Modern aka “Next-generation” sequencing

Sequence alignment - part 1

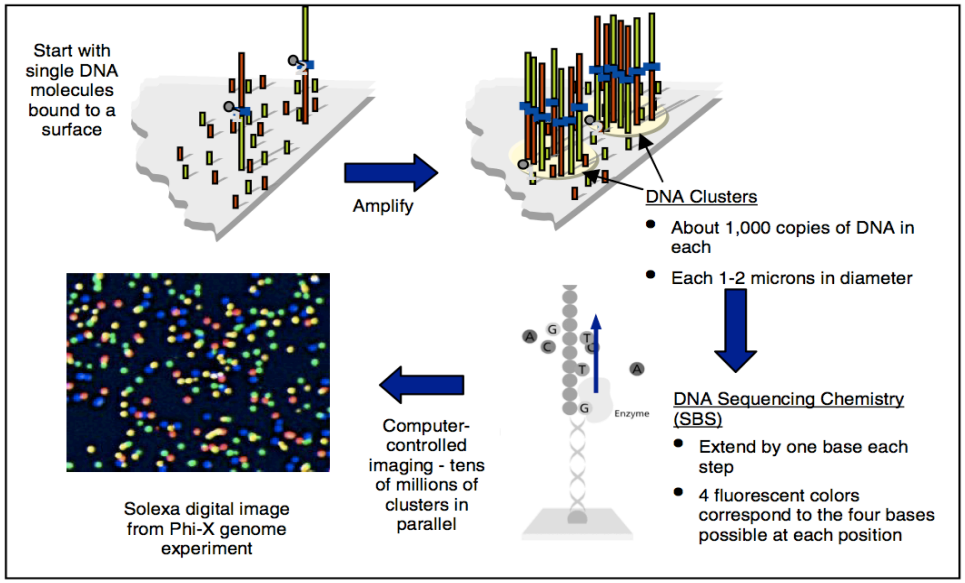
1

Performance of available ‘next generation’ sequencing platforms

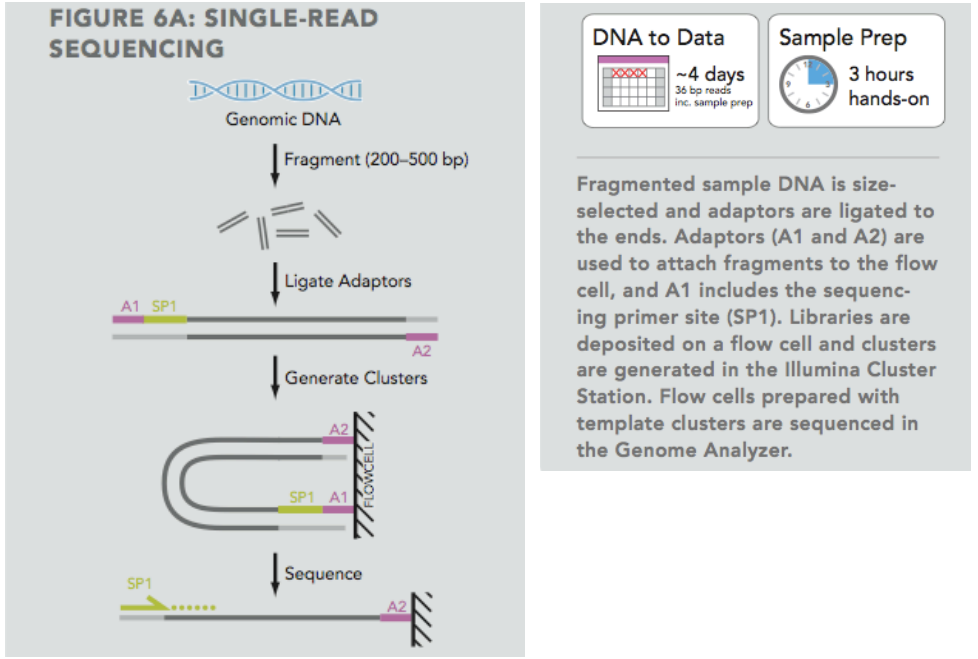
Platform	Read Length (bases)	Reads/ run	GBases/ single end run	Paired end?	Mate pair?
Sanger (for comparison)	400-900	NA	NA	NA	NA
454	450-700	1m	0.5-0.7	yes	yes
SOLiD	30-50	1,200-1,400m	70	short	yes
Solexa / Illumina	100-150	3,000m	300	yes	yes
Ion Torrent	100-400	80m	30	no	no
Pacific Biosystems	5,000-20,000	50,000	0.4	NA	NA

2

Illumina (Solexa) sequencing

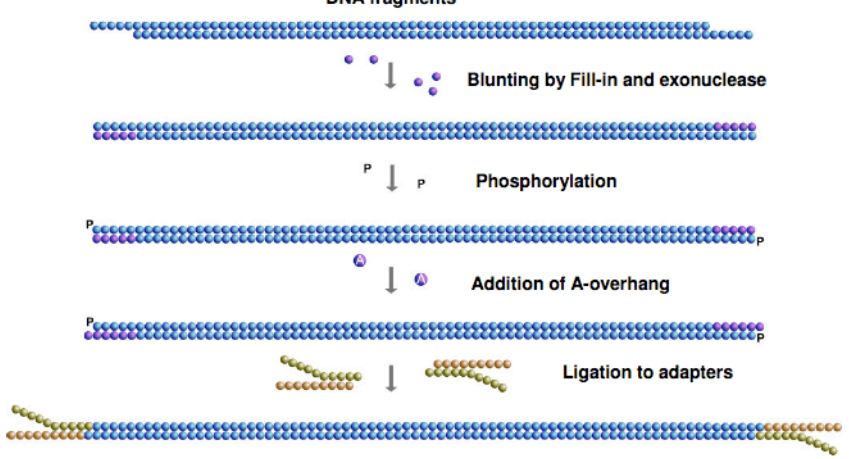


3



4

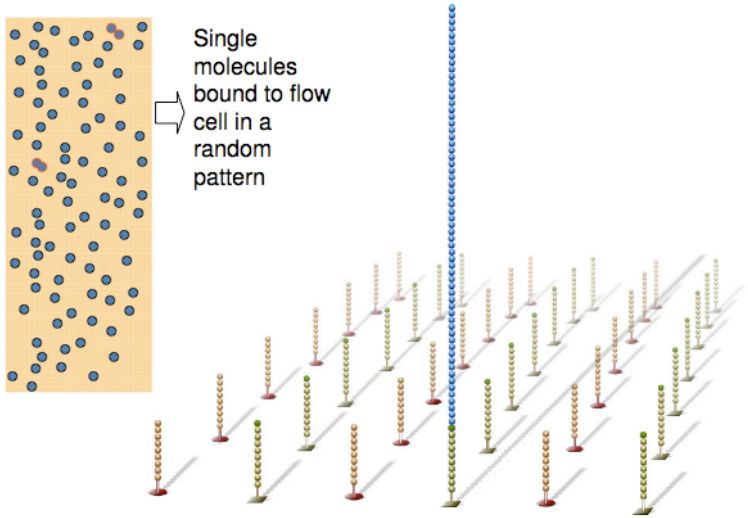
Genomic DNA Library Prep



5

Cluster Generation

Covalently-Bound Spatially Separated Single Molecules

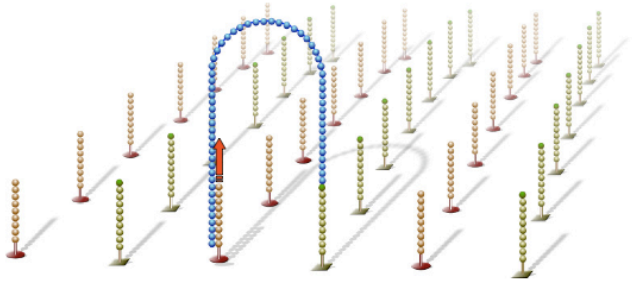


6

Cluster Generation

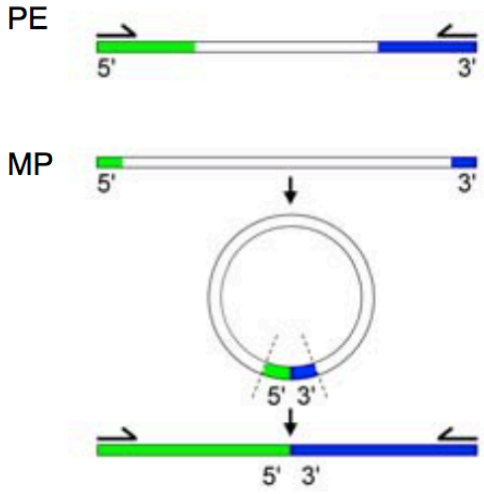
Bridge Amplification

- Single-strand flips over to hybridize to adjacent primers to form a bridge
- Hybridized primer is extended by polymerases



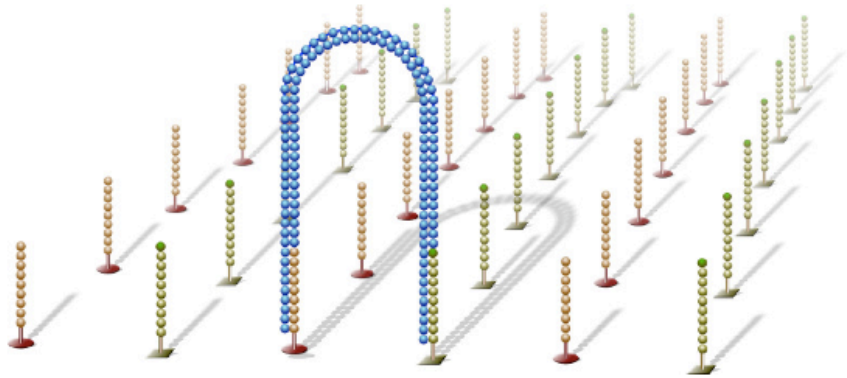
7

Paired Ends and Mate-pairs



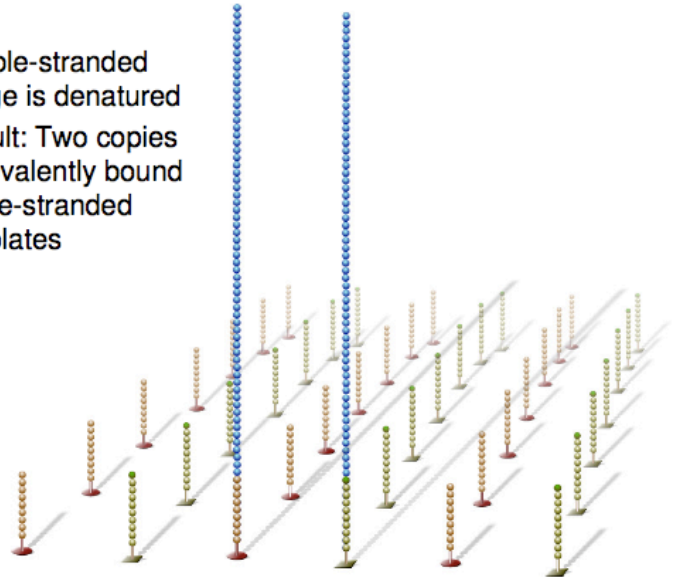
8

- Double-stranded bridge is formed



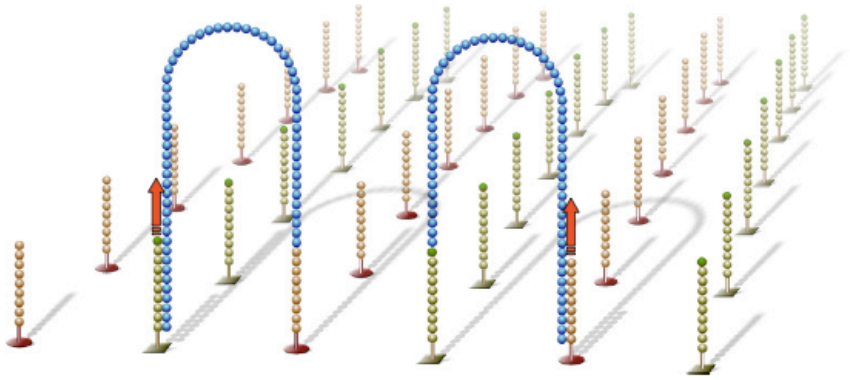
9

- Double-stranded bridge is denatured
- Result: Two copies of covalently bound single-stranded templates



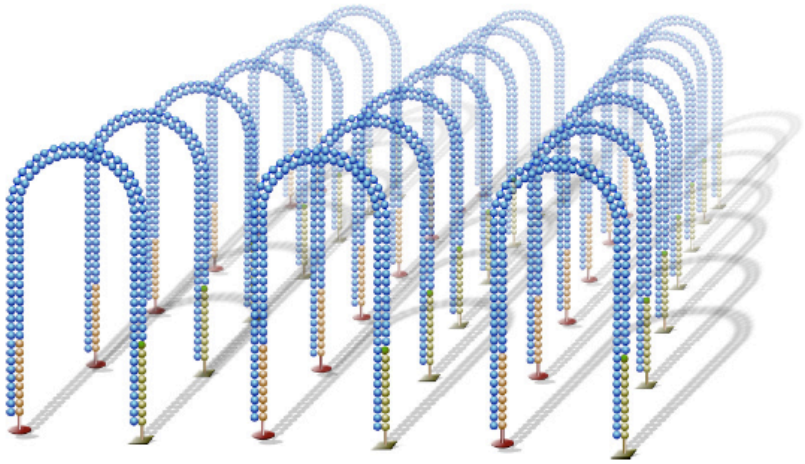
10

- Single-strands flip over to hybridize to adjacent primers to form bridges
- Hybridized primer is extended by polymerase



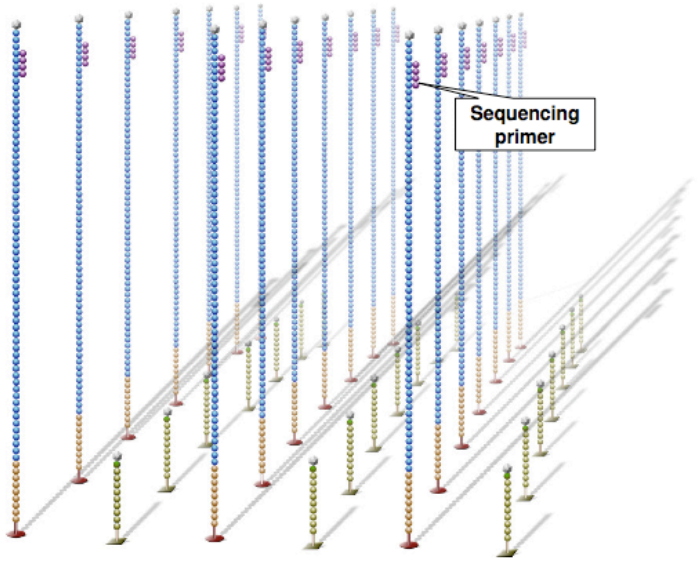
11

- Bridge amplification cycle repeated until multiple bridges are formed

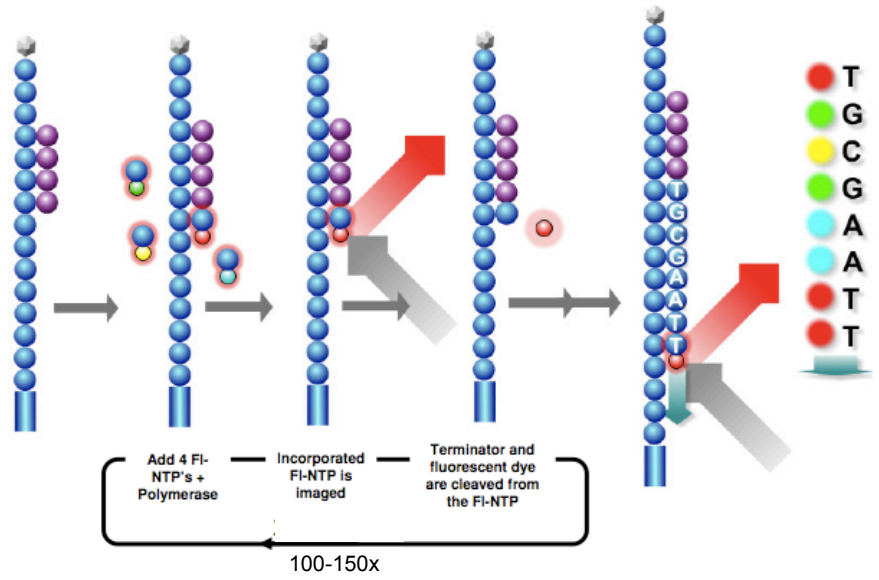


12

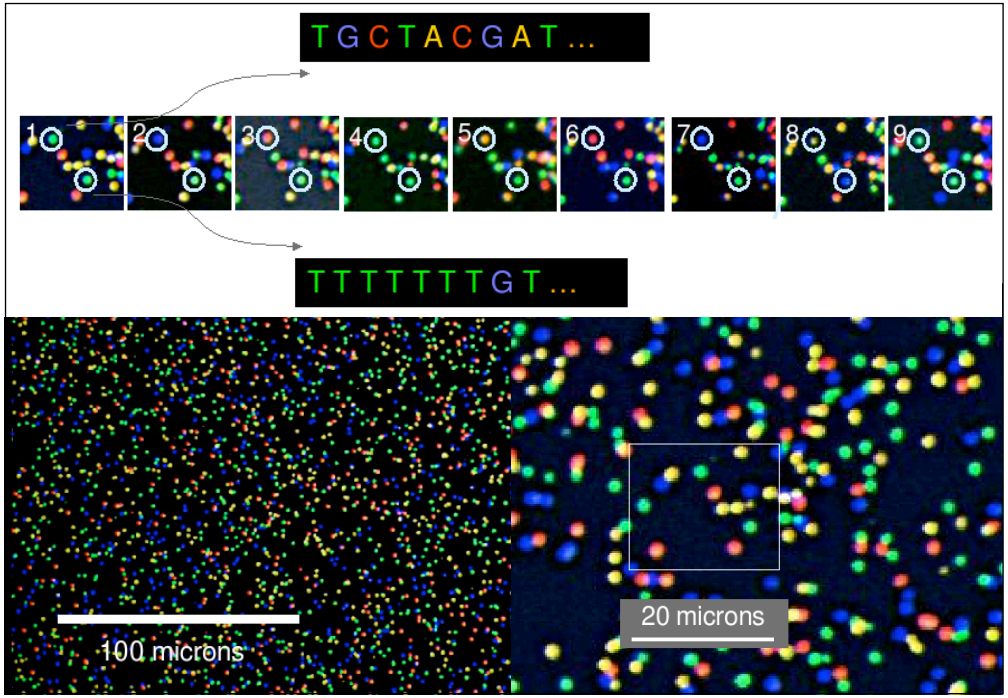
Sequencing



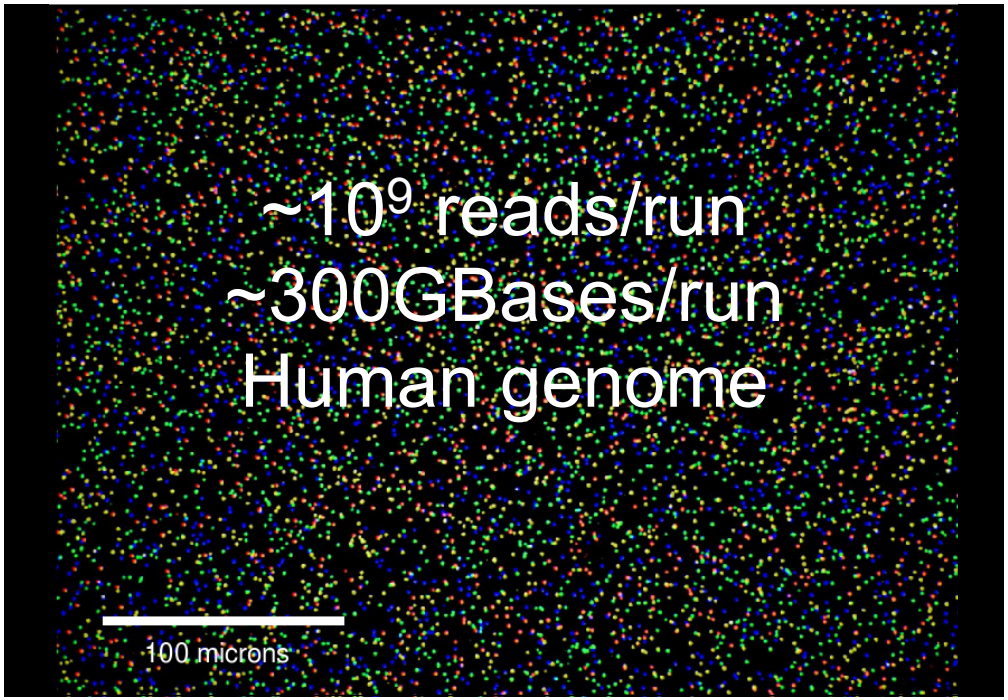
13



14

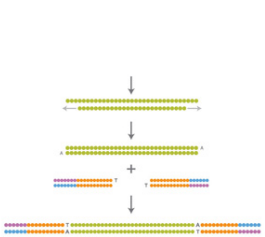

TGCTACGAT...
1
2
3
4
5
6
7
8
9
TTTTTTTGT...
100 microns
20 microns

15


~10^9 reads/run
~300GBases/run
Human genome
100 microns

16

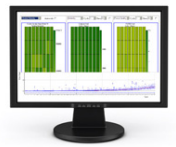
$3x10^9$ reads/ run yielding 300+Gbases of data (~8 days)



Library Preparation
<6 h (<3 h hands-on)

Cluster Generation
<4 h (<10 min hands-on)

Sequencing by Synthesis
1.5-8 days (<10 min hands-on)



RTA v1.7, CASAVA v1.7
2 days (30 min hands-on)

Single Molecule Real Time sequencing

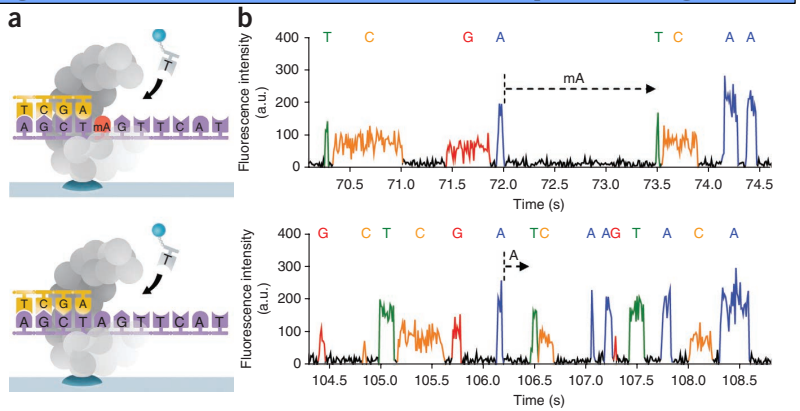


Figure 1 | Principle and corresponding example of detecting DNA methylation during SMRT sequencing. (**a**) Schematics of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenine. (**b**) Typical SMRT sequencing fluorescence traces for samples in **a**. Letters above the fluorescence trace pulses indicate the identity of the nucleotide incorporated into the growing complementary strand. Dashed arrows indicate the IPD before incorporation of the cognate thymine. For this typical example, the IPD is about five times larger for mA in the template compared to adenine.

Flusberg et al *Nature Methods* 7, 461 - 465 (2010)

Direct detection of DNA methylation during single-molecule, real-time sequencing

Why align sequences?

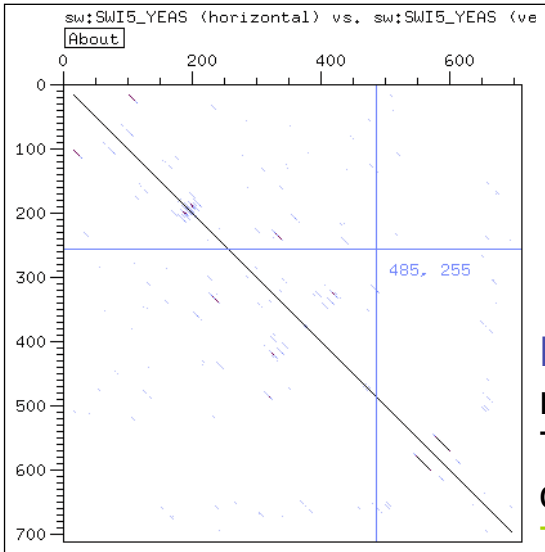
Find overlapping sequences to allow sequence assembly

Find where PCR primers might be mis-priming in a genome

Identify families of related protein sequences

Infer function of protein sequences, assuming related sequences have related functions.

Gene annotation:
- matching mRNA to a genome sequence
- matching protein sequences to a genome

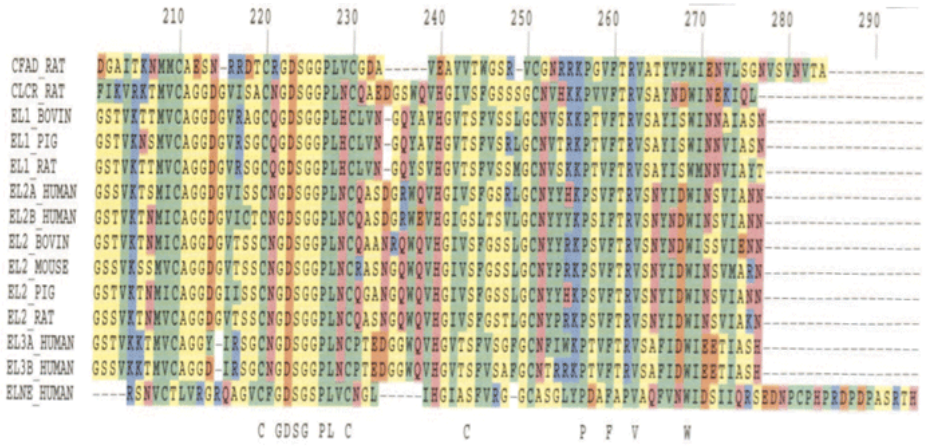


Dotter: Compares each residue in one sequence To every residue in the other.

Time, memory

(length M) x (length N)

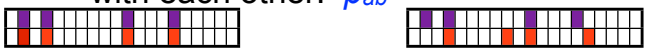
O (M x N)



Scoring using log likelihoods

From a large set of high quality *ungapped* protein sequence alignments, for pairs of aligned sequences:

- measure background frequencies of residues *a* and *b*: q_a, q_b.
- measure frequency with which *a* and *b* are found aligned with each other: p_{ab}



Log likelihood: score(*a*,*b*) = log ($p_{ab} / q_a q_b$)

score 0 if aligned as often as expected
score positive if preferentially aligned
score negative if alignment is avoided

Rounded to nearest integer for computational efficiency

Where did the BLOSUM62 alignment score matrix come from?
Eddy SR.
Nat Biotechnol. 2004 Aug;22(8):1035-6. Review. PMID: 15286655

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**

Figure 2.2 *The* BLOSUM50 *substitution matrix. The log-odds values have been scaled and rounded to the nearest integer for purposes of computational efficiency. Entries on the main diagonal for identical residue pairs are highlighted in bold.*

Human and yeast cdc2 genes

```
cdc2.aln

CLUSTAL    1 W
HSCDC2     1          MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELR-
SPCDC2     1          MENYQKVEKIGEGTYGVVYKARHKLSGRIVAMKKIRLEDESEGVPSTAIREISLLKEVND


CLUSTAL   69
HSCDC2    60          ---HPNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPG--QYMDSSLVKSYLYQILQ
SPCDC2    61          ENNRSNCVRLLDILHAESKLYLVFEFLDMDLKKYMDRISETGATSLDPRLVQKFTYQLVN
          17

CLUSTAL  137
HSCDC2   115          GIVFCHSRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEV
SPCDC2   121          GVNFCHSRRIIHRDLKPQNLLIDKEGNLKLADFGLARSFGVPLRNYTHEIVTLWYRAPEV
          81

CLUSTAL  205
HSCDC2   175          LLGSARYSTPVDIWSIGTIFAELATKKPLFHGDSEIDQLFRIFRALGTPNNEVWPEVESL
SPCDC2   181          LLGSRHYSTGVDIWSVGCIFAEMIRRSPLFPGDSEIDEIFKIFQVLGTPNEEVWPGVTLL
         149

CLUSTAL  273
HSCDC2   235          QDYKNTFPKWKPGSLASHVKNLDENGLDLLSKMLIYDPAKRISGKMALNHPYFNDLDNQI
SPCDC2   241          QDYKSTFPRWKRMDLHKVVPNGEEDAIELLSAMLVYDPAHRISAKRALQQNYLRDFH---
         217

CLUSTAL  341              351
HSCDC2   295          KKM 297
         285              295
```

Gap penalties

Linear:

total penalty = -d * g

where *g* is length of gap and d is the per-residue gap penalty

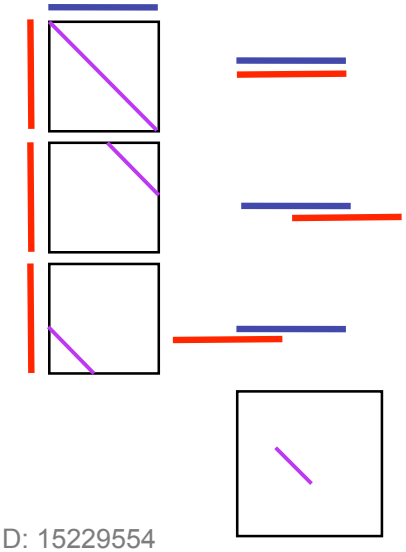
Affine:

total penalty = -d - e(g-1)

where *e* is the gap extension penalty and *e* < *d*

Dynamic Programming

Given a scoring scheme for aligning residues and gaps, DP algorithm guarantees the best (sub)sequence alignment



What is dynamic programming? Eddy SR.
Nat Biotechnol. 2004 Jul;22(7):909-10. PMID: 15229554