creature and the frillneck lizard for your local natural history museum. The museum wants a display of your findings using model creatures complete with a drawing of the most likely phylogenetic tree. The order of genes will be given to you later.

Photograph by Miklos Schiberna (Own work) [Public domain], via Wikimedia Commons.

You might first be thinking, “What is a phylogenetic tree?” and then be trying to figure our how you would create one. In this unit you will learn about sequencing sections of chromosomes and analyzing inversions of these sequences. You will discover how to create a phylogenetic tree.

# Unit Goals and Objectives

Goal**:** Students will understand the role of chromosome inversion mutations in evolution.
Objectives:

- Define a chromosome inversion mutation.
- Diagram an inversion event.
- List the 3 possible outcomes of a genetic inversion.
- Explain why each outcome could arise.
- Calculate the number of inversions necessary to transform one sequence into another.
- Calculate the most likely phylogeny of a set of organisms using genome inversion data.

Goal: Students will understand algorithms and algorithmic thinking, particularly in the context of optimizing some value (in this case, minimizing inversions).
Objectives:

- Explain key aspects of algorithms and develop an algorithm to solve inversion problems.
- Apply a given algorithm to inversion problems.
- Compare algorithms' efficiencies.

Goal: Students will understand the need for combining biological and mathematical approaches to answering some questions.
Objectives:

- Explain how the study of evolution requires both biological and mathematical understanding.
- Generate examples of other situations (not the study of evolution) where a combination of biology and mathematics is needed.

## Lesson 1 Subsequence Inversions

> *The surface of our planet is populated by living things—curious, intricately organized chemical factories that take in matter from their surroundings and use these raw materials to generate copies of themselves. The living organisms appear extraordinarily diverse in almost every way. What could be more different than a tiger and a piece of seaweed, or a bacterium and a tree? Yet our ancestors, knowing nothing of cells or DNA, saw that all these things had something in common. They called that something "life," marveled at it, struggled to define it, and despaired of explaining what it was or how it worked in terms that relate to nonliving matter.*[1]

In this unit, we will work on the problem of determining how closely related two different **species** of animals are. Because the genes located on the chromosomes of animals can be represented mathematically as **sequences** of numbers, we will focus on the related question of how closely related two numerical sequences are. Consider the three pairs of sequences shown below.

| | | |
|---|---|---|
| 1 3 2 4 5 6 | 3 2 1 5 6 4 | 3 4 6 5 2 1 |
| 1 2 3 4 5 6 | 1 2 3 4 5 6 | 1 2 3 4 5 6 |

Many people might say that the pair of sequences of numbers in the left column above are much more similar than the pair of sequences in the right column above. How can we measure how closely related two given sequences are? What is the significance of this question for biology and for mathematics? These are the questions we explore in this unit.

Note: The sequential ordering of a group of numbers beginning with 1 is called the **identity** (or normal) **sequence**. The 2nd sequence in each pair above is the identity sequence.

### Inversions

One way to determine how closely related two sequences are is to use the method of subsequence inversions. A **subsequence** is a sequence of two or more adjacent items that are usually a smaller portion of a larger sequence; however, sometimes the entire sequence might be considered a subsequence of itself. An **inversion** is the reversing of a subsequence within a sequence, or the reversing of the entire sequence.

We count the number of subsequence inversions required to change the **initial sequence** into the **target sequence**.

Example 1: Start with the initial sequence 1 3 2 4 5 6. Invert the subsequence 3 2 to 2 3 to reach the target sequence 1 2 3 4 5 6. Just one inversion!
Example 2: Now start with 3 2 1 5 6 4. We could invert the subsequence 3 2 1 to produce the new sequence 1 2 3 5 6 4. Next, we could invert the subsequence 5 6 to produce the new sequence 1 2 3 6 5 4. Finally, we could invert the subsequence 6 5 4 to produce the identity (or normal) sequence 1 2 3 4 5 6. We used three inversions to change 3 2 1 5 6 4 into 1 2 3 4 5 6.

## *ACTIVITY 1-1* Playing Card Inversions

**Objective:** Use the minimum numbers of subsequence inversions to change the initial sequence of cards to the identity (normal) sequence.

**Materials:**

Numbered index cards (from 1 to 10) or playing cards of one suit (from Ace to 10)
Handout GI-H2: Playing Card Inversions Activity Worksheet
Handout GI-H3: Score Table

**Round 1**. Use only the cards from 1 to 6 (or Ace to 6). Shuffle the cards and have each player choose one card without looking. The player choosing the higher number (Player A) shuffles the cards and randomly places them in a sequence. The other player (Player B) must use subsequence inversions to change the random arrangement of cards into the identity (or normal) ordering of 1 2 3 4 5 6. Keep track of the inversions in the score table. The number of subsequence inversions used by Player B is Player B's score for the first round. Then Player B shuffles the cards and gives Player A a random sequence of the six cards. Player A changes the sequence into the identity ordering by using sequential inversions. The number of inversions used is Player A's score. After each player has had three sequences, add up the scores. The lowest score wins.

**Round 2.** Now play a few more times but with sequences of different lengths (choose lengths between 3 and 10).

Draw a score table as shown below to keep track of the lengths of your sequences and the numbers of inversions you use. Pay attention to the least and greatest numbers of inversions different pairs of sequences need to transform one sequence of the pair into the other sequence of the pair.

| Length | Initial Sequence | Inversions | Target Sequence | # of Inversions |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

**SCORE: _____**

## *ACTIVITY 1-2 Reverso!*

**Objective:** *Reverso* is an interactive computer game based on sequences of colors. The goal is to change one sequence of colors into another sequence called the identity (or target) sequence.

**Materials:**

Computer with Reverso game downloaded
Handout GI-H3: Reverso! Activity Worksheet
Handout GI-H4: Score Table

**Trial Round:** After launching the applet, minimize other windows so that you can see the workspace window and the initial window asking you how long of a sequence you want to work with (between 6 and 14). Choose a length of 6. You begin with two rows of colored tiles. The top row is the target sequence. The second row is the sequence you need to transform into the target sequence. Click on two tiles that mark the beginning and end of the subsequence you want to invert. X's appear on the two tiles, if you make a mistake, clicking again on a tile removes its X. Once you have two X's, click on "Go!" Another row of tiles will appear that results from making the subsequence inversion you specified. When you succeed, a message appears telling you how many swaps (or inversions) you made. Select "New Game" from the "home" menu to continue.

**Round 1:** Participants in the group take turns using a sequence of length 6. Participants keep track of their inversions in the score table below. The number of inversions used to reach the target sequence is the player's score. After each player has had three sequences, add up the scores. The lowest score wins.

**Round 2:** Now play a few more times but with sequences of different lengths (choose lengths between 3 and 10).

**Round 3:** You can replay the game with a specific sequence. Challenge your teammates to solve one of their sequences in fewer inversions to lower your score.

Draw a score table as shown below to keep track of the lengths of your sequences and the numbers of inversions you use. Pay attention to the least and greatest numbers of inversions different pairs of sequences need to transform one sequence of the pair into the other sequence of the pair.

| Length | Initial Sequence | Inversions | Target Sequence | # of Inversions |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

**SCORE: _____**

**Biology Background**

The **genome** is the entirety of the biological information of an organism. And, although living things are seem to vary infinitely when viewed from the outside, the inside mechanical structure of all living things are fundamentally similar. **Cells** are the building blocks of all living things. Cells are controlled by the genetic information contained within their **DNA**. DNA is coiled and packed into a structure called a **chromosome**. A segment of the DNA/chromosome that codes for a particular **protein** is called a **gene**. The chromosome has many genes lined up in a particular order. Another piece of the chromosome is called the **centromere**. A centromere is visualized as a constriction observed in mitotic chromosomes. The location of the centromere can be useful in determining if an inversion has occurred. **Mutations** can occur that change the order of the genes and the centromere along a chromosome. These "chromosomal mutations" include insertions, duplications, deletions, translocations and inversions. This unit is only interested in inversions. An inversion occurs when a single gene or a group of genes detach from the DNA strand, rotate 180°, and reattach to the strand.

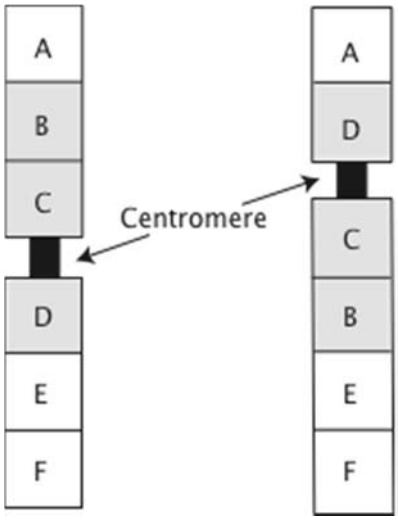



**Figure 1.1:** Section of Chromosome

Figure 1.1 shows a section of a chromosome (with letters representing arbitrary regions containing many genes). It shows the subsequence that has been inverted, thus moving the centromere. Any genes that were within this section of the chromosome have now been moved to a new position with the exception of the potential gene at the 'pivot point.' Illustrated a different way with numbers, let us start with the gene sequence 123456. If we invert the strip '234' of the chromosome then we have the new alignment 143256. If the centromere was within the 234 segment then it will move positions, otherwise no change in centromere location will be noted. Note that genes outside the inversion area remain unchanged in their global location on the chromosome but may now be next to a gene they were not adjacent to before. In the above example, gene 1 started next to gene 2. After the inversion of 234, gene 1 is now adjacent to gene 4, even though gene 1 was not involved in the inversion.

The change in location of the centromere can indicate that an inversion has occurred. If you know the original chromosome has the centromere in the center and, after **DNA replication**, a copy has the centromere in a different location you can deduce that a mutation event has occurred. Again, this can be seen in Figure 1.1.

Figure 1.2 below shows all the different human chromosomes—note the centromere location is quite different in any given chromosome. The fact that it is not in the center is suggestive that inversions occurred throughout **evolution** to shift its position, and therefore the relative gene positions, which led to the human species.

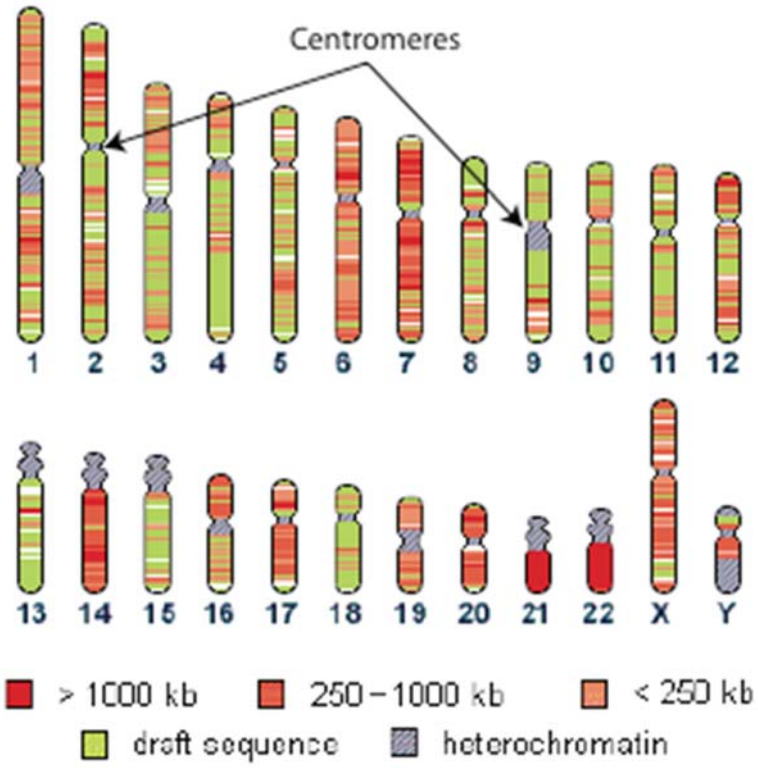



**Figure 1.2:** Human Chromosomes
Source: www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/ chromosome_ideograms.gif. Public domain

Chromosomal inversions can occur in any cell in the human body (or in any living creature) but the most important events occur in **sex cells** (sperm or egg cells). These changes exhibit their full effect on any offspring formed from these inverted sperm or egg chromosomes. In other words, an inversion that occurs in a skin cell only affects that skin cell and its descendants, but the other trillion cells in your body remain normal. However, if the inversion occurs in a sperm or egg cell, when the cells combine to form a single cell **zygote** every new cell created in that offspring will contain that inverted sequence.

Chromosomal inversion events can result in 3 basic outcomes for the offspring:

Advantageous - perhaps due to the activation of a gene that was previously inhibited due to its position on the chromosome or deactivation of a gene that was previously active.

Example: In an arctic environment, if the inversion deactivates a gene that codes for black fur then it causes the animal to have white fur thus blending better with its environment.

Disadvantageous – same as above but either the environment was different so the color change was not preferred or the inactivated gene coded for a protein that is essential for the survival of the organism. Without the essential protein the animal will not survive. Therefore disadvantageous changes can result in an offspring that does not survive or an offspring with less chance to survive in that particular environment.

Example: Hemophilia, a disease where an individual continues to bleed when injured, can be caused by genetic inversions on the X-chromosome. Another example of a disease that is caused by an inversion on the X-chromosome is Complete Androgen Insensitivity Syndrome (CAIS). This is where a person's cells do not respond to testosterone (a type of androgen hormone) so the person has the external anatomy of a female, but the XY chromosomes of a male. The mutation on the X-chromosome disrupts the gene that codes for androgen receptors, so the person does not make the receptor to cause the cells to react to the presence of testosterone.

No effect – this could occur for 2 reasons

a. The activation or deactivation of a gene was not important to the organism's survival.

b. The inversion caused no change in the gene activity. The genes are simply in a new location but work exactly the same as before the inversion.

If you are interested, search the Internet for ***Joint Genome Institute: Sequencing Targets and Associated Diseases*** to find diseases that have been mapped to mutations of chromosome 5, 16 and 19. Not all of these diseases are caused by an inversion mutation but, nonetheless, the image is an excellent visual to demonstrate the importance of mutations and genetic study.

**Questions for Discussion**

1. What is a gene? What does a gene do?

2. What is the term used when a gene mutates by rotating 180º?

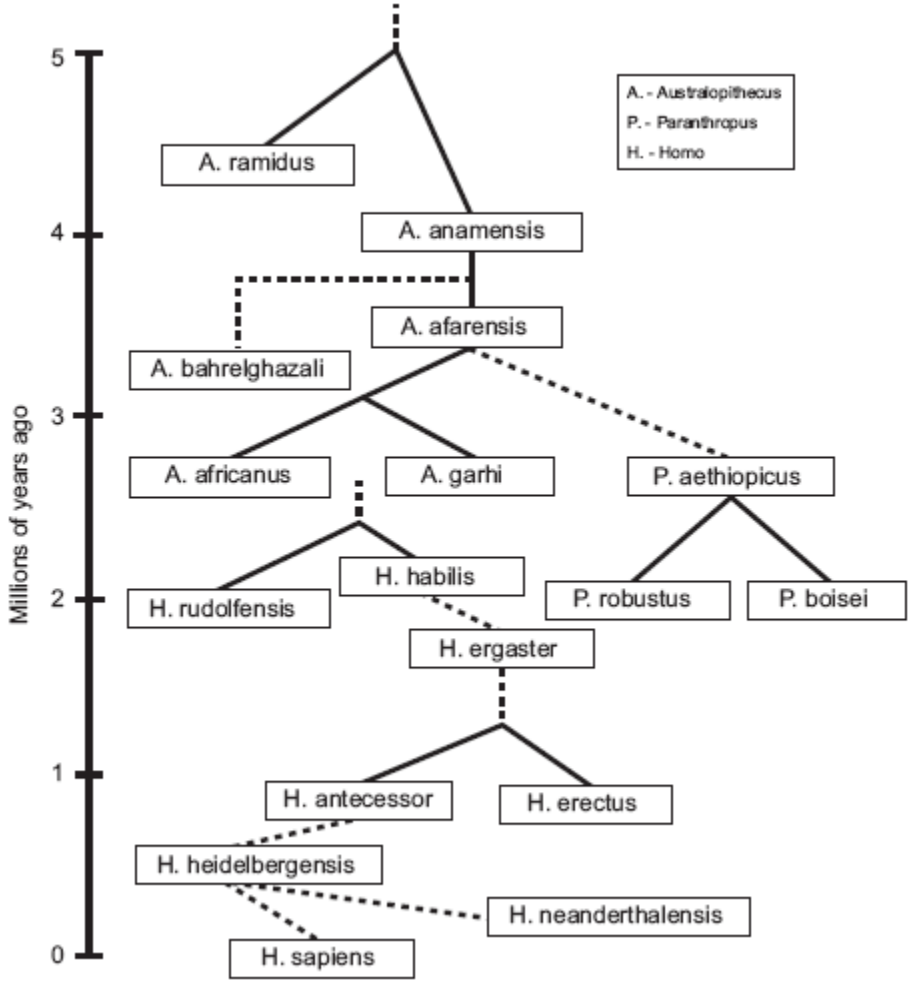
3. What are the three possible outcomes of a mutation?

4. Why is a mutation in a sex cell more significant than a mutation in any other type of cell?

5. Explain the relation between **centromere** location and evolution.

**Practice**

1. Ancestors. Use the diagram below to answer questions a – g.




a. Which organism is the oldest?

b. Which two are more closely related: *H. rudolfensis* and *H. erectus*, or *P. robustus* and *P.boisei?*

c. Who is the direct ancestor of *P. boisei*? Of *Au. Garhi*?

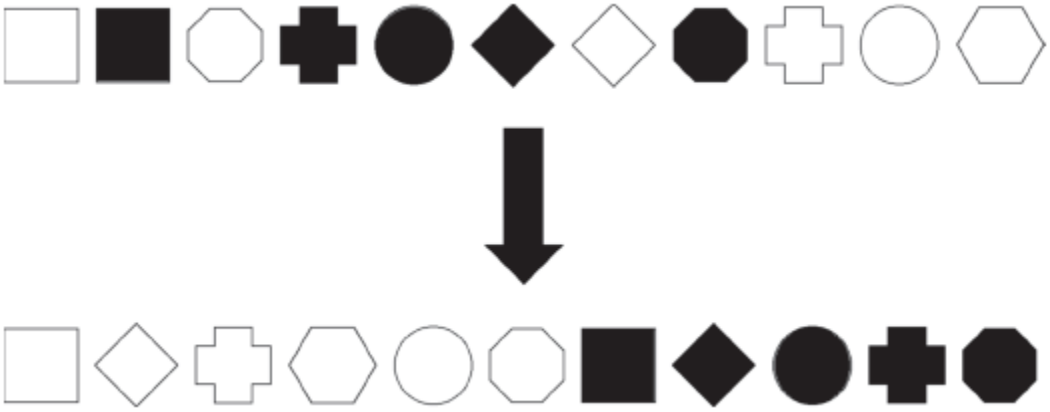
d. What does it mean if a species has no descendants?

e. What do we call the type of diagram above?

f. How do you think decisions were made regarding where to place the species?

g. What other ways could they have compared these species to place them on the tree? What way do you think would give the most accurate diagram?

2. Inversion Puzzle. Convert the top sequence of shapes into the bottom target sequence using only inversions.



a. How many inversions did you need to make?

b. Do you think anyone could make the conversion with fewer inversions? Why or why not?

## Lesson 2 Chromosomal Inversions and Evolution

When genes are inverted in an organism, they actually rotate. For example, in the sequence MOXIN, if the MOX is rotated 180° it would actually lead to XOWIN. Note the sequence of 3 letters is reversed and then flipped so that the M is now 'upside down' to become the W. (Conveniently, the X and O look the same when rotated.)

### Inversions and Rotations

In order to take into account both inversions and rotations, we use negative and positive numbers. Consider the following example.

| I | W | O | S | S | I | N | O |
|---|---|---|---|---|---|---|---|
| -3 | -2 | -1 | 4 | 5 | 6 | -8 | -7 |

| O | M | I | S | S | I | N | O |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | -8 | -7 |

| O | M | I | S | S | I | O | N |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

Ignoring the sign of each number, notice the **breakpoints** in the original sequence. Breakpoints are places in the sequence where the numbers are not sequential. In the original sequence this occurs between -1 and 4, as well as between 6 and -8. Note that the inversions were made at the breakpoints. If you know the location of the breakpoints, is the inversion easier to complete?

The negative signs denote a gene that is positioned 'backwards.' We can see this notation in the actual version of the mouse-to-human rearrangement shown below.

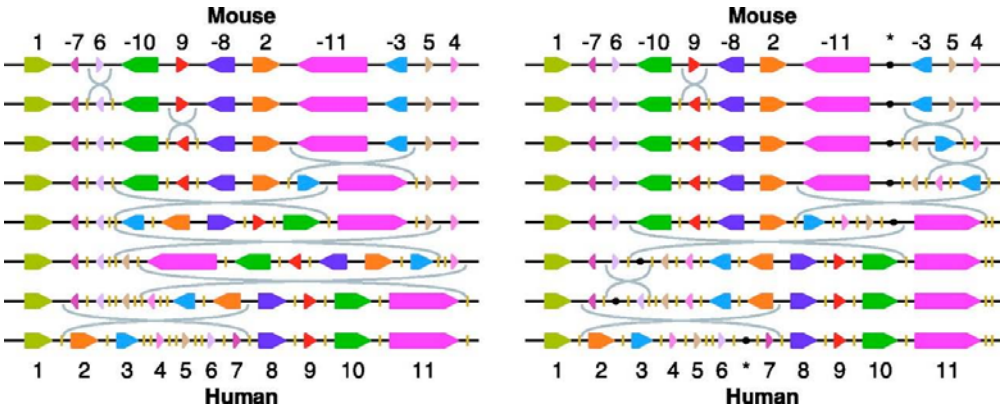



**Figure 2.1:** Examples of Mouse to Human X-chromosome Inversion

Figure 2.1 shows two ways a mouse X-chromosome can be converted to a human X-chromosome in 7 steps by simply using inversions, occurrences of which might be separated by thousands of years.

Since we are just beginning our study of genetic inversions, we will omit the rotation aspect of the inversions and only focus on the inversions of the kind we did in Lesson 1. This unit's version of the mouse to human conversion disregards these reversals of direction and simply concentrates on the sequence of numbers. In reality, a sequence of 1 2 3 4 -5 is not the same as 1 2 3 4 5. It will require one more inversion of the -5 to obtain the +5 direction. For now, we will ignore the final inversion.

**Making Connections: Inversions, Chromosomes, Mutation, and Evolution**

Tracking chromosomal inversions is important biologically because it can help determine **evolutionary phylogeny** (the evolutionary development or history of related species). Mutations are the driving force of evolution and genetic inversions are one form of these mutations. By determining the most efficient series of inversions to lead from the given sequence to the identity sequence, one can infer the minimum number of evolutionary steps between two organisms and the evolutionary order of a group of organisms connected by those inversions.

> Example: Given the identity sequence 1 2 3 4 5 for one species and a second species' sequence of 3 2 1 4 5, it can be seen that a single inversion might separate these two organisms. Therefore, it would be deduced that the two species are closely related evolutionarily. On the other hand, 5 2 3 1 4 would require three inversions to obtain the identity sequence and is therefore a more distant relative than the single inversion organism. One inversion sequence to obtain this is: 5 2 3 1 4 → 1 3 2 5 4 → 1 2 3 5 4 → 1 2 3 4 5.

Additionally, a chromosomal inversion can actually break in the middle of a gene, thus rendering that gene inoperable. For this unit it is assumed that the chromosomal breaks do not break within a gene but rather shift complete genes in their entirety. And, as was discussed above, in this unit, we also do not pay attention to the changes in orientation (due to rotation) that accompany chromosomal inversions. In fact, some rotations do not affect gene function, so this simplification is warranted in many cases.

**Questions for Discussion**

1. If a sequence of numbers represents the genes on a species' chromosome, then how is the number of inversions required to change a chromosome from one species into another species' chromosome related to the similarity of the two species?

2. Can any sequence be transformed into the identity sequence using only inversions? Encourage students to develop a convincing argument for answering this question in the affirmative, or challenge them to provide a counter example. Encourage them to draw upon all the sequences that they have worked with in previous activities. A counter example would be a sequence that cannot be changed into the identity sequence by inversions.

**Algorithms**

Mathematics and science make use of a variety of **algorithms.** An algorithm is a step-by-step set of rules used to solve a given problem in a finite number of steps. Following an algorithm is often more successful and less frustrating than using trial and error.

## *ACTIVITY* 2-1 Paper Folding Algorithm

**Objective**: Follow the steps of an algorithm
**Materials**:
Handout GI-H5: Paper Folding Algorithm Activity Worksheet
Paper, Scissors, Ruler

Try to follow this simple origami algorithm:

1. Cut a circle out of a piece of paper.

2. Draw a diameter of the circle and label the endpoints A and B.

3. Fold the paper so that point A touches the center of the circle.

4. Fold the paper along the chord formed by the point B and the point where the previous fold meets the circumference (on either side)

5. Repeat step 4 from point B to the point where the previous fold meets the circumference on the other side. Your circle has now become what shape?

6. Fold one vertex of the triangle to what was the center of the circle.

7. Repeat step 6 for the other two vertices of the triangle. Now what has your circle become?

Notice that this algorithm is very general. It will work the same way on a circle of any size. The algorithm wouldn't be very useful if it had to be changed for circles of different sizes. Good algorithms are general, efficient, and precise—and they always work, if followed correctly!

**Questions for Discussion**

3. What are the key elements of a good algorithm?

4. What are some examples of algorithms you have used in the past?

**Practice**

1. Now that you have followed an algorithm, you will create one for someone else to follow. Take another circle of paper and fold it into some new shape of your own design. As you fold the shape, write down precise directions for each step so that someone else could follow them and make the same shape that you made—without knowing what it is ahead of time!

2. After your algorithm is complete, exchange algorithms with a partner—but do NOT let your partner see your completed shape! Your partner must follow your algorithm as best as possible WITHOUT seeing your shape, and WITHOUT any help or hints from you. When you are each done following each other's algorithms, compare shapes and discuss how successful you were at both writing and following your algorithms.

3. Write an inversion algorithm to transform any given sequence into its identity sequence. Your algorithm should work on sequences of any length and transform them into the sequence: 1 2 3 4 5 ... *n* (*n* is the largest number in the given sequence.) Make sure your algorithm is general, precise, and most of all, make sure it works!

## Lesson 3 An Improved Inversion Algorithm

Are all inversions algorithms the same? If not, are some algorithms more efficient than others? What kind of algorithm would be most useful for studying how different species' chromosomes are related? In this lesson we address these questions.

### *ACTIVITY 3-1* Comparing Algorithms

**Objectives:** Identify important characteristics of an algorithm.
Compare and evaluate algorithms.

**Materials**:
Poster board or chart paper (optional)
Handout GI-H6: Comparing Algorithms Worksheet

1. Participants within the group exchange inversion algorithms from Lesson 2 Practice item 3. Participants individually apply their new algorithm to transform the sequences below to the identity sequence. Record each intermediate step and the number of inversions required to reach the target sequence.

2 3 4 5 6 1 4 3 6 5 2 1 1 2 3 5 4 6

2. Identify a few things you like about the algorithm and a few things that you might improve.

3. As a group, discuss all of the algorithms and choose the one you think is the best. What makes it better than the others? What is your definition of best?

4. Revise your group's "best" algorithm to make it even better. Perhaps you can combine ideas from several of your group's algorithms to produce a new improved algorithm. Be prepared to present and test your group's algorithm.

**How Many Inversions?**

In the mathematical process of **optimization** we usually try to either maximize or minimize some entity. An algorithm that efficiently transforms any sequence into the identity sequence using a minimum number of inversions is an optimal algorithm for comparing chromosomes. Before we look for the minimum number of inversions, let's bound our problem by considering the maximum number of inversions that might be required.

As you think about the maximum number of inversions required, use a table such as the one below to organize and record your data from the inversions you have completed in Lessons 2 and 3.

| **Sequence Length (*n*)** | **2** | **3** | **4** | **5** | **6** | ***n*** |
|---|---|---|---|---|---|---|
| **Maximum # of Inversions** | | | | | | |

**Questions for Discussion**

1. What is the maximum number of inversions required to transform any given sequence to the identity sequence or a target sequence?

2. Justify your answer to #1.

3. One of your friends completes a subsequence inversion on the sequence 1 5 3 4 2 6 using 7 inversions. Is this possible? Explain.

**Good, Better, Best**

In looking for a more efficient algorithm, we consider the concepts of **breakpoints** and **strips**. A breakpoint in a sequence of *n* numbers occurs at the following places:

- before the first element of a sequence, unless the first element is 1
- after the last element of a sequence, unless the last element is *n*
- between any two nonconsecutive numbers (examples: 4 7 or 5 1)

The sequence 3 4 8 7 6 1 2 5 has five breakpoints, the sequence 3 2 5 1 4 6 also has five breakpoints and the sequence 3 4 7 6 5 1 2 8 has four.

Determine the breakpoints in the sequences below. The first one is done for you.

|3 4|8 7 6|1 2|5|  3 25146  3 47 6 51 28

Breakpoints separate a sequence into subsequences called **strips**. A strip is labeled **increasing** if it contains two or more elements that increase by 1 when read from left to right (examples: 3 4 5 or 1 2). A strip is labeled **decreasing** if it contains two or more elements that decrease by 1 (example: 7 6 5 4 or 4 3). A strip with exactly one element and appearing at the beginning or end of the sequence is considered an **increasing** strip. All other strips with exactly one element are considered **decreasing** strips.

For example, reading left to right, the first sequence below contains 3 increasing strips & 1 decreasing strip. Label the decreasing and increasing strips in the sequences below with a D or an I. The first one is done for you.

I D I I
|3 4|8 7 6|1 2|5|  3 25146  3 47 6 51 28

We use the concepts of breakpoints and strips to develop a new algorithm that will specify which inversions to perform according to a series of rules. For example, in the first sequence above, the algorithm would instruct us to invert the sequence consisting of the strips 1 2 and 5 to yield the new sequence | 3 4 | 8 7 6 5 | 2 1 | which is closer to the identity sequence because it has one fewer breakpoint.

## Improved Inversion Algorithm

**1**. Mark all the breakpoints in the sequence and label the strips decreasing or increasing. Stop when the sequence is the identity sequence (1 2 3 4 5 … *n*). It will have no breakpoints.

**2**. If there is at least one decreasing strip, find the decreasing strip with the smallest number. Call this number *x* and do *one* of the following, considering the options in order:

a. If *x* is 1 and is not in the first position, invert the subsequence beginning with the first position and ending with 1. This inversion will put 1 in the first position. Return to step 1 with your new sequence.

b. If (a) is not possible, invert the strip (or group of adjacent strips) that results in *x* and *x* – 1 being adjacent. Note: sometimes you will invert a subsequence ending with *x* and other times you will invert a subsequence ending with *x* –1. Return to step 1 with your new sequence.

**3**. If there is no decreasing strip, create one by doing ***one*** of the following, considering the three options in order:

a. If 1 is not the first element, invert the first increasing strip (of length 2 or more). Return to step 1 with your new sequence.

b. If (a) is not possible, and *n* is not the last element, invert the last increasing strip (of length 2 or more). Return to step 1 with your new sequence.

c. If (b) is not possible, invert the strip between the first and second breakpoints located after 1. Return to step 1 with your new sequence.

**Notes**: The subsequences you will be inverting will always consist of one or more complete strips as defined by breakpoints. In following the algorithm, never create a new breakpoint by breaking a strip and do not perform an inversion of a subsequence unless the subsequence consists of one or more complete strips.

**Practice**

Use the Improved Inversion Algorithm to transform the sequences below into the identity sequence. Mark the breakpoints and label the increasing and decreasing strips. Also, next to each new sequence, write down which step of the algorithm is being applied during each inversion (e.g. 2a or 3b).

1. 5 4 3 1 2

2. 5 6 1 2 3 4

3. 4 6 3 2 1 5

4. 1 3 4 5 2 6

5. 1 4 6 7 5 3 2

6. 1 4 5 2 3 6

7. How does the algorithm affect the number of breakpoints?

8. How often are the different steps (e.g. 2b or 3a or 3c) used? Are some used more than others?

9. What questions do you have about the algorithm?

## Lesson 4 Using and Analyzing the Improved Inversion Algorithm

Was the improved algorithm more efficient in inverting sequences? Is it possible to use the algorithm and perform different inversions than another student? Compare your homework solutions with another student.

**Improved Algorithm**

## *ACTIVITY 4-1* Using the Improved Algorithm

**Objectives:** Identify breakpoints and increasing and decreasing strips. Apply the improved inversion algorithm.

**Materials:**

Handout GI-H7: Using the Improved Algorithm Activity Worksheet
Lesson 3 Practice problems

1. Compare your homework solutions and identify any errors and correct them by explaining any differences in how you and your partner used the algorithm.

2. Complete the table below, by inserting breakpoints, labeling strips as increasing or decreasing, and counting strips. Make up two sequences of your own that you think are interesting and write them in rows E and F.

| Sequence Name | Initial Sequence Add Breakpoints & Labels | # of Increasing Strips | # of Decreasing Strips |
|---|---|---|---|
| A | 1 4 3 2 | | |
| B | 1 2 3 4 5 | | |
| C | 4 3 5 2 1 6 | | |
| D | 3 4 6 5 12 8 9 11 7 1 2 10 | | |
| E | | | |
| F | | | |

**Questions for Discussion**

1. In step 2b of the improved algorithm, $x$ and $x - 1$ become adjacent. What would happen if $x$ and $x - 1$ were adjacent to begin with?

2. Explain why step 2 always decreases the number of breakpoints by at least 1. In what case(s) does step 2 decrease the number of breakpoints by more than 1? More than 2?

3. What is the largest number of breakpoints that one step can eliminate?

The **lower bound** is the smallest possible number of inversions necessary to invert a sequence. In mathematics we use a bracket notation with the bottom part of the bracket missing to represent the **ceiling function**. The ceiling function means "the closest integer greater than or equal to $c$". For example, if $c = 5$, then $\lceil c \rceil = 5$. And if $c = 3.5$, then $\lceil 3.5 \rceil = 4$. It is a way of rounding up to the nearest integer, if a number is not already an integer.

If $b$ is the number of breakpoints, explain why the lower bound is equal to $\lceil b/2 \rceil$.

**Practice**

Use the Improved Inversion Algorithm to transform the following sequences into the identity sequence. Note the total number of breakpoints at each step.

1. 5 1 2 3 4 6

2. 4 5 2 1 6 3 8 7 9

3. 1 3 2 5 7 4 6 8

4. 1 2 3 6 7 8 4 5

5. Calculate the lower bound and upper board for the inversion distance in each sequence below. The upper bound is the largest possible number of non-repetitive inversions needed.

a. 3 2 5 6 7 4 1 8

b. 3 2 5 7 4 6 8 1

c. 4 5 2 1 6 3 8 7 9

d. 3 2 1 5 4 8 9 7 10 12 6 11

**Evolutionary Separation**

Knowing the minimum number of inversions that separate two chromosomal sequences and inferring that this mathematical separation is also the biological **evolutionary separation** between the chromosomes allows us to construct the possible **phylogenetic trees** that connect the two species. A phylogenetic tree is a branching or tree diagram showing an inferred evolutionary relationship among species.

In constructing a phylogenetic tree, the nodes in the tree represent the organisms and a line connects two nodes only if the organisms are separated by one genetic inversion. The tree is oriented such that the most recently appearing organisms are located at the bottom of the tree; so

as one moves up a tree one is moving into the evolutionary past. In this unit, we do not pay attention to the lengths of the lines. The lengths of the lines represent time and would require additional information about the species to properly specify their exact positions on a timeline.

The tree with two species can only be drawn in one way with one node above the other as in the upper left of Figure 4.1 below. Trees with three or more nodes could be represented by a variety of trees as shown below.

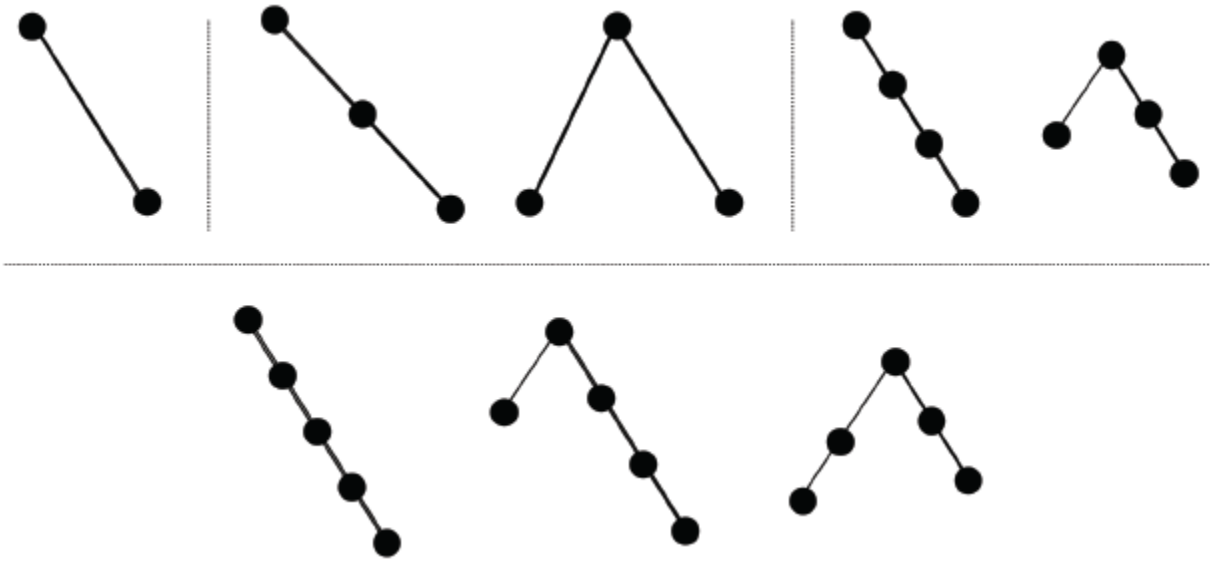


**Figure 4.1:** Possible relative arrangements of 2, 3, 4, and 5 nodes in a phylogenetic tree. (Note: The lengths of the lines are not to scale and are not important in this unit.)

The key to inferring which tree most likely represents the actual evolutionary relationship of the organisms involved depends upon what is known about the organisms and upon the relative chances of a series of mutations all leading to one single lineage or to a tree with multiple branching points.

Consider the two sequences: 4 3 1 2 5 and 1 2 3 4 5. Using the Improved Algorithm reveals that just two inversions connect these two sequences: 4 3 1 2 5 → 4 3 2 1 5 → 1 2 3 4 5. There are three possibilities for how the species represented by these sequences might be related. Two trees would have no branching points (left two trees in Figure 4.2) and would depend only on which species appeared first in evolutionary history. The other tree would have two twigs and one branching point (right tree in Figure 4.2).

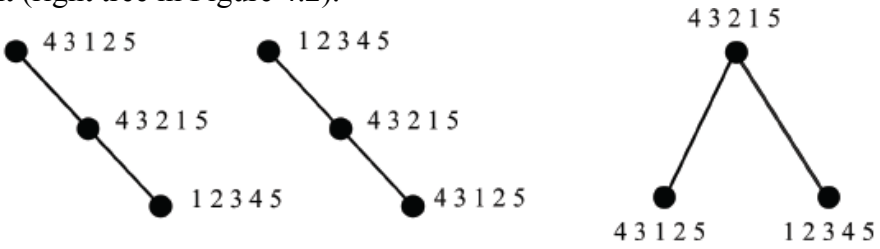



**Figure 4.2:** Possible Trees for Example Sequence

Note that placing 4 3 2 1 5 at either end of a tree makes no sense because 4 3 2 1 5 is the node that connects each of the other two nodes by one inversion. Linking 1 2 3 4 5 and 4 3 1 2 5 violates our principle of assuming that only nodes separated by a single inversion should be joined by a line.

The tree at right in Figure 4.2 is the most likely relationship between the species. It is more likely that one species gave rise to two viable other species, each differing from the ancestor by one inversion (or mutation), than that one species gave rise to another which added an additional mutation and still yielded a viable species. In this simplistic model, it is important to note that *either* tree potentially could be the correct phylogeny; in a more complete and complex model, biologists would use additional data about the animals to gain confidence in which phylogeny they believe to be correct.

Note also that in addition to species 1 2 3 4 5, which is always assumed to be a living or extant, as we also assume about the species 4 3 1 2 5, the linking species 4 3 2 1 5 might be living or extinct. Whether this intermediate species is extant or extinct is independent of which of the phylogenetic trees we assume to be correct. We would have to find information on whether species are extinct or not from other sources.

**Practice**

Transform each of the following sequences into the identity sequence and draw and label the most likely phylogenetic trees for the resulting sequences. Assume that the identity sequence is the most recently evolved living species.

6. 1 2 3 6 5 4

7. 5 4 3 1 2

**Glossary**

**Algorithm** – a step-by-step set of rules used to solve a given problem in a finite number of steps.

**Breakpoint** – a location in a sequence between two non-consecutive numbers, at the beginning of a sequence when the first number is not in the first place or at the end of a sequence when the last numbers is not in the last place.

**Ceiling function** – a mathematical function denoted by $\lceil x \rceil$ that gives the smallest integer greater than or equal to *x*. In other words, the function rounds x up to the nearest integer ≥ *x*.

**Cell** – smallest unit of life; controlled by the genetic information contained within DNA.

**Centromere** – region of a chromosome to which the spindle fibers become associated during cell division. Often constricted in mitotic chromosomes. The location of the centromere can be useful in determining if an inversion has occurred.

**Chromosome** – a long segment of DNA that is coiled and packed around proteins; a portion of the genome carrying many genes.

**DNA** – deoxyribonucleic acid; the double strand helix shaped molecule in living organisms that contains inherited information made up of a sequence of nucleotides referred to as A, T, C, G.

**DNA replication** – the process whereby the genetic information is duplicated. Each strand in double-stranded DNA serves as a template for the construction of a complimentary strand.

**Evolution** – a change over time; biologically speaking, it is the production of new forms of life over time.

**Evolutionary separation** – the number of mutations that separate different species.

**Evolutionary phylogeny** – a reconstruction, usually in a tree model, of a possibility of the evolutionary relationships among species.

**Gene** – a unit of hereditary information located on a chromosome and composed of DNA nucleotides that code for a protein; several genes line up in a particular order on one chromosome.

**Genome** – the complete DNA information for an organism; all the chromosomes of an organism.

**Identity (or normal) sequence** – the sequential ordering of a group of numbers beginning with 1.

**Initial sequence** – a given or starting sequence prior to any inversions or mutations.

**Inversion** – (mathematics) the reversing of a subsequence within a sequence, or the reversing of the entire sequence; (biology) when a single gene or a group of genes detach from DNA, rotate 180°, and reattach.

**Lower bound** – the least possible value of some unknown quantity.

**Mathematical model** – a mathematical representation of a real-life situation or problem.

**Mutation** – a change in the DNA of an organism; the result may be beneficial, harmful, or have no effect on the organism.

**Optimization** – finding the best solution; minimizing or maximizing some quantity or quality subject to other constraints.

**Phylogenetic trees** – a branching diagram that represents possible evolutionary relationships among species.

**Protein** – an organic compound made of one or more chains of amino acid subunits; the order of amino acids is coded for by the order of nucleotides in DNA.

**Sequence** – a listing of numbers or objects in a specified order.

**Sex cell** – a spermatozoon from a male or an ovum from a female (sperm or egg).

**Species** – in terms of classifying organisms, a group of organisms that can interbreed and produce fertile offspring.

**Strip** – a subsequence of a sequence.

**Subsequence** – a sequence of two or more adjacent items that are a subset of a larger sequence or the entire sequence itself.

**Target sequence** – the final desired sequence of a series of inversions.

**Upper bound** – the largest possible value of some unknown quantity.

**Zygote** – a fertilized egg produced by the union of male and female reproductive cells; it will develop into an embryo.

## References

**[1]** Alberts, B., Johnson, A., Lewis, J., et al. (2002). *Molecular biology of the cell* (4th ed.). New York: Garland Science. Found at http://www.ncbi.nlm.nih.gov/books/NBK21077/.