DNA Sequencing

Guan-Shieng Huang

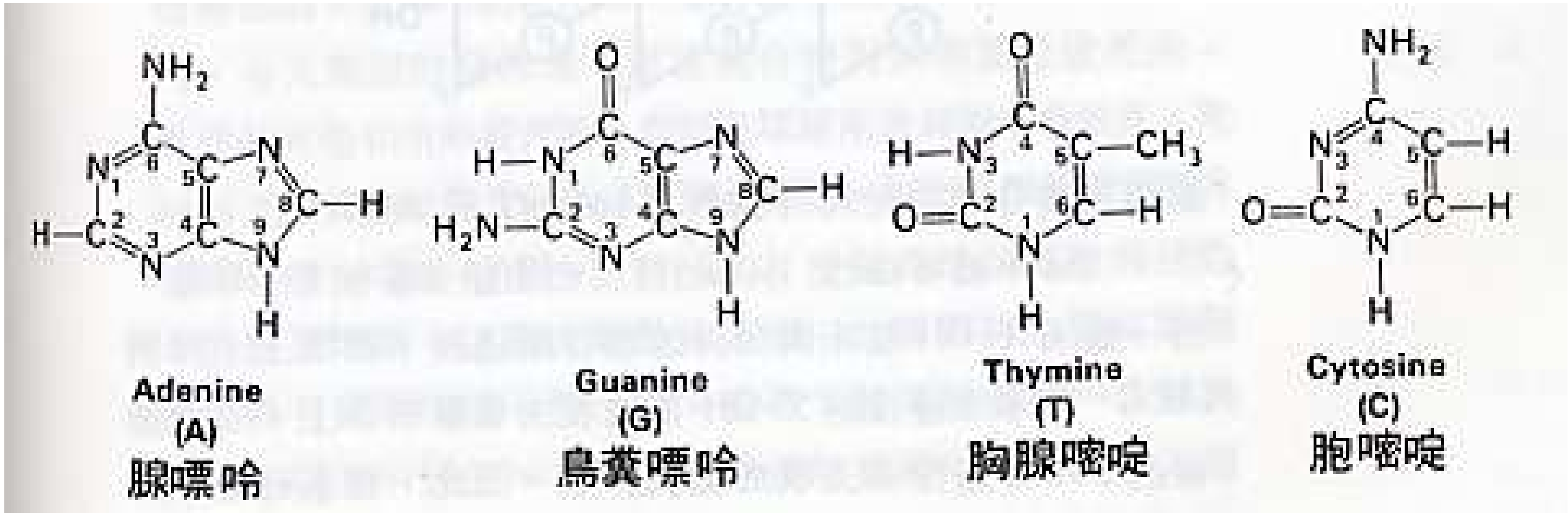
Apr. 1, 2003

What is DNA Sequencing

To obtain the linear structure of a DNA sequence.

DNA—deoxyribonucleic acid

DNA is constructed from four kinds of building blocks:



Adenine (A) 腺嘌呤

Guanine (G) 鳥糞嘌呤

Thymine (T) 胸腺嘧啶

Cytosine (C) 胞嘧啶

Two single strands of DNA combine to form a double helix.

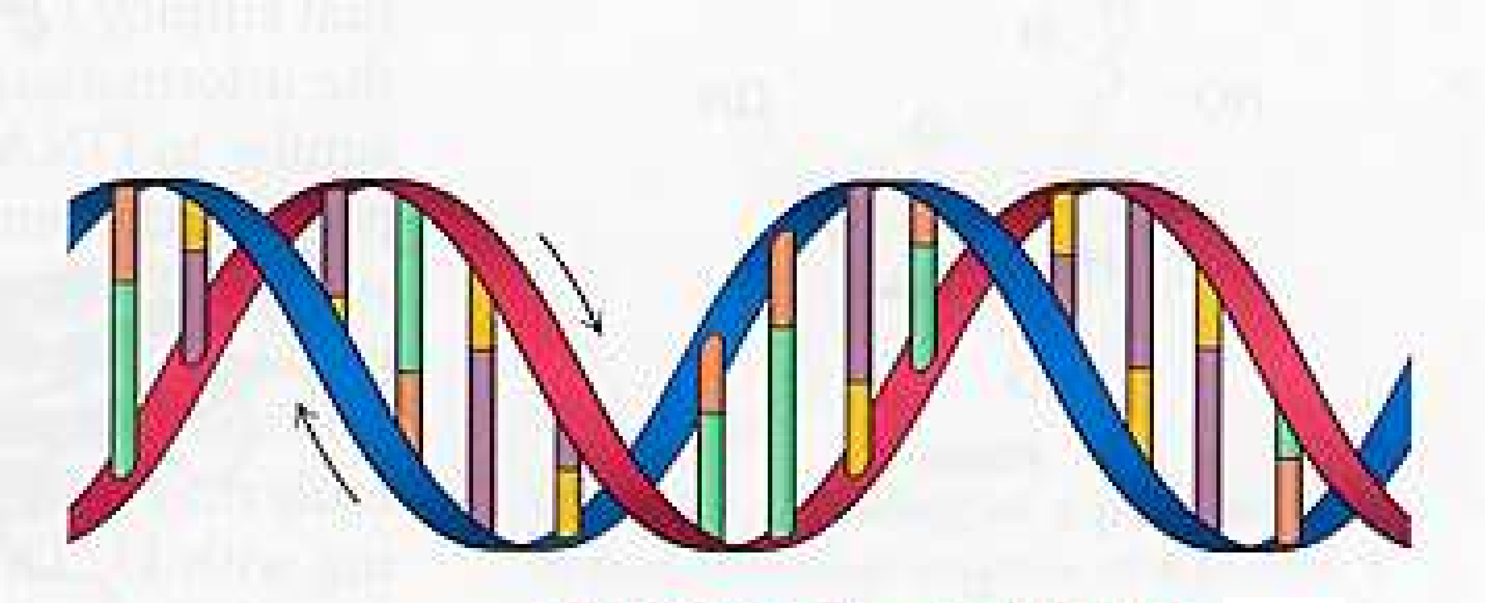
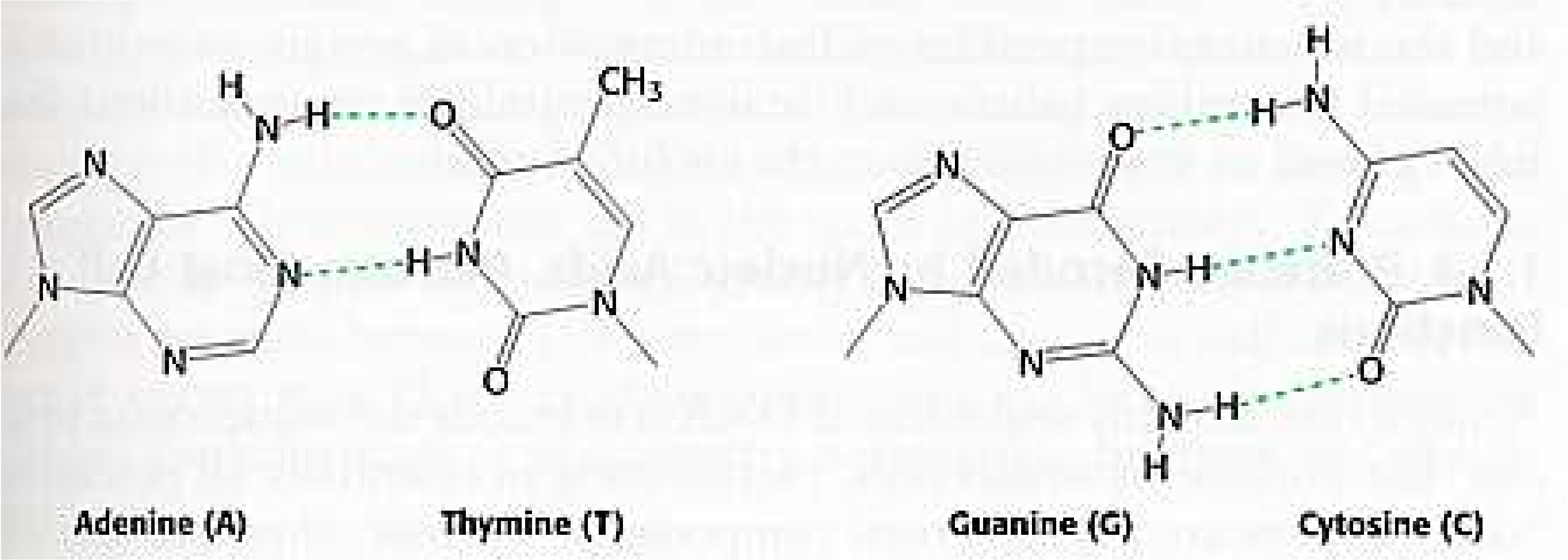
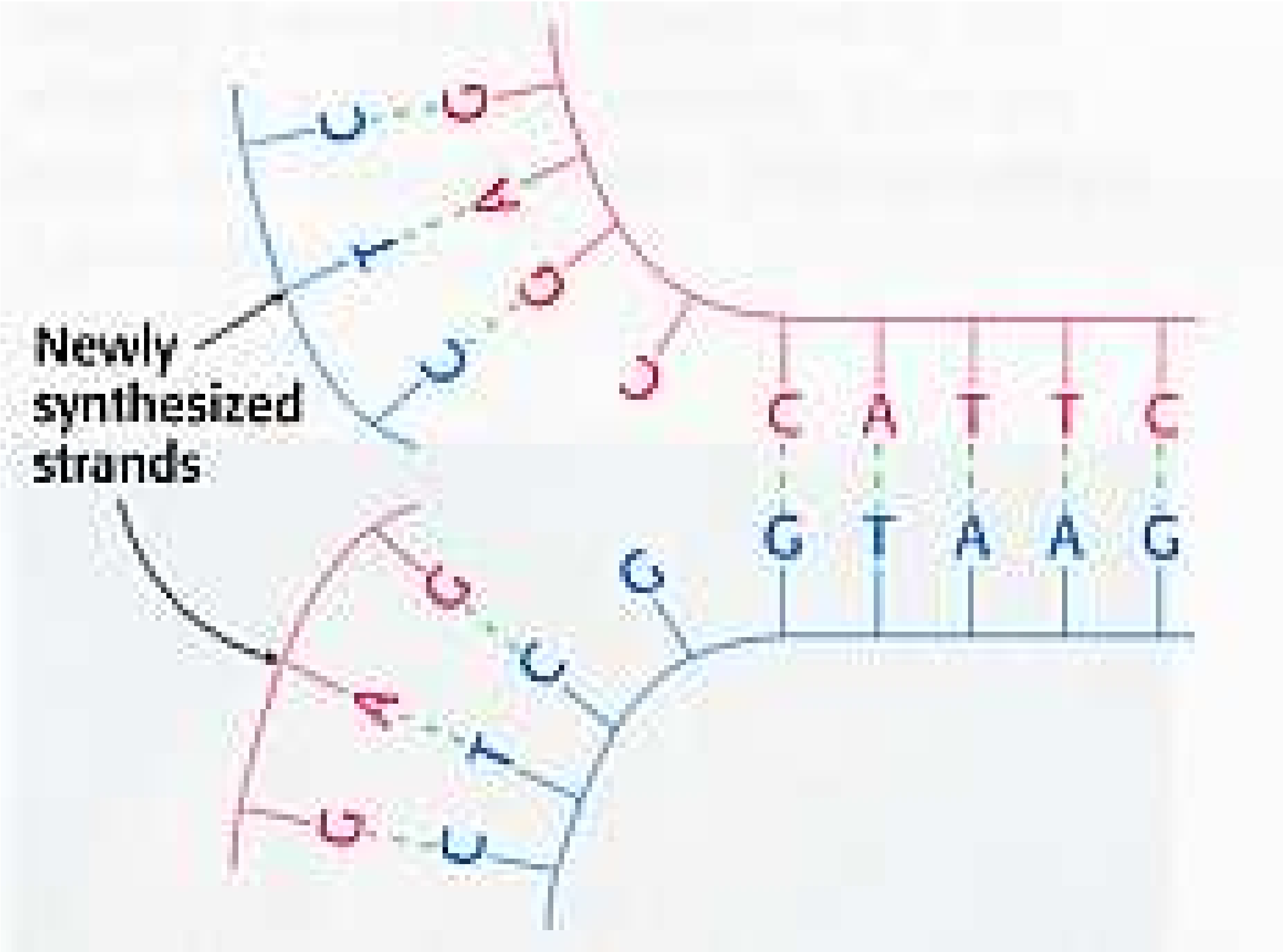


FIGURE 1.2 The double helix.

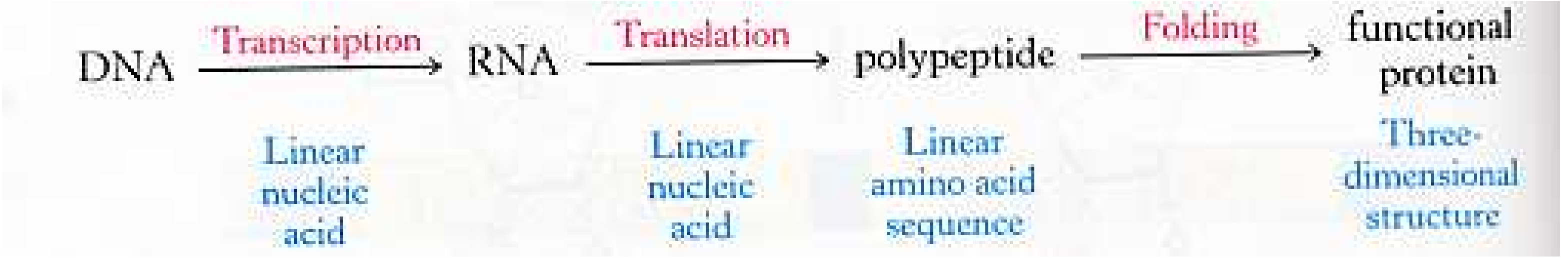
Watson-Crick base pairs



DNA replication



DNA is a stable storage form for genetic information.



Sanger Dideoxy Method

- The most popular method to determine the sequence of DNA.
- Invented in 1977.

1. **Given any DNA fragment, if we can measure the length of all of its prefixes ending at A (and G, C, T, resp.), the fragment is determined.**

   ```
   CTTAAGCGATTA ...
   CTTA
   CTTAA
   CTTAAGCGA
   CTTAAGCGATTA
   ```

2. **Controlled interruption of enzymatic replication:**

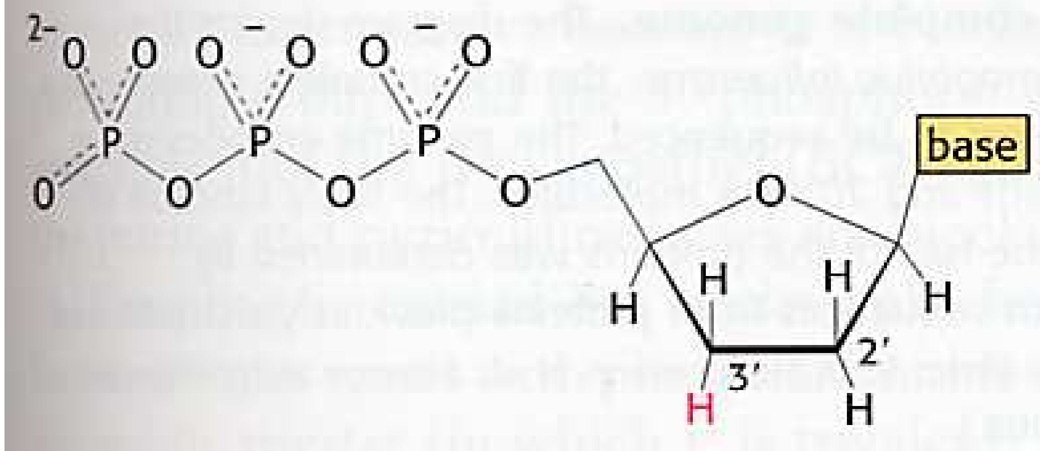
   (a) **DNA polymerase:** 催化DNA之合成

   (b) 放射性**dATP, dTTP, dCTP, dGTP: provide resource for the extension**

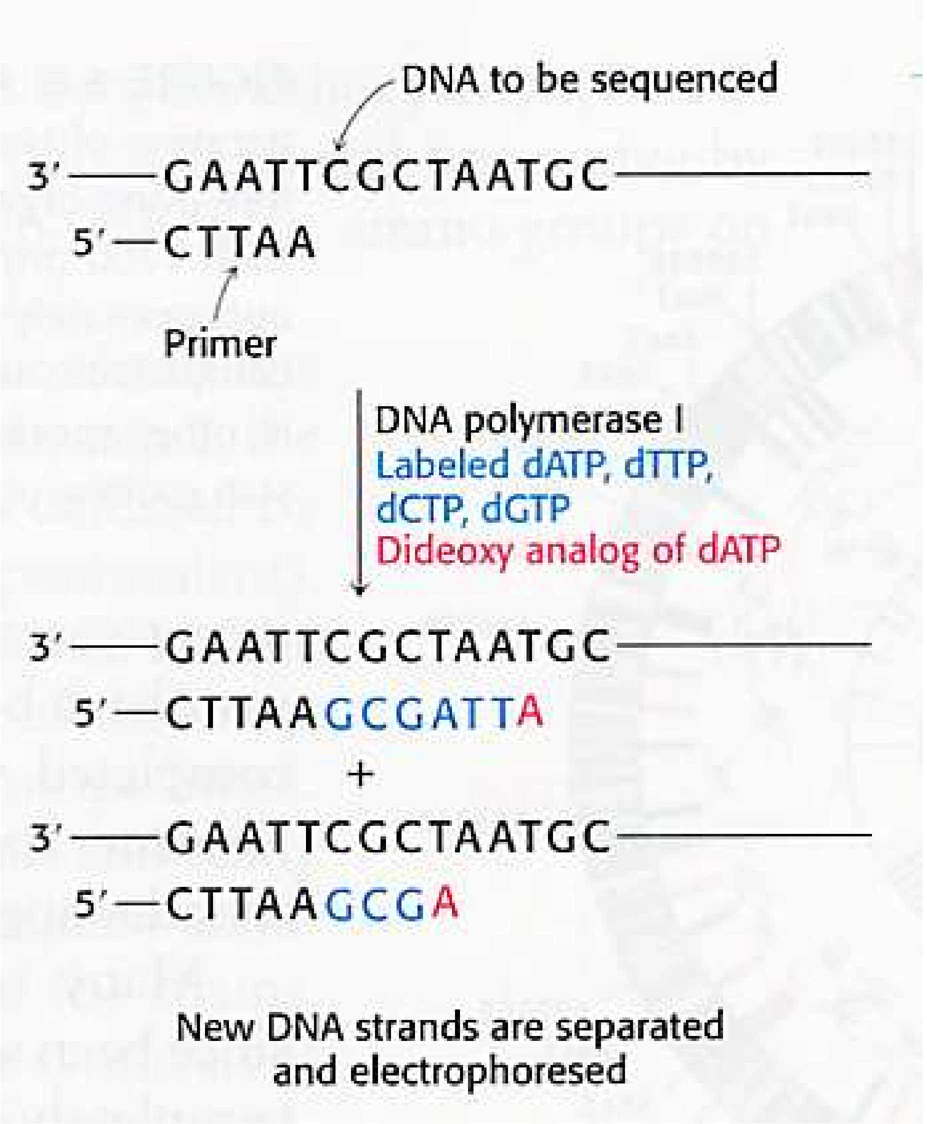
(c) primer: trigger the duplication

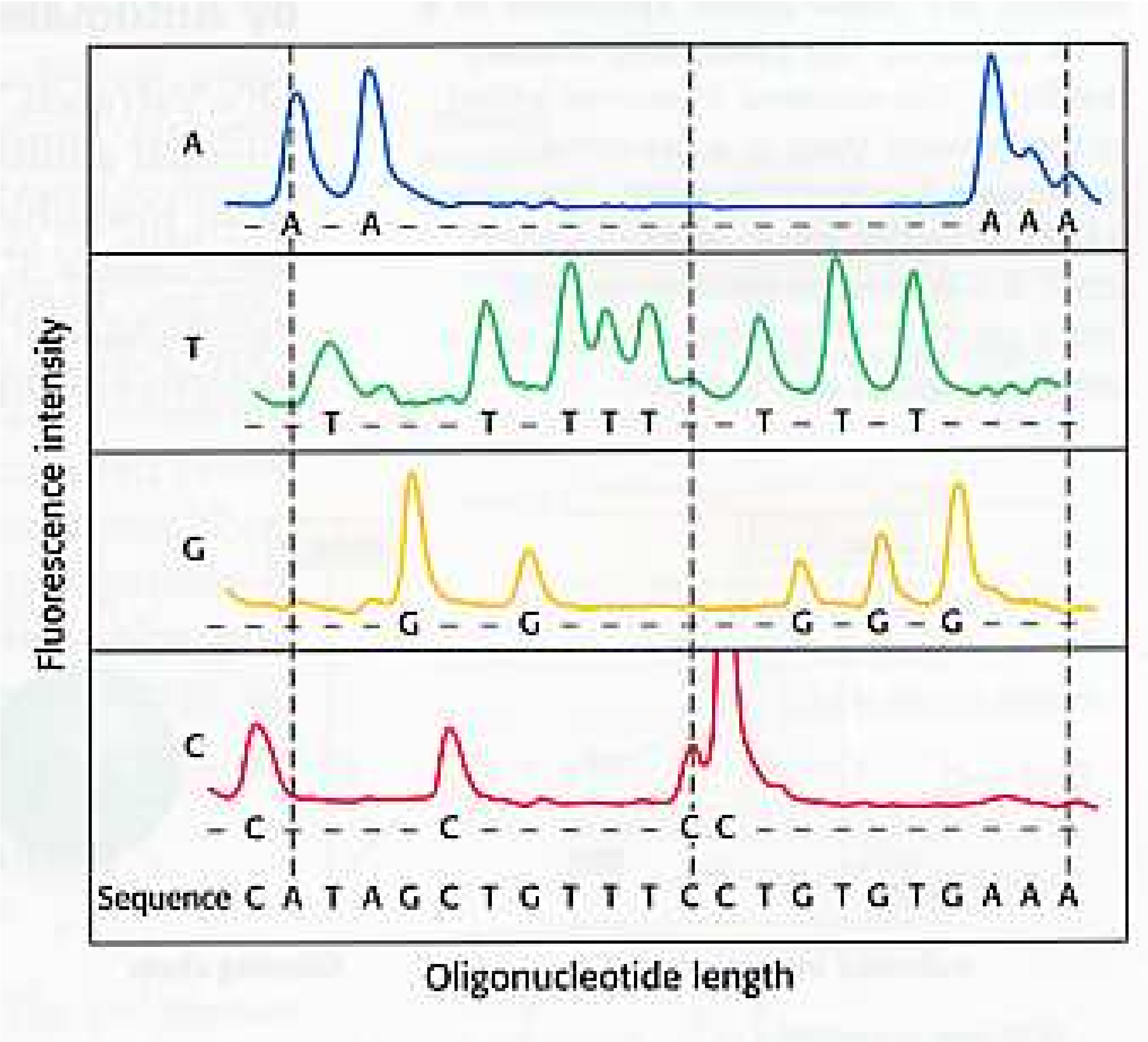
(d) dideoxy analog of dATP: terminate the extension of A.

3. electrophoresis: separate the replications by length



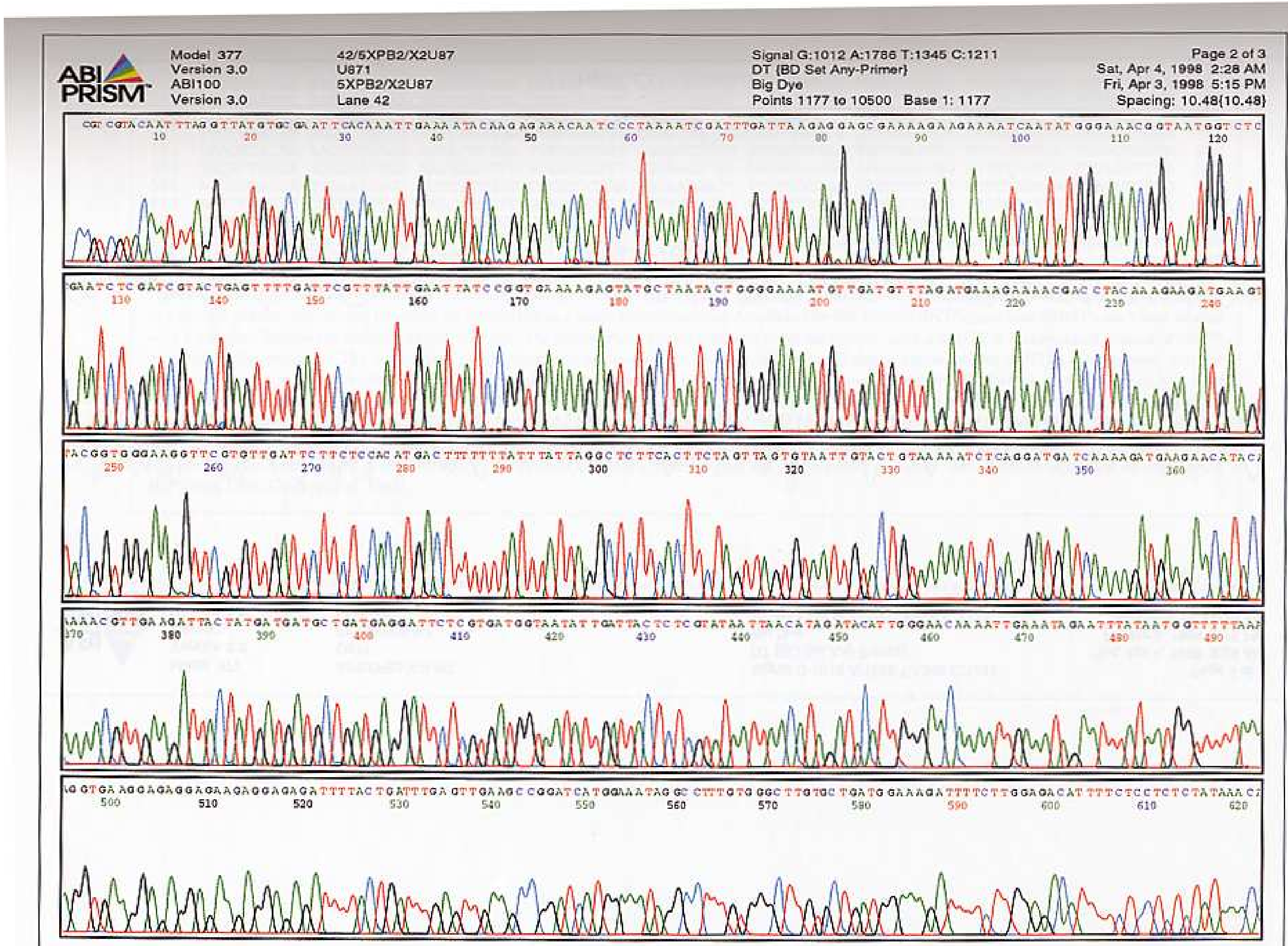
2′, 3′-Dideoxy analog


DNA to be sequenced
3′—GAATTCGCTAATGC—
5′—CTTAA
Primer
DNA polymerase I
Labeled dATP, dTTP,
dCTP, dGTP
Dideoxy analog of dATP
3′—GAATTCGCTAATGC—
5′—CTTAAGCGATTA
+
3′—GAATTCGCTAATGC—
5′—CTTAAGCGA
New DNA strands are separated
and electrophoresed


Fluorescence intensity
A
– A – A – – – – – – – – – – – – – – – – – A A A
T
– – T – – – T – T T T – T – T – T – – – –
G
– – – – – G – – G – – – – – G – G – G – – –
C
– C – – – – – C – – – – – – – C C – – – – – – – – –
Sequence C A T A G C T G T T T C C T G T G T G A A A
Oligonucleotide length

High-throughput Sequencing

1. **automation of Sanger sequencing**
   - (a) **Four-color fluorescent dyes have replaced the radioactive label.**
   - (b) **Reads greater that 800 bp are possible, though $500 \sim 700$ is more common.**
   - (c) **Applied Biosystem's ABI Prism™ 3700: six 96-well plates per day** $(96 \times 6 \times 800 \sim 0.5M)$
   - (d) **Amersham Pharmacia's Mega BASE 1000™**



ABI PRISM

Model 377
Version 3.0
ABI100
Version 3.0

42/5XPB2/X2U87
U871
5XPB2/X2U87
Lane 42

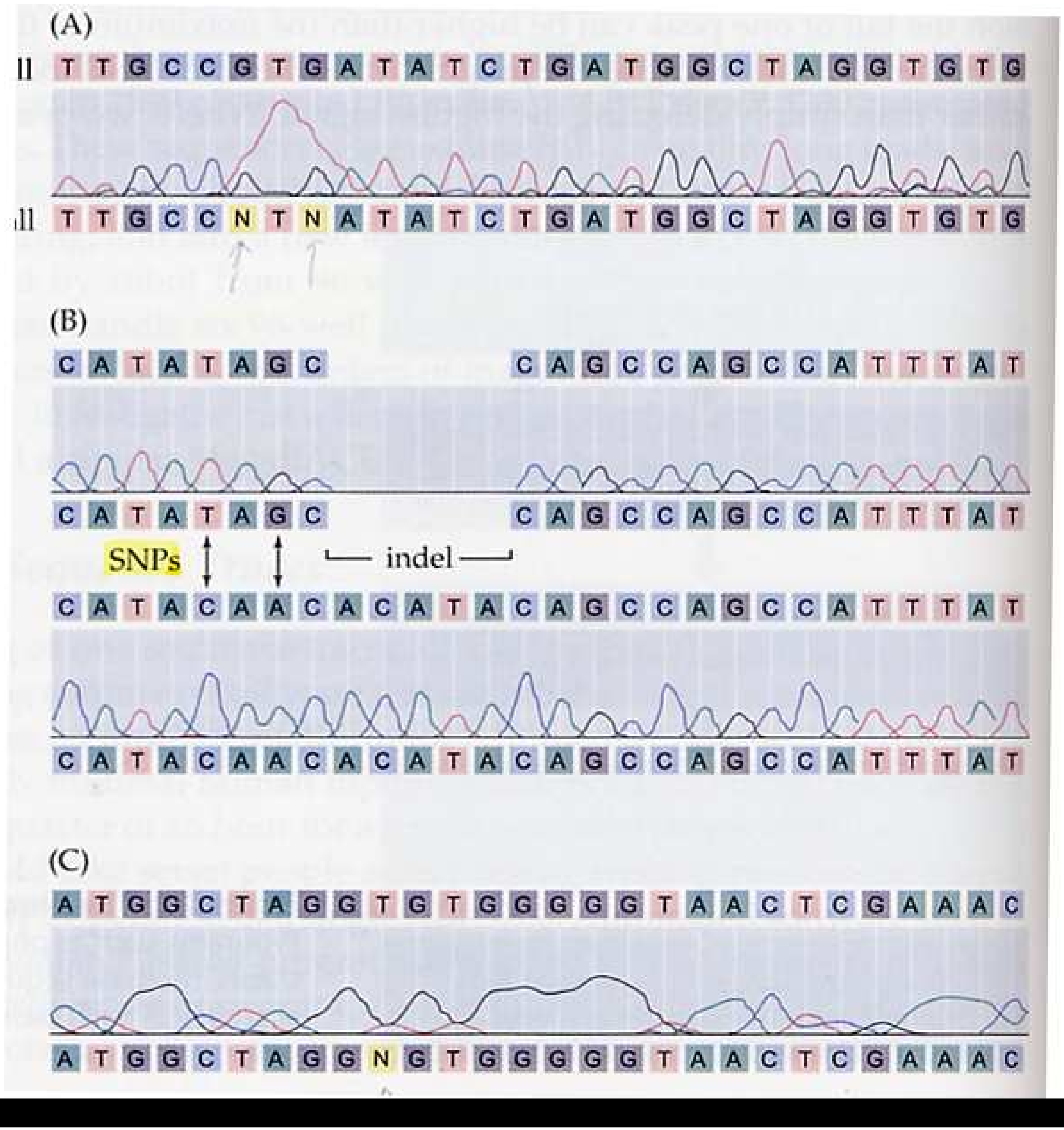
Signal G:1012 A:1786 T:1345 C:1211
DT {BD Set Any-Primer}
Big Dye
Points 1177 to 10500 Base 1: 1177

Page 2 of 3
Sat, Apr 4, 1998 2:28 AM
Fri, Apr 3, 1998 5:15 PM
Spacing: 10.48{10.48}

Figure 2.2. *Continued.*

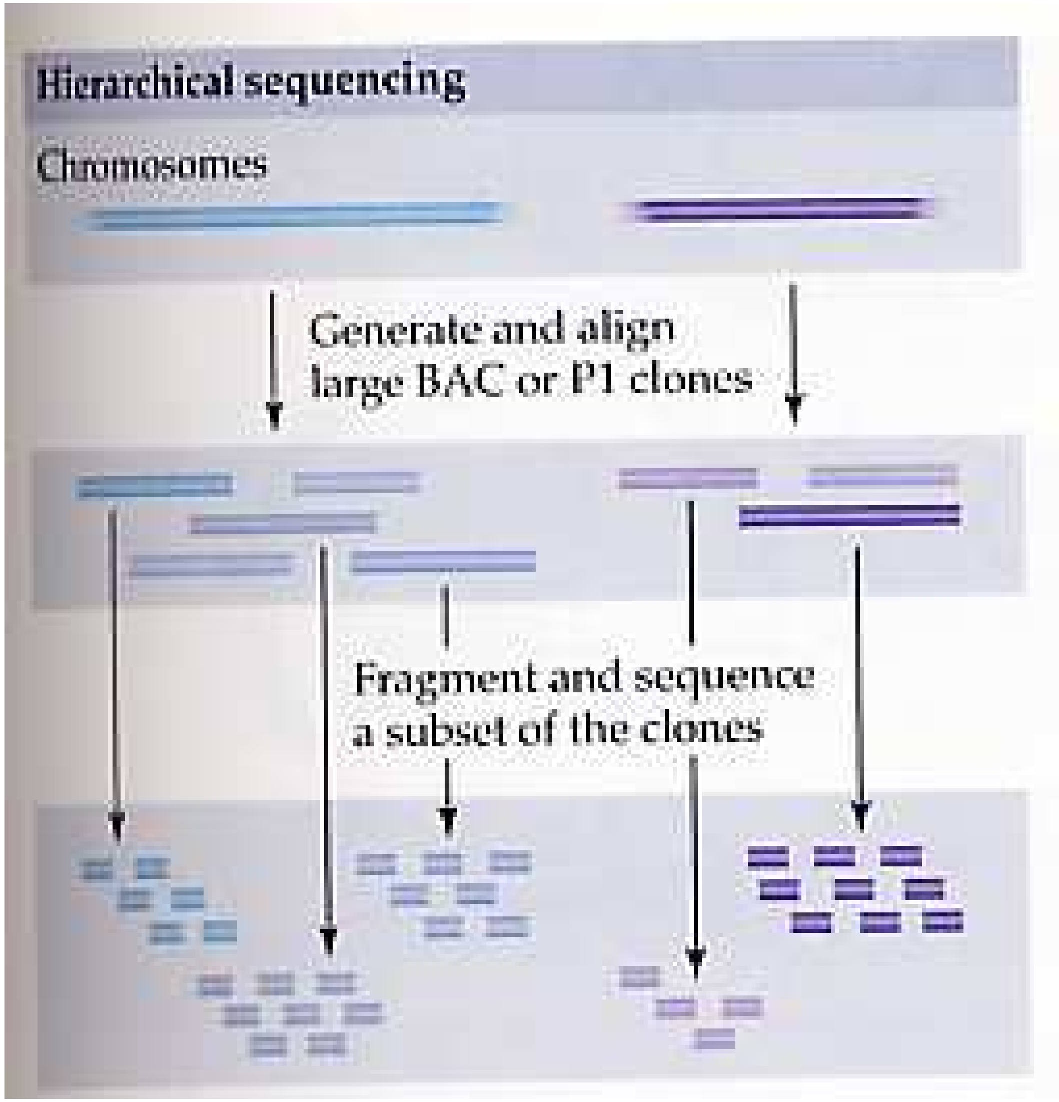
Reading Sequence Traces

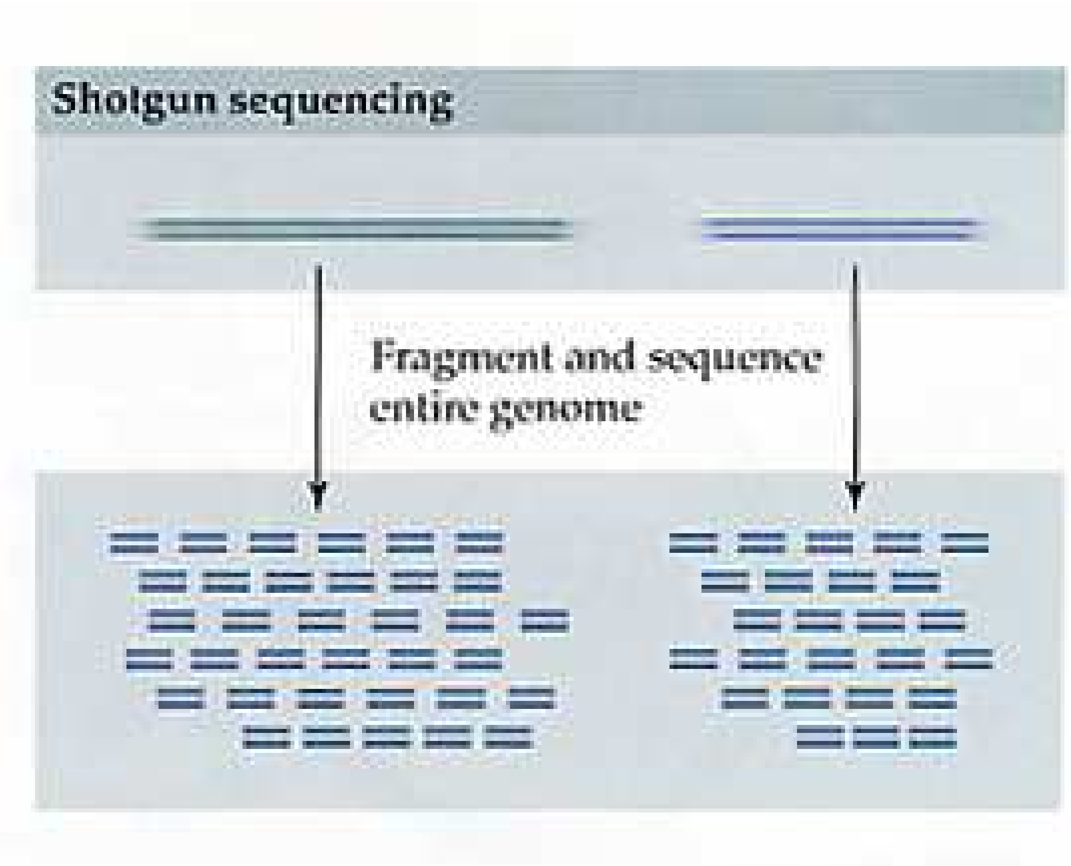
- base-calling
- phred program (http://www.phrap.org)
  Developed at the University of Washinton in 1998, can convert traces (analog signals) into sequences (digital signals).
- < 50: noisy
  > 800: signals declined


(A)
TTGCCGTGATATCTGATGGCTAGGTGTG
TTGCCNTNATATCTGATGGCTAGGTGTG
(B)
CATATAGC
CAGCCAGCCATTTAT
CATATAGC
CAGCCAGCCATTTAT
SNPs
indel
CATACAACACATACAGCCAGCCATTTAT
CATACAACACATACAGCCAGCCATTTAT
(C)
ATGGCTAGGTGTGGGGGTAACTCGAAAC
ATGGCTAGGNGTGGGGGTAACTCGAAAC

Genome Sequending

- **hierachical sequencing**
- **shotgun sequencing**


Hierarchical sequencing
Chromosomes
Generate and align large BAC or P1 clones
Fragment and sequence a subset of the clones



Shotgun sequencing

Contig Assembly

Assemble short DNA fragments (reads) to the entire DNA sequence of the clone:

- Overlap: finding potentially overlapping fragments
- Layout: finding the order of reads
- Consensus: deriving the DNA sequence from the layout

Overlap

- **Find the best match between the suffix of one segment and the prefix of another.**
- **Fragment may contain** $1 \sim 3\%$ **errors.**

$\Longrightarrow$ **probability & dynamic programming**

Layout

The difficulty is deciding whether two fragments with a good overlap really overlap.

- **Long repeated regions:**

$$\underbrace{RXRR}_{f_1}\underbrace{RRZ}_{f_2} \Rightarrow RXRRZ$$

- **Chimeras: two fragments are mistakenly joined end-to-end and are interpreted as a contiguous region.**

- **Basecalling errors: the name (i.e. A, G, C, T) of an individual base is reported incorrectly in a fragment.**

Greedy Algorithm to Contig Assembly

1. Start with each fragment as a contig.
2. Repeatly merge the two configs with the best overlap.

Remark: This algorithm fails in the presence of repeats.

Shortest Common Superstring

Given a set of substrings, find the shortest string that contains all of these substrings.

- **NP-complete (Gallant et al, 1980)**
- **It is conjectured that the Greedy Algorithm is a 2-approximation. (Open Problem)**

AGCGCGC, $\underline{\texttt{CGCGCG}}$, GCGCGCT

optimal: AG$\underline{\texttt{CGCGCG}}$CT

greedy: AGCGCGCT$\underline{\texttt{CGCGCG}}$

- **Also Fails on repeats.**

Repeats

- **Alu repeats** (300 **bp**) $\sim 1M$
- **LINE repeats** (1000 **bp**) ~ 200000
- $\sim 25\%$ **of human genes are present in at least two copies**

Assembly Programs

- Phrap (G. Green, U. Washington)
- FAK (G. Myers, U. Arizona)
- CAP (X. Huang, Mich. Tech. U.)
- TIGR Assembler (G. Sutton, TIGR)

SARS

雞鴨體內的冠狀病毒突變種,
可能是人與畜的病毒基因發生重組所導致.