# Sequence Day

Martin Asser Hansen
2015-02-12

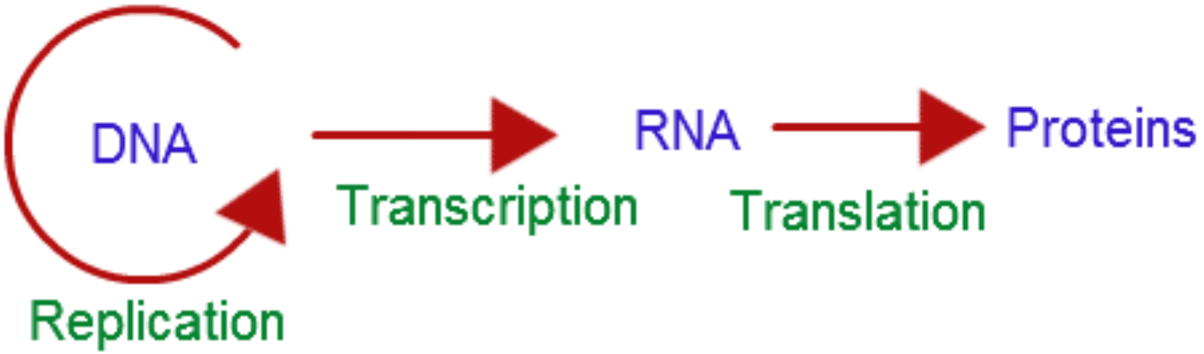
# Overview

- Projects
- Sequences
  - DNA
  - RNA
  - Protein
- PCR
- Sequencing
  - Shotgun
  - Amplicon
- Data

# Projects

- scan_for_matches
- KMC REGEX
- Clustering

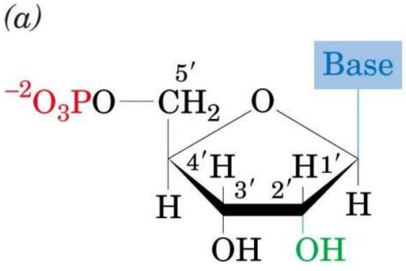
# The Central Dogma



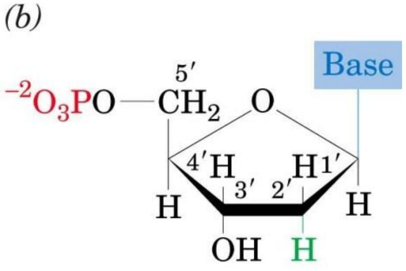
ATGATGCATCGT AUGAUGCAUCGU MASWPQCPYKEDEN

`strcmp()`

# RNA vs DNA



**Ribonucleotides**



**Deoxyribonucleotides**

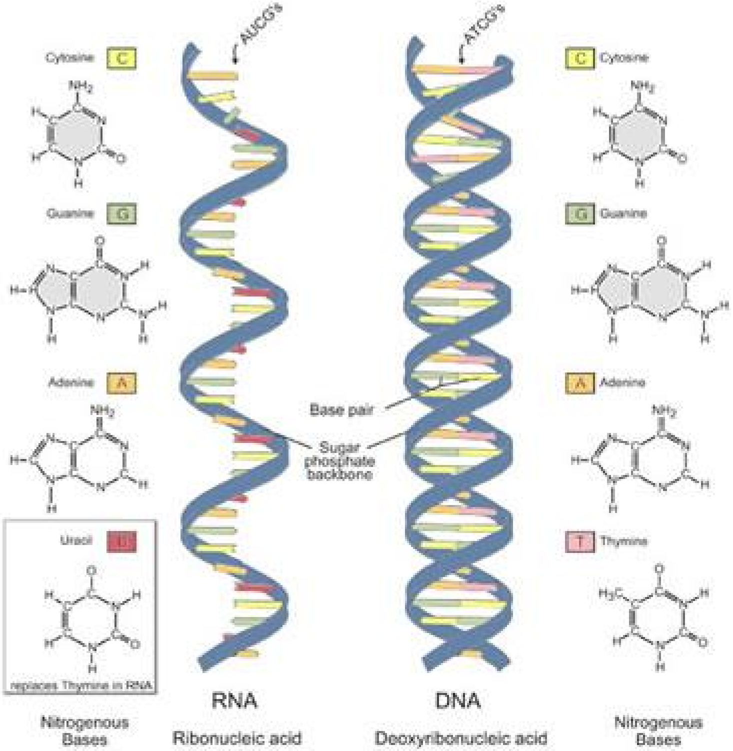
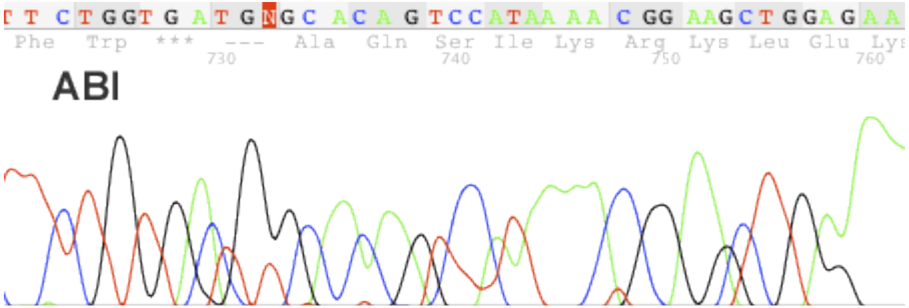
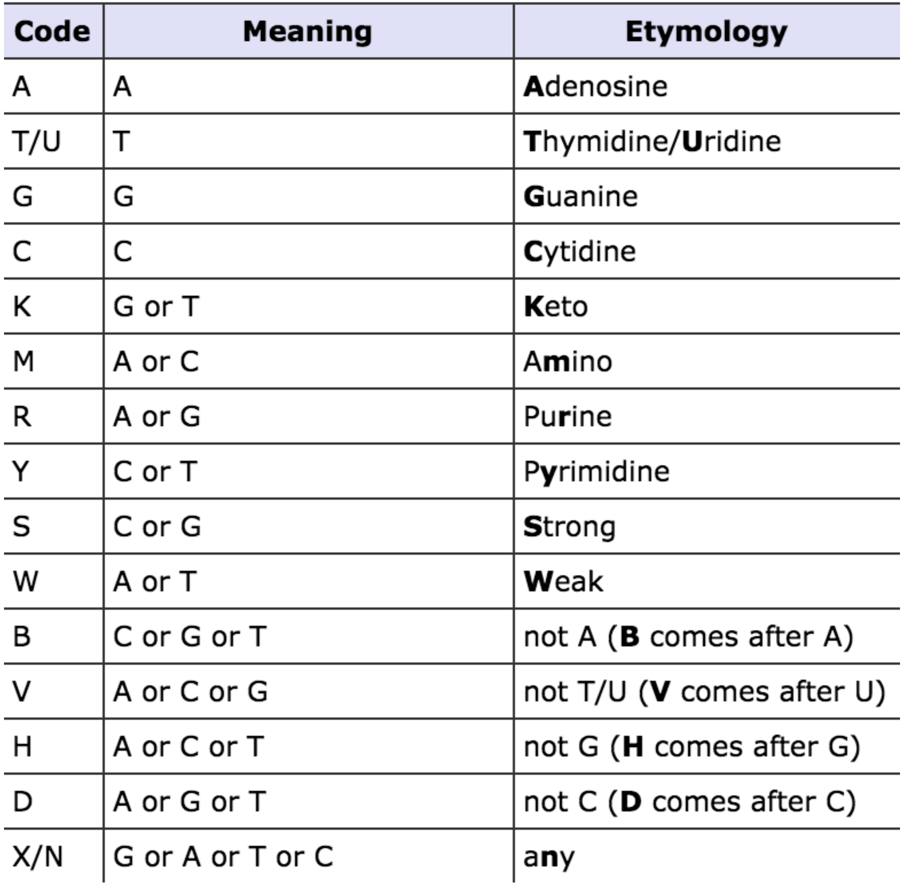


Image adapted from: National Human Genome Research Institute. Talking Glossary of Genetic Terms. Available at: www.genome.gov/Pages/Hyperion//DIR/VIP/Glossary/Illustration/rna.shtml.

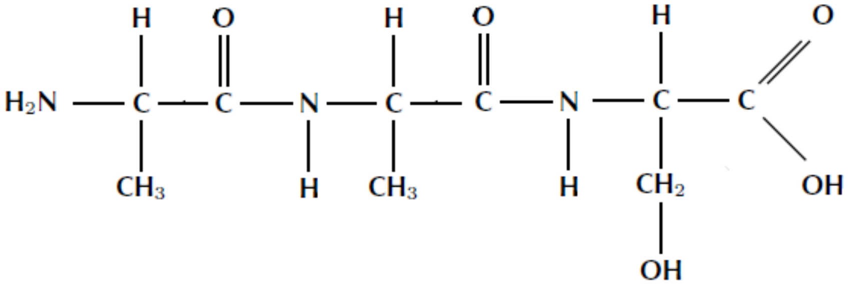
# Ambiguity codes



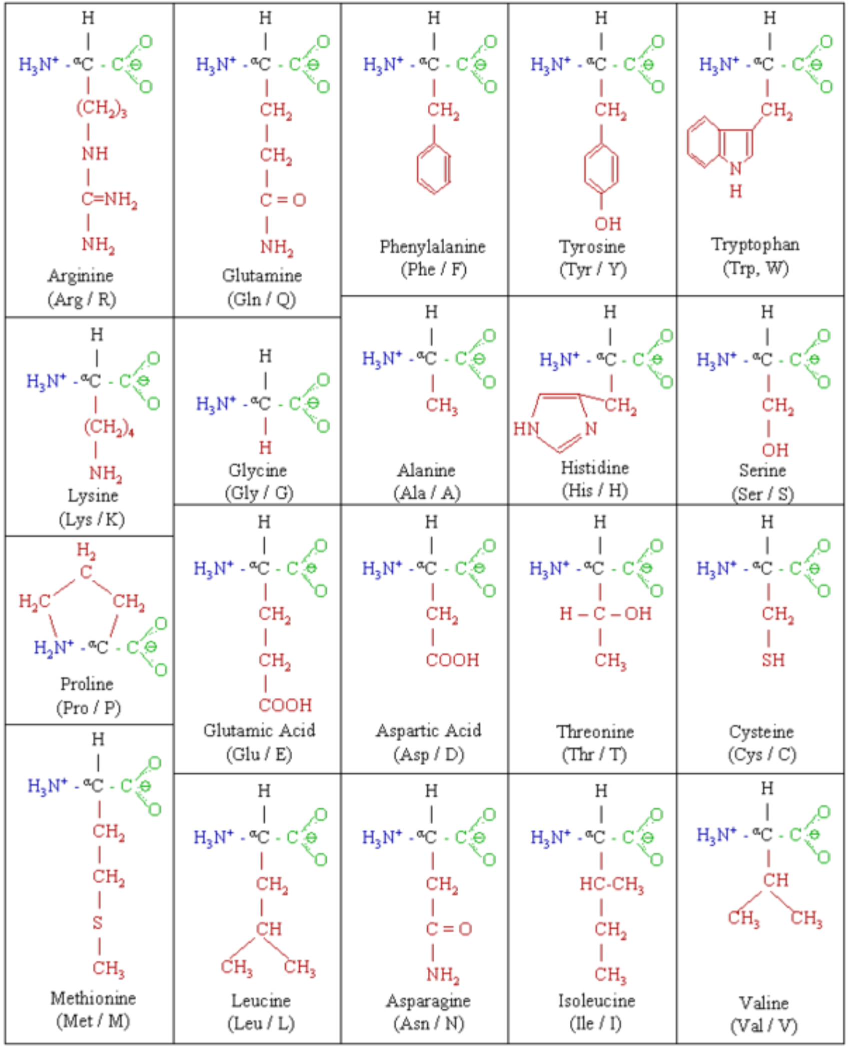


| Code | Meaning | Etymology |
|---|---|---|
| A | A | **A**denosine |
| T/U | T | **T**hymidine/**U**ridine |
| G | G | **G**uanine |
| C | C | **C**ytidine |
| K | G or T | **K**eto |
| M | A or C | A**m**ino |
| R | A or G | Pu**r**ine |
| Y | C or T | P**y**rimidine |
| S | C or G | **S**trong |
| W | A or T | **W**eak |
| B | C or G or T | not A (**B** comes after A) |
| V | A or C or G | not T/U (**V** comes after U) |
| H | A or C or T | not G (**H** comes after G) |
| D | A or G or T | not C (**D** comes after C) |
| X/N | G or A or T or C | a**n**y |

# Protein



AAS

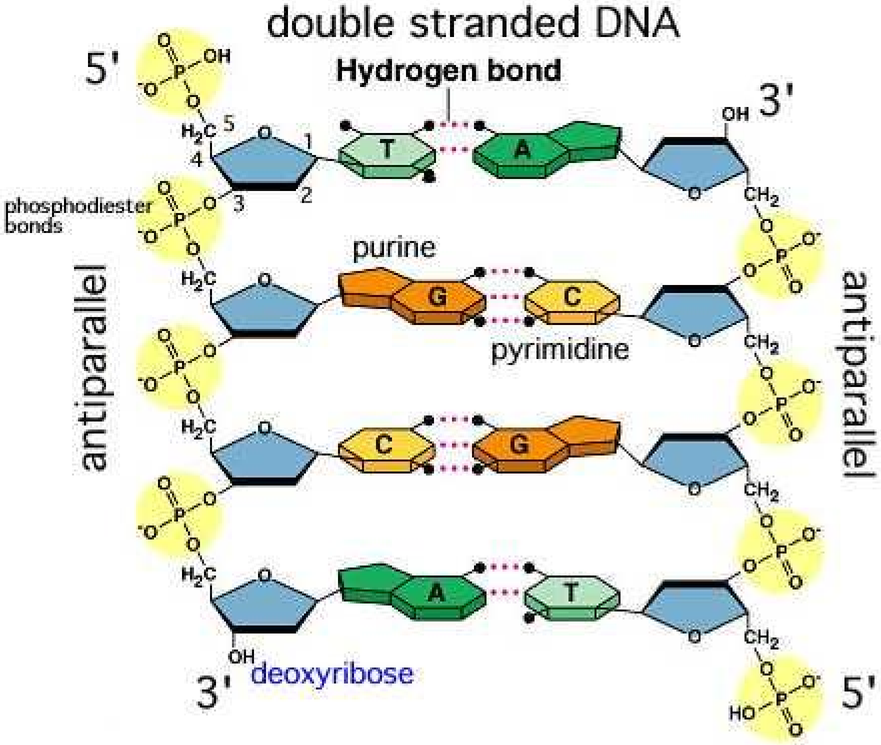

Arginine
(Arg / R)
Glutamine
(Gln / Q)
Phenylalanine
(Phe / F)
Tyrosine
(Tyr / Y)
Tryptophan
(Trp, W)
Lysine
(Lys / K)
Glycine
(Gly / G)
Alanine
(Ala / A)
Histidine
(His / H)
Serine
(Ser / S)
Proline
(Pro / P)
Glutamic Acid
(Glu / E)
Aspartic Acid
(Asp / D)
Threonine
(Thr / T)
Cysteine
(Cys / C)
Methionine
(Met / M)
Leucine
(Leu / L)
Asparagine
(Asn / N)
Isoleucine
(Ile / I)
Valine
(Val / V)

# Hydrogen Bonds

```
5’-TGCA-3’
   ||||
3’-ACGT-5’

   TGCA
   ||||
   ACGT

   TGCA
```



# Mismatches, Insertions, Deletions

```
TGTA      TG-A      TGCA
|| |      || |      | ||
TGCA      TGGA      T-CA
```

# Alignment and Similarity

```
   AC-TGAACTACG
   || ||| || ||
ATCACGTGATCT-CGAT
```

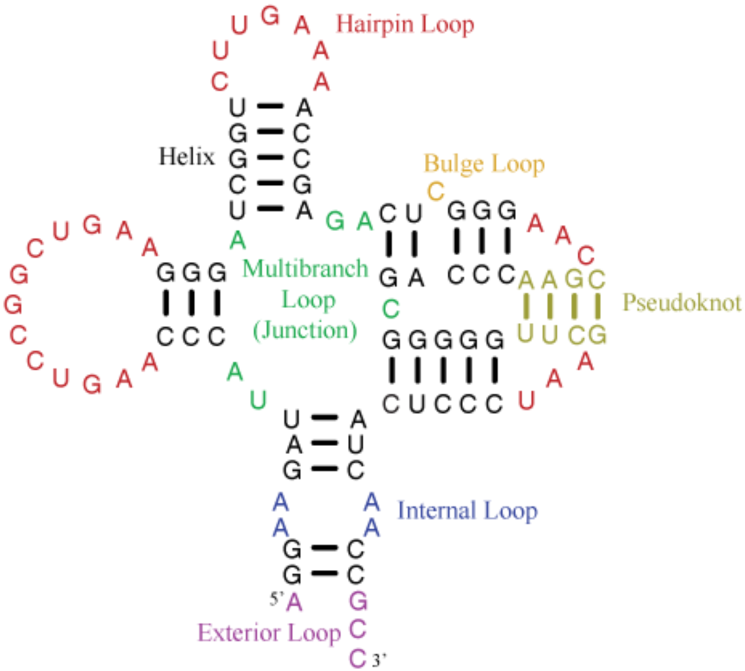
# Primary Structure

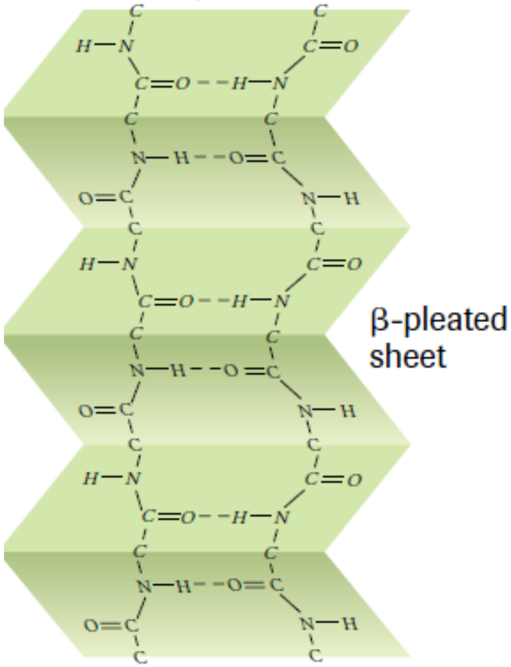
ATGATGCATCGT

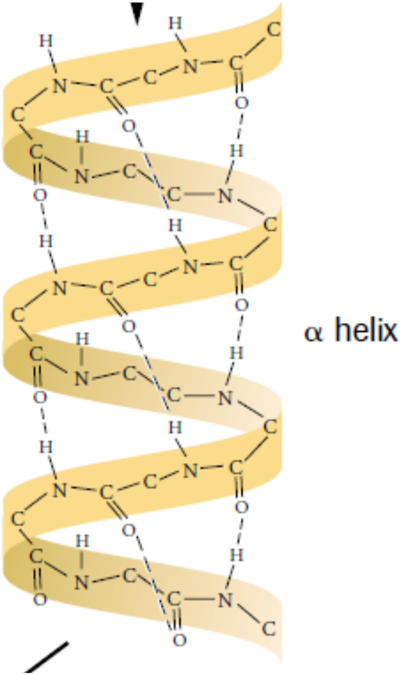
AUGAUGCAUCGU

MASWPQCPYKEDEN

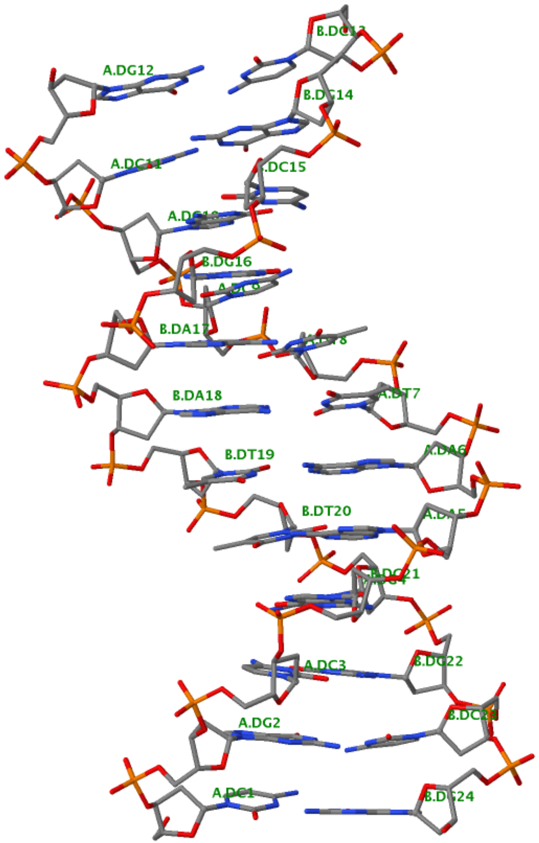
# Secondary Structure

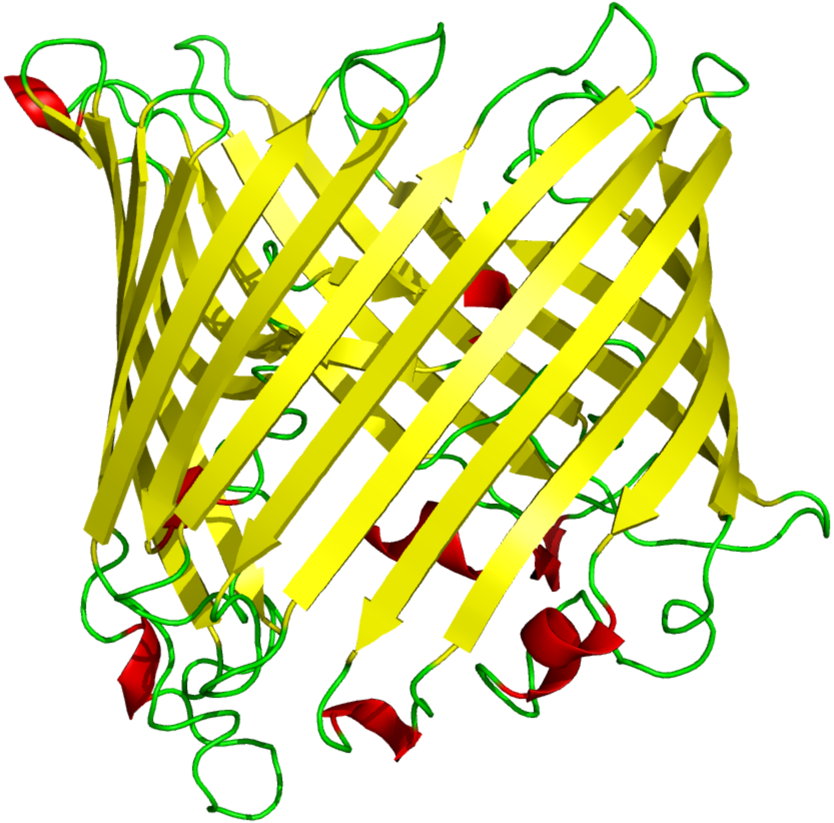




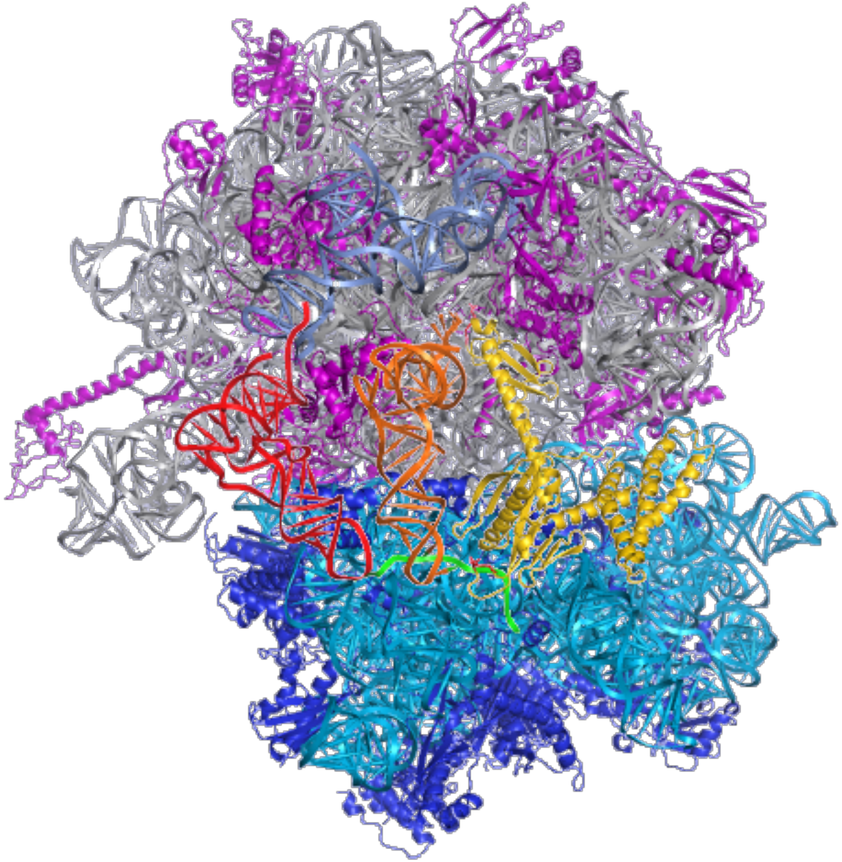


# Tertiary Structure





# Quaternary Structure

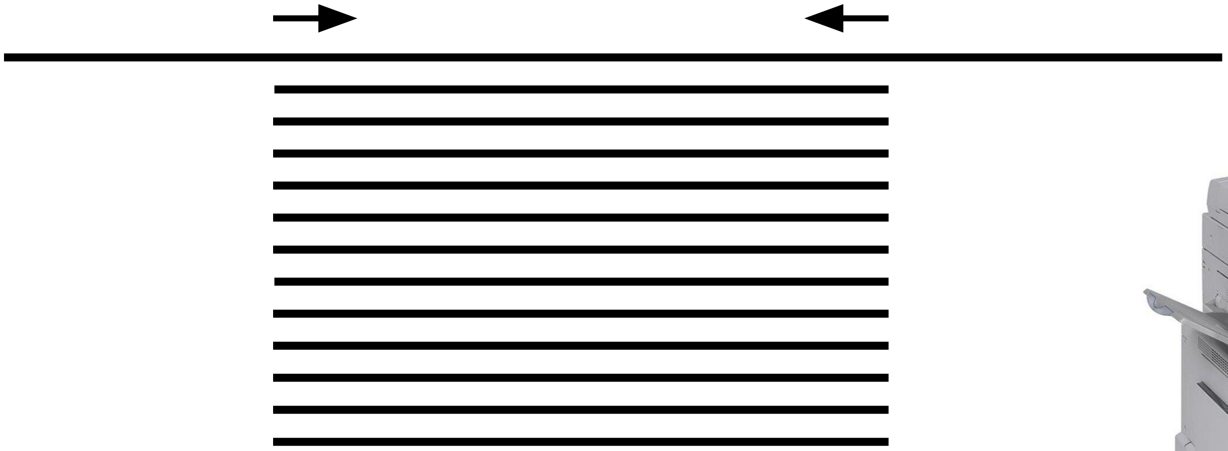


# Primer Binding and Extension

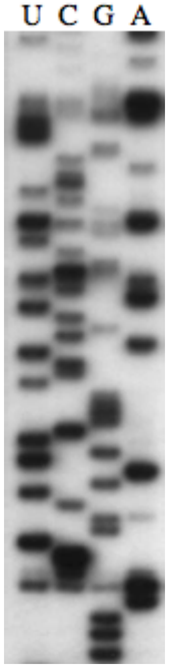
```
                          TGATGCTGTCGTAGT

          TGATGCTGTCGTAGTCGTAGCTGATCGATGCTGCCCATG
          |||||||||||||||||||||||||||||||||||||||
ACTAGCATCGACTACGACAGCATCAGCATCGACTAGCTACGACGGGTAC
```

# PCR

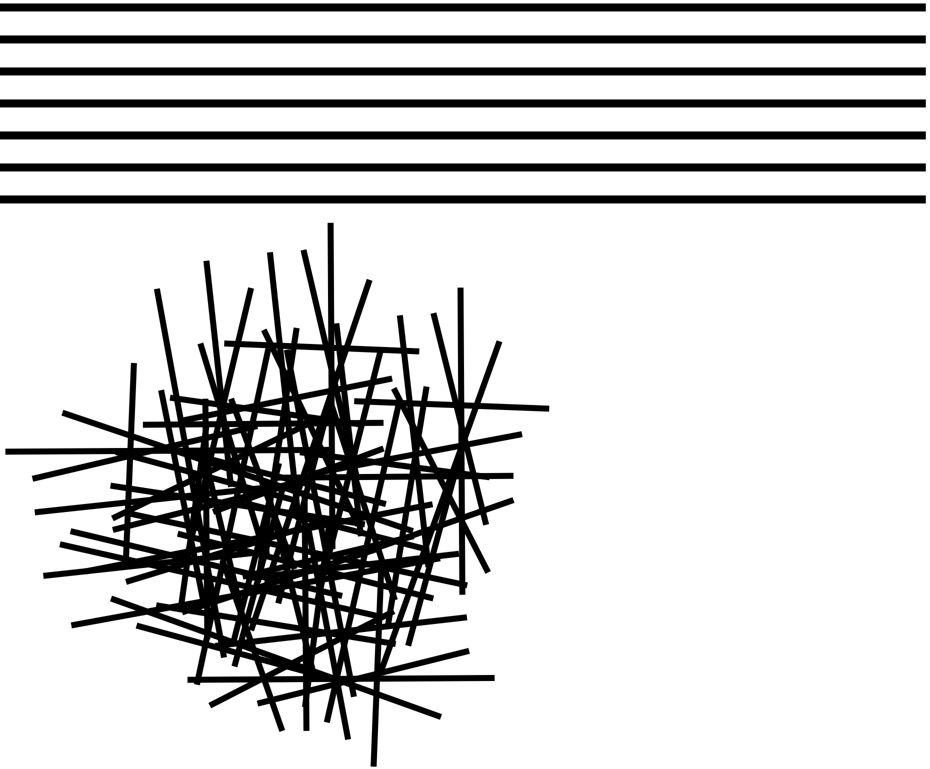


# Sequencing

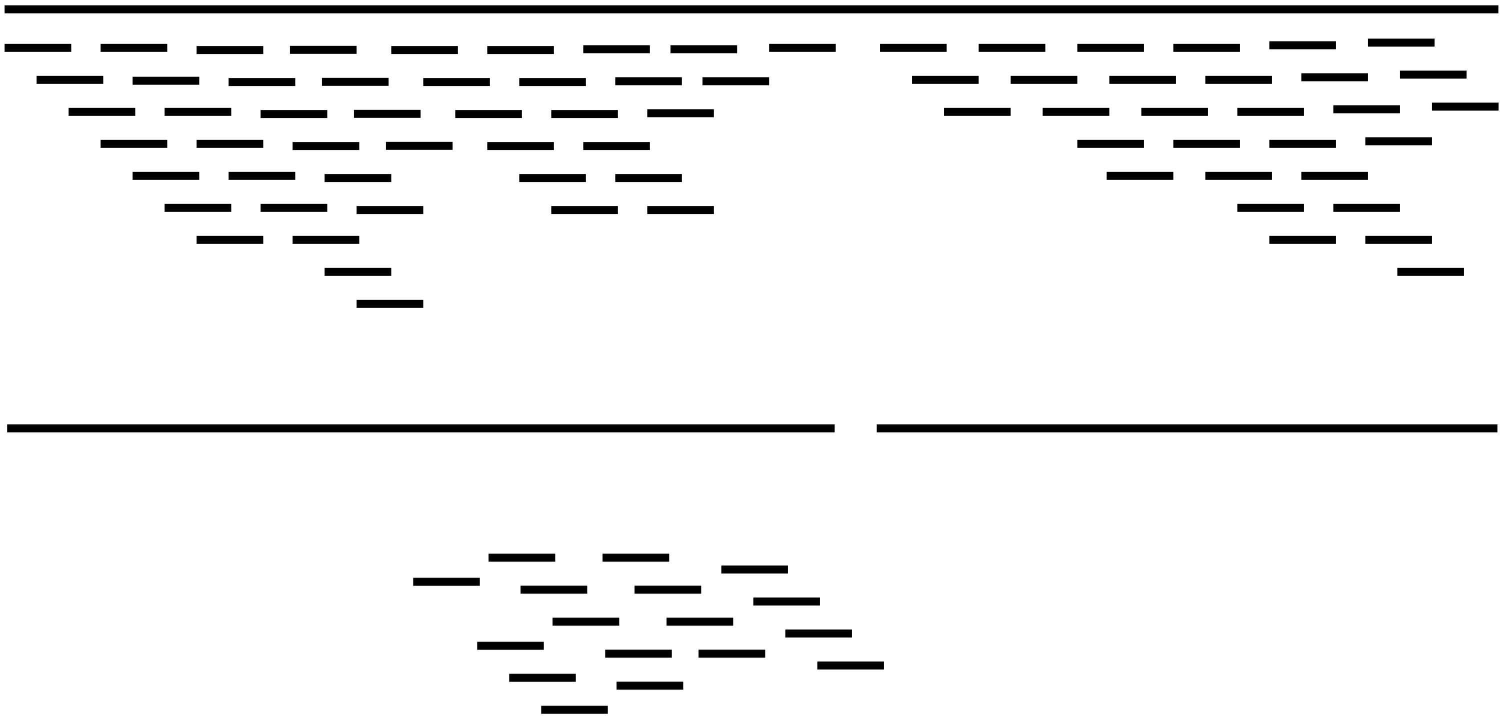


3.6Tb of data in six days

# Shotgun Sequencing

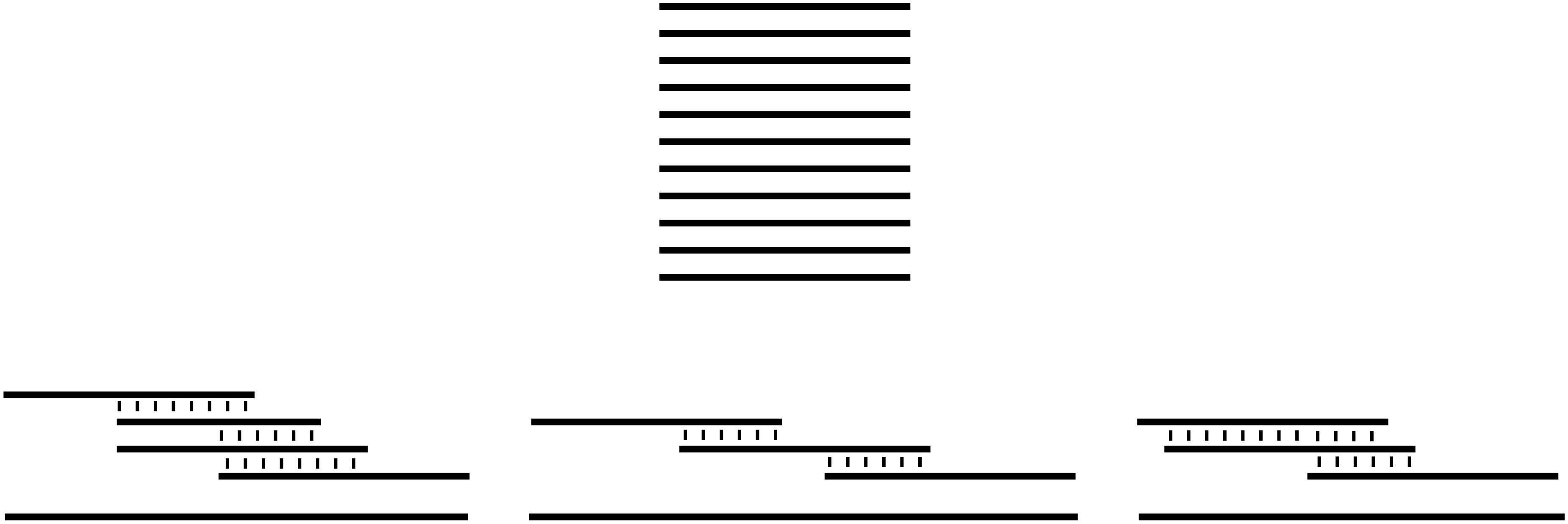


# Mapping Assembly

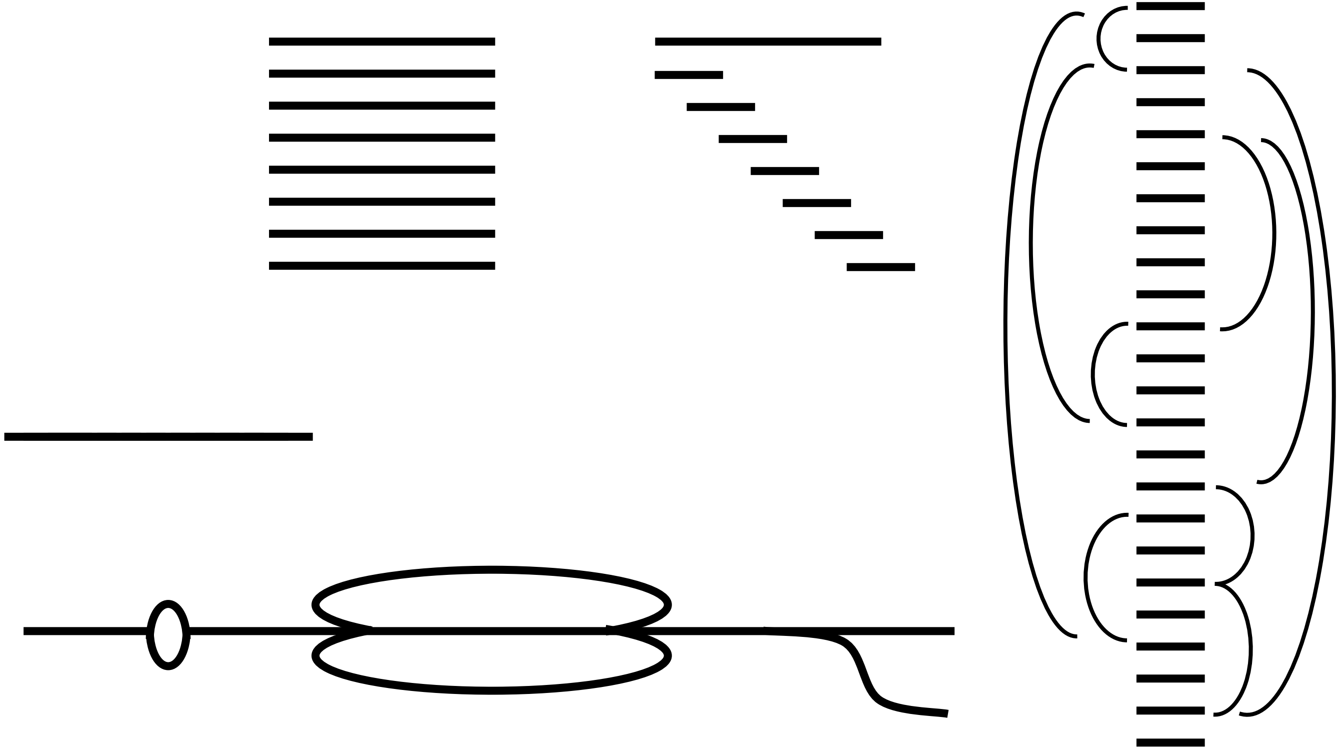


# De Novo Assembly

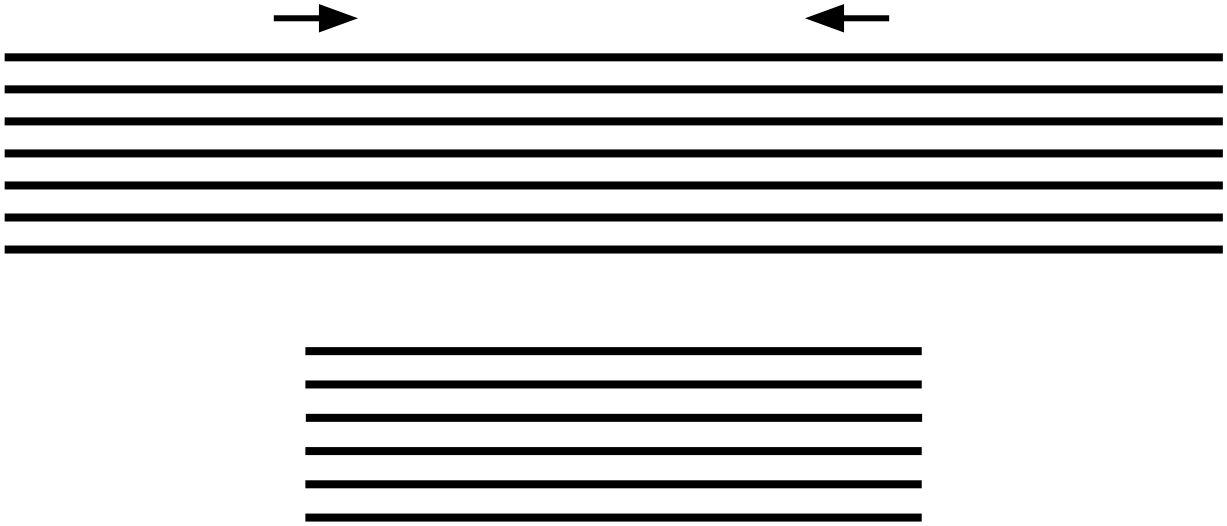
# Overlap Consensus



# De Bruijn Graph



# Amplicon Sequencing



# FASTA format (.fna, .faa, .fa, .fasta)

```
>id0
AGTAGTAGT
ATGAGAGAT
ATGT
>id1
AGTAGTAGG
ATGATGA
```

# FASTQ format (.fq, .fastq)

```
@id0
ATCGACTGCA
+
!babc@!h!1
@id1
TGGTAGTAGT
+
bab$#@11@a
```

# Real Life Data

- http://hgdownload.cse.ucsc.edu/goldenPath/hg18/chromosomes/
- ftp://ftp_20150211_13111:f7nFq+q+fGT+@ftp.dna.ku.dk*
    - HiSeq shotgun data 1 sample, ~60M x 70b
    - MiSeq amplicon data ~400 samples, 56-60K x ~250b
    - scan_for_matches patterns and examples

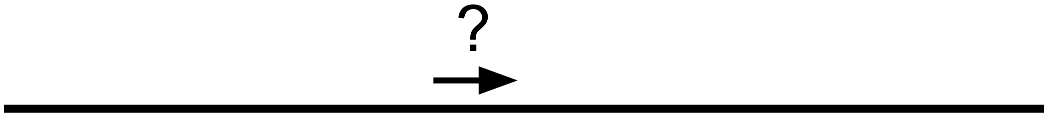
* Will self destruct in 2015-02-11 + 60 days

# The Human Genome

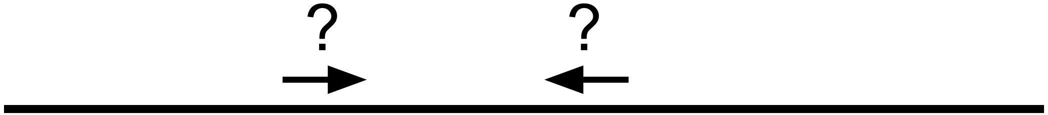
| Chr | Size |
|---|---|
| 1 | 249,250,621 |
| 2 | 243,199,373 |
| 3 | 198,022,430 |
| 4 | 191,154,276 |
| 5 | 180,915,260 |
| 6 | 171,115,067 |
| 7 | 159,138,663 |
| 8 | 146,364,022 |
| 9 | 141,213,431 |
| 10 | 135,534,747 |
| 11 | 135,006,516 |
| 12 | 133,851,895 |
| 13 | 115,169,878 |
| 14 | 107,349,540 |
| 15 | 102,531,392 |
| 16 | 90,354,753 |
| 17 | 81,195,210 |
| 18 | 78,077,248 |
| 19 | 59,128,983 |
| 20 | 63,025,520 |
| 21 | 48,129,895 |
| 22 | 51,304,566 |
| X | 155,270,560 |
| Y | 59,373,566 |
| MT | 16,569 |

# Projects

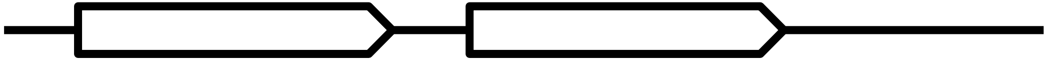
# scan_for_matches



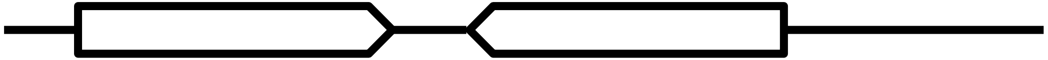
```
ATGTGTWSTTGCGT[2,1,1]
```



```
ATGTGTWSTTGCGT[2,1,1] 2000...3000 GGACTAGCTACGATC[2,1,1]
```

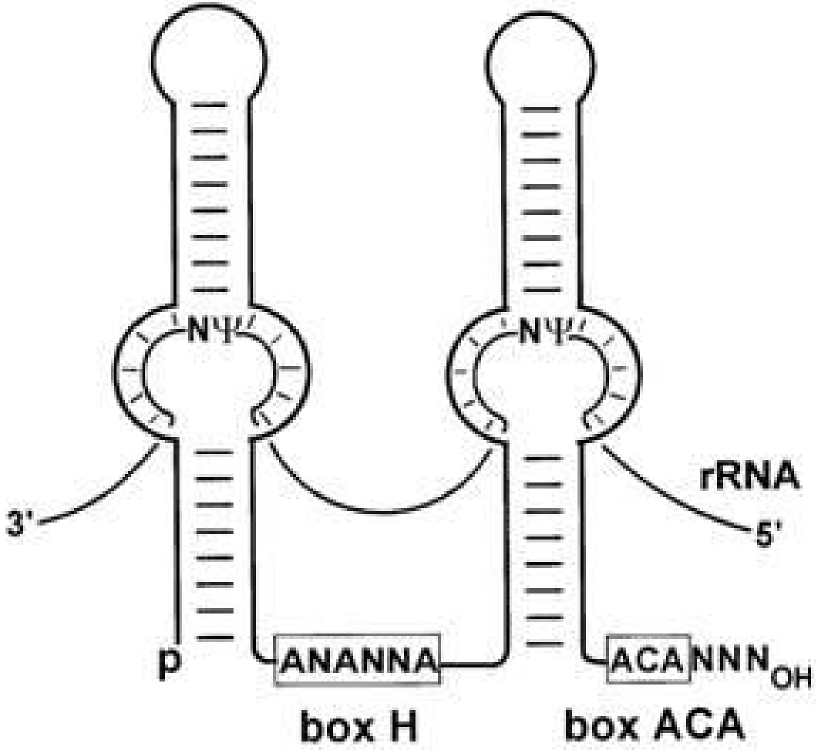


```
p1=10...50 2000...3000 p1[2,1,1]
```



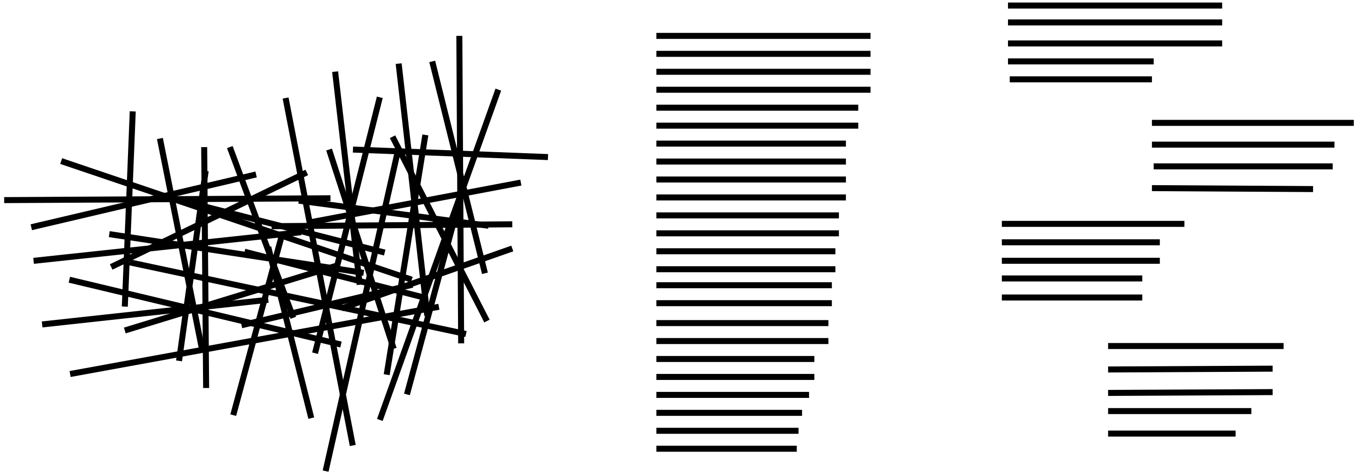
```
p1=10...50 2000...3000 ~p1[2,1,1]
```

```
p1=4...8
4...8
p2=4...8
4...4
~p2
4...8
~p1
0...4
ANANNA
0...4
p3=4...8
4...8
p4=4...8
4...4
~p4
4...8
~p3
0...4
ACANNN
```



# KMC REGEX

# Clustering



???

Martin Asser Hansen
maahansen@bio.ku.dk

http://www1.bio.ku.dk/microbiology/