

# I. Retrieving DNA Sequence Information

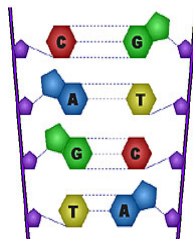
## DNA Sequencing

### Lecture 6

Instructor: Su-In Lee

## DNA sequencing

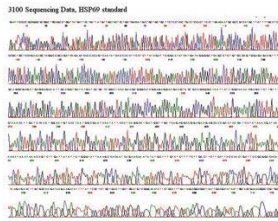
- A process of determining the exact order of the bases A, T, C and G in a piece of DNA.
- Why important? Blueprint of life; personalized medicine



```
...ACGTGACTGAGGACCGTG  
CGACTGAGACTGACTGGGT  
CTAGCTAGACTACGTTTAA  
TATATATATACGTCGTCGT  
ACTGATGACTAGATTACAG  
ACTGATTTAGATACTGAC  
TGATTTTAAAAAATATT...
```

## DNA sequencing

- Goal
  - Find the complete sequence of A, C, G, T's in DNA.
- Challenges
  - There is no machine that takes long DNA as an input, and gives the complete sequence as output.
  - Can only sequence ~900 letters at a time.

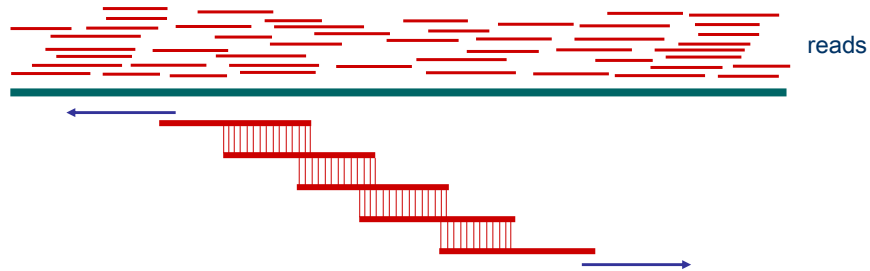


## Shotgun sequencing

### Genomic segment

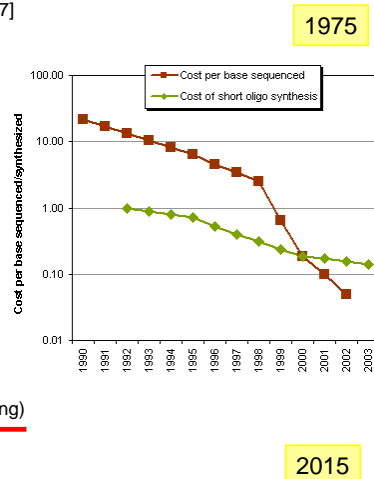


- Fragment assembly

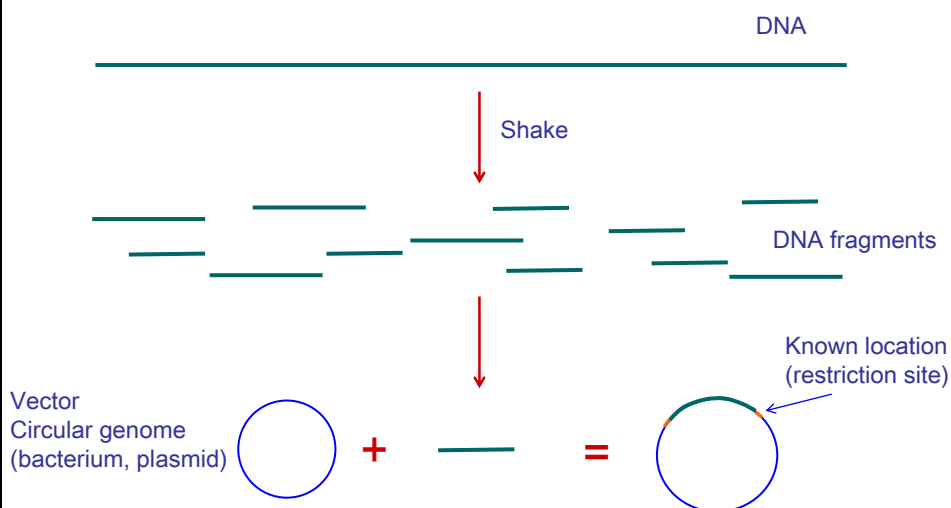


## DNA sequencing – past, current and future

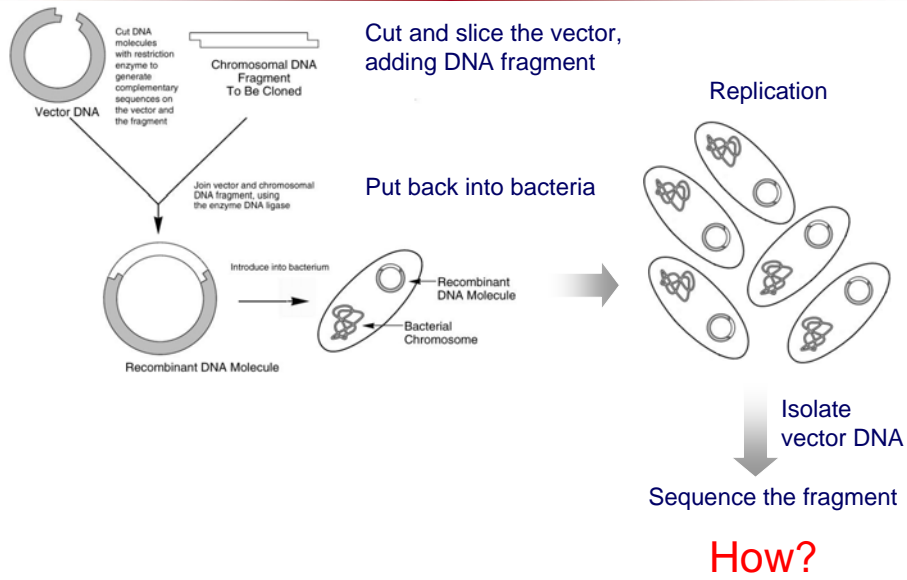
- Gel electrophoresis based sequencing
  - 300–500 bases of a DNA strand
  - Maxam-Gilbert sequencing [Maxam & Gilbert, 1977]
  - Chain-termination methods (“Sanger method”) [Sanger et al, 1975]: Nobel prize in Chemistry
- The first fully sequenced DNA-based genome
  - Bacteriophage phi X 174 of 5,375 nucleotides [Sanger et al, 1977]
- First Shotgun projects
  - 6,569 bp human mitochondrial DNA [Anderson et al, 1981]
  - 48,502 bp Bacteriophage  $\lambda$  nucleotide sequence [Sanger et al, 1982]
  - 172,282 bp sequence of Epstein-Barr virus [Baer et al, 1984]
- Whole genome sequencing
  - Hierarchical sequencing (Physical mapping, Walking)
  - Whole-genome shotgun sequencing ←
- New sequencing technologies
  - Pyrosequencing, single-molecule methods; New assembly techniques



## DNA Sequencing – vectors

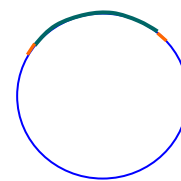


# Cloning



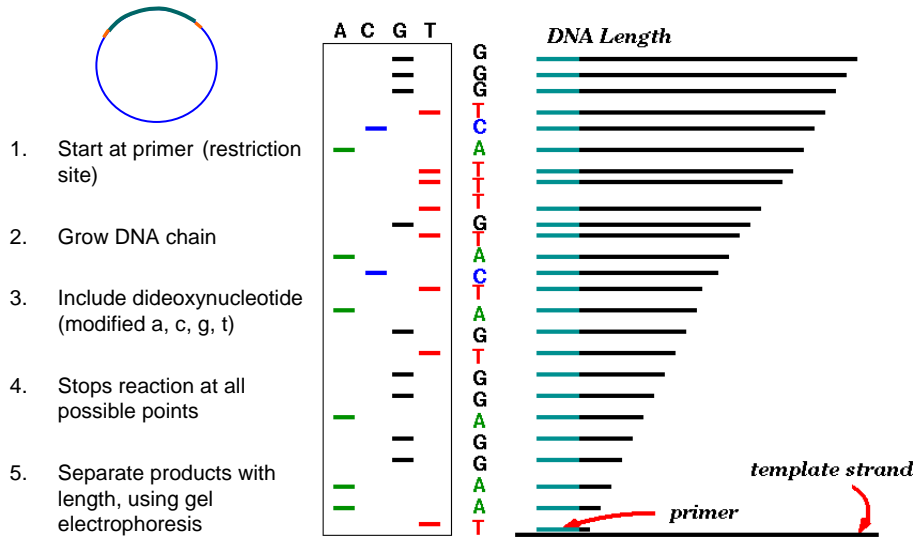
# Different types of vectors

<u>VECTOR</u>	<u>Size of insert</u>
Plasmid	2,000-10,000 Can control the size
Cosmid	40,000
BAC (Bacterial Artificial Chromosome)	70,000-300,000
YAC (Yeast Artificial Chromosome)	> 300,000 Not used much recently



## Chain-termination method

- [Animation](#)

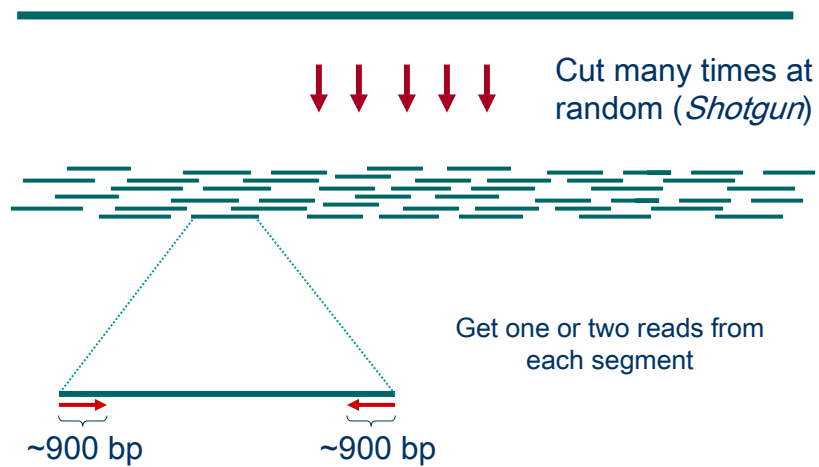


Computational Molecular Biology and Genomics, Spring 2009

9

## Method to sequence longer regions

### Genomic segment

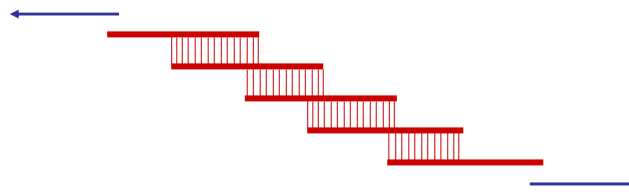


Computational Molecular Biology and Genomics, Spring 2009

10

## Reconstructing the sequence

- Fragment assembly



Cover region with high redundancy

Overlap & extend reads to reconstruct the original genomic region

## Definition of coverage



Length of genomic segment: **G**

Number of reads: **N**

Length of each read: **L**

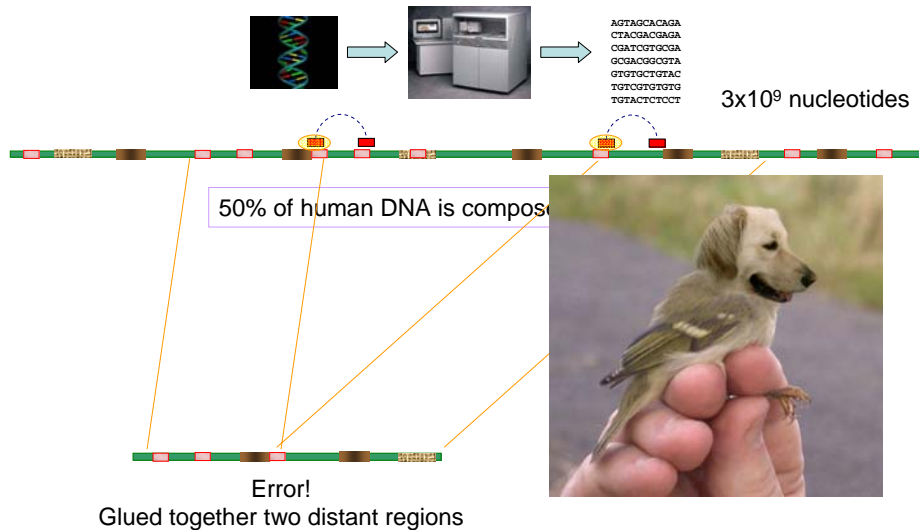
**Definition:** Coverage  $C = N L / G$

How much coverage is enough?

**Lander-Waterman model:**  $\text{Prob}[\text{not covered bp}] = e^{-C}$

Assuming uniform distribution of reads,  $C=10$  results in 1 gapped region / 1,000,000 nucleotides

## Challenges – Repeats!



Computational Molecular Biology and Genomics, Spring 2009

13

## Repeats

Bacterial genomes: 5%

Mammals: 50%

- Repeat types

- Low-complexity DNA

- e.g. ATATATATACATA...

- Microsatellite repeats  $(a_1 \dots a_k)^N$  where  $k \sim 3-6$

- e.g. CAGCACTAGCAGCACCCAG

- Common Repeat Families

- SINE (Short Interspersed Nuclear Elements)

- e.g. ALU: ~300-long,  $10^6$  copies

- LINE (Long Interspersed Nuclear Elements)

- e.g. ~500-5,000-long, 200,000 copies

- MIR (Mammalian Interspersed Repeat)

- LTR/Retroviral

- Other

- Genes that are duplicated & then diverge (paralogs)

- Recent duplications, ~100,000-long, very similar copies

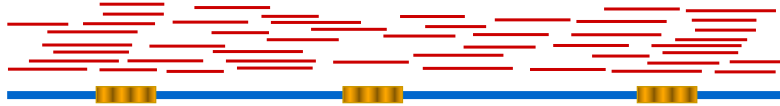
Computational Molecular Biology and Genomics, Spring 2009

14

## What can we do about repeats?

Two main approaches:

- Cluster the reads

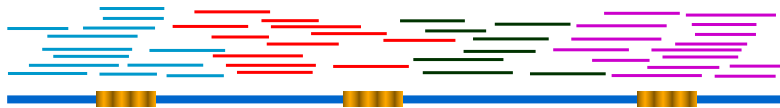


- Link the reads

## What can we do about repeats?

Two main approaches:

- Cluster the reads

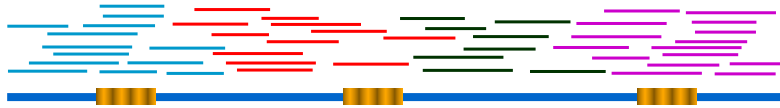


- Link the reads

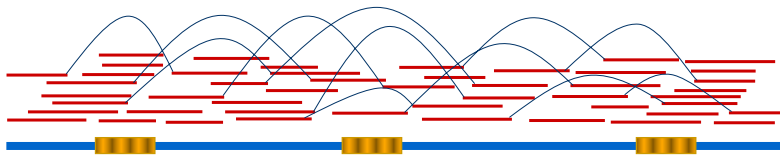
## What can we do about repeats?

Two main approaches:

- Cluster the reads



- Link the reads

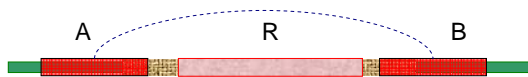
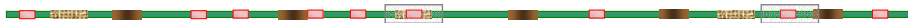


## Sequencing and Fragment Assembly

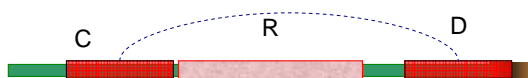


```
AGTAGCACAGA
CTACGACGAGA
CGATCGTGCGA
GCGACGGCGTA
GTGTCTGTATC
TGTCTGTGTGT
TGTACTCTCCT
```

$3 \times 10^9$  nucleotides

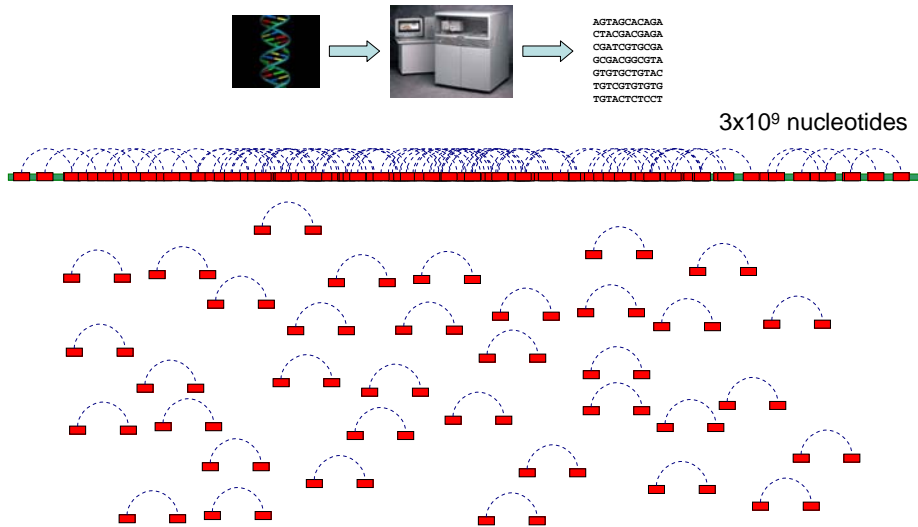


ARB, CRD



or  
~~ARD, CRB ?~~

## Sequencing and fragment assembly



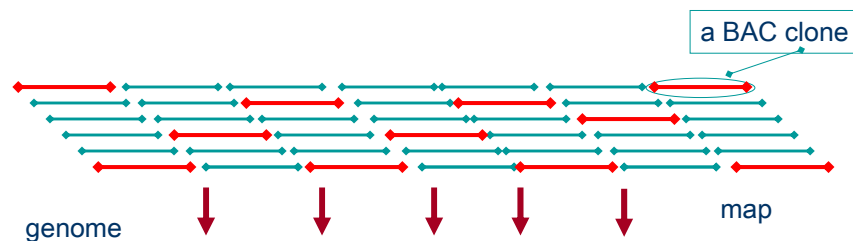
## Strategies for whole-genome sequencing

- Hierarchical – **Clone-by-clone**
  - Break the genome into long pieces
  - Map each long pieces onto the genome
  - Sequence each piece with shotgun
  - Example: Yeast, worm, rat, human
- Online version of Clone-by-clone – **Walking**
  - Break the genome into long pieces
  - Start sequencing each piece with shotgun
  - Construct the map as you go
  - Example: Rice genome
- Whole genome shotgun
  - One large shotgun pass on the whole genome
  - Example: Drosophila, Human (Celera), Neurospora, Mouse, Rat and Dog



# Hierarchical Sequencing

## Hierarchical Sequencing Strategy



1. Obtain a large collection of BAC clones
2. Map them onto the genome – **Physical Mapping**
3. Select a minimum tiling path
4. Sequence each clone in the path with shotgun
5. Assemble
6. Put everything together

## Methods of physical mapping



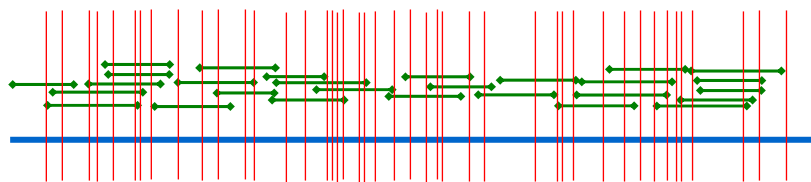
### Goal:

Make a map of the locations of each clone relative to one another  
Use the map to select a minimal set of clones to sequence

### Methods:

- Restriction mapping (digestion)
  - mapping of restriction sites of a cutting enzyme based on lengths of fragments
- Hybridization
  - mapping clones based on hybridization data with probes

## Physical mapping I – Digestion



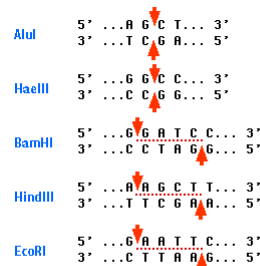
Restriction enzymes cut DNA where specific words appear

### Partial digestion:

1. Cut each clone separately with an enzyme
2. Run fragments on a gel and measure length
3. Clones  $C_a$ ,  $C_b$  have fragments of length  $\{l_i, l_j, l_k\} \Rightarrow$  overlap

### Double digestion:

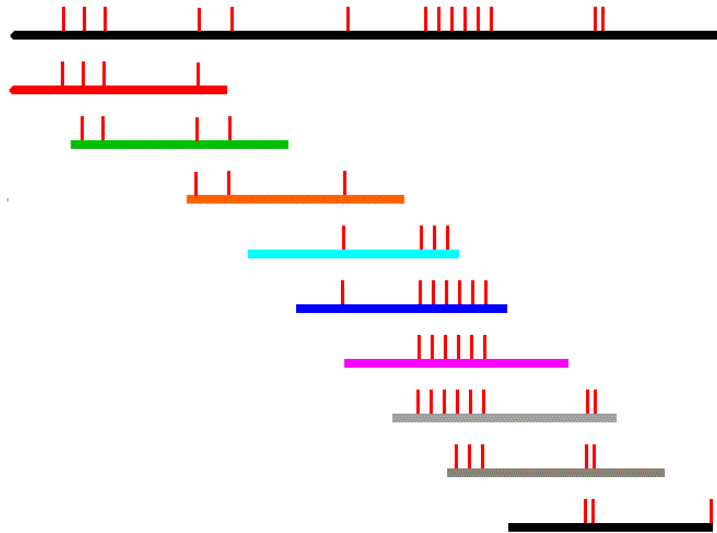
Cut with enzyme A, enzyme B, then enzymes A + B



**AluI** and **HaeIII** produce blunt ends

**BamHI** **HindIII** and **EcoRI** produce "sticky" ends

## Hybridization

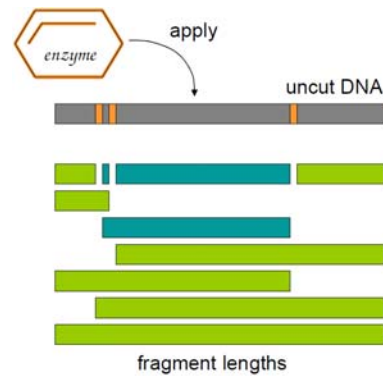


## Physical mapping I – Digestion

- Restriction map
    - The locations of the restriction sites of a given enzyme on the target DNA
  - Restriction mapping problem
    - Given a subset  $E \subseteq \Delta X$ , construct  $X$  from  $E$ .
    - $\Delta X = \{ |x_i - x_j| : x_i, x_j \in X \}$
    - $|\Delta X| ?$
1. Partial digest problem (PDP): one enzyme ( $E = \Delta X$ )
  2. Double digest problem (DDP): two enzymes

## Partial digest

- One restriction enzyme only
- Obtain fragments of all possible lengths



## Partial digest problem

- Given  $E = \Delta X$ , construct  $X$  from  $E$ .
- Example:  $E = \{3, 11, 17, 19, 8, 14, 16, 6, 8, 2\}$



- No polynomial-time algorithm is yet known
- Experimental errors
  - There is uncertainty in length measurement by gel electrophoresis (5% error)
  - May lost some fragments in the digestion process (gaps occur)

## Partial digestion problem

- No polynomial time algorithm known
- Not known to be NP-complete
- Practical Backtracking algorithm by Skiena et al, 1998

## The algorithm

- Find the longest distance in  $\Delta X$ , and delete that distance from  $\Delta X$
- Repeatedly position the longest remaining distance of  $\Delta X$ 
  - Determine between two possible positions (one of the two outermost points) for the point
  - For each of these two points, check whether all the distances from the position to the points already positioned are in  $\Delta X$
  - If they are, delete all those distances from  $\Delta X$  and proceed
  - Backtrack if they are not for both of the two positions.

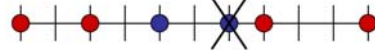
## Example

- $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$

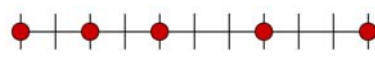
1.  $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$



2.  $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$     4.  $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$

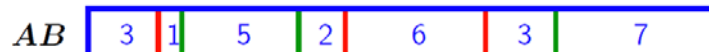


3.  $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$     5.  $\Delta X = \{2, 2, 3, 3, 4, 5, 6, 7, 8, 10\}$



## Double digest problem

- Example:  $A = \{3, 6, 8, 10\}$ ,  $B = \{4, 5, 7, 11\}$  and  $AB = \{1, 2, 3, 3, 5, 6, 7\}$



- Many equivalent solutions
- NP-complete (by Goldstein and Waterman, 1987)

## Acknowledgement

---

- This set of slides is based on the slides by:
  - Serafim Batzoglou (Stanford Univ)
  - Richard C.T. Lee (National Chi Nan Univ)