CAP 5510: Introduction to Bioinformatics CGS 5166: Bioinformatics Tools

Giri Narasimhan ECS 254; Phone: x3748 giri@cis.fiu.edu www.cis.fiu.edu/~giri/teach/BioinfF18.html

PCR and Sequencing



Polymerase Chain Reaction (PCR)

□ For testing, large amount of DNA is needed

Identifying individuals for forensic purposes

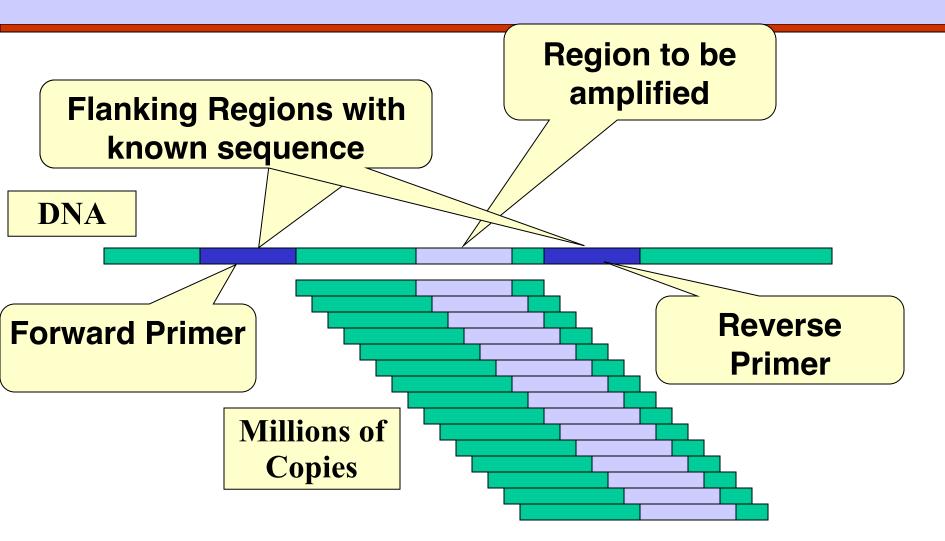
>(0.1 microliter of saliva contains enough epithelial cells)

Identifying pathogens (viruses and/or bacteria)

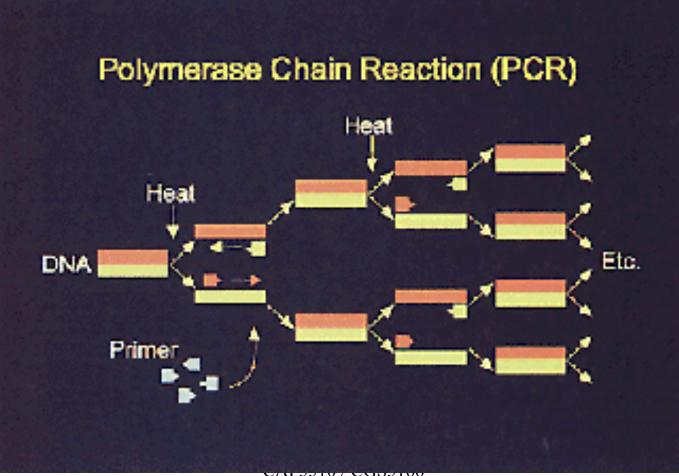
PCR is a technique to amplify the number of copies of a specific region of DNA.

- Useful when exact DNA sequence is unknown
- Need to know "flanking" sequences
- Primers designed from "flanking" sequences

PCR

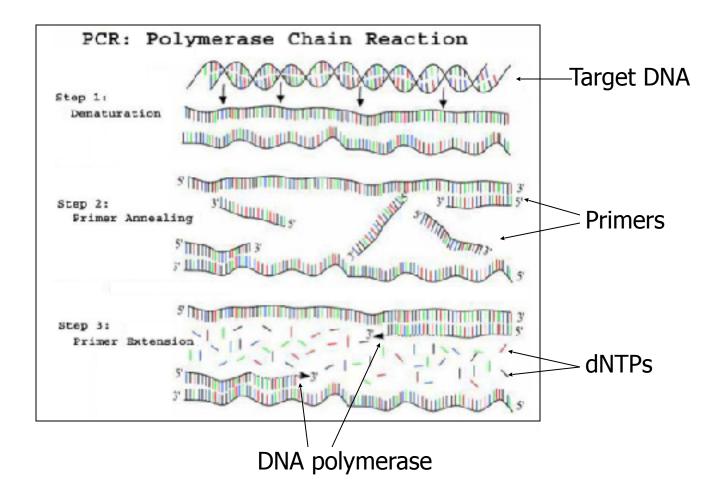


PCR

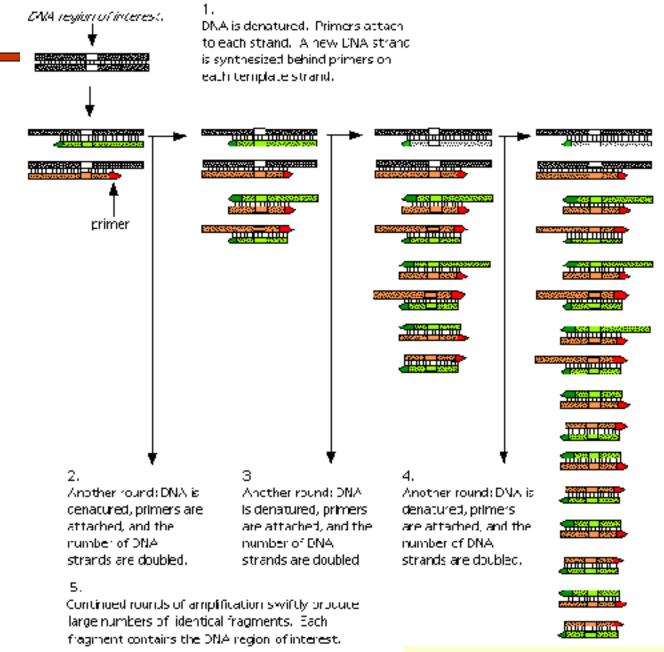


9/17/18

Schematic outline of a typical PCR cycle



POLYMERASE CHAIN REACTION

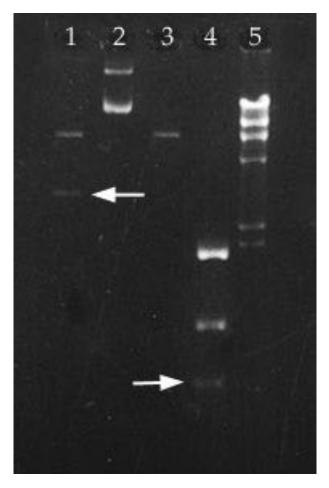


9/17/18

Gel Electrophoresis

Used to measure the lengths of DNA fragments.
 When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

Gel Pictures



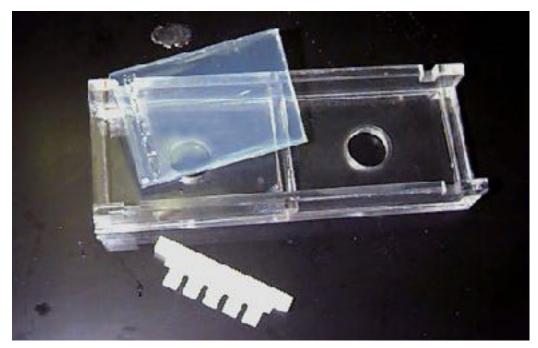
CAP5510 / CGS5166

Gel Electrophoresis: Measure sizes of fragments

- The phosphate backbone makes DNA a highly negatively charged molecule.
- □ DNA can be separated according to its size.
- Gel: allow hot 1% solution of purifed agarose to cool and solidify/polymerize.
- DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- Varying concentration of agarose makes different pore sizes & results.
- Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.

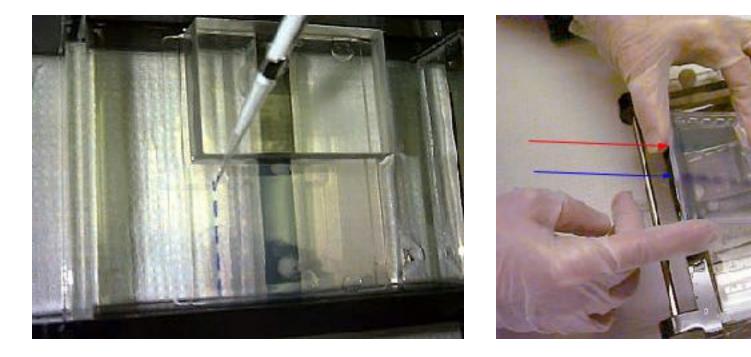
Gel Electrophoresis

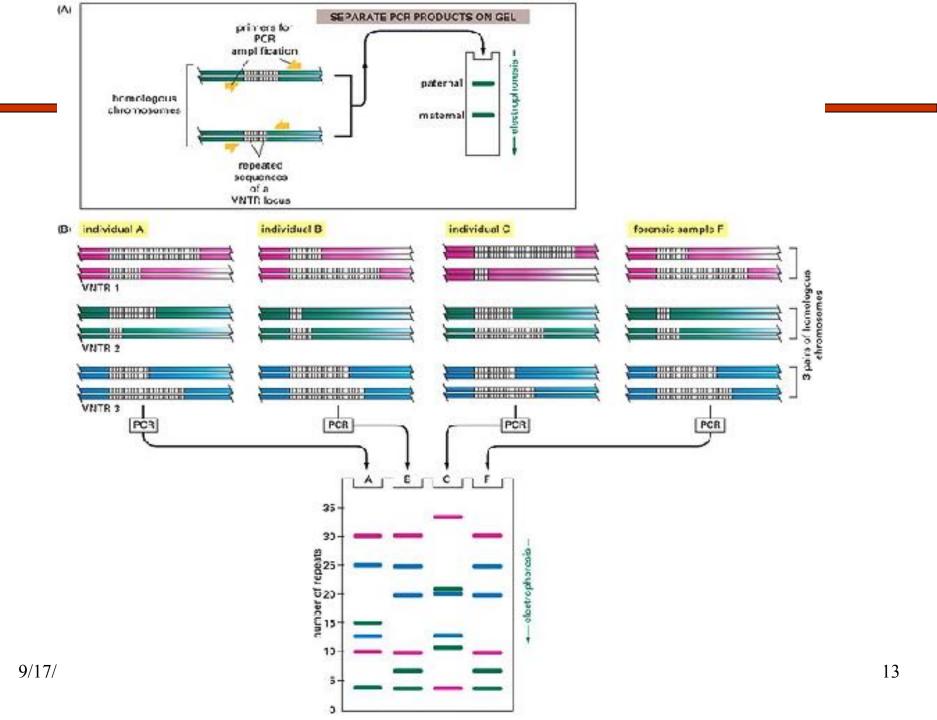




CAP5510 / CGS5166

Gel Electrophoresis





Sequencing



Why sequencing?

Useful for further study:

- Locate gene sequences, regulatory elements
- Compare sequences to find similarities
- Identify mutations
- Use it as a basis for further experiments

Next 4 slides contains material prepared by Dr. Stan Metzenberg. Also see: http://stat-www.berkeley.edu/users/terry/Classes/s260.1998/Week8b/week8b/node9.html

History

Two methods independently developed in 1974
 Maxam & Gilbert method
 Sanger method: became the standard
 Nobel Prize in 1980

Original Sanger Method

(Labeled) Primer is annealed to template strand of denatured DNA. This primer is specifically constructed so that its 3' end is located next to the DNA sequence of interest. Once the primer is attached to the DNA, the solution is divided into four tubes labeled "G", "A", "T" and "C". Then reagents are added to these samples as follows:

"G" tube: ddGTP, DNA polymerase, and all 4 dNTPs

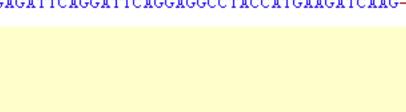
*A" tube: ddATP, DNA polymerase, and all 4 dNTPs

"T" tube: ddTTP, DNA polymerase, and all 4 dNTPs

"C" tube: ddCTP, DNA polymerase, and all 4 dNTPs

- DNA is synthesized, & nucleotides are added to growing chain by the DNA polymerase. Occasionally, a ddNTP is incorporated in place of a dNTP, and the chain is terminated. Then run a gel.
- □ All sequences in a tube have same prefix and same last nucleotide.
- http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/ Animations/DNA/WTDV026689.htm

9/17/18









5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG 3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG 3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

3 - - GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5

Example of sequences seen in gel from "G" tube:

5'-GAATGTCCTTTCTCTAAGTCCTAAG

Sanger Method

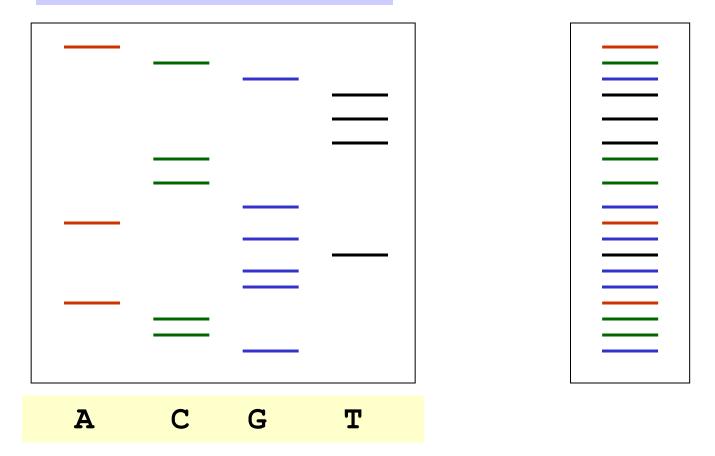
Modified Sanger

Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color dye



Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA



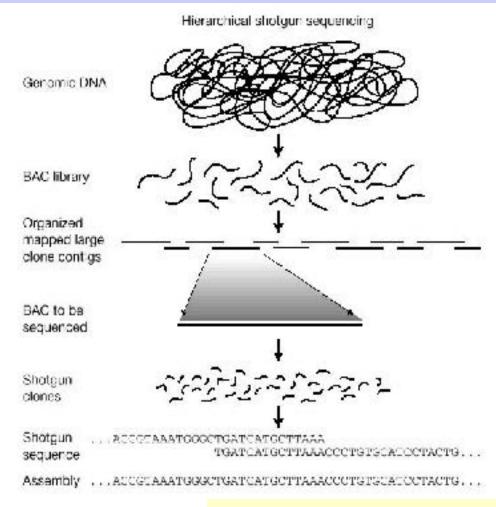
Sequencing

545 CHAPTER THIRTEEN Sequence Assembly and Finishing Mathicels

le grif bil yf wdwrei Godyne 1114-017 wrei	Link 95 Pilmin 0TS/30P0Pi((30x3).neb Cirminit: 177(01	Sporing: 15.05 Sporie C. 128: A 3854 (0.2538 7:3619 Base: 903	Sel sava Sel test rete: S. Le 2014 Gel dat time 4: 121
M.N.A.GECATOFCC	20 100 1 10 0 10 0 10 0 10 0 10 10 10 10		CA C CO SAAA AA & CTO TATAA
	0.00 0.00	1 1 1 1 1 1 1 1	
10 100 100	CN Devance Manual Col	and water water	surprise and
TA AA ACA A AG C CTA & CT N	R DAAT FRANKER ARRANGE AND TO AGE		TD # 9 0 A A A T 64 T T 16 6 A A
mmmmmm	hundrahuman	MWWWWWWWWWWWW	W. Monte Manal
180 190 190 190 190 190 190 190 190 190 19	1 TATATT 10 TO TO TO TO TA FARAGE	20 XM TGT GAC/ GT A TAFT TEAA A COA C TG 1	TTAAA AGAAAC TEAAATAA
in hard			
ana manimum	WWWWWWWWWWWWW	1940 MANAWAWAWAWAWAWAWAWAWAWAWAWAWAWAWAWAWAWA	MWWWWWWW
LAT TYALATA CO I TA IS	17AAAATAE 00 1 10 2014 013 40164	0 00 0 TAMAAA AGT 6 00 0 000 6A 400 6A	NO SHO AND ANA THEAT 2 CO
MAMANANAN	Innaria Aremanalarleava	Conservation and Conservation of the Conservat	M. Com Anna Man
AN AT TABOA JOAT CO F CT	S CAD & AD G AA + TO A AA & AD CA CT & 1	MAL MILLING FLAMMALLA FLAMMA	WAW TUTA VIANA MAN
		A A A A	04040 JOD JOD 400 2 JOB 2
WWWWWWW	WWWWWWWWWW	WAVAVAW/W/W/W/W/	Mannuhandhing
AD ADTO ANTO O OAC TOOL	40 600 400 400 400 400 ACTA NACA	CRACAT ATA AAATA ACCTTACTCCAT	AATO GALACE SH
A marcallon harry	on An Annalsha barran a	-0	A . Arrenter
C WYYYYXXX W XXXXX	TWAN CALARIAN AN TANKA	SPAPHAR WARANT ANT ANT ANT	WWWWWWWWWWW
AATOGEAASTTOFTER	AT C TAT TA ACT "T TARCET AT AT	A SCETANIGEACTATT P MIAGAGGAA	
mommen	Margan Margan	montontonton	mannon
GUI UAAAAT TASTGAA	0 AGAS & CFC TCA 6 & SPT 8 C TA ADT 1	ATOACT TATAAT TAACCARGA TGT FAT	GIICTIGADITITITGATI
A municipal Ar	Anna Maria	and man and the	
Man Obally	W MUN HOUR MOUNT	hall month man king	Wand Wand

FIGURE 13.3 A sample chromatogram, as viewed with the strace program (Ewing, 2002). Signal intensities corresponding to fragments ending with A (green). C (blue), G (black) and T (red) are shown out to approximately 722 bases.

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

9/17/18

CAP5510 / CGS5166

Sequencing

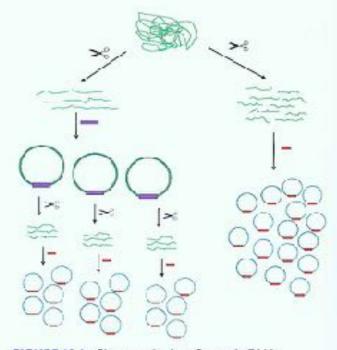
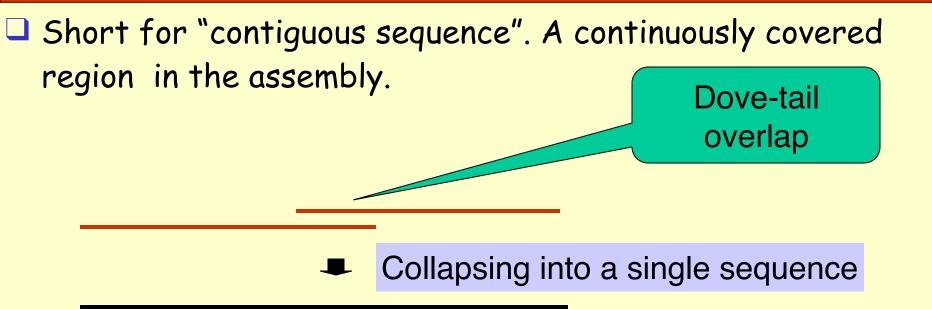


FIGURE 13.1 Shotgun cloning. Genomic DNA requercing begins with kolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 KL, then ligated to a suitable vector, such as a BAC vector shown in blue. Individually isolated clones in turn are sheared independently, generating fragments of approximately 4 Kb, which are then ligated to a small-scale vector, typically a plasmid (red bar) suitable for sequencing reactions. The whole genome shotgun approach bypasses the intermediate large-insert clone and generates large numbers of small fragments, typically 4 Kb and 10 Kb.

Sequencing: Generate Contigs



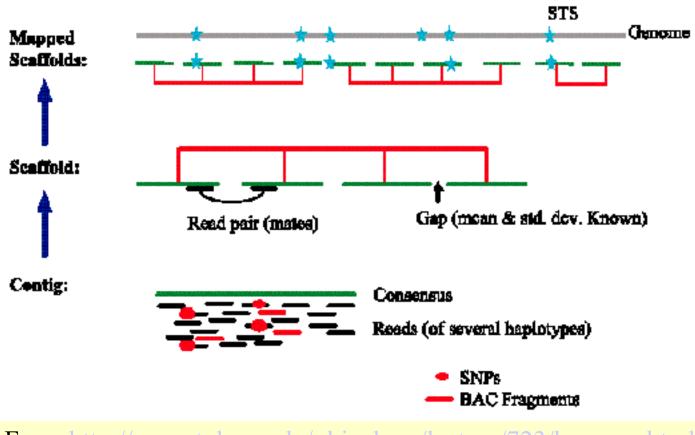
 Jang W et al (1999) Making effective use of human genomic sequence data. Trends Genet. 15(7): 284-6.
 Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11(9): 1541-8.

Paired Reads

Scaffold (supercontig): formed when two contigs with no sequence overlap can be linked

- Data from paired end reads help create scaffolds with known gaps
 - If two reads end up in two different contigs, then we can link contigs to form scaffold.

Shotgun Sequencing

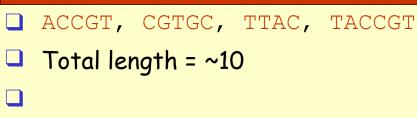


From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Human Genome Project

- Many videos available on youtube.com, dnatube.com, and elsewhere.
- Find some and watch them.

Assembly: Simple Example



- --ACCGT--
- ----CGTGC
- **TTAC**-----
- -TACCGT-
- TTACCGTGC

Assembly: Complications

Errors in input sequence fragments (~3%) Indels or substitutions
Contamination by host DNA
Chimeric fragments (joining of non-contiguous fragments)
Unknown orientation
Repeats (long repeats)
Fragment contained in a repeat
Repeat copies not exact copies
Inherently ambiguous assemblies possible
Inverted repeats
Inadequate Coverage

Assembly: Complications

- w = AGTATTGGCAATC
- z = AATCGATG
- $\mu = ATGCAAACCT$
- x = CCTTTTGG
- $\gamma = TTGGCAATCACT$

AGTATTGGCAATCAATCG	ATG
	ATGCAAACCT
TTGGCAATCACT	CCTTTTGG
AGTATTGGCAATCACTAATCG	ATGCAAACCTITTGG

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

AGTATTGGCAATC-----CCTTTTGG-----------TTGGCAATCACT -----ATGCAAACCT------AGTATTGGCAATCGATGCAAACCTTPTGGCAATCACT

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

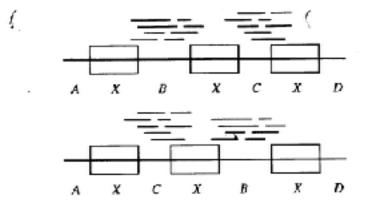
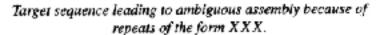
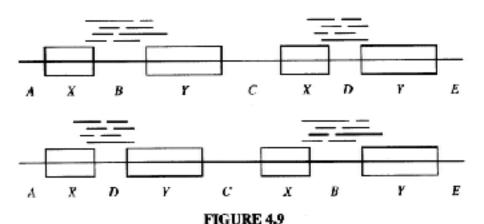
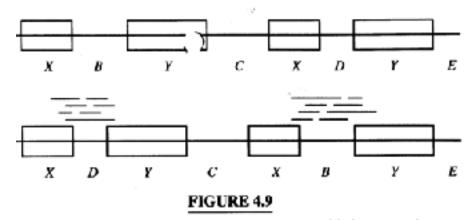


FIGURE 4.8





Target sequence leading to ambiguous assembly because of repeats of the form XYXY.



Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

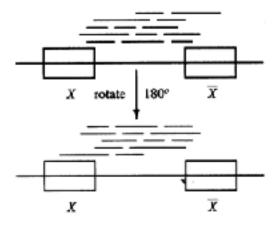


FIGURE 4.10

Target sequence with inverted repeat. The region marked \overline{X} is the reverse complement of the region marked X.