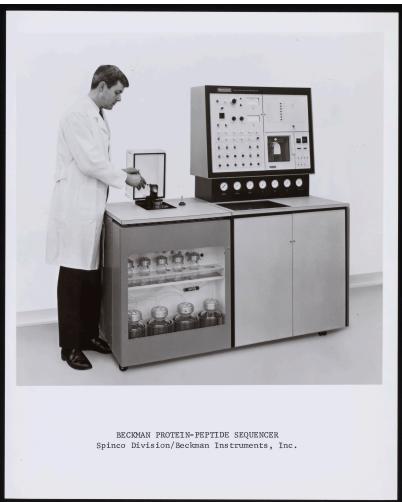
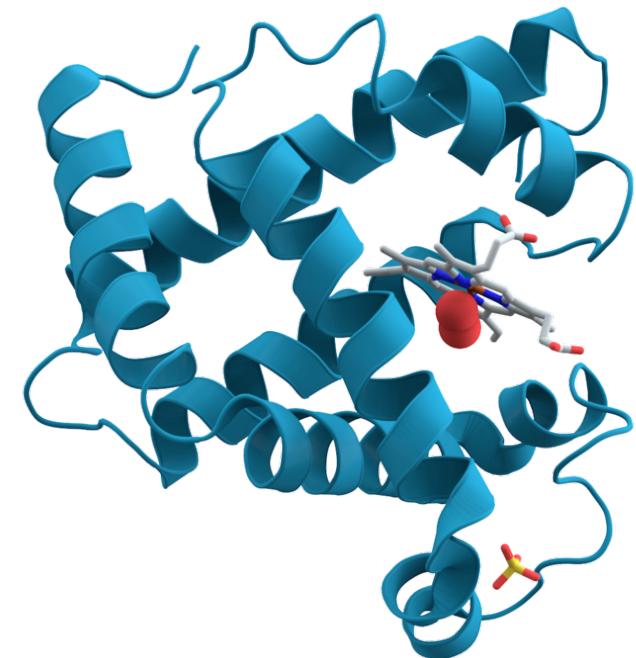


Protein bioinformatics: introduction



Episode in the series on
protein bioinformatics

Mikael Bodén

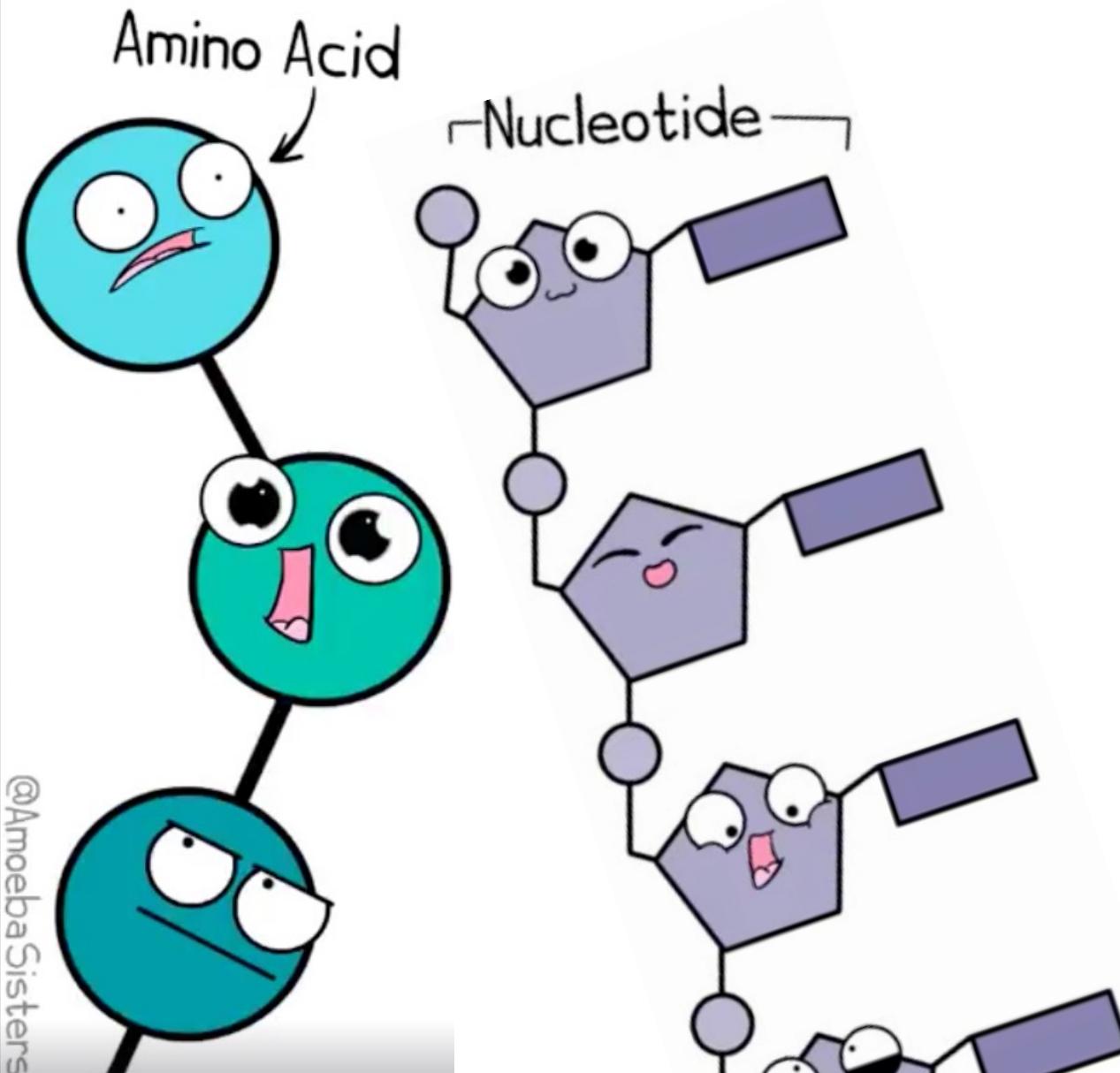


Part 1: Protein sequence

Motivation

Amino acids

Sequence to structure



Structure of the RNA-dependent RNA polymerase from COVID-19 virus

Yan Gao^{1,2*}, Liming Yan^{1*}, Yucen Huang¹✉, Fengjiang Liu²✉, Yao Zhao², Lin Cao³, Tao Wang¹, Qianqian Sun², Zhenhua Ming⁴, Lianqi Zhang¹, Ji Ge¹, Litao Zheng¹, Ying Zhang¹, Haofeng Wang^{2,5}, Yan Zhu², Chen Zhu², Tianyu Hu², Tian Hua², Bing Zhang², Xiuna Yang², Jun Li², Haitao Yang², Zhijie Liu², Wenqing Xu², Luke W. Guddat⁶, Quan Wang²†, Zhiyong Lou¹†, Zihe Rao^{1,2,3,7}†

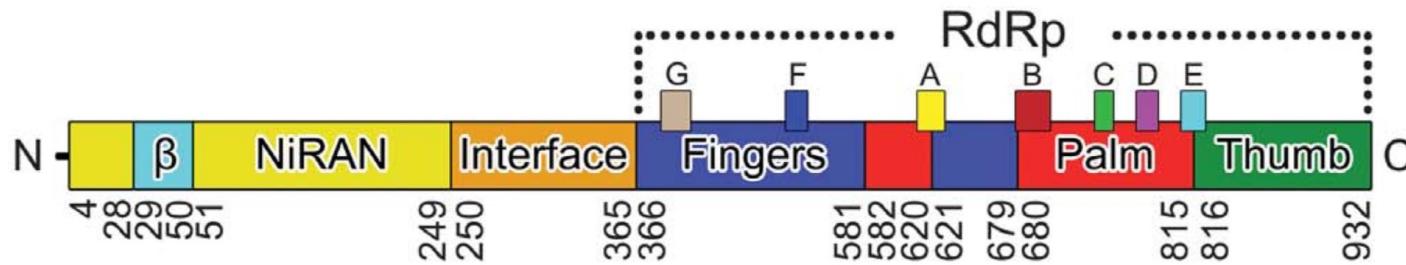
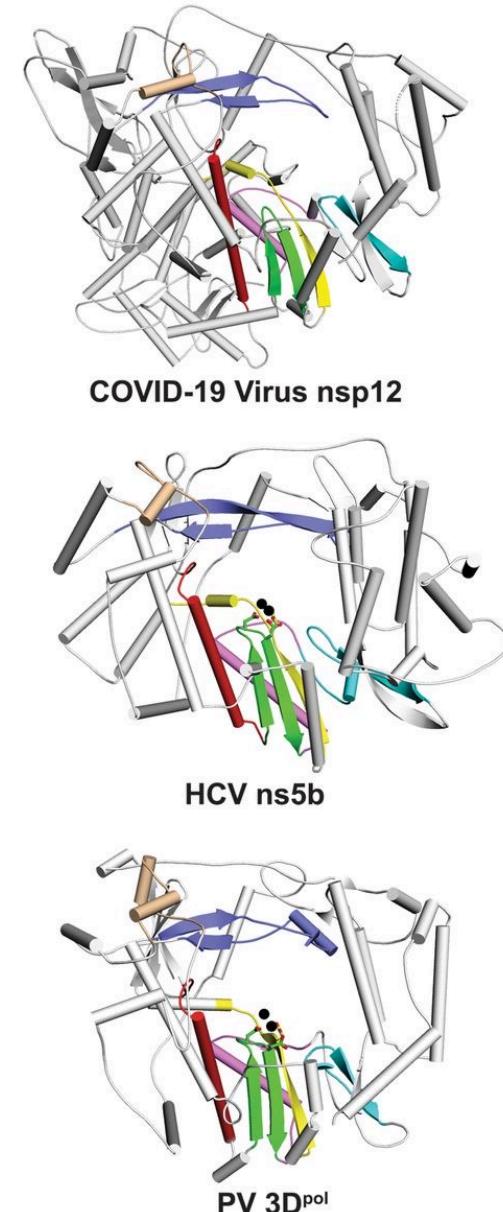


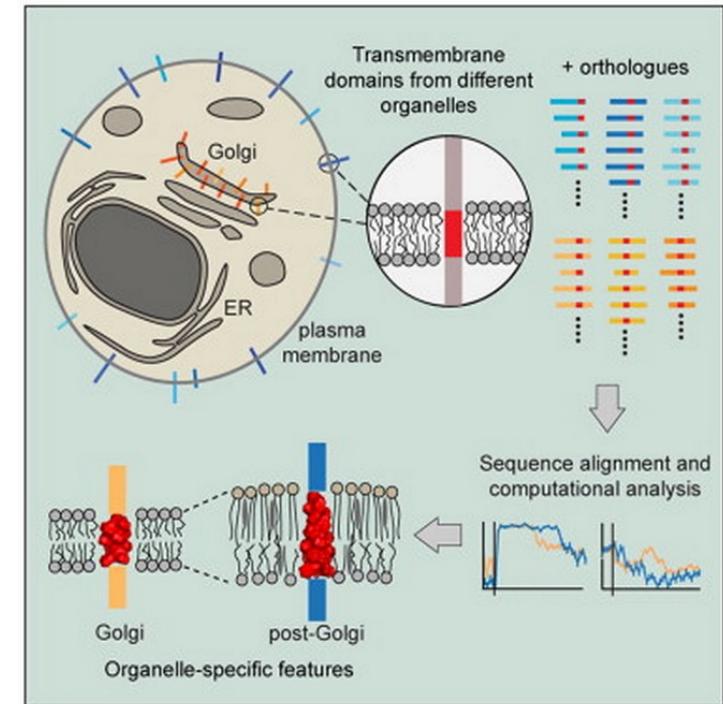
Fig. 1. Structure of COVID-19 virus nsp12

(A) Domain organization of COVID-19 virus nsp12. The interdomain borders are labeled with residue numbers. The N-terminal portion with no cryo-EM map density and the C-terminal residues that cannot be observed in the map are not included in the assignment. The polymerase motifs are colored as follows: motif A, yellow; motif B, red; motif C, green; motif D, violet; motif E, cyan; motif F, blue; and motif G, light brown.

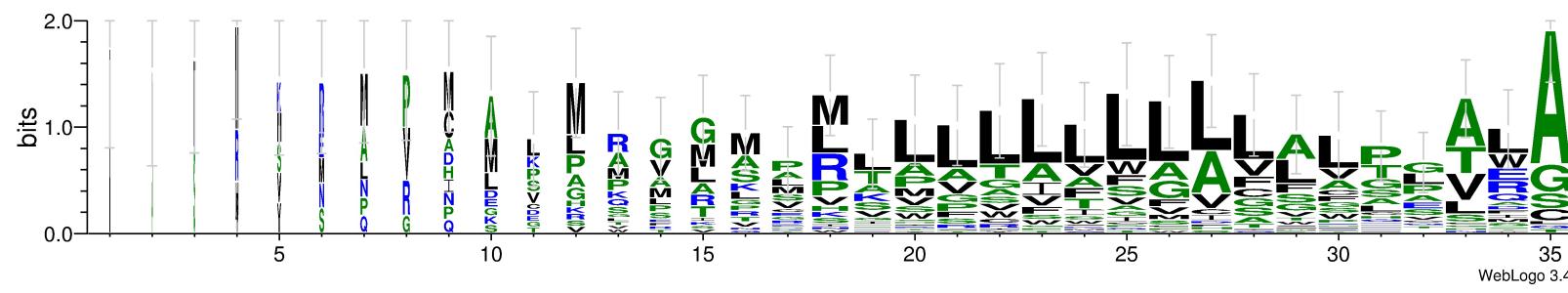


The secretory pathway

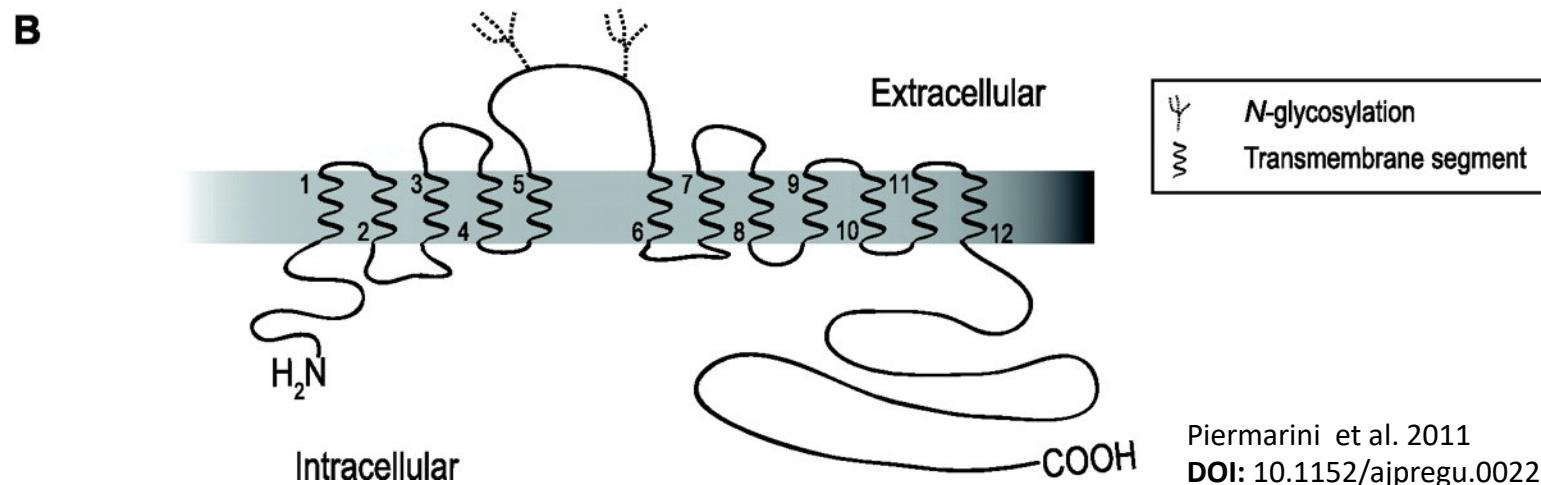
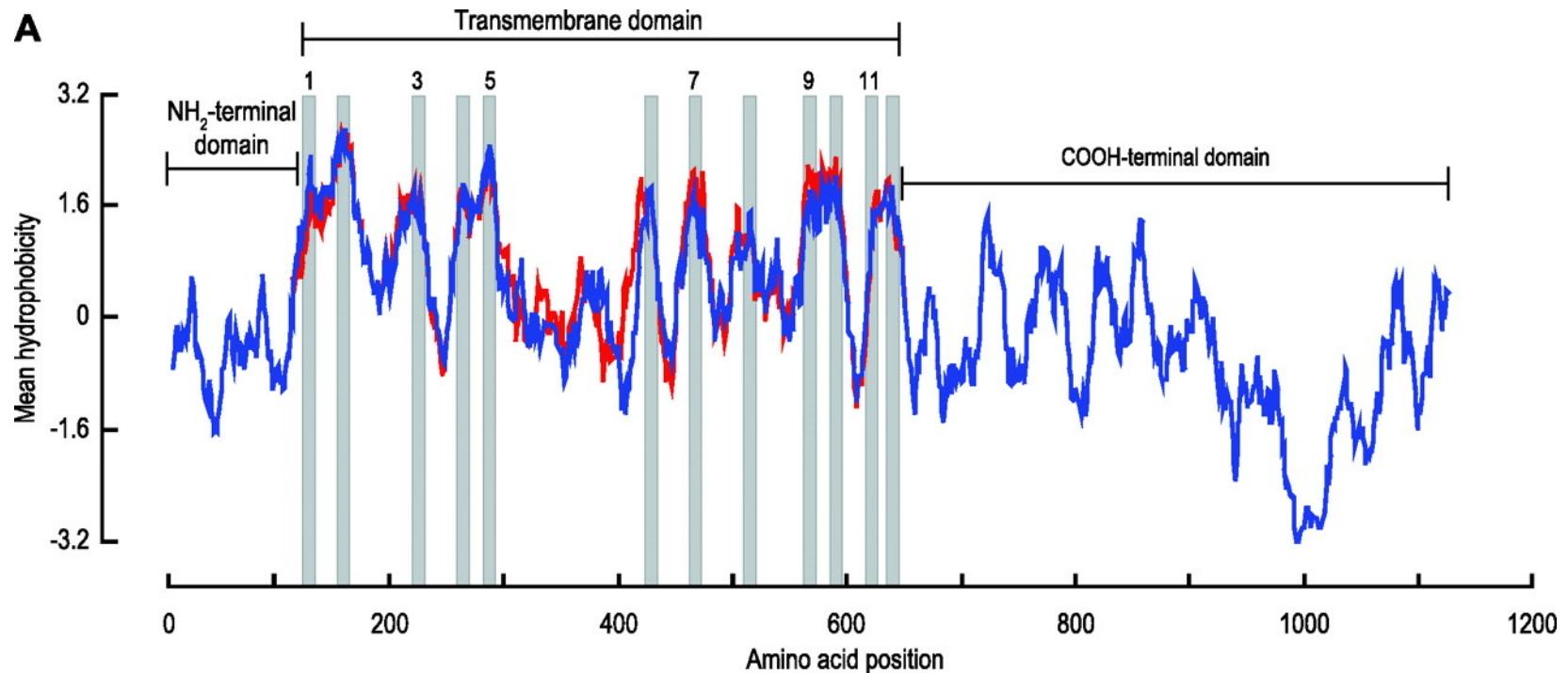
- Signal and transmembrane sequences have
 - core hydrophobic region
 - Propensity to form “helices”



Sequence logo of human signal peptides

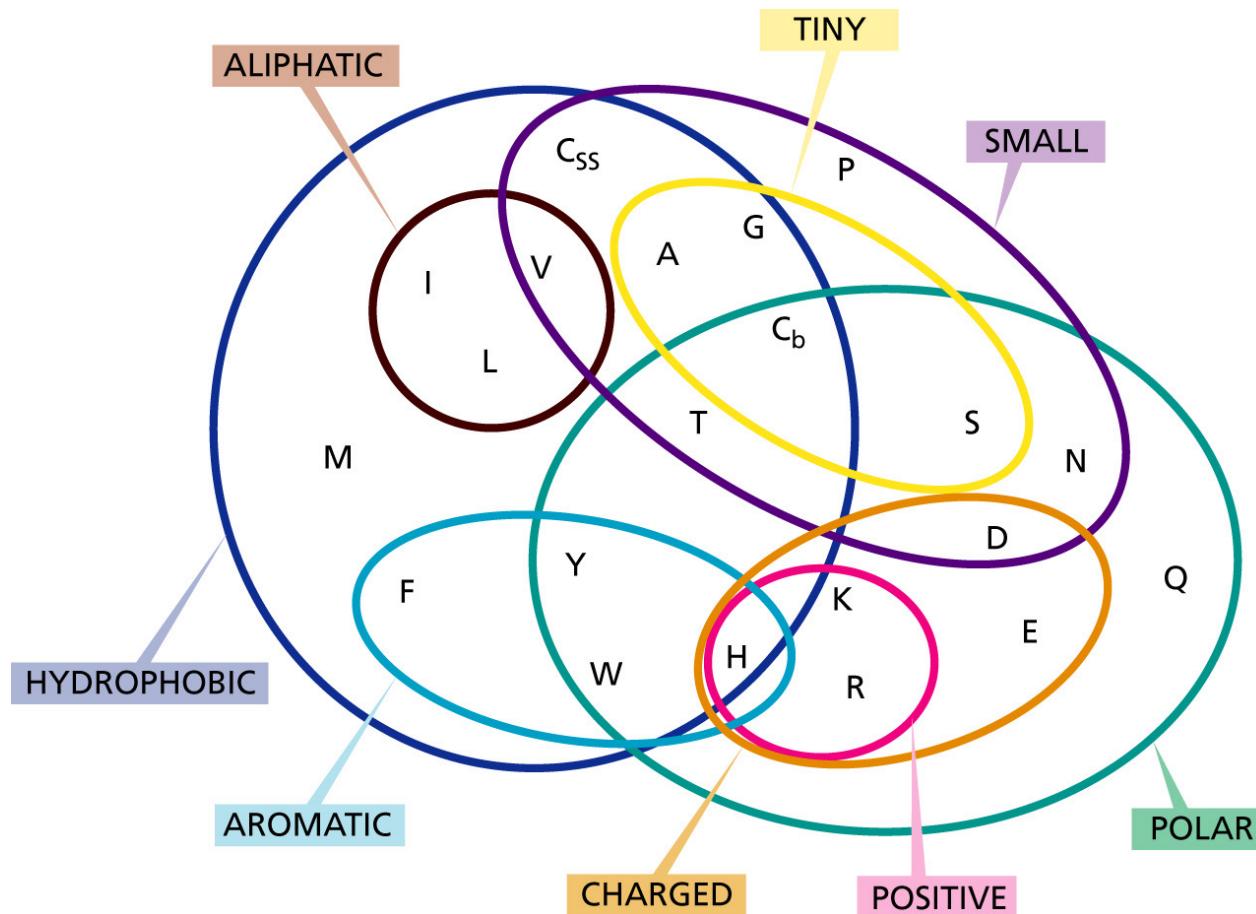


Generated by Sarah Kempe (BINF6000 in 2015)

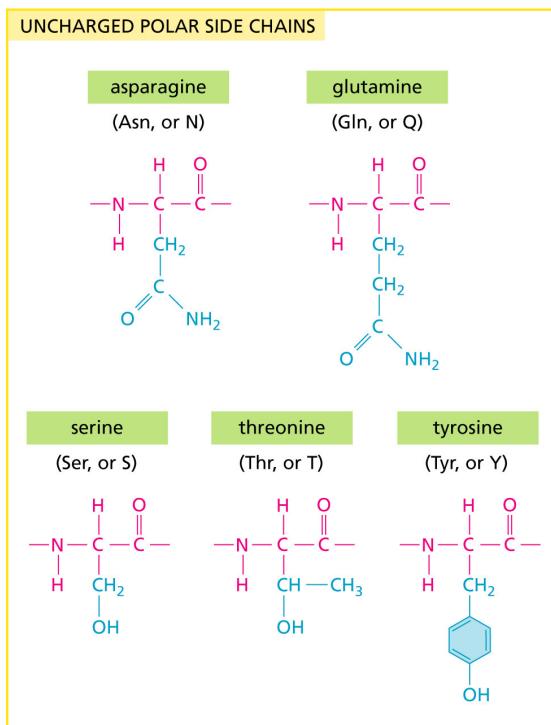
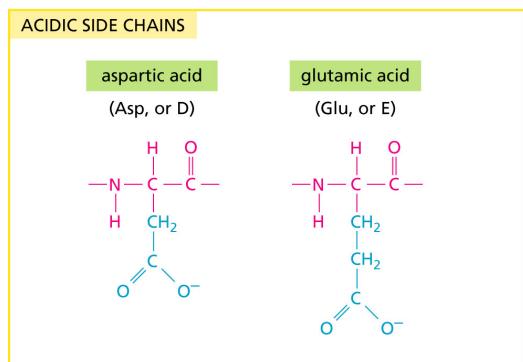
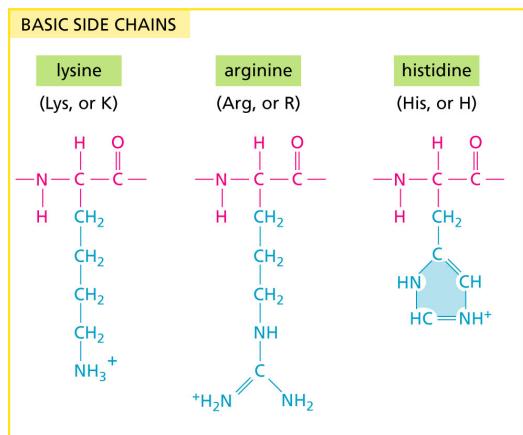


Piermarini et al. 2011
DOI: 10.1152/ajpregu.00223.2011

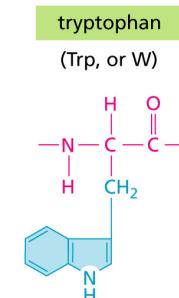
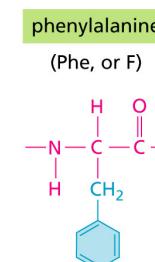
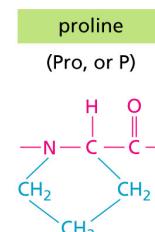
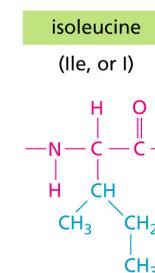
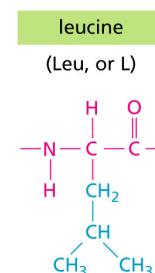
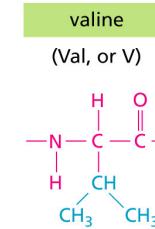
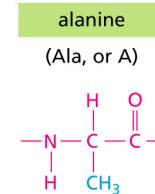
Protein parts are physiologically diverse



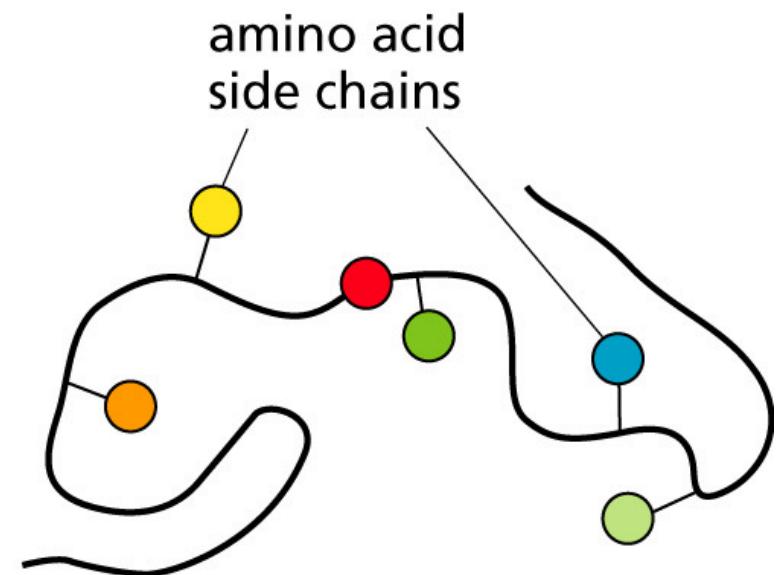
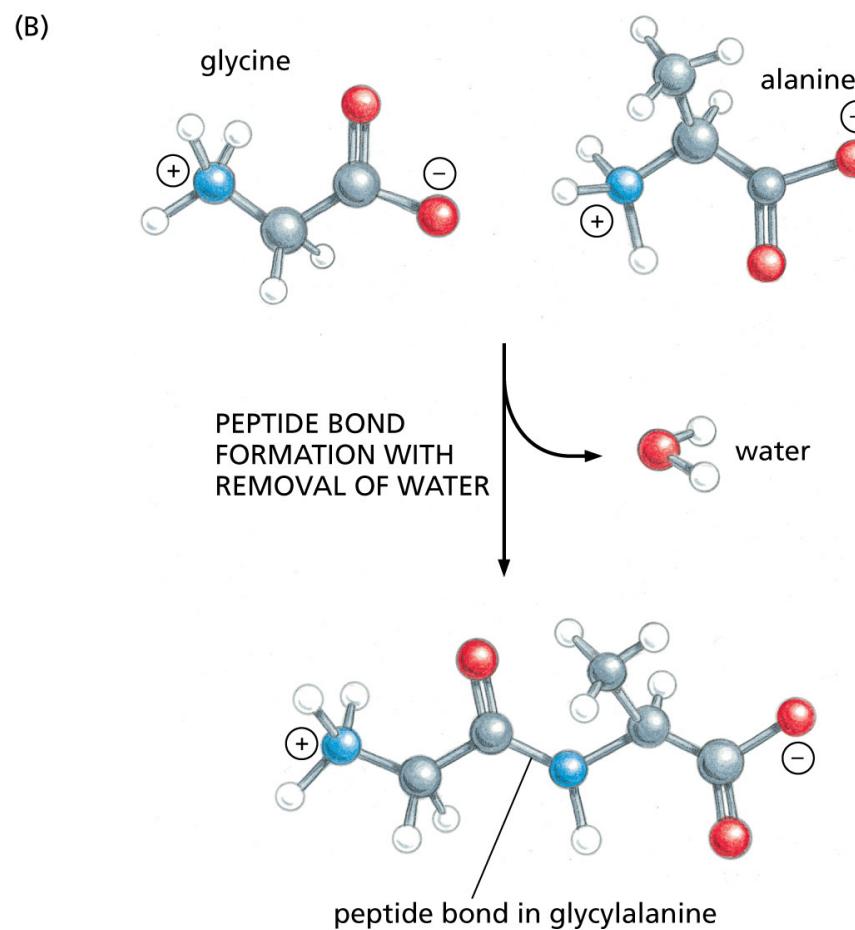
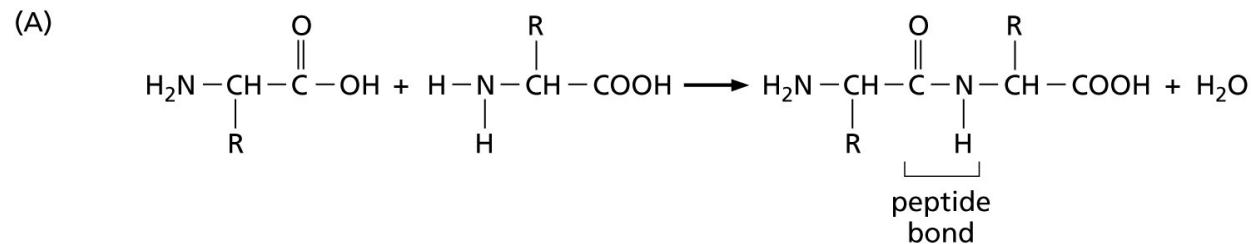
Amino acids



NONPOLAR SIDE CHAINS



Zvelebil and Baum Fig. 2.3



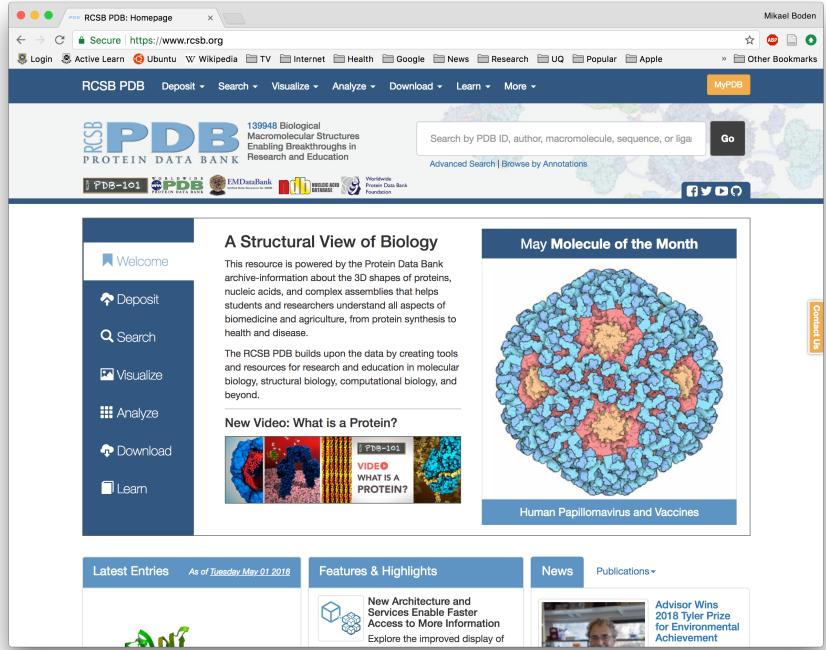
Zvelebil and Baum Fig. 2.4 p 31 and Fig. 2.14 p 41.

Protein sequence

- Translated by ribosomes
- Composed of amino acids
 - “starting” at the amino-terminal (N-terminal)
 - “ending” at the carboxyl-terminal (C-terminal)
- Mass spectrometry can “sequence” proteins
- Post-translational modifications, e.g.
 - Glycosylation
 - Methylation
 - Acetylation
 - Ubiquitination
 - Sumoylation
- Proteins are sometimes cleaved, degraded and ligated



End part 1:
Protein
sequence



Part 2: Protein structure

Primary, secondary and tertiary structure

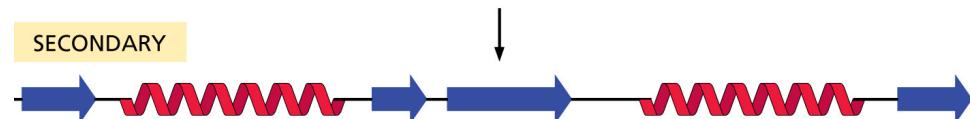
α -helix, β -sheet

Prediction of secondary structure, Chou-Fasman propensity and algorithm

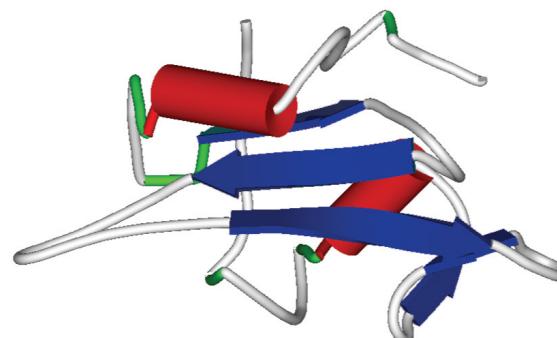
PRIMARY

N terminus—...MYCATISEATINGFISHANDMEATANDWATER...—C terminus

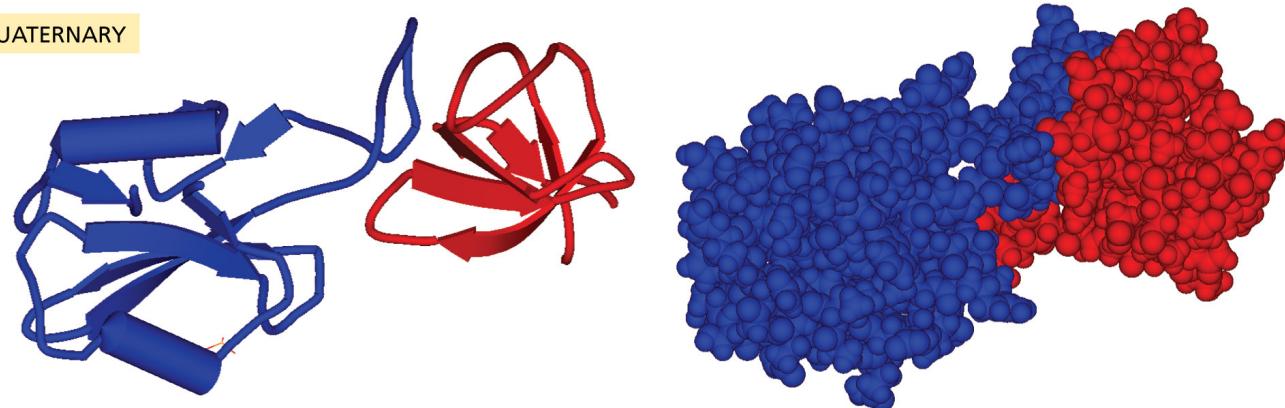
SECONDARY



TERTIARY

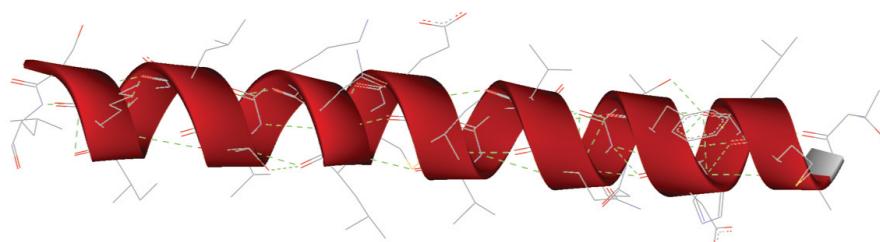


QUATERNARY

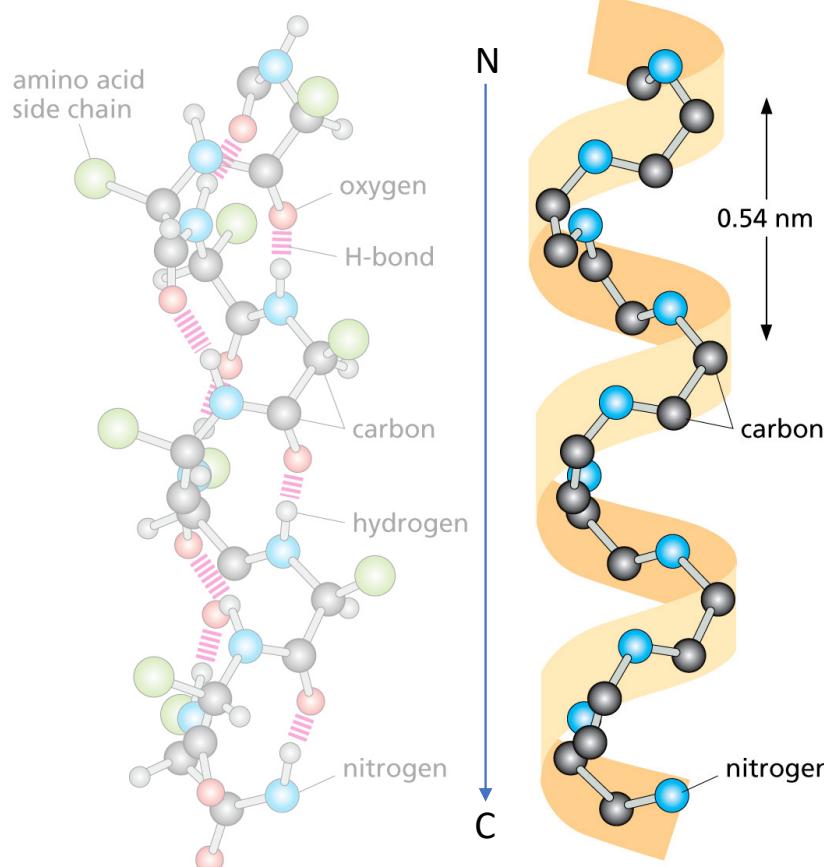


α -helix

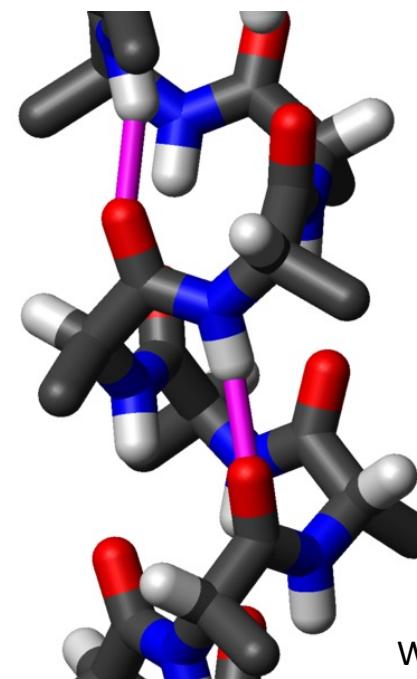
(A)



(B)



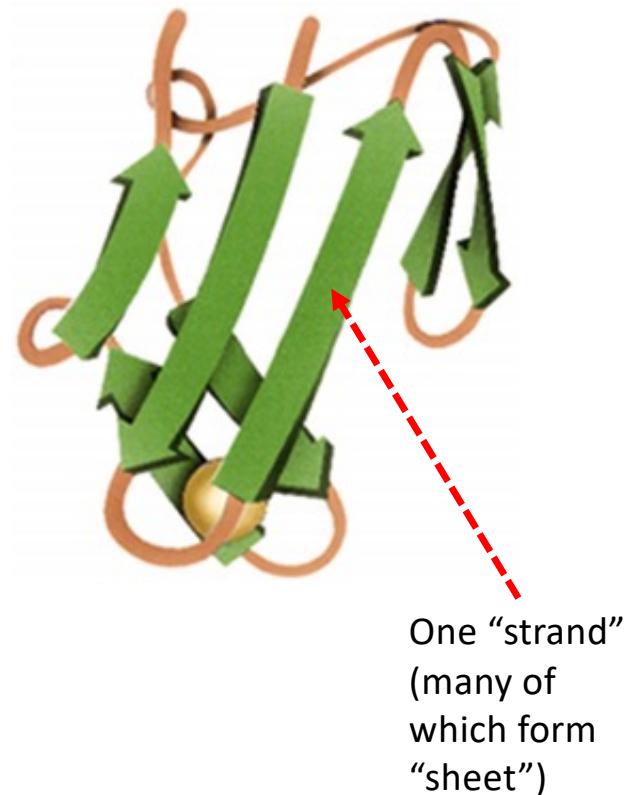
- 3.6 amino acid (residues) per turn
- O(i) hydrogen binds to N($i+4$)
[where i is a residue position in the sequence]



Wikipedia

β -sheet

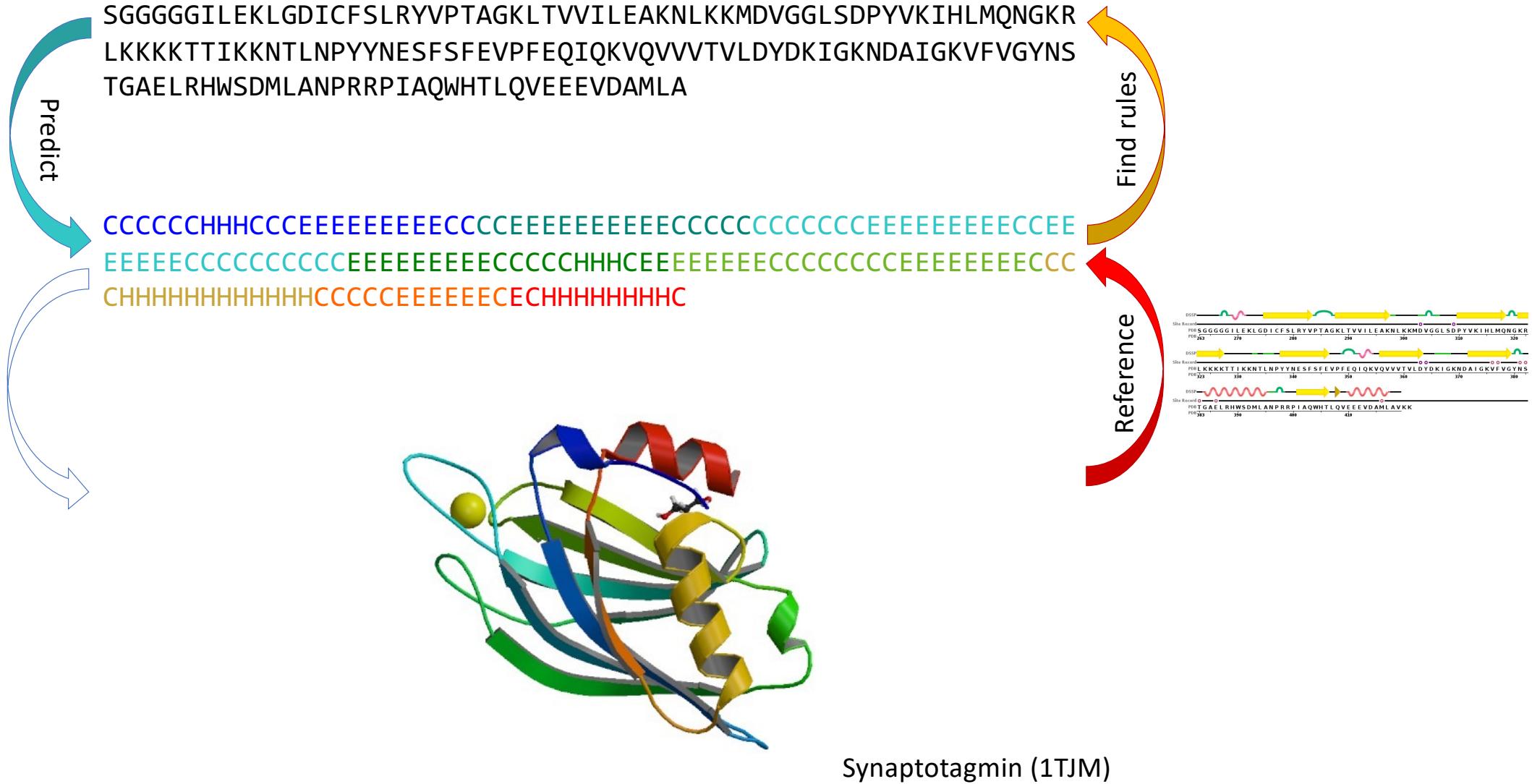
- Shape maintained by intra-molecular H bonding between chains (β -strands)
- Longer range interactions often involved
- Sheets are usually curved and can even form barrels



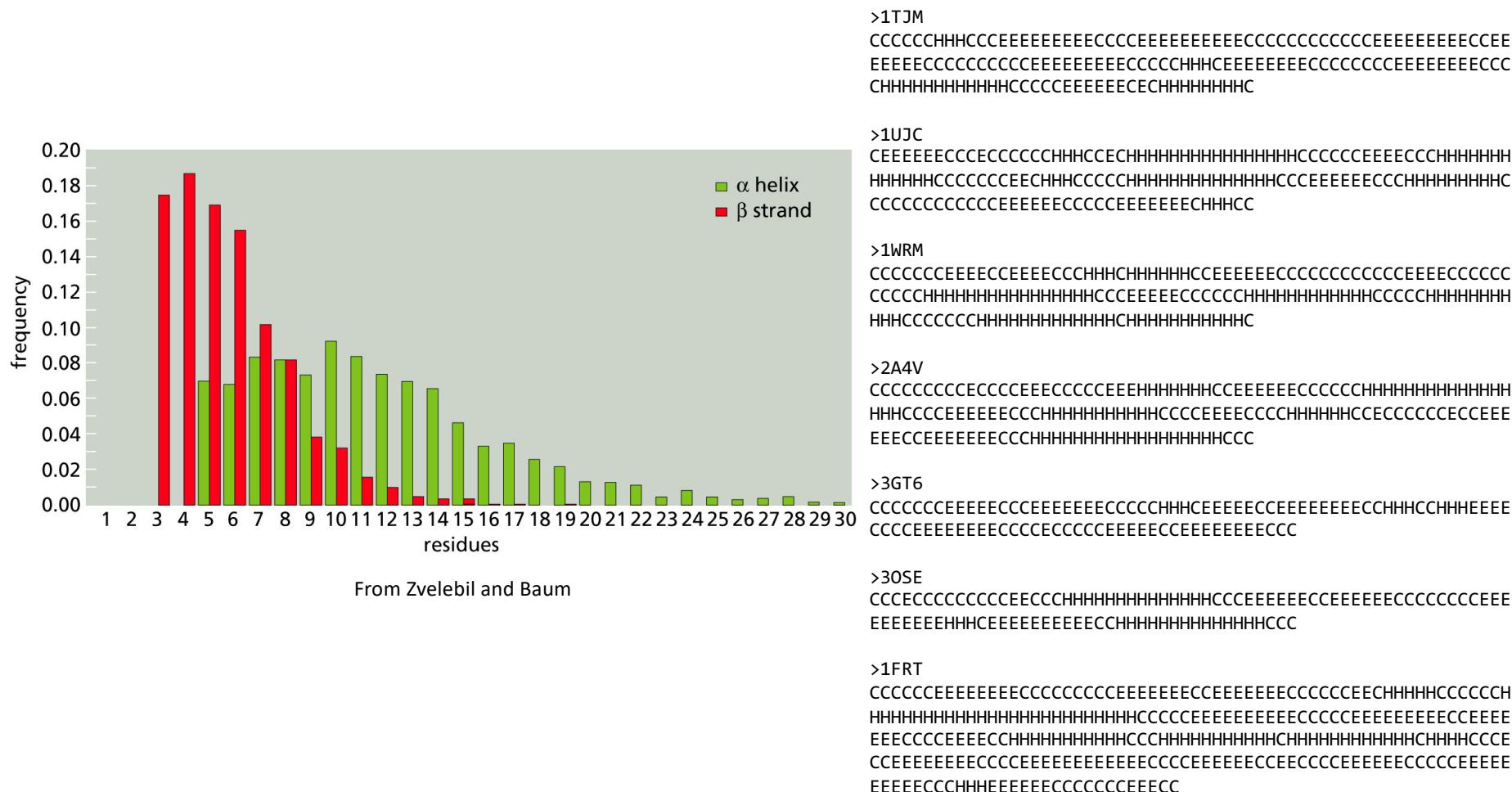
Secondary structure states

- Secondary structure are defined in terms of 3, 4 or even 7 classes, which include
 - α -helix (**H**)
 - β -strand (**E**)
 - β -turn (**T**) (sometimes excluded)
 - random coil (**C**)

B (Beta-bridge)
E (Beta-strand)
G (Helix-3)
H (Alpha-helix)
I (Helix-5)
S (Bend)
T (Turn)



Question: what can we learn from a set of proteins with each residue labeled as having a secondary structure?



Question: what can we learn from a set of proteins with each residue labeled as having a secondary structure?

>1TJM
SGGGGGILEKLGDICFSLRYVPTAGKLTVVILEAKNLKKMDVGLSDPYVKIHLMQNGKR
LKKKTTIKKNTLPYNNESFSFEVPFEQIQKVQVVTVLDYDKGKNDAGKVFVGYN
TGAELRHWSDMANPRRPIAQWHTLQVEEVDAMLA

>1UJC
MQFIMRHDAALDAASDSVRPLTTNGCDESRLMANWLKGQKVIEIRVLVSPFLRAEQL
EEVGDCNLPSAELVPELTPCGDVGLVSAYLQALTNEGVASVLVISHLPLVGYLV
AELCPGETPPMFTTSAIASVTLDESGNGTFNWQMSPCNLK

>1WRM
GPMGNGMNKILPGLYIGNFKDARDAEQLSKNKVTIILSVHDSARPMLEGVKYL
CIPAADSPSQLTRHFKEISKIFIHECRRLRGESCLVHCLAGVRSVTLVI
YIMTVTDGFEDALHTVRAGRSCANPNVGFQRQLQEFEKHEVHQYRQLKEEY

>2A4V
DVNELEIGDPIDPSLLNEDNDSISLKKITENNRRVVFVYPRASTPGSTRQASGRDN
YQELKEYAAVFGLSADSVTQSQQKFQSKQNLPYHLLSDPKREFIGLLGAK
KTPLSGSIRSHFIFVDGKLKFKRVKISPEVSNDAKKEVLEVAEKFKE

>3GT6
TAQWVPRVDIKEEVNHFVYADLPGIDPSQIEVQMDKGILSIRGERKSESSTETERFS
ERRYGSFHRRFALPDSADADGITAAGRNGVLEIRIPKRPA

>3OSE
KPRSLRFTWSMKTSSMDPNDMMREIRKVLDA
NCDYEQKERFLLFCVHG
DARQDSL
VQWE
MEVCKLPLRLSNGVRFKRISGTSIAFKNIASKIAN
ELKL

>1FRT
AEPRPLPMYH
LAAVSDLSTGLPSFWATGWLG
AQQLTYNNLRQEADPCGA
WI
WENQVSWY
WEKETTD
LKSKEQLF
LEAIRT
LENQ
INGFTLQ
QGLL
GCE
LAPD
NSL
PTAV
FALNG
EEFM
RFNP
RTGN
WSGE
WPET
DIV
GNL
WMK
P
QPE
A
R
K
E
S
F
L
T
S
C
P
E
R
L
L
G
H
L
R
G
R
Q
N
L
E
W
K
P
S
M
R
L
K
A
R
P
G
N
S
G
S
V
L
T
C
A
A
F
S
F
Y
P
P
E
L
K
F
R
F
L
R
N
G
L
A
S
G
S
G
N
C
S
T
G
P
N
G
D
G
S
F
H
A
W
S
L
L
E
V
K
R
G
D
E
H
Y
Q
C
O
V
E
H
E
G
L
A
Q
P
L
T
V
D
L

>1TJM
CCCCCCCCHHCCCEEEEEEECCCCCEEEEEEECCCCCCCCCCCCCEEEEEEECCCC
EEEECCCCCCCCCEEEEEEECCCCCCCCHHCEEEEEEECCCCCCCCCEEEEEEECCCC
CHHHHHHHHHHHHHCCCCCEEEEECECHHHHHHHHH

>1UJC
CEEEEEECCECCCCCHHHCECHHHHHHHHHHHHHCCCCCEEEECCCCHHHHHHHH
HHHHHHCCCCCEEECHHHCCCCHHHHHHHHHHHHCCCCCEEEECCCCHHHHHHHH
CCCCCCCCCCCCCEEEEECCCCCEEEEEECCHHHCC

>1WRM
CCCCCCCCCEECCCEEECCCCHHCHHHHHHHCCCCCEEEECCCCCCCCCEEEECCCC
CCCCHHHHHHHHHHHHHHHHHHCCCCCEEEECCCCHHHHHHHHHHHHCCCCHHHHHHHH
HHHCCCCCCCCHH

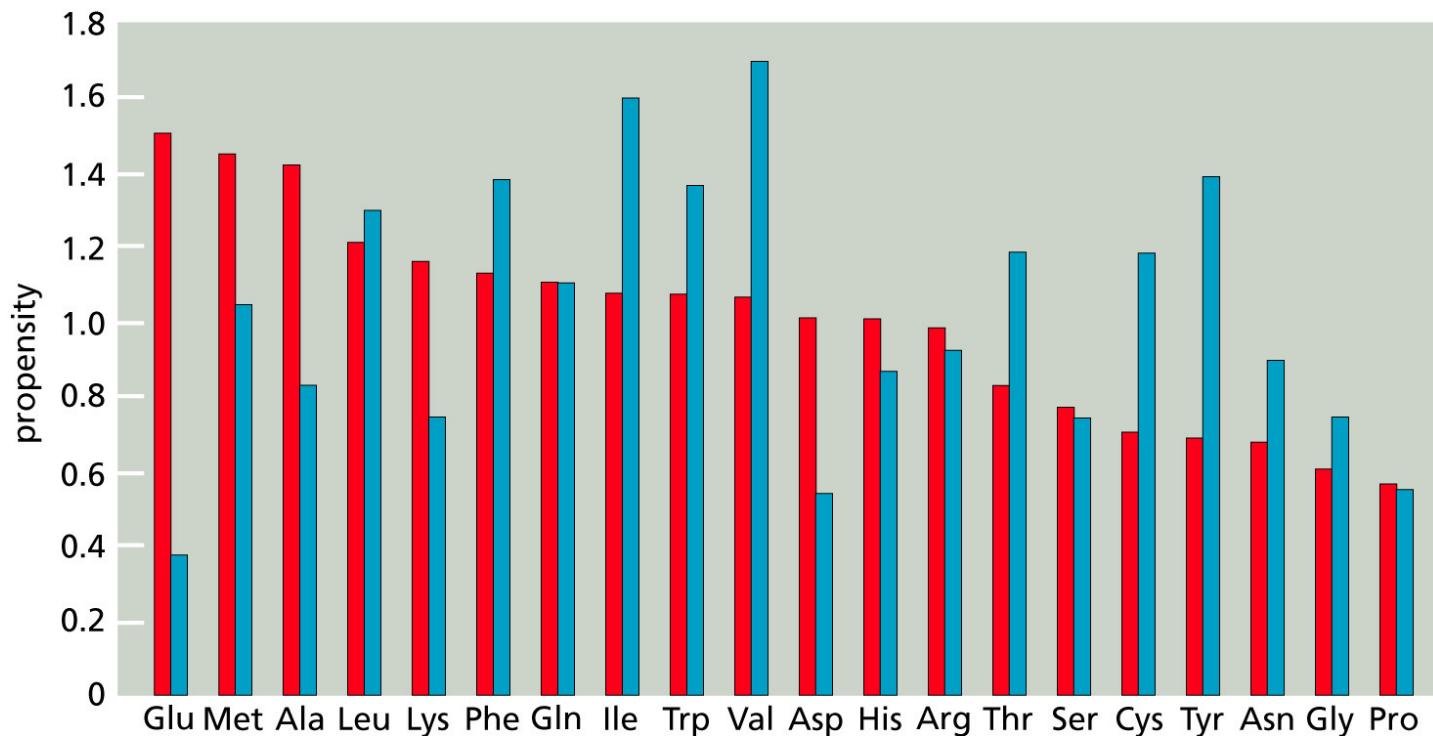
>2A4V
CCCCCCCCCECCCCEEECCCCCEEEHHHHHHCCCCCEEEECCCCHHHHHHHHHHHHHH
HHHCCCCCEEEECCHHHHHHHHHHHHHHHHHCCCCCEEEECCCCHHHHHHCCCCCECCEEE
EEECCCCCCCCHH

>3GT6
CCCCCCCCCEEECCCCCEEECCCCHHCEEEECCCEEEEEECCCHHHCCHHHEEEE
CCCCCCCCCEEECCCCCEEEECCEEEEEECC

>3OSE
CCCECCCCCCCCCEEECCCCHHHHHHHHHHHHHHHHCCCCCEEEECCCEEEEEECCCC
EEEEEEHHCEEEEEECCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

>1FRT
CCCCCCCCCCCCCCCCCEEEEEECCCEEEEEECCCCCEEECCCCCEEECCCCCCCC
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCEEEEEECCCCCEEEEEECCCC
EEECCCCCEEECCCCHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHH
CCCCCCCCCCCCCEEEEEECCCCCEEEEEECCCEEECCCCCEEEEEECCCCCCCC
EEEECCCCHHHEEEEEECCCCCCCCCEEECCCC

Chou and Fasman propensities (P)



$$P_{s,a} = \frac{p_{s,a}}{p_a}$$

See Zvelebil and Baum

Chou and Fasman propensities (P)

Amino acid	α -helix		β -strand	
	Designation	P	Designation	P
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74
Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

F = strong former

f = weak former

I = indifferent

b = weak breaker

B = strong breaker



Chou-Fasman algorithm

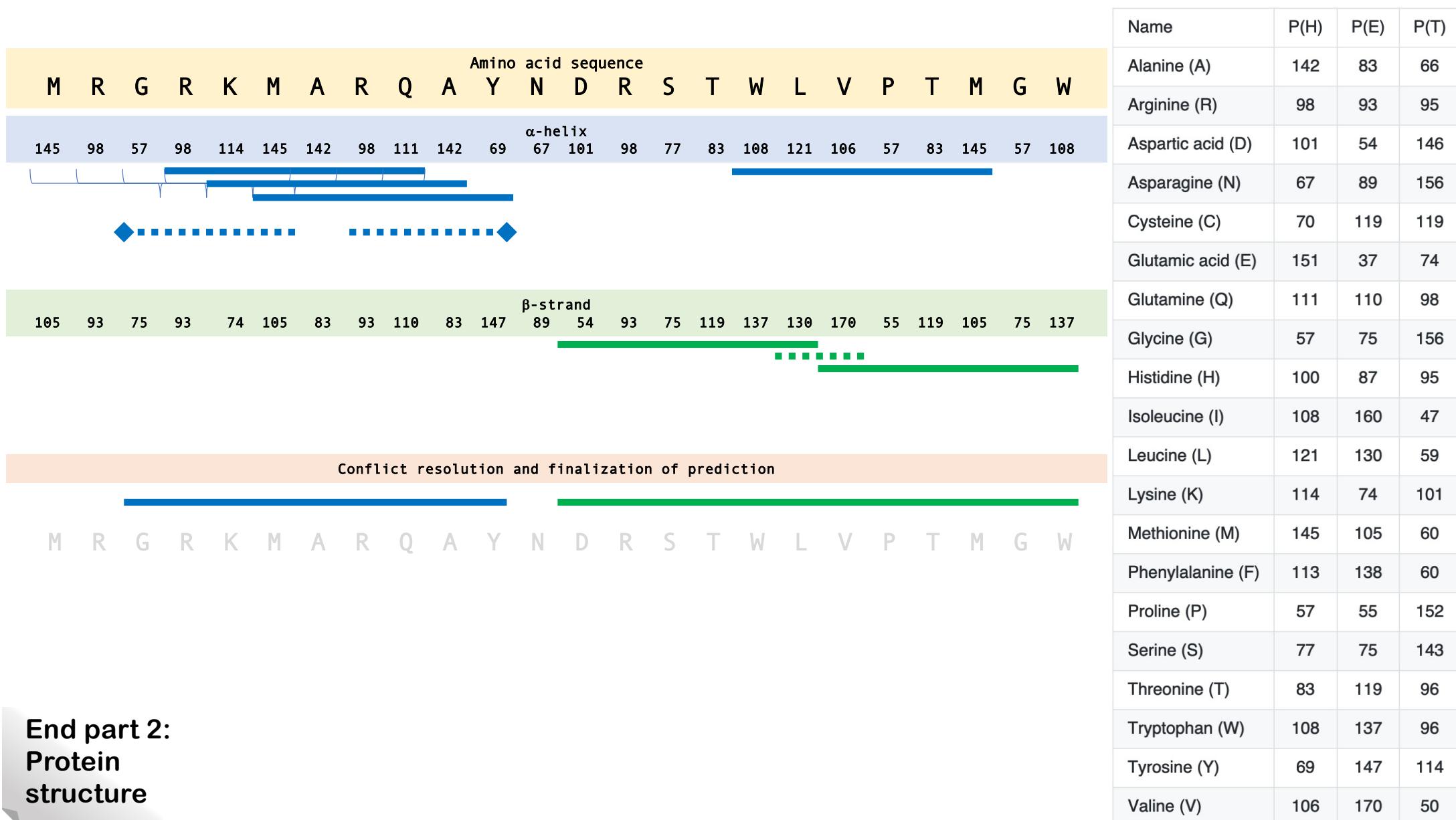
- Helix, Strand
 1. Scan for window of 6 residues, **mark state** when score > 100 for *at least* 4 residues (for helix state) or 3 residues (for strand state)
 2. Propagate mark in both directions until 4 (or 3) residue window has *average* propensity < 100
 3. Move forward and repeat

GHWIATRGQLIREAYEDYRHFSSECPFIP

- Conflict solution

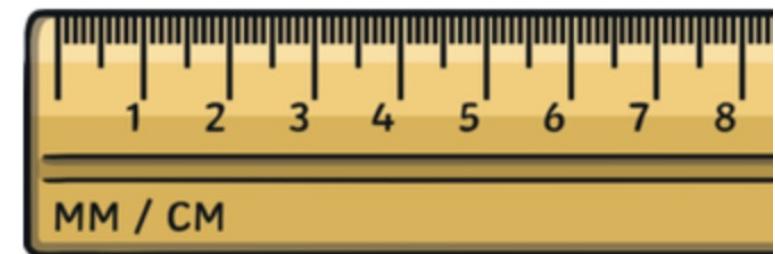
Any region containing overlapping alpha-helical and beta-strand assignments are taken to be

 - helical if the average $P(\text{helix}) > P(\text{strand})$
 - strand if the average $P(\text{strand}) > P(\text{helix})$
- Accuracy approx. 50% - 60%



**End part 2:
Protein
structure**

	Helix prediction	Strand prediction	Coil prediction
Helix observation	523	131	83
Strand observation	42	509	102
Coil observation	31	23	389



Part 3: Metrics

Classification, confusion matrix and threshold

Sensitivity and specificity

Multi-class problems, accuracy and Q_k

Classification metrics: terminology

- *true positives*, $tp(C)$, is the number of test samples in class C , predicted to be in class C
- *true negatives*, $tn(C)$, is the number of test samples not in class C , predicted *not* to be in class C
- *false negatives*, $fn(C)$, is the number of test samples in class C , predicted *not* to be in class C , and
- *false positives*, $fp(C)$, which is the number of test samples not in class C , predicted to be in class C

Classification metrics: confusion matrix

	Positive prediction	Negative prediction
Positive observation	123 (TP)	31 (FN)
Negative observation	15 (FP)	207 (TN)

Observed	Prediction
True	False
False	False
True	True
False	False
True	True
True	False
True	False
False	True
False	False
True	True

Outcome C=True
FN
TN
TP
TN
TP
FN
FN
FP
TN
TP



Test metrics

	Positive prediction	Negative prediction
Positive observation	3 (TP)	3 (FN)
Negative observation	1 (FP)	3 (TN)

Observed	Prediction
True	0.18
False	0.25
True	0.78
False	0.11
True	0.55
True	0.28
True	0.45
False	0.61
False	0.22
True	0.62

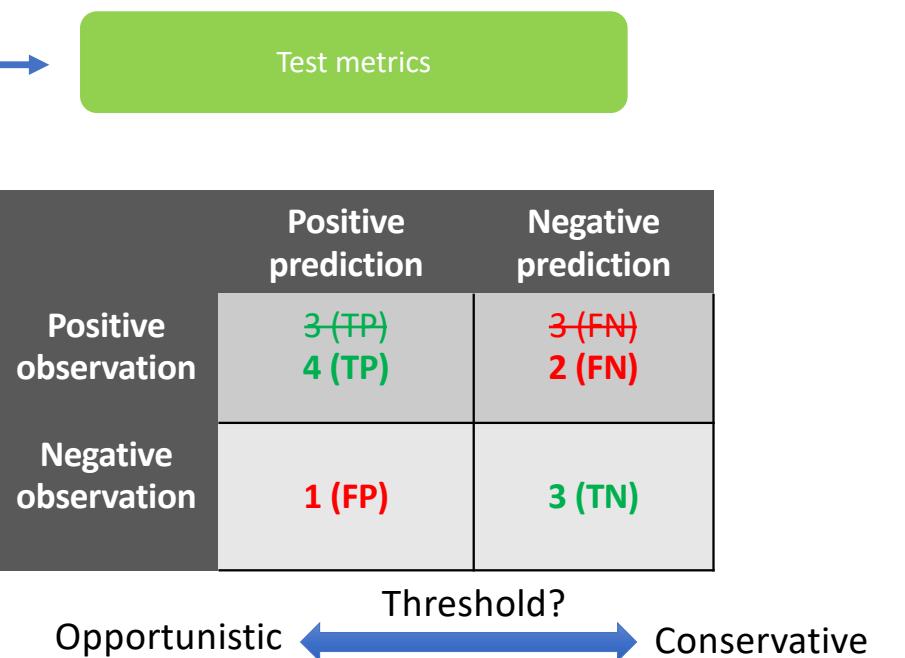
Outcome C=True $\theta=0.50$		Test metrics	
		Positive prediction	Negative prediction
Positive observation	TP	3 (TP)	3 (FN)
	FP	1 (FP)	3 (TN)
Negative observation	FN		
	TN		

Threshold?
Opportunistic ← → Conservative

Observed	Prediction
True	0.18
False	0.25
True	0.78
False	0.11
True	0.55
True	0.28
True	0.45
False	0.61
False	0.22
True	0.62

Outcome $C=\text{True } \theta=0.40$

FN
TN
TP
TN
TP
FN
TP
FP
TN
TP



Classification metrics: more of them

- Sensitivity

$$\frac{tp}{tp + fn}$$

- Specificity

$$\frac{tn}{tn + fp}$$

- Recall

$$\frac{tp}{tp + fn}$$

- Precision

$$\frac{tp}{tp + fp}$$

```
j = sstr[888]
s = seqs[888]
for i in range(len(s) - 2):
    print("Observed:", j[i+1], "Predicted:", nb[s[i:i+3]])
```

```
Observed: C Predicted: < B=0.00 C=0.51 E=0.20 G=0.00 H=0.29 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.43 E=0.20 G=0.00 H=0.37 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.43 E=0.18 G=0.00 H=0.39 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.38 E=0.12 G=0.00 H=0.50 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.28 E=0.28 G=0.00 H=0.44 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.25 E=0.38 G=0.00 H=0.37 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.36 E=0.24 G=0.00 H=0.40 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.29 E=0.16 G=0.00 H=0.55 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.25 E=0.14 G=0.00 H=0.61 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.32 E=0.13 G=0.00 H=0.55 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.27 E=0.17 G=0.00 H=0.56 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.35 E=0.22 G=0.00 H=0.43 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.30 E=0.33 G=0.00 H=0.37 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.23 E=0.27 G=0.00 H=0.50 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.20 E=0.17 G=0.00 H=0.63 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.73 E=0.09 G=0.00 H=0.18 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.54 E=0.06 G=0.00 H=0.39 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.42 E=0.23 G=0.00 H=0.35 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.16 E=0.27 G=0.00 H=0.57 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.25 E=0.23 G=0.00 H=0.52 I=0.00 S=0.00 T=0.00 >
Observed: H Predicted: < B=0.00 C=0.42 E=0.11 G=0.00 H=0.46 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.64 E=0.10 G=0.00 H=0.26 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.61 E=0.11 G=0.00 H=0.28 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.85 E=0.09 G=0.00 H=0.06 I=0.00 S=0.00 T=0.00 >
Observed: C Predicted: < B=0.00 C=0.81 E=0.06 G=0.00 H=0.13 I=0.00 S=0.00 T=0.00 >
```

Classification metrics: multi-state confusion matrix

	Helix prediction	Strand prediction	Coil prediction
Helix observation	523	131	83
Strand observation	42	509	102
Coil observation	31	23	389

Classification metrics: Q_k

- **Accuracy** of k -class problems

$$Q_k = \frac{\sum_{j=1}^k tp(j)}{\sum_{j=1}^k tp(j) + fn(j)} \cdot 100$$

We use the quantities *true positives*, $tp(C)$, which is the number of test samples in class C predicted to be in class C , *true negatives*, $tn(C)$, which is the number of test samples not in class C predicted not to be in class C , *false negatives*, $fn(C)$, which is the number of test samples in class C predicted not to be in class C , and *false positives*, $fp(C)$, which is the number of test samples not in class C predicted to be in class C .

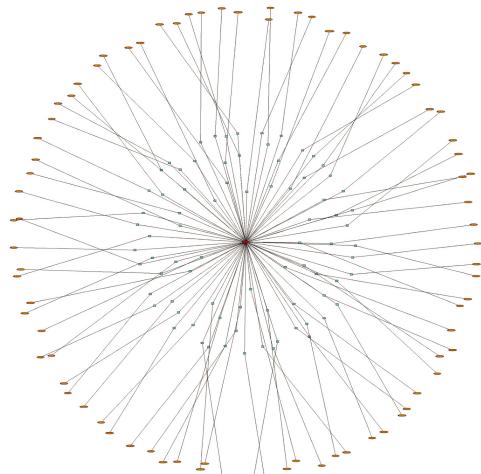
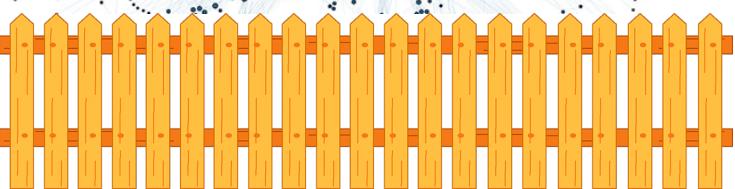
- When $k=3$: accuracy for three-class prediction problem (e.g., Helix, Beta, Coil)
- Percentage of correct secondary structure class predictions

$$Q_3 = \frac{\text{number of residues correctly predicted}}{\text{total number of residues}}$$

VLHQASGNSQLFGSDVTVPGATNAEQAR amino acid sequence 29 residues long
HHHHHCCCCEEEECCCEECCCCHHHHHH actual secondary structure
CHHHCCCCEEEECCCCCCEECCCHHHHHH prediction 1: $Q_3 = 22/29 = 76\%$: useful
HHHHHCCCCHHHHCCCHHHCCCCCHHHHHH prediction 2: $Q_3 = 22/29 = 76\%$: terrible



Proteins are part of a biological system

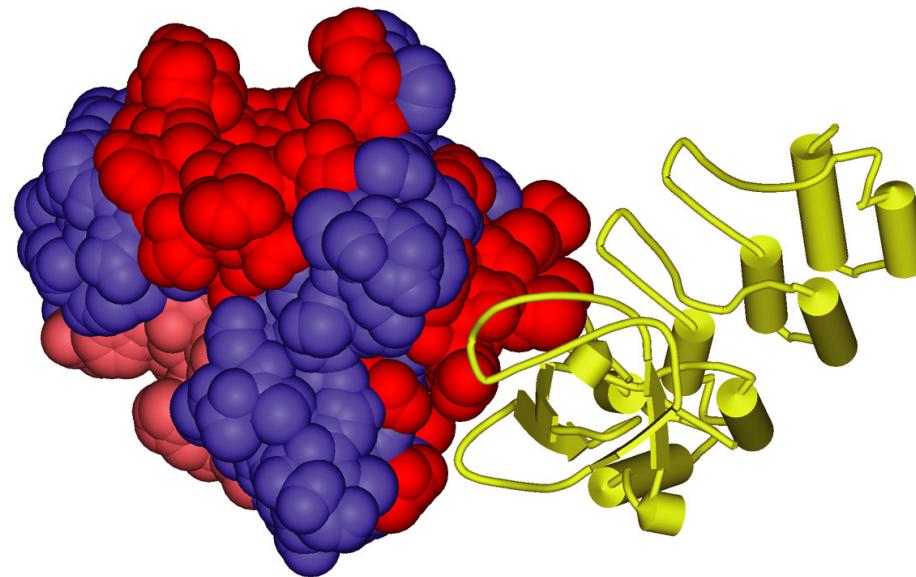


Episode in the series on protein bioinformatics

Function, interaction, localisation and systems of proteins

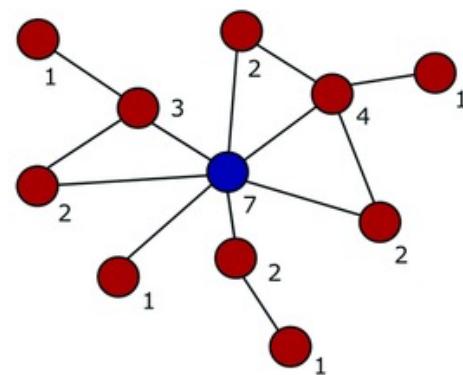
- interpreting sites in protein sequence and structure that are relevant to system role,
- connecting components in ways that biological evidence indicate, and
- analysing patterns of connectivity—biological networks and pathways

Example: p53 and its binding domains

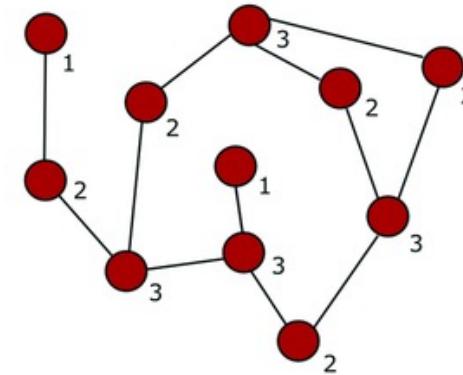


Biological networks are not random

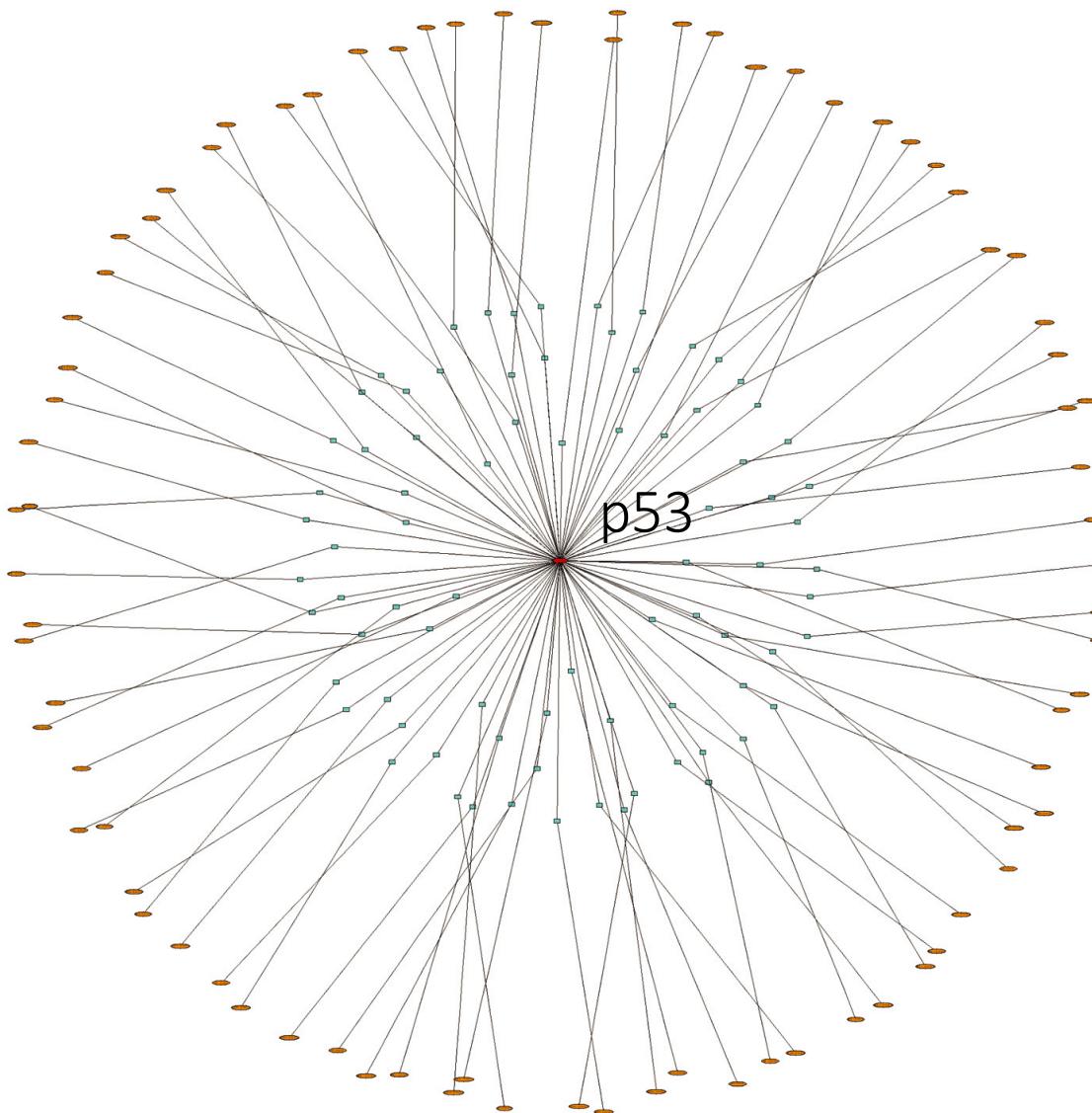
A Scale-free network



B Random network

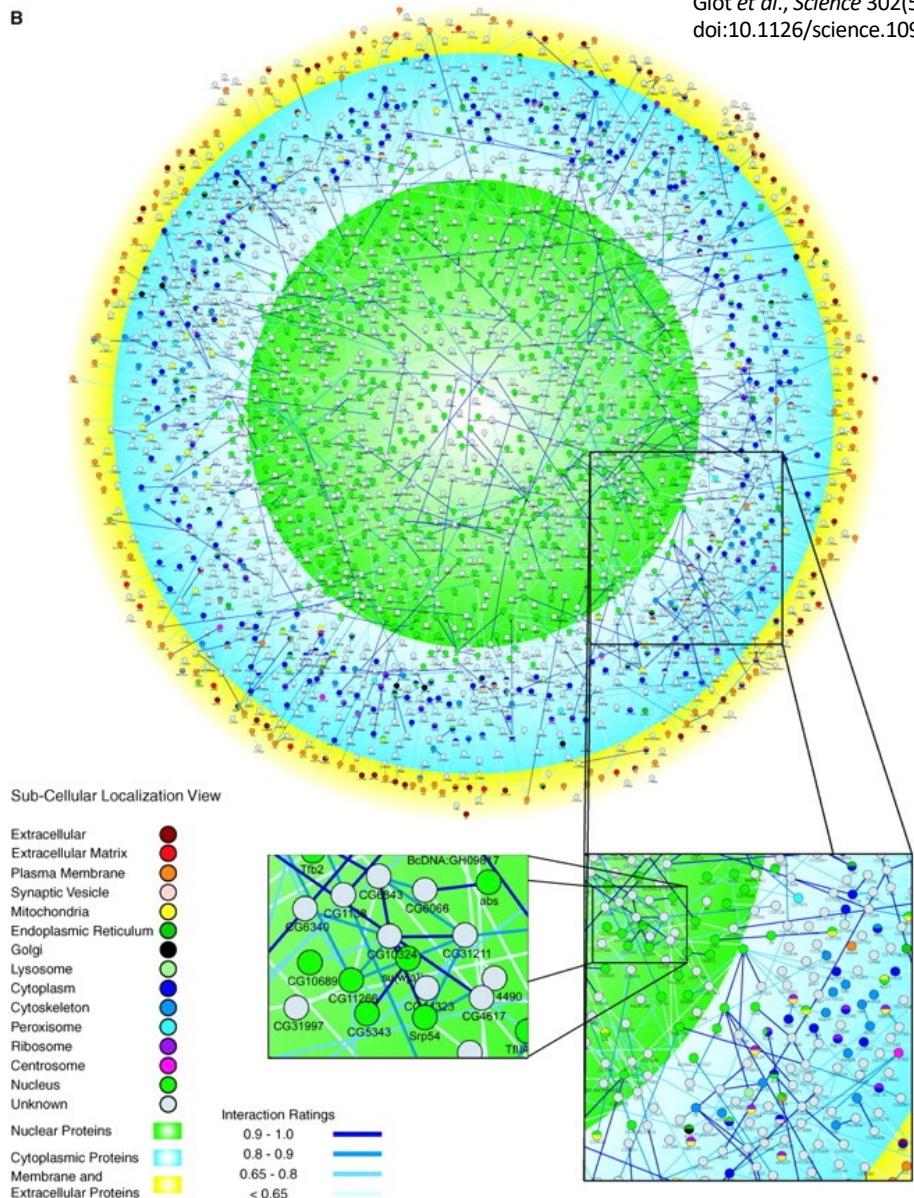


Networks are an abstraction but have properties

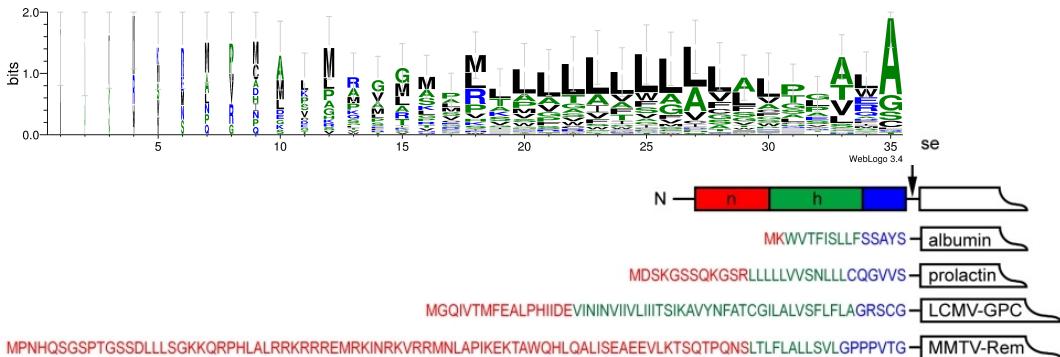


B

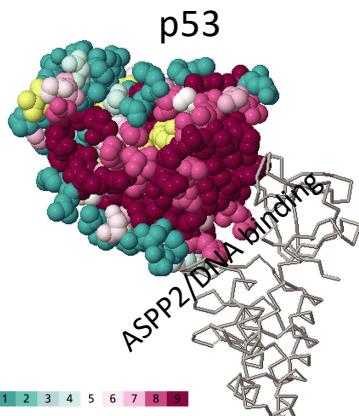
Giot et al., *Science* 302(5651)
doi:10.1126/science.1090289



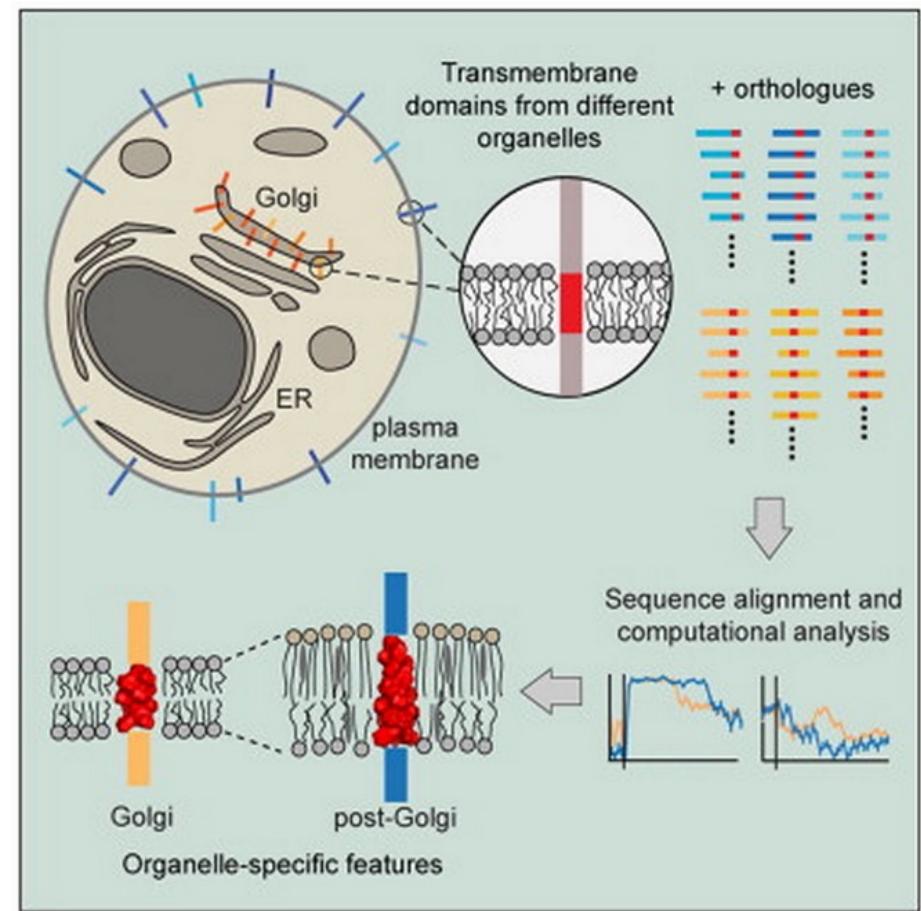
- Most analyses are focused on individual components and their properties—in spatial and temporal isolation
- Molecular function is often determined from properties of *multiple* components
- Biological processes are supported by the *interaction* and *abundance* of components
- How do we bring this information together?



Protein localisation signals (here signal peptides)



- Conservation (maybe spread in sequence but come together in structure)
- Many same-charged residues (electro-static interaction)
- Hydrophobic patch (unusual at surface; interaction by hydrophobic forces)
- Aromatic residues
- Mainly β sheets or long loops



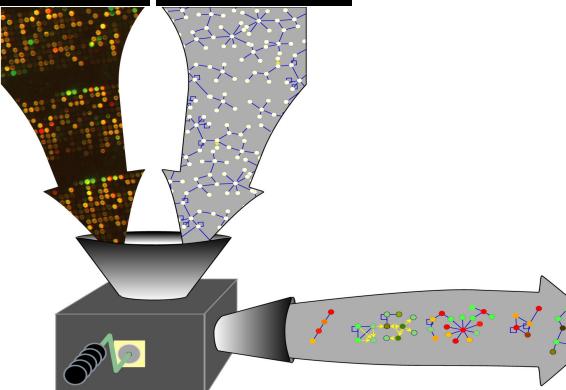
Expression @ condition

	X_1	X_2	...	X_m
G_1	+			+
G_2	+	+		
G_3		+		
G_4	+	+		
...	...			
G_n				

Co-expression

	G_1	G_2	...	G_n
G_1	+			
G_2	+	+		
G_3		+	+	
G_4	+	+		+
...	...			
G_n				

Expression Interaction

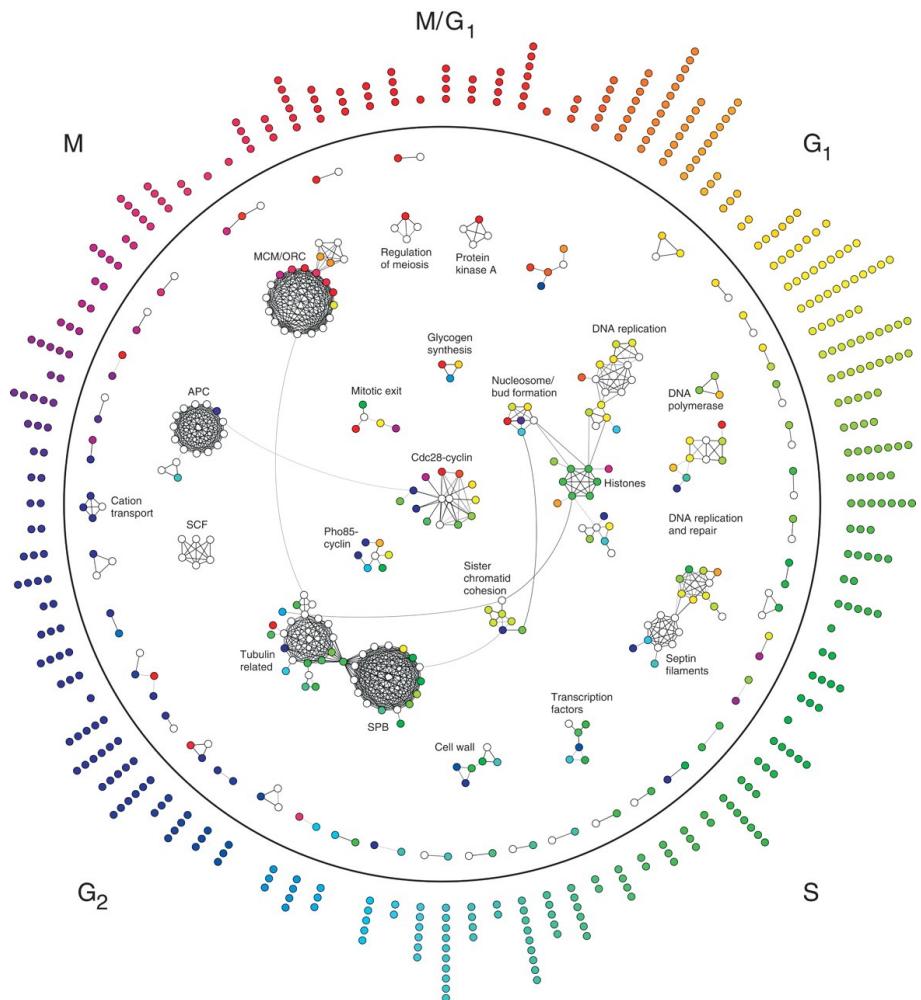
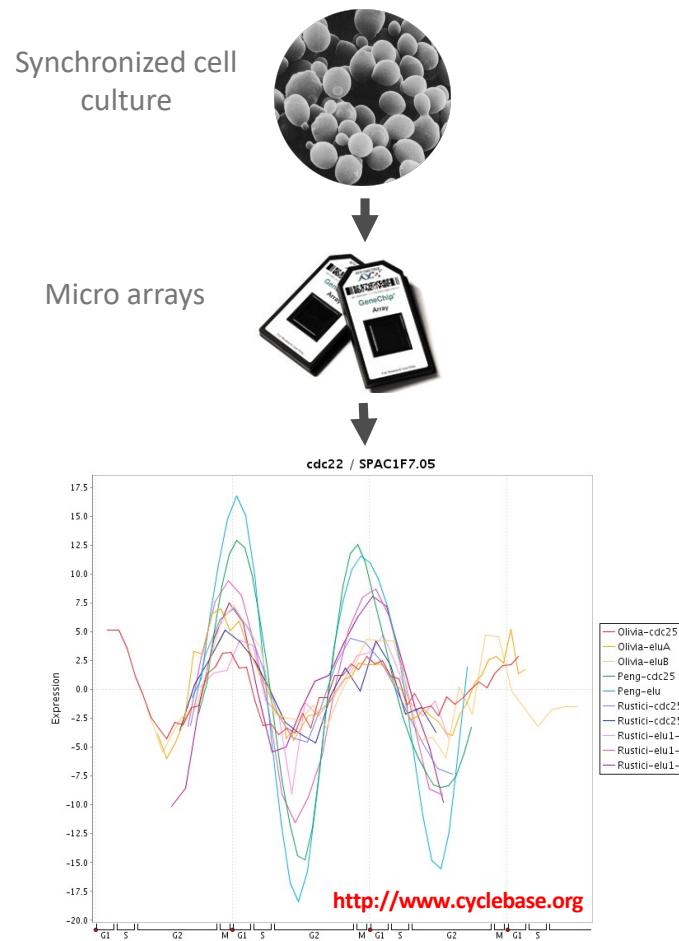


Biological system

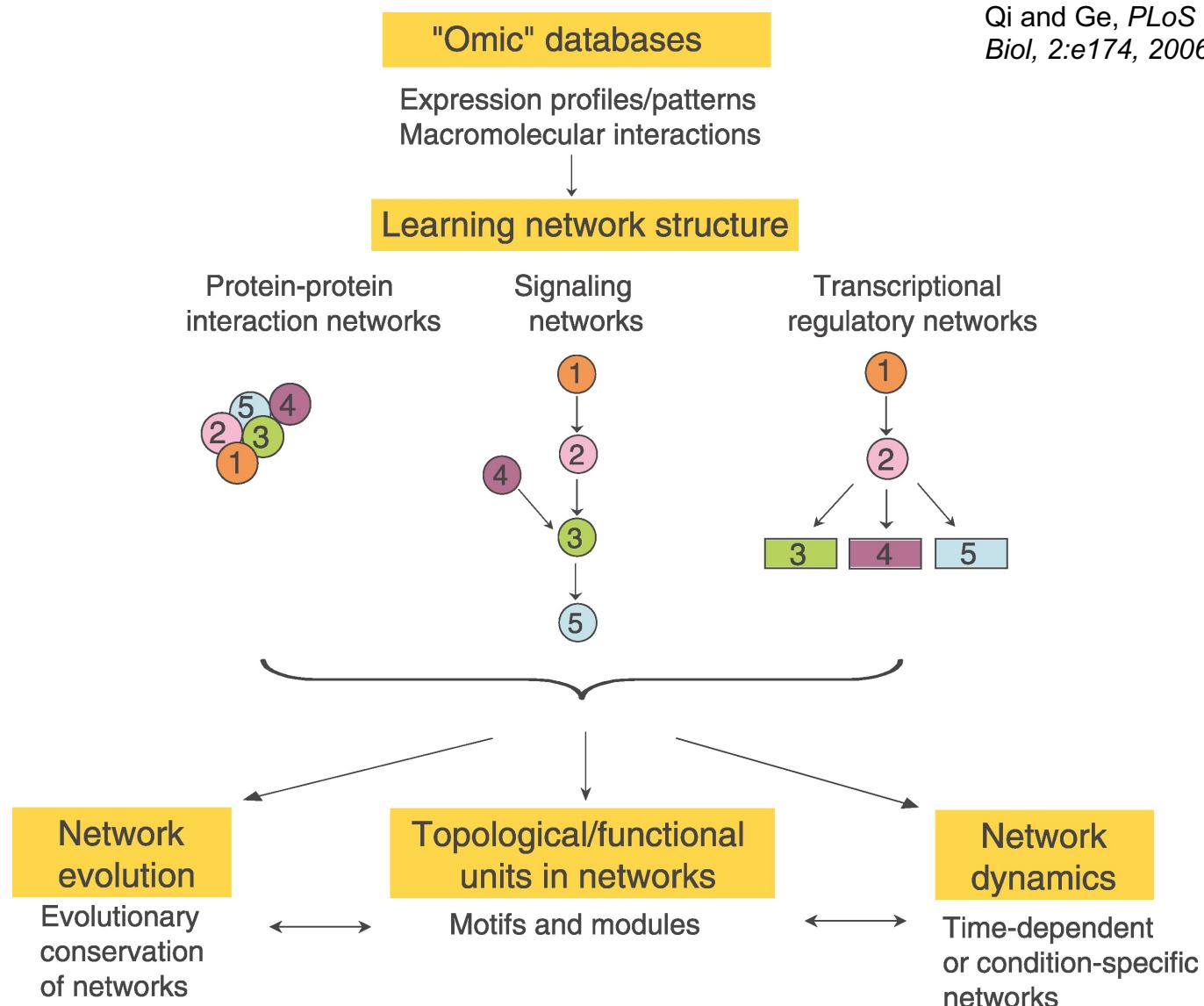
Interaction

	G_1	G_2	...	G_n
G_1				
G_2	Yes			
G_3				
G_4		Yes		
...	...			
G_n				

Creating a yeast cell cycle interactome



Qi and Ge, *PLoS Comp Biol*, 2:e174, 2006.



Protein bioinformatics: week 12

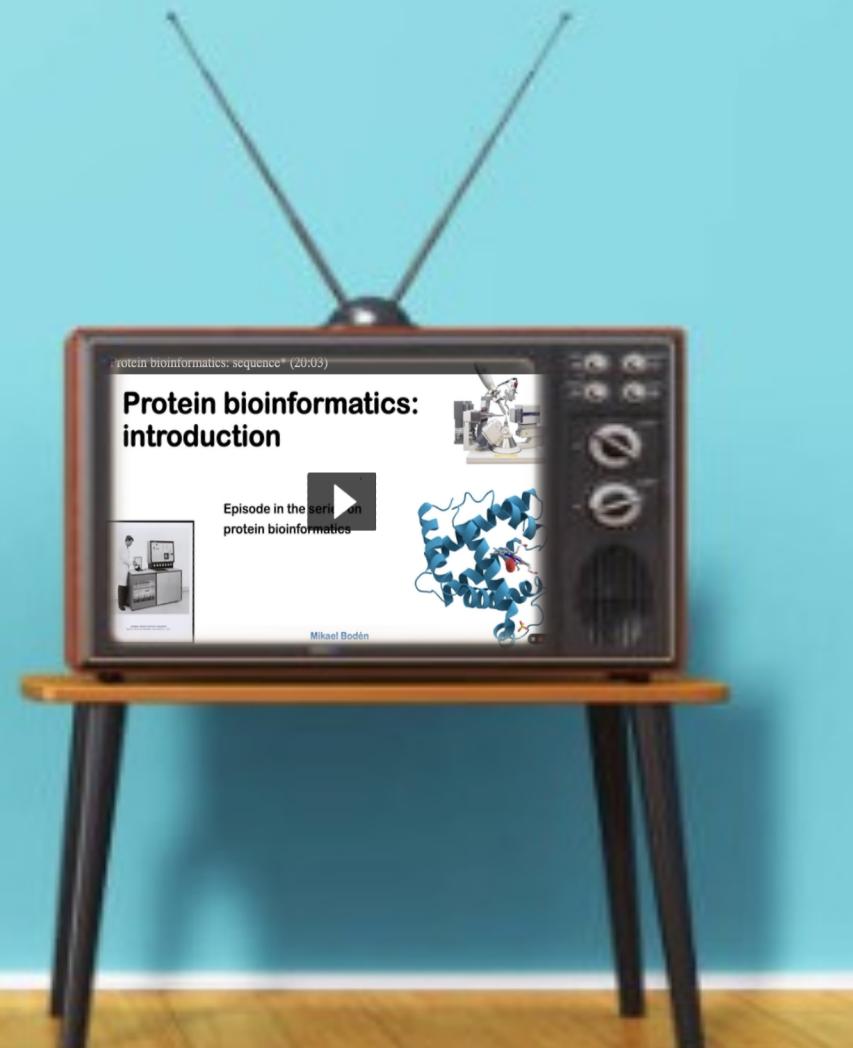
Watch the recordings

Introduction (3 parts)

- Protein sequence, structure, prediction and metrics

Systems (1 part)

- Protein function, localization, interaction, biological networks and data integration



Mikael Bodén

Biological networks

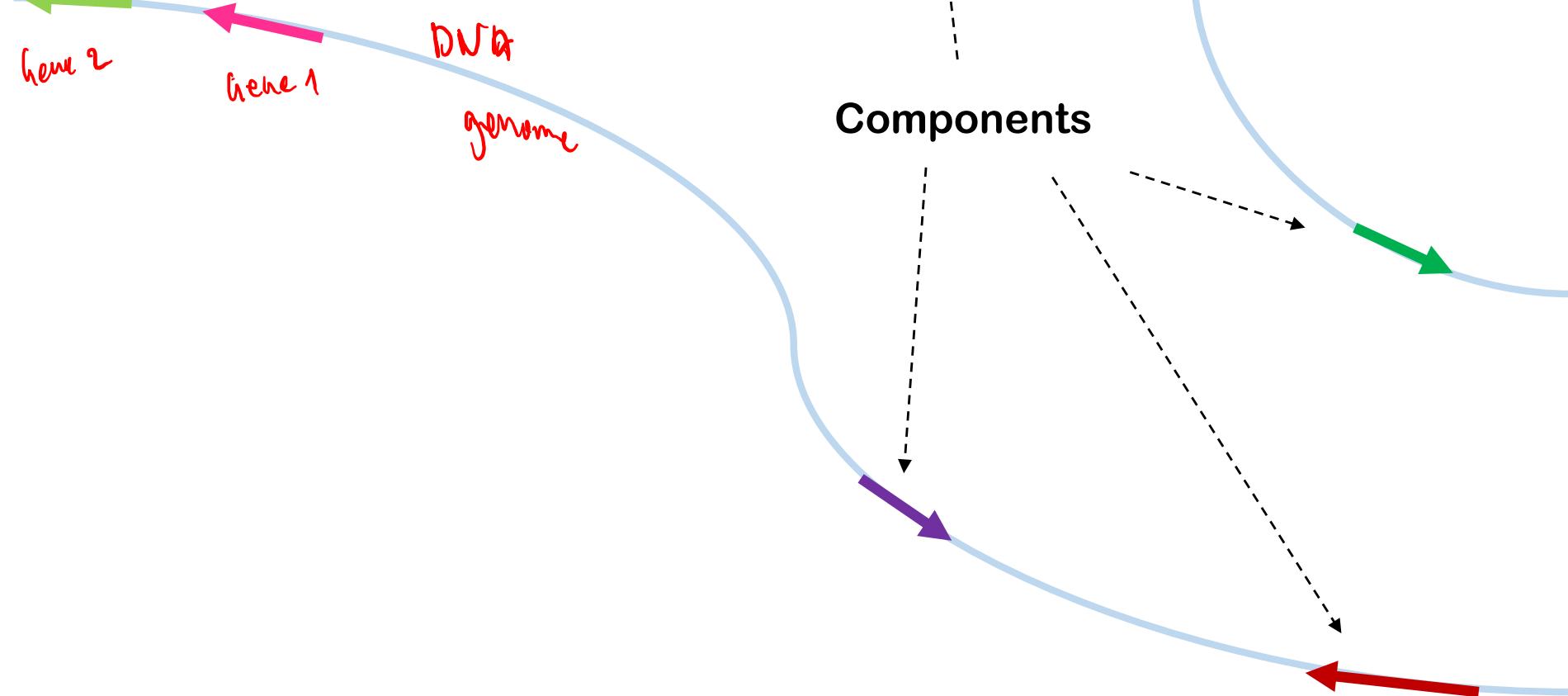
- Genome-wide assays
 - contain a catalog of “system components” (each data set)
 - indicate “relationships” (data sets jointly)
can be used to model “biological systems”
- Biological networks are used to represent
 - physical interactions and complexes
 - protein-protein, protein-DNA, protein-RNA, RNA-DNA, ...
 - permanent and transient interactions
 - cellular pathways (logical interactions)
 - metabolic/enzyme reaction
 - transcriptional/regulatory networks
- Biological networks can
 - integrate data types
 - Expression, co-expression
 - Subcellular localization
 - ...

What is a regulatory network?

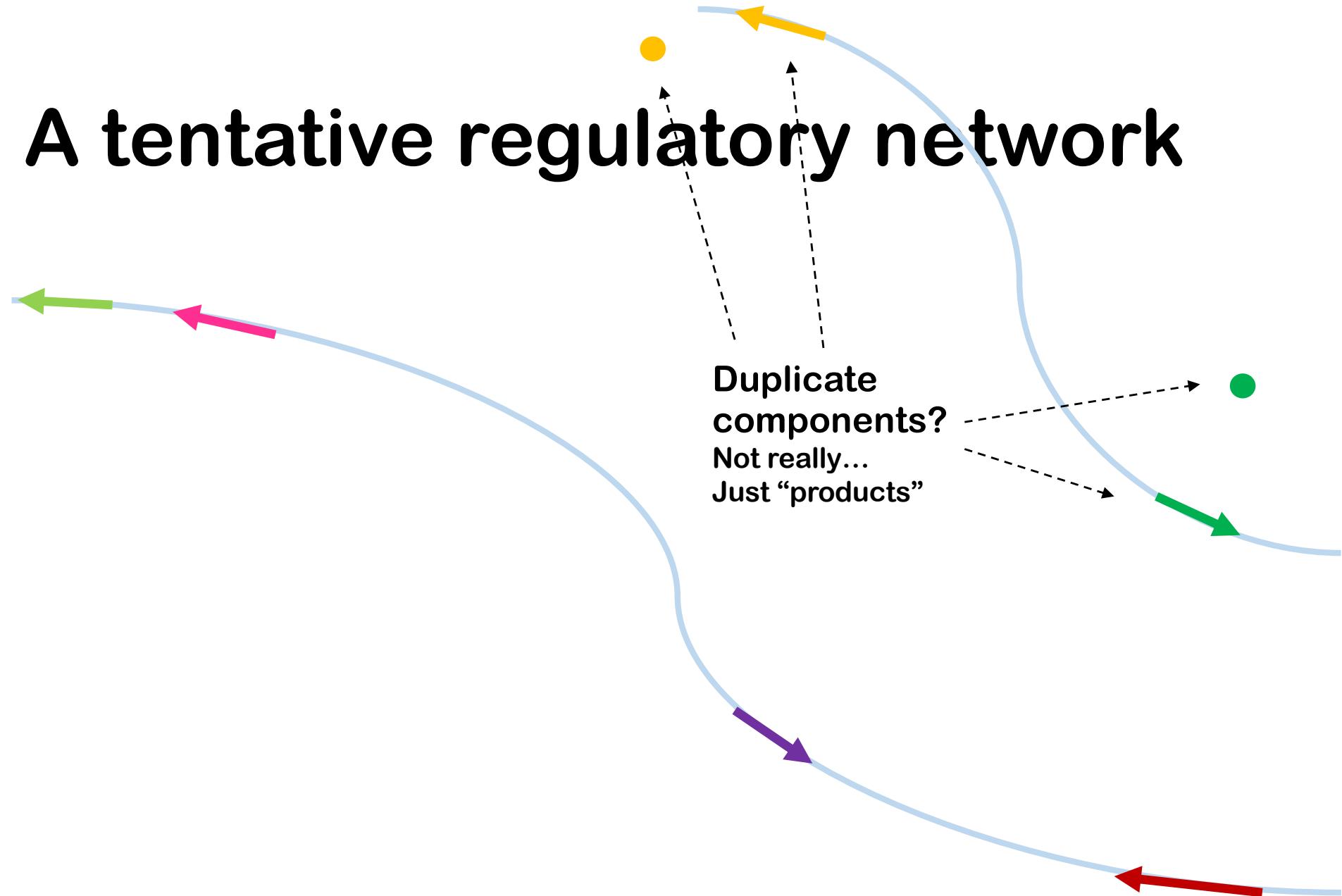
- A. Components are genes, they are linked if they share function
- B. Components are proteins, they are linked if they physically interact
- C. Components are RNA, they are linked if they bind to mRNA
- D. Components are genes, they are linked if the product of gene 1 binds and modulates gene 2's expression
- E. Components are proteins, they are linked if protein 1 is facebook friends with protein 2

Note: multiple options may be true

A tentative regulatory network



A tentative regulatory network

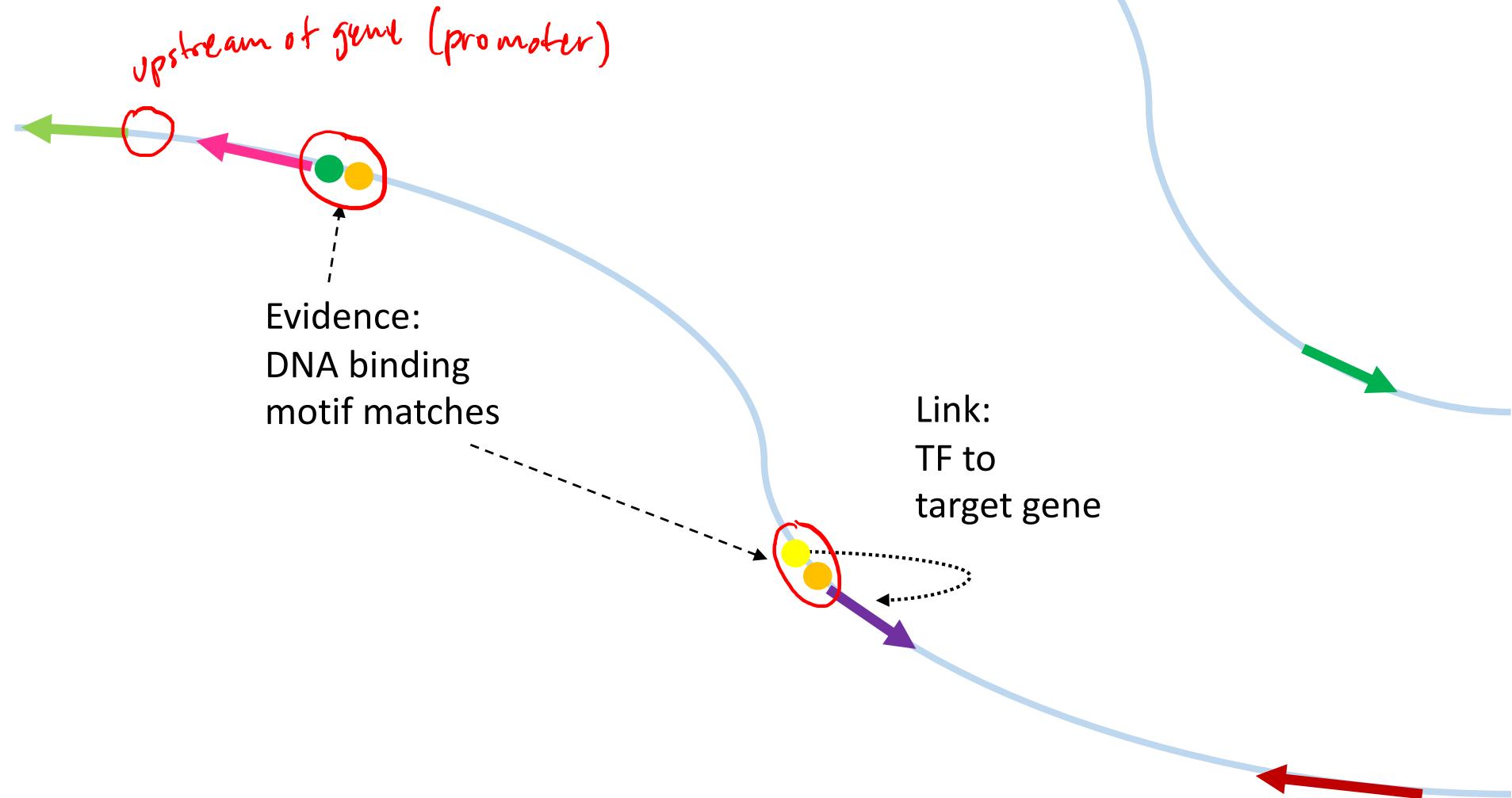


What evidence/data can we use to link the genes?

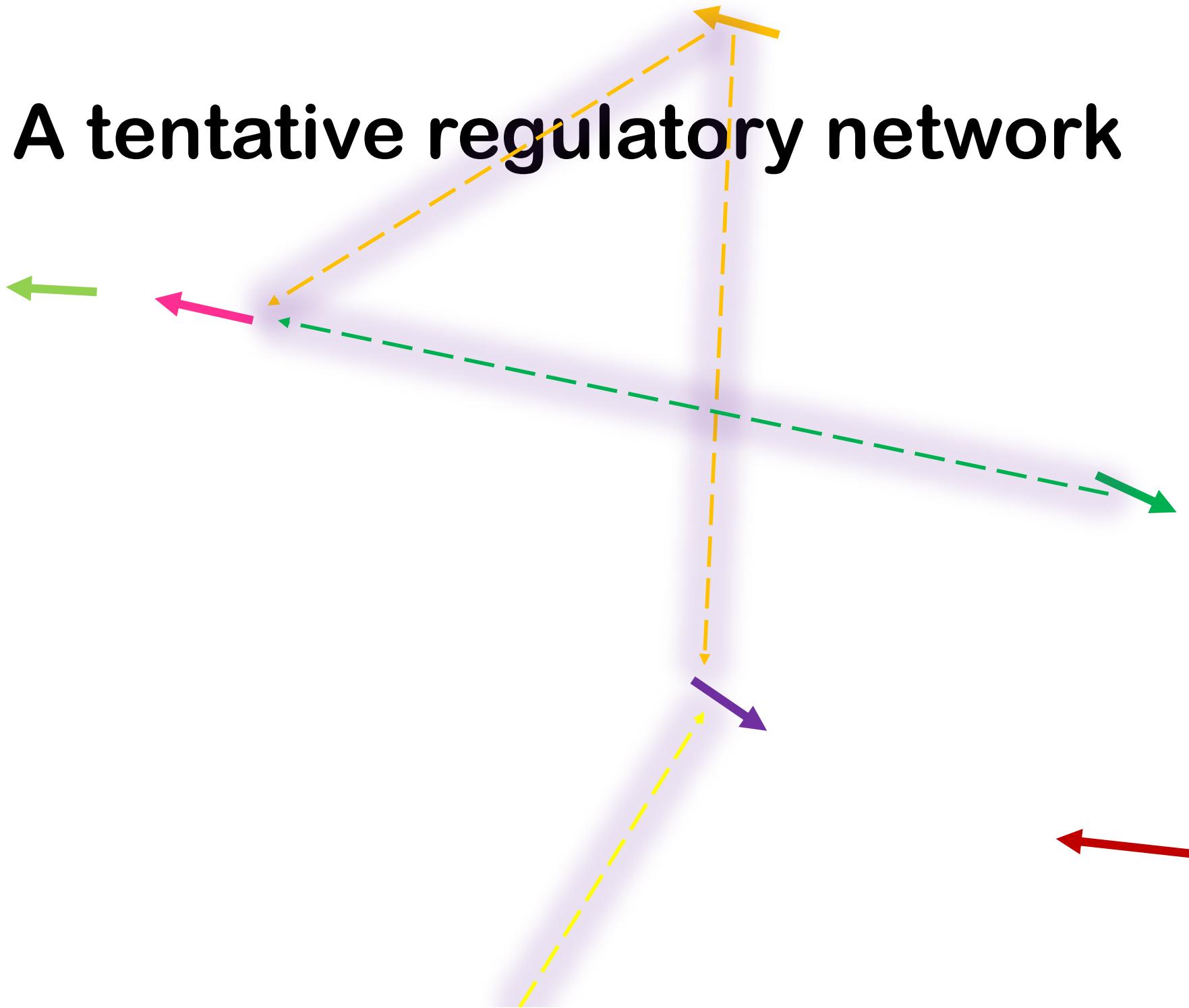
- A. Co-expression of gene
- B. Protein-protein interaction (pairs of gene products)
- C. Co-localization of proteins (pairs of gene products)
- D. Search signalling pathway database for source protein to target gene association
- E. Transcription factor DNA binding motif at target gene promoter

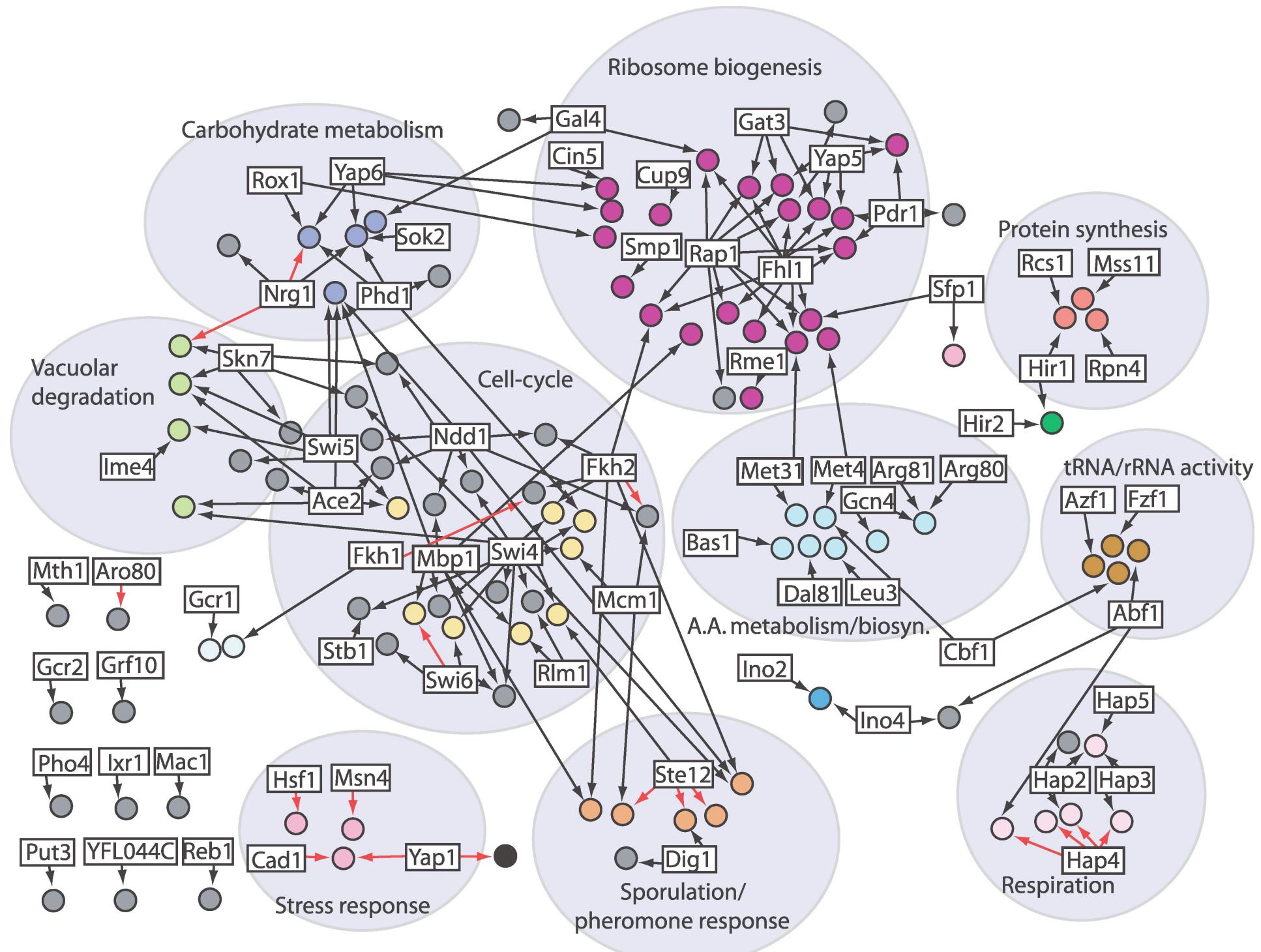
Note: multiple options may be true

A tentative regulatory network



A tentative regulatory network



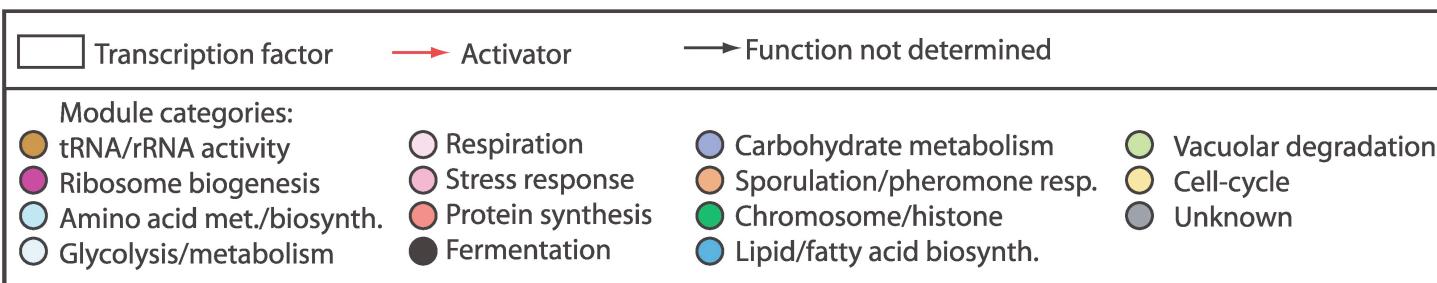
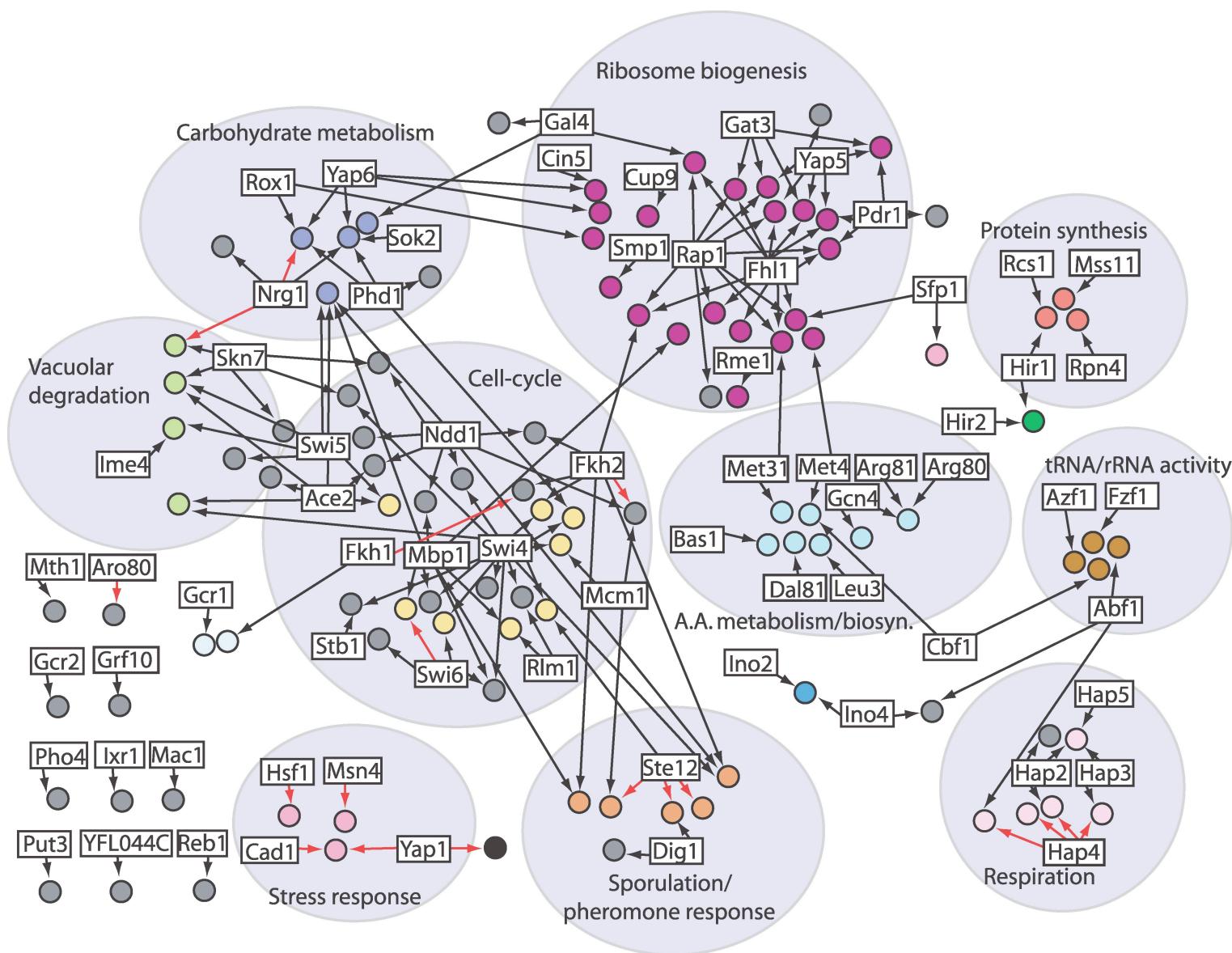


Expression @ condition

	x_1	x_2	...	x_m
G_1	✓	✓	✗	✗
G_2	✗	✗	✓	✗
G_3	✓	✓	✗	✗
G_4				
...			...	
G_n				

Co-expression

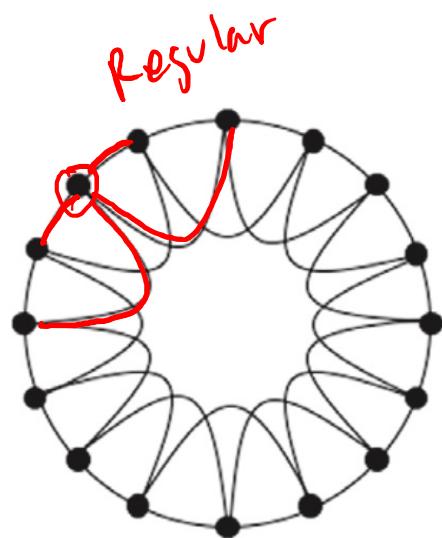
	G_1	G_2	...	G_n
G_1	l			
G_2	-l			
G_3	l			
G_4				
...			...	
G_n				



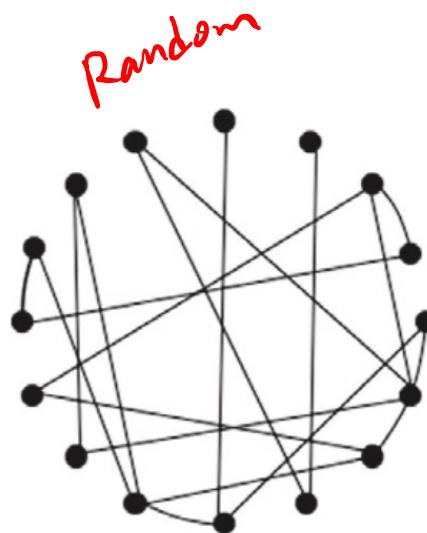
Yeast
TF/DNA
binding
motifs
combined
with
expression
data

- Modular
- Function specific structure
- Robust and redundant
- Predictive

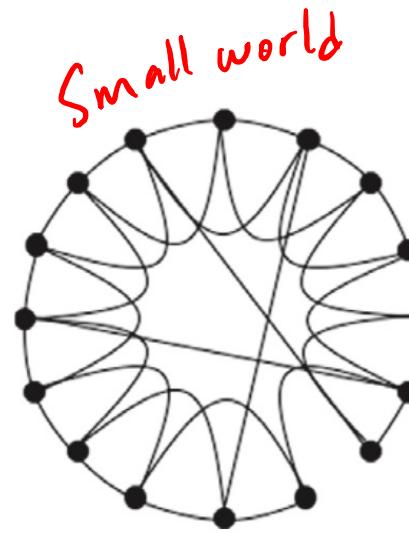
Reproduced from Bar-Joseph et al., *Nat Biotechnol* 21:1337-1342 (2003)



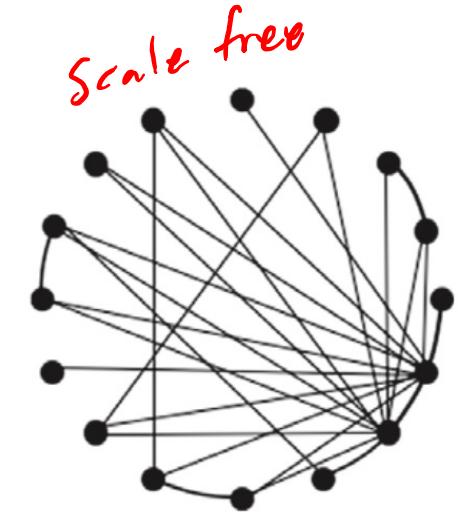
A



B



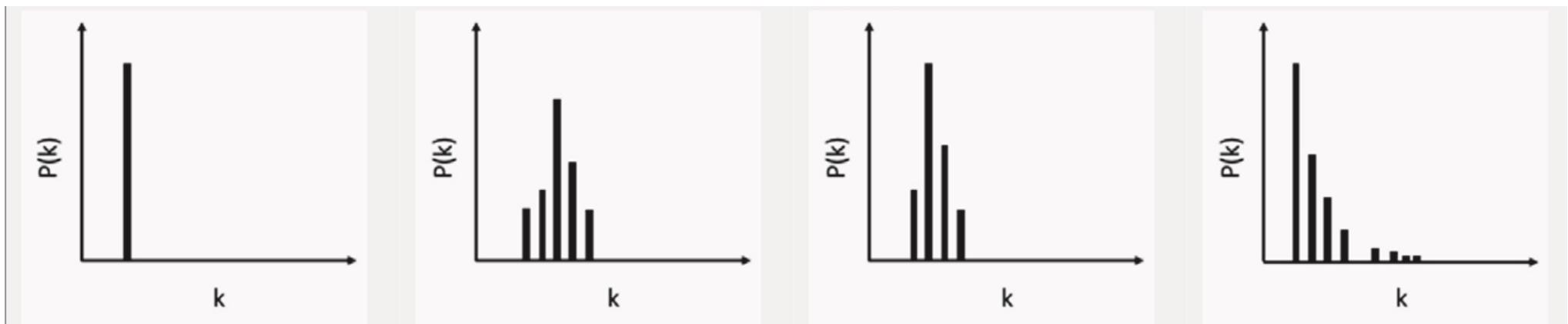
C



D

Which network has a topology of a typical biological network?

What is the degree distribution of a typical biological network?



A

B

C

D

Questions so far?

About biological networks and systems

Final exam 2018: Protein bioinformatics



The Chou-Fasman propensity (P) values for each amino acid to form an α -helix or a β -strand, respectively is shown in the table (right). A larger value denotes higher propensity.

Based on your understanding of the formation of protein secondary structure and the propensity table, predict the secondary structure class (**α -helix**, **β -strand** or **coil**) at the highlighted position of the following amino acid sequences. You do not need to use Chou-Fasman's algorithm, but similar principles should apply in determining your answers.

(a) His – Lys – Glu – Ile – Cys – Leu – Pro
– Ile – Val – Phe – Lys – Asp

...

(d) Justify the predictions for (a)-(c) by explaining what the table is based on and the strategy with which predictions were made. Calculations are not required.

Amino acid	α -helix		β -strand	
	Designation	P	Designation	P
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74
Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

(a)

His – Lys – Glu – Ile – Cys – Leu – Pro – Ile – Val – Phe – Lys – Asp

α
 β

0.70 1.21 0.57 1.08 1.06 1.13
1.19 1.30 0.55 1.60 1.70 1.38

Amino acid	α -helix		β -strand	
	Designation	P	Designation	P
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74

Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

(a)

His – Lys – Glu – Ile – Cys – Leu – **Pro** – Ile – Val – Phe – Lys – Asp
 1.00 1.16 1.51 1.08 0.70 1.21 0.57 1.08 1.06 1.13 1.16 1.01
 0.87 0.74 0.37 1.60 1.19 1.30 0.55 1.60 1.70 1.38 0.74 0.54

Amino acid	α -helix		β -strand	
	Designation	P	Designation	P
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74

Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

(b)

Arg – Pro – Met – Ala – Lys – Thr – Gln – Ala – Phe – Cys – Gly

0.98 0.57 1.45 1.42 1.16 0.83 1.11 1.42 1.13 0.70 0.61

0.93 0.55 1.05 0.83 0.74 1.19 1.10 0.83 1.38 1.19 0.75

Amino acid	α-helix		β-strand	
	Designation	P	Designation	P
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74

Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

(c)

Pro – Gly – Cys – His – Pro – Ser – Tyr – Ala

0.57 0.61 0.70 1.00 0.57 0.77 0.69 1.42

0.55 0.75 1.19 0.87 0.55 0.75 1.40 0.83

Amino acid	α-helix		β-strand	
	Designation	P	Designation	P
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74

Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	B	0.57	B	0.55
Gln	f	1.11	f	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	B	0.69	F	1.4

The correct answers were (a) beta, (b) alpha, (c) coil

What *sensitivity* to the class alpha helix do I get by predicting *alpha helix* for all three questions?

- (a) observing not α predicting α FP
- (b) observing α predicting α TP
- (c) observing not α predicting α FP

- Sensitivity
$$\frac{tp}{tp + fn}$$
- Specificity
$$\frac{tn}{tn + fp}$$

↓
Two decimal places multiplied by 100

e.g. 100 50 33 0
correct

The correct answers were (a) *beta*, (b) *alpha*, (c) *coil*

What *specificity* to the class *alpha helix* do I get by predicting *alpha helix* for all three questions?

(a) observing not α
predicting α

FP

(b) observing α
predicting α

TP

(c) observing not α
predicting α

FP

• Sensitivity

$$\frac{tp}{tp + fn}$$

• Specificity

$$\frac{tn}{tn + fp} \quad \frac{0}{0+2}$$

multiply by 100

e.g.

100 50 33 0

The correct answers were (a) *beta*, (b) *alpha*, (c) *coil*

What *accuracy* (Q_3) do I get by predicting *alpha helix* for all three questions?

	Helix prediction	Strand prediction	Coil prediction
Helix observation	1	0	0
Strand observation	1	0	0
Coil observation	1	0	0

$$Q_k = \frac{\sum_{j=1}^k tp(j)}{\sum_{j=1}^k tp(j) + fn(j)} \cdot 100$$

0.33 $\frac{1}{3}$

Questions so far?

About prediction, protein sequence and structure