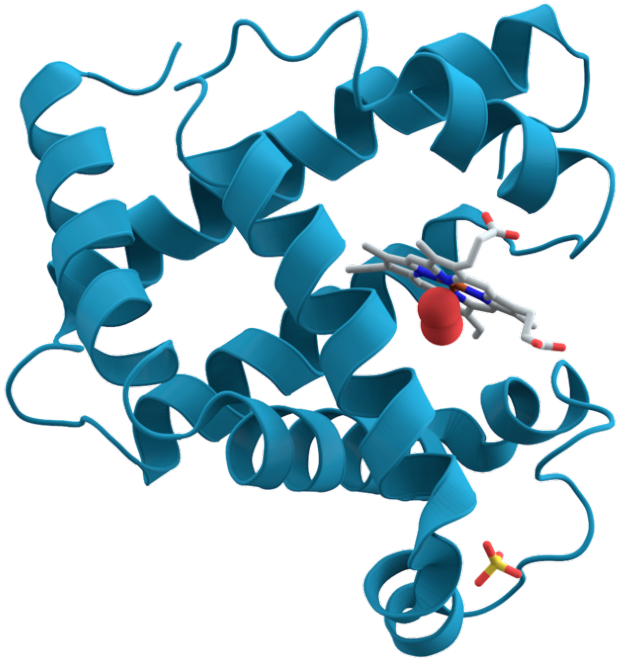
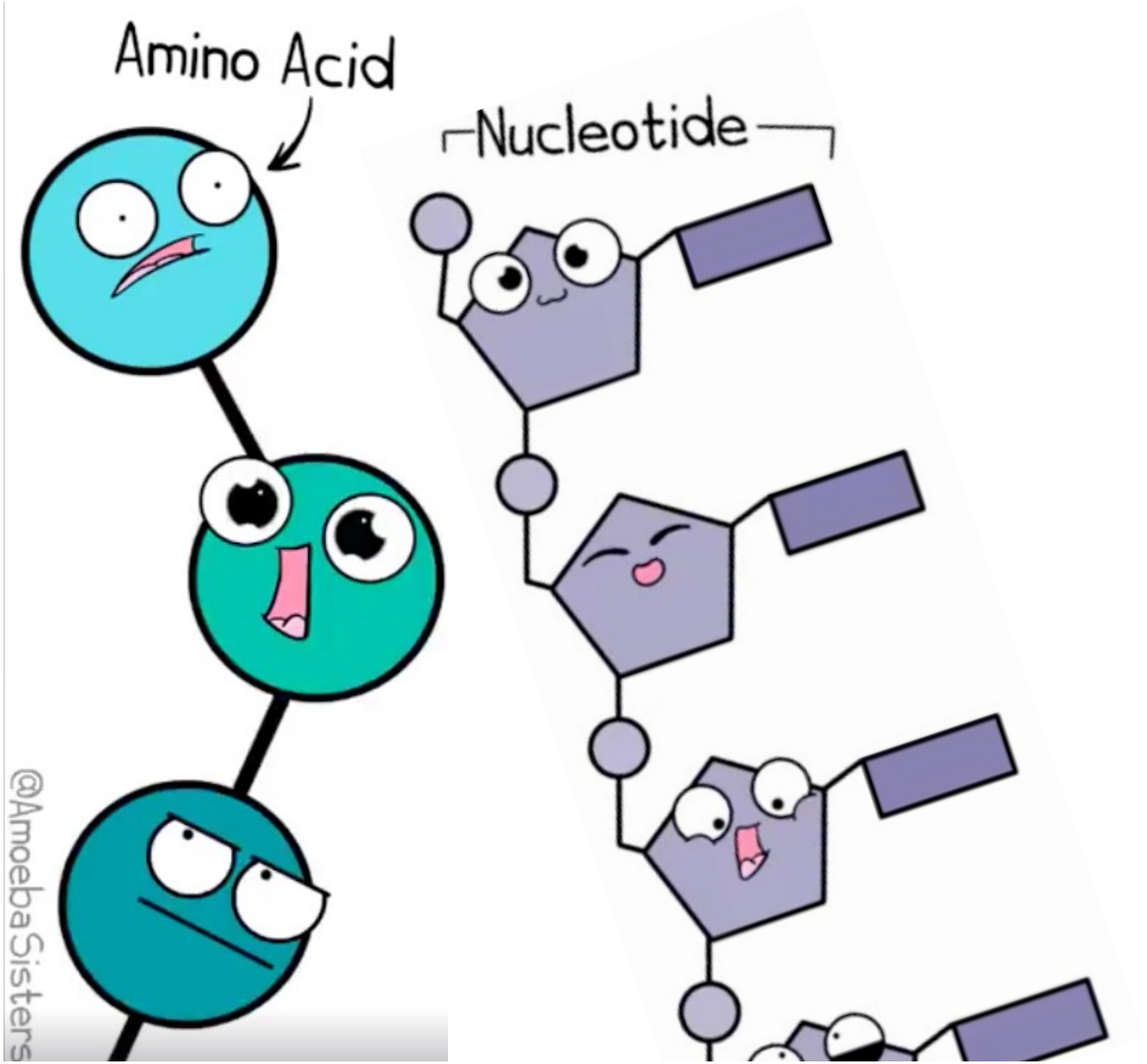
Protein bioinformatics: introduction

Episode in the series on protein bioinformatics



Mikael Bodén



Part 1: Protein sequence

Motivation

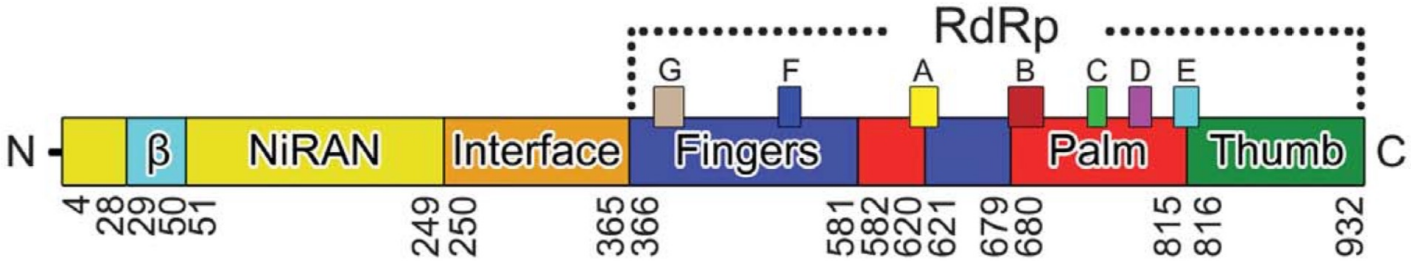
Amino acids

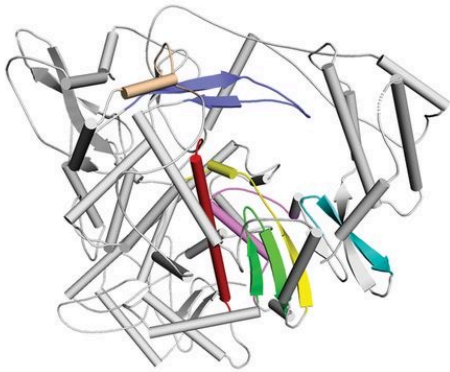
Sequence to structure

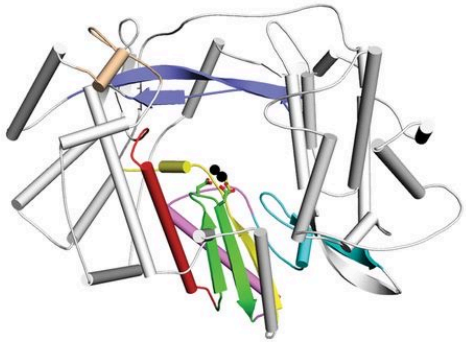
CORONAVIRUS *Science* **368**, 779–782 (2020) 15 May 2020

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

Yan Gao[1,2]*, Liming Yan[1]*, Yucen Huang[1]*, Fengjiang Liu[2]*, Yao Zhao[2], Lin Cao[3], Tao Wang[1], Qianqian Sun[2], Zhenhua Ming[4], Lianqi Zhang[1], Ji Ge[1], Litao Zheng[1], Ying Zhang[1], Haofeng Wang[2,5], Yan Zhu[2], Chen Zhu[2], Tianyu Hu[2], Tian Hua[2], Bing Zhang[2], Xiuna Yang[2], Jun Li[2], Haitao Yang[2], Zhijie Liu[2], Wenqing Xu[2], Luke W. Guddat[6], Quan Wang[2]†, Zhiyong Lou[1]†, Zihe Rao[1,2,3,7]†







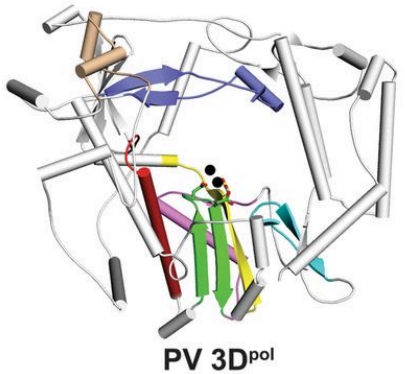
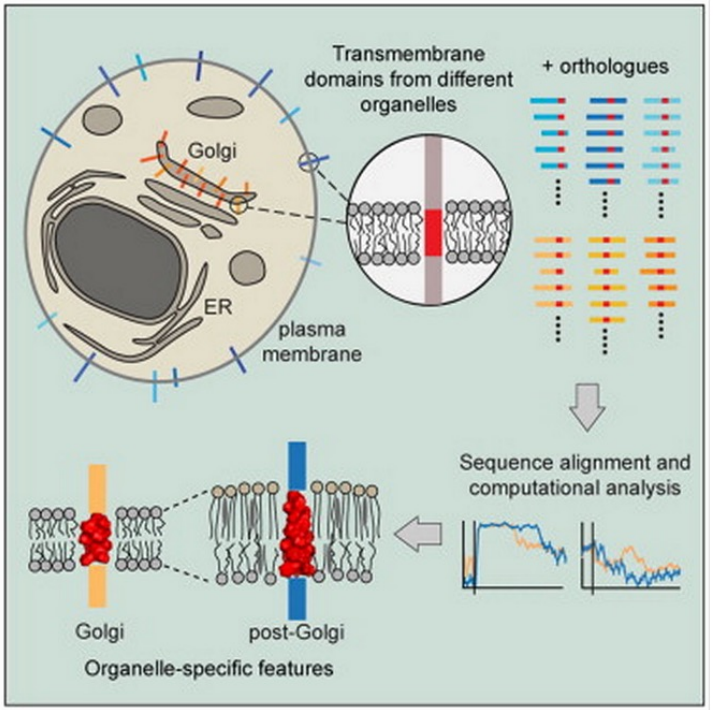


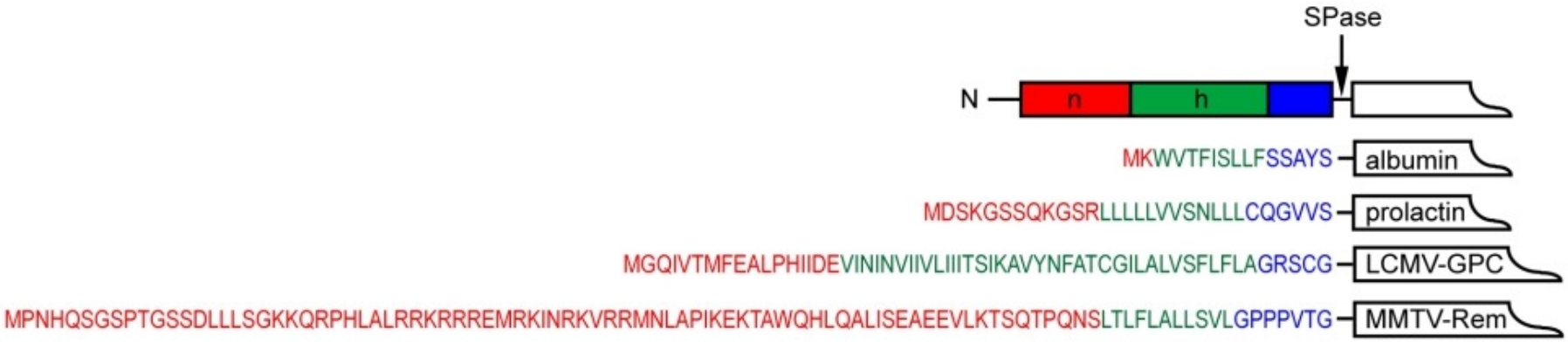
Fig. 1. Structure of COVID-19 virus nsp12

(**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.

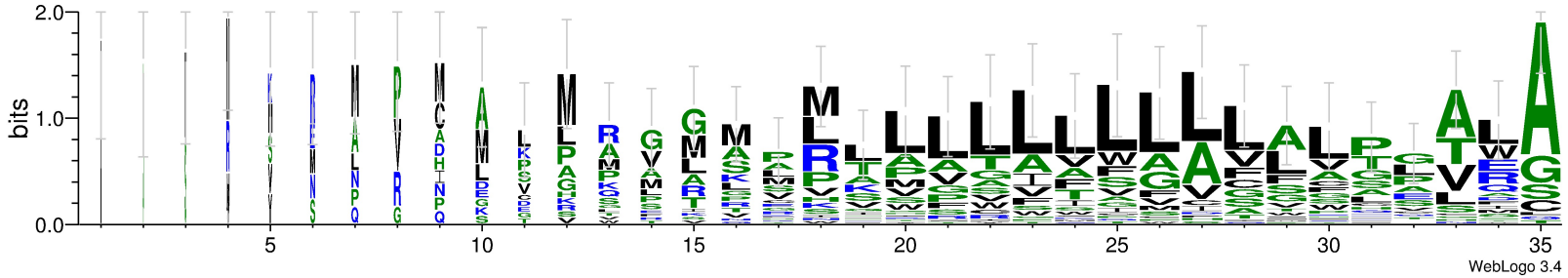
The secretory pathway

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

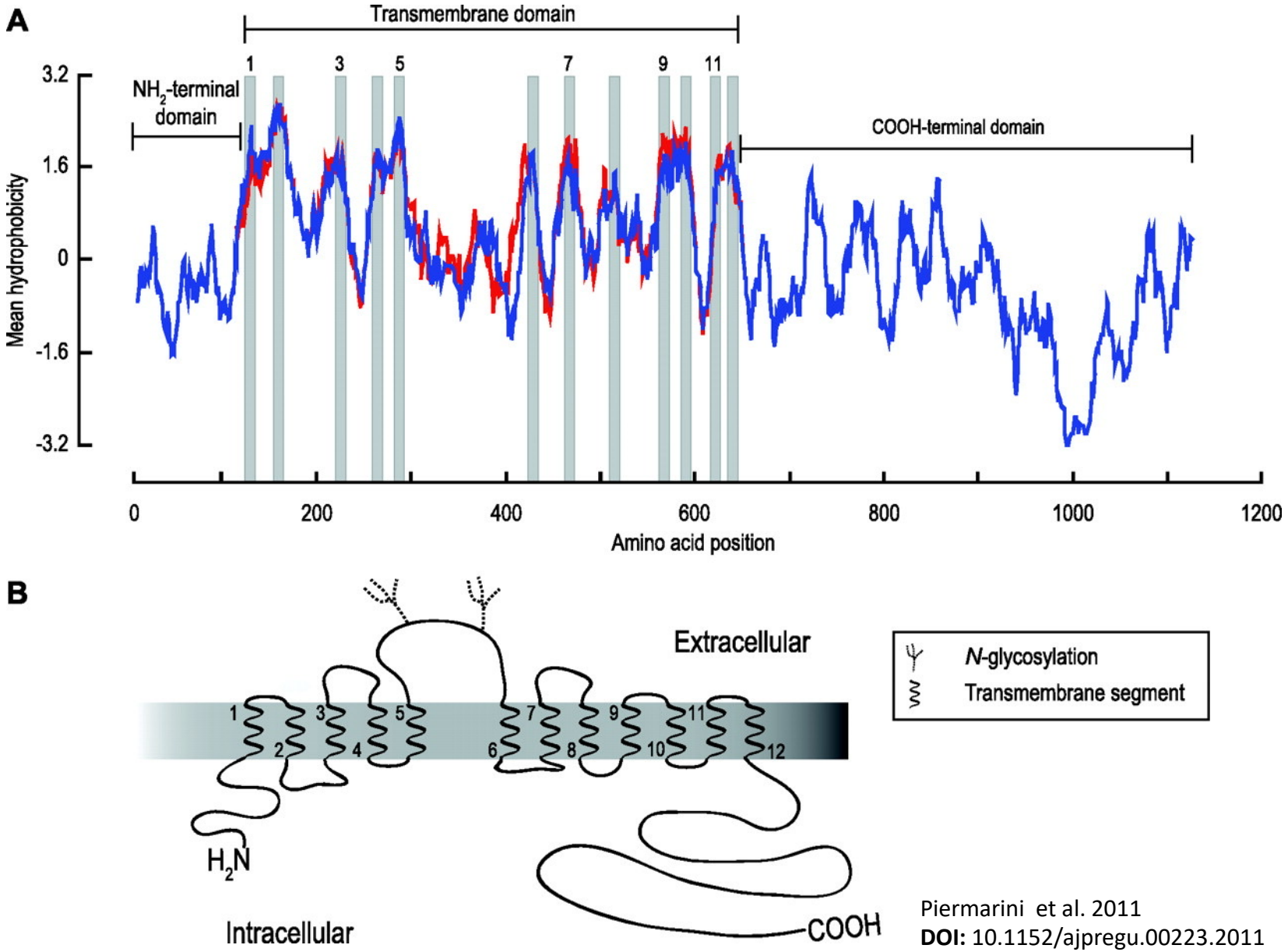




Sequence logo of human signal peptides

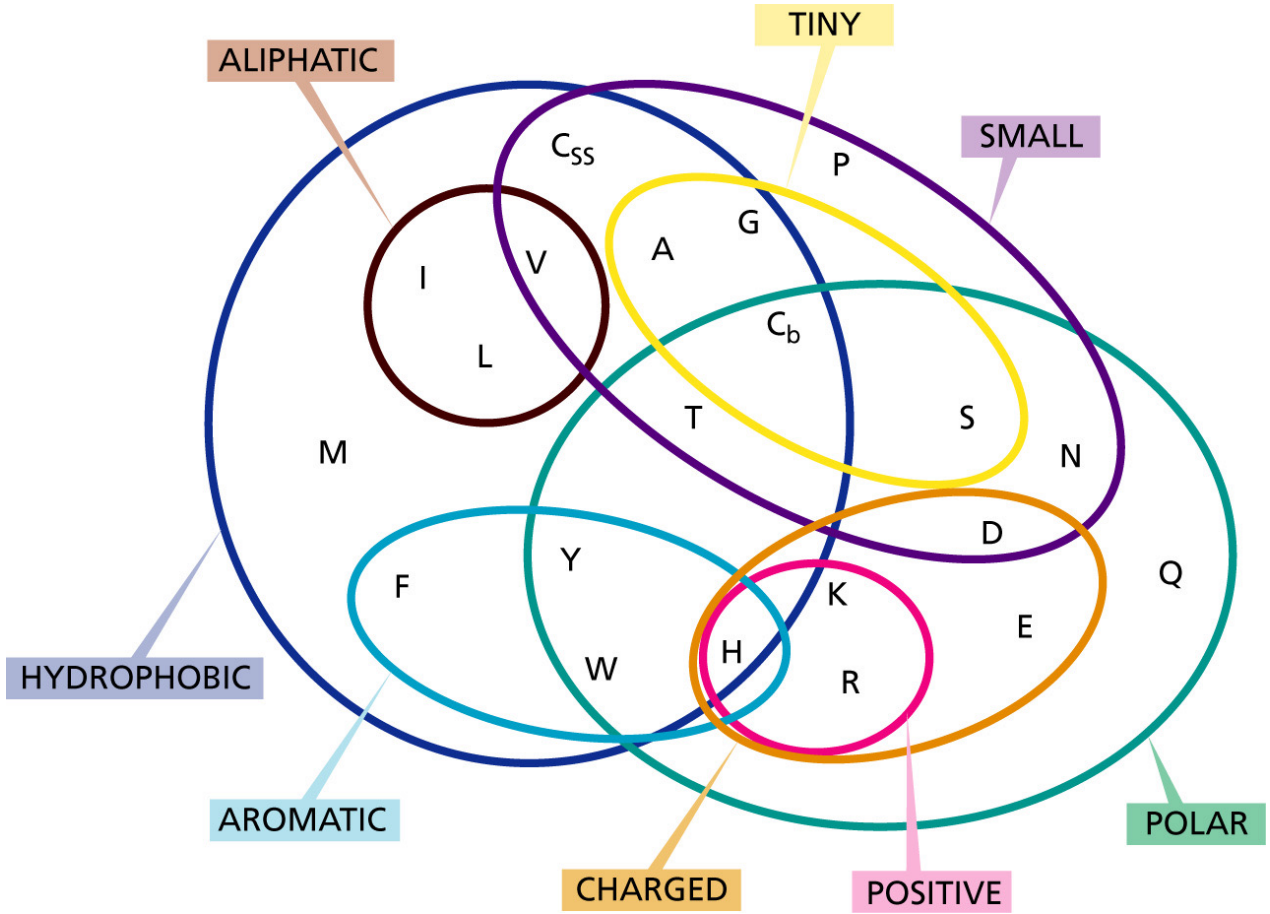


Generated by Sarah Kempe (BINF6000 in 2015)

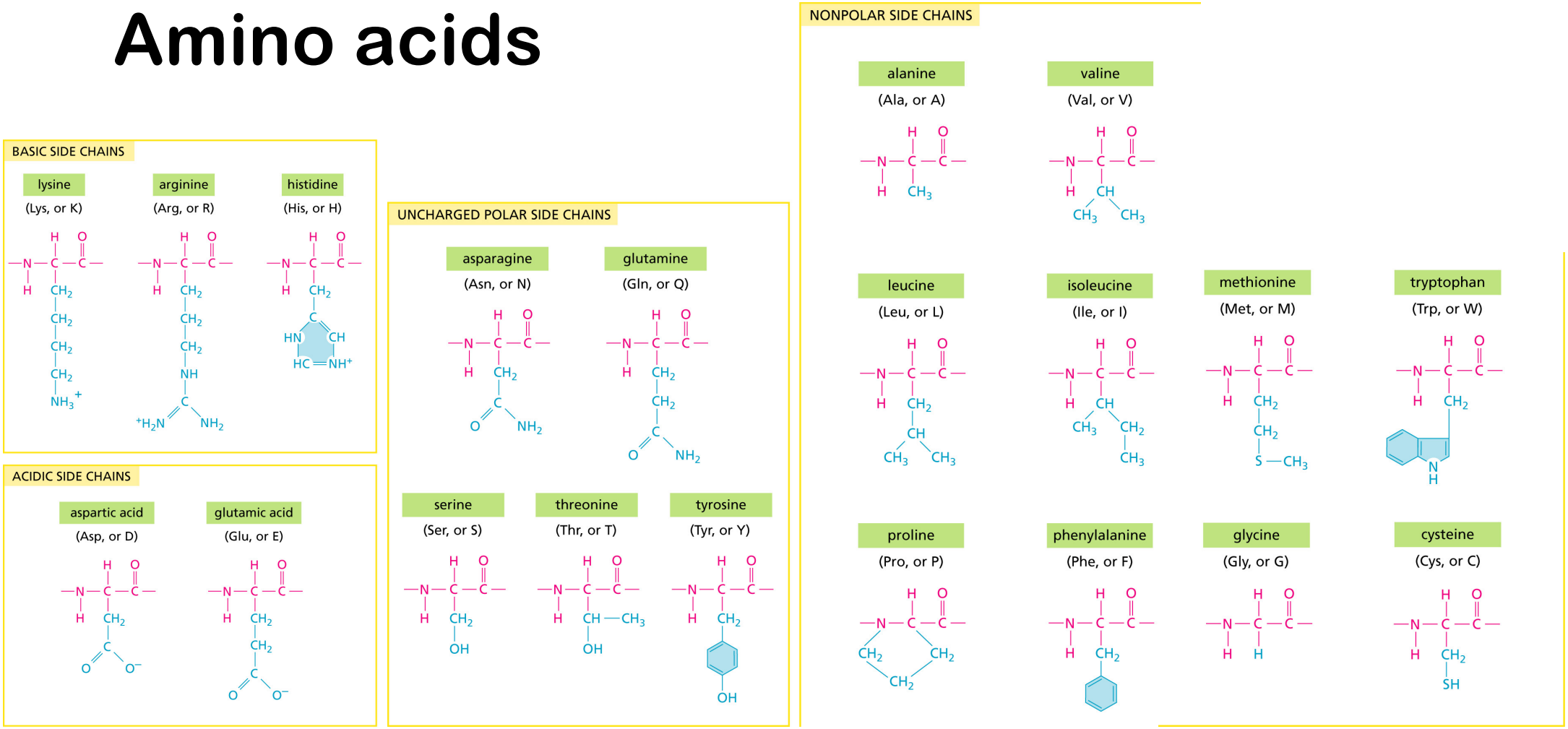


Piermarini et al. 2011
DOI: 10.1152/ajpregu.00223.2011

Protein parts are physiologically diverse

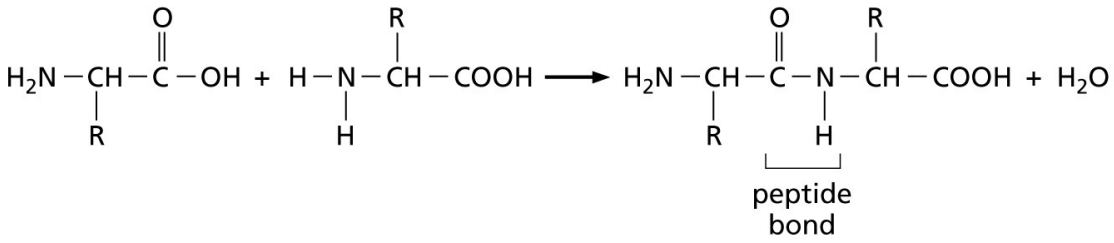


Amino acids

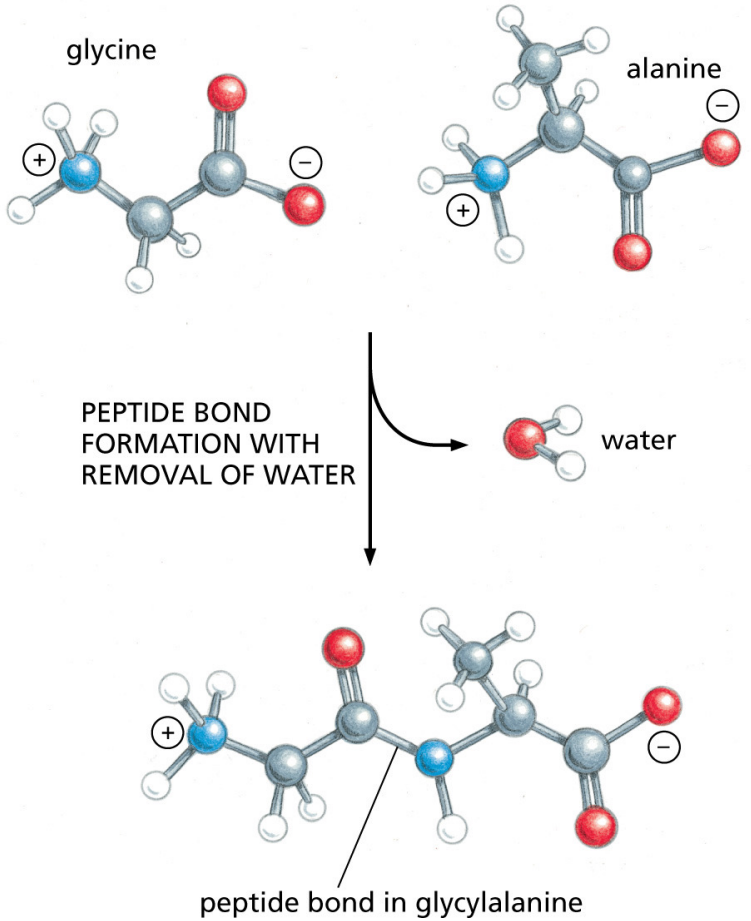


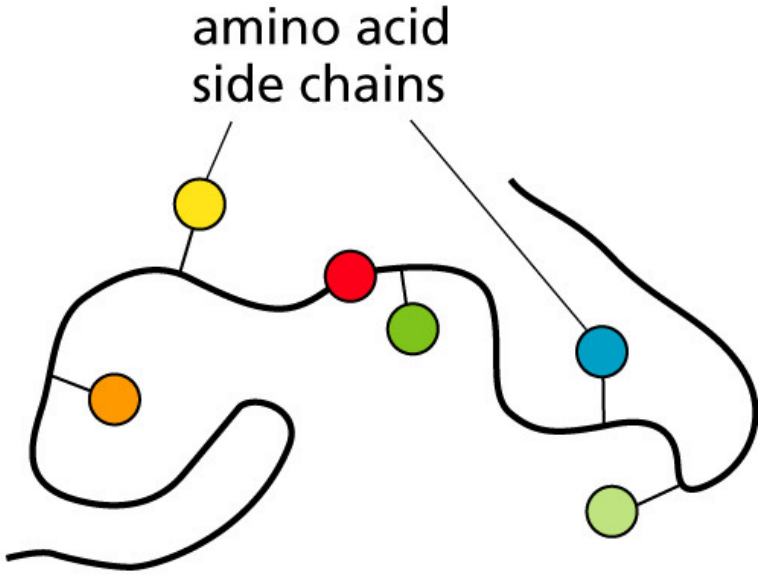
Zvelebil and Baum Fig. 2.3

(A)



(B)





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

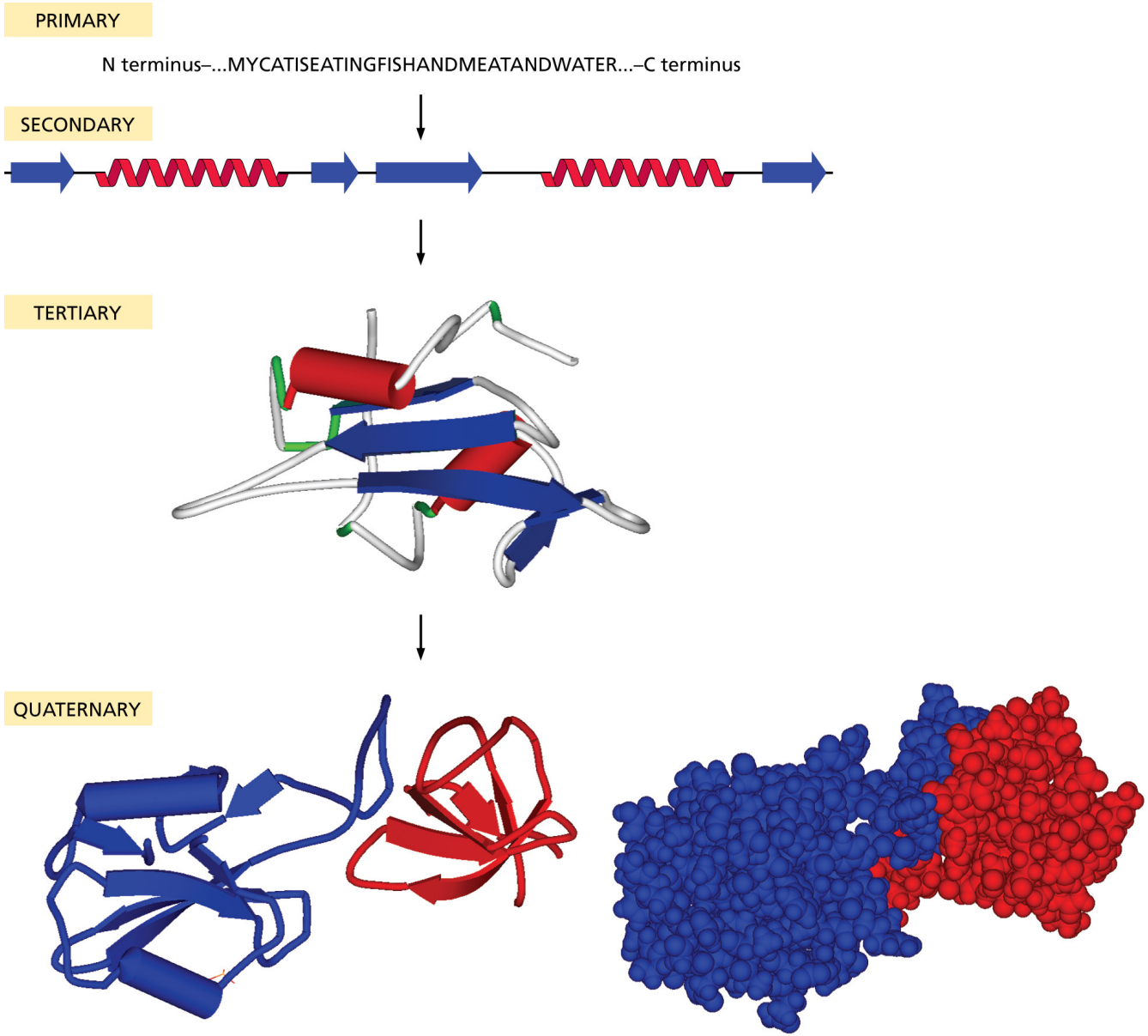
End part 1: Protein sequence

Part 2: Protein structure

Primary, secondary and tertiary structure

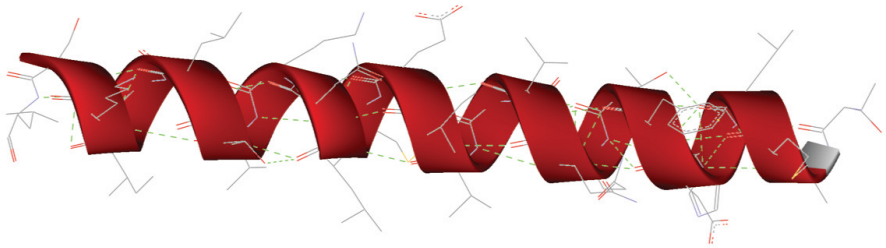
α-helix, β-sheet

Prediction of secondary structure, Chou-Fasman propensity and algorithm

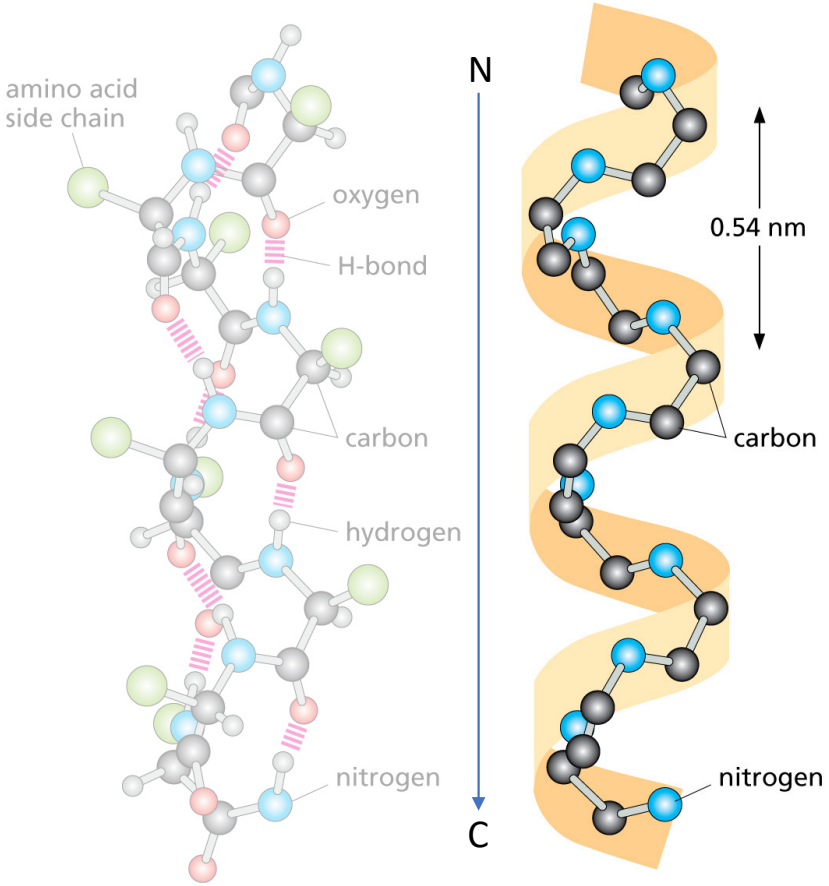

PRIMARY
N terminus–...MYCATISEATINGFISHANDMEATANDWATER...–C terminus
SECONDARY
TERTIARY
QUATERNARY

α-helix

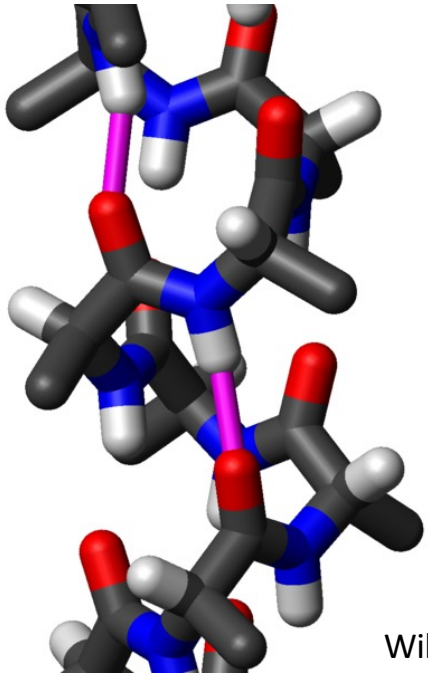
(A)



(B)



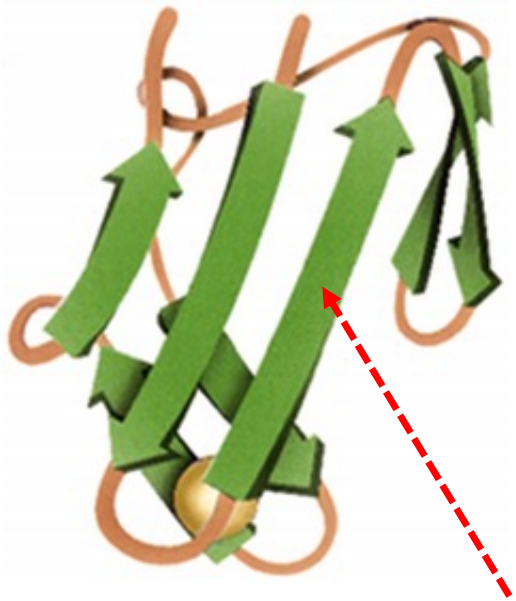
- 3.6 amino acid (residues) per turn
- O(*i*) hydrogen binds to N(*i*+4)
  [where *i* is a residue
  position in the sequence]



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



One “strand” (many of which form “sheet”)

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)

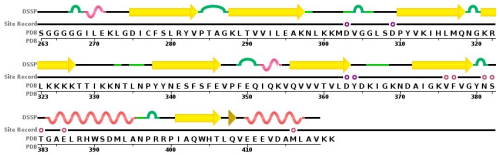
SGGGGGILEKLGDICFSLRYVPTAGKLTVVILEAKNLKKMDVGGLSDPYVKIHLMQNGKR
LKKKKTTIKKNTLNPYYNESFSFEVPFEQIQKVQVVVTVLDYDKIGKNDAIGKVFVGYNS
TGAELRHWSDMLANPRRPIAQWHTLQVEEEVDAMLA

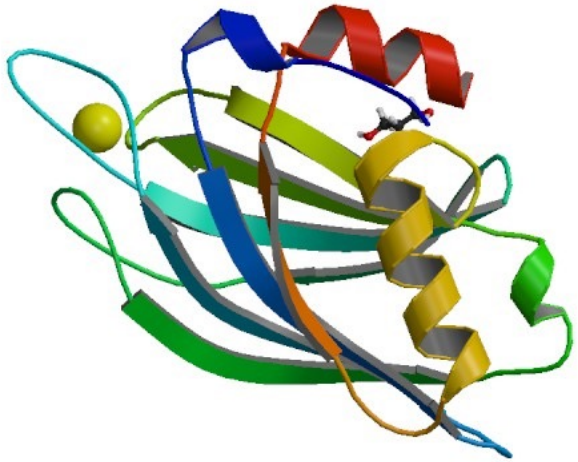
Predict

CCCCCCHHHCCCEEEEEEEEECCCCEEEEEEEEEEECCCCCCCCCCCCEEEEEEEEEECCEE
EEEEECCCCCCCCCCCEEEEEEEEECCCCCHHHCEEEEEEECCCCCCCCCEEEEEEEECCC
CHHHHHHHHHHHHCCCCCEEEEEECECHHHHHHHHC

Find rules

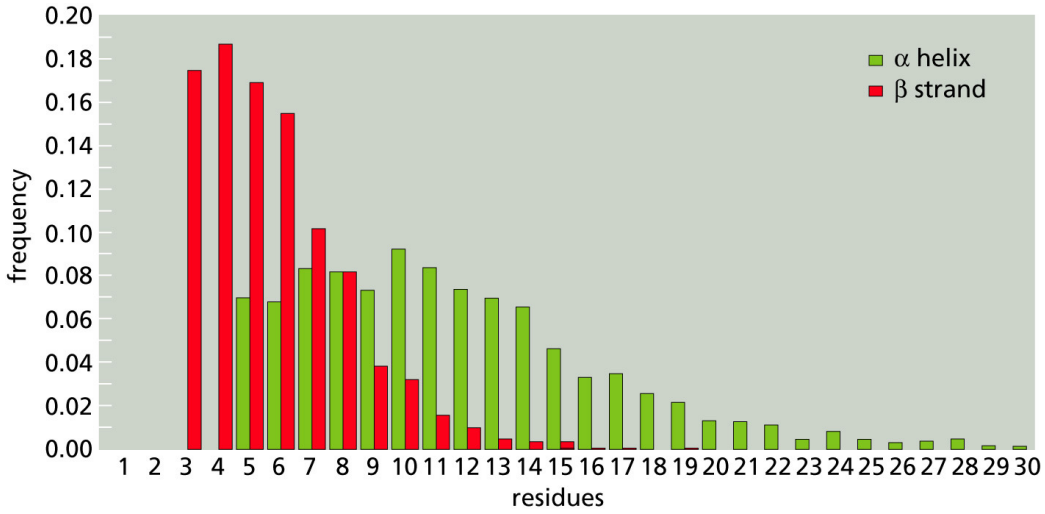
Reference





Synaptotagmin (1TJM)

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



From Zvelebil and Baum

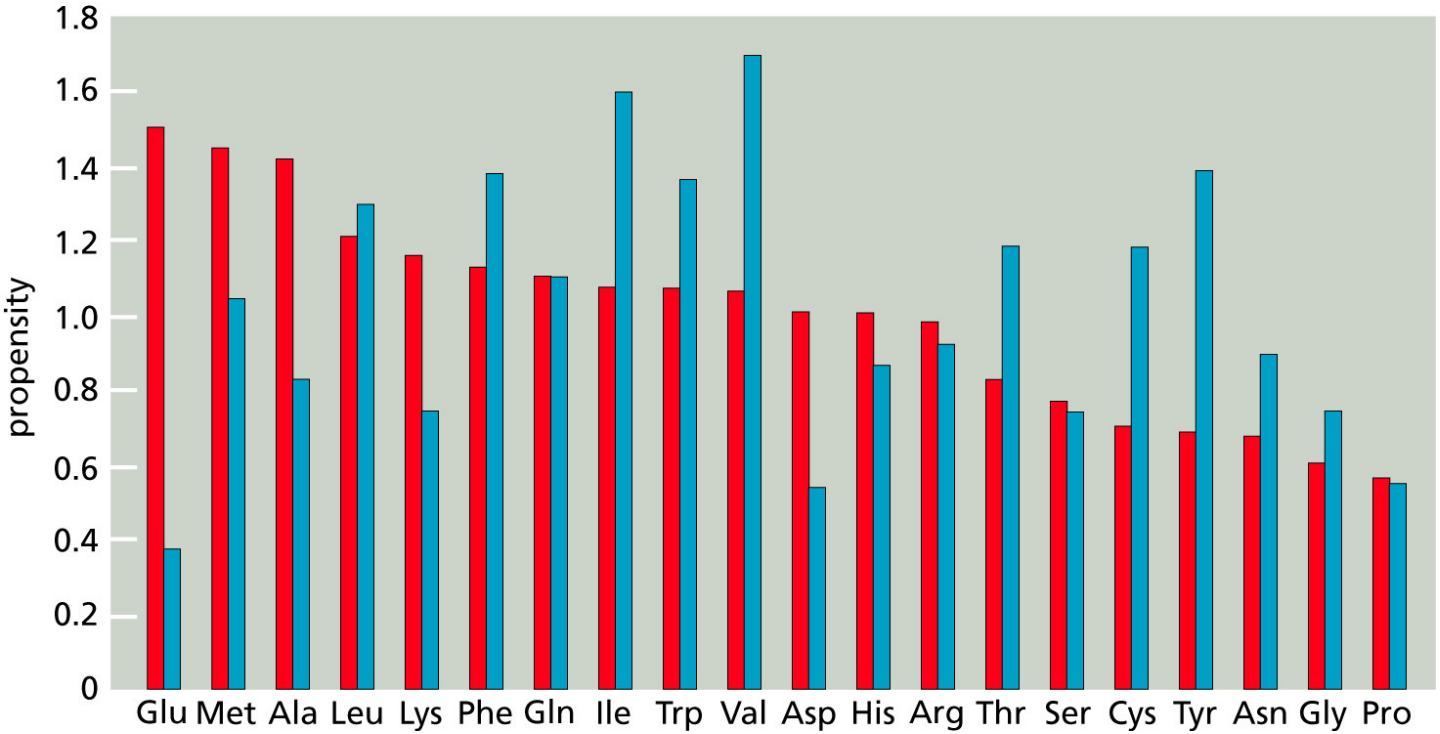
```
>1TJM
CCCCCCHHHCCCEEEEEEEEECCCCEEEEEEEEEECCCCCCCCCCCCEEEEEEEEECCEE
EEEEECCCCCCCCCCEEEEEEEEECCCCCHHHCEEEEEEECCCCCCCCEEEEEEEECCC
CHHHHHHHHHHHHCCCCCEEEEECECHHHHHHHHC

>1UJC
CEEEEEECCCECCCCCCHHHCCECHHHHHHHHHHHHHHHHHCCCCCCEEEECCCHHHHHHH
HHHHHHCCCCCCCEECHHHCCCCCHHHHHHHHHHHHHHHCCCEEEEEECCCHHHHHHHHC
CCCCCCCCCCCCEEEEEECCCCCEEEEEEECHHHCC

>1WRM
CCCCCCCEEEECCEEEECCCHHHCHHHHHHCCEEEEEECCCCCCCCCCCCEEEECCCCCC
CCCCCHHHHHHHHHHHHHHHHCCCEEEEECCCCCCHHHHHHHHHHHHCCCCCHHHHHHH
HHHCCCCCCCHHHHHHHHHHHHCHHHHHHHHHHC

>2A4V
CCCCCCCCCECCCCEEECCCCCEEEHHHHHHHCCEEEEEECCCCCCHHHHHHHHHHHHH
HHHCCCCEEEEEECCCHHHHHHHHHHCCCCEEECCCCHHHHHHCCECCCCCCECCEEE
EEECCEEEEEEECCCHHHHHHHHHHHHHHHHCCC

>3GT6
CCCCCCCEEEEECCCEEEEEEECCCCCHHHCEEEECCEEEEEEECCHHHCCHHHEEEE
CCCCEEEEEEEECCCCECCCCCEEEEECCEEEEEECCC

>3OSE
CCCECCCCCCCCCEECCCHHHHHHHHHHHHHHCCCEEEEECCEEEEECCCCCCCCEEE
EEEEEEHHHCEEEEEEEEECCHHHHHHHHHHHHHCCC

>1FRT
CCCCCCEEEEEEECCCCCCCCCEEEEEECCEEEEEECCCCCCEECHHHHHCCCCCCH
HHHHHHHHHHHHHHHHHHHHHHHHHCCCCCEEEEEEEECCCCCEEEEEEEECCEEEE
EEECCCCEEEECCHHHHHHHHHHCCCHHHHHHHHHHCHHHHHHHHHHCHHHCCCE
CCEEEEEEECCCCEEEEEEEEEECCCCEEEEECCEECCCCEEEEECCCCCEEEEE
EEEECCCHHHEEEEECCCCCCEEECC
```

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

```
>1TJM
SGGGGGILEKLGDICFSLRYVPTAGKLTVVILEAKNLKKMDVGGLSDPYVKIHLMQNGKR
LKKKKTTIKKNTLNPYYNESFSFEVPFEQIQKVQVVVTVLDYDKIGKNDAIGKVFVGYNS
TGAELRHWSDMLANPRRPIAQWHTLQVEEEVDAMLA

>1UJC
MQVFIMRHGDAALDAASDSVRPLTTNGCDESRLMANWLKGQKVEIERVLVSPFLRAEQTL
EEVGDCLNLPSSAEVLPELTPCGDVGLVSAYLQALTNEGVASVLVISHLPLVGYLVAELC
PGETPPMFTTSAIASVTLDESGNGTFNWQMSPCNLK

>1WRM
GPMGNGMNKILPGLYIGNFKDARDAEQLSKNKVTHILSVHDSARPMLEGVKYLCIPAADS
PSQNLTRHFKESIKFIHECRLRGESCLVHCLAGVSRSVTLVIAYIMTVTDFGWEDALHTV
RAGRSCANPNVGFQRQLQEFEKHEVHQYRQWLKEEY

>2A4V
DVNELEIGDPIPDLSLLNEDNDSISLKKITENNRVVVFFVYPRASTPGSTRQASGFRDNY
QELKEYAAVFGLSADSVTSQKKFQSKQNLPYHLLSDPKREFIGLLGAKKTPLSGSIRSHF
IFVDGKLKFKRVKISPEVSVNDAKKEVLEVAEKFKE

>3GT6
TAQWVPRVDIKEEVNHFVLYADLPGIDPSQIEVQMDKGILSIRGERKSESSTETERFSRI
ERRYGSFHRRFALPDSADADGITAAGRNGVLEIRIPKRPA

>3OSE
KPRSLRFTWSMKTTSSMDPNDMMREIRKVLDANNCDYEQKERFLLFCVHGDARQDSLVQW
EMEVCKLPRLSLNGVRFKRISGTSIAFKNIASKIANELKL

>1FRT
AEPRLPLMYHLAAVSDLSTGLPSFWATGWLGAQQYLTYNNLRQEADPCGAWIWENQVSWY
WEKETTDLKSKEQLFLEAIRTLENQINGTFTLQGLLGCELAPDNSSLPTAVFALNGEEFM
RFNPRTGNWSGEWPETDIVGNLWMKQPEAARKESEFLLTSCPERLLGHLERGRQNLEWKE
PPSMRLKARPGNSGSSVLTCAAFSFYPPELKFRFLRNGLASGSGNCSTGPNGDGSFHAWS
LLEVKRGDEHHYQCQVEHEGLAQPLTVDL
```

```
>1TJM
CCCCCCHHHCCCEEEEEEEEECCCCEEEEEEEEECCCCCCCCCCCCEEEEEEEEECCEE
EEEEECCCCCCCCCCEEEEEEEEECCCCCHHHCEEEEEEEECCCCCCCCEEEEEEEECCC
CHHHHHHHHHHHHCCCCCEEEEEECECHHHHHHHHC

>1UJC
CEEEEEECCCECCCCCCHHHCCECHHHHHHHHHHHHHHHHHCCCCCCEEEECCCHHHHHHH
HHHHHHCCCCCCCCEECHHHCCCCCHHHHHHHHHHHHHHHCCCEEEEEECCCHHHHHHHHC
CCCCCCCCCCCCCEEEEEECCCCCEEEEEEECHHHCC

>1WRM
CCCCCCCEEEECCEEEECCCHHHCHHHHHHCCEEEEEECCCCCCCCCCCCEEEECCCCCC
CCCCCHHHHHHHHHHHHHHHHCCCEEEEECCCCCCHHHHHHHHHHHHCCCCCHHHHHHHH
HHHCCCCCCCHHHHHHHHHHHHCHHHHHHHHHHC

>2A4V
CCCCCCCCCECCCCEEECCCCCEEEHHHHHHHCCEEEEEECCCCCCHHHHHHHHHHHHH
HHHCCCCEEEEEECCCHHHHHHHHHHCCCCEEEECCCCHHHHHHCCECCCCCCECCEEE
EEECCEEEEEEECCCHHHHHHHHHHHHHHHHHCCC

>3GT6
CCCCCCCEEEEECCCEEEEEEECCCCCHHHCEEEECCEEEEEEECCHHHCCHHHEEEE
CCCCEEEEEEEECCCCECCCCCEEEEECCEEEEEEECCC

>3OSE
CCCECCCCCCCCCEECCCHHHHHHHHHHHHHHCCCEEEEECCEEEEECCCCCCCCEEE
EEEEEEHHHCEEEEEEEECCHHHHHHHHHHHHHHCCC

>1FRT
CCCCCCEEEEEEECCCCCCCCCEEEEEECCEEEEEECCCCCCEECHHHHHCCCCCCH
HHHHHHHHHHHHHHHHHHHHHHHHHCCCCEEEEEEEEECCCCCEEEEEEEECCEEEE
EEECCCCEEEECCHHHHHHHHHHCCCHHHHHHHHHHCHHHHHHHHHHCHHHCCCE
CCEEEEEEECCCCEEEEEEEEEECCCCEEEEECCEECCCCEEEEEECCCCCEEEEE
EEEECCCHHHEEEEECCCCCCCEEECC
```

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

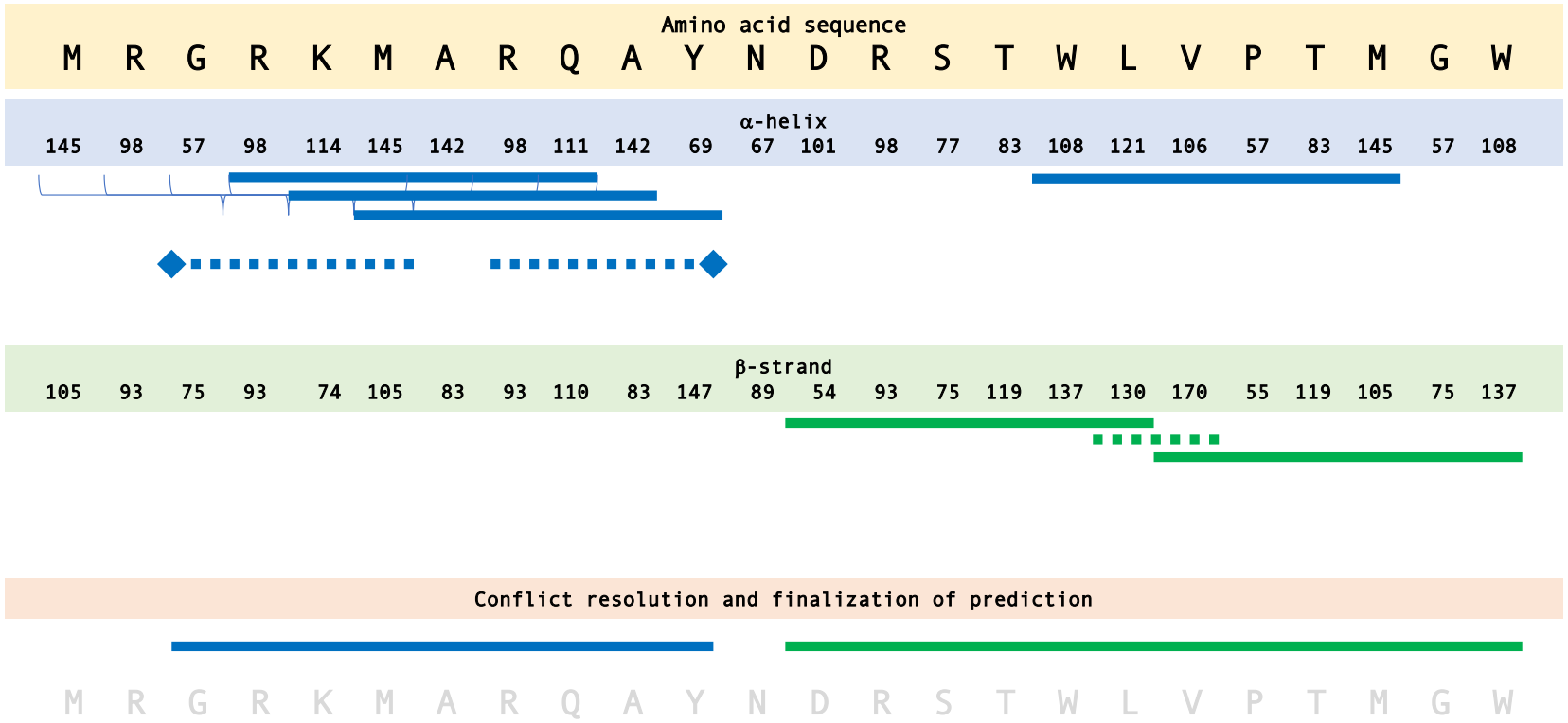
- Helix, Strand
  1. Scan for window of 6 residues, mark state when score > 100 for *at least* 4 residues (for helix state) or 3 residues (for strand state)
  2. Propagate mark in both directions until 4 (or 3) residue window has *average* propensity < 100
  3. Move forward and repeat

GHWIATRGQLIREAYEDYRHFSSECPFIP

- Conflict solution

  Any region containing overlapping alpha-helical and beta-strand assignments are taken to be
  - helical if the average *P*(helix) > *P*(strand)
  - strand if the average *P*(strand) > *P*(helix)

- Accuracy approx. 50% - 60%



End part 2: Protein structure

Name	P(H)	P(E)	P(T)
Alanine (A)	142	83	66
Arginine (R)	98	93	95
Aspartic acid (D)	101	54	146
Asparagine (N)	67	89	156
Cysteine (C)	70	119	119
Glutamic acid (E)	151	37	74
Glutamine (Q)	111	110	98
Glycine (G)	57	75	156
Histidine (H)	100	87	95
Isoleucine (I)	108	160	47
Leucine (L)	121	130	59
Lysine (K)	114	74	101
Methionine (M)	145	105	60
Phenylalanine (F)	113	138	60
Proline (P)	57	55	152
Serine (S)	77	75	143
Threonine (T)	83	119	96
Tryptophan (W)	108	137	96
Tyrosine (Y)	69	147	114
Valine (V)	106	170	50

	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 θ=0.50
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)

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=True θ=**0.40**
FN
TN
TP
TN
TP
FN
TP
FP
TN
TP

Test metrics

	Positive prediction	Negative prediction
Positive observation	~~3 (TP)~~ **4 (TP)**	~~3 (FN)~~ **2 (FN)**
Negative observation	**1 (FP)**	**3 (TN)**

Threshold?

Opportunistic ⟷ Conservative

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

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

VLHQASGNSVILFGSDVTVPGATNAEQAR	amino acid sequence 29 residues long
HHHHHCCCCEEEECCCEEECCCCCHHHHH	actual secondary structure
CHHHCCCCEEEECCCCCEEECCCHHHHHH	prediction 1: **Q_3 = 22/29 = 76%:** useful
HHHHHCCCCHHHHCCCHHHCCCCCHHHHH	prediction 2: **Q_3 = 22/29 = 76%:** terrible