ASSEMBLING A SEQUENCE

A HIGH SCHOOL ACTIVITY IN CONTIGUOUS DNA SEQUENCE CONSTRUCTION

Michael Grupe
Lutheran High School North
St. Louis, Missouri

2005 Summer Research Fellowships for Science Teachers

Sponsored by
Howard Hughes Medical Institute
Washington University Science Outreach

ASSEMBLING A SEQUENCE
Teacher Manual

LESSON OVERVIEW

Assembling a Sequence is designed to help students better visualize how a large contiguous DNA sequence can be constructed from smaller overlapping DNA sequences. This activity illustrates one of the last steps in sequencing projects such as the Human Genome Project. Through readings, a hands-on paper activity, and completion of a worksheet, students will become more familiar with the concept of the construction of a contiguous sequence. They will also become aware of some of the potential problems inherent in this process. Students will read about polymerase chain reactions and fluorescent-labeled dye terminator reactions. They will organize DNA fragments by comparing nucleotide sequences and will determine which fragments best give complete coverage of this part of the genome.

TIMELINE

The background information should be read before the activity, assigned as homework or read as a class. The paper model activity requires about 30 minutes to complete. The questions should be completed when the paper model activity is completed. The questions can be assigned as homework or can be discussed by the class when everyone has completed the paper model activity.

MATERIALS

Per group of 2-4 students:

1 page of DNA sequences (either sequences 1-10 or 11-20)
4 pages of legal size paper or 5 pages of 8.5” x 11” paper
Scissors
Tape or glue

ADVANCE PREPARATION

You will need to print one page of DNA sequences for each team. Sequences 1-10 are from the first half of the contiguous sequence. Sequences 11-20 are from the last half of the contiguous sequence. Give half the class the 1-10 sheet and the other half of the class the 11-20 sheet. Later have each team with a completed sequence locate a team that has the adjacent contiguous sequence.

1	tggtggtctacccttggacccagaggttctttgagtcctttggggatctgtc
2	gggcaaccctaaggtgaaggctcatggcaagaaagtgctcggtgccttt
3	tgccgttactgccctgtggggcaaggtgaacgtggatgaag
4	aacagacaccatggtgcatctgactcctgaggagaagtctgccgttac
5	gcaaggtgaacgtggatgaagttggtggtgaggccctgggcaggctgctggtggt
6	aagtgctcggtgcctttagtgatggcctggctcacctgga
7	ggacccagaggttctttgagtcctttggggatctgtccactcctgatg
8	atggtgcatctgactcctgaggagaagtctgccgttactgccctgtggg
9	gaaggctcatggcaagaaagtgctcggtgcctttagtgatggcctggctcacctgg
10	ggggatctgtccactcctgatgctgttatgggcaaccctaaggtgaa

11	ggctggtgtggctaatgccctggcccacaagtatcactaagctcgctttct
12	gctaatgccctggcccacaagtatcactaagctcgctttcttgctg
13	gtgcaggctgcctatcagaaagtggtggctggtgtggctaatg
14	gtgctggcccatcactttggcaaagaattcaccccacc
15	tggctcacctggacaacctcaagggcacctttgccacactgagtg
16	tggatcctgagaacttcaggctcctgggcaacgtgctggtctgtgtg
17	ggcacctttgccacactgagtgagctgcactgtgacaagctgcacgtggatcc
18	agtgagctgcactgtgacaagctgcacgtggatcctgagaacttcaggctcct
19	ggcccatcactttggcaaagaattcaccccaccagtgcaggctgcctatcaga
20	ctgggcaacgtgctggtctgtgtgctggcccatc

HINTS AND TROUBLESHOOTING

- At some point you might want to discuss with the class the reason for requiring an overlap of at least 5 nucleotides. The probability of a particular 5-base sequence existing randomly is much less than the probability of a particular 2- or 3-base sequence existing randomly. (The probability of a particular 2-base sequence is 1 out of 16. The probability of a particular 3-base sequence is 1 out of 64. The probability of a particular 5-base sequence is 1 out of 1024.) It is not reasonable to think that two fragments overlap because they each contain the same 2- or 3-base sequence.

- The time required for the paper model activity will decrease as the number of students in the team increases.

- Students should not fix their sequences to the paper background until all ten sequence locations have been identified. Only then will they know how to arrange the sequences on the background paper.

ADDITIONAL RESOURCES

- A high school biology genetics curriculum is available at http://www.so.wustl.edu/science_outreach/curriculum/genetics.html

- More information about sequencing a genome is available at: http://www.nslc.wustl.edu/elgin/genomics

- Micklos, David A., ed. DNA Science: A First Course. Second Edition. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2003.

ADDITIONAL INFORMATION

There are major differences between sequence assembly in a real genomic sequencing project and what students will do in this activity. Sequences that must be aligned are actually about ten times longer than the sequences that students will work with in this activity. The sequences have been made shorter to make it easier for students to recognize overlapping regions. Also the students will make a visual comparison of the sequences in order to construct a contiguous sequence. In real assembly work, computers look for overlapping regions because it would be so difficult for a person to find common sequences in such large fragments of DNA.

Despite these differences, students should be able to better visualize how long sequences are constructed from shorter fragments. The overlapping sequences can also be used to illustrate how sequencers choose a "golden pathway" of fragments to cover the entire length of the genome in the initial stages of sequencing a genome.

The sequence that was used for this activity is the normal human betaglobin gene. The sequence includes 10 bases upstream of the start DNA codon (ATG) and 16 bases downstream of the DNA stop codon (TAA). The questions at the end of the activity include the opportunity for the class to focus more on the features of this gene and the mutation in the seventh codon that leads to a glutamic acid to valine change in betaglobin. The result of this amino acid change is the genetic condition sickle cell anemia.

SOLUTIONS

The correct order of the first ten fragments is:

Position #	1	2	3	4	5	6	7	8	9	10
Fragment #	4	8	3	5	1	7	10	2	9	6

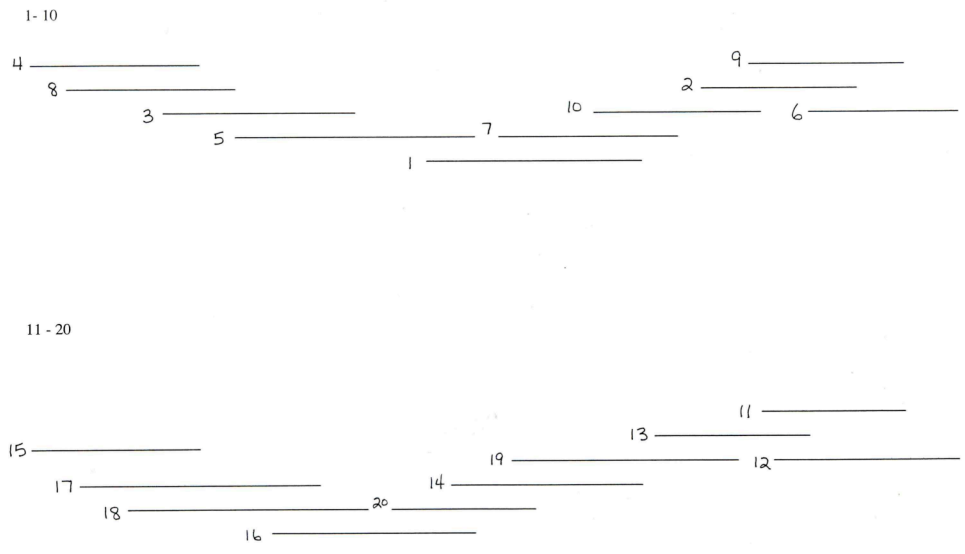
The correct order of the second ten fragments is:

Position #	1	2	3	4	5	6	7	8	9	10
Fragment #	15	17	18	16	20	14	19	13	11	12

These two sets of contiguous sequences overlap at the following base sequence:

t g g c t c a c c t g g a

The fragments should be fixed to the background paper as shown below.



ANSWERS TO DISCUSSION QUESTIONS

1. In this activity we consider a matching 5-base sequence on two strands to be a reliable overlap. Why? To answer this question let's calculate some probabilities.

    a. What is the probability that any 2-base sequence in one DNA strand will be identical to any 2-base sequence in a second DNA strand? To answer this, multiply the probability of the first bases of two 2-base sequences being the same (1/4) times the probability of the second bases of the 2-base sequences being the same (1/4).

    1/16

    b. What is the probability that two strands will contain the same 3-base sequence? Since this is a 3-base sequence, you must multiply the probability that the bases in each of the three positions will match.

    1/64

    c. Finally, what is the probability that two strands will contain the same 5-base sequence?

    1/1024

    d. A typical sequencing reaction will produce at least 300 bases of good sequence. Let's say you are comparing two strands that contain 320 different 2-base sequences. How many matching 2-base sequences would you expect to find in these two strands? (Multiply 320 by the probability you calculated in question 1a.)

    20

    e. How many matching 3-base sequences would you expect to find in two strands that contain 320 different 3-base sequences? (320 x 1b probability)

    5

    f. How many matching 5-base sequences would you expect to find in two strands that contain 320 different 5-base sequences? (320 x 1c probability)

    3/8 probability

    g. Is it likely that you would randomly find even one matching 5-base sequence in two strands that are not overlapping strands?

    No is it unlikely that you would find even one matching sequence.

Obviously, overlapping strands have the same sequence because they are the same sequence. We look for at least 5-base sequences that match because it is unlikely that two strands of about 300 bases would have matching 5-base sequences just by random chance. Does this mean that two strands are definitely overlapping fragments if they have at least 5 bases in a row that match?

2. How do you know when you have found the end of a contiguous sequence?

   There will be no fragments that have any matching sequence with either end of a contiguous sequence.

3. What was the number order of fragments in your contiguous sequence?

   Strands 1-10 4 8 3 5 1 7 10 2 9 6

   Strands 11-20 15 17 18 16 20 14 19 13 11 12

4. Which fragments would you use to read the entire sequence with 1X coverage? (Avoid including fragments that are unnecessary in determining the sequence.)

   4 3 5 1 10 2 6 (8, 7 and 9 are not needed for 1X coverage.)

   15 17 16 20 19 13 12 (18, 14, and 11 are not needed for 1X coverage.)

5. The following questions focus on the gene that was used for this activity. (Look at the sequence containing strands 1-10.) This gene codes for the production of human beta-hemoglobin (betaglobin). This polypeptide is part of the larger hemoglobin molecule that is found in blood.

   a. Copy the first 20 bases of the DNA contiguous sequence that was made from strands 1-10. Find the DNA start codon (a t g) and put a square around it.

      a a c a g a c a a c c **a t g** g t g c a t

   b. Use a translation table to determine the first eight amino acids that are coded for by this gene. You need to change t to u because your translation table translates mRNA codons. You may use the accepted amino acid abbreviations found in the following table:

G- Glycine (Gly)	**P**- Proline (Pro)	**A**- Alanine (Ala)
V- Valine (Val)	**L**- Leucine (Leu)	**I**- Isoleucine (Ile)
M- Methionine (Met)	**C**- Cysteine (Cys)	**F**- Phenylalanine (Phe)
Y- Tyrosine (Tyr)	**W**- Tryptophan (Trp)	**H**- Histidine (His)
K- Lysine (Lys)	**R**- Arginine (Arg)	**Q**- Glutamine (Gln)

N- Asparagine (Asn) **E**- Glutamic Acid (Glu) **D**- Aspartic Acid (Asp)
S- Serine (Ser) **T**- Threonine (Thr)

M(Met) V(Val) H(His) L(Leu) T(Thr) P(Pro) E(Glu) E(Glu)

c. The genetic disorder sickle cell anemia is caused by a point mutation in the first part of the gene. This DNA sequence contains the mutation. Locate it and describe what point mutation has occurred.

a t g g t g c a t c t g a c t c c t g **t** g g a g

Base #20 is t instead of a. A substitution has occurred.

d. What is the effect of this mutation on the betaglobin polypeptide?

Glutamic acid in the seventh amino acid position is changed to valine.

e. Do some research to find the function of hemoglobin in humans.

Hemoglobin is carried by red blood cells. Hemoglobin is an oxygen acceptor. It picks up oxygen in the lungs and drops oxygen off as blood travels through body tissue capillaries.

f. Do some research to determine how a change of a single amino acid can have such an extreme effect on a polypeptide chain.

Glutamic acid is an acidic amino acid. Valine is a neutral, non-polar amino acid. The tertiary structure (3-D shape) of a protein is caused by its amino acid sequence. Different amino acid sequences lead to different protein conformations because of the interactions between the amino acids. A change in one amino acid in a polypeptide can change some of these interactions which would in turn change the three dimensional structure of the protein. A change in protein shape may lead to a change in protein function.

g. Do some research to learn about sickle cell anemia. What symptoms does it cause?

Hemoglobin molecules with this mutation cluster together, causing red blood cells to take on a rigid, sickled shape. These cells cannot pass through capillaries as easily because of their loss of flexibility. Small clots often form that prevent blood flow through capillary beds. The symptoms may include fatigue, breathlessness, rapid heart rate, delayed growth and puberty, susceptibility to infections, ulcers on the lower legs, jaundice, attacks of abdominal pain, weakness, joint pain, fever, vomiting, bloody urination, excessive thirst, chest pain and decreased fertility. Eventually organs may be damaged by repeated sickle cell crises.

ASSEMBLING A SEQUENCE

Student Manual

BACKGROUND INFORMATION

An organism's genome is all of the DNA in a somatic cell of that organism. The human genome is contained in 23 pairs of chromosomes—22 autosomal pairs and 1 pair of sex chromosomes, and in the DNA found inside the mitochondria. One set of human chromosomes contains a tremendous amount of DNA—over 3 billion nucleotide pairs. Even the smallest genomes, those of bacteria, contain no fewer than half a million base pairs.

Sequencing the genome of an organism is a very complex task, given the huge amount of DNA found in a cell. First, a genome must be cut by restriction enzymes into pieces that are from about 50,000 to 250,000 base pairs (50 kb to 250 kb) in length. These restriction fragments are cloned into bacterial artificial chromosomes (BACs). BAC fragments are mapped to identify the position of each fragment relative to all other fragments. Analysis of these fragment maps ensures that every part of the genome is being selected for sequencing. The fragment in a BAC is randomly fragmented into smaller pieces for sequencing. These fragments are about 2 to 3 kb in length.

Sequencing of a DNA fragment can be done by producing copies of portions of the fragment using the polymerase chain reaction (PCR). As copying proceeds from a primer, fluorescent dye labels are randomly added to the end of the growing fragment copy, stopping the reaction. Each of the four different terminator nucleotides is labeled with a different color. If enough copies are made, every possible length of fragment copy will be produced, from a copy that is a few bases long all the way up to a full-length copy of the fragment. The fragment copies run on a sequencing gel past a laser light, the shortest fragments passing the laser first followed by larger and larger fragments. The laser causes each terminator dye to fluoresce and the color of the emitted light is interpreted. An emitted green light signals an adenine terminal nucleotide, red signals thymine, blue signals cytosine and yellow signals a terminal guanine. A computer printout displays the entire sequence of nucleotide bases in the fragment being sequenced. (If you want to know more about this process you can go to http://www.nslc.wustl.edu/elgin/genomics and click on "Sequencing a Genome." Go to the Video Tour and select Production: Part II. About two minutes into Part II is an explanation of what happens in sequencing.)

Generally a sequence of 300-400 bases can be read with a high degree of reliability. A computer aligns each fragment sequence with other fragment sequences to make longer and longer sequences. Since there are two strands to a DNA molecule each strand is sequenced (from opposite ends of the double strand) in hopes of covering a 2-3 kb fragment from end to end. Unfortunately, not all sequence data is easily read and not all regions of a 2-3 kb fragment provide sequence data. Certain regions may have to be sequenced again to make sure that all gaps have been filled in and that all sequence data is accurate and reliable.

ACTIVITY OVERVIEW

There are two parts to this activity. First you will play the role of a computer that aligns fragment pieces by comparing their sequences. Then you will align your aligned sequences with another team's aligned sequences. From these aligned sequences you can read the entire sequence of the gene that was selected for this activity. Hopefully you will begin to get the idea of larger and larger contiguous sequences taking shape as more and more of the genome is sequenced.

MATERIALS

Per group of 2-4 students:

1 page of DNA sequences (either sequences 1-10 or 11-20)
4 pages of legal size paper or 5 pages of 8.5" x 11" paper
Scissors
Tape or glue

PROCEDURE: PART 1

1. Tape or glue the blank sheets of paper end to end. This will be the background to which you will attach your completed DNA sequence. Write the names of all teams members in the upper right corner of this sheet.

2. Cut out the ten DNA sequence strips from the page provided by your teacher.

3. Look for regions of two strips that contain identical sequence. When you find two such strips, lay them on your desk or table so that the matching sequences of the two strips are aligned. (You should have a minimum of five matching bases to consider the two sequences a reliable overlap.)

4. Continue to add additional strips to the growing alignment until all ten strips have been used.

5. When all ten strips have been placed in the correct position, tape or glue the ten strips to the background paper that you constructed in step 1. Plan how you will attach the ten strips so that:
    - The completed arrangement fits on the background paper
    - Matching sequences of adjacent strips line up vertically

PROCEDURE: PART 2

6. When fragment sequences have been aligned, it should be possible to locate these contiguous sequences with other contiguous sequences in order to build larger and larger sequences. Other teams in your class have sequenced a different BAC fragment that overlaps with your fragment. Find the overlapping region on your two fragments. You do not need to connect them.

7. Write the names of the members of your team on your sequence background paper.

8. Answer the questions assigned by your teacher.

Name _____ Class Hour _____ Date _____

DISCUSSION QUESTIONS

1. In this activity we consider a matching 5-base sequence on two strands to be a reliable overlap. Why? To answer this question let's calculate some probabilities.

    a. What is the probability that any 2-base sequence in one DNA strand will be identical to any 2-base sequence in a second DNA strand? To answer this, multiply the probability of the first bases of two 2-base sequences being the same (1/4) times the probability of the second bases of the 2-base sequences being the same (1/4).

    b. What is the probability that two strands will contain the same 3-base sequence? Since this is a 3-base sequence, you must multiply the probability that the bases in each of the three positions will match.

    c. Finally, what is the probability that two strands will contain the same 5-base sequence?

    d. A typical sequencing reaction will produce at least 300 bases of good sequence. Let's say you are comparing two strands that contain 320 different 2-base sequences. How many matching 2-base sequences would you expect to find in these two strands? (Multiply 320 by the probability you calculated in question 1a.)

    e. How many matching 3-base sequences would you expect to find in two strands that contain 320 different 3-base sequences? (320 x 1b probability)

    f. How many matching 5-base sequences would you expect to find in two strands that contain 320 different 5-base sequences? (320 x 1c probability)

    g. Is it likely that you would randomly find even one matching 5-base sequence in two strands that are not overlapping strands?

Obviously, overlapping strands have the same sequence because they are the same sequence. We look for at least 5-base sequences that match because it is unlikely that two strands of about 300 bases would have matching 5-base sequences just by random chance. Does this mean that two strands are definitely overlapping fragments if they have at least 5 bases in a row that match?

2. How do you know when you have found the end of a contiguous sequence?

3. What was the number order of fragments in your contiguous sequence?

4. Which fragments would you use to read the entire sequence with 1X coverage? (Avoid including fragments that are unnecessary in determining the sequence.)

5. The following questions focus on the gene that was used for this activity. This gene codes for the production of human beta-hemoglobin (betaglobin). This polypeptide is part of the larger hemoglobin molecule that is found in blood.

    a. Copy the first 20 bases of the DNA contiguous sequence that was made from strands 1-10. Find the DNA start codon (a t g) and put a square around it.

    b. Use a translation table to determine the first eight amino acids that are coded for by this gene. You need to change t to u because your translation table translates mRNA codons. You may use the accepted amino acid abbreviations found in the following table:

G- Glycine (Gly)	**P**- Proline (Pro)	**A**- Alanine (Ala)
V- Valine (Val)	**L**- Leucine (Leu)	**I**- Isoleucine (Ile)
M- Methionine (Met)	**C**- Cysteine (Cys)	**F**- Phenylalanine (Phe)
Y- Tyrosine (Tyr)	**W**- Tryptophan (Trp)	**H**- Histidine (His)
K- Lysine (Lys)	**R**- Arginine (Arg)	**Q**- Glutamine (Gln)
N- Asparagine (Asn)	**E**- Glutamic Acid (Glu)	**D**- Aspartic Acid (Asp)
S- Serine (Ser)	**T**- Threonine (Thr)	

c. The genetic disorder sickle cell anemia is caused by a point mutation in the first part of the gene. This DNA sequence contains the mutation. Locate it and describe what point mutation has occurred.

a t g g t g c a t c t g a c t c c t g t g g a g

d. What is the effect of this mutation on the betaglobin polypeptide?

e. Do some research to find the function of hemoglobin in humans.

f. Do some research to determine how a change of a single amino acid can have such an extreme effect on a polypeptide chain.

g. Do some research to learn about sickle cell anemia. What symptoms does it cause?