Alignment

Lecture 7
Sept 14, 2016

Announcements

- Codes??
- No class next Wednesday
- Reading presentations

Announcements

	Student 1	Student 2	Student 3
Week 4			
Week 5			
Week 6			
Week 7			
Week 8			
Week 9			
Week 10			
Week 11			
Week 12			
Week 13			
Week 14			
Week 15			

PAM1 MATRIX

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

PAM versus Divergence

PAM0 PAM1 PAM250 PAM∞

Less divergent More divergent

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp

PAM250

	A	R	N	D	C	Q	E	G	H	I
A	13	6	9	9	5	8	9	12	6	8
R	3	17	4	3	2	5	3	2	6	3
N	4	4	6	7	2	5	6	4	6	3
D	5	4	8	11	1	7	10	5	6	3
C	2	1	1	1	52	1	1	2	2	1
Q	3	5	5	6	1	10	7	3	7	3
E	5	4	7	11	1	9	12	5	6	3
G	12	5	10	10	4	7	9	27	5	3
H	2	5	5	4	2	7	4	2	15	2
I	3	2	2	2	2	2	2	2	2	10
L	6	4	4	3	2	6	4	3	5	15
K	6	18	10	8	2	10	8	5	8	3
M	1	1	1	1	0	1	1	1	1	5
F	2	1	2	1	1	1	1	1	3	6
P	7	5	5	4	3	5	4	5	5	2
S	9	6	8	7	7	6	7	9	6	4
T	8	5	6	6	4	5	5	6	4	5
W	0	2	0	0	0	0	0	0	1	0
Y	1	1	2	1	3	1	1	1	3	1
V	7	4	4	4	4	4	4	5	4	29

FIGURE 3.13. The PAM250 mutation probability matrix. From Dayhoff (1978, p. 350, fig. 83). At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to Fig. 3.11, and the columns sum to 100. Used with permission.

From mutational probability to scoring matrices

$$s_{i,j} = 10 * \log_{10}\left(\frac{q_{i,j}}{p_i}\right)$$

	A	R	N	D	C	Q
A	13	6	9	9	5	8
R	3	17	4	3	2	5
N	4	4	6	7	2	5
D	5	4	8	11	1	7

TABLE 3-2 Normalized Frequencies of Amino Acid

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

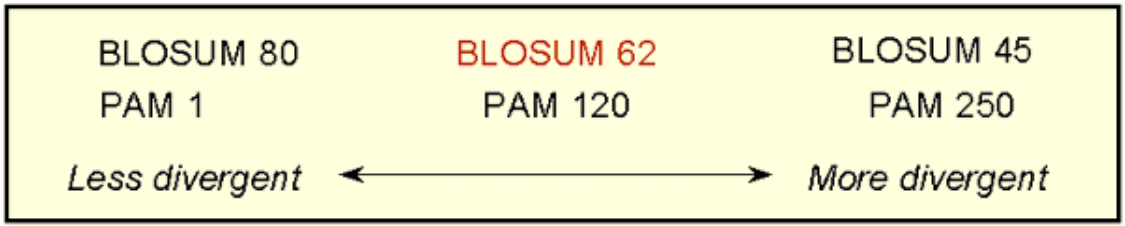
These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence
Source: From Dayhoff (1978). Used with permission.

Alignment – think BLAST

Q	ANCQE	versus	ANCQE
D	ANCGE		ANCHE

Blosum Matrix

BLOSUM MATRIX



Advanced Search

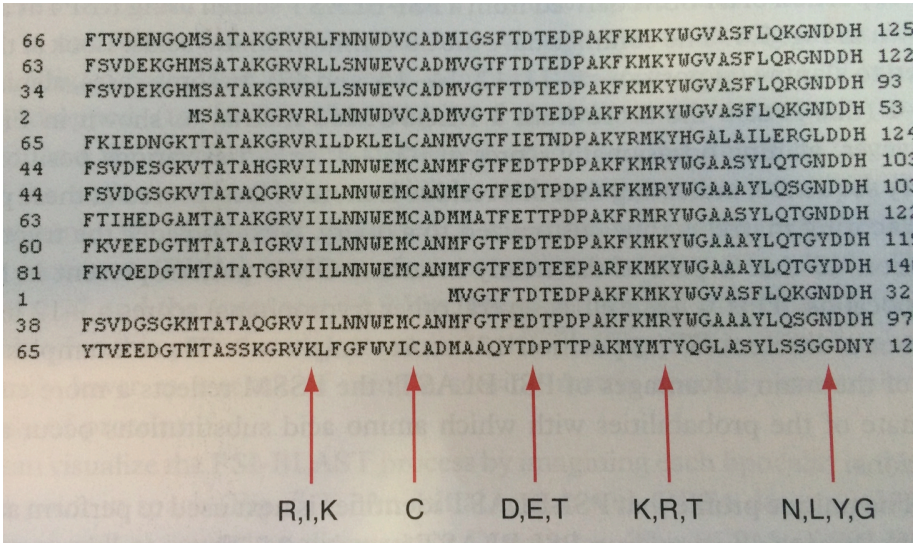
Advanced Search

PSI-BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi?&PROGRAM=blastp&QUERY=AAG22855.1

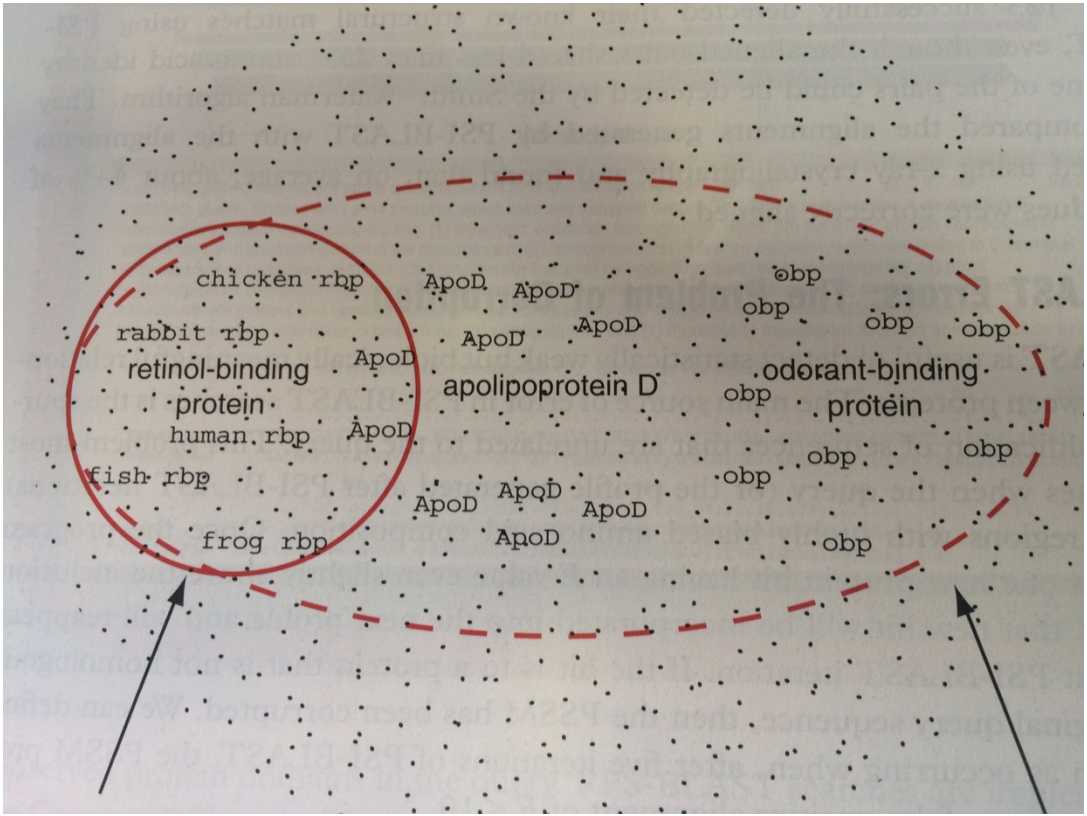
Advanced Search

PSI-BLAST



Advanced Search

PSI-BLAST



Advanced Search

PHI-BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi?&PROGRAM=blastp&QUERY=AAG22855.1

AAG22855.1

Hidden Markov Model

Hidden Markov Model

(a)

1D8U	HAMSV
1OJ6A	HIRKV
2hhbB	HGKKV
1FSL	HAEKL
2MM1	HGATV

(b)

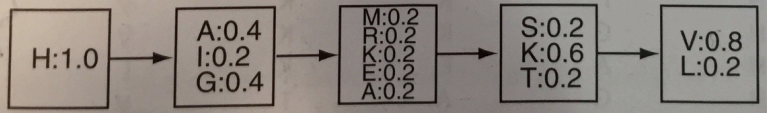
Probability	position 1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
p(E)			0.2		
p(A)			0.2		
p(S)				0.2	
p(K)				0.6	
p(T)				0.2	
p(V)					0.8
p(L)					0.2

(c)

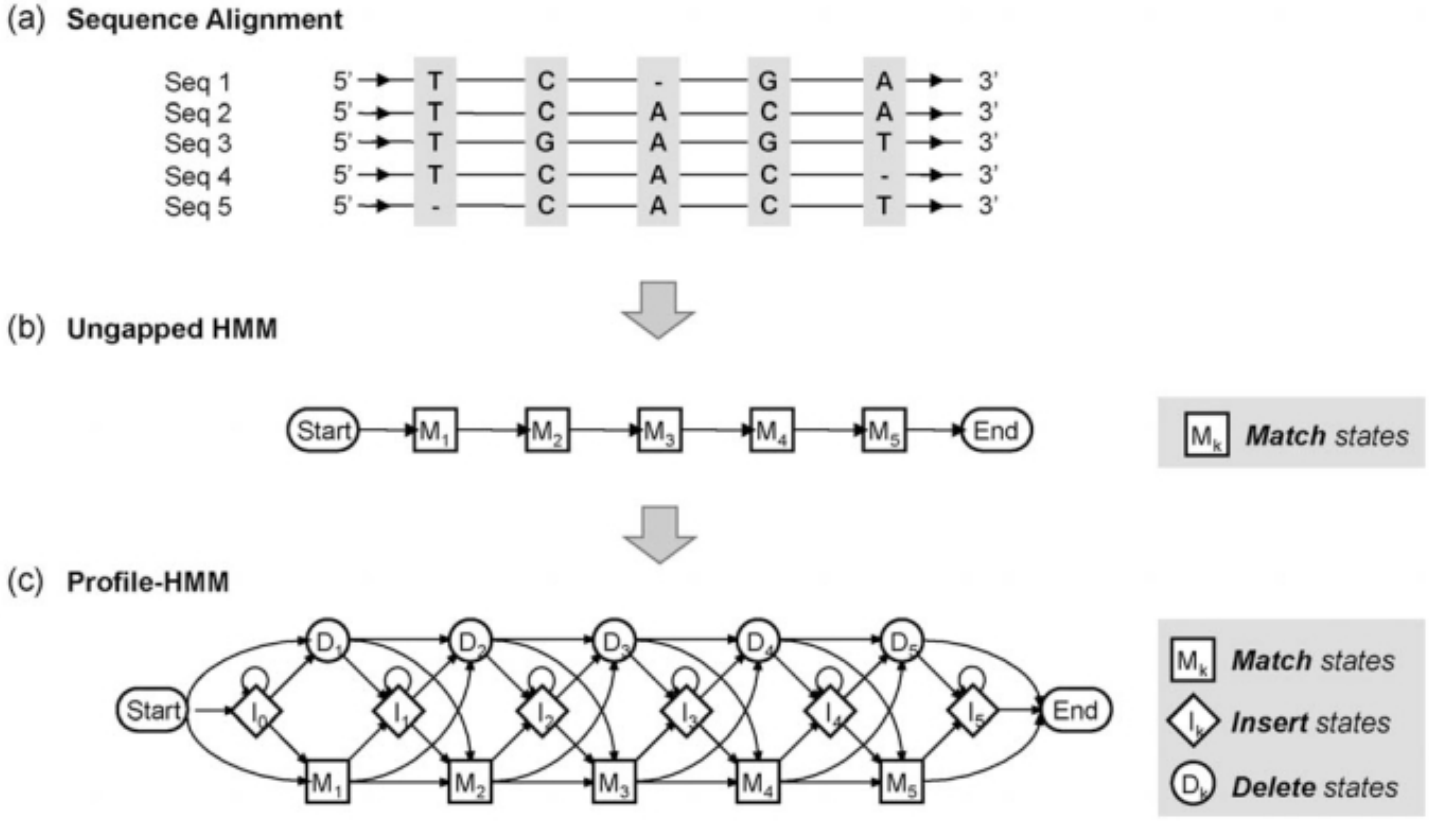
$p(\text{HARTV}) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128$

$\text{Log odds score} = \ln(1.0) + \ln(0.4) + \ln(0.2) + \ln(0.2) + \ln(0.8) = -4.4$

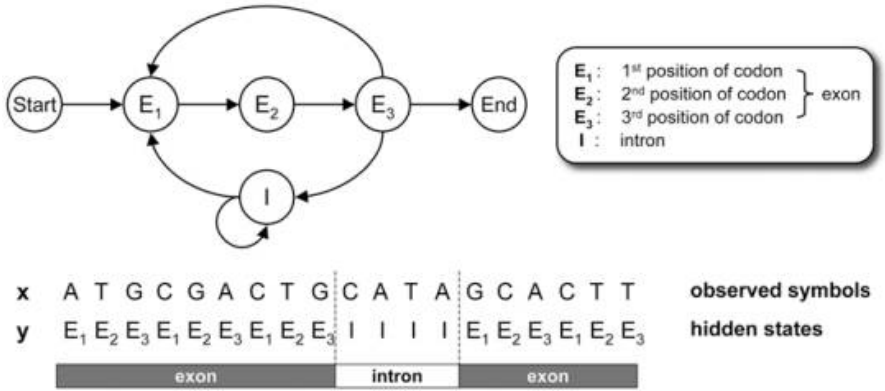
(d)



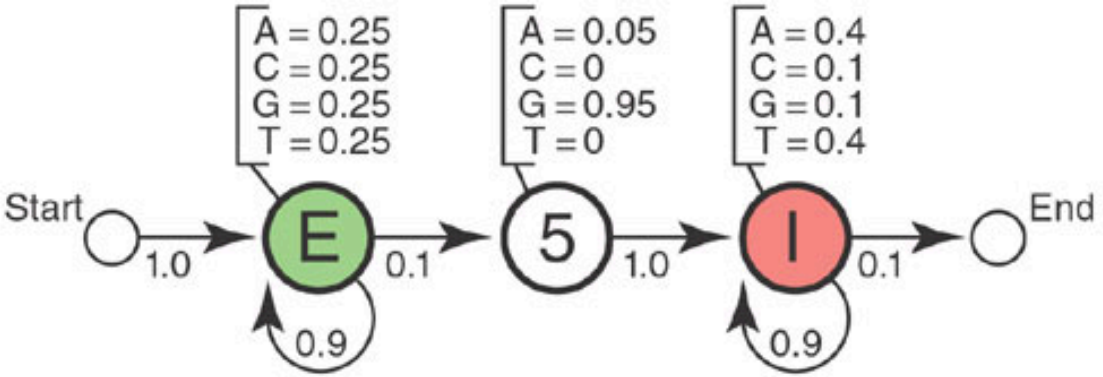
HIDDEN MARKOV MODEL



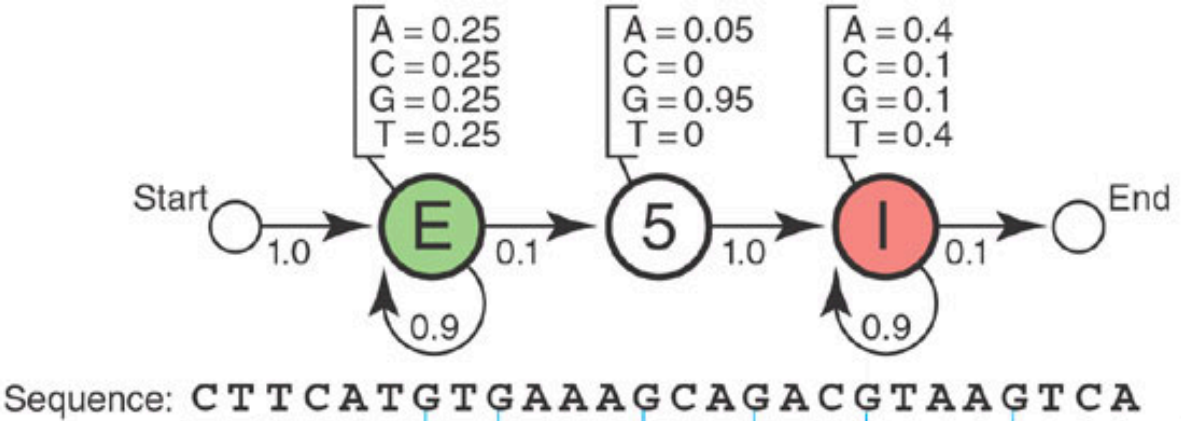
HIDDEN MARKOV MODEL



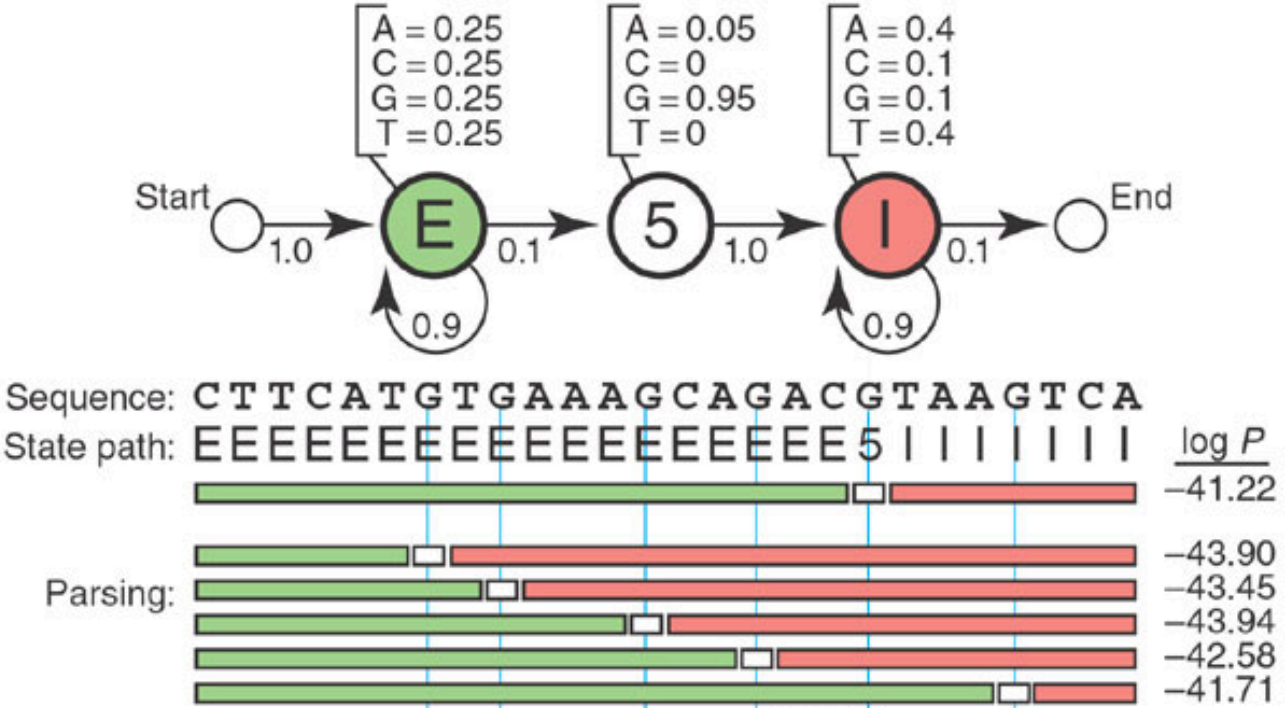
Hidden Markov Model



Hidden Markov Model



Hidden Markov Model



fastQ format and Illumina sequence data

fastQ

Description

Uses

fastQ

@HSQ-7001360:67:H88RHADXX:1:1101:1448:2158 1:N:0:CAGATC
ATCTATCTGAGACTGATACGCCTTCGGCTTAATTTATACAAG
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIF

fastQ

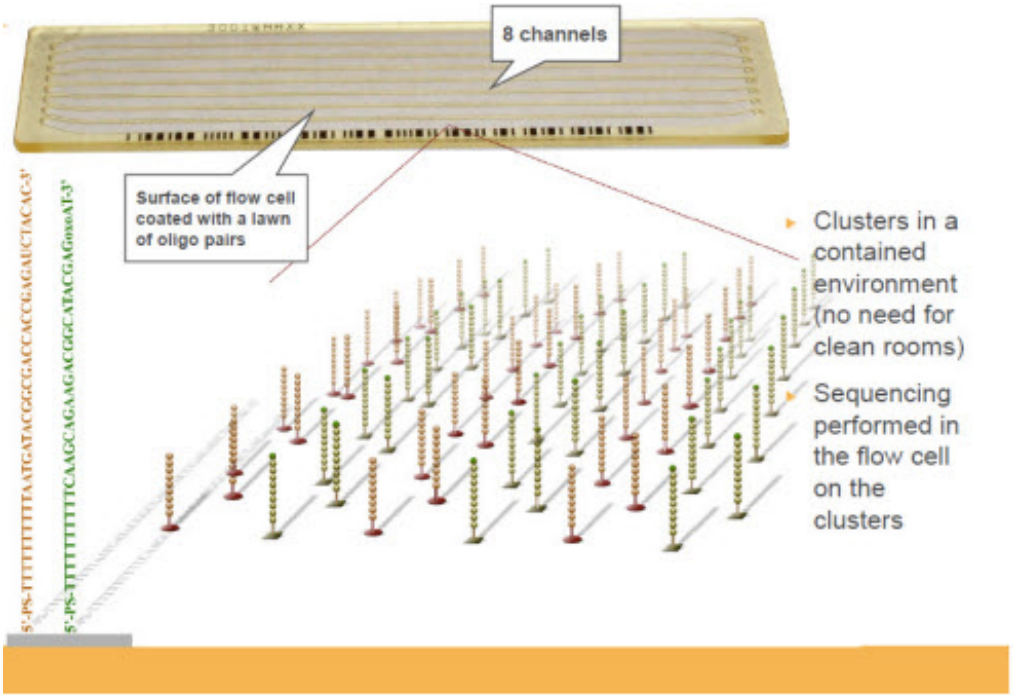
@HSQ-7001360:67:H88RHADXX:1:1101:1448:2158 1:N:0:CAGATC

- HSQ-7001360= Instrument name
- 67= run ID
- H88RHADXX=Flowcell ID
- 1=lane 1
- 1101=tile number
- 1448= x coordinate
- 2158= y coordinate
- 1=left read
- N=not filtered
- 0=control bit -> (not used anymore)
- CAGATC= adapter sequence

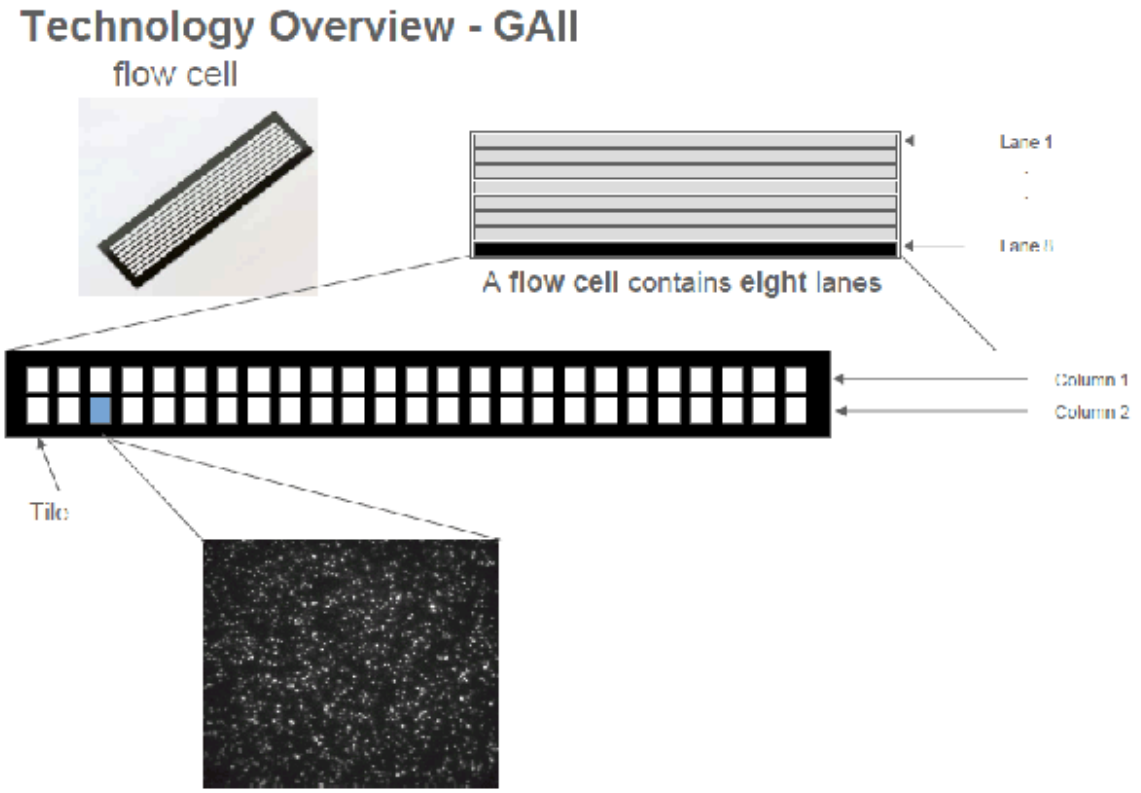
fastQ -> Illumina Seq

https://youtu.be/womKfikWlxM

Illumina Seq



Illumina Seq



Illumina Seq

@HSQ-7001360:67:H88RHADXX:1:1101:1448:2158 1:N:0:CAGATC

- 1101=tile number
- 1448= x coordinate
- 2158= y coordinate

Illumina Seq

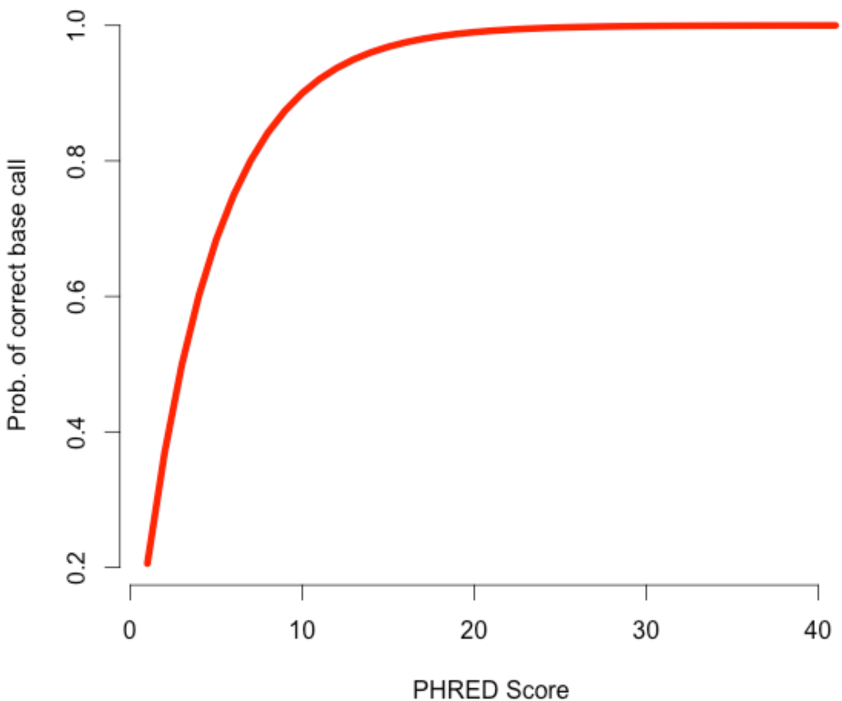
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFF

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL...................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

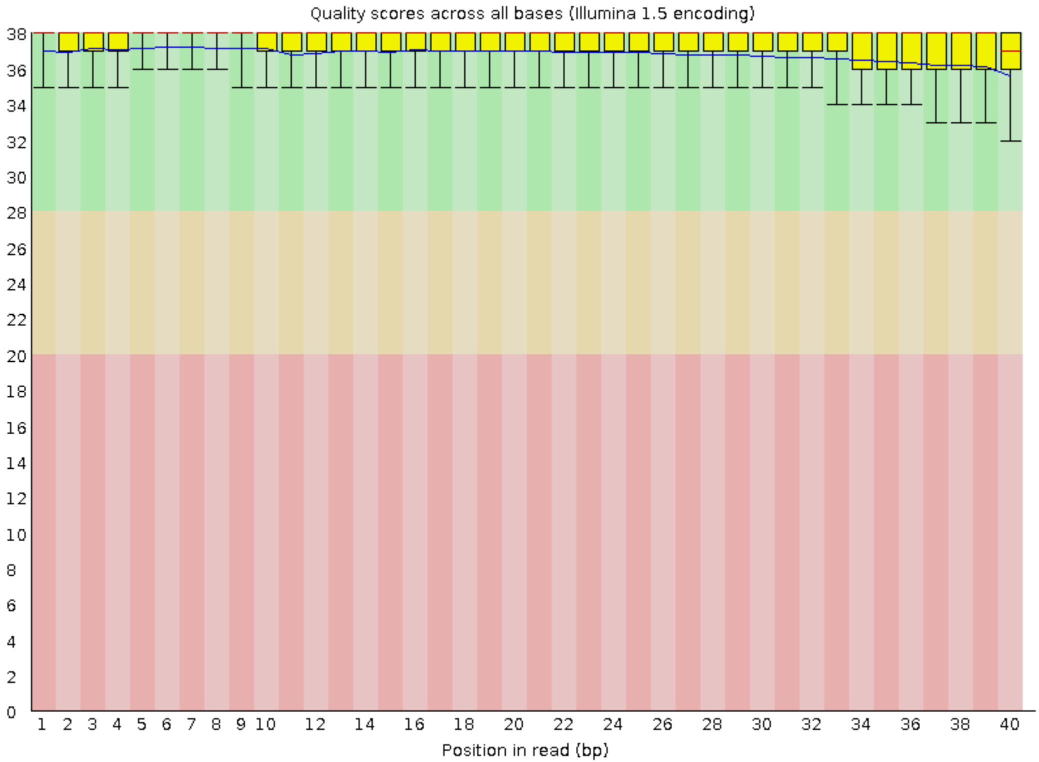
Illumina Seq

BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFF



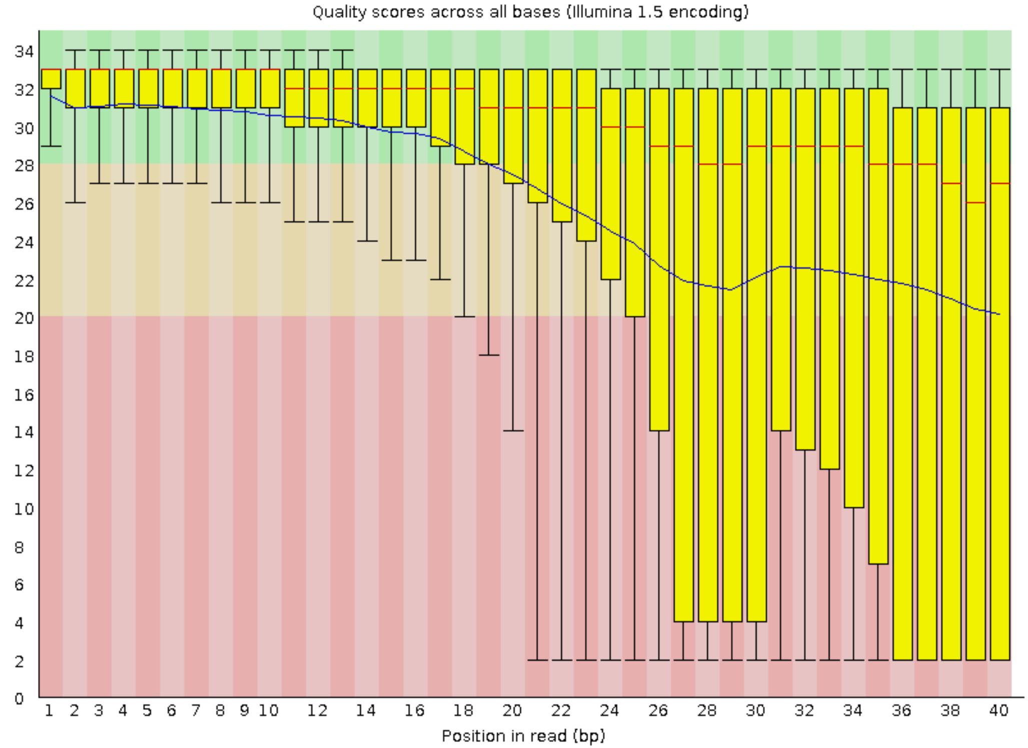
Illumina Seq

BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFF



Illumina Seq

BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFF



Illumina Seq

Figure 4. Paired-End Sequencing and Alignment