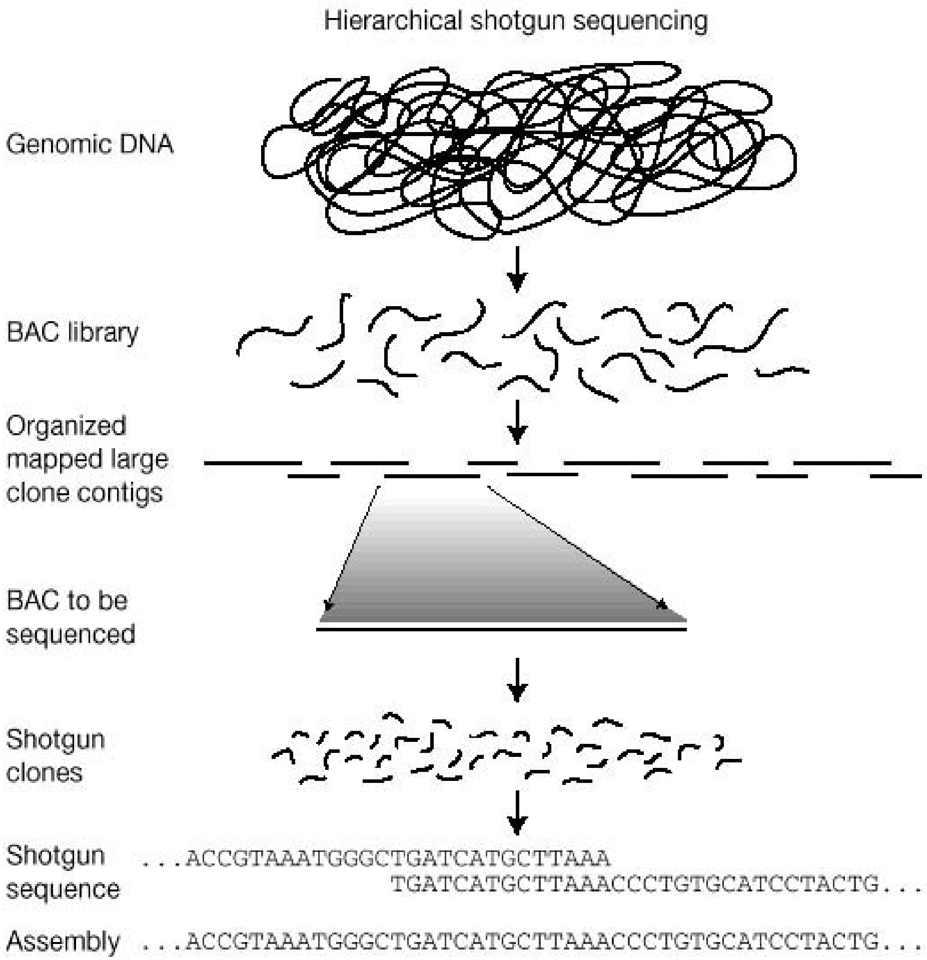
Sequencing

Shotgun Sequencing


Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

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

Sequencing

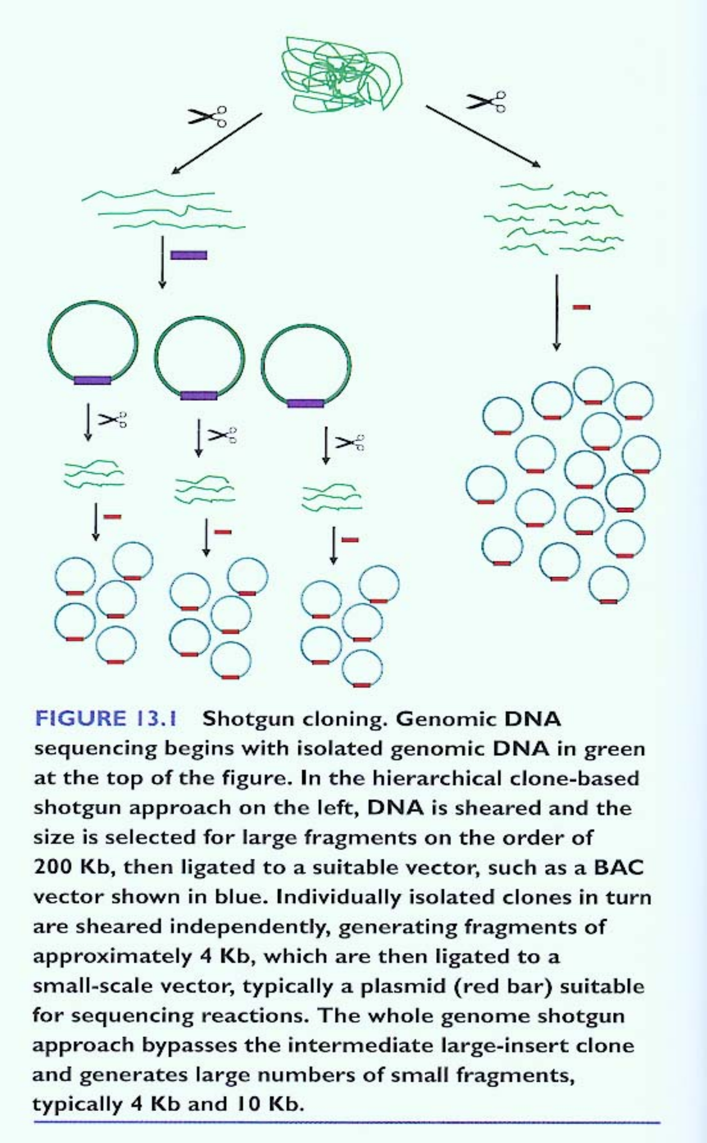
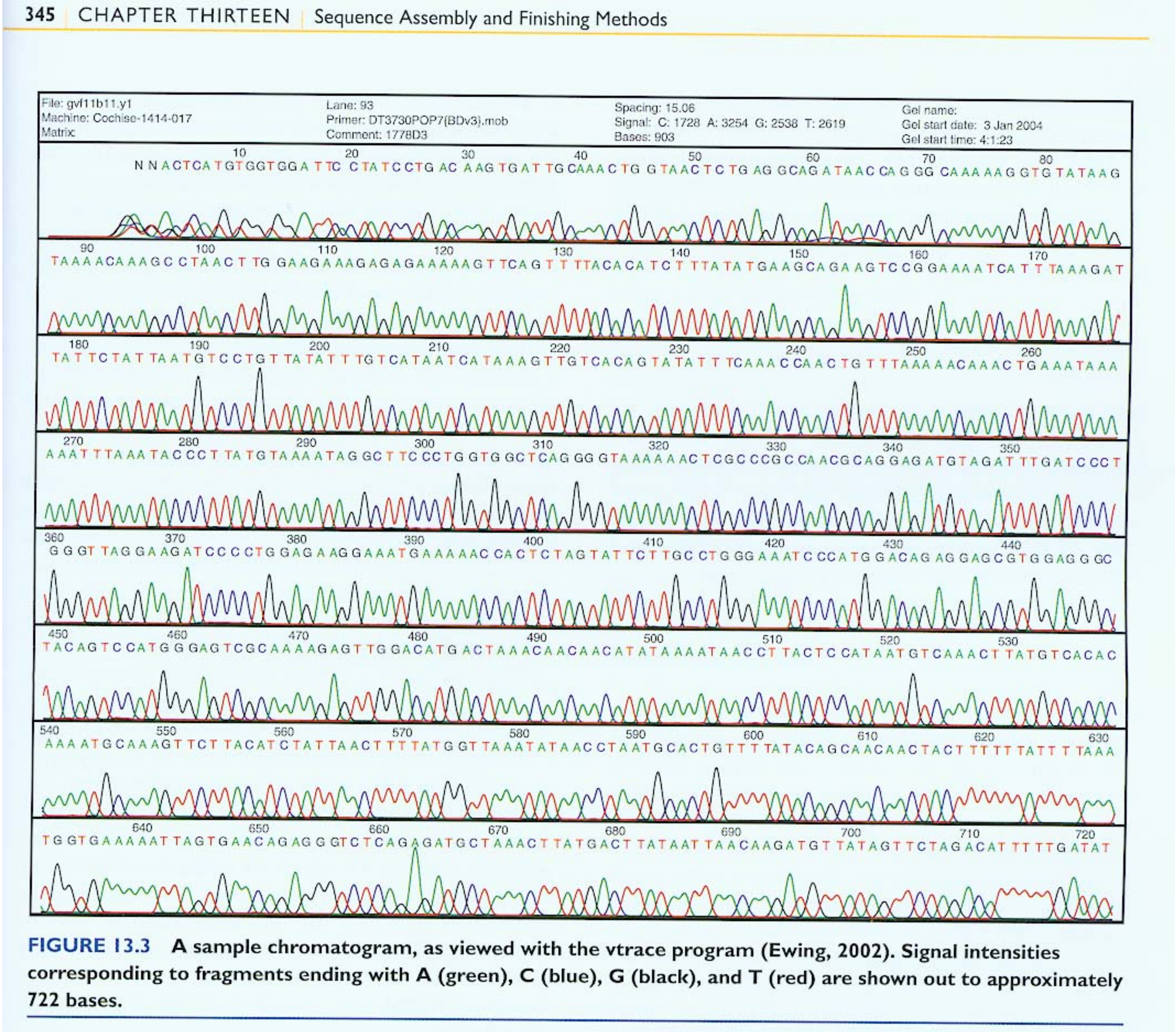


FIGURE 13.1 Shotgun cloning. Genomic DNA sequencing begins with isolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 Kb, then ligated to a suitable vector, 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.

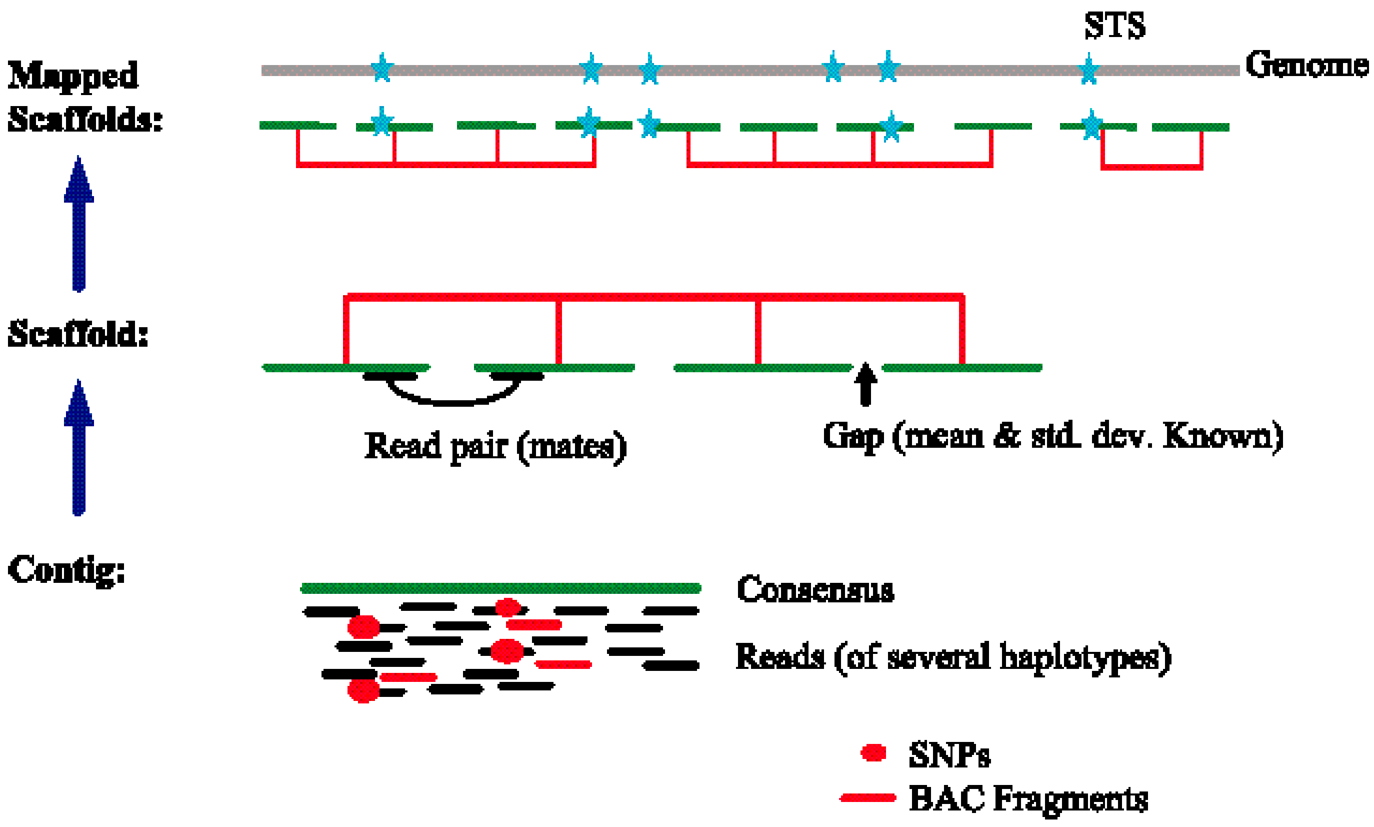
Sequencing



345 | CHAPTER THIRTEEN | Sequence Assembly and Finishing Methods

FIGURE 13.3 **A sample chromatogram, as viewed with the vtrace 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

Human Genome Project

Play the Sequencing Video:

- Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe
- Then run it on your PC.

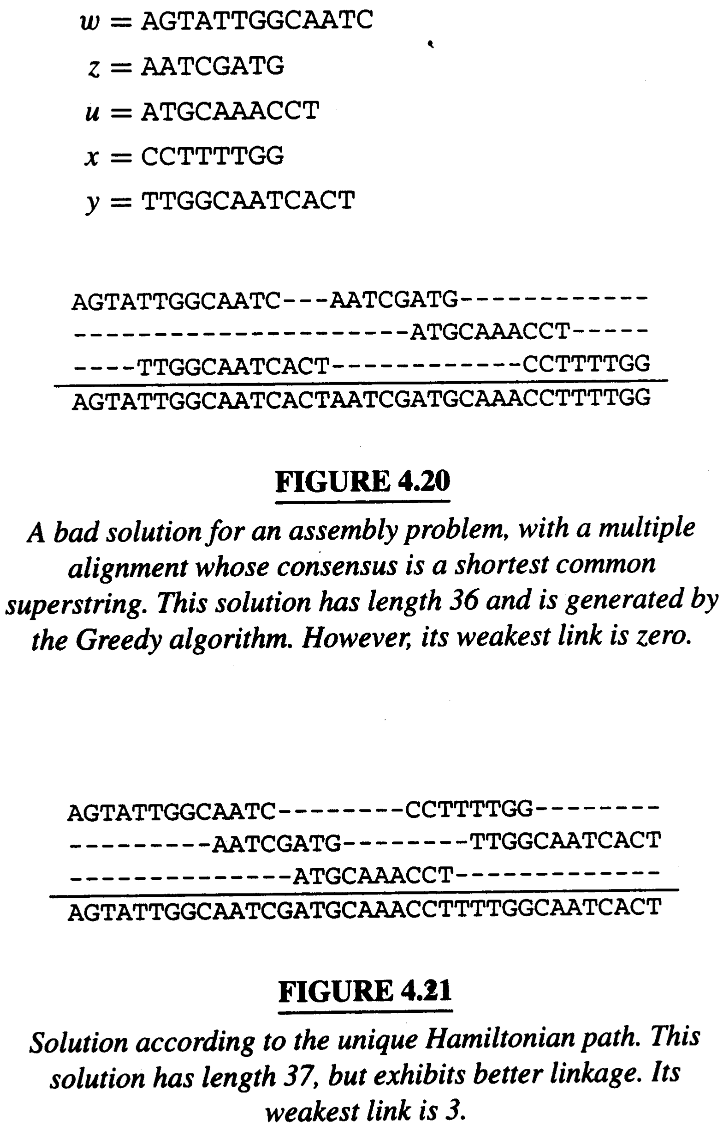
Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-
  - **--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

u = ATGCAAACCT

x = CCTTTTGG

y = TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
---------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

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--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT-------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

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

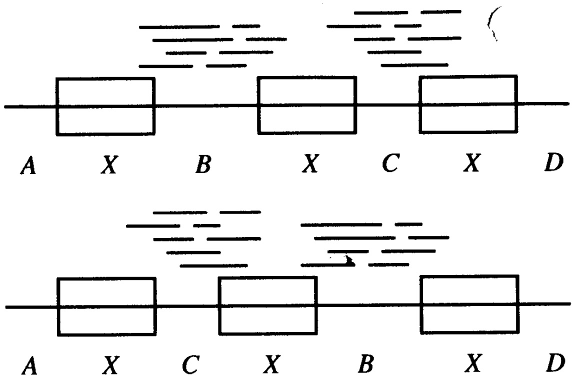


FIGURE 4.8

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

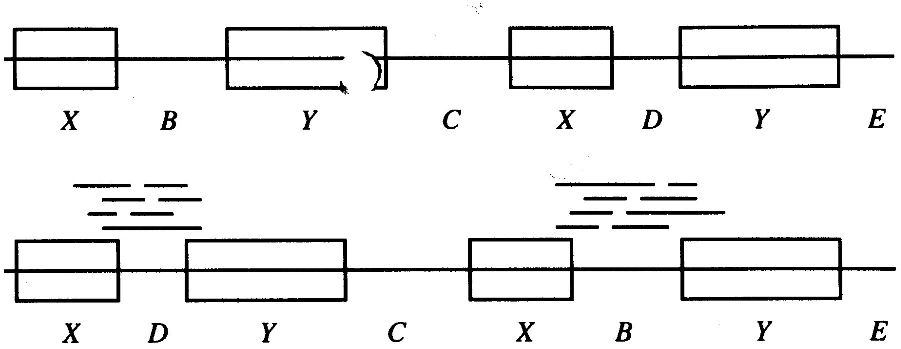


FIGURE 4.9

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

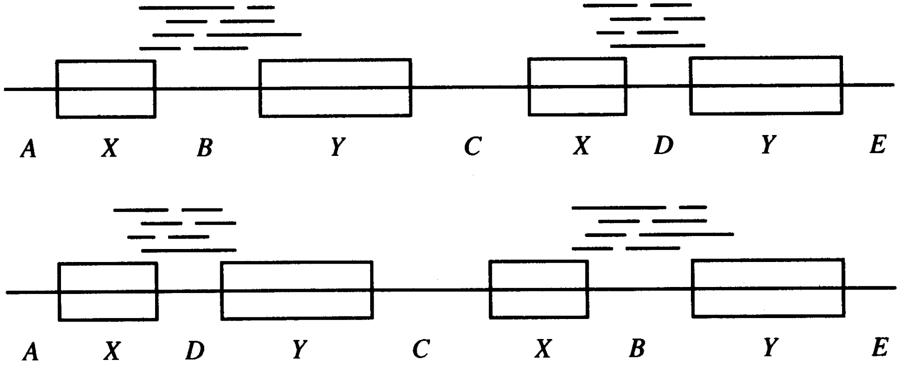


FIGURE 4.9

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

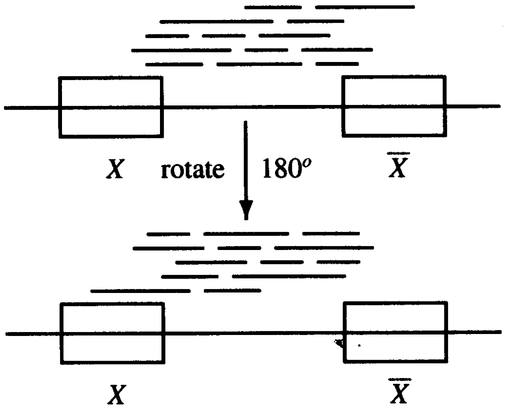


FIGURE 4.10

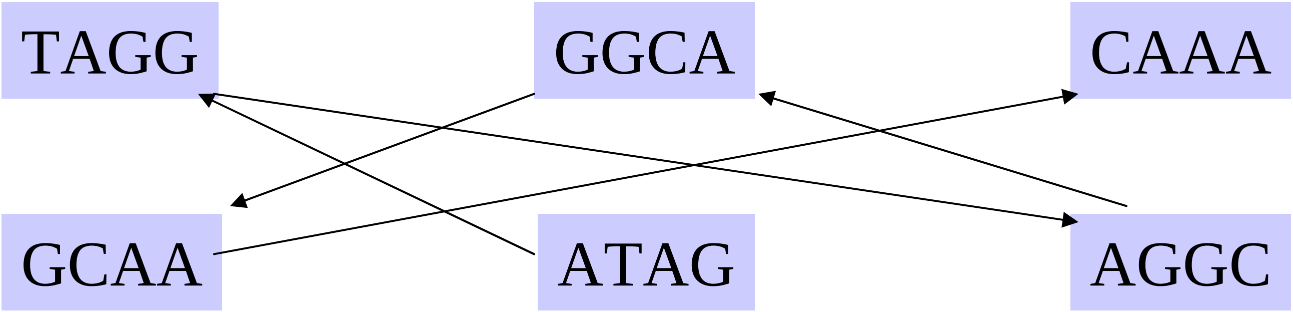
Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.

Miscellaneous

- Contig: A continuously covered region in the assembly.
- Other sequencing methods:
  - Sequencing by Hybridization (SBH)
  - Dual end sequencing
  - Chromosome Walking (see page 5-6 of Pevzner's text).

SBH

- Suppose that the only length 4 fragments that hybridize to S are: **TAGG, GGCA, CAAA, GCAA, ATAG, AGGC**. Then what is S, if it is of length ~9?



TAGG

GGCA

CAAA

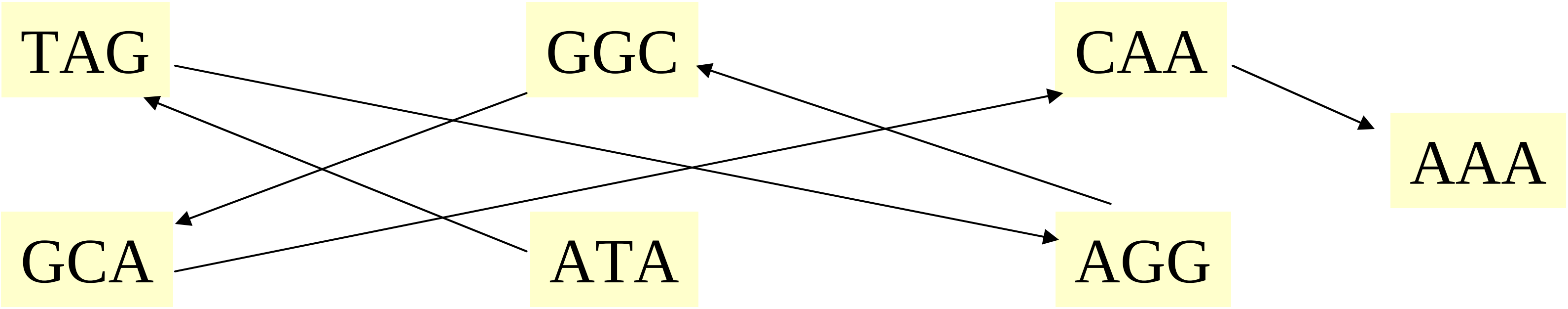
GCAA

ATAG

AGGC

Hamiltonian Path Problem

SBH



TAG

GGC

CAA

AAA

GCA

ATA

AGG

Eulerian Path Problem

Assembly Software

- Parallel EST alignment engine (http://corba.ebi.ac.uk/EST") with a CORBA interface to alignment database. Can perform ad hoc assemblies. Can act as foundation for CORBA-based EST assembly and editing package. [Parsons, EBI]
- Software using multiple alternative sequence assembly "engines" writing to a common format file [Staden, Cambridge] (http://www.mrc-lmb.cam.ac.uk/pubseq/index.html).
- Phrap,(http://bozeman.genome.washington.edu/phrap.docs/phrap.html)
- Assembler (TIGR) for EST and Microbial whole-genome assembly (http://www.tigr.org/softlab/)
- FAK2 and FAKtory (http://www.cs.arizona.edu/people/gene/) [Myers]
- GCG (http://www.gcg.com)
- Falcon [Gryan, Harvard] fast (rascal.med.harvard.edu/gryan/falcon/)
- SPACE, SPASS [Lawrence Berkeley Labs] (http://www-hgc.lbl.gov/inf/space.html)
- CAP 2 [Huang] (http://www.tigem.it/ASSEMBLY/capdoc.html)