Sequence Comparison MSA (Part 2)

patterns/motifs

曹志伟

Recap

- **Biological Databases**
- **Sequence analysis**
- **Sequence comparisons**
  - local and global alignment
  - Substitution matrices
- **BLAST and its uses**
  - Query sequence database;
  - How different BLAST variants allow searching different types of sequences
  - Why apparent significance from BLAST depends on the size of the database
- MSA:
  - building up a family of related sequences from a set of pairwise alignments

是否有consensus？

- Multiple Sequence Alignment

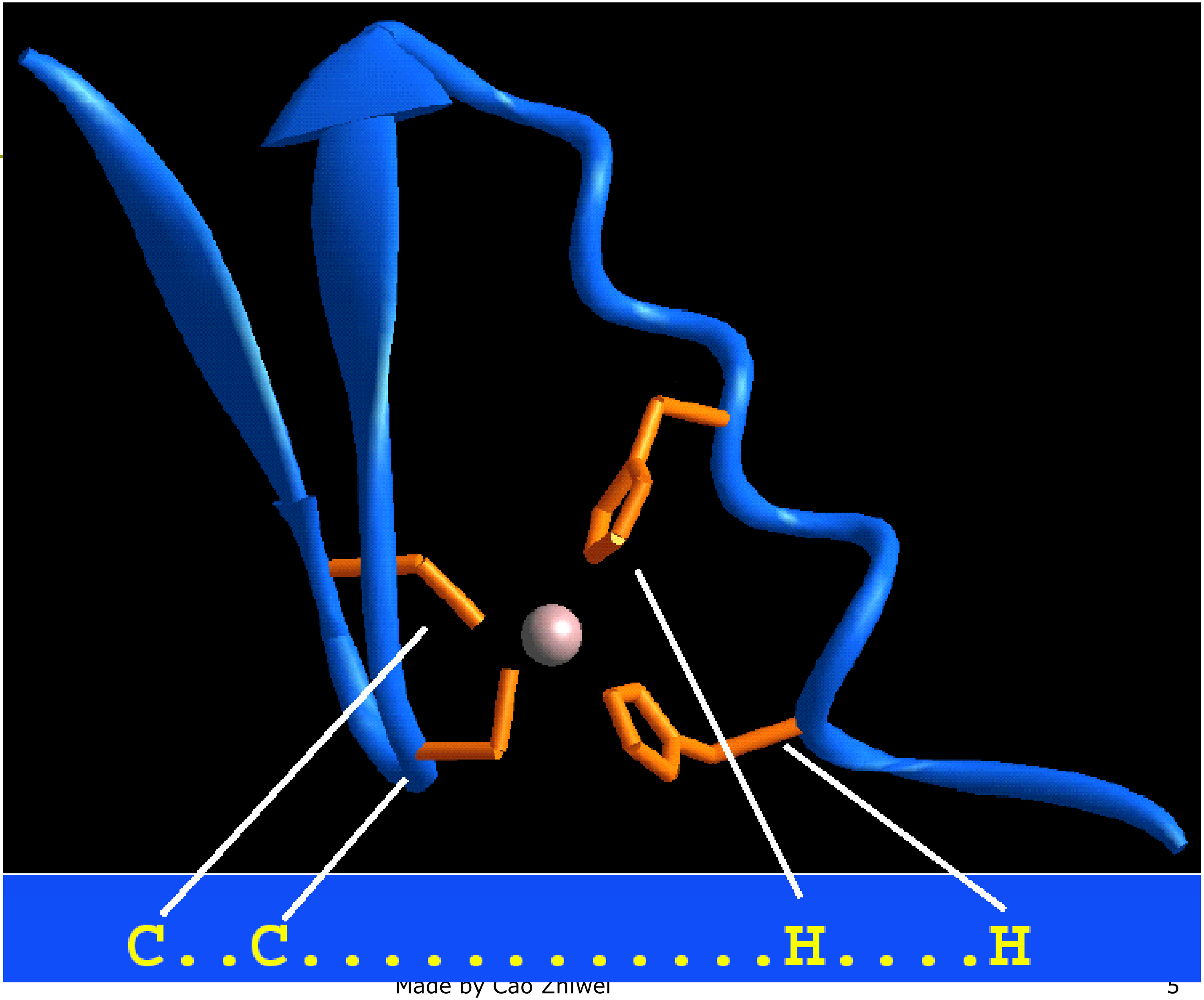
```
             ....|....| ....|....| ....|....| ....|....| ....|....|
                665        675        685        695        705
Sp1          ACTCPYCKDS EGRGSG---- DPGKKKQHIC HIQGCGKVYG KTSHLRAHLR
Sp2          ACTCPNCKDG EKRS------ GEQGKKKHVC HIPDCGKTFR KTSLLRAHVR
Sp3          ACTCPNCKEG GGRGTN---- -LGKKKQHIC HIPGCGKVYG KTSHLRAHLR
Sp4          ACSCPNCREG EGRGSN---- EPGKKKQHIC HIEGCGKVYG KTSHLRAHLR
DrosBtd      RCTCPNCTNE MSGLPPIVGP DERGRKQHIC HIPGCERLYG KASHLKTHLR
DrosSp       TCDCPNCQEA ERLGPAGV-- HLRKKNIHSC HIPGCGKVYG KTSHLKAHLR
CeT22C8.5    RCTCPNCKAI KHG------- DRGSQHTHLC SVPGCGKTYK KTSHLRAHLR
Y40B1A.4     PQISLKKKIF FFIFSNFR-- GDGKSRIHIC HL--CNKTYG KTSHLRAHLR
```

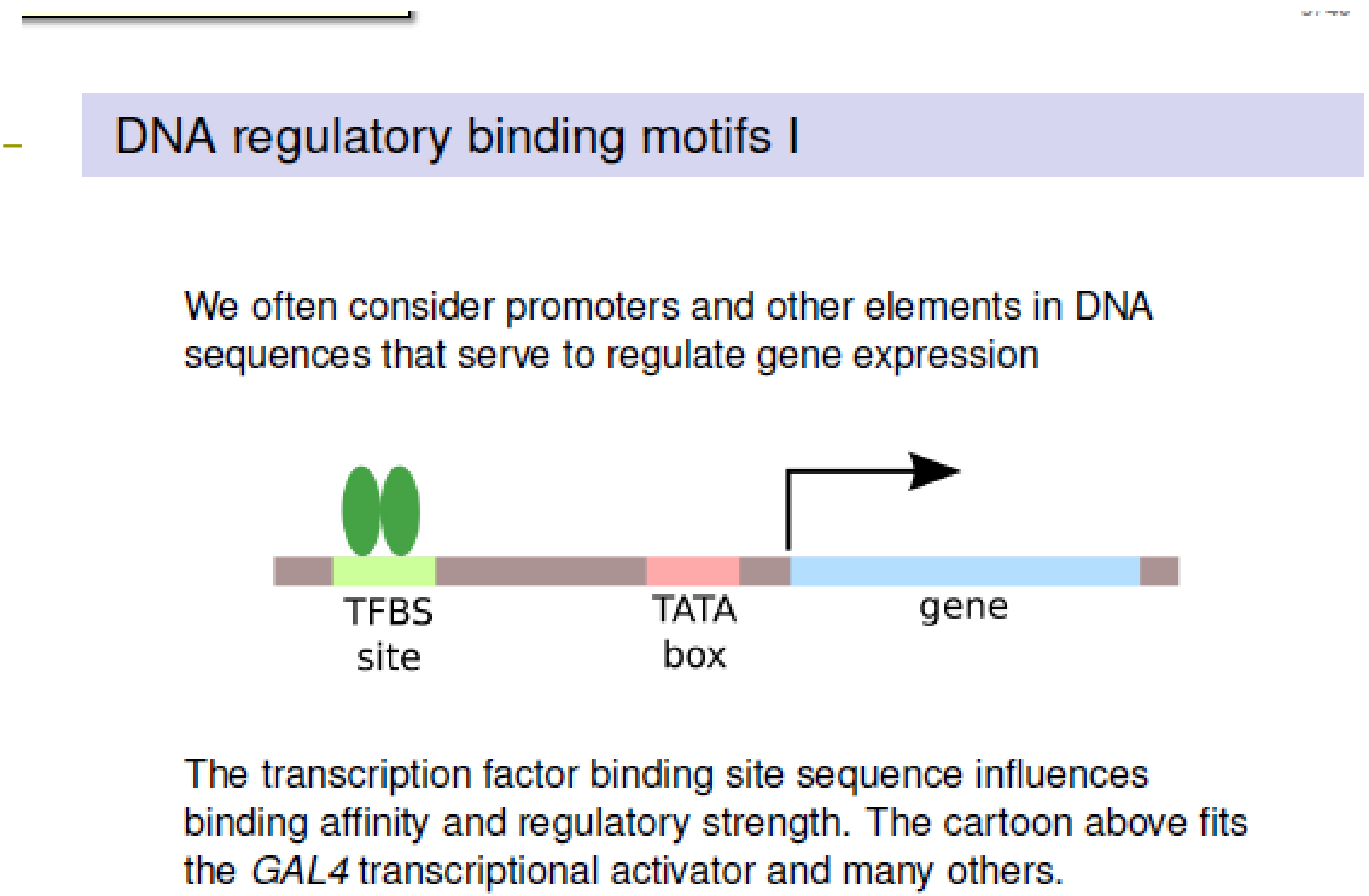
C_2H_2 Zinc finger motif

```
               ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
                  665        675        685        695        705        715
Sp1            ACTCPYCKDS EGRGSG---- DPGKKKQHIC HIQGCGKVYG KTSHLRAHLR WHTGERPFMC
Sp2            ACTCPNCKDG EKRS------ GEQGKKKHVC HIPDCGKTFR KTSLLRAHVR LHTGERPFVC
Sp3            ACTCPNCKEG GGRGTN---- -LGKKKQHIC HIPGCGKVYG KTSHLRAHLR WHSGERPFVC
Sp4            ACSCPNCREG EGRGSN---- EPGKKKQHIC HIEGCGKVYG KTSHLRAHLR WHTGERPFIC
DrosBtd        RCTCPNCTNE MSGLPPIVGP DERGRKQHIC HIPGCERLYG KASHLKTHLR WHTGERPFLC
DrosSp         TCDCPNCQEA ERLGPAGV-- HLRKKNIHSC HIPGCGKVYG KTSHLKAHLR WHTGERPFVC
CeT22C8.5      RCTCPNCKAI KHG------- DRGSQHTHLC SVPGCGKTYK KTSHLRAHLR KHTGDRPFVC
Y40B1A.4       PQISLKKKIF FFIFSNFR-- GDGKSRIHIC HL--CNKTYG KTSHLRAHLR GHAGNKPFAC
```

Prosite pattern

C-x(2,4)-C-x(12)-H-x(3)-H


C..C............H....H

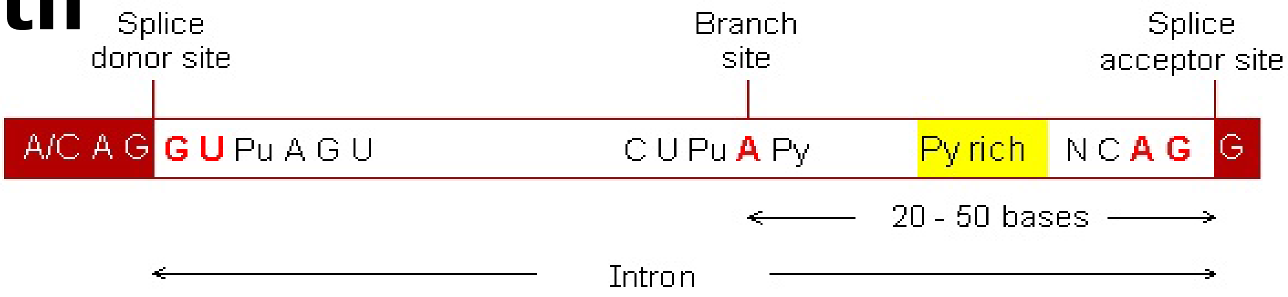


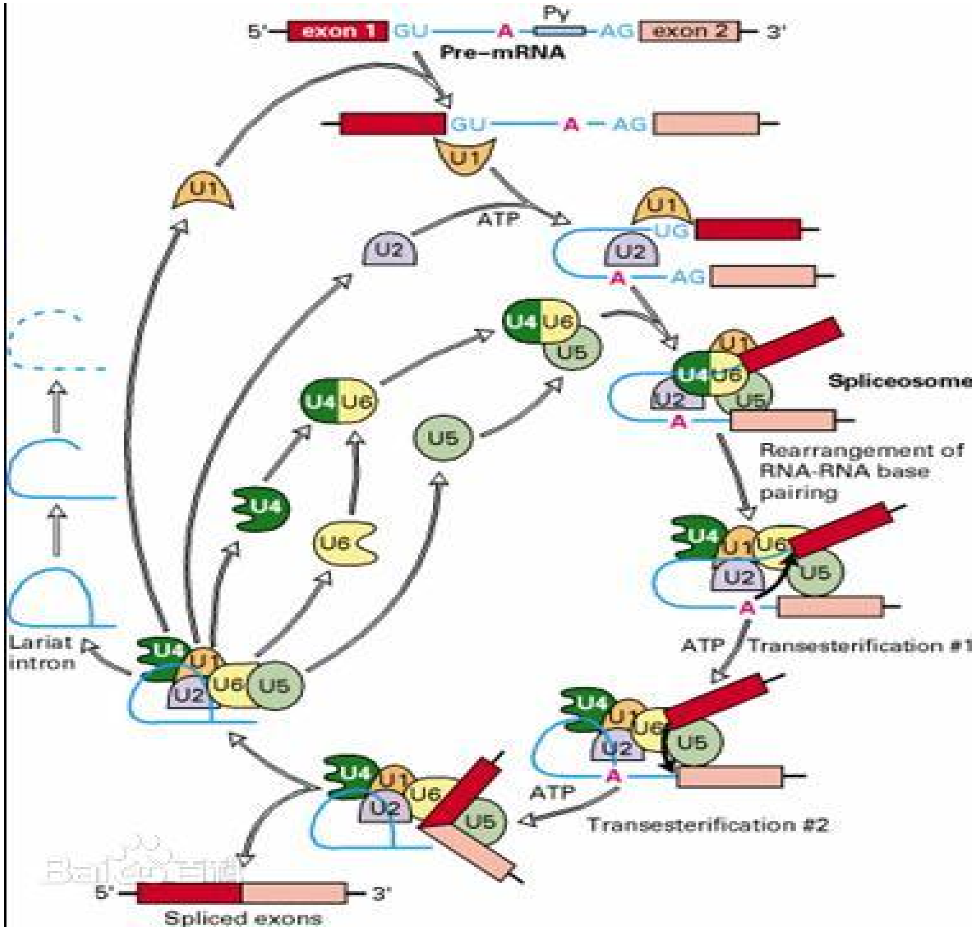
DNA regulatory binding motifs I

We often consider promoters and other elements in DNA sequences that serve to regulate gene expression

The transcription factor binding site sequence influences binding affinity and regulatory strength. The cartoon above fits the *GAL4* transcriptional activator and many others.

RNA motif

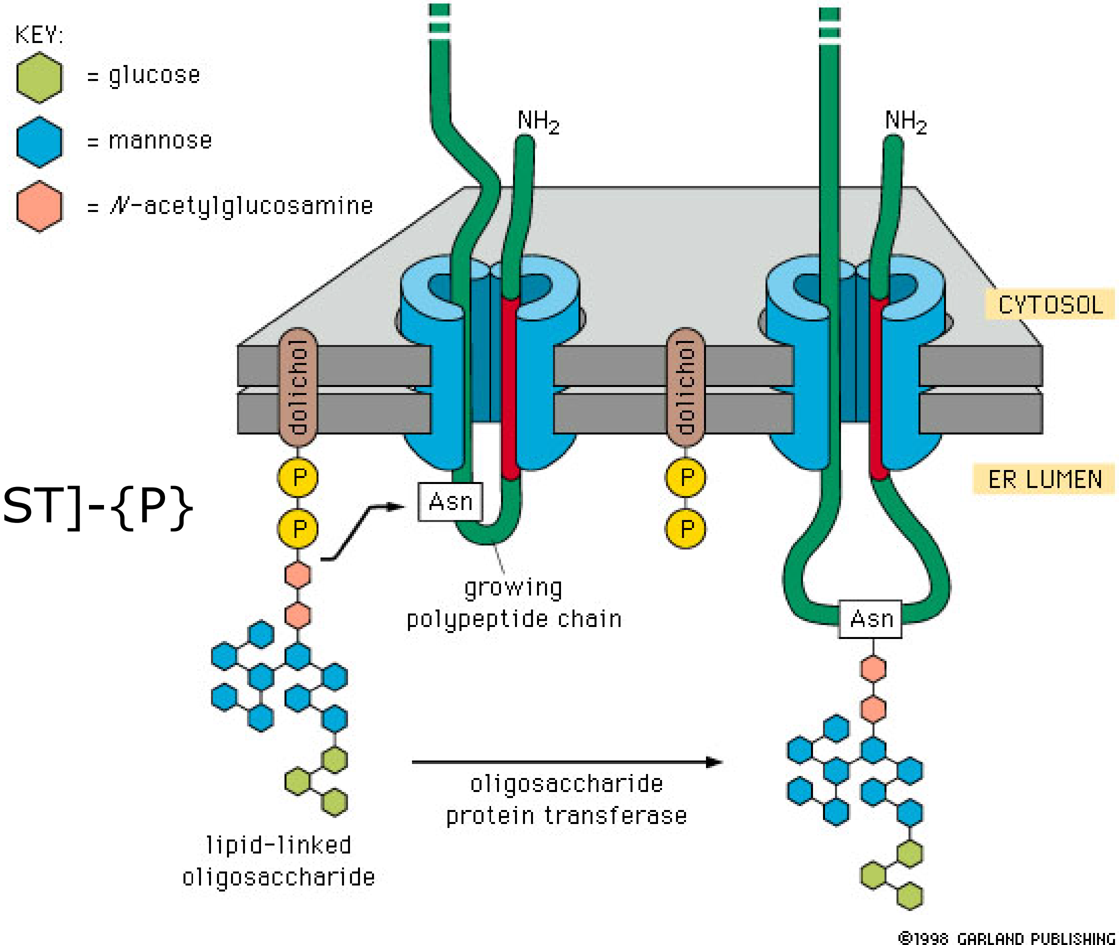

Splice donor site
Branch site
Splice acceptor site
A/C A G
G U Pu A G U
C U Pu A Py
Py rich
N C A G
G
20 - 50 bases
Intron


5'
exon 1
GU
A
Py
AG
exon 2
3'
Pre-mRNA
U1
U2
ATP
U4
U6
U5
Spliceosome
Rearrangement of RNA-RNA base pairing
ATP
Transesterification #1
ATP
Transesterification #2
Lariat intron
Spliced exons

Protein motif

Many eukaryotic proteins translated at the rough endoplasmic reticulum are subject to N-linked glycosylation

PROSITE pattern: N-{P}-[ST]-{P}



Some key terms

- Motif
  - Common elements shared by a group of sequences.
  - Indicative of functional or evolutionary relationship.

  - Eg. N-Glycosylation site, N-{P}-[ST]-{P}

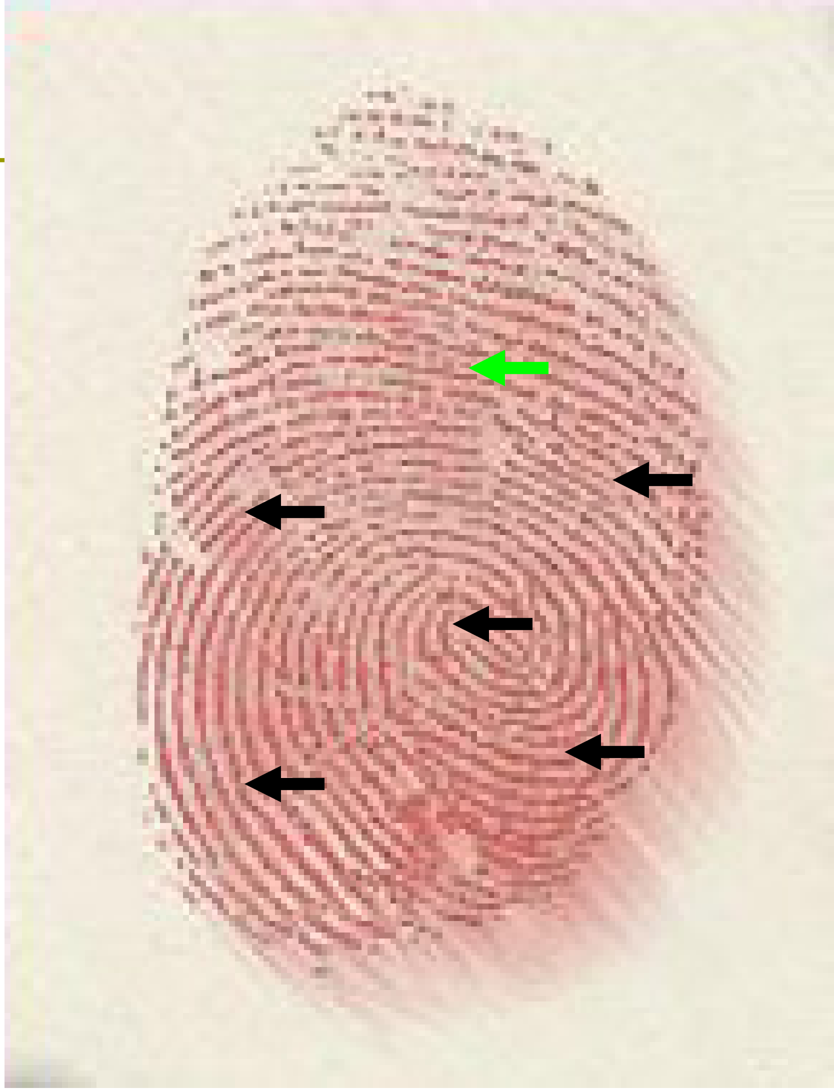
- Pattern
  - “A consistent, characteristic form, style, or method, as a composite of traits or features characteristic of an individual or a group.” (dictionary.com)
  - A physical expression of a motif.
  - Many forms of expression.

- Profile： A quantitative description of a pattern or motif

- using a position dependent scoring system

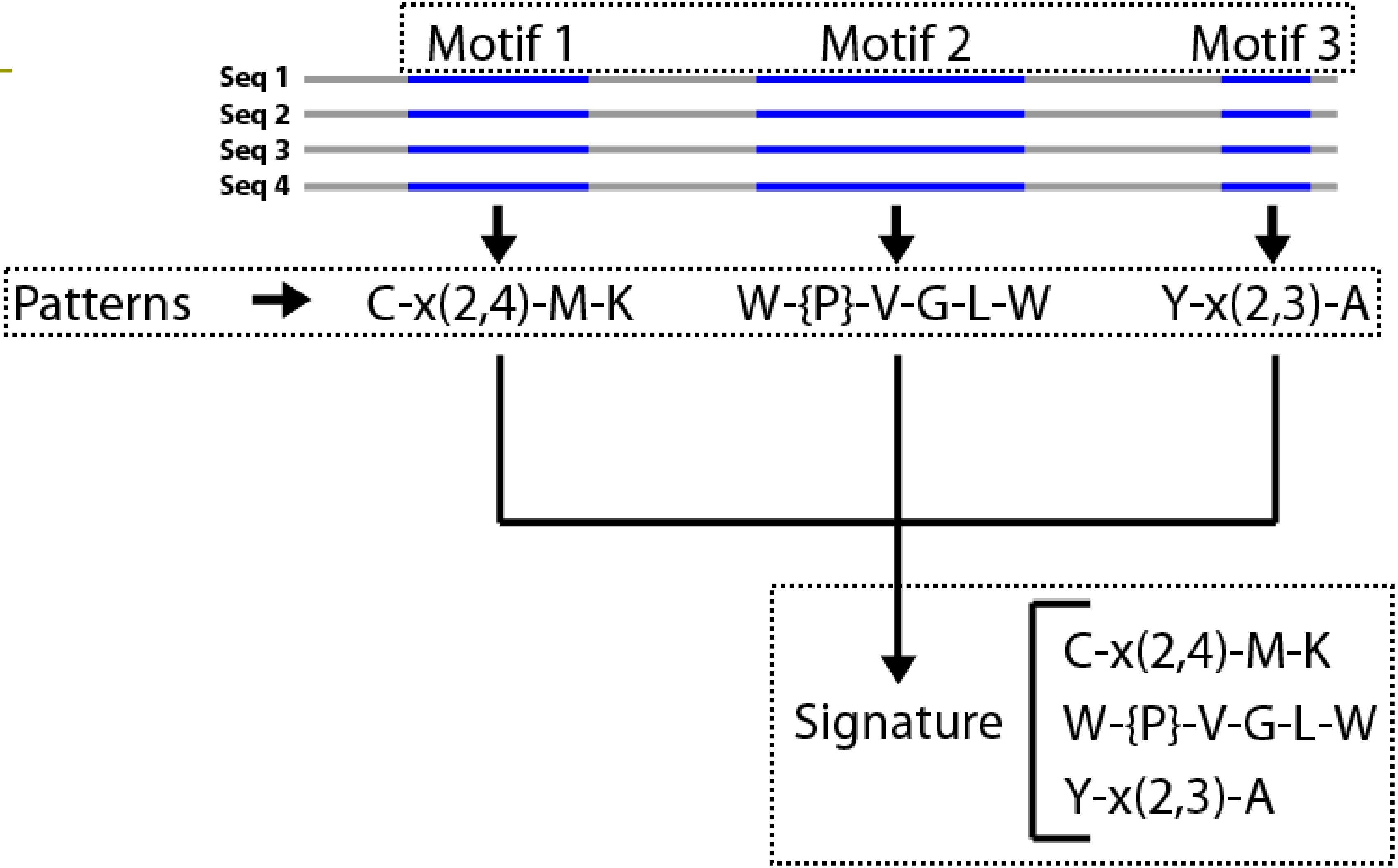
- patterns :formal, qualitative description of sequence motifs,
- profiles : quantitative models.

- Signature/Print
  - A set of patterns that defines a group of sequences having a certain common characteristic.
  - Bacterial Rhodopsin (2 patterns)
    - R-Y-x-[DT]-W-x-[LIVMF]-[ST]-T-P-[LIVM](3)
    - [FYIV]-x-[FYVG]-[LIVM]-D-[LIVMF]-x-[STA]-K-x(2)-[FY]



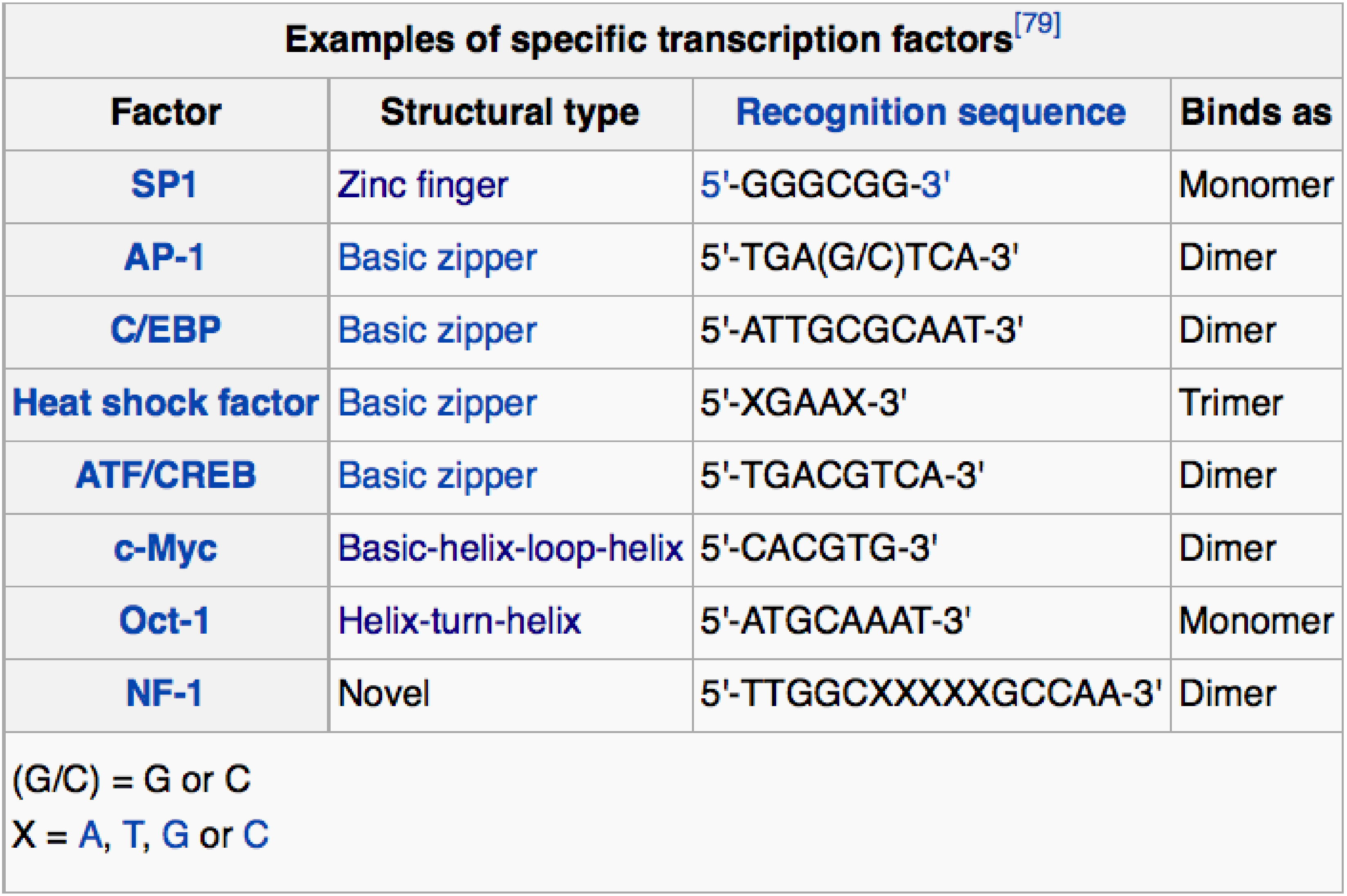
A single point is not indicative of identity.

But many points allow for identification.


Motif 1
Motif 2
Motif 3
Seq 1
Seq 2
Seq 3
Seq 4
Patterns
C-x(2,4)-M-K
W-{P}-V-G-L-W
Y-x(2,3)-A
Signature
C-x(2,4)-M-K
W-{P}-V-G-L-W
Y-x(2,3)-A

Types of Patterns

- DNA
  - Restriction Endonuclease sites
  - DNA binding motifs
  - Transcription Factor binding sites
  - Splicing site motifs
  - Other signals



Examples of specific transcription factors[79]

Factor	Structural type	Recognition sequence	Binds as
SP1	Zinc finger	5'-GGGCGG-3'	Monomer
AP-1	Basic zipper	5'-TGA(G/C)TCA-3'	Dimer
C/EBP	Basic zipper	5'-ATTGCGCAAT-3'	Dimer
Heat shock factor	Basic zipper	5'-XGAAX-3'	Trimer
ATF/CREB	Basic zipper	5'-TGACGTCA-3'	Dimer
c-Myc	Basic-helix-loop-helix	5'-CACGTG-3'	Dimer
Oct-1	Helix-turn-helix	5'-ATGCAAAT-3'	Monomer
NF-1	Novel	5'-TTGGCXXXXXGCCAA-3'	Dimer

(G/C) = G or C

X = A, T, G or C

- Protein
  - Sequence motifs
    - Zinc finger
    - SH2 domains
  - Structural patterns
- Structural motif != sequence motif
  - structural motifs do not need to share similar sequence in order to adopt similar structure

Motif…

- How should we describe them?
- How can we know whether or not a given sequence fits a given motif?
- How can we discover new motifs?
- How can we detect our new motif in sequence databases?
- How can we test a new sequence against known motifs?

Representations

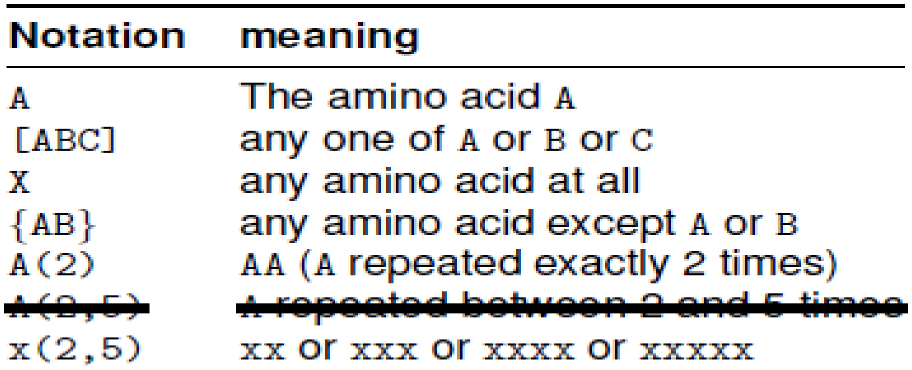
- Regular Expression (RE)
- Prosite Patterns
- Profiles (PSSM)
- Hidden Markov Models (HMM)

PROSITE

□PROSITE is an online resource describing protein domains,families and functional sites as well as **patterns and profiles**

□ PROSITE patterns are **regular expressions** to describe a set of sequences.

Pattern notation for PROSITE entries

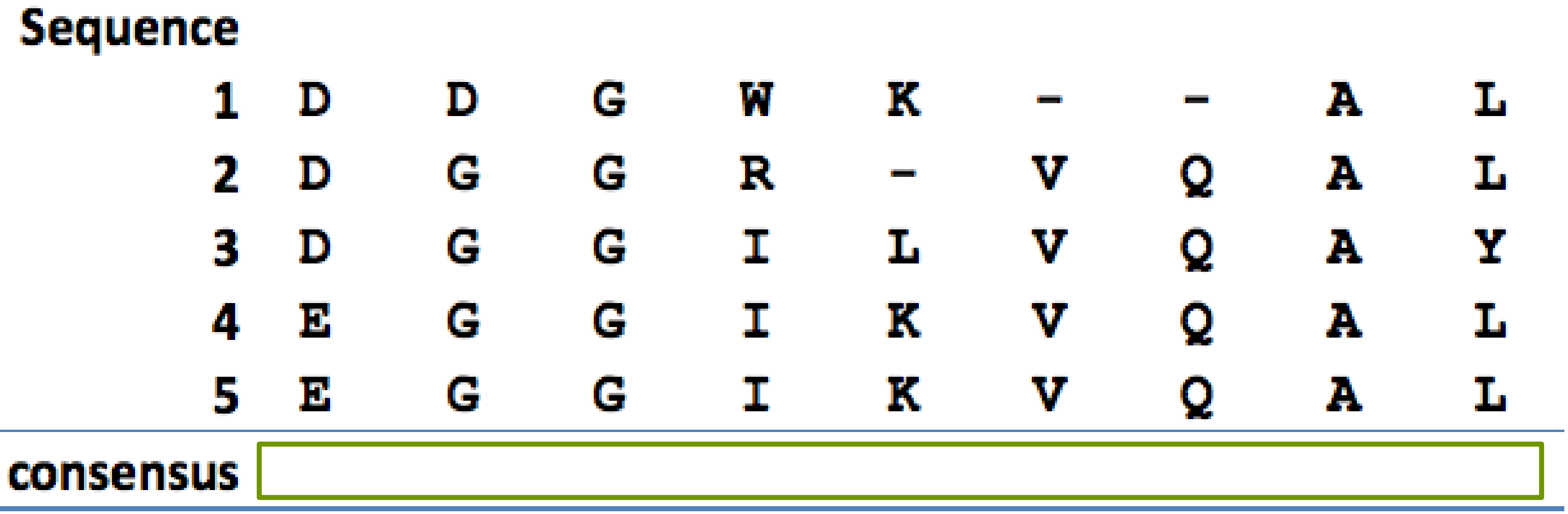


Notation	meaning
A	The amino acid A
[ABC]	any one of A or B or C
X	any amino acid at all
{AB}	any amino acid except A or B
A(2)	AA (A repeated exactly 2 times)
~~A(2,5)~~	~~A repeated between 2 and 5 times~~
x(2,5)	xx or xxx or xxxx or xxxxx

Additional comments

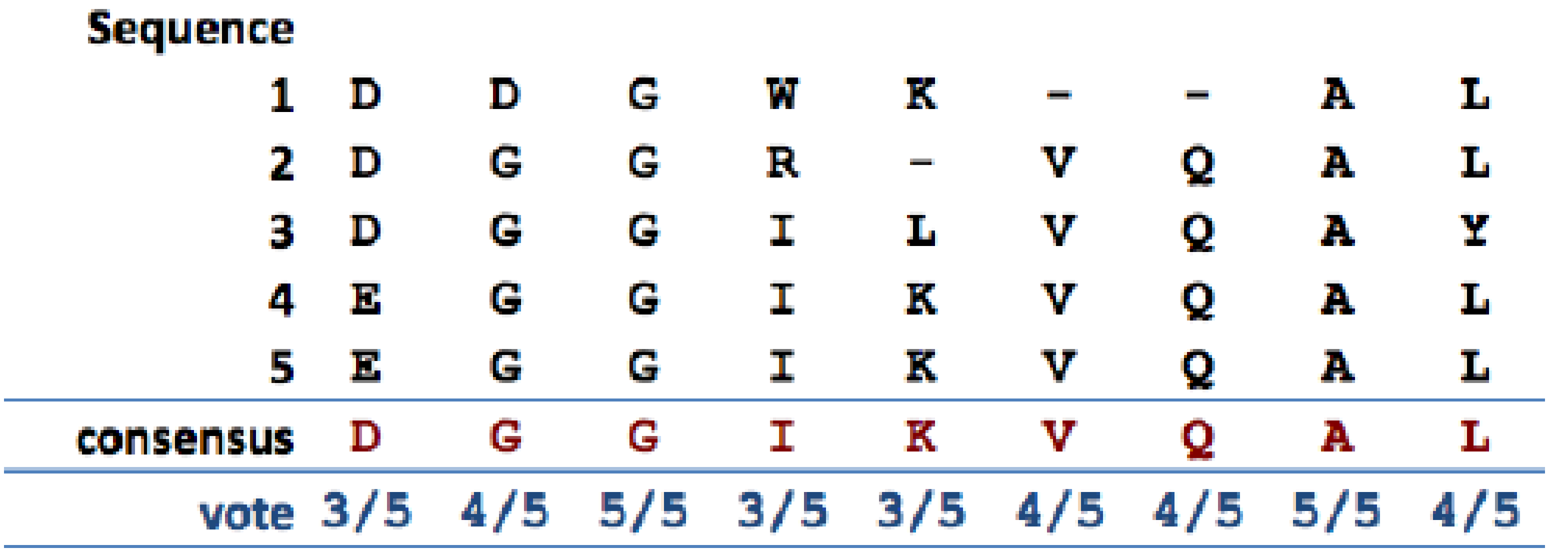
- The one-letter abbreviation for amino acids is used in all PROSITE patterns
- Adjacent amino acids in PROSITE notation are separated by a hyphen (-). This is how gaps are represented in an MSA.
- How would you represent gaps in an MSA using PROSITE patterns?

Consensus… try to write



Sequence									
1	D	D	G	W	K	-	-	A	L
2	D	G	G	R	-	V	Q	A	L
3	D	G	G	I	L	V	Q	A	Y
4	E	G	G	I	K	V	Q	A	L
5	E	G	G	I	K	V	Q	A	L
consensus									

Problems in deriving consensus



Sequence									
1	D	D	G	W	K	-	-	A	L
2	D	G	G	R	-	V	Q	A	L
3	D	G	G	I	L	V	Q	A	Y
4	E	G	G	I	K	V	Q	A	L
5	E	G	G	I	K	V	Q	A	L
consensus	D	G	G	I	K	V	Q	A	L
vote	3/5	4/5	5/5	3/5	3/5	4/5	4/5	5/5	4/5

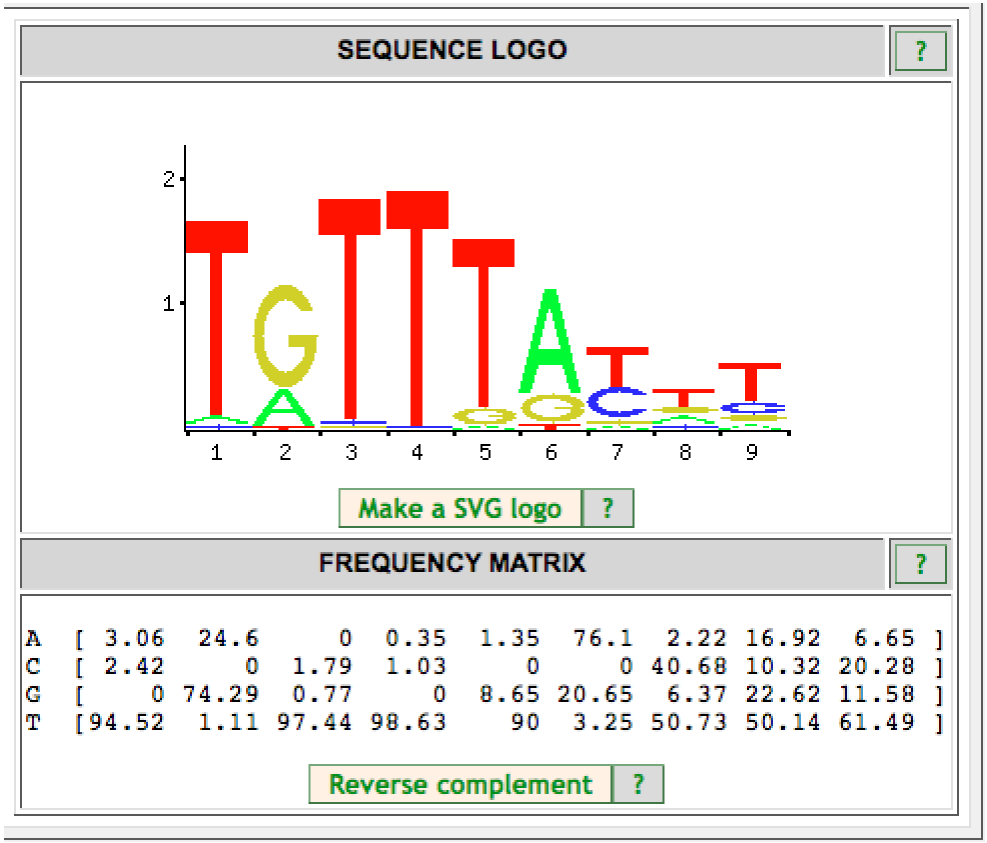
- Fits the pattern [DE]-[DG]-G-[WRI]-[KLVQ](1,3)-A-[LY]

- Does not tell us anything about the frequency of occurrence

Limitations of consensus

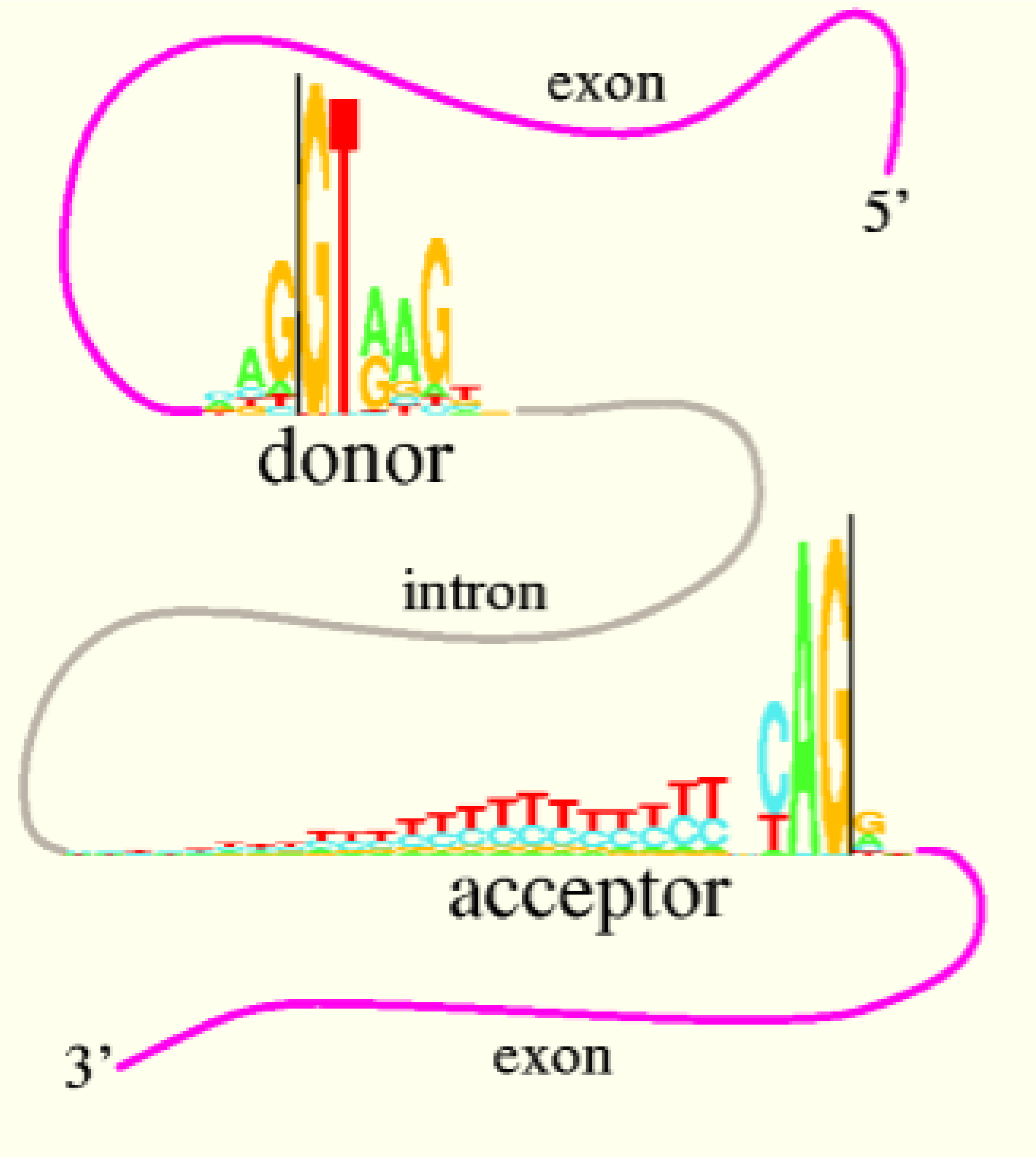
- What if the consensus is weak?
  - Minority, even large minorities
- What if the consensus is nonexistent?
  - Suppose any amino acid can be at a particular position?
  - A degenerate consensus sequence can handle this in many cases, such as the HindII example above.
- How can a consensus handle variable-length gaps?
- Consensus sequences are most useful for highly conserved sequence patterns

Position Frequency Matrix (PFM)



```
A  [ 3.06  24.6     0  0.35  1.35  76.1  2.22 16.92  6.65 ]
C  [ 2.42     0  1.79  1.03     0     0 40.68 10.32 20.28 ]
G  [    0 74.29  0.77     0  8.65 20.65  6.37 22.62 11.58 ]
T  [94.52  1.11 97.44 98.63    90  3.25 50.73 50.14 61.49 ]
```

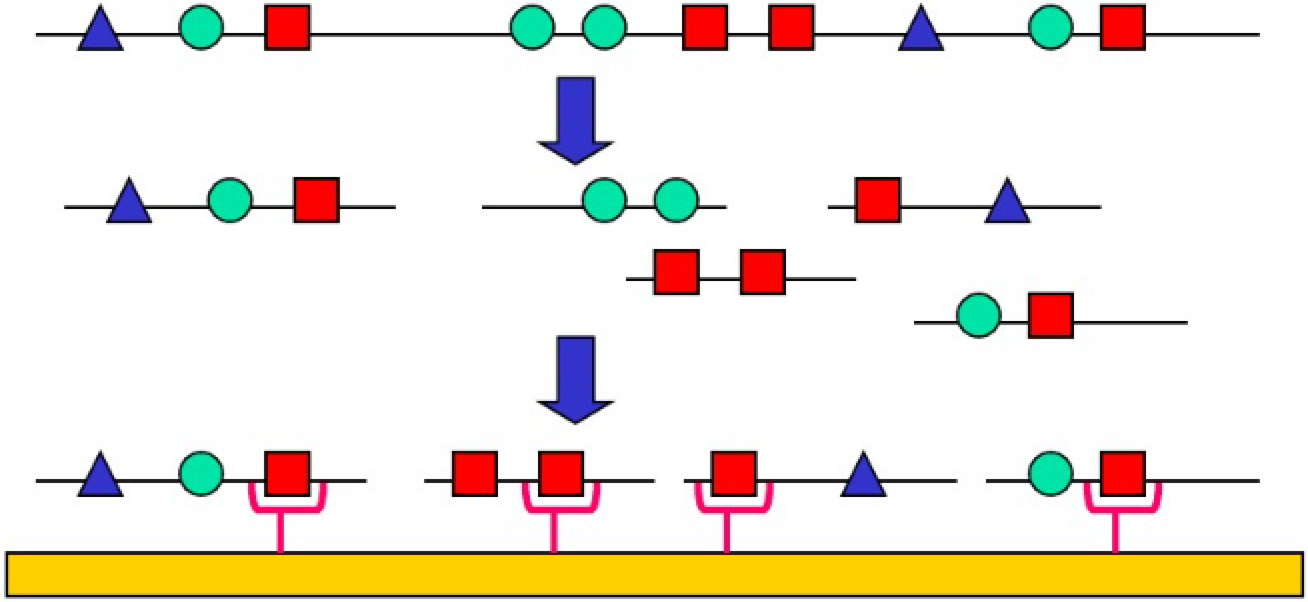
JASPAR: A TF binding site profile database



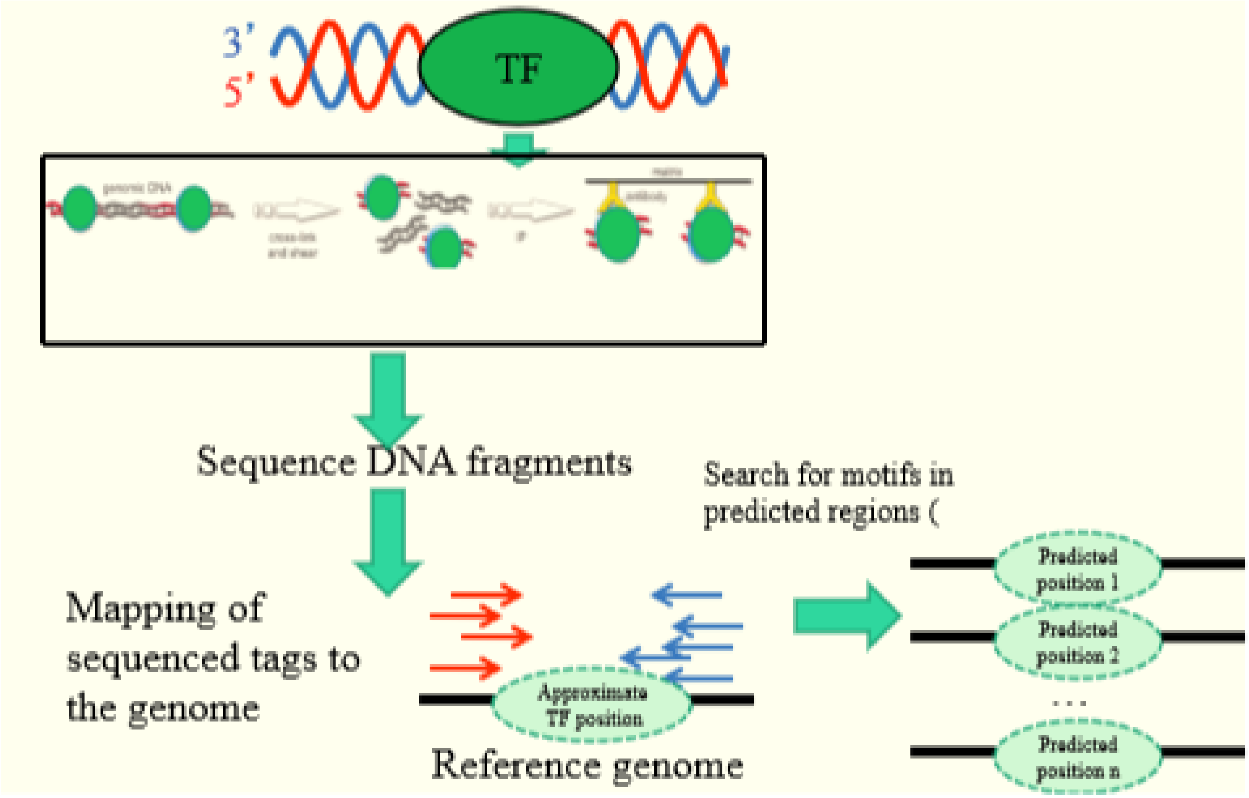
Made by Cao Zhiwei

HOW TO PROVE A MOTIF (chip-seq)

- Detect the interaction between protein (transcription factor) and DNA.



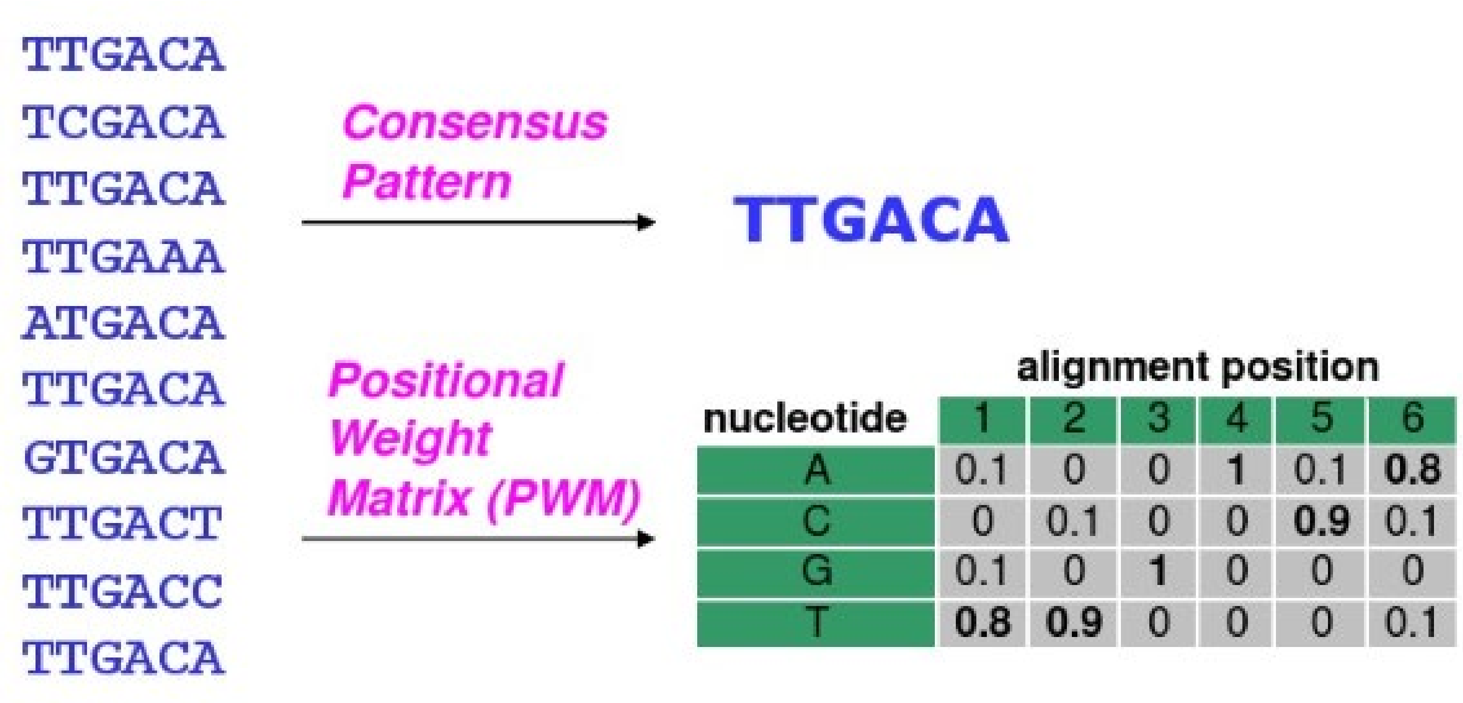
Made by Cao Zhiwei


3'
5'
TF
genomic DNA
cross-link and shear
IP
matrix
antibody
Sequence DNA fragments
Search for motifs in predicted regions (
Mapping of sequenced tags to the genome
Approximate TF position
Reference genome
Predicted position 1
Predicted position 2
. . .
Predicted position n

Sequence alignment has its limitations

- Substitution matrices (BLOSUM, etc.) represents each change as independent of position
  - For example, a Ser -> Ala substitution is given the same penalty no matter where it occurs

- We can see that this is not always a good representation of reality
  - E.g., sometimes a Ser -> Ala substitution may destroy the function of a protein.
  - In that case, should it have the same penalty in sequence alignment as an ordinary Ser -> Ala substitution?



TTGACA
TCGACA
TTGACA
TTGAAA
ATGACA
TTGACA
GTGACA
TTGACT
TTGACC
TTGACA

Consensus Pattern → **TTGACA**

Positional Weight Matrix (PWM) →

	alignment position					
nucleotide	1	2	3	4	5	6
A	0.1	0	0	**1**	0.1	**0.8**
C	0	0.1	0	0	**0.9**	0.1
G	0.1	0	**1**	0	0	0
T	**0.8**	**0.9**	0	0	0	0.1

Position Specific Scoring Matrix (PSSM)

- In the real world of protein sequences:
  - we might not want to use a strict consensus
  - we might not want to use a yes/no criterion like patterns
  - we might not have enough observations to describe all possibilities well by a PFM
- What do we do?
  - Rather than using frequencies, we would like to use likelihoods of occurrence at each position.

PSSM

- Position-specific table or matrix containing comparison information for aligned sequences.
- Columns represent positions in sequences
- Rows contain score for alignment of positions with each residue
  - Used to find sequences similar to alignment rather than one sequences

Summary

✓Interesting biology is encoded in sequence motifs
✓ Position matters within sequence motifs
✓ Because position matters, pairwise alignment has a hard time detecting motifs.
✓ Patterns can be used to define motifs qualitatively
✓ Profiles can be used to score motifs quantitatively
✓PSSM can be used to quantitatively score motifs at specific sites
✓ Motifs can be used to search sequence Databases
✓ Sequences can be used to search pattern and profile databases

课堂作业

- 网上自学，**PSSM** 构建原理
- 浏览 **Prosite** 网站