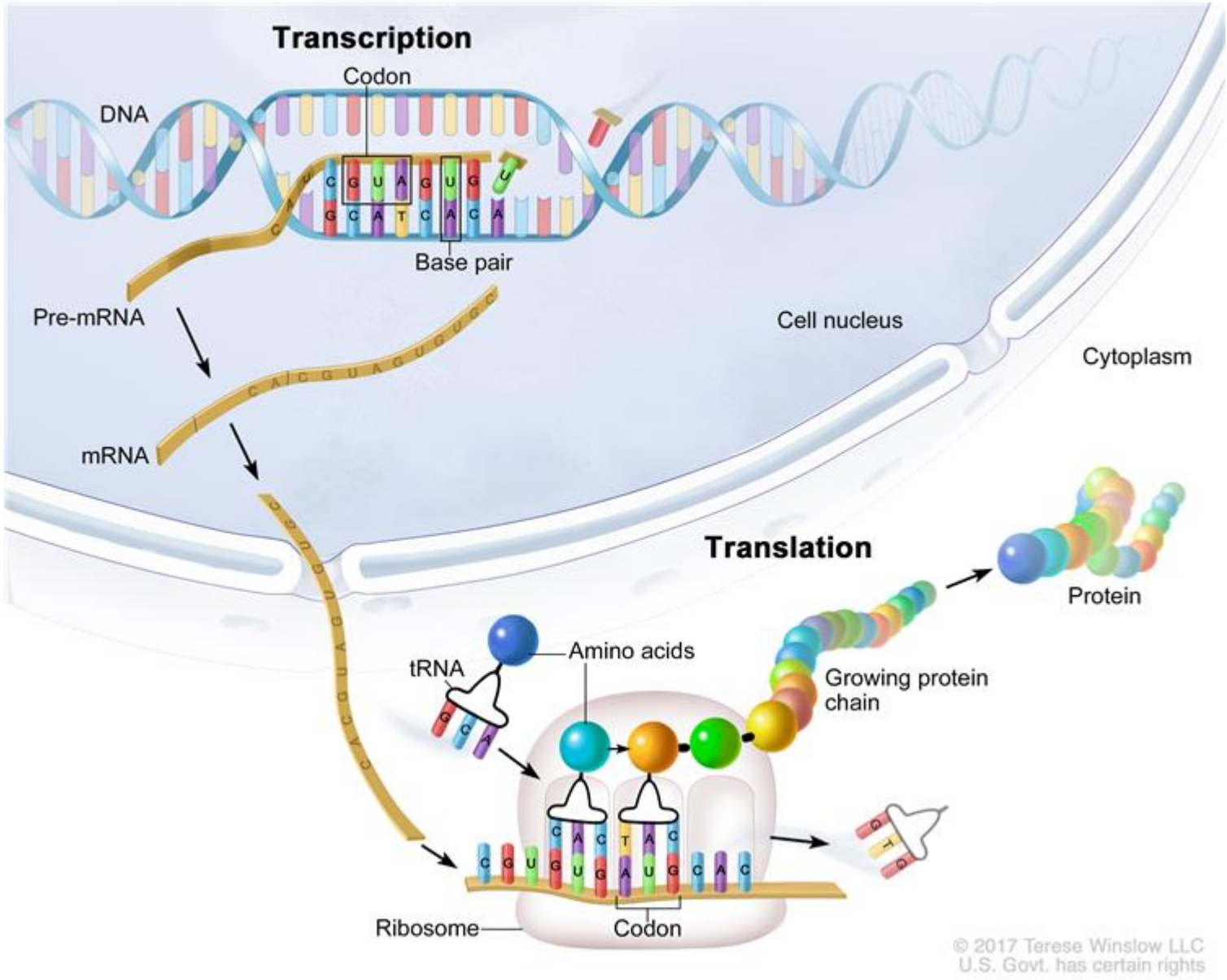
轉錄因子與DNA結合問題

張文綺

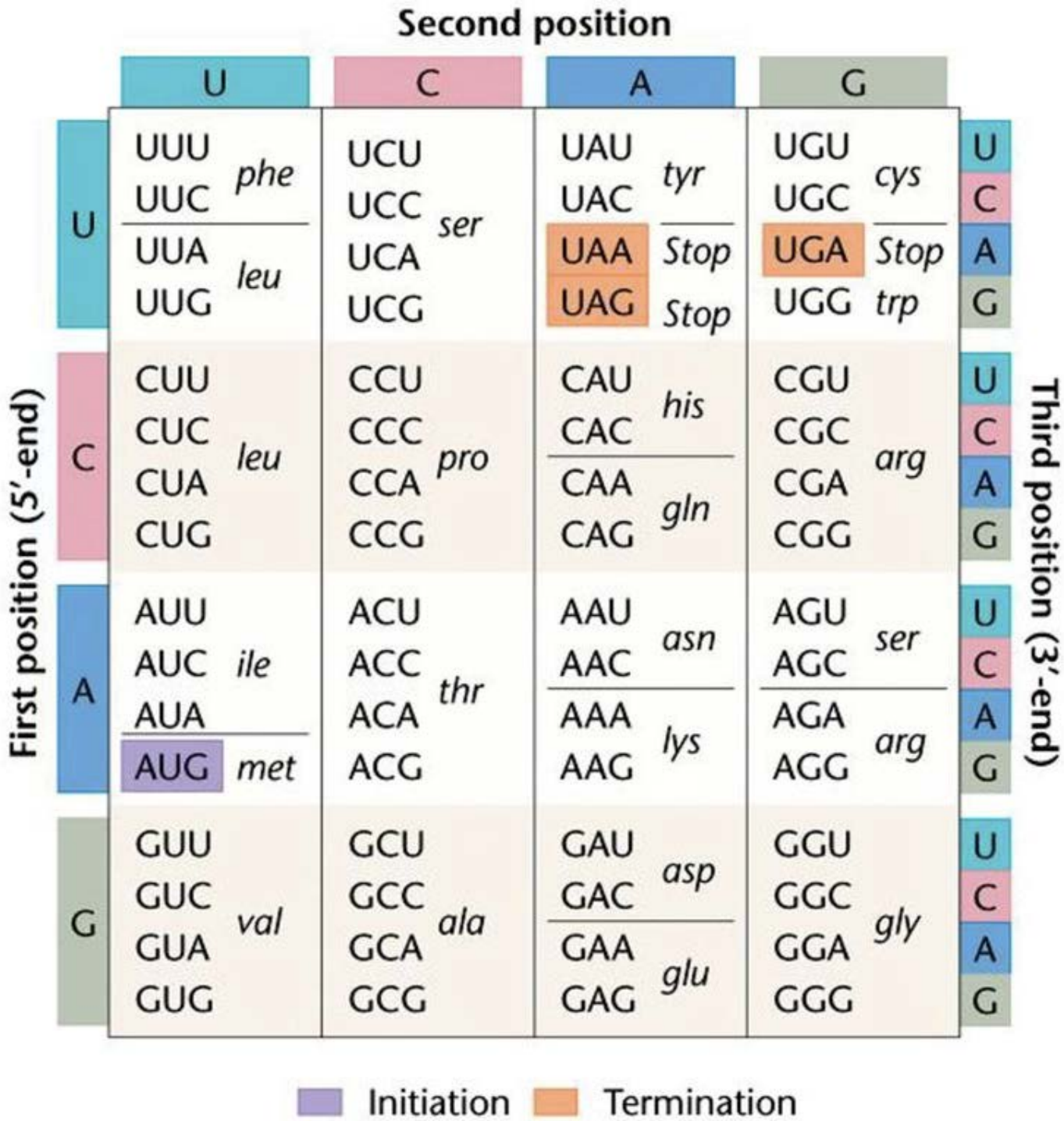
sarah321@mail.ncku.edu.tw

Central dogma

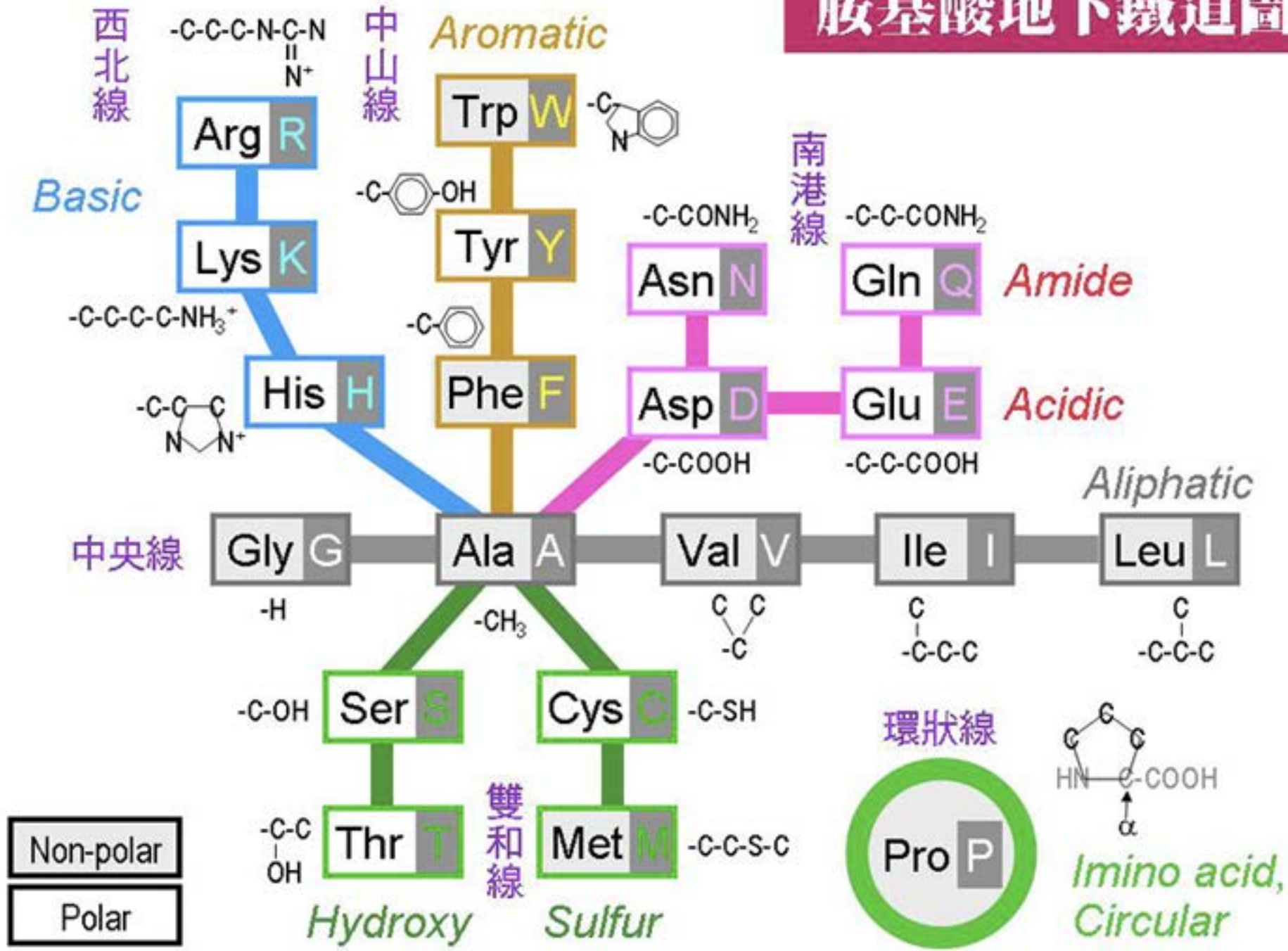
- The ‘Central Dogma’ is the process by which the instructions in DNA are converted into a functional product. It was first proposed in 1958 by Francis Crick, discoverer of the structure of DNA.
- The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid (Francis Crick, 1957).



(https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription
https://www.yourgenome.org/facts/what-is-the-central-dogma)

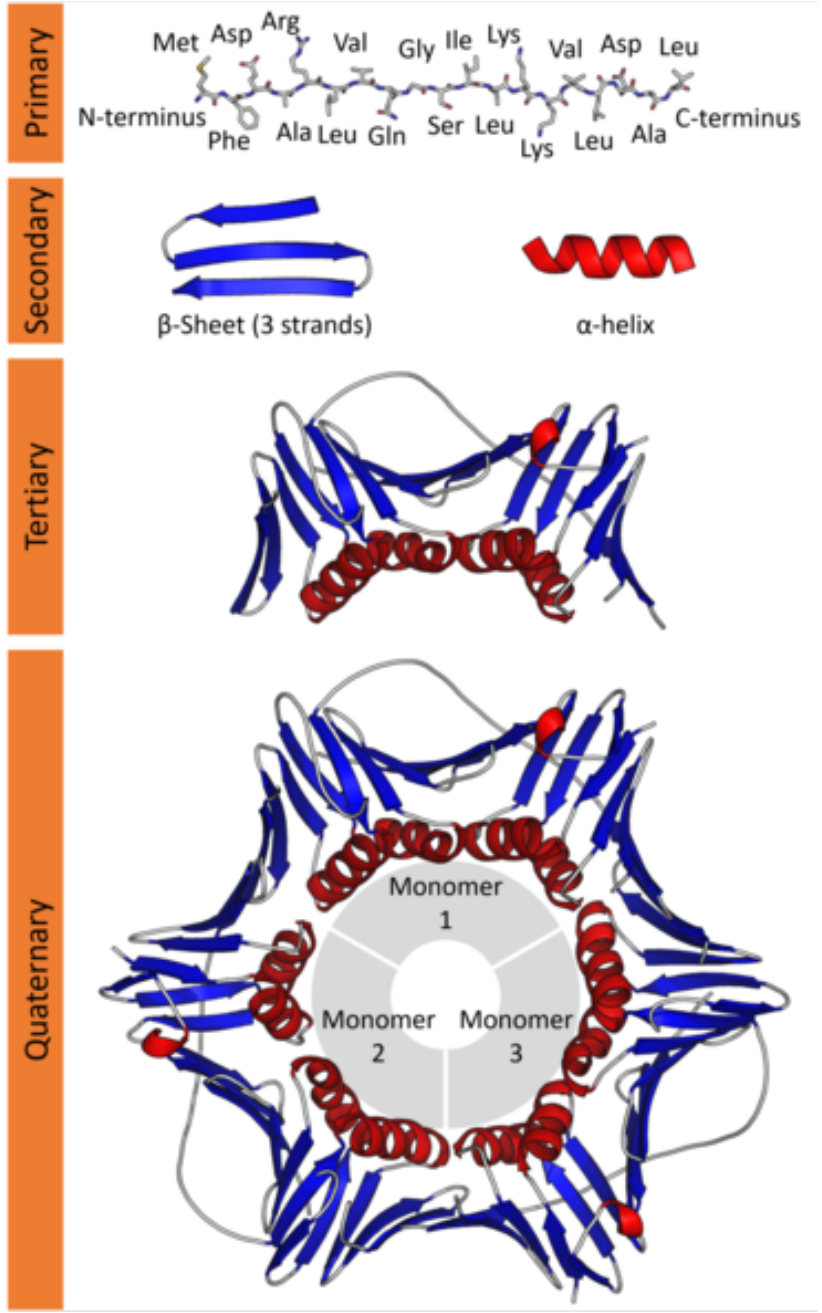

Second position
U
C
A
G
First position (5′-end)
Third position (3′-end)
U
C
A
G
UUU phe
UUC
UUA leu
UUG
UCU ser
UCC
UCA
UCG
UAU tyr
UAC
UAA Stop
UAG Stop
UGU cys
UGC
UGA Stop
UGG trp
CUU leu
CUC
CUA
CUG
CCU pro
CCC
CCA
CCG
CAU his
CAC
CAA gln
CAG
CGU arg
CGC
CGA
CGG
AUU ile
AUC
AUA
AUG met
ACU thr
ACC
ACA
ACG
AAU asn
AAC
AAA lys
AAG
AGU ser
AGC
AGA arg
AGG
GUU val
GUC
GUA
GUG
GCU ala
GCC
GCA
GCG
GAU asp
GAC
GAA glu
GAG
GGU gly
GGC
GGA
GGG
Initiation
Termination

胺基酸地下鐵道圖


西北線
中山線
南港線
中央線
雙和線
環狀線
Aromatic
Basic
Amide
Acidic
Aliphatic
Hydroxy
Sulfur
Imino acid, Circular
Arg R
Lys K
His H
Trp W
Tyr Y
Phe F
Asn N
Gln Q
Asp D
Glu E
Gly G
Ala A
Val V
Ile I
Leu L
Ser S
Thr T
Cys C
Met M
Pro P
-C-C-C-N-C-N
-C-OH
-C-CONH2
-C-C-CONH2
-C-C-C-C-NH3+
-C-COOH
-C-C-COOH
-H
-CH3
-C-SH
-C-C-S-C
HN
C-COOH
α
Non-polar
Polar

Protein structure

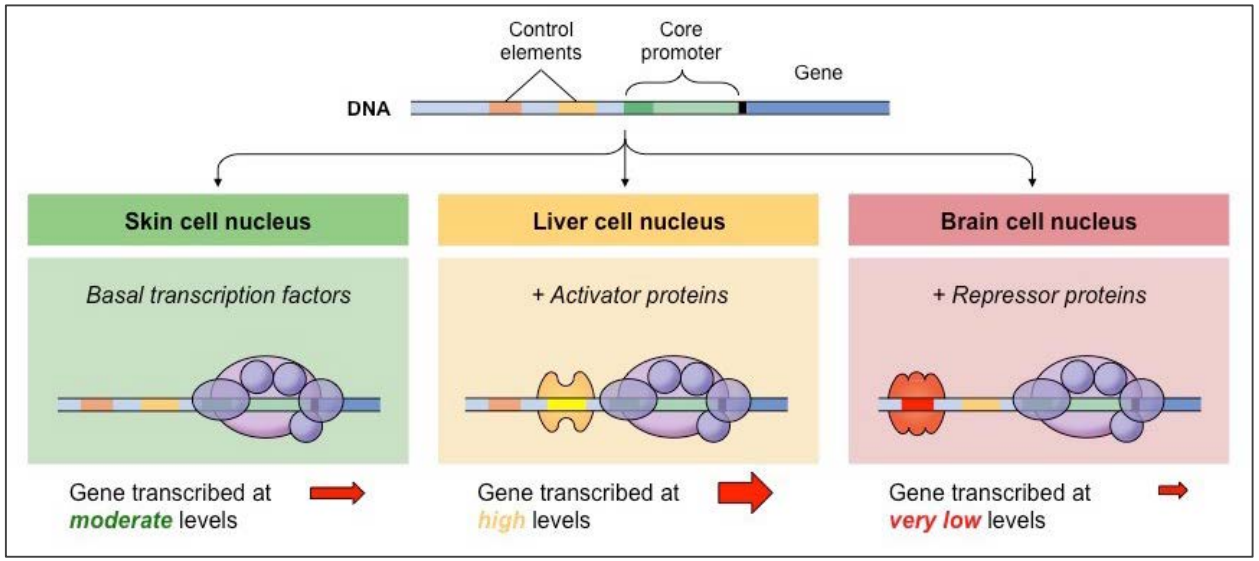
- Protein structure is crucial to protein function.
- The proper structure and function are required for the protein folding from a polypeptide chain to quaternary structure.
- Primary structure:
  the sequence of amino acids in the polypeptide chain
- Secondary structure:
  local folded structures that form within a polypeptide due to interactions between atoms of the backbone
- Tertiary structure:
  three-dimensional structure of monomeric and multimeric protein molecules
- Quaternary structure:
  the aggregation of two or more individual polypeptide chains that operate as a single functional unit

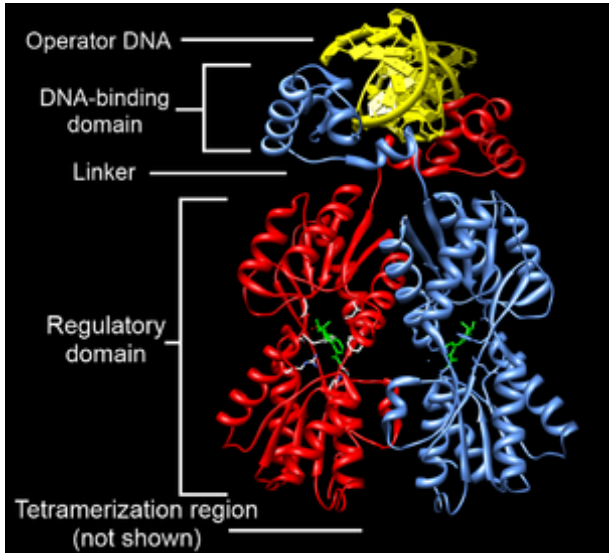


(https://en.wikipedia.org/wiki/Protein_structure
https://www.ncbi.nlm.nih.gov/books/NBK9843/
https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure)

Transcription Factors（轉錄因子）

- Transcription factors (TFs) are proteins that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.



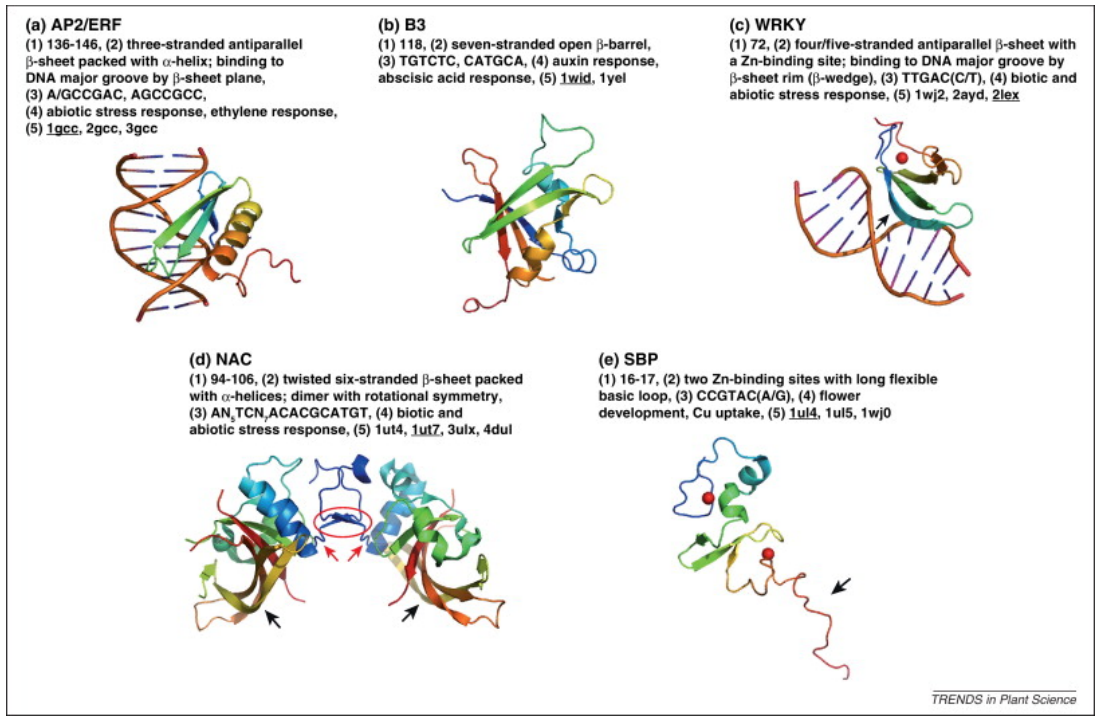


- Transcription factors contain two domains:
  - DNA-binding domain (DBD): attaches to specific sequences of DNA
  - Activation domain (AD):

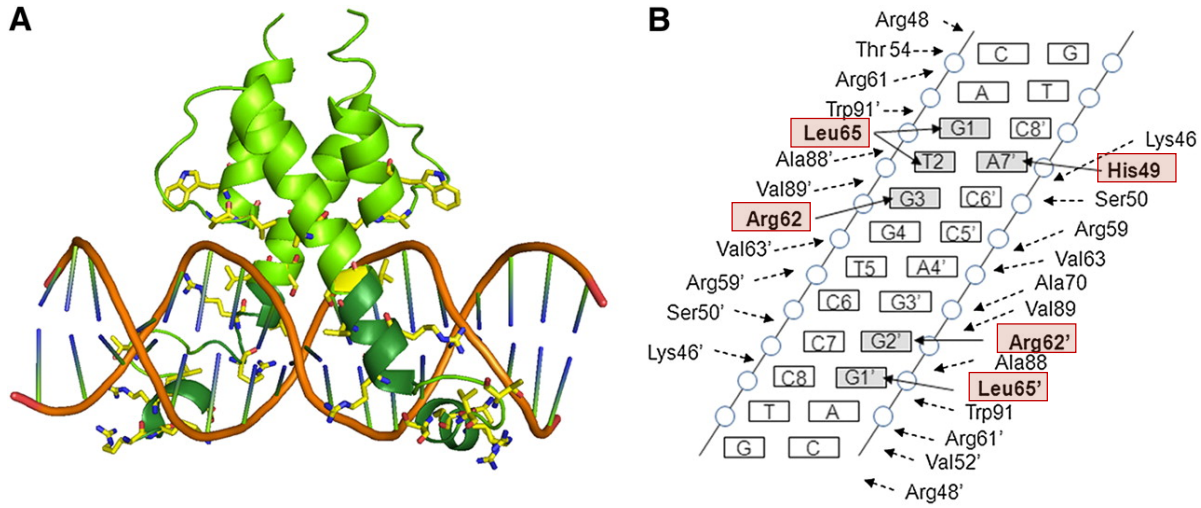
(https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/72-transcription-and-gene/gene-expression.html
https://en.wikipedia.org/wiki/Transcription_factor)

DNA-binding domain (DBD)

- TFs are classified into families according to their DBDs.
- Proteins with similar DBD sequences tend to bind very similar DNA sequences.
- Nuclear magnetic resonance spectroscopy and X-ray crystallography are applied to determine 3D structures of the DBDs-DNA Complex.



Structures of the DBDs of plant-specific TFs (P. Aggarwal, *et al.,* 2010)



TCP transcription factors (P. Aggarwal, *et al.,* 2010)

Motivations and problems

- Only a handful of TFs have been studied their DNA binding patterns.
  - only < 2% of eukaryotic TFs (M. T. Weirauch, *et. al,* 2014)
  - About 57% of *Arabidopsis thaliana* TFs (Data from PlantPAN 3.0 and PlantTFDB v4.0)
- Experimental protein structure determination is hard.

- Whether we can predict DNA binding sites by using polypeptide sequences or DNA-binding domain?
- Whether we can determine the key amino acids essential for DNA recognition from known TF-DNA pairs?
- Are there unknown features in polypeptide sequences which can be used to illustrate the interaction between TF and DNA?

TF families in Plants

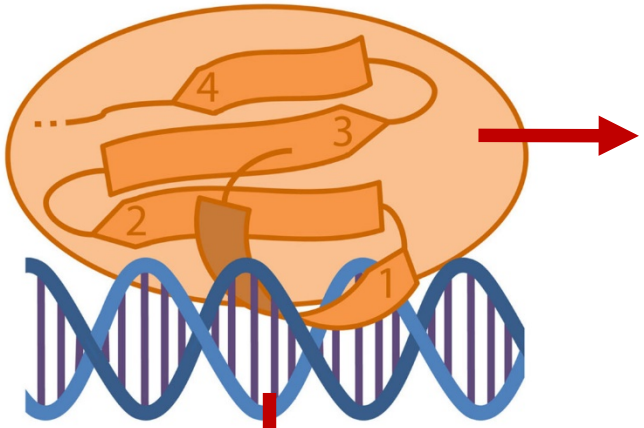
Alpha-amylase (4)	AP2 (2304)	ARF (78)	ARID (5)	ARR-B (19)
AT-Hook (121)	Aux/IAA (1)	B3 (408)	BBR-BPC (15)	BES1 (22)
Bet_v_1 (2)	bHLH (1497)	bZIP (1048)	C2C2COlike (2)	C2H2 (552)
C3H (6)	C3H Zinc finger (140)	CAMTA (14)	CG-1 (44)	CO-like (31)
CPP (19)	CSD (9)	Cupin_1 (1)	DBB (20)	Dehydrin (1)
Dof (101)	E2F (42)	E2F/DP (27)	EIL (15)	EIN3 (107)
ERF (330)	FAR1 (37)	G2-like (83)	GATA (483)	GeBP (23)
GRAS (139)	GRF (34)	HB-other (26)	HB-PHD (6)	HD-ZIP (252)
Homeodomain (1055)	HSF (74)	LBD (75)	LEA type 1 (1)	LEA_5 (1)
LFY (5)	LIM (2)	LOB (138)	LSD (14)	Lyase_aromatic (2)
M-type (46)	MADF (51)	MADS box (714)	MIKC (101)	mTERF (1)
MYB (235)	MYB-related (170)	Myb/SANT (1242)	NAC (974)	NAM (808)
NF-X1 (6)	NF-YA (25)	NF-YB (32)	NF-YC (24)	Nin-like (30)
PLATZ (1)	PsaH (1)	RAV (9)	Ribosomal protein L21P (1)	S1Fa-like (7)
SAP (1)	SBP (384)	Sox (24)	SRS (17)	STAT (1)
Storekeeper (13)	TALE (60)	TBP (58)	TCP (497)	TCR (55)
tify (1)	Trihelix (59)	trp (4)	Tryp_alpha_amyl (1)	VOZ (8)
Whirly (5)	WOX (39)	WRC (1)	WRKY (1733)	YABBY (25)
ZF-HD (48)	(Others)[1] (55)			

1: (Others):TFs without family information.

TF Browse by Families in PlantPAN 3.0 (http://plantpan.itps.ncku.edu.tw/TFsearch.php)

A case study

WRKY family (as a case study)



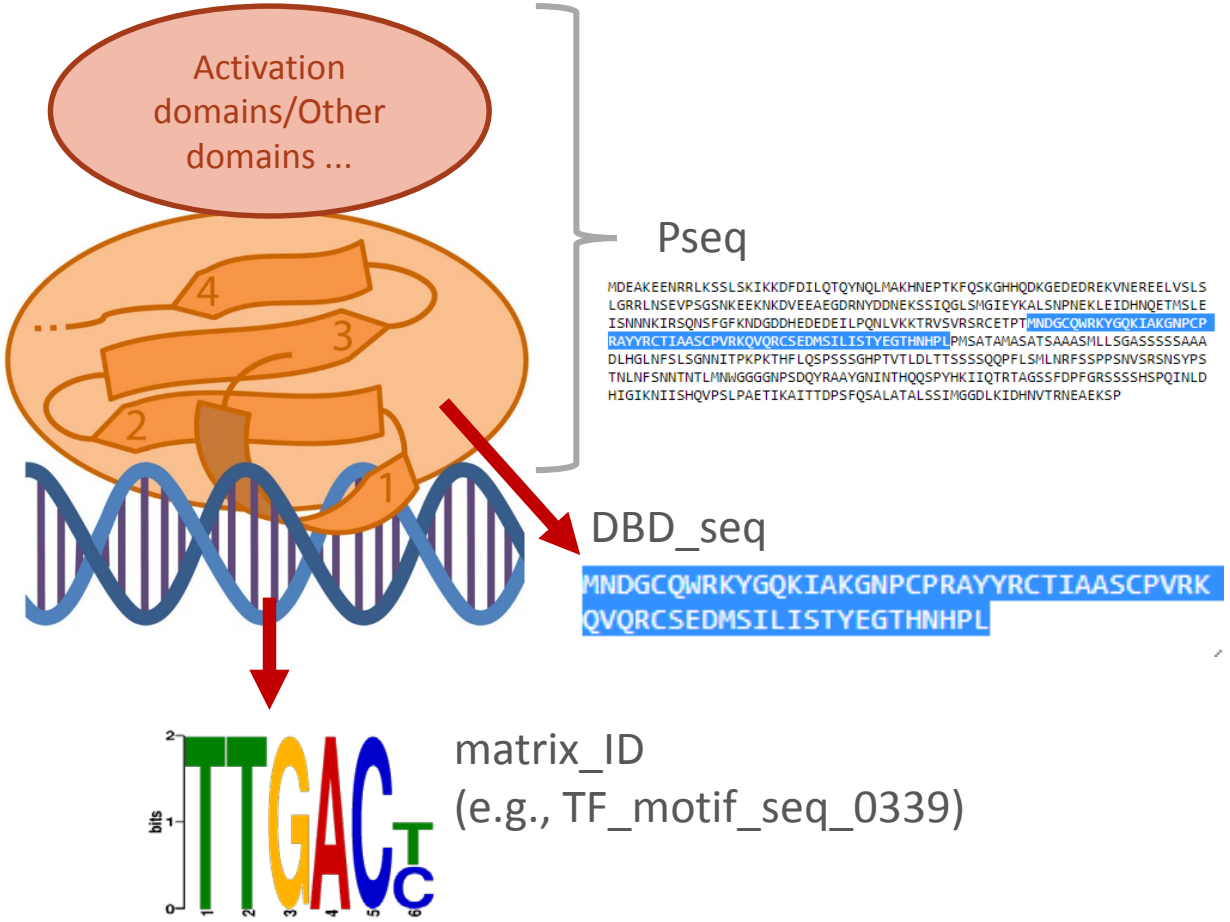
DNA Binding Domain

Binding DNA
(Sequence logo)

Structures of the WRKY DBD-DNA complex (C. M. Llorca, *et al.,* 2014)

- The WRKY domain binds to a so called W-box 5′-TTGAC(C/T)-3′ in the promoters of target genes.
- The binding DNA is highly conserved in WRKY family.
- The DNA binding domain of WRKY family is called the WRKY domain.
  - Almost invariant amino acid sequence at the N-terminus
  - About 60 residues in length
  - Four-stranded antiparallel β-sheet

Positive set

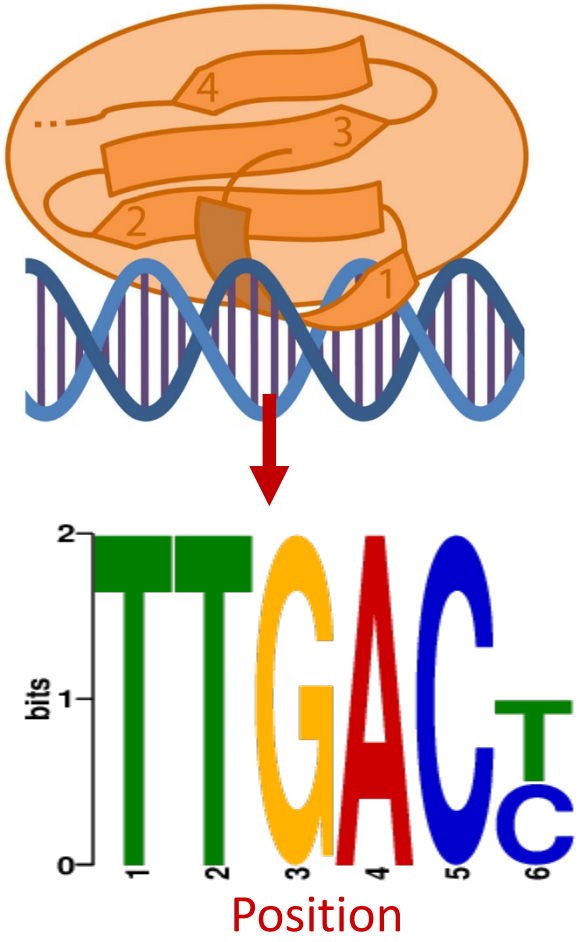


- File Name: WRKY_info_table_positive
- Column Name:
  - TF identifier (TF_ID)
  - Protein sequence identifier (Pseq_ID)
  - Protein primary sequence (Pseq)
  - DNA- binding domain sequence (DBD_seq)
  - Binding matrix identifier (matrix_ID)

Note: Please don't use matrix_ID as a feature.

```
TF_ID   Pseq_ID Pseq    DBD_seq matrix_ID
AT1G13960       TFprotseq_12499 MSEKEEAPSTSKSTGAPSRPTLSLPPRPFSEMFFNGGVGFSPGPMTLVSNMFPD
SDEFRSFSQLLAGAMSSPATAAAAAAAATASDYQRLGEGTNSSSGDVDPRFKQNRPTGLMISQSQSPSMFTVPPGLSPAMLLDSPS
FLGLFSPVQGSYGMTHQQALAQVTAQAVQANANMQPQTEYPPPSQVQSFSSGQAQIPTSAPLPAQRETSDVTIIEHRSQQPLNVDK
PADDGYNWRKYGQKQVKGSEFPRSYYKCTNPGCPVKKKVERSLDGQVTEIIYKGQHNHEPPQNTKRGNKDNTANINGSSINNNRGS
SELGASQFQTNSSNKTKREQHEAVSQATTTEHLSEASDGEEVGNGETDVREKDENEPDPKRRSTEVRISEPAPAASHRTVTEPRII
VQTTSEVDLLDDGYRWRKYGQKVVKGNPYPRSYYKCTTPGCGVRKHVERAATDPKAVVTTYEGKHNHDLPAAKSSSHAAAAAQLRP
DNRPGGLANLNQQQQQQPVARLRLKEEQTT      ADDGYNWRKYGQKQVKGSEFPRSYYKCTNPGCPVKKKVERSLDGQVTEII
YKGQHNHEP     TFmatrixID_0663
AT1G13960       TFprotseq_12499 MSEKEEAPSTSKSTGAPSRPTLSLPPRPFSEMFFNGGVGFSPGPMTLVSNMFPD
SDEFRSFSQLLAGAMSSPATAAAAAAAATASDYQRLGEGTNSSSGDVDPRFKQNRPTGLMISQSQSPSMFTVPPGLSPAMLLDSPS
FLGLFSPVQGSYGMTHQQALAQVTAQAVQANANMQPQTEYPPPSQVQSFSSGQAQIPTSAPLPAQRETSDVTIIEHRSQQPLNVDK
PADDGYNWRKYGQKQVKGSEFPRSYYKCTNPGCPVKKKVERSLDGQVTEIIYKGQHNHEPPQNTKRGNKDNTANINGSSINNNRGS
SELGASQFQTNSSNKTKREQHEAVSQATTTEHLSEASDGEEVGNGETDVREKDENEPDPKRRSTEVRISEPAPAASHRTVTEPRII
VQTTSEVDLLDDGYRWRKYGQKVVKGNPYPRSYYKCTTPGCGVRKHVERAATDPKAVVTTYEGKHNHDLPAAKSSSHAAAAAQLRP
DNRPGGLANLNQQQQQQPVARLRLKEEQTT      ADDGYNWRKYGQKQVKGSEFPRSYYKCTNPGCPVKKKVERSLDGQVTEII
YKGQHNHEP     TFmatrixID_0359
```

Matrices file



- File Name: All_matrices.meme
- Matrices format
  - position-specific weight matrix
  - MEME format (http://meme-suite.org/doc/meme-format.html#min_motif_name)

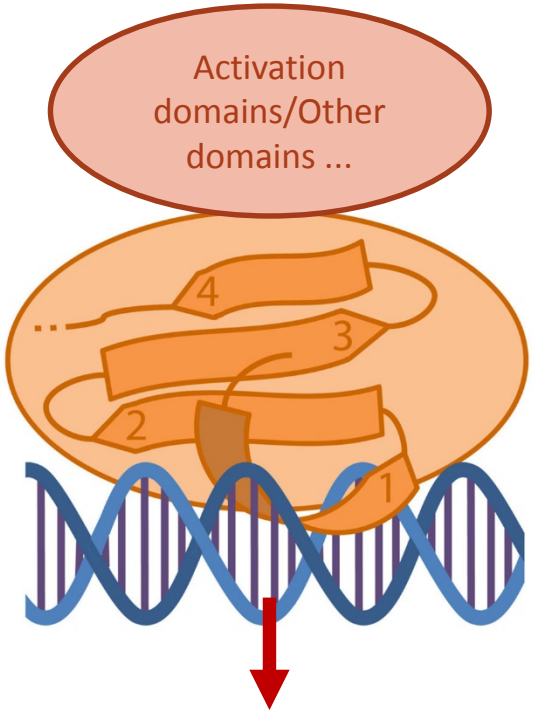
matrix_ID

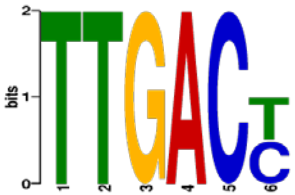
```
MOTIF TF_motif_seq_0339

letter-probability matrix: alength= 4 w= 6 nsites= 1 E= 0
  0.000000        0.000000        0.000000        1.000000
  0.000000        0.000000        0.000000        1.000000
  0.000000        0.000000        1.000000        0.000000
  1.000000        0.000000        0.000000        0.000000
  0.000000        1.000000        0.000000        0.000000
  0.000000        0.500000        0.000000        0.500000
```

Position: 1 2 3 4 5 6

A C G T

Negative set -- 1/3

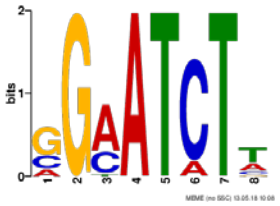


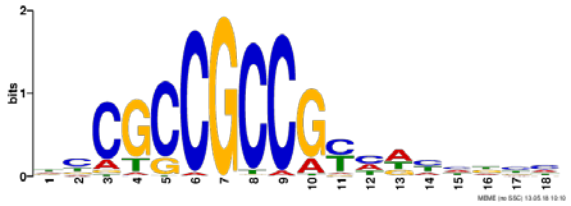
WRKY 

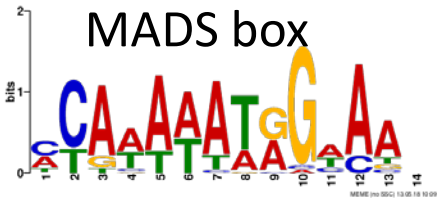
matrix_ID
(e.g., TF_motif_seq_0339)

Modification

- File Name: WRKY_info_table_negative_one
- Modification:
  - Using matrices from other TF families
  - Randomly selecting matrices belong to other families

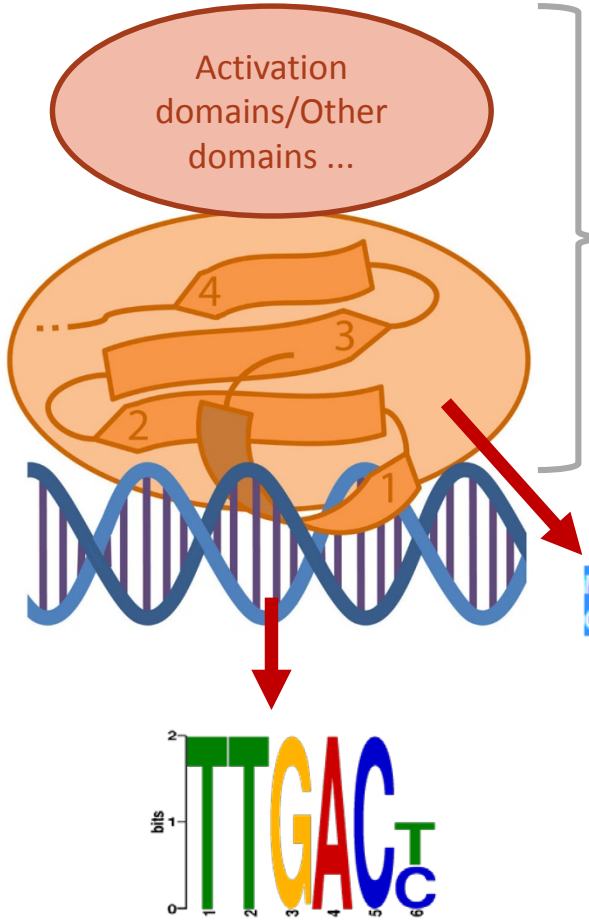
Myb/SANT 

AP2; ERF 

MADS box 

...

Negative set -- 2/3



Pseq

MDEAKEENRRLKSSLSKIKKDFDILQTQYNQLMAKHNEPTKFQSKGHHQDKGEDEDREKVNEREELVSLS
LGRRLNSEVPSGSNKEEKNKDVEEAEGDRNYDDNEKSSIQGLSMGIEYKALSNPNEKLEIDHNQETMSLE
ISNNNKIRSQNSFGFKNDGDDHEDEDEILPQNLVKKTRVSVRSRCETPTMNDGCQWRKYGQKIAKGNPCP
RAYYRCTIAASCPVRKQVQRCSEDMSILISTYEGTHNHPLPMSATAMASATSAAASMLLSGASSSSSAAA
DLHGLNFSLSGNNITPKPKTHFLQSPSSSGHPTVTLDLTTSSSSQQPFLSMLNRFSSPPSNVSRSNSYPS
TNLNFSNNTNTLMNWGGGGNPSDQYRAAYGNINTHQQSPYHKIIQTRTAGSSFDPFGRSSSSHSPQINLD
HIGIKNIISHQVPSLPAETIKAITTDPSFQSALATALSSIMGGDLKIDHNVTRNEAEKSP

DBD_seq

MNDGCQWRKYGQKIAKGNPCPRAYYRCTIAASCPVRK
QVQRCSEDMSILISTYEGTHNHPL

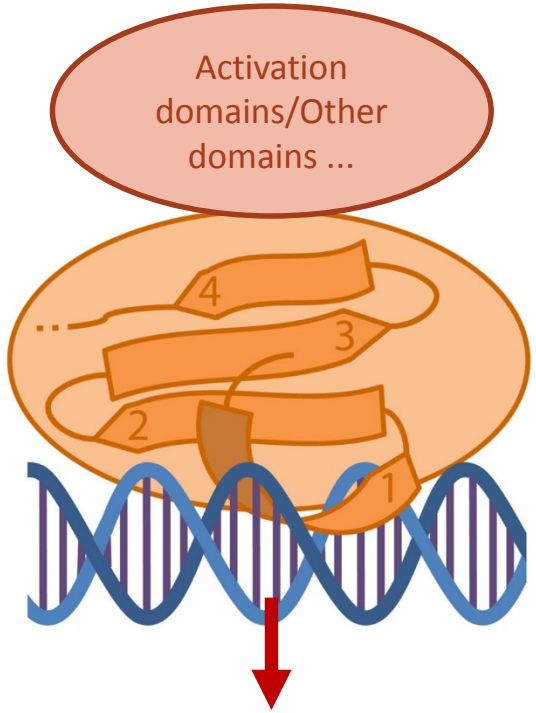
Modification

MDEAKEENRRLKSSLSKIKKDFDILQTQYNQLMAKHNEPTKFQSKGHHQDKGEDEDREKVNEREELVSLS
LGRRLNSEVPSGSNKEEKNKDVEEAEGDRNYDDNEKSSIQGLSMGIEYKALSNPNEKLEIDHNQETMSLE
ISNNNKIRSQNSFGFKNDGDDHEDEDEILPQNLVKKTRVSVRSRCETPTMNDGCQWRKYGQKIAKGNPCP
RAYYRCTIAASCPVRKQVQRCSEDMSILISTYEGTHNHPLPMSATAMASATSAAASMLLSGASSSSSAAA
DLHGLNFSLSGNNITPKPKTHFLQSPSSSGHPTVTLDLTTSSSSQQPFLSMLNRFSSPPSNVSRSNSYPS
TNLNFSNNTNTLMNWGGGGNPSDQYRAAYGNINTHQQSPYHKIIQTRTAGSSFDPFGRSSSSHSPQINLD
HIGIKNIISHQVPSLPAETIKAITTDPSFQSALATALSSIMGGDLKIDHNVTRNEAEKSP

- File Name: WRKY_info_table_negative_two
- Modification:
  - Using Non DNA-binding domain sequence
  - Randomly selecting the region outside of DNA-binding domain
  - The length is equal to DNA-binding domains'

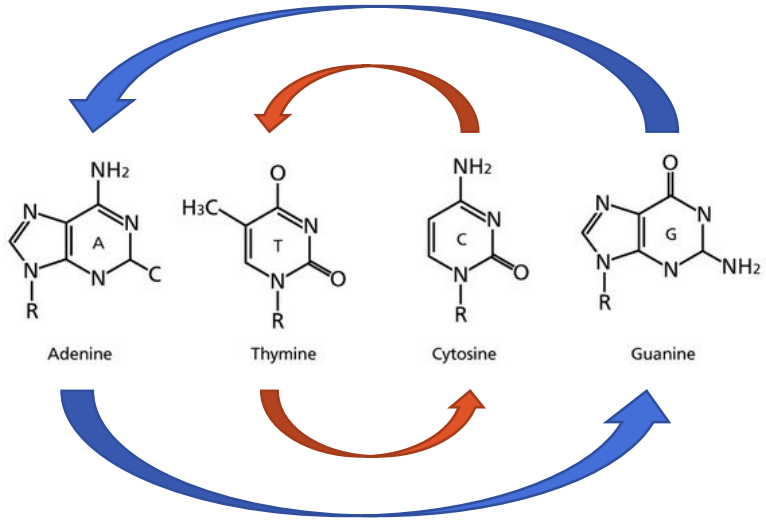
Negative set -- 3/3

- File Name: WRKY_info_table_negative_three
- Modification:
  - Change the core sequence of matrices
  - Weight (possibility) ≥ 0.8
  - Change the nucleotide:
    - A → G
    - G → A
    - T → C
    - C → T



matrix_ID
(e.g., TF_motif_seq_0339)

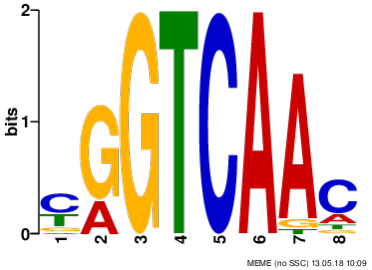
Modification



Problems You may face!

- Different length for each protein/DNA-binding domain sequence
- Different length for each binding matrix

TFmatrixID_0449:
Length: 8 bp
Core regions: 3rd - 7th



TFmatrixID_0534:
Length: 11 bp
Core regions: 5th - 11th