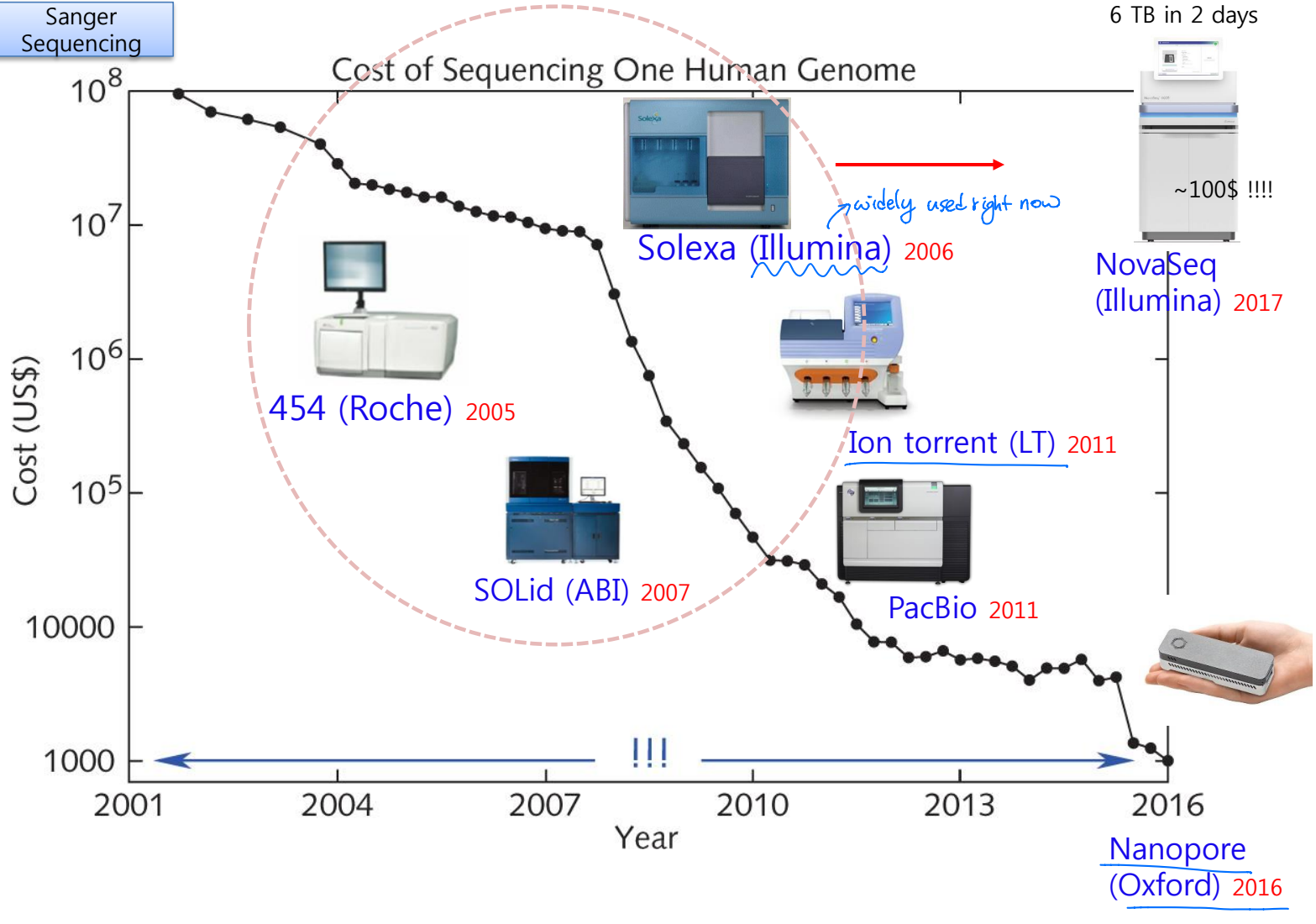
Summary & More

: Next-Generation Sequencing (NGS)

Sung Wook Chi
Division of Life Sciences, Korea University

Next-generation Sequencing (NGS)

Sanger Sequencing

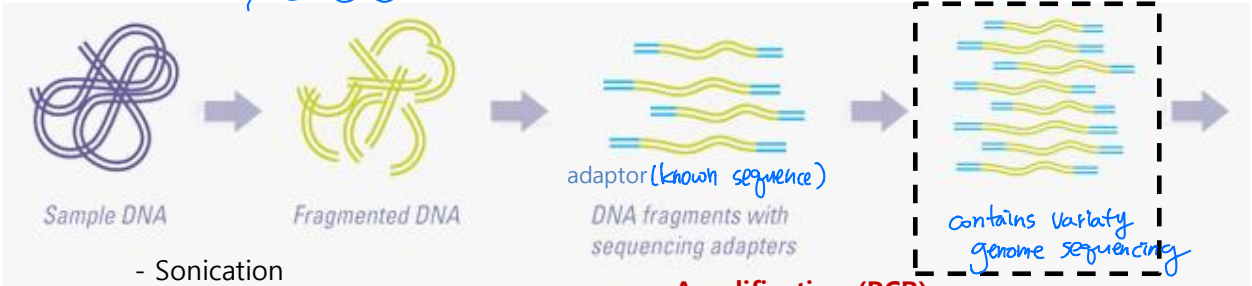


DNA Sequencing by NGS (flow & principle)

Preparation

Genome → **Fragmentation** → **Adaptor ligation** → **Sequencing Library**

genome을 분석하는 library



- Sonication
- DNase

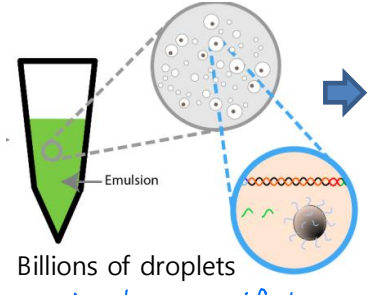
Amplification (PCR)

- Depending on NGS platform

NGS sequencing platform (Sequencer)

454 (Roche) 2005

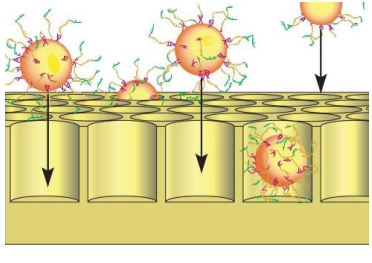
Emulsion PCR



Billions of droplets

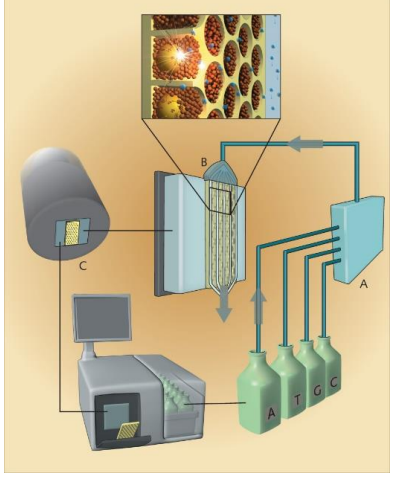
independently amplified in separated droplet

Bead deposit



Pyrosequencing

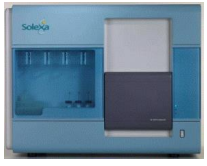
detect pyro phosphate → light



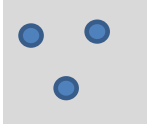
Illumina NGS platform (Solexa)

SBS
(sequencing by synthesis)
utilize polymerase

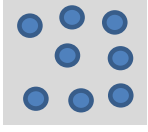
Library loading



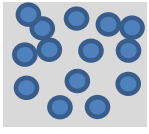
attach template



Sparse
(low number of reads)

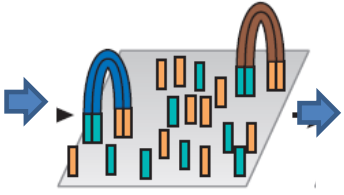


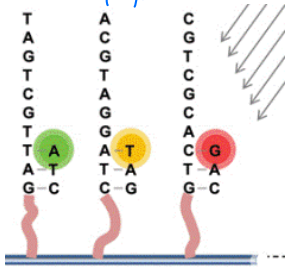
Optimal



crowded
(Interfering of signal: low quality)

Bridge amplification





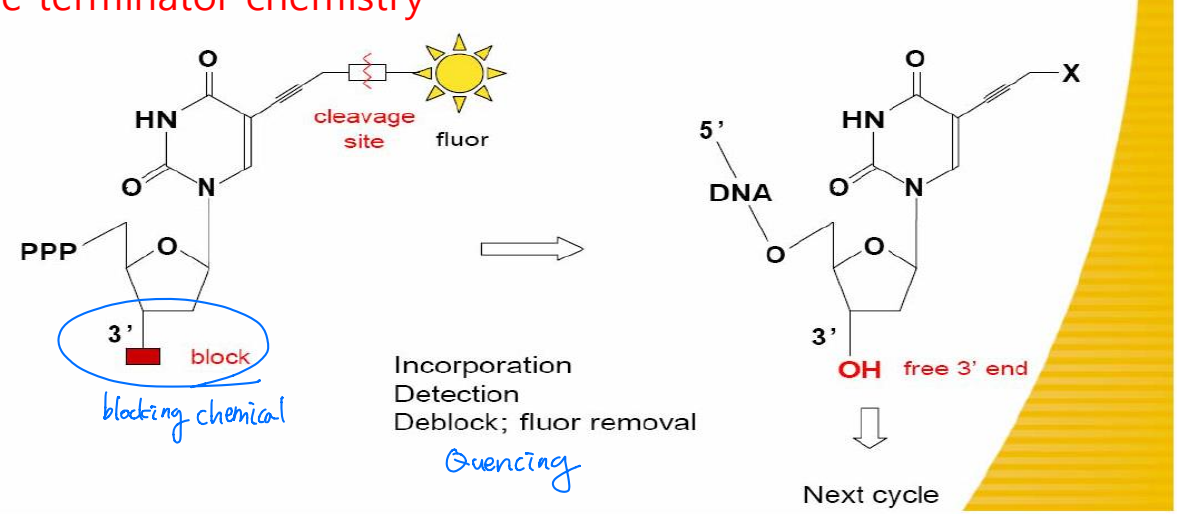
Solid phase amplification
(primer is already attatched on plate)

Reversible termination chemistry

Solexa (Illumina)
2006

Loading (spread) your library on glass plate (flow cell)

Reversible terminator chemistry

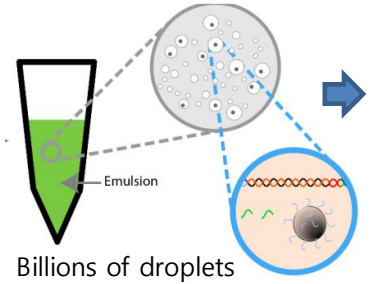


2nd Generation of NGS sequencer

Ion torrent (LT) 2011

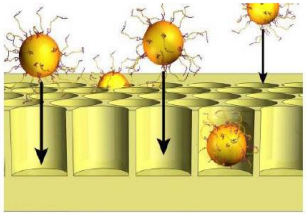
454경 배수

Emulsion PCR

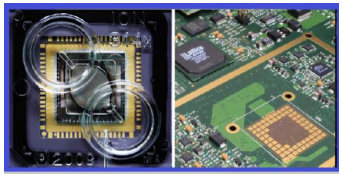


Billions of droplets

Bead deposit



Semiconductor
: PH detection

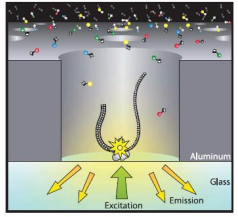


Electric signal

PacBio 2011

Single molecule

Real time sequencing



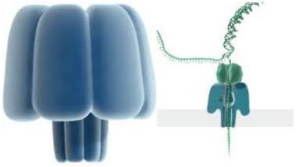
DNA polymerase

Light (recording color change)
(image process)

Nanopore
(Oxford) 2016

Single molecule

Real time sequencing

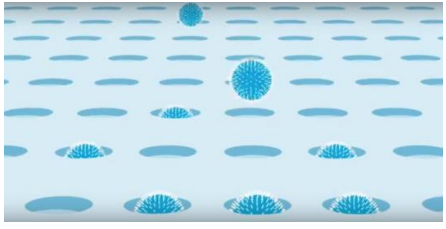


nanopore

Electric signal

Ion torrent/proton: semiconductor sequencing

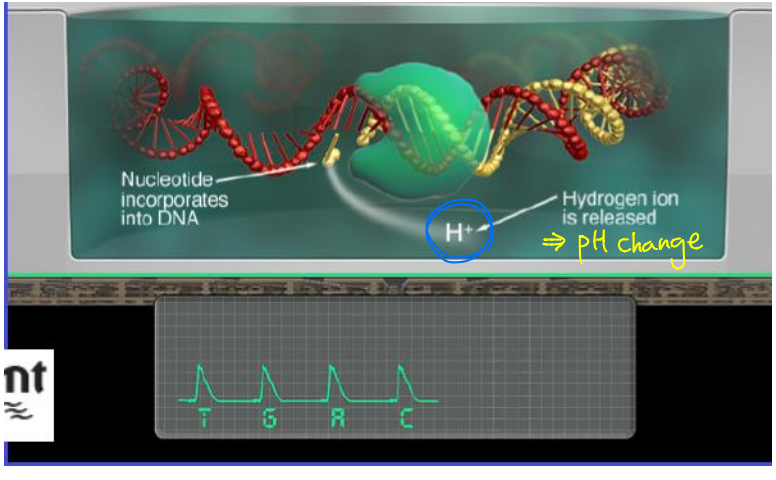
Emulsion PCR



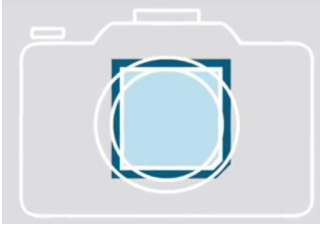
Electric signal
(No camera !!, low price)

1.5 million pH meters semiconductor chip

Ion torrent



Ion proton



80 million pH meters semiconductor chip

→ camera가 pixel을 detect → signal
하는것과 같은 chip이지만,
pH정보 detect 하도록 개발된 chip

monocolor

DNA polymerase reaction

Release of proton (H+) > change in pH > detecting as electric signal

454랑 같음 (A만 넣고 signal 나오는지 보고, →washing → T만 넣고 signal 나오는지 보고)
AA2개면 pH change signal이 2배!

장점: 염기정보를 빠르게 추정가능

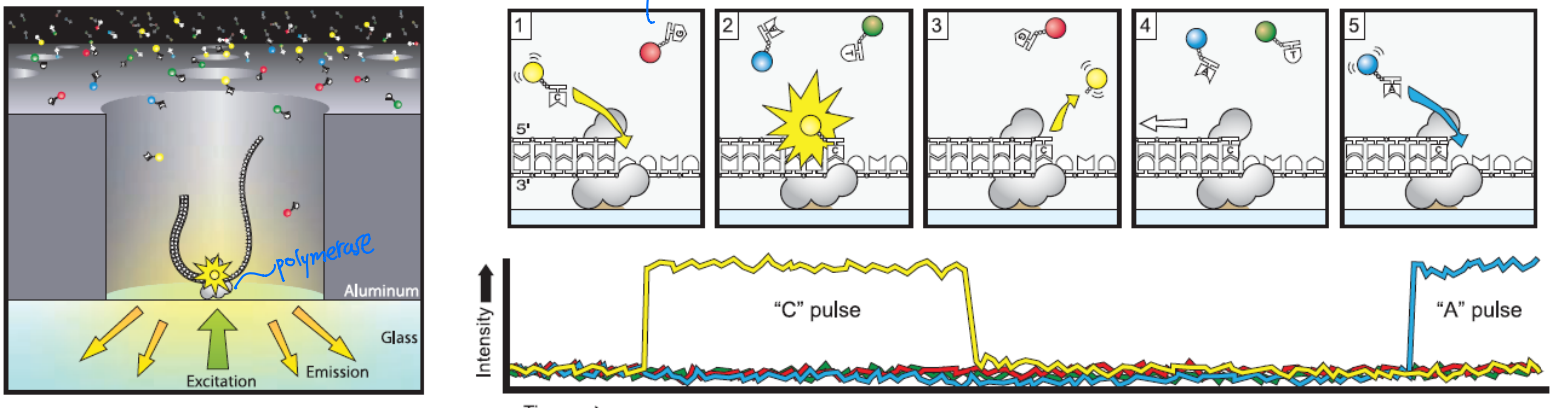
https://www.youtube.com/watch?v=WYBzbxIfuKs 이해하기좋음

(The same problems in
- Emulsion PCR, Pyrosequencing

Pacific Biosciences

No amplification (Single molecule)
Long reads (3kb ~ 15kb) ,
Fast (real time, 30 min)

but **low accuracy** (95%)
low throughput (70,000 reads)

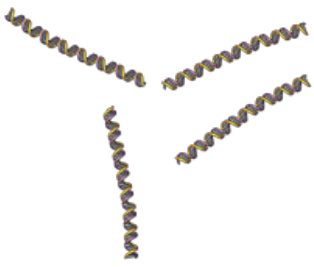


Single Molecule, Real Time Sequencing

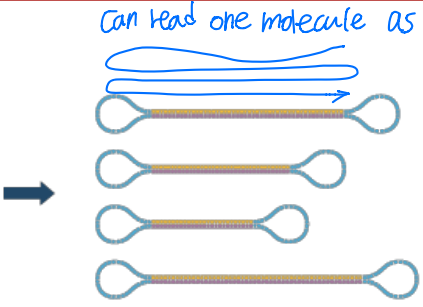
https://www.youtube.com/watch?v=v8p4ph2MAvI

Useful for long read sequencing
(resequencing genomes)

SMRT (Single Molecule RealTime) Sequncing : PacBio



modified



can read one molecule as many as we want

repeatable reading 가능

Library Preparation

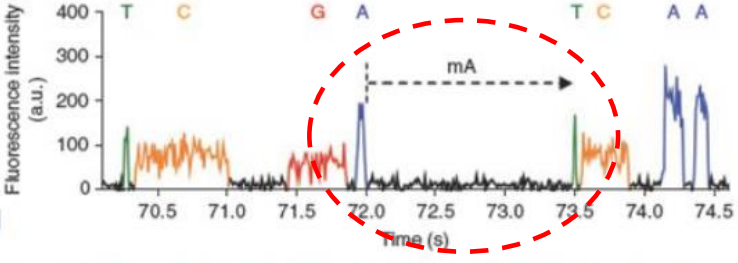
- Universal SMRTbell template accepts insert sizes from 250 bp to 40 kb

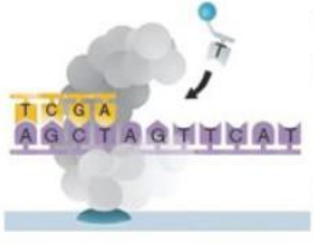
long read 가능

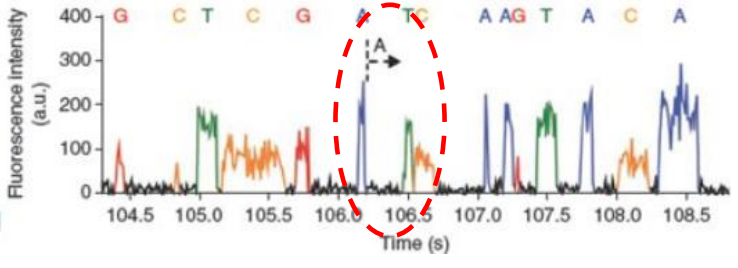
Redundant Sequencing Reading

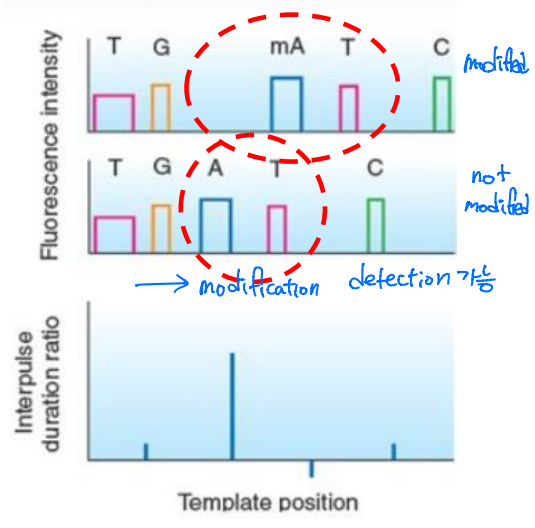
Consensus accuracies > 99.999%

Realtime direct dectection of modified DNA





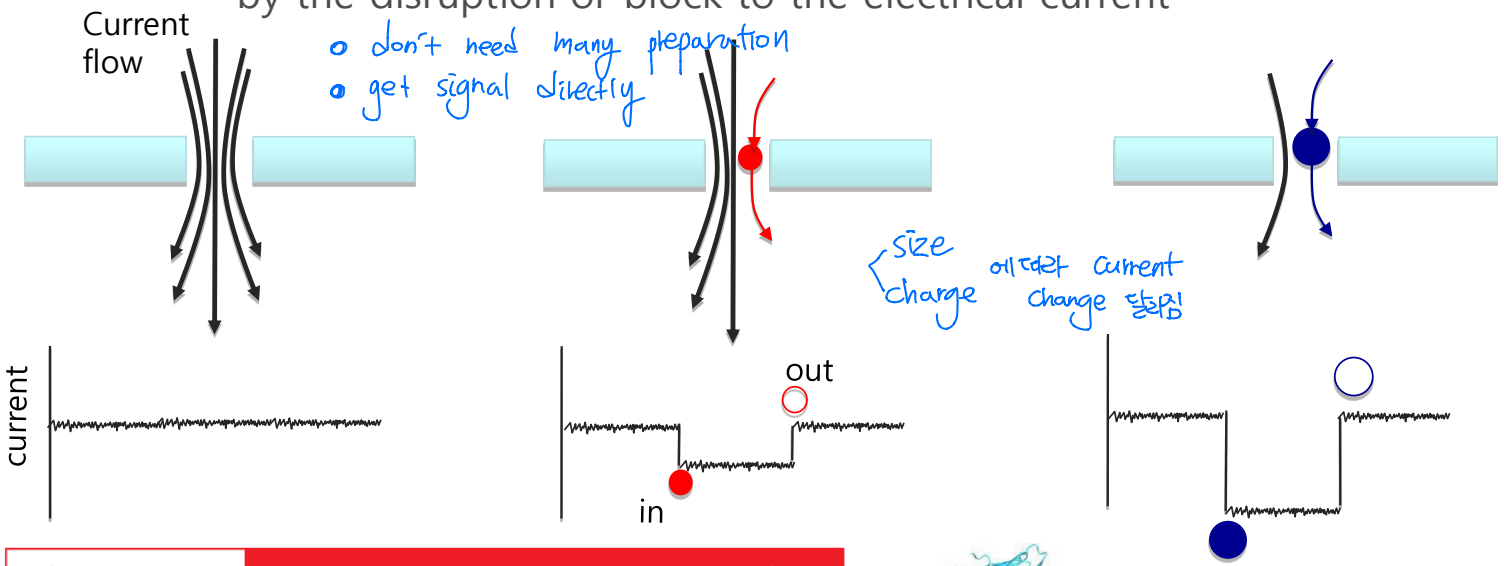




Oxford Nanopore

Electronic & Single molecule Real time

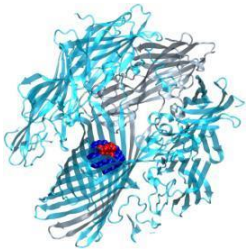
- Nanopore = 'very small hole' (channel을 응용해서 만든것)
- Electrical current flows through the hole
- Introduce analyte of interest into the hole ➔ identify "analyte" by the disruption or block to the electrical current



nature nanotechnology
ARTICLES
PUBLISHED ONLINE: XX XX 2009 | DOI: 10.1038/NNANO.2009.12

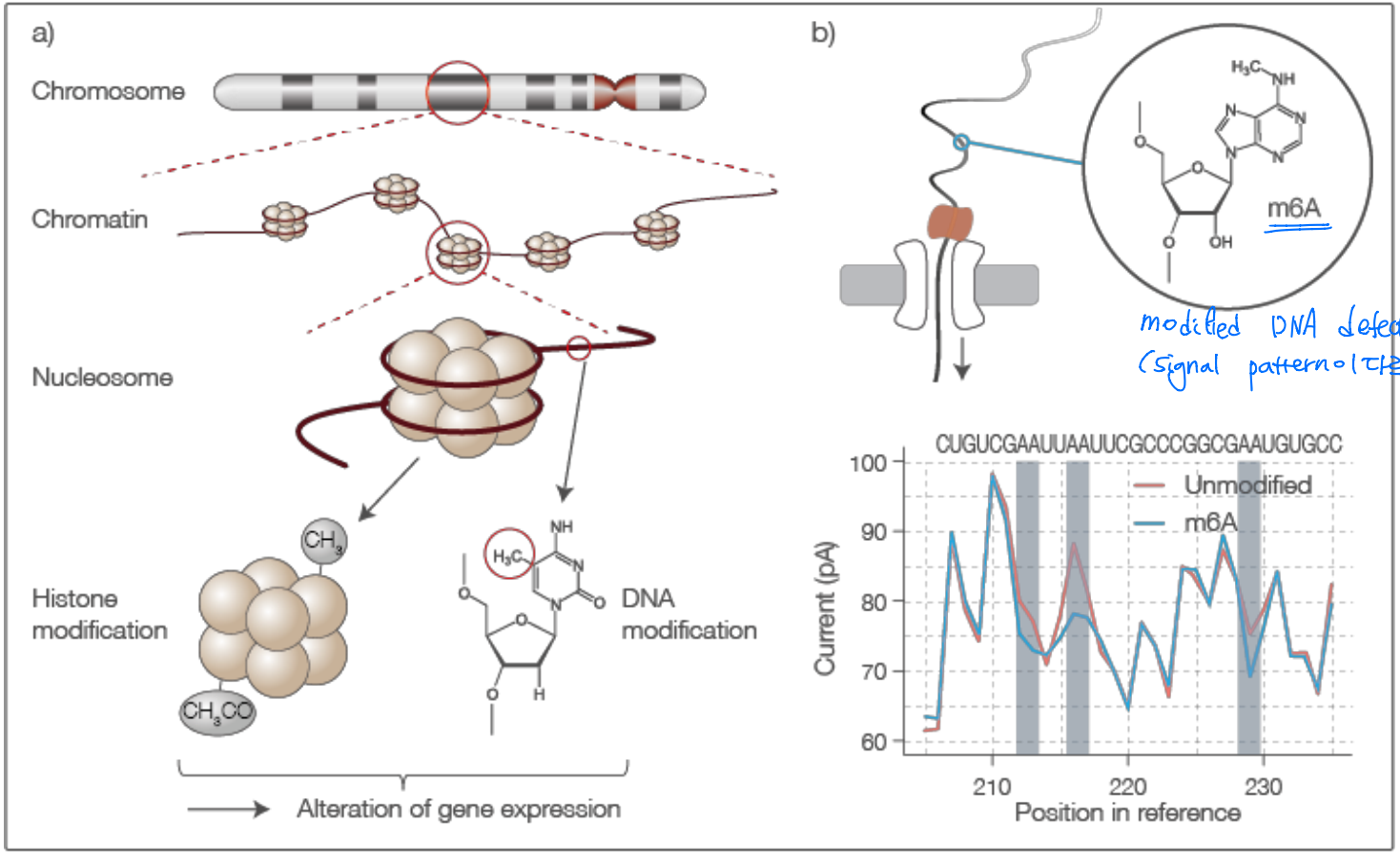
Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke[1], Hai-Chen Wu[2], Lakmal Jayasinghe[1,2], Alpesh Patel[1], Stuart Reid[1] and Hagan Bayley[2]*



https://www.youtube.com/watch?v=BNz880V52rQ

Direct detection of modified DNA/RNA by Oxford Nanopore

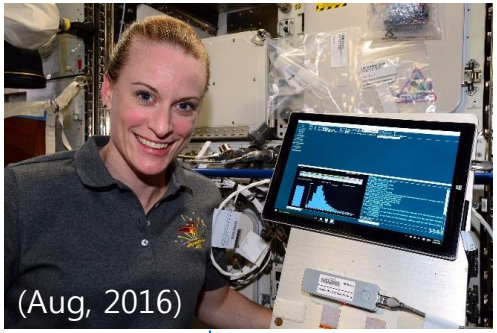


Realtime Portable Sequencer: Oxford Nanopore

Oxford Nanopore (MinION)

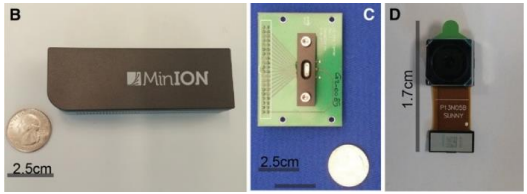
아프리카 등에서 사용

9th NASA/SpaceX commercial cargo



space ship에 실려서 우주에서 사용

minimizing



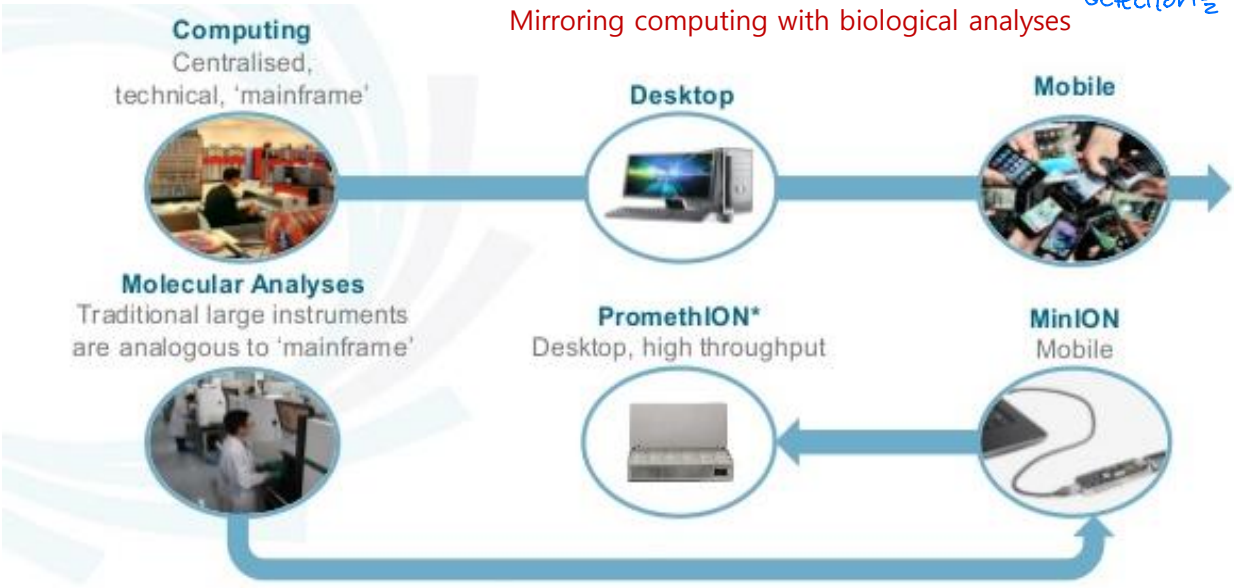
IoT (Internet of Thing)

⇒ IoLT (Internet of Living Thing)

ex) micro-organism 이 갖는 realtime detection을 통한 예측효과

Creation of internet of living things (DNA)

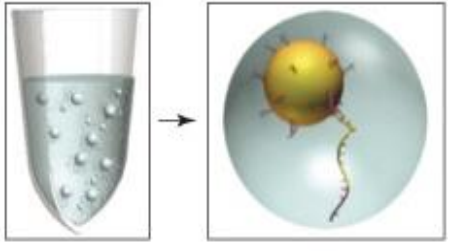
Mirroring computing with biological analyses



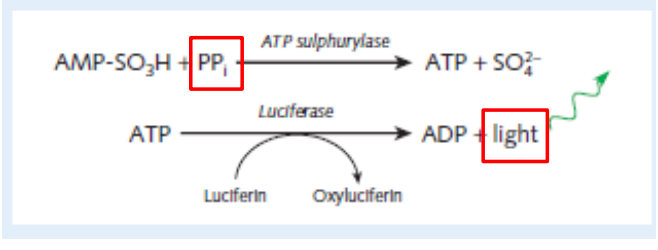
NGS platform comparison

Platform	Amplification	Chemistry	Detection	description
454	emPCR	Pyrosequencing	Image (mono)	Low throuput, long read (~400), ~0.35 day
Abi SOLiD	emPCR	Seq by ligation	Image (color)	High throuput, short read (~50), **~2weeks**
Illumina /Solexa	Solid-phase	Reversible teminator	Image (color)	High throuput, short read (~50), ~1 week
Ion torrent /proton	emPCR	Pyrosequencing	H+ (pH)	low-High throuput, short read (~50), 5 hours
Pacific Bioscience	Single molecule	realtime, polymerase	Image (color)	**longer read (3-15kb), Accuracy (~95%)**, 30min
Oxford Nanopore	Single molecule	realtime, polymerase	H+ (pH)	?

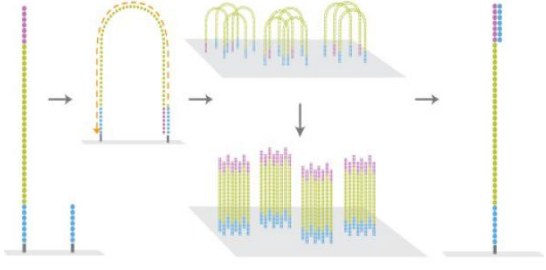
Emulsion PCR



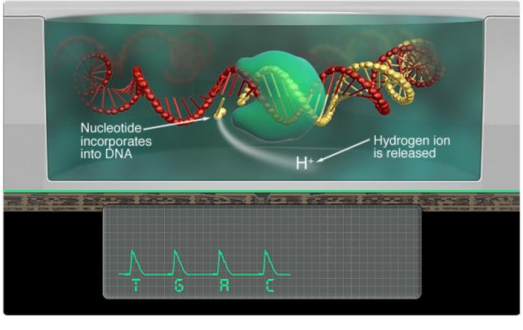
pyrosequencing



Solid-phase (bridge amplification)



Proton (pH meter)

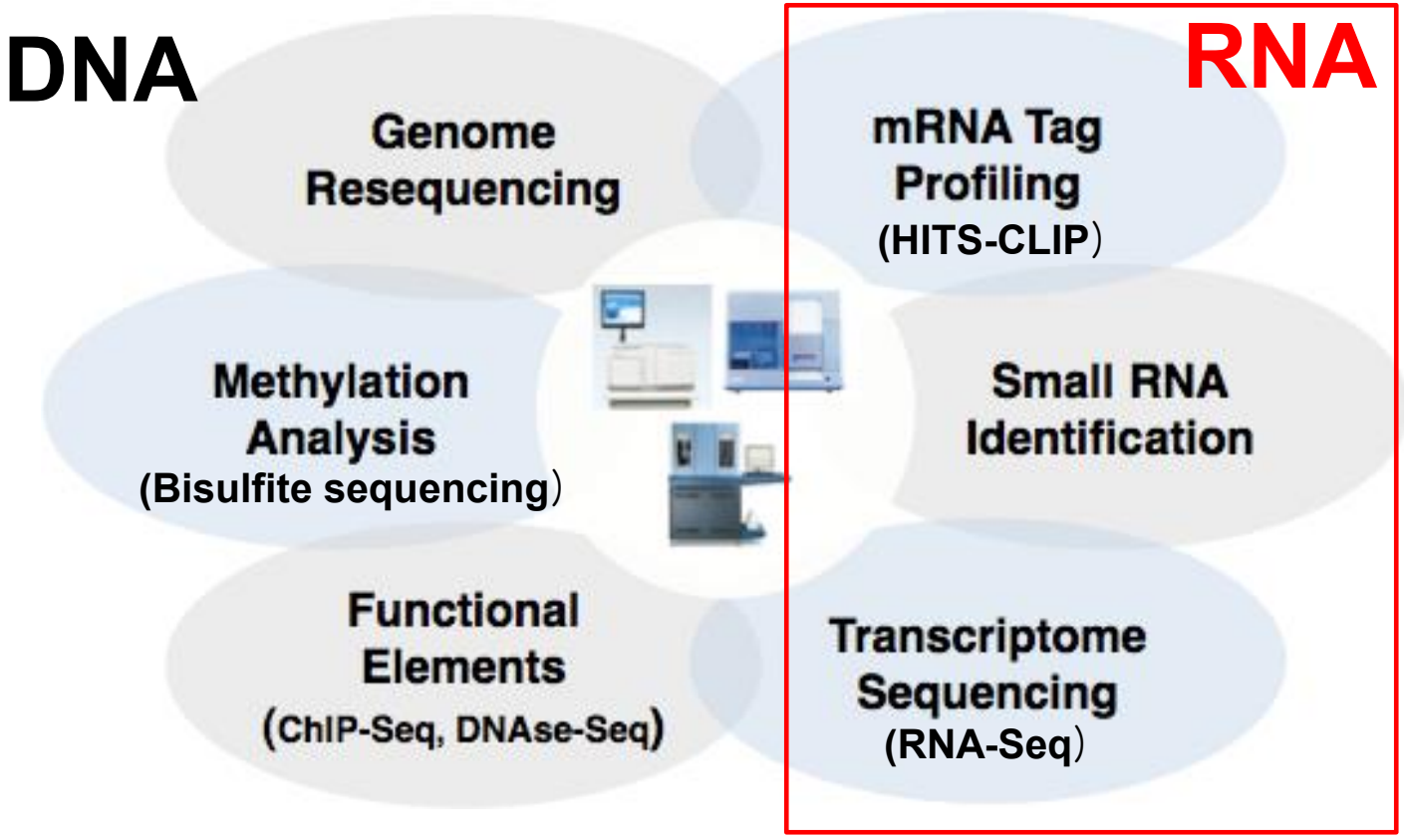


NGS platform comparison

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (US$)
Sanger ABI 3730×1	First	600–1000	0.001	96	0.5–3 h	500
Ion Torrent	Second	200	1	8.2×10^7	2–4 h	0.1
454 (Roche) GS FLX+	Second	700	1	1×10^6	23 h	8.57
Illumina HiSeq 2500 (High Output)	Second	2 × 125	0.1	8×10^9 (paired)	7–60 h	0.03
Illumina HiSeq 2500 (Rapid Run)	Second	2 × 250	0.1	1.2×10^9 (paired)	1–6 days	0.04
SOLiD 5500×l	Second	2 × 60	5	8×10^8	6 days	0.11
PacBio RS II: P6-C4	Third	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80
Oxford Nanopore MinION	Third	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90

From: Rhoads, A. and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics, *Proteomics & Bioinformatics*, **13**, 278–289. (This article gives citations of sources of data.)

Application of Next-Generation Sequencing

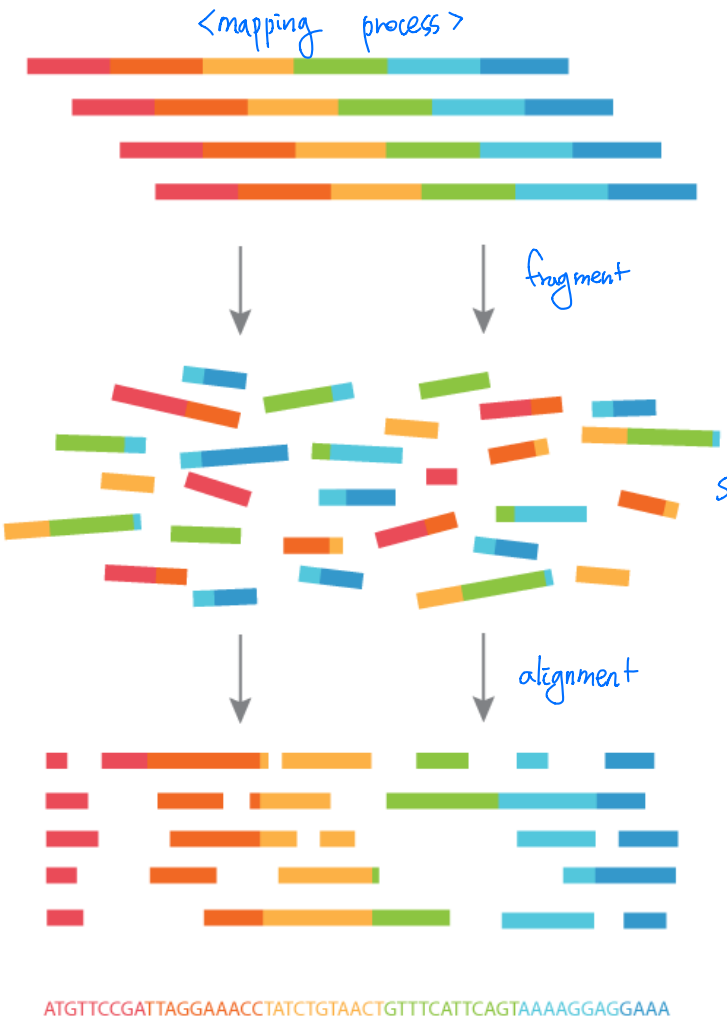


Genome Annotation

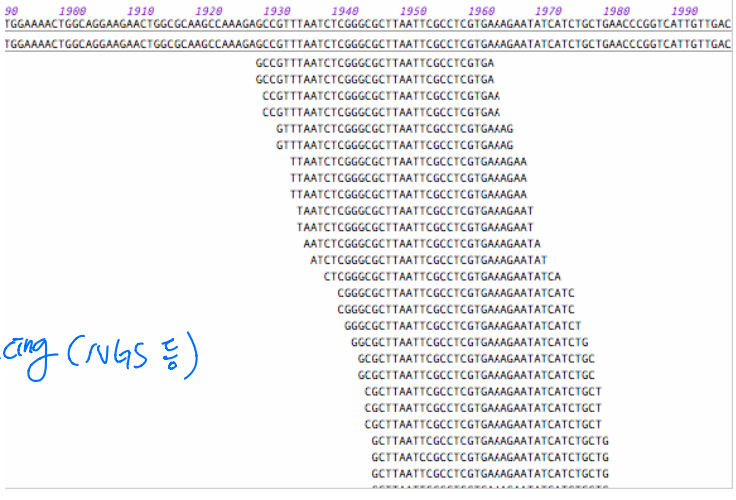
- Sequence alignment

Sung Wook Chi
Division of Life Sciences, Korea University

Mapping reads (NGS) : Sequence Alignment



↳ mapping our sequence into known sequence



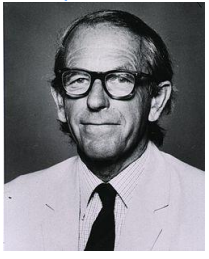
Pairwise sequence alignment

Sequence Alignment : history

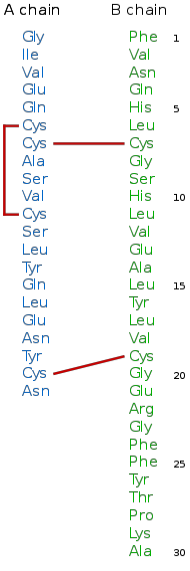
Amino acid sequence

insulin (1951)

first sequenced amino acid



Fred Sanger
(1918-2013)



Protein sequence atlas (1960)

mother of bioinformatics

Margaret Dayhoff
(1925-1983)

Cytochrome protein
(Families & Superfamilies)

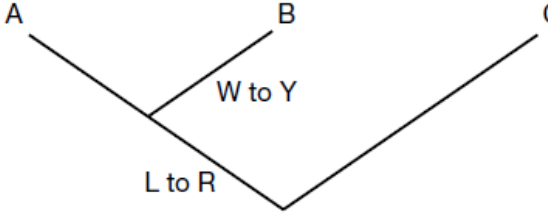
→ So many isoform & spread out through species
→ homolog 들을 모두 찾아냄

A	A	W	T	V	A	S	A	V	R	T	S	I
B	A	Y	T	V	A	A	A	V	R	T	S	I
C	A	W	T	V	A	A	A	V	L	T	S	I

→ sequence 약간은 다르지만 same job
→ biology system이 약간의 mutation은 allow

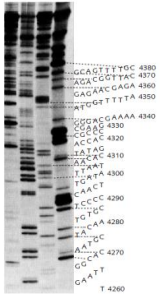
- Phylogenetic tree
:Evolutional and functional relationship
:Substitution matrix (PAM), one letter code

accepted mutation frequency
(allowed mutation through evolution)



- DNA sequence (Sanger Sequencing)
(bacteriophage, 1977)

Fred Sanger



Sequence analysis > Sequence alignment

Biological question from sequence alignment

ACGCTGA

(do same function)
Common Ancestor

ACTGT

changed through evolution process

don't exactly know what happened

Evolutional Changes

alignment can tell you about history

ACGCTGA

2 insertion, 1 substitution

```
ACGCTGA
A--CTGT
```

Sequence alignment 1

ACTGT

ACGCTGA

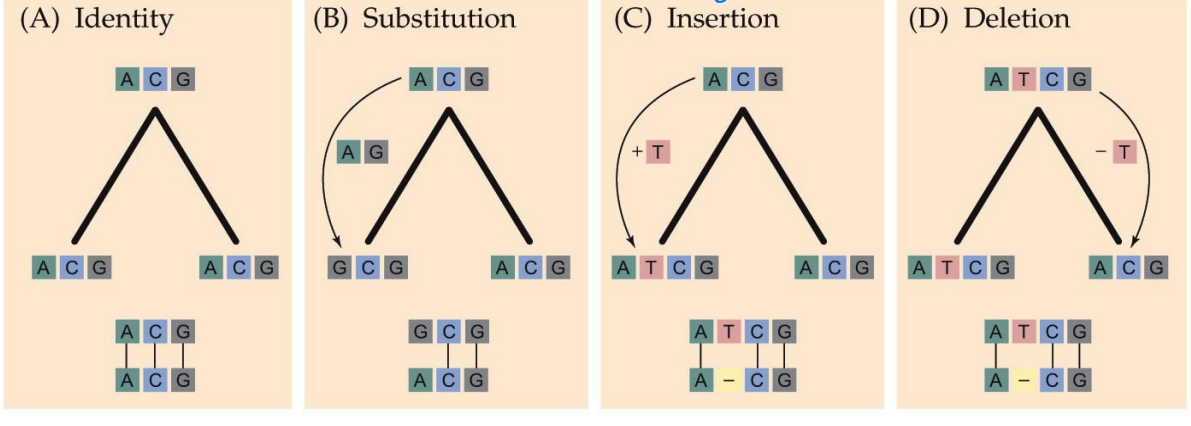
2 substitution, 2 insertion

```
ACGCTGA
ACTGT--
```

Sequence alignment 2

ACTGT

↝ alignment 에 따라 어떻게 변했는지가 달라짐



tell you most likely change

What is the best optimal alignment?

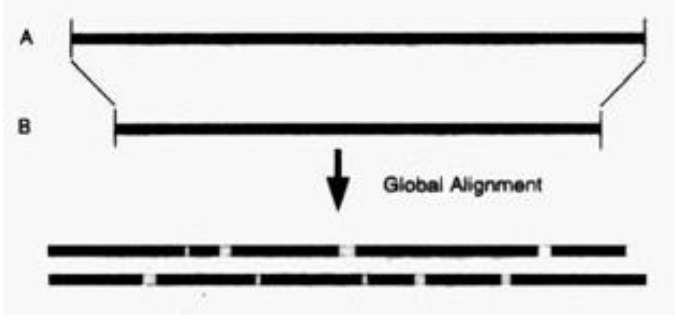
Which alignment is biologically relevant?

: purpose of alignment

Two types of sequence alignment : local & global alignment

- **Sequence Alignment**

  - Pairwise alignment
  - Multiple Sequence alignment



Global Alignment same start / last position

–Suite for similar sequences

–Nearly equal legnth

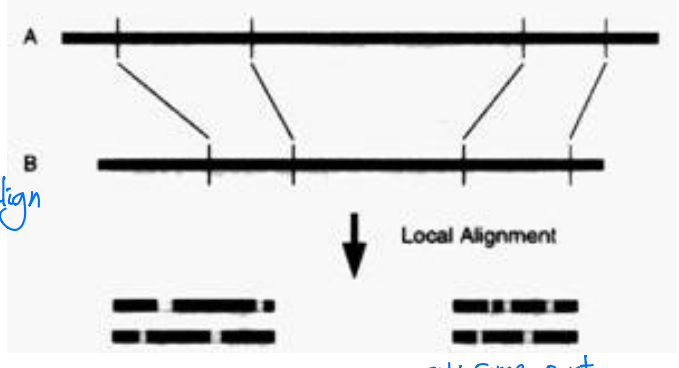
– Overall similarity is detected

Local Alignment regardless of start / last position → local part가 very similar 하면 align

–Isolate regions in sequences

–Suitable for database searching

–Easy to detect repeats



only some part

ex) kinase ⇒ 전체 염기서열 but same job → same sequence 부분적으로 (local 이 high similarity.)

High sequence similarity

> Homolog > same function

	Alignment 1	Alignment 2
Sequence 1	ACGCTGA	ACGCTGA
Sequence 2	A--CTGT	ACTGT--

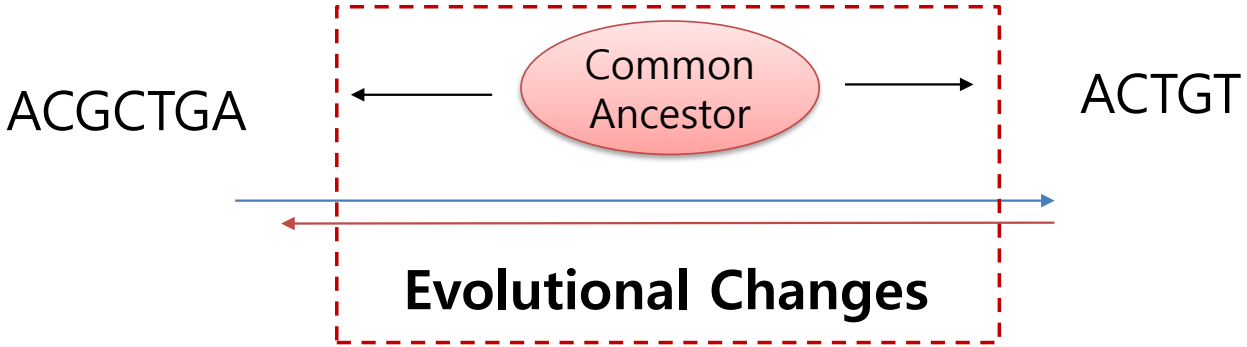
Conserved region of sequence

> Similar function

Global alignment

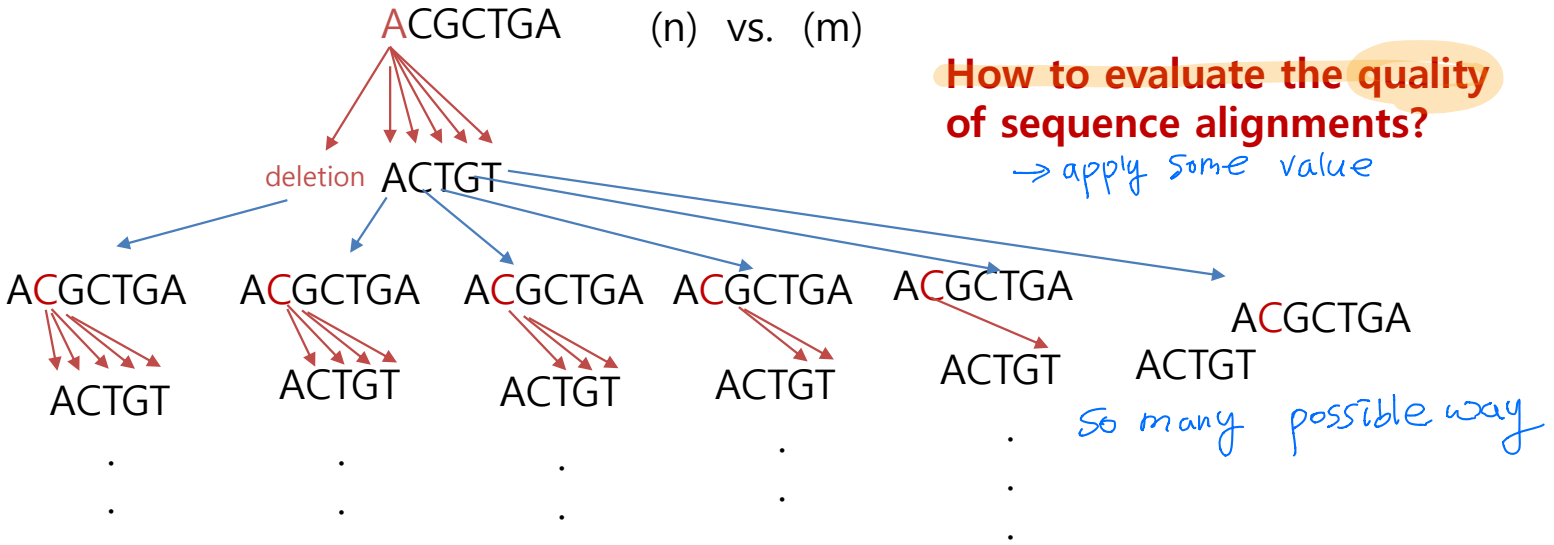
Local alignment

Sequence Alignment : Bioinformatics methods



All combination of sequence alignments: Which one is the optimal ?

all possible combination 경우의 수



How to evaluate the quality of sequence alignments?

→ apply some value

So many possible way

Need to have similarity or distance value for evaluating sequence alignments.

How to measure sequence similarity : distance

how different these two seq

- (Method 1) Counting identical letters on each position

→ **Hamming distance** : Number of position with mismatch characters

*distance는 작을수록 good
score는 클수록 good

- (Method 2) Inserting gaps to maximize the number of identical letters

→ **Edit distance** (Levenshtein distance) : How many opperations are required ?

```
A T C C G A T
  |   |   | |
T G C – A T A T
```

```
A T C C G A T
    |     | |
T G C A T A T
```

- x="TGCATAT" (m=7), y="ATCCGAT" (n=7)

TGCATAT → ATGCATAT : insertion of "A" (1)
ATGCATAT → ATCCATAT : substitution of "G" with "C" (2)
ATCCATAT → ATCCGATAT : insertion of "G" (3)
ATCCGATAT → ATCCGATT : deletion of "A" (4)
ATCCGATT → ATCCGAT : deletion of "T" (5)

✓ Hamming distance = 3 better

Edit distance = 5

- x="TGCATAT" (m=7), y="ATCCGAT" (n=7)

TGCATAT → ATGCATAT : insertion of "A" (1)
ATGCATAT → ATGCAAT : deletion of "T" (2)
ATGCAAT → ATCCAAT : substitute of "G" with "C" (3)
ATCCAAT → ATCCGAT : substitute of "A" with "G" (4)

✓ Hamming distance = 4

Edit distance = 4 better

How to measure sequence similarity : Score

give score　minus score

Score = (match or mismatch penalty) – gap penalty

Match = 3 , Mismatch = -1

G→A : 퓨린→퓨린 Gap= -2

higher score than 퓨린→피리미딘

	A	G	T	C
A	20	10	5	5
G	10	20	5	5
T	5	5	20	10
C	5	5	10	20

Affine gap penalty

(gap opening + (gap extension) x (n-1))

ACGCTGA

A--CTGT

Gap opening, Gap extension penalty

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

estimated by frequency ⇒ more frequently observed in same family ⇒ higher score

rare observed ⇒ high penalty

Substitution matrix

ORGANISM A　A W T V A S A V R T S I

ORGANISM B　A Y T V A A A V R T S I

ORGANISM C　A W T V A A A V L T S I

Log-odds score

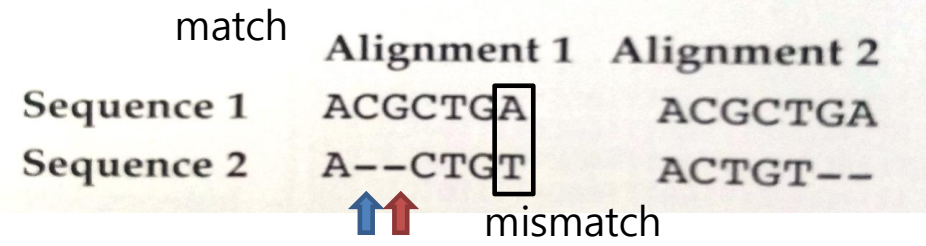
$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{observed\ frequency}{expected\ frequency}$$

-BLOSUM62: Pairs frequencies were counted between segments less than 62% identical.

The PAM1 matrix estimates what rate of substitution would be expected if 1% of the amino acids had changed

Sequence Alignment

Pairwise sequence alignment



Optimal sequence alignment

depending on score, program can decide which alignment is better

1. Evaluation of sequence similarity : **Distance**, **Score**
2. Performing optimal sequence alignment search: **Dynamic Programming**

alignment 가능한 모든 경우의 program 이용

➢ **Global Alignment**

```
TCAG--T-GTCGAAGT-TA
|  ||   |  |  |  |  ||  |
T-AGGCTAG-C-A-GTGTA
```

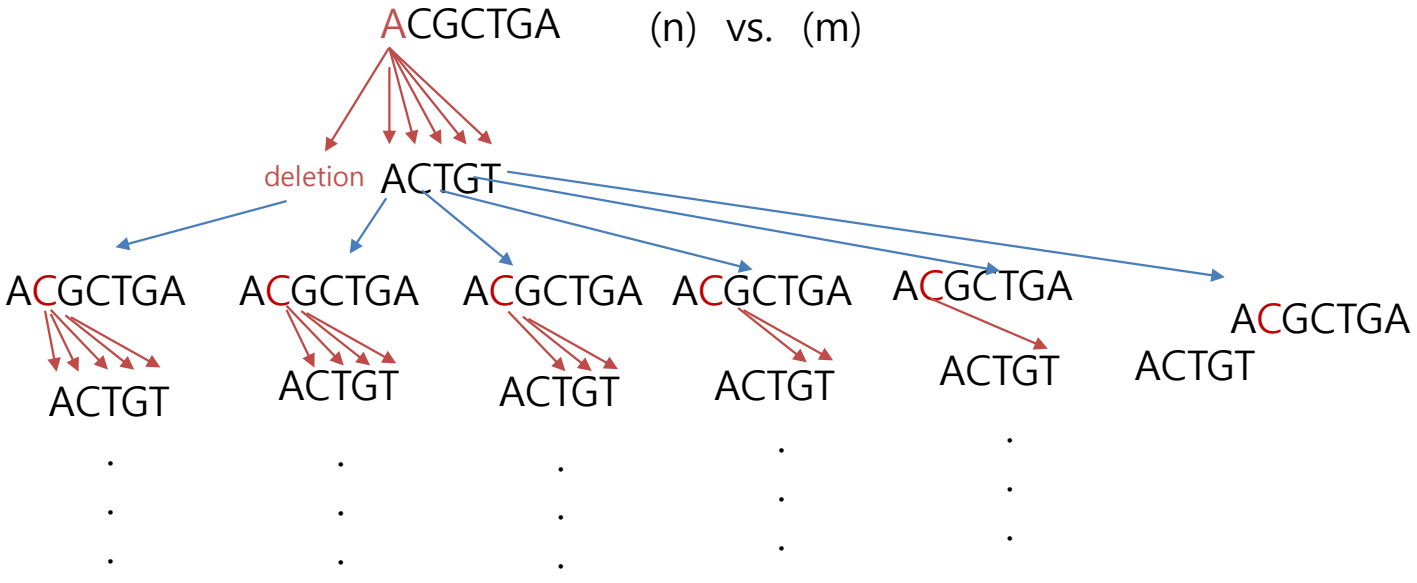
➢ **Local Alignment**

```
        TCAGTGTCGAAGTTA
          ||||||
TAGGCTAGCAGTGTA
```

Sequence Alignment : Find the optimal alignment

All combination of sequence alignments:

Too many operations, and storage !!!

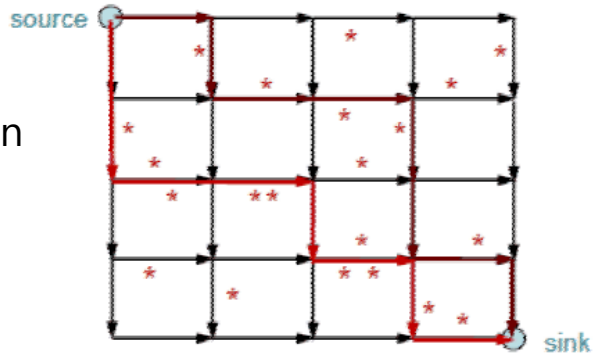


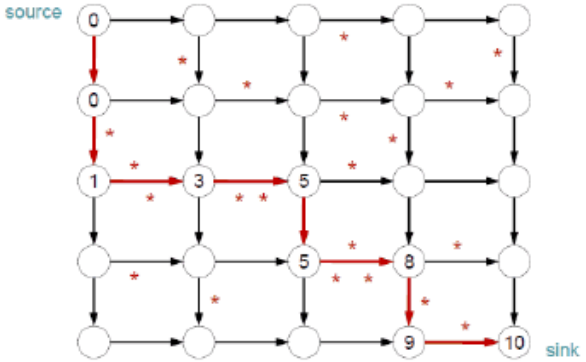
Solution : The longest Common Path Problem

Dynamic Programming

Finding optimal alignment: finding shortest distance

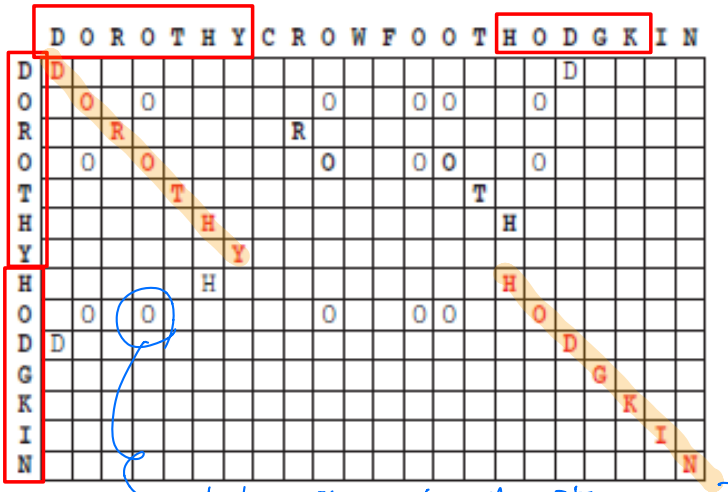
Manhattan Tourist Problem



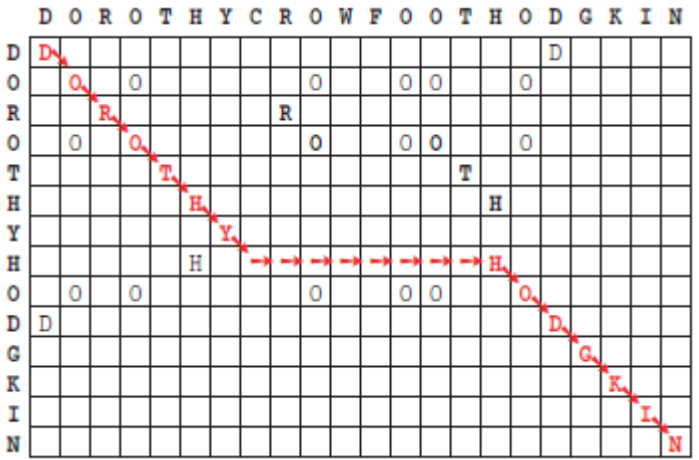


DOROTHY--------HODGKIN
DOROTHYCROWFOOTHODGKIN

- A dotplot shows perspicuously the quality and distribution of the pattern of similarity between two sequences. Each possible alignment of the two sequences corresponds to a path through the dotplot, from upper left to lower right.



marked position : x와 y가 같음

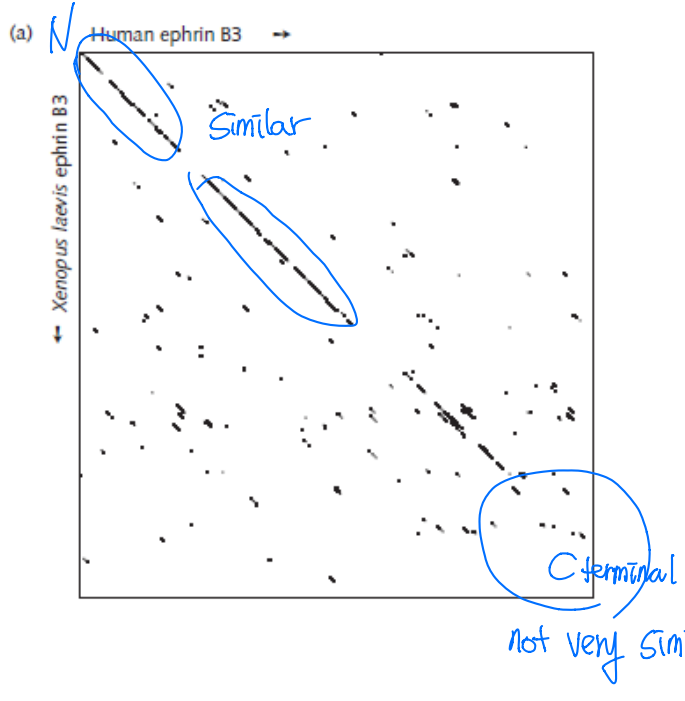


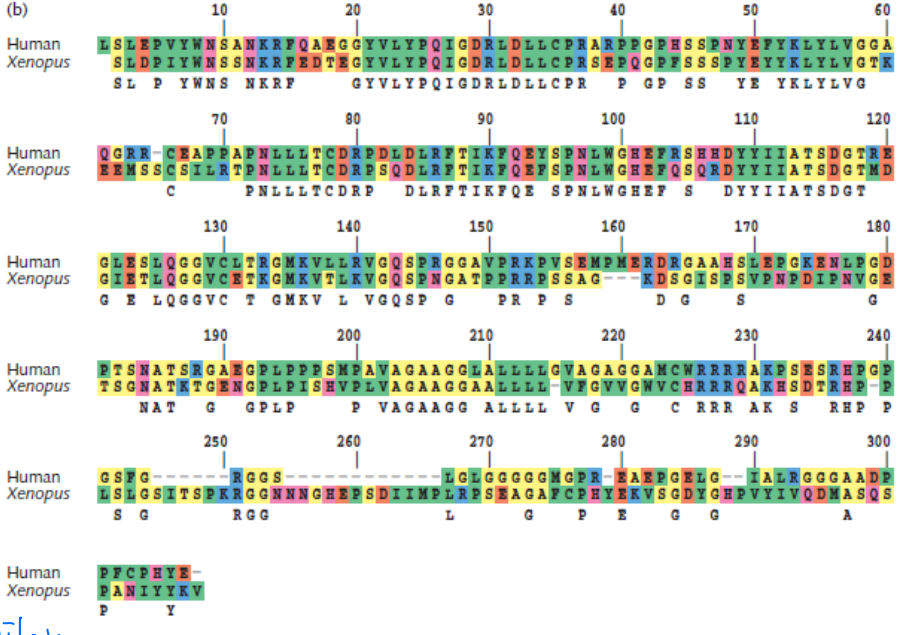
경치는 원점을 떠나 지나가도록 하되 maximum score 얻은수 있음

Sequence Alignment : Dot plot

2차원 map
& match → mark as dot

Relationships between the sequences of ephrin B3 proteins from human and *Xenopus laevis*.

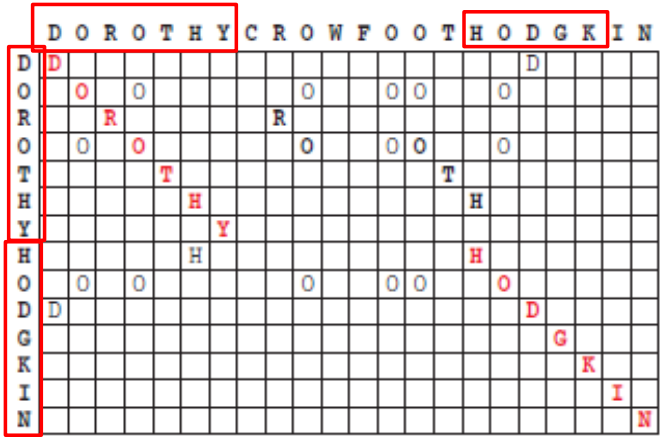




⇒ easy to visualize similarity

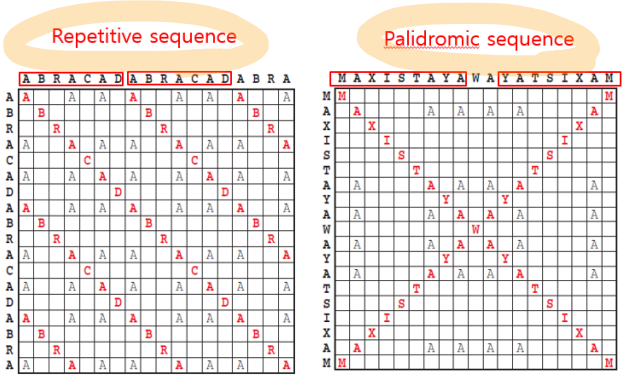
Dynamic Programming

- Dot Plot



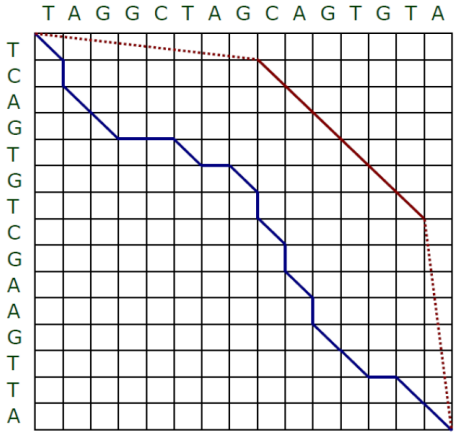
DOROTHY--------HODGKIN

DOROTHYCROWFOOTHODGKIN



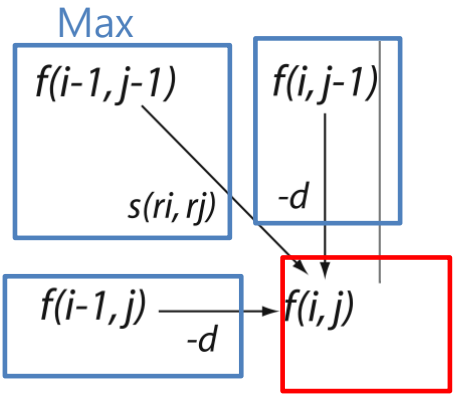
trace path to get high score

-Dynamic Programming:



Divide the problem into subproblems. 재귀적으로 다루는 방법

Previous state > Max Score > **current state (Max Score)**



➢ **Global Alignment** — **Needleman-Wunsch**

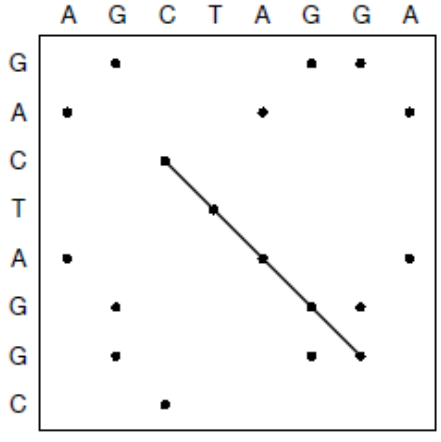
```
T C A G - - T - G T C G A A G T - T A
|   | |   |   |   |   |   | |   |
T - A G G C T A G - C - A - G T G T A
```

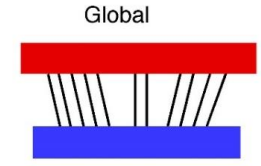
➢ **Local Alignment**

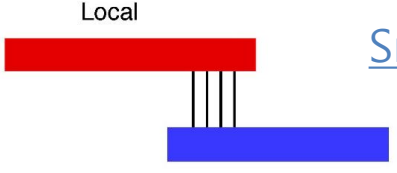
```
T C A G T G T C G A A G T T A
|   |   |   |   |   |
T A G G C T A G C A G T G T A
```

Smith-Waterman

Sequence alignment and database search







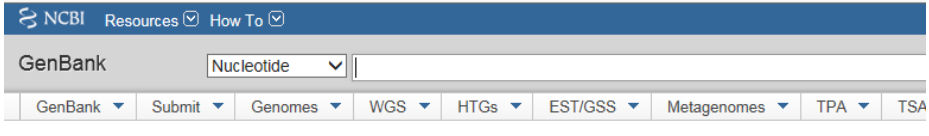
Needleman-Wunsch algorithm(1970)

```
Global FTFTALILLAVAV
       F--TAL-LLA-AV
```

Smith-Waterman algorithm (1981)

```
Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

GenBank (1982-) → first database collecting sequence data



Temple Smith

Michael Waterman

Walter Goad

David Lipman

NCBI
National Center for
Biotechnology Information

BLAST (1981)

Database search

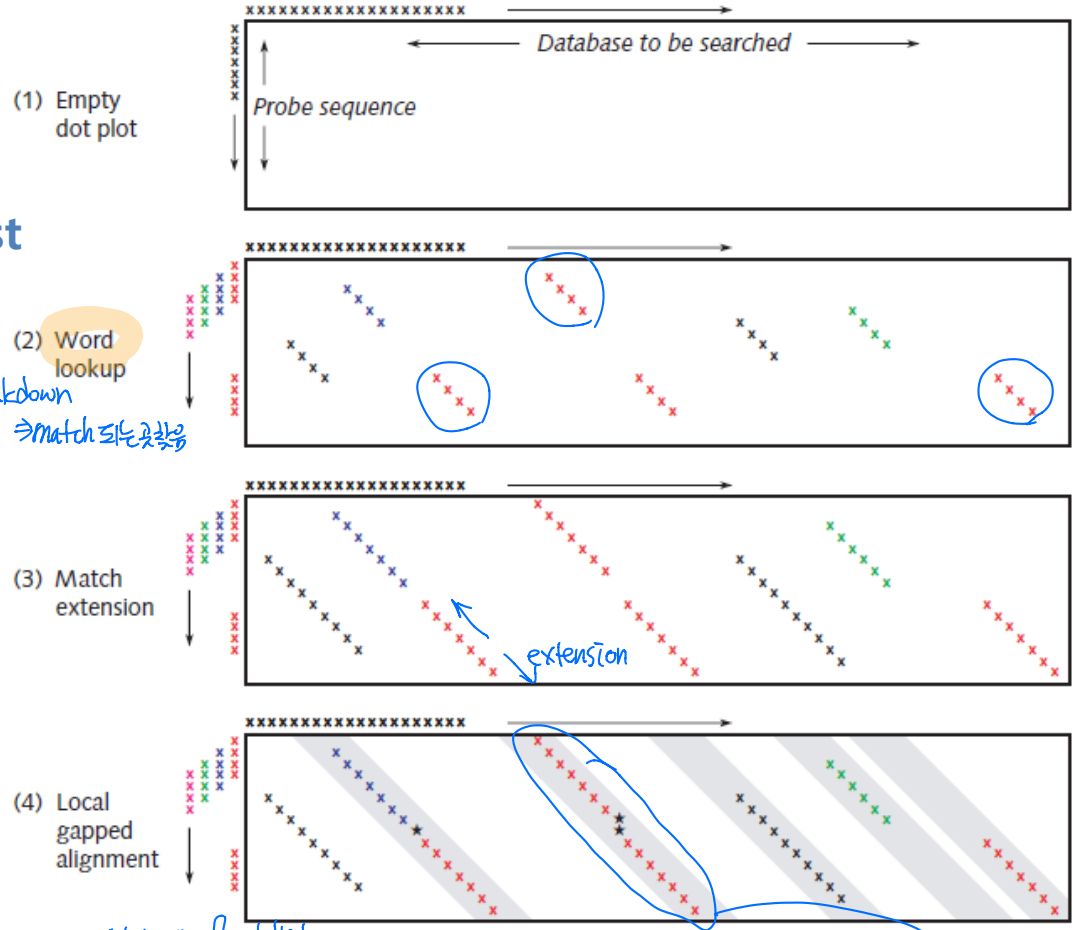
Alignment of sequence to sequences in GenBank database

Sequence Database Search: BLAST

like finding text from textbook with index

BLAST: **B**asic **L**ocal **A**lignment **S**earch **T**ool,

book : NCBI genebank



1. List

내 sequence를 3~4 nt로 breakdown
⇒ match되는 곳찾음

2. Scan

3. Extend

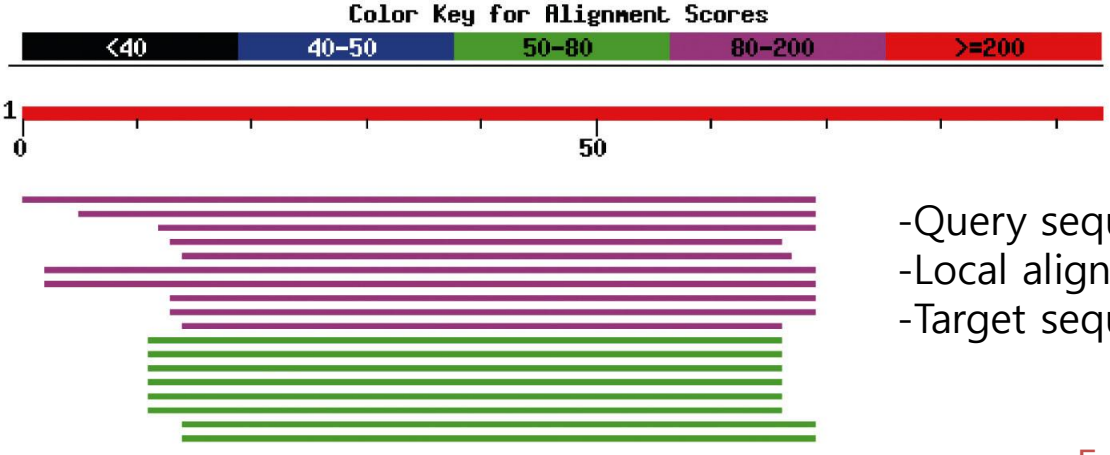
if mismatch가 계속 나와서 score drops
⇒ stop extension

4. Statistical significance

이정도 score가 되면 significant한지
∴ random에 의한건 아닌지

내 sequence랑 match 될 확률이 높은 sequence

BLAST

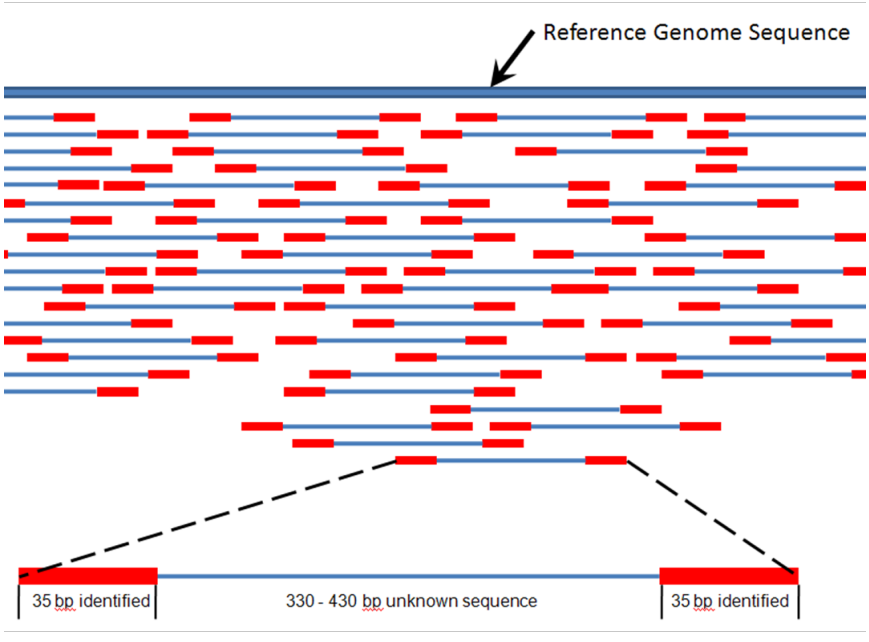


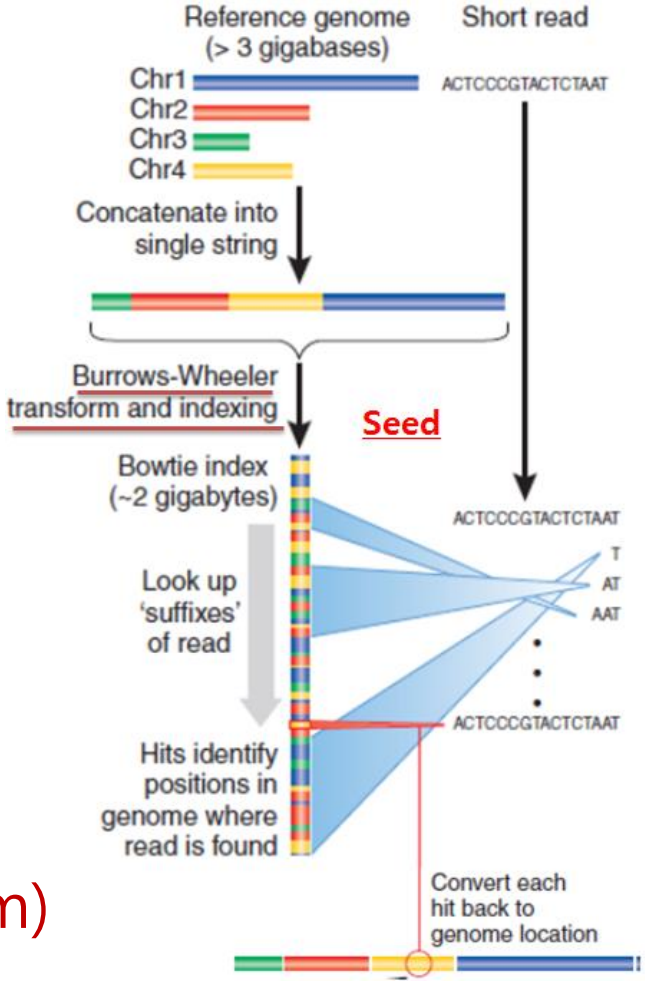
-Query sequence
-Local alignment
-Target sequence

E-Value

Sequences producing significant alignments:		Score (bits)	E Value
gi|256517|gb|S45649.1|S45649	16S rRNA [Mastotermes electrod...	139	5e-33
gi|12005612|gb|AF246514.1|AF246514	Drosophila ornatipennis ...	86	7e-17
gi|11119031|gb|AF304735.1|AF304735	Sphyracephala bipunctipe...	84	3e-16
gi|3552018|gb|AF086859.1|AF086859	Mystacinobia zealandica 1...	84	3e-16
gi|256518|gb|S45650.1|S45650	16S rRNA [Mastotermes darwinie...	84	3e-16
gi|15341487|gb|AF403473.1|AF403473	This canus 16S ribosomal...	82	1e-15
gi|15341483|gb|AF403469.1|AF403469	Icaridion debile 16S rib...	82	1e-15
gi|13435200|ref|NC_002697.1|	Chrysomya chloropyga mitochond...	82	1e-15
gi|13384216|gb|AF352790.1|AF352790	Chrysomya chloropyga mit...	82	1e-15
gi|3552016|gb|AF086857.1|AF086857	Calliphora quadrimaculata...	82	1e-15
gi|15341485|gb|AF403471.1|AF403471	Malacomyia sciomyzina 16...	80	4e-15
gi|15341463|gb|AF403449.1|AF403449	Helcomyza mirabilis 16S ...	80	4e-15

Mapping millions of short reads (NGS) on genome





reduce size of file

BWT (Burrows-Wheeler transform) algorithm