Lecture 4
Sequence Analysis

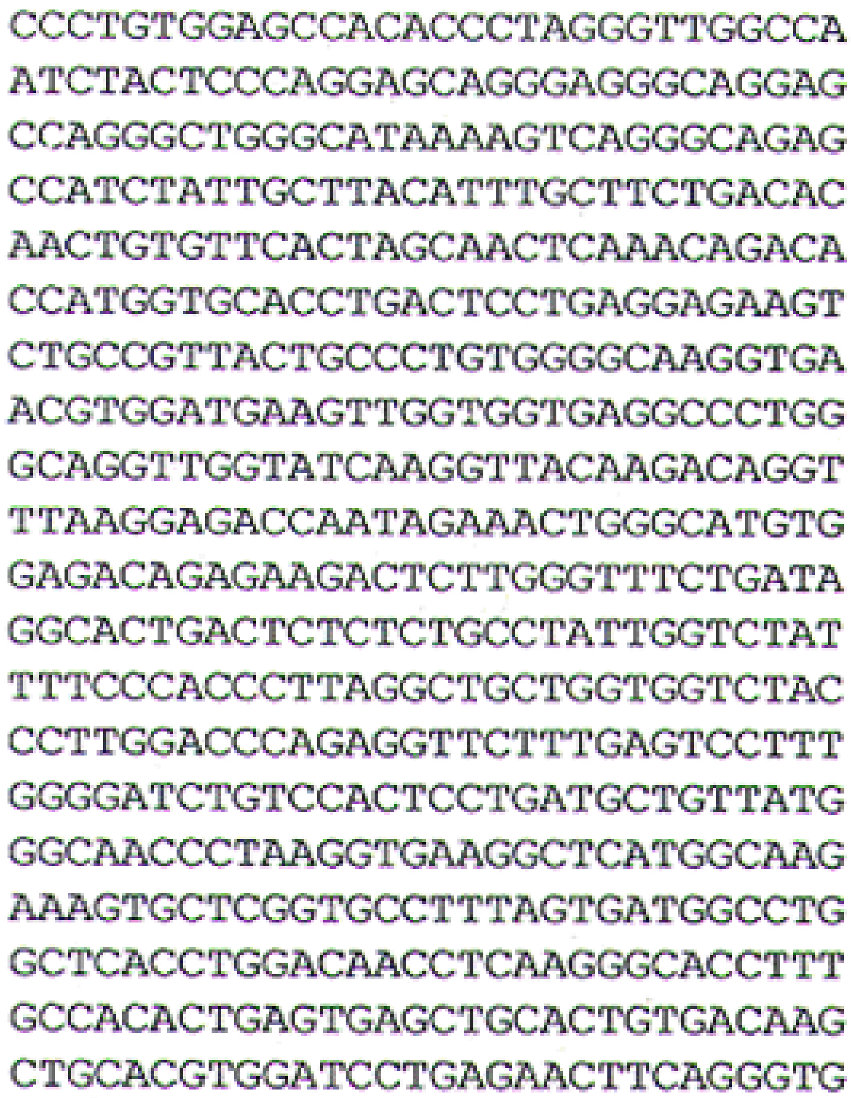
Part I – DNA sequence analysis

Part II – Protein sequence analysis

What is sequence analysis?

- Nucleic acids: DNA and RNA
- Proteins:

5’



```
CCCTGTGGAGCCACACCCTAGGGTTGGCCA
ATCTACTCCCAGGAGCAGGGAGGGCAGGAG
CCAGGGCTGGGCATAAAAGTCAGGGCAGAG
CCATCTATTGCTTACATTTGCTTCTGACAC
AACTGTGTTCACTAGCAACTCAAACAGACA
CCATGGTGCACCTGACTCCTGAGGAGAAGT
CTGCCGTTACTGCCCTGTGGGGCAAGGTGA
ACGTGGATGAAGTTGGTGGTGAGGCCCTGG
GCAGGTTGGTATCAAGGTTACAAGACAGGT
TTAAGGAGACCAATAGAAACTGGGCATGTG
GAGACAGAGAAGACTCTTGGGTTTCTGATA
GGCACTGACTCTCTCTGCCTATTGGTCTAT
TTTCCCACCCTTAGGCTGCTGGTGGTCTAC
CCTTGGACCCAGAGGTTCTTTGAGTCCTTT
GGGGATCTGTCCACTCCTGATGCTGTTATG
GGCAACCCTAAGGTGAAGGCTCATGGCAAG
AAAGTGCTCGGTGCCTTTAGTGATGGCCTG
GCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAG
CTGCACGTGGATCCTGAGAACTTCAGGGTG
```

3’

What Do You Want to Know?

Why do sequence analysis?

- Genetic information carrier
  - DNA or RNA → Protein
- Genetic information carried
  - Sequence
- Hence:

Y（Life） = *f* (Sequence)

Digital Archives

Why do sequence analysis?

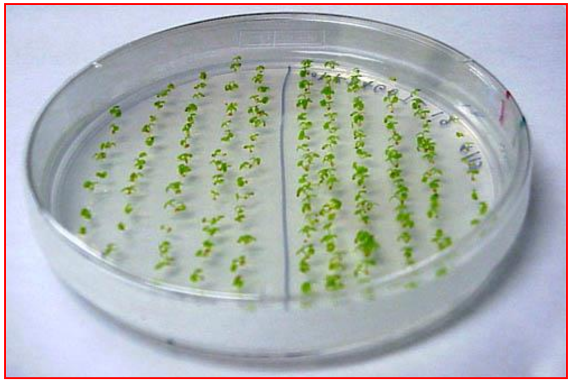
- Human Genome Project (completed in 2001).

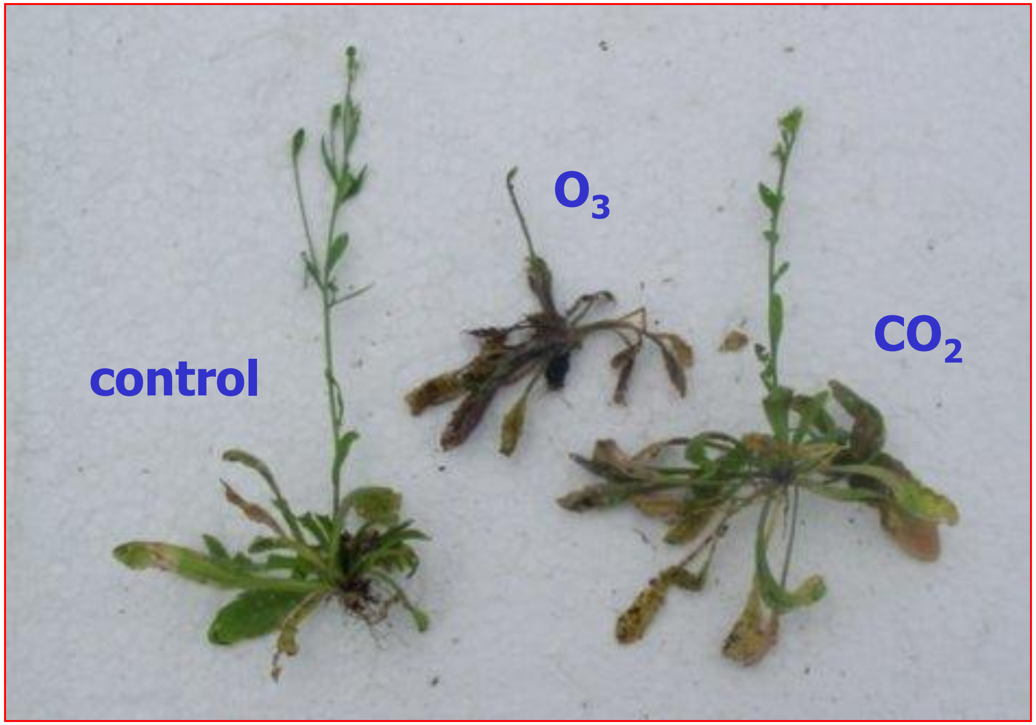
- High throughput technology…

Arabidopsis – 拟南芥 model plant

small, fast, prolific,
mutants,
lines, ecotypes,
genome sequence

Field on
a dish!





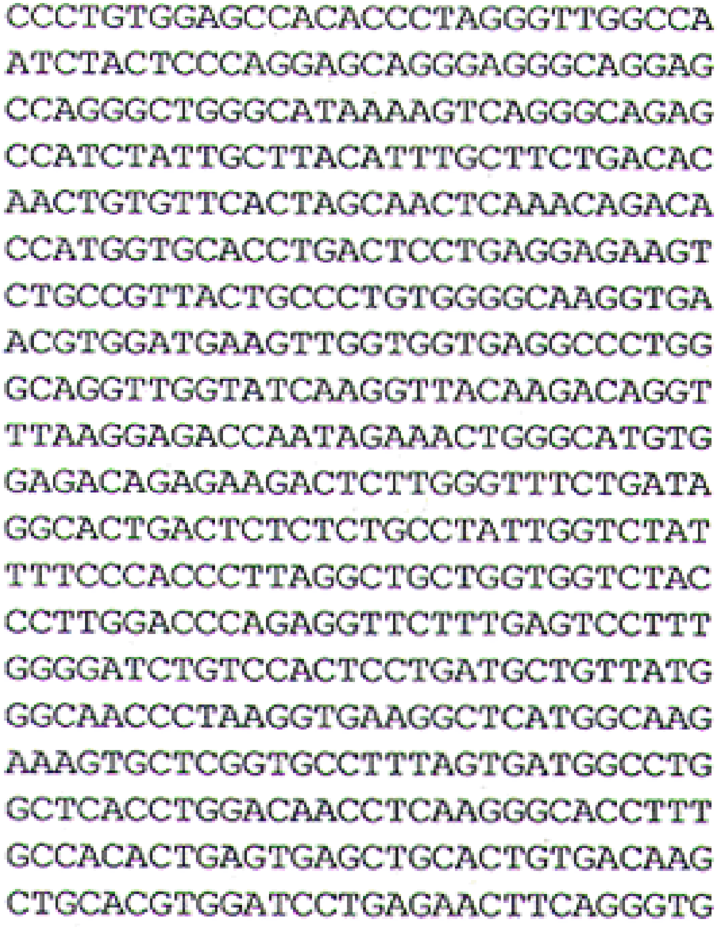
Columbia grown in Soy-FACE

Y（Life-i） =ώ *f* (Sequence) ±ß

When you have a sequence

What Do You Want to Know?

5’



```
CCCTGTGGAGCCACACCCTAGGGTTGGCCA
ATCTACTCCCAGGAGCAGGGAGGGCAGGAG
CCAGGGCTGGGCATAAAAGTCAGGGCAGAG
CCATCTATTGCTTACATTTGCTTCTGACAC
AACTGTGTTCACTAGCAACTCAAACAGACA
CCATGGTGCACCTGACTCCTGAGGAGAAGT
CTGCCGTTACTGCCCTGTGGGGCAAGGTGA
ACGTGGATGAAGTTGGTGGTGAGGCCCTGG
GCAGGTTGGTATCAAGGTTACAAGACAGGT
TTAAGGAGACCAATAGAAACTGGGCATGTG
GAGACAGAGAAGACTCTTGGGTTTCTGATA
GGCACTGACTCTCTCTGCCTATTGGTCTAT
TTTCCCACCCTTAGGCTGCTGGTGGTCTAC
CCTTGGACCCAGAGGTTCTTTGAGTCCTTT
GGGGATCTGTCCACTCCTGATGCTGTTATG
GGCAACCCTAAGGTGAAGGCTCATGGCAAG
AAAGTGCTCGGTGCCTTTAGTGATGGCCTG
GCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAG
CTGCACGTGGATCCTGAGAACTTCAGGGTG
```

3’

一场关于蘑菇的官司

When you have a sequence

- Is it likely to be a gene?
- What is the possible expression level?
- What is the possible protein product?
- Can we get the protein product?
- Can we figure out the key residue in the protein product?
- ……

Part I – DNA sequence analysis

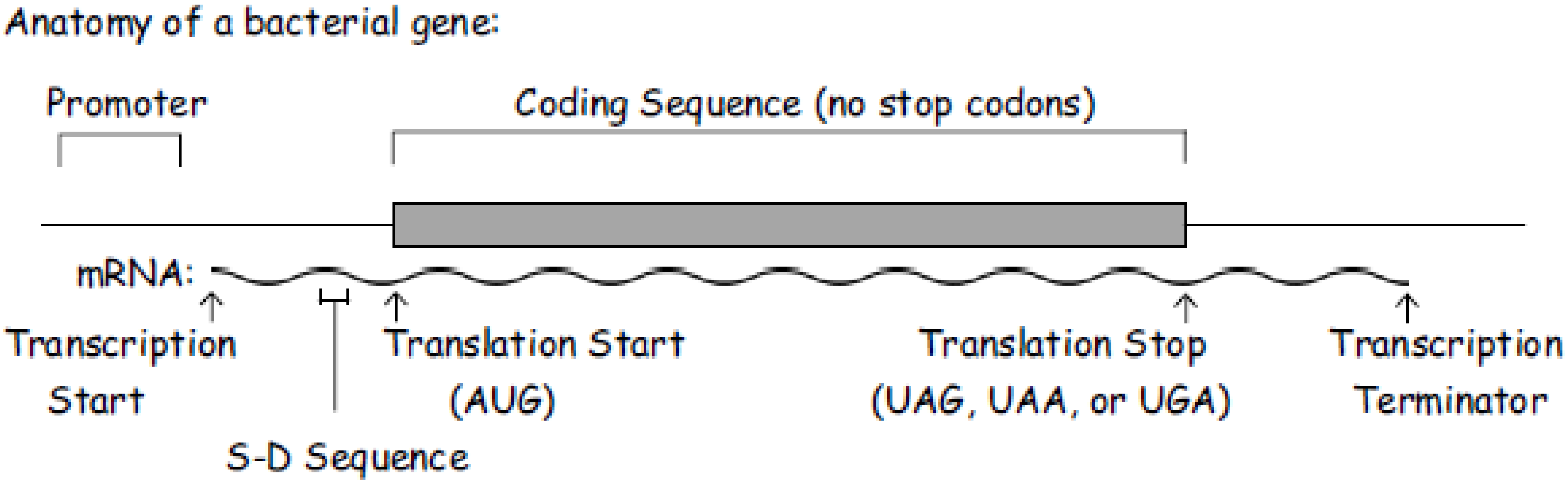
DNA sequence analysis

Related to genes:

1. GC content
2. Pattern analysis
3. Gene finding(Open Reading Frame detection)
4. Translation
5. Mutation
6. Primer design
7. Restriction map

- ……

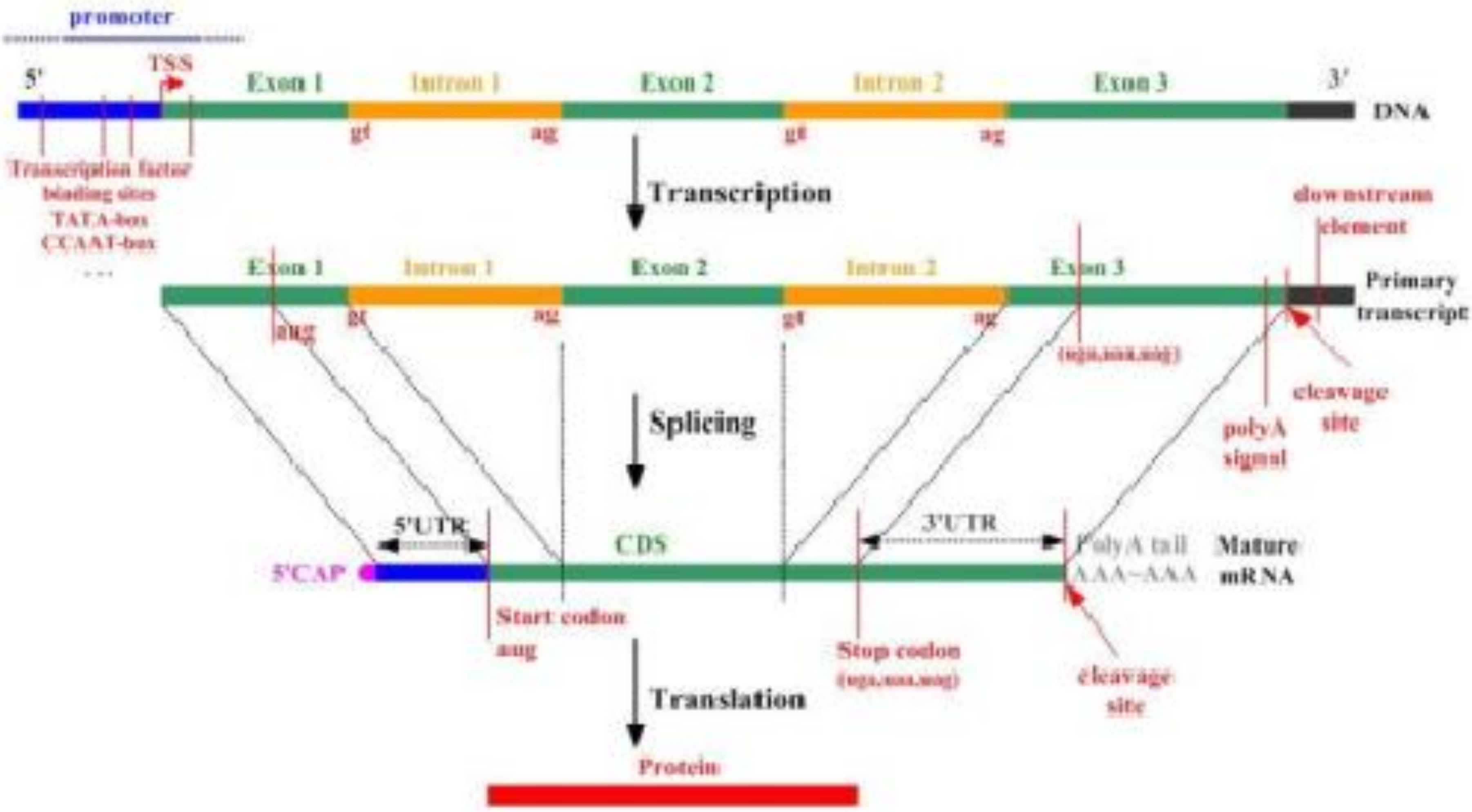
Gene Structure of Bacterial



Anatomy of a bacterial gene:
Promoter
Coding Sequence (no stop codons)
mRNA:
Transcription Start
S-D Sequence
Translation Start (AUG)
Translation Stop (UAG, UAA, or UGA)
Transcription Terminator

Sequence Element	Function
Promoter	To target RNA polymerase to DNA and to start transcription of a mRNA copy of the gene sequence.
Transcription terminator	To instruct RNA polymerase to stop transcription.
Shine-Dalgarno sequence and translation start	S-D sequence in mRNA will load ribosomes to begin translation. Translation almost always begins at an AUG codon in the mRNA (an ATG in the DNA becomes an AUG in the mRNA copy). Synthesis of the protein thus begins with a methionine.
Coding Sequence	Once translation starts, the coding sequence is translated by the ribosome along with tRNAs which read three bases at a time in linear sequence. Amino acids will be incorporated into the growing polypeptide chain according to the genetic code.
Translation Stop	When one of the three stop codons [UAG (amber), UAA (ochre), or UGA] is encountered during translation, the polypeptide will be released from the ribosome.

Gene


promoter
5'
TSS
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
3'
DNA
gt
ag
gt
ag
Transcription factor
binding sites
TATA-box
CCAAT-box
...
Transcription
downstream
element
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Primary
transcript
aug
gt
ag
gt
ag
(uga,uaa,uag)
Splicing
cleavage
site
polyA
signal
5'UTR
CDS
3'UTR
PolyA tail
AAA~AAA
Mature
mRNA
5'CAP
Start codon
aug
Stop codon
(uga,uaa,uag)
cleavage
site
Translation
Protein

1. GC content

- Stability
  - GC: 3 hydrogen bonds
  - AT: 2 hydrogen bonds
- GC rich fragment

  → Gene…?
- Gene regulation? Mutation?

calculation:

GC content：

$$\boxed{\frac{G+C}{A+T+G+C} \times 100}$$

AT/GC ratio：

$$\boxed{\frac{A+T}{G+C}}$$

GC content variation

1. 酵母 (*Saccharomyces cerevisiae*)： 38%,
2. 人类: ~40%

3. 天蓝色链霉菌 (*Streptomyces coelicolor*)*:?*
4. 恶性疟原虫(*Plasmodium falciparum*)： ？

（AT-rich instead of GC-poor）

2. CpG island

- **CpG islands** or **CG islands** (CGI) are genomic regions that contain a high frequency of CpG sites.

- In a CpG site, both C and G are found on the same strand of DNA or RNA and are connected by a phosphodiester bond .

- . The "p" in CpG refers to the phosphodiester bond .This is a covalent bond between atoms, stable and permanent as opposed to the three hydrogen bonds established after base-pairing of C and G in opposite strands of DNA.

2. CpG island

- Related to methylation
- Associated with genes which are frequently switched on
- Estimate: ½ mammalian gene have CpG island
- Most mammalian housekeeping genes have CpG island at 5′ end

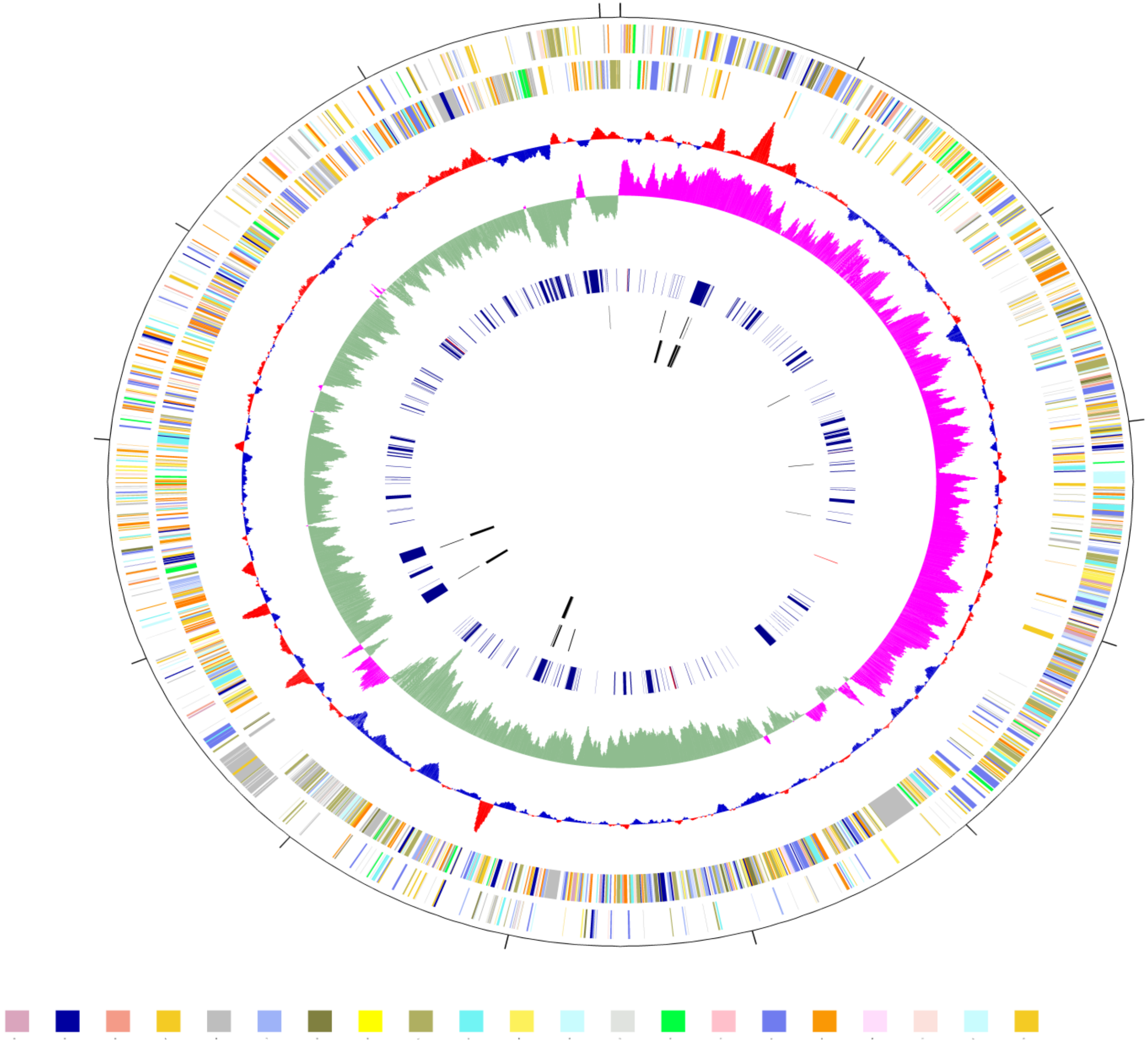
CpG岛查找

- 用一个长度（eg.200bp）的窗口移过序列，每次移一个碱基对，进行计算。

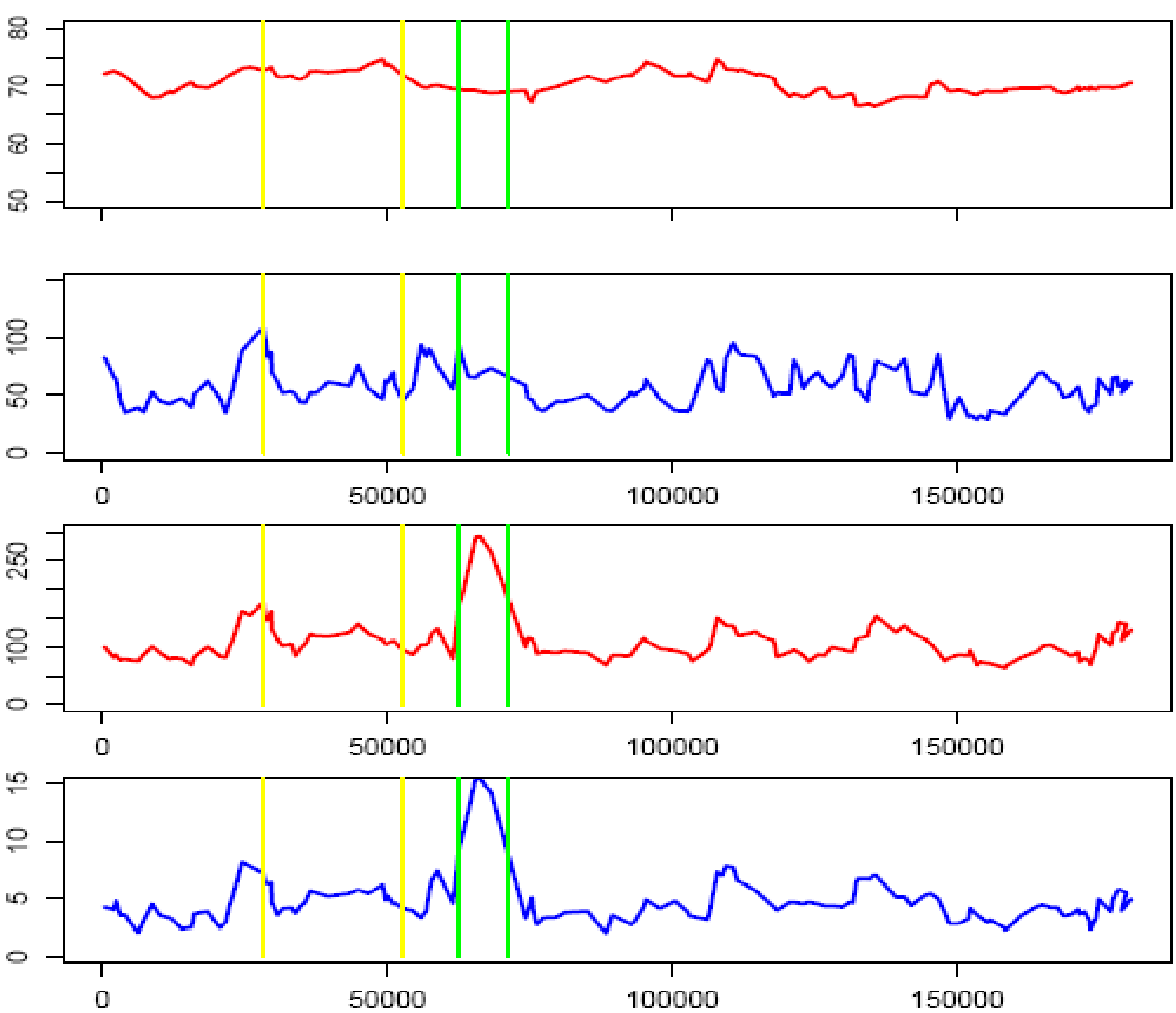
- CpG岛: GC含量大于50%,且出现 CpG 的概率高与随机的一段序列区域。

- CpG岛主要位于基因的启动子和第一外显子区域，约有60％以上基因的启动子含有CpG岛。
- CpG甲基化的研究在肿瘤的研究中有着非常主要的地位。通过基因启动子区及附近区域CpG岛胞嘧啶的甲基化可以在转录水平调节基因的表达，从而引起相应基因沉默，去甲基化又可恢复其表达。

Genome Annotation

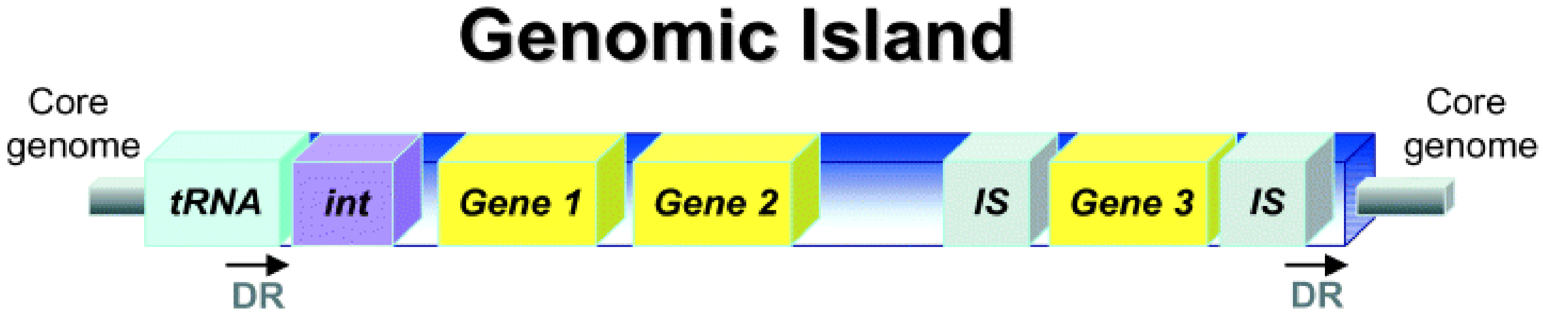


Genomic islands



Characteristics of genomics islands

Genomic Island



Functions encoded by Genomic Islands:

Pathogenicity, Iron Uptake, Secondary Metabolism,
Antibiotic Resistance, Secretion, Degradation of Xenobiotics, Symbiosis

Impact of pathogenicity islands in bacterial diagnostics. *Apmis* **112** (11-12), 930-936.

Characterization of anomalous Genomic islands

- Genomic characterization:
  - **Compositional contrasts (standard method)**
  - **Codon usage contrasts**
  - **Amino acid contrasts**
  - 。。。.
- IS sequences, transposonase, tRNA

Genomic Islands in *Leptospira interrogans* (*钩端螺旋体*）

	GC	Codon usage	AA usage	IS	transposase	tRNA	repeat
Genome	35.00	0.011	0.0065				
element I	**39.90**	**0.023**	**0.012**	Y	Y	N	N
element II	35.10	**0.027**	**0.017**	Y	Y	Y	Y
element III	36.70	**0.023**	**0.014**	Y	Y	N	Y
element IV	**40.95**	0.019	0.0066	Y	Y	Y	Y
element V	37.00	**0.021**	**0.013**	Y	N	Y	Y

3. Pattern analysis

- Patterns in the sequence
- Associated with certain biological function
  - Transcription factor binding
  - Transcription starting
  - Transcription ending
  - Splicing
  - ……

4. Gene finding

- A kind of pattern search
- Gene structure
  - Promoter, Exon, Intron
  - Promoter: TATA box (TATAAT)
  - Exon: Open Reading Frame (ORF)
  - Intron: Only eukaryotes, have splicing signal
  - Other motifs

Gene finding

- Prokaryotes
  - No intron
  - Long open reading frame
  - High density
  - Easy to detect
- Eukaryotes
  - Have intron
  - Combination of short open reading frames
  - Low density
  - Hard to detect

5. Translation

- Six reading frames
- Open reading frame (ORF)
  - Start codon
  - Stop codon
  - Certain length
- Tools: ShowORF

Conceptual translation

+1 AATGGCAATCCGCGTAGACTAGGCA

+2 AATGGCAATCCGCGTAGACTAGGCA

+3 AATGGCAATCCGCGTAGACTAGGCA

5’ AATGGCAATCCGCGTAGACTAGGCA 3’

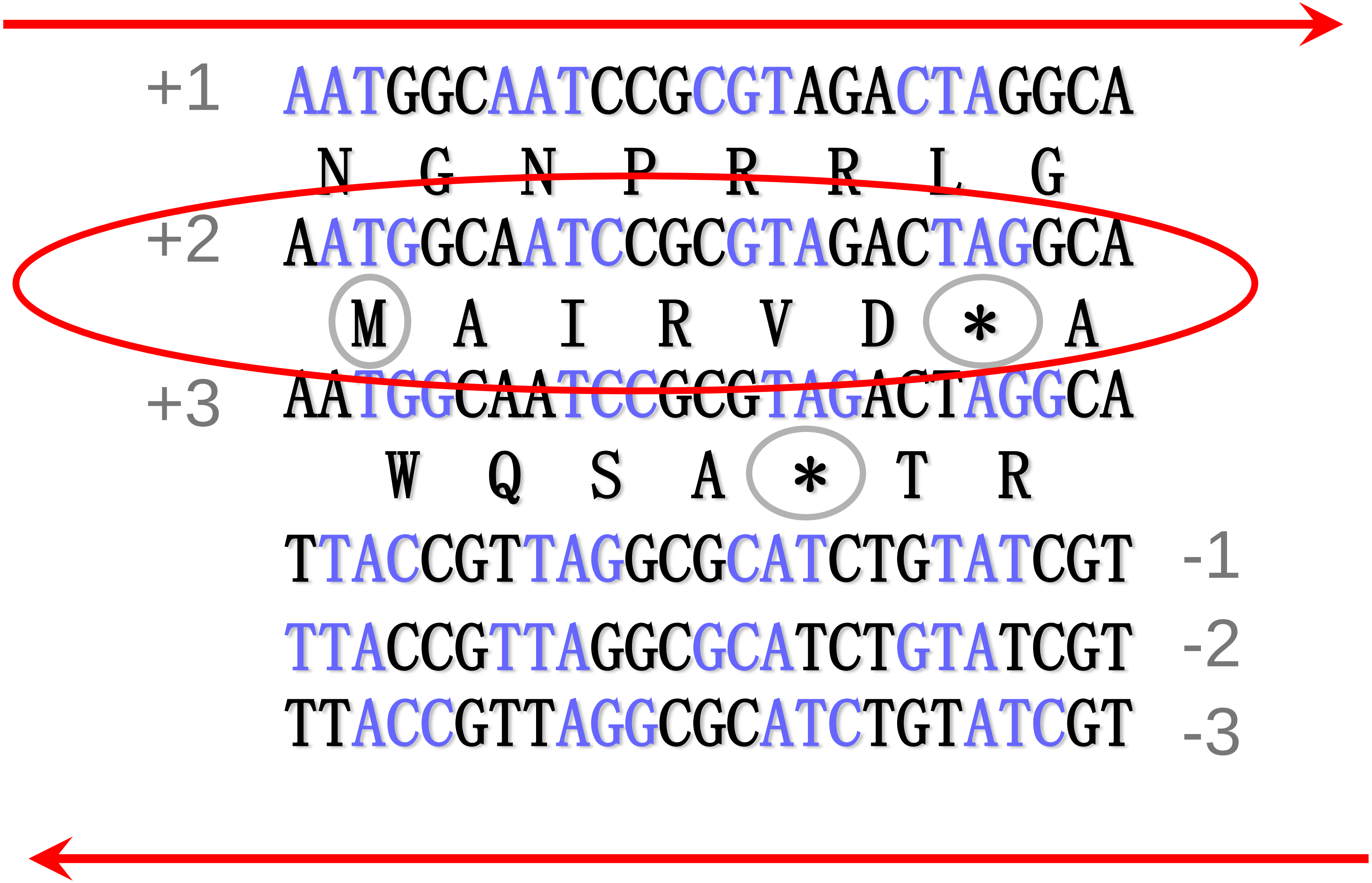
3’ TTACCGTTAGGCGCATCTGTATCGT 5’

TTACCGTTAGGCGCATCTGTATCGT -1

TTACCGTTAGGCGCATCTGTATCGT -2

TTACCGTTAGGCGCATCTGTATCGT -3

Six reading frames



```
+1   AATGGCAATCCGCGTAGACTAGGCA
      N  G  N  P  R  R  L  G
+2   AATGGCAATCCGCGTAGACTAGGCA
       M  A  I  R  V  D  *  A
+3   AATGGCAATCCGCGTAGACTAGGCA
        W  Q  S  A  *  T  R
     TTACCGTTAGGCGCATCTGTATCGT   -1
     TTACCGTTAGGCGCATCTGTATCGT   -2
     TTACCGTTAGGCGCATCTGTATCGT   -3
```

课堂作业

- Coding sequence 1200 BP
- Length of protein?
- Molecular weight?

6. Primer design

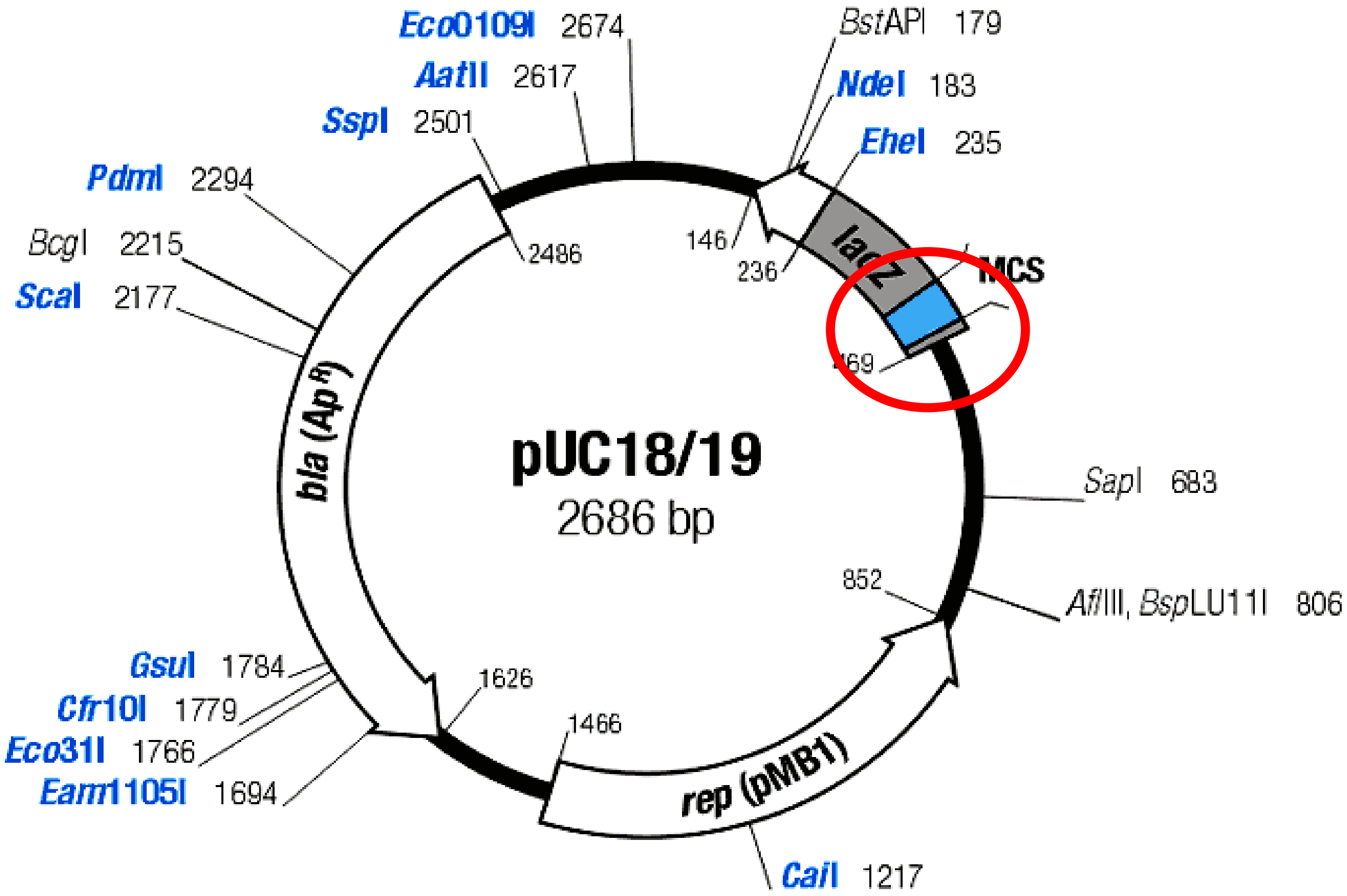
- Design primers only from accurate sequence data
- Restrict your search to regions that best reflect your goals
- Locate candidate primers
- Verification of your choice

Primer design

- Mispriming areas
- Primer length: 18-30 (Usually)
- Annealing Temperature (55 - 75 C)
- GC content: 35% - 65% (usually)
- Avoid regions of secondary structure
- 100% complimentarity is not necessary
- Avoid self-complimentarity

7. Restriction map

- Restriction enzyme
  - Recognize a pattern
  - Recognition site V.S. Cutting site
- Select restriction enzyme to get a fragment of sequence
- Rebuild the sequence to create or invalidate a restriction site
- Tools: Omiga, remap, bioedit


Eco0109I 2674
AatII 2617
SspI 2501
PdmI 2294
BcgI 2215
ScaI 2177
BstAPI 179
NdeI 183
EheI 235
146
236
2486
lacZ
MCS
469
pUC18/19
2686 bp
bla (Ap^R)
SapI 683
852
AflIII, BspLU11I 806
GsuI 1784
Cfr10I 1779
Eco31I 1766
Eam1105I 1694
1626
1466
rep (pMB1)
CaiI 1217

8. Mutation & variation

- Can be generated by PCR
  - Primers that not perfectly match
- Frame shift mutation
  - Insertion
  - Deletion
- Substitution
  - Normal
  - Silent

Mutation

- 新冠　突变啦！　突变啦！

- 突变与进化

  $\approx 10^{-6}$ /generation/gene

http://news.sciencenet.cn/htmlnews/2020/3/436836.shtm?from=timeline&isappinstalled=0

Organism – organ – tissue – cell – compartment

Euchromatin & heterochromatin – *常染色质和异染色质*

gene islands – gene

“

Promoters – 5’-regulatory (untranslated = UTR) –

introns & exons – mature coding region –

3’-regulatory (UTR) regions

”

Life *in silico*? Sure! And then: Design from Scratch

Controls for Gene Expression – many Switchboards

- Chromatin condensation state
- Local chromatin environment
- Transcription initiation
- Transcript elongation
- mRNA splicing
- mRNA export
- mRNA place in the cell
- RNA half-life
- Killer microRNAs
- Ribosome loading
- Protein transport/targeting
- Protein modifications
- Protein turnover

Levels of regulation that affect what we call "gene expression"

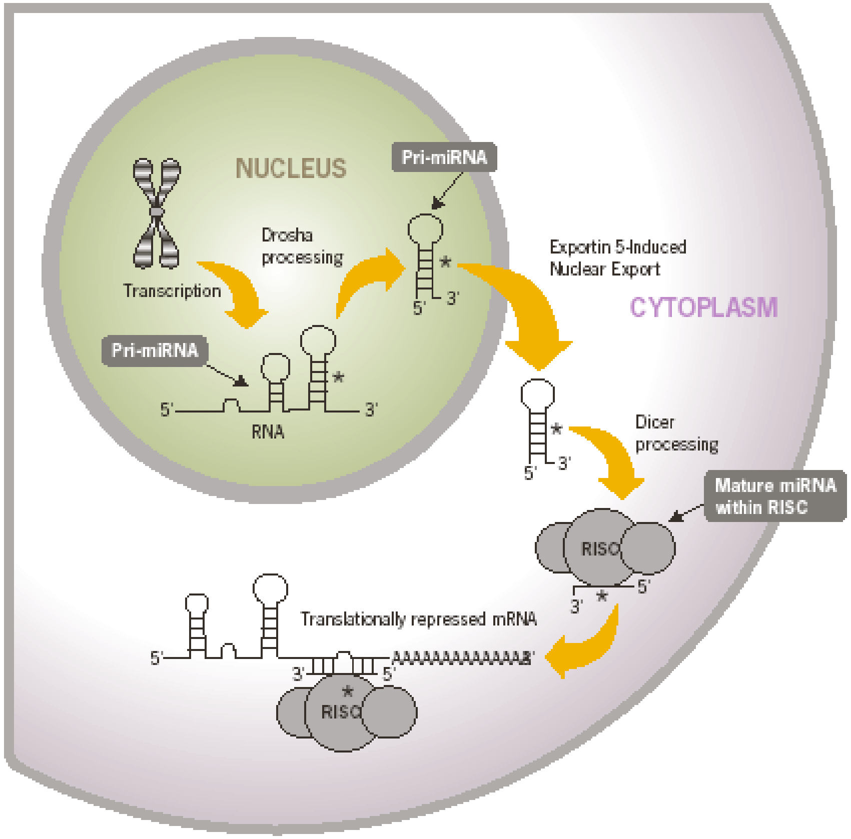
The Plant Transcriptome

years ago, we did not know that

such a control system existed!

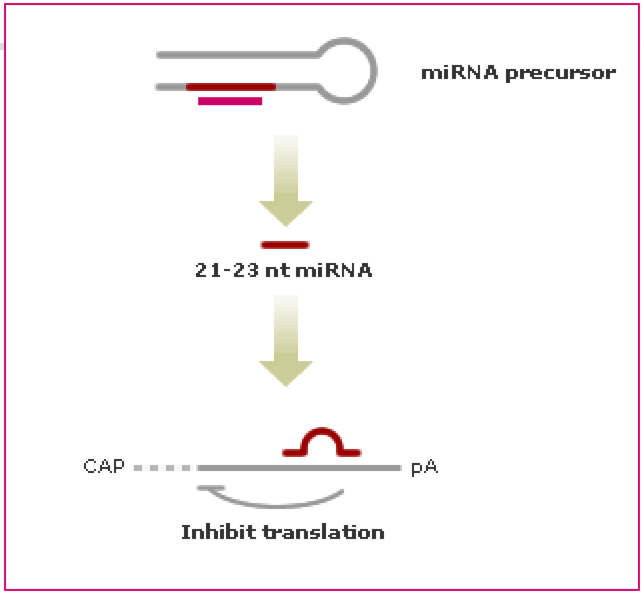
Killer RNAs

(there are micro-genes)



miRNA Processing and Activity.

microRNAs



Result: no protein
-
i.e., gene is essentially "silenced"

Genome annotation --1

- Is it a gene?
  - Not sure, but have some confidence
- What is the expression level if it is a gene?
  - Determined by the promoter and other upper stream elements

Genome annotation --2

- What is the possible product of this gene?
  - It is likely to be ….
  - This conceptual translation is in open reading frame ……
- Can we get the gene product?
  - If expression level high: Directly separate
  - If expression level low: Clone it

Genome annotation --3

- Can we get the protein product?
  - Clone it and use a bacteria to express it
- Can we figure out the key residue in the protein product?
  - Guess the important residue
  - Mutate the residue to see whether the activity loses

Summary

- Life is largely determined by nucleotide sequences
- Sequence analysis reveals patterns with biological significance
- Sequence analysis helps the design of wet-lab experiments

- Next part will be on protein sequence analysis