Structural Bioinformatics
Lecture 4

Prediction of structural features from sequence

UNIVERSITÄT
DES
SAARLANDES

Proteins

Sequence  **Structure**  **Function**

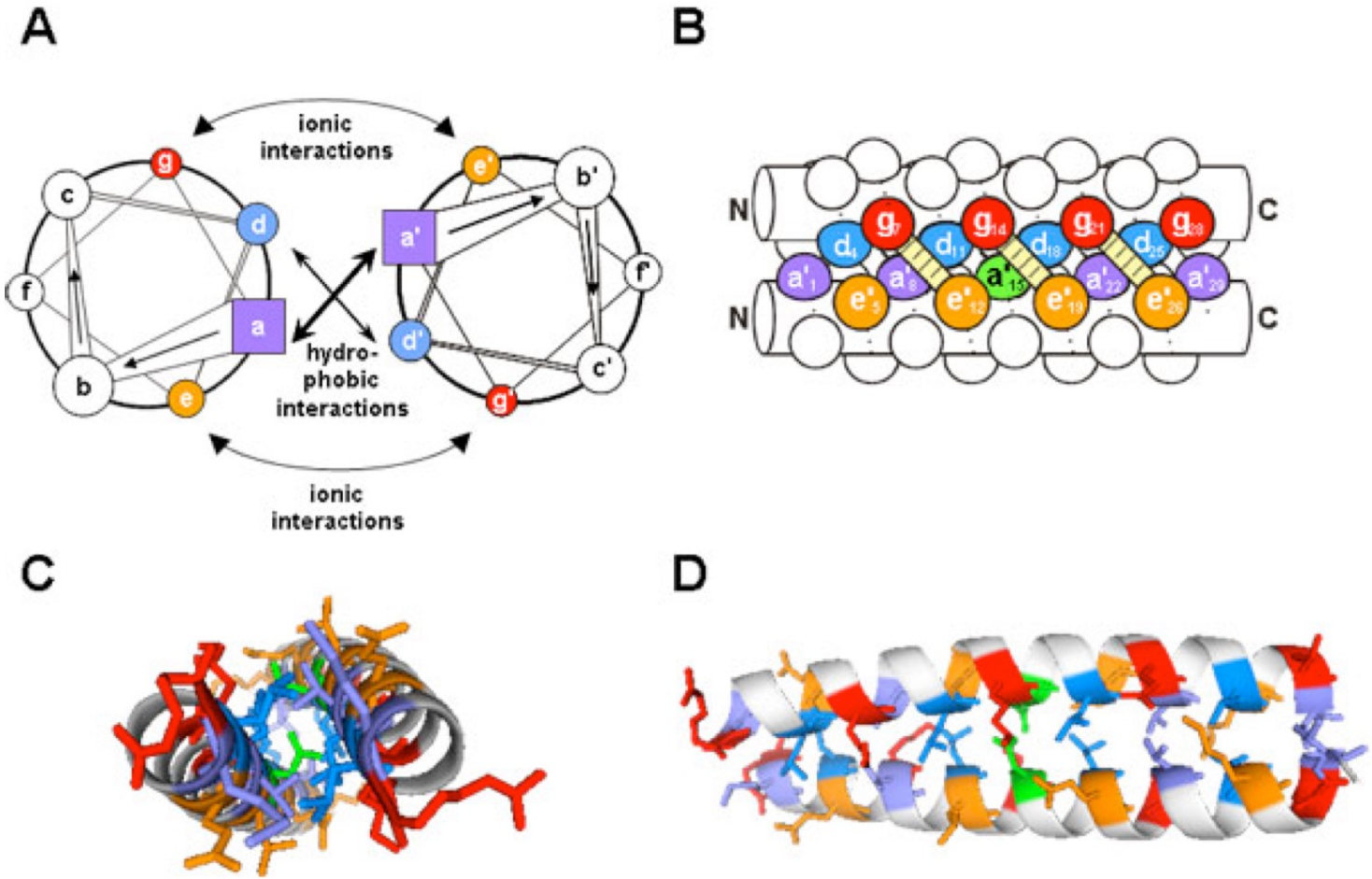
Outline

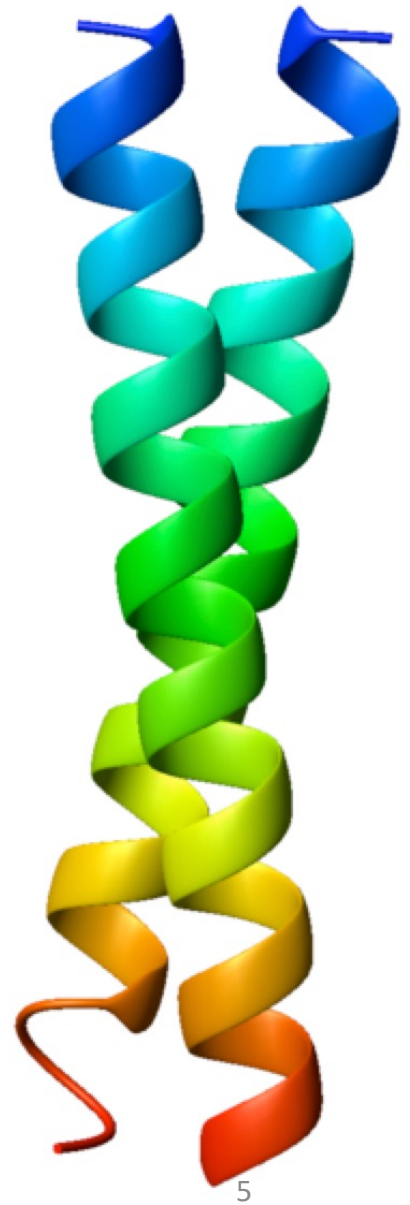
- Coiled-coil regions
- Membrane proteins
- Prediction of membrane-embedded segments
- Prediction of secondary structure

Predicting coiled-coil regions

What are coiled-coils?

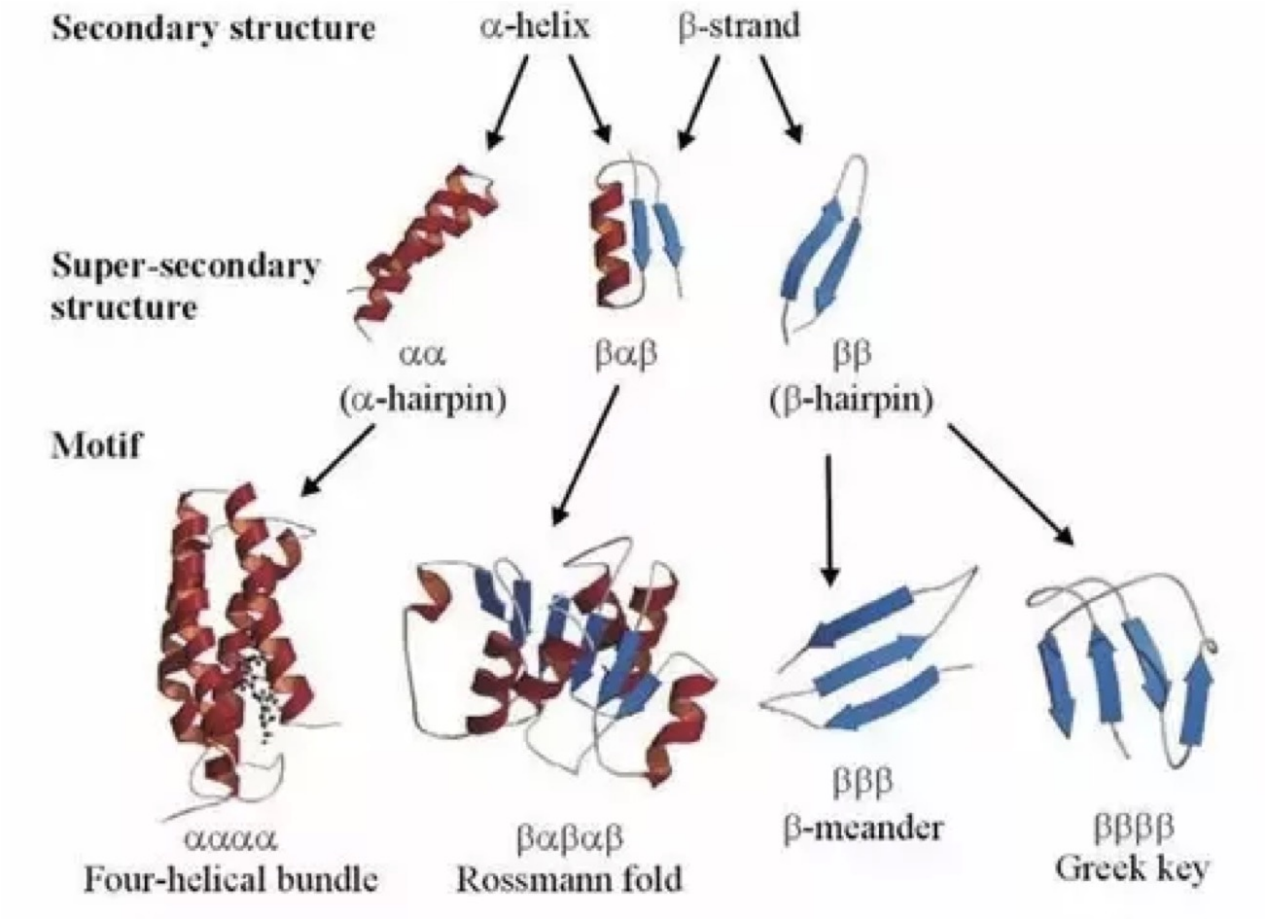
- 2 to 7 α-helices coiled together (**super-secondary structure**)



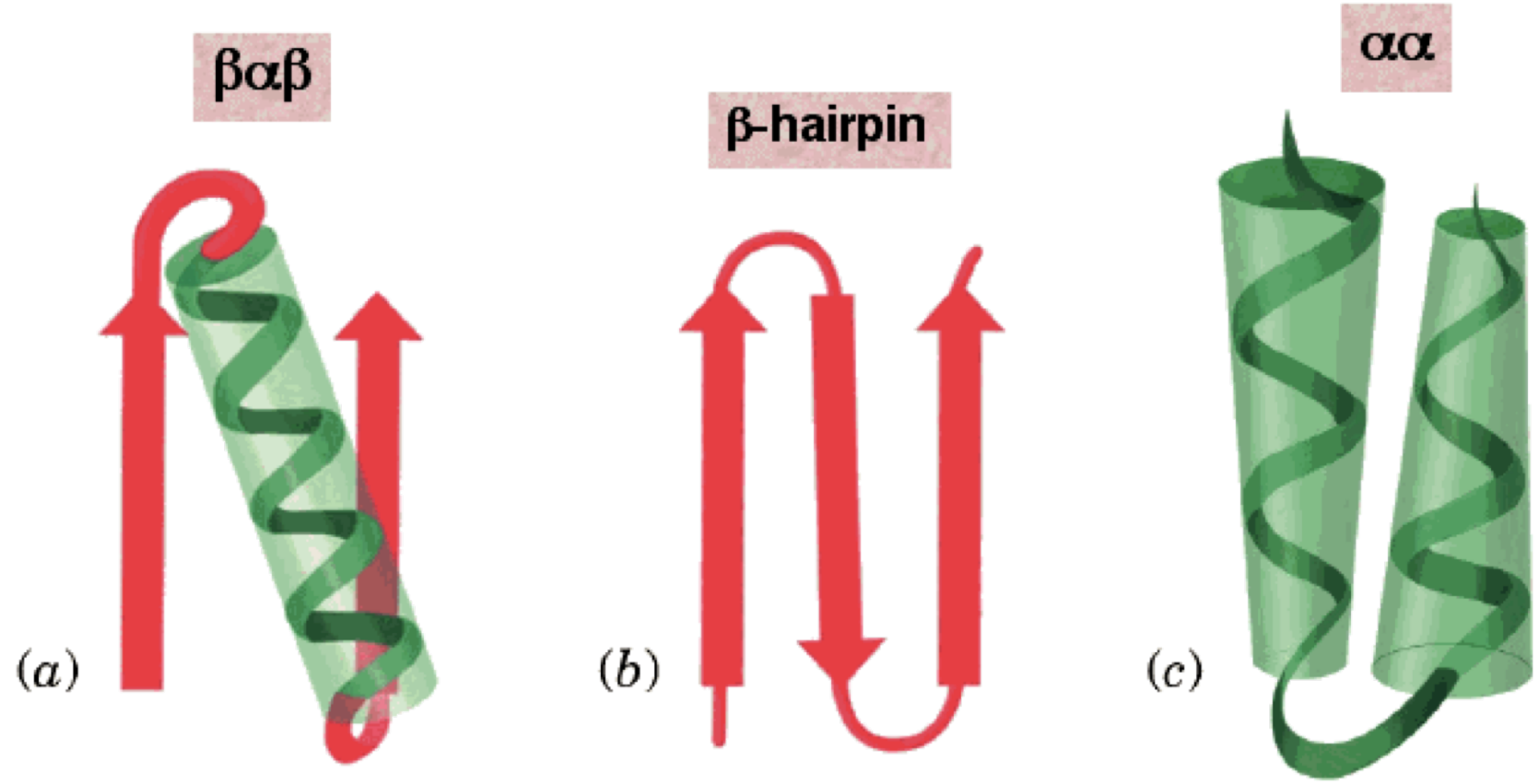


Detour: super-secondary structure

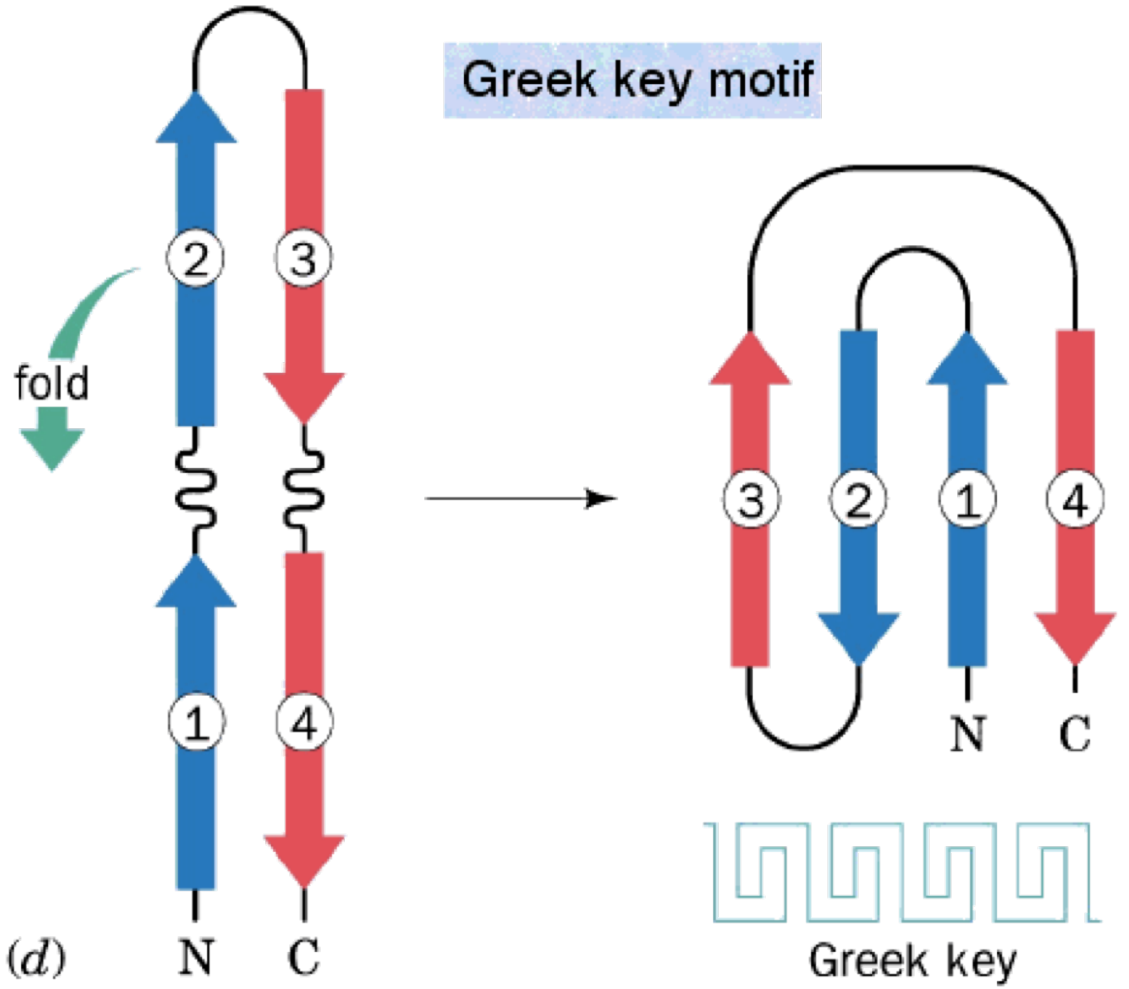
- Halfway between secondary and tertiary structure: spatial arrangement of several secondary structure elements (usually, continuous)

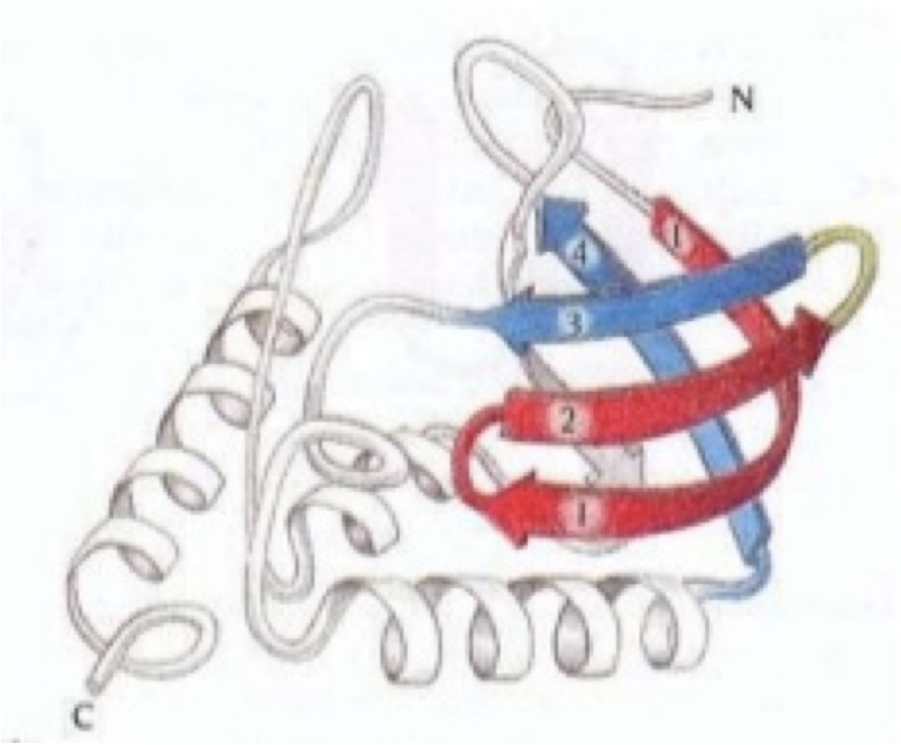


Super-secondary structure motifs



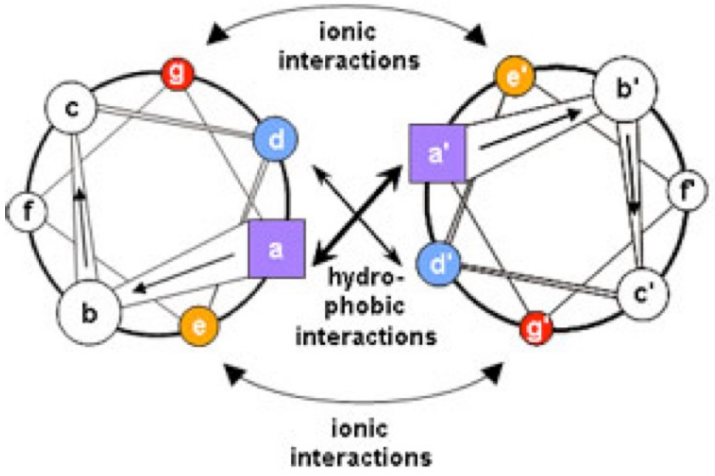
Super-secondary structure motifs

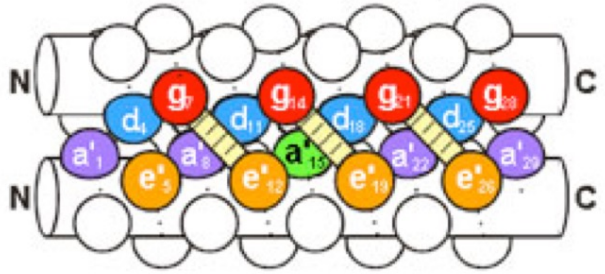




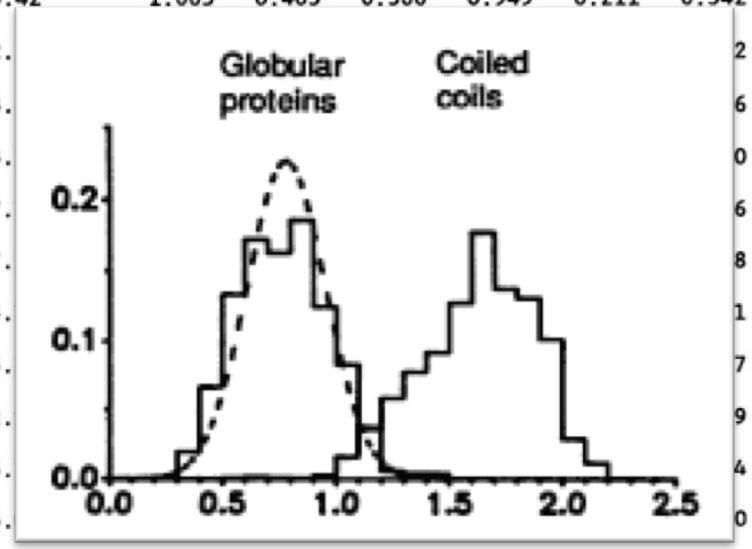
Coiled-coil prediction: COILS (Lupas et al., 1991)

- The 7-residues (**heptad**) pattern is very strong
- A matrix for relative occurrence of all aa. in known coiled-coil regions
- Sliding-window





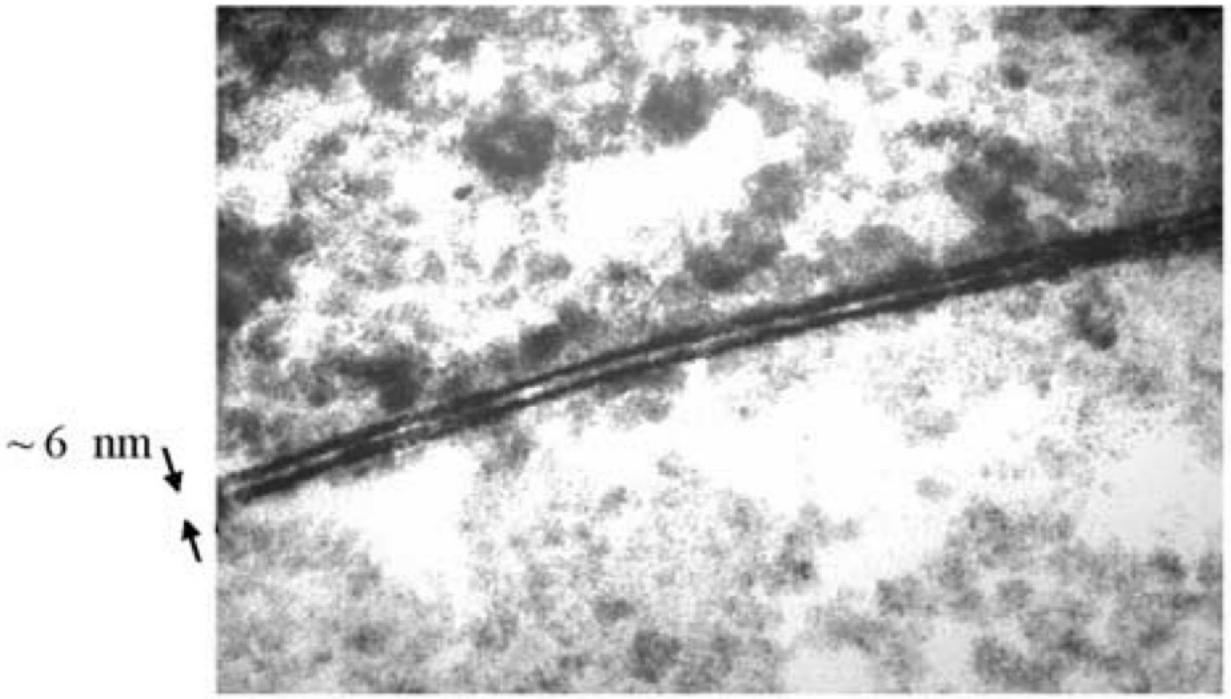
Residue	Frequency in GenBank (%)	Relative occurrence at position						
		a	b	c	d	e	f	g
Leu	9.33	3.167	0.297	0.398	3.902	0.585	0.501	0.483
Ile	5.35	2.597	0.098	0.345	0.894	0.514	0.471	0.431
Val	6.42	1.665	0.403	0.386	0.949	0.211	0.342	0.360
Met	2.						2	0.663
Phe	3.						6	0.013
Tyr	3.						0	0.155
Gly	7.						6	0.156
Ala	7.						8	0.877
Lys	5.						1	2.795
Arg	5.						7	1.798
His	2.						9	0.213
Glu	6.						4	3.048
Asp	5.						0	1.448
Gln	4.27	0.179	2.114	1.778	0.631	2.550	1.578	2.526
Asn	4.25	0.835	1.475	1.534	0.039	1.722	2.456	2.280
Ser	7.28	0.382	0.583	1.052	0.419	0.525	0.916	0.628
Thr	5.97	0.169	0.702	0.955	0.654	0.791	0.843	0.647
Cys	1.86	0.824	0.022	0.308	0.152	0.180	0.156	0.044
Trp	1.41	0.240	0	0	0.456	0.019	0	0
Pro	5.28	0	0.008	0	0.013	0	0	0



Membrane proteins

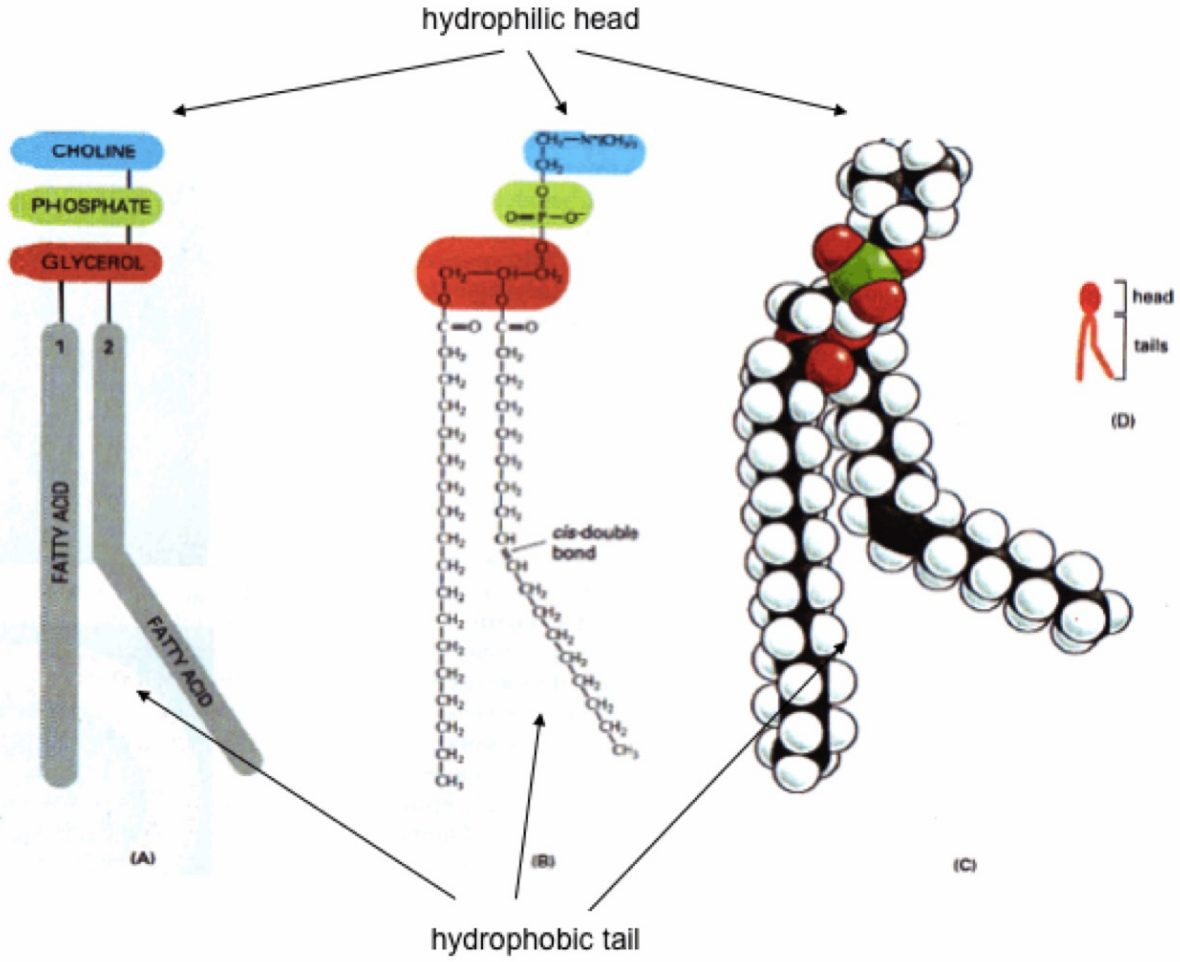
Biological membranes

- Maintain compartmentalisation
- Cell membranes
  - Plasma membrane: encloses the cell
  - In eukaryotic cells:
    - ER
    - Golgi apparatus
    - Mitochondia
    - Nucleus
    - …



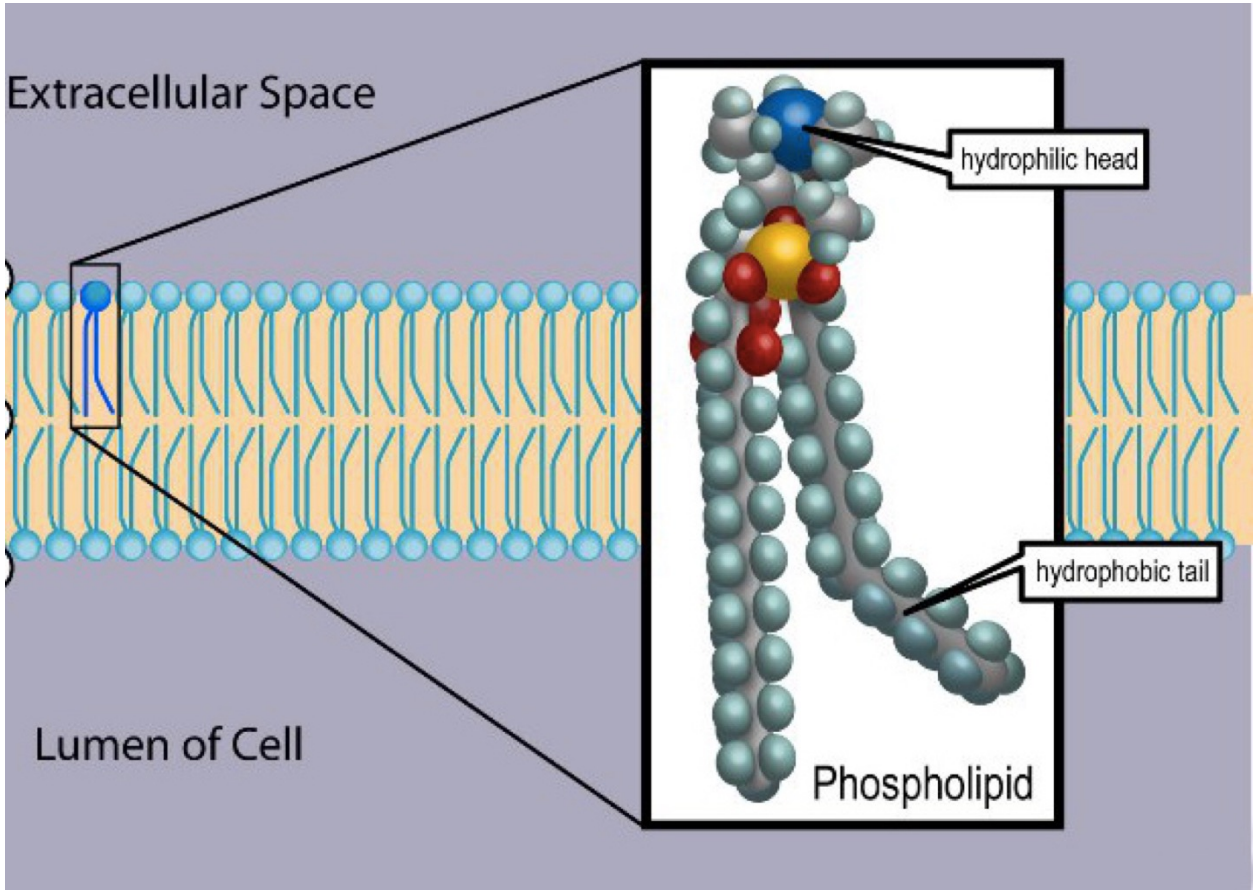
Lipids

- Lipids are **amphipathic** molecules: have a hydrophobic and a hydrophilic part

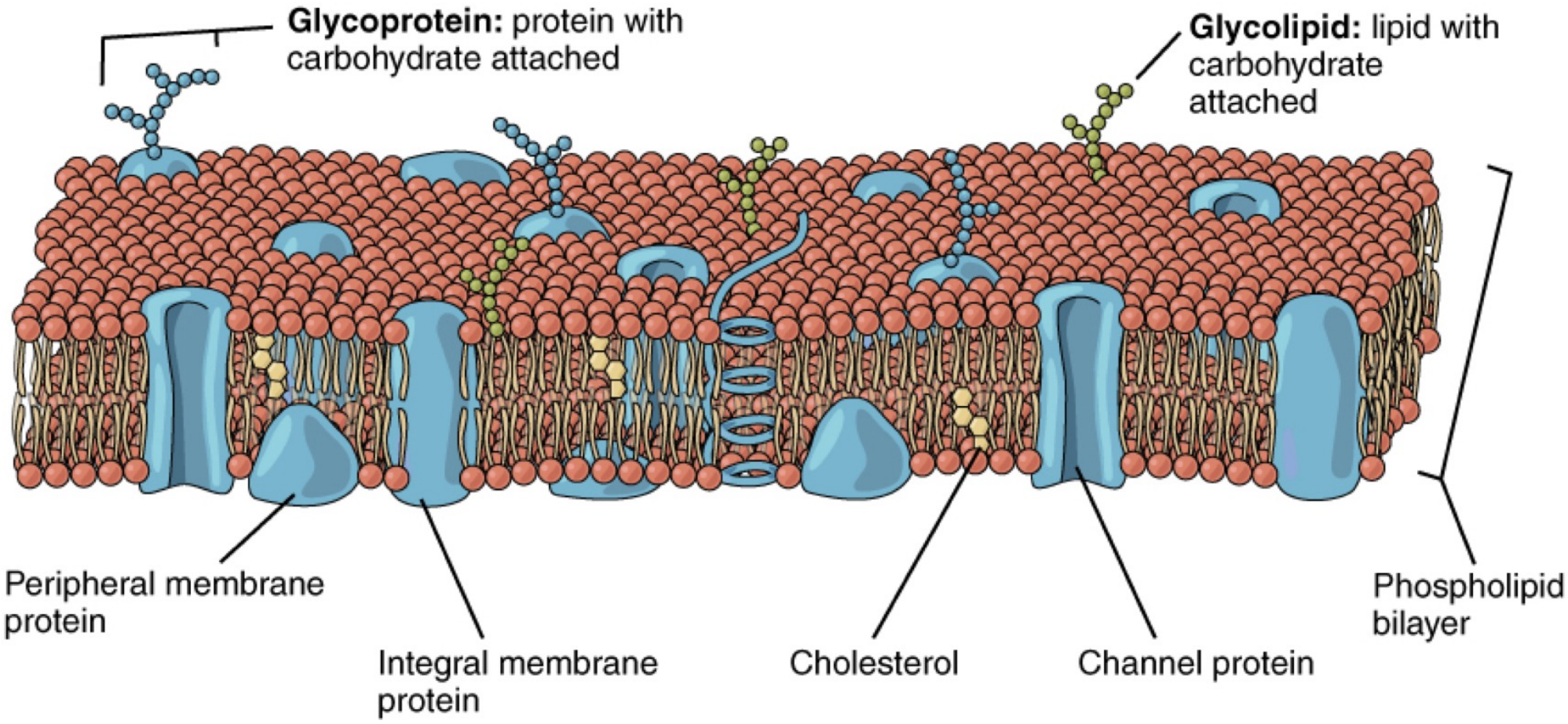


Lipid bilayer

- Lipids are **amphipathic** molecules: have a hydrophobic and a hydrophilic part
  => tend to form bilayers



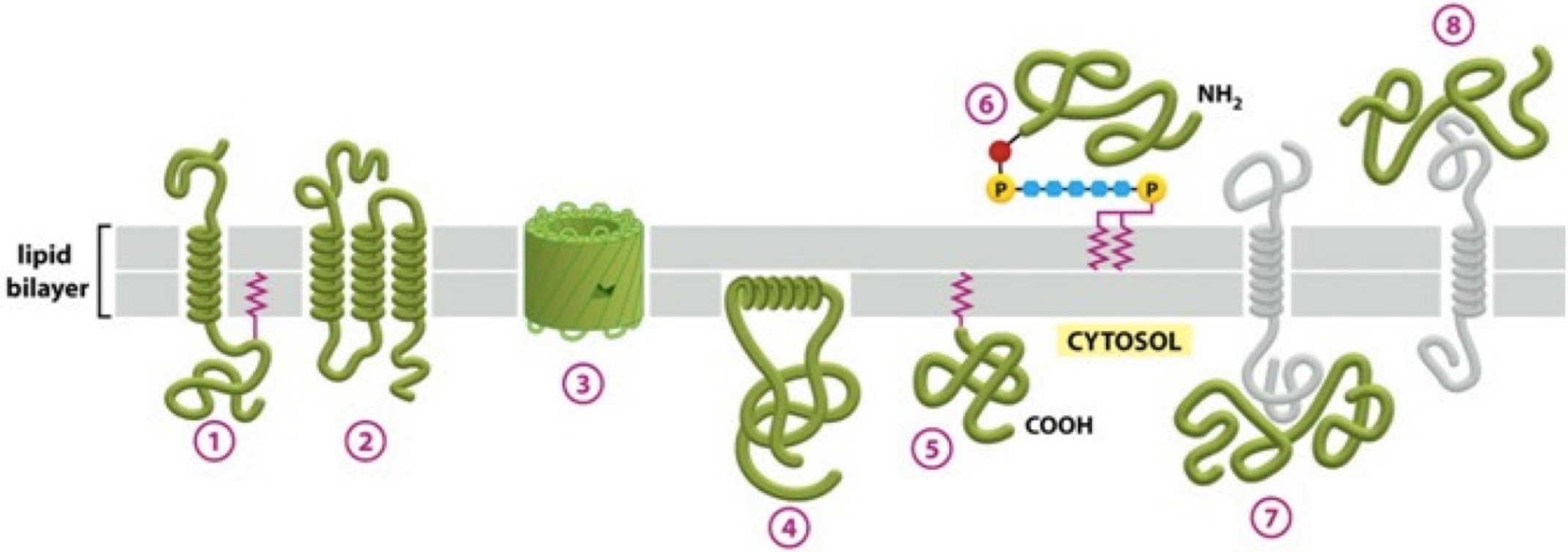
Lipid bilayer and membrane proteins



Membrane proteins can be associated with the lipid bilayer in different ways

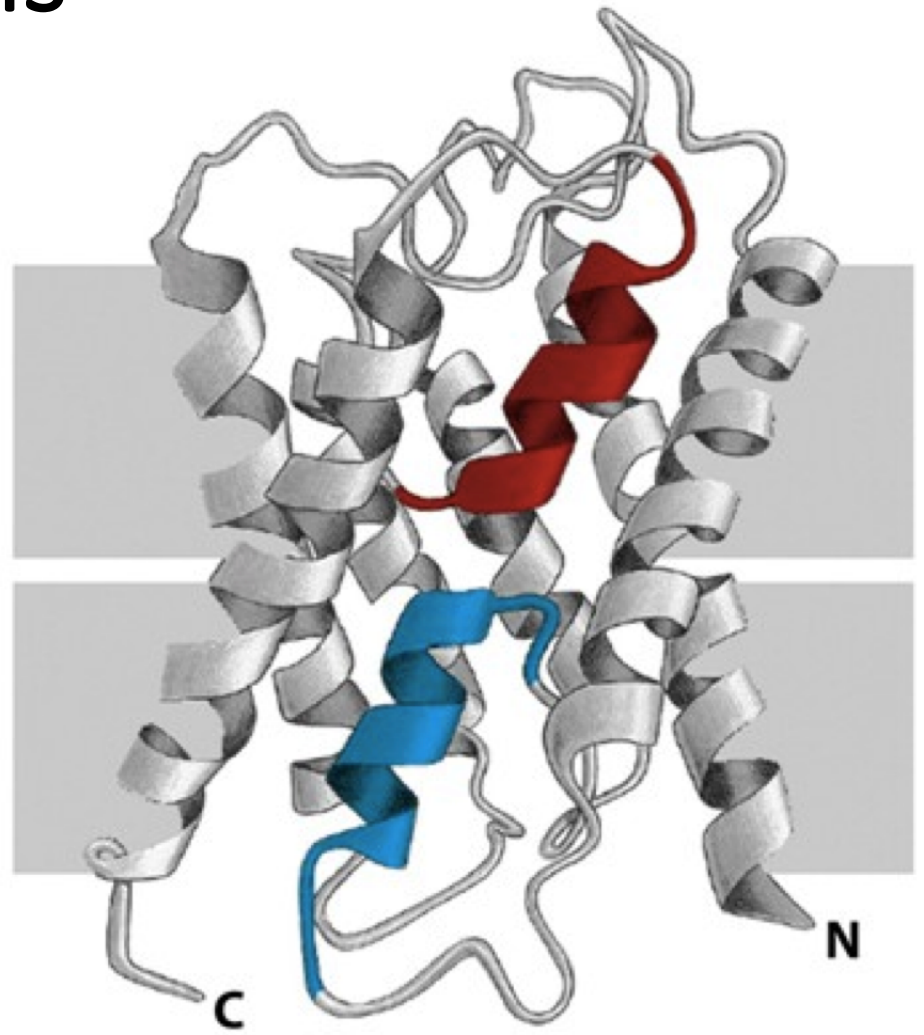
Membrane proteins

- **Membrane proteins** can be associated with the lipid bilayer in different ways
- Membrane proteins perform most membrane-specific functions

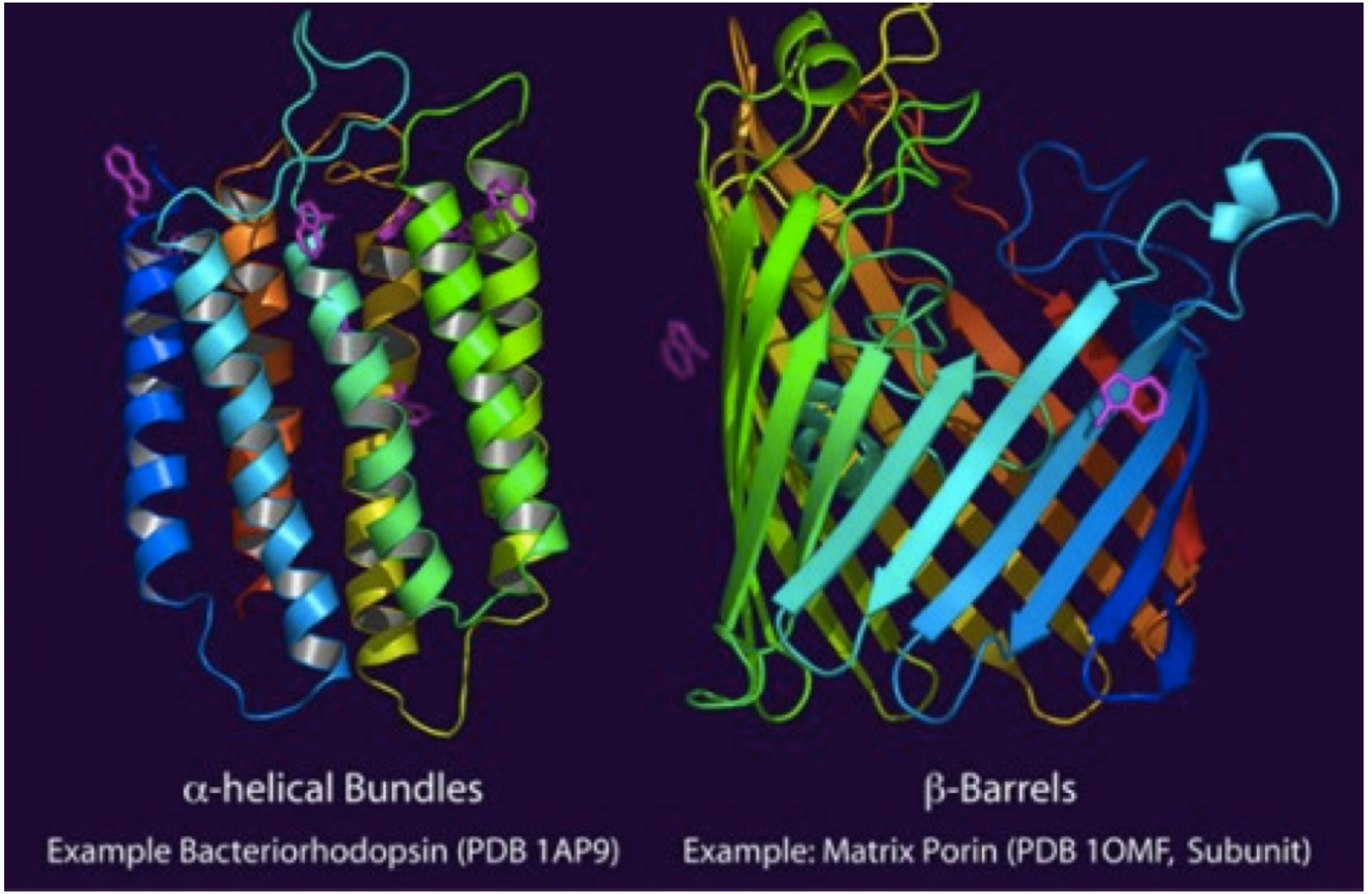


Transmembrane protein regions

- Some proteins have segments that span the membrane — **transmembrane (TM) regions**
- Proteins can have one (**single-pass**, **anchored**) or many TM segments (**multi-pass**)
- In multi-pass proteins, TM regions are connected by loops or larger structures
- TM regions are more hydrophobic than extracellular and intracellular loops

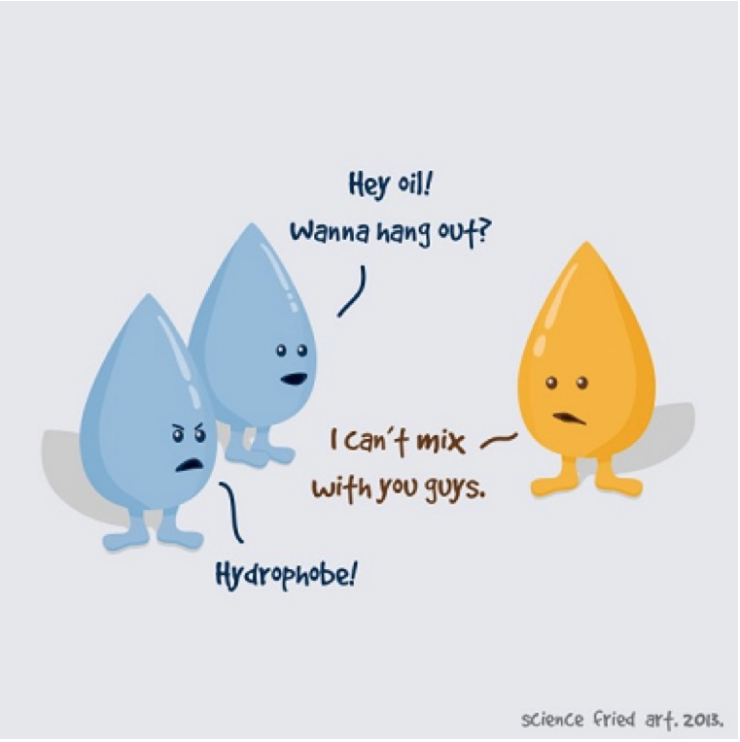


Two major types of (multi-pass) transmembrane proteins



Hydrophobicity

- Nonpolar substances tend to aggregate in an aqueous solution and exclude water



Protein hydrophobic core

- Hydrophobic amino acids tend to be excluded from the protein surface => pack against each other in its core

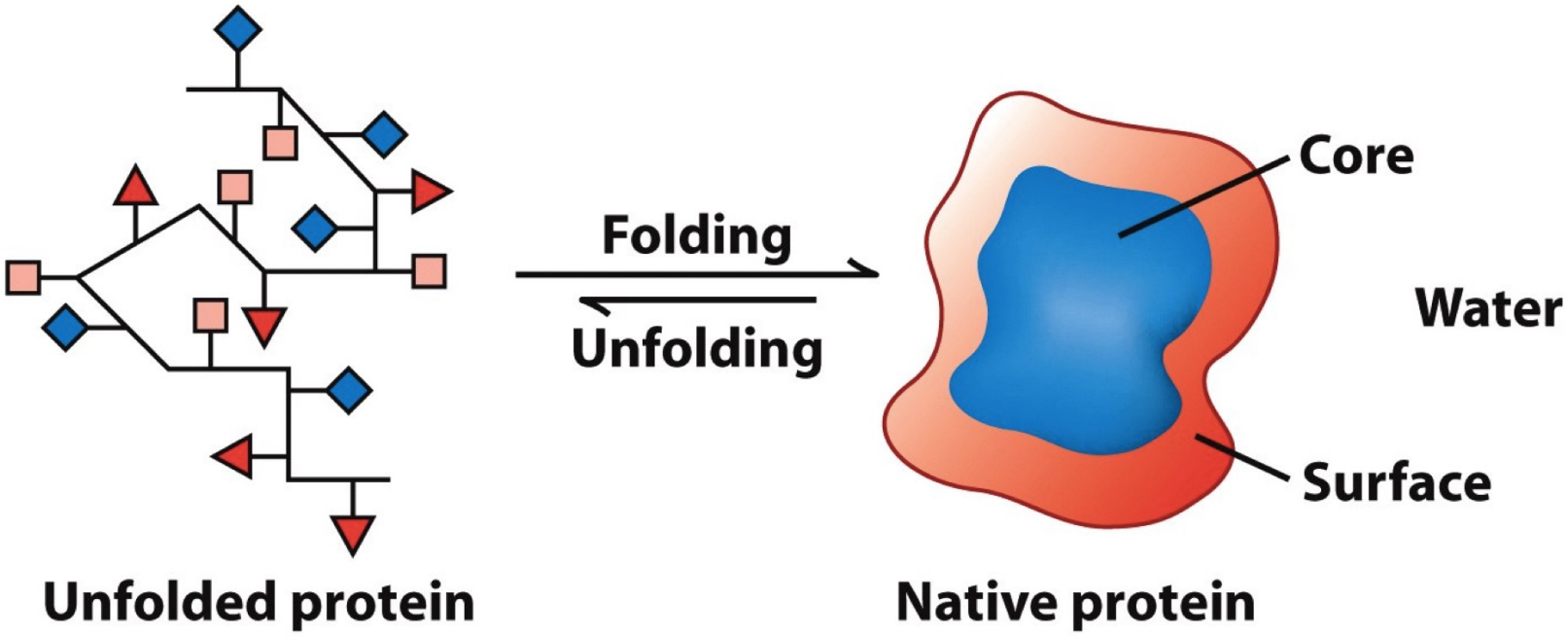


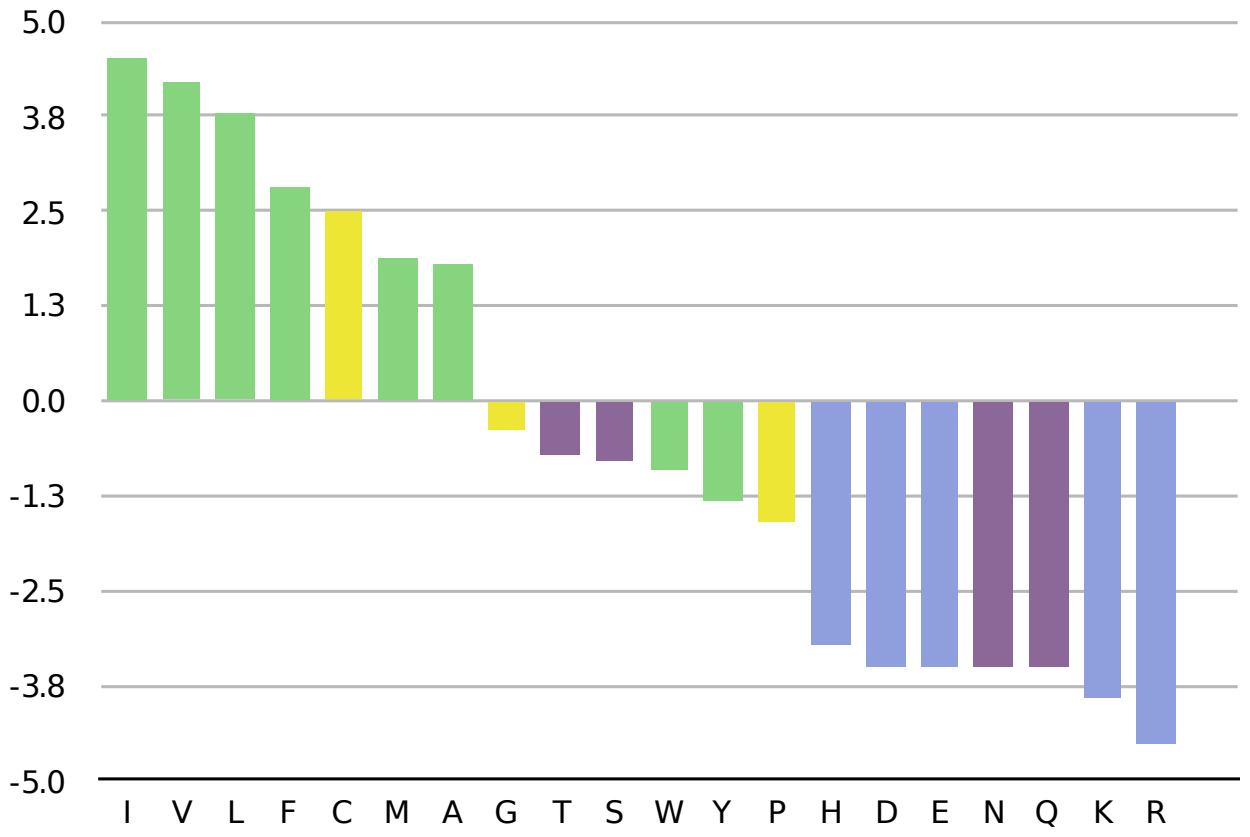
Figure 3-7
Molecular Cell Biology, Sixth Edition

Hydropathy scale

- Partitioning method:
  - two immiscible liquids (water + something that favours hydrophobic molecules — solvent)
  - measure free energy of transfer of each amino acid between phases
- (Other methods exist)

Kyte-Doolittle scale (Kyte & Doolittle, 1982)

Side-chain	Hydropathy index	$\Delta G^{\circ}_{transfer}$ (water–vapor)[a]	Fraction of side-chains 100% buried[b]	Fraction of side-chains 95% buried[c]
Isoleucine	4·5	4·4	4·5	5·2
Valine	4·2	4·2	4·3	4·2
Leucine	3·8	4·5	3·2	2·8
Phenylalanine	2·8	2·5	2·5	3·5
Cysteine/cystine	2·5	1·9	6·0	3·2
Methionine	1·9	1·9	1·0	1·9
Alanine	1·8	3·9	5·3	1·6
Glycine	−0·4	—	4·2	1·3
Threonine	−0·7	−0·6	−0·5	−1·0
Tryptophan	−0·9	−0·9	−2·4	−0·3
Serine	−0·8	−0·8	−0·7	−1·0
Tyrosine	−1·3	−1·1	−3·3	−2·2
Proline	−1·6	—	−2·4	−1·8
Histidine	−3·2	−4·2	−3·6	−1·9
Glutamic acid	−3·5	−3·9	−2·8	−1·7
Glutamine	−3·5	−3·5	−4·0	−3·6
Aspartic acid	−3·5	−4·5	−2·5	−2·3
Asparagine	−3·5	−3·8	−3·1	−2·7
Lysine	−3·9	−3·2	—	−4·2
Arginine	−4·5	—	—	—



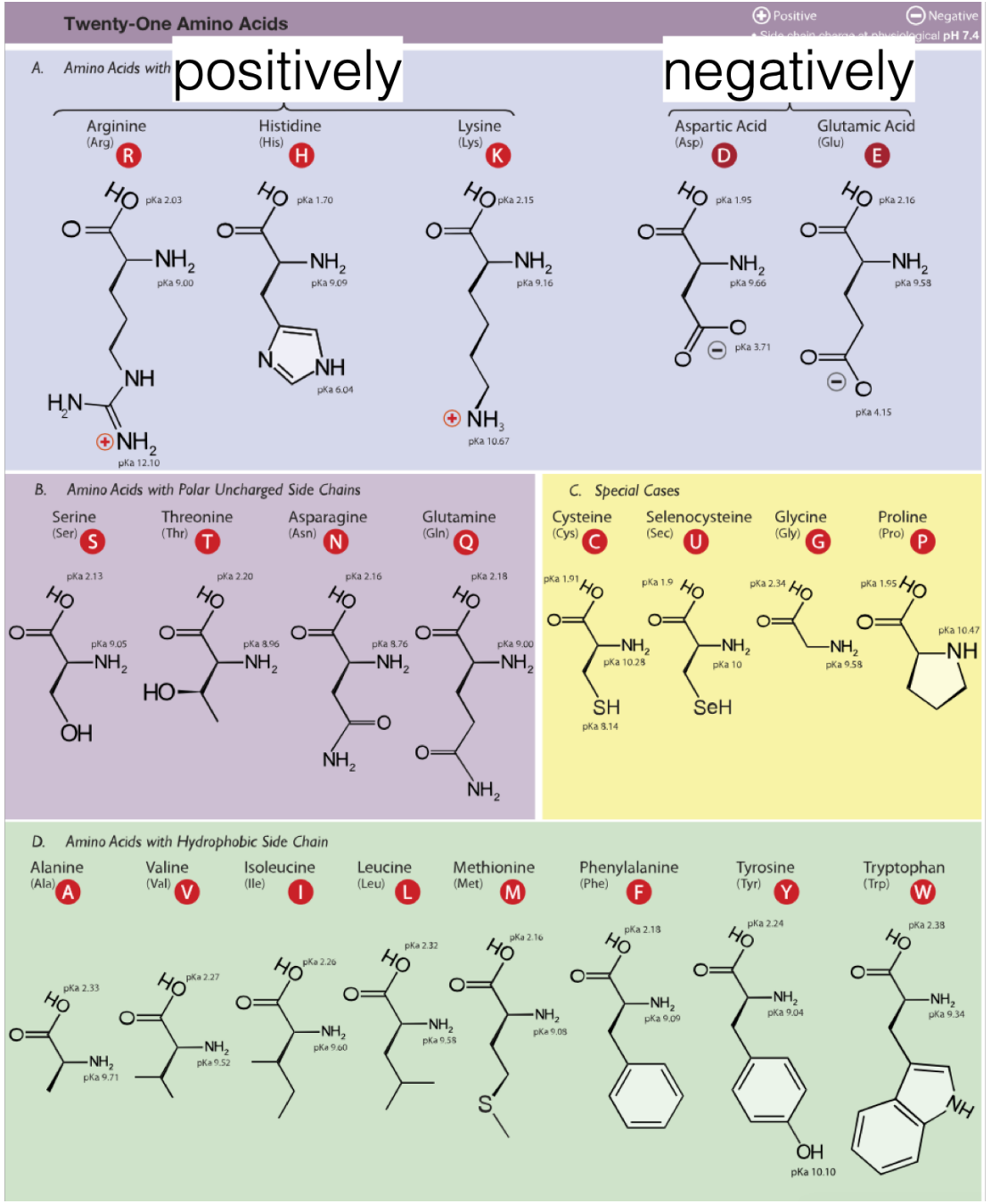
Amino acids have different chemical properties

charged

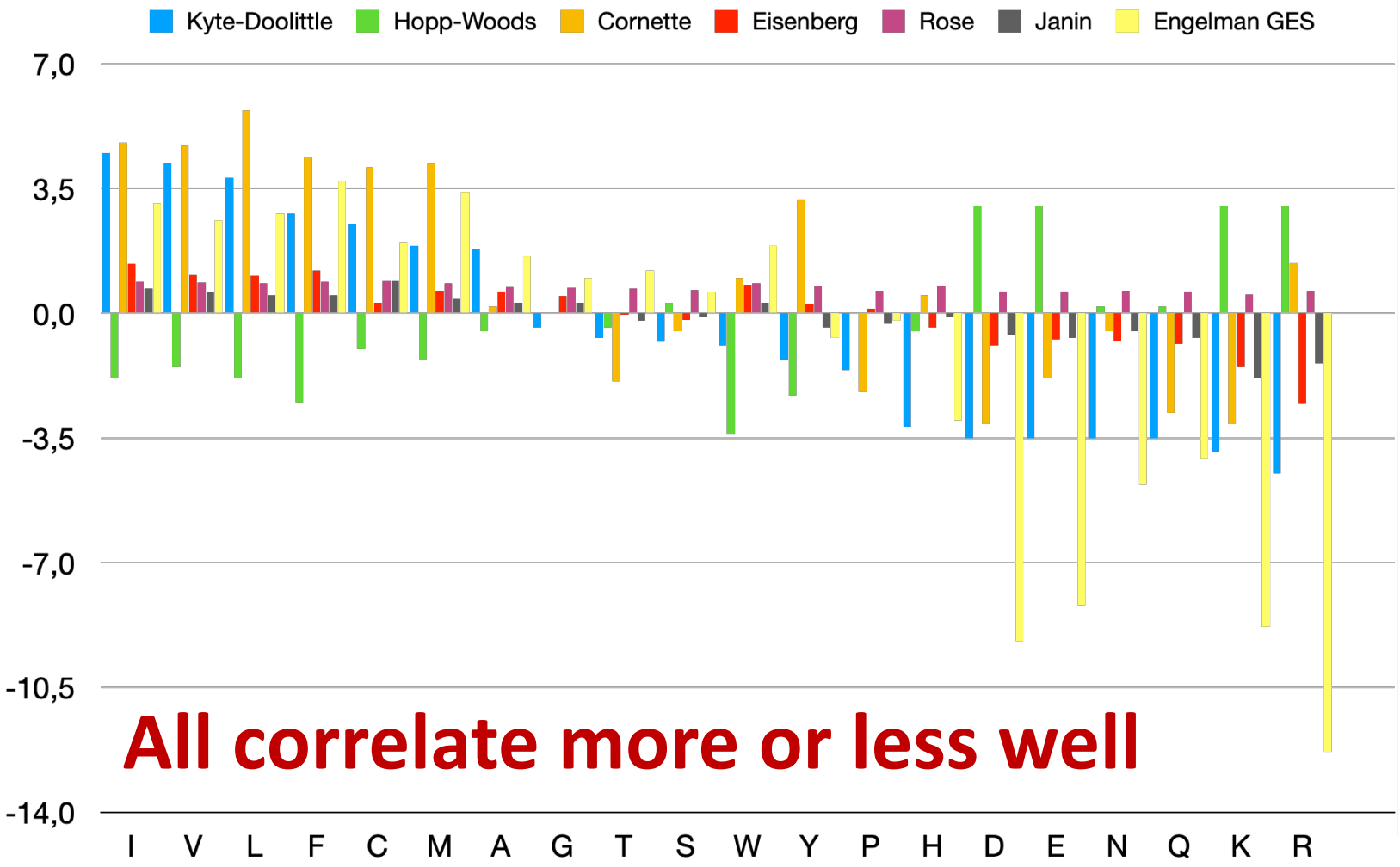
polar

hydrophobic

special



Other hydropathy scales



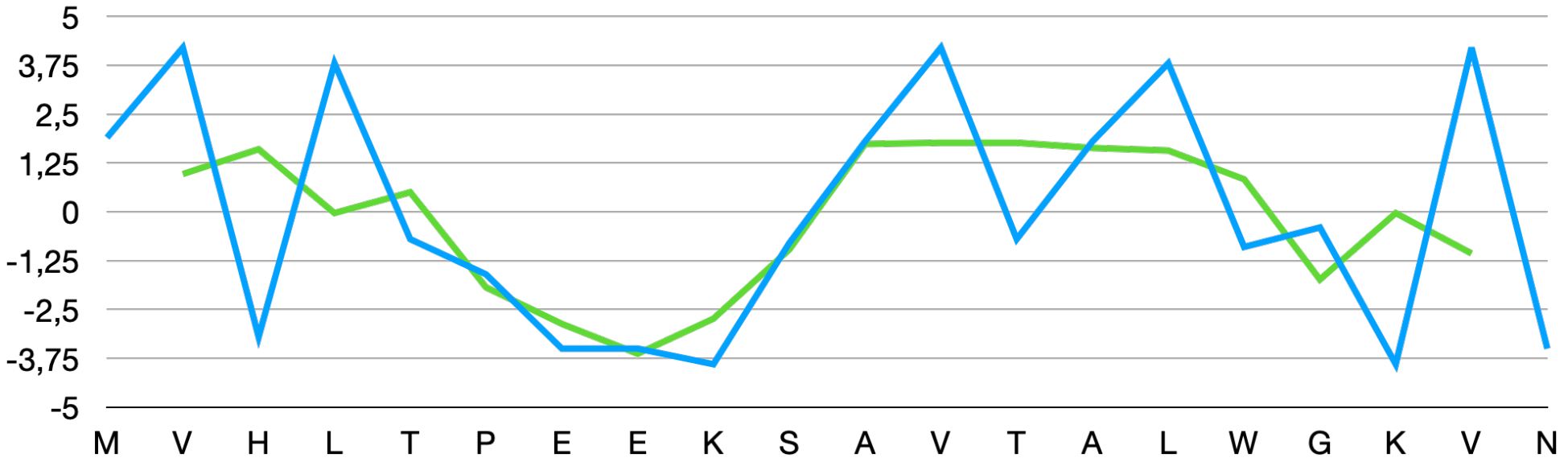
All correlate more or less well

Kyte and Doolittle (1)	Rose, et al (2)	Wolfenden, et al (3)	Janin (1979) (4)
Ile Val Leu Phe Cys Met,Ala Gly Thr,Ser Trp,Tyr Pro His Asn,Gln Asp,Glu Lys Arg	Cys Phe,Ile Val Leu,Met,Trp His Tyr Ala Gly Thr Ser Pro,Arg Asn Gln,Asp,Glu Lys	Gly,Leu,Ile Val,ala Phe Cys Met Thr,Ser Trp,Tyr Asp,Lys,Gln Glu,His Asp Arg	Cys Ile Val Leu,Phe Met Ala,Gly,Trp His,Ser Thr Pro Tyr Asn Asp Gln,Glu Arg Lys

Calculating hydrophobicity

- Human haemoglobin

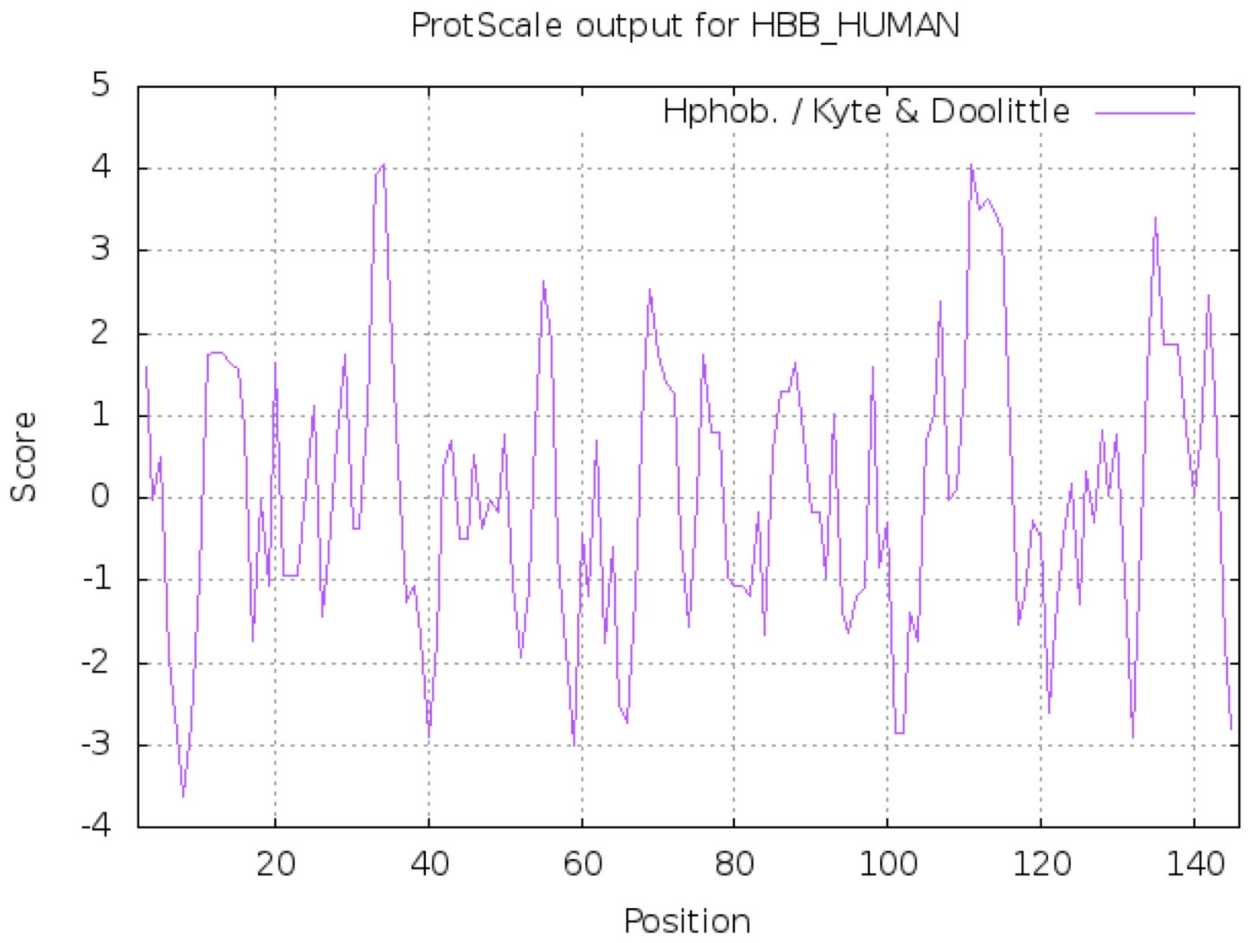
```
         10         20         30         40         50         60         70
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK VKAHGKKVLG
         80         90        100        110        120        130        140
AFSDGLAHLD NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG KEFTPPVQAA YQKVVAGVAN ALAHKYH
```



Average over 3 positions

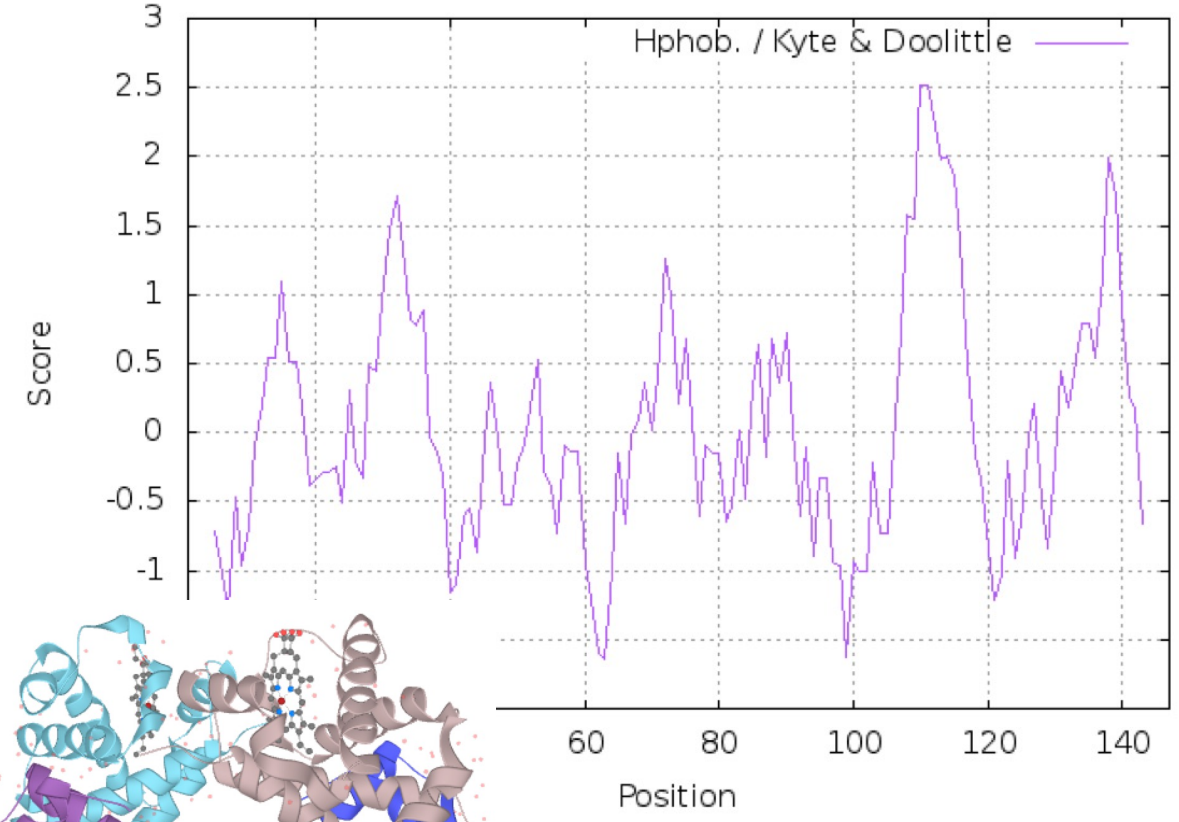
M	1,9	
V	4,2	0,96667
H	-3,2	1,6
L	3,8	-0,03333
T	-0,7	0,5
P	-1,6	-1,93333
E	-3,5	-2,86667
E	-3,5	-3,63333
K	-3,9	-2,73333
S	-0,8	-0,96667
A	1,8	1,73333
V	4,2	1,76667
T	-0,7	1,76667
A	1,8	1,63333
L	3,8	1,56667
W	-0,9	0,83333
G	-0,4	-1,73333
K	-3,9	-0,03333
V	4,2	-1,06667
N	-3,5	

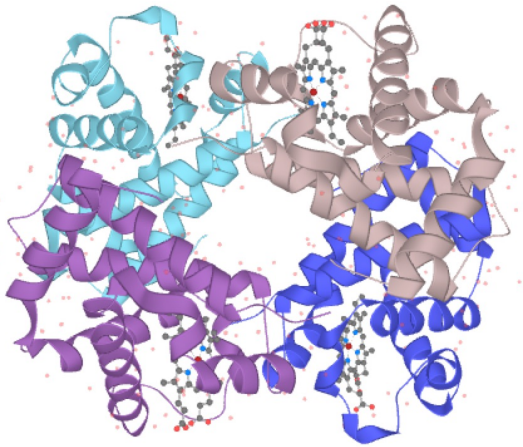
Calculating hydrophobicity



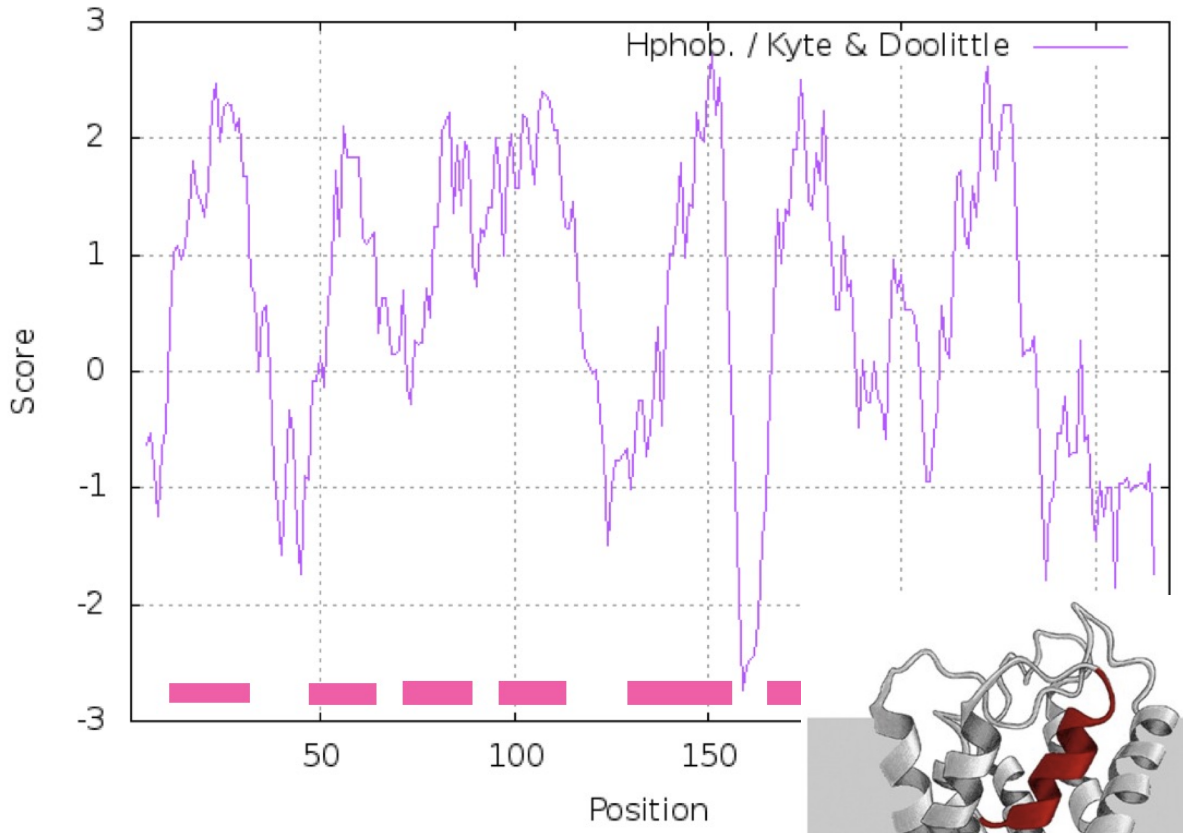
Globular vs. membrane proteins

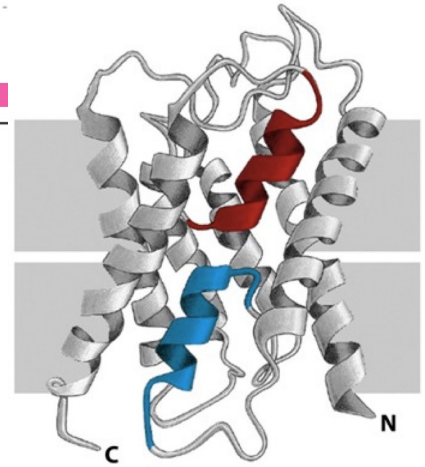
Prc Hemoglobin MAN





ProtSc: Aquaporin MAN



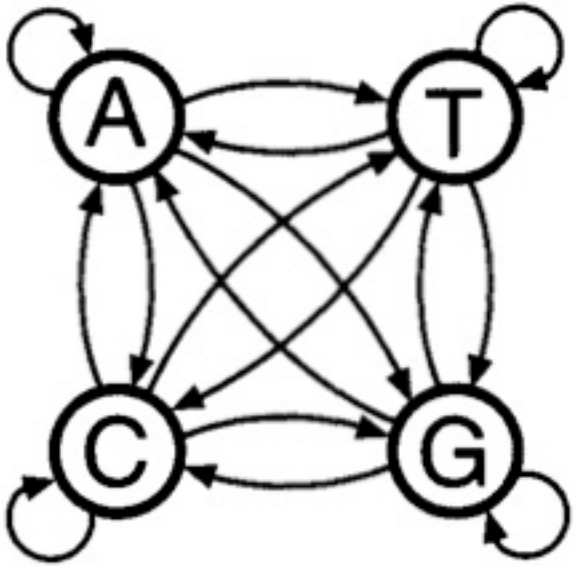


How do we know where hydrophobicity if "high enough"?

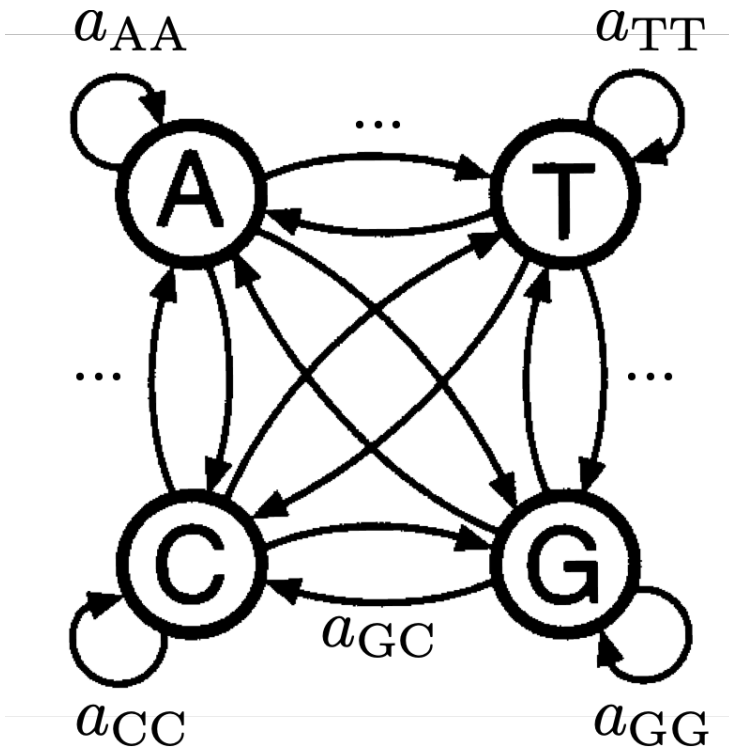
Hidden Markov models

CS detour: Markov chains

- **Markov property**: Current state of the model depends on its previous state(s)
- Example: in the genome of vertebrate, C's are avoided before G's (CpG context)
  - C's before G's get methylated, which leads to functional consequences
  - in some regions, CpG context is needed for functional reasons
- We can model this sequence with a probabilistic model that is aware of its previous state — **Markov chain**



Markov chains cont'd: parameters



- In the CpG example, letters: **states**
- We create a sequence as we travel between states
- Parameters: **transition probabilities**

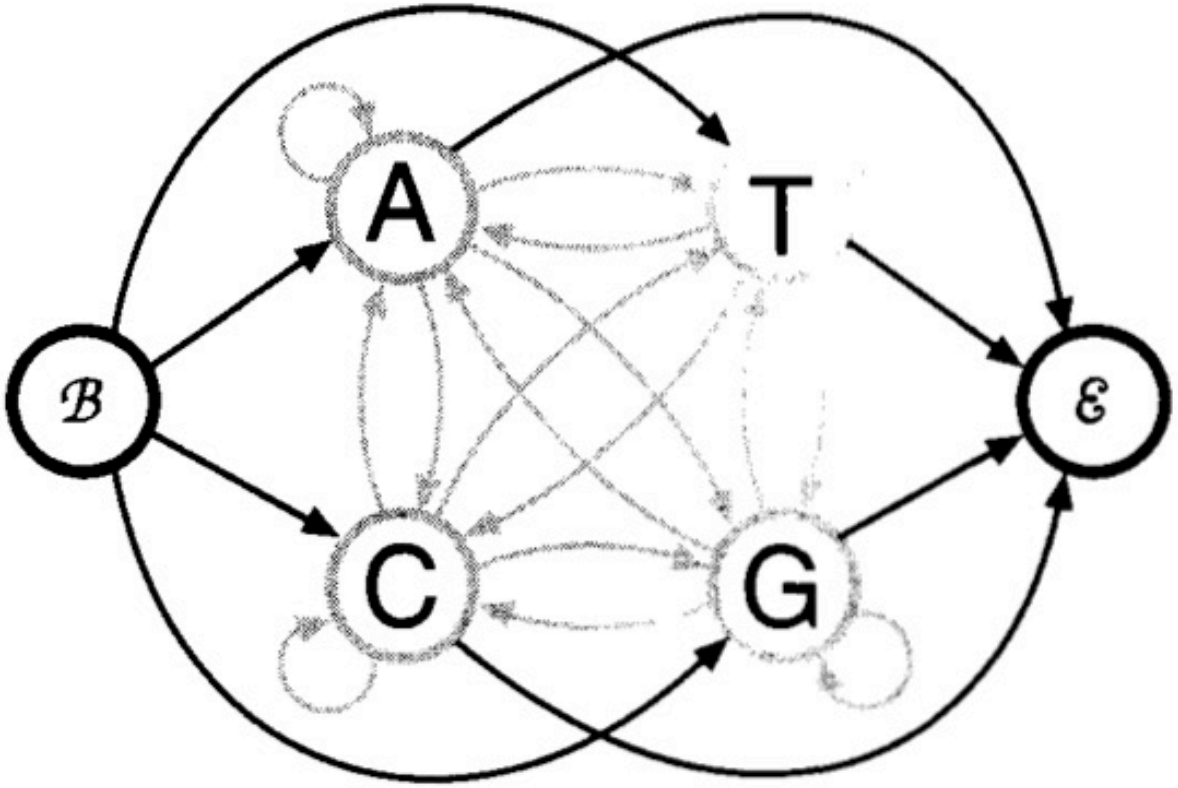
$$a_{st} = P(x_i = s | x_{i-1} = t)$$

- Probability of a sequence X:

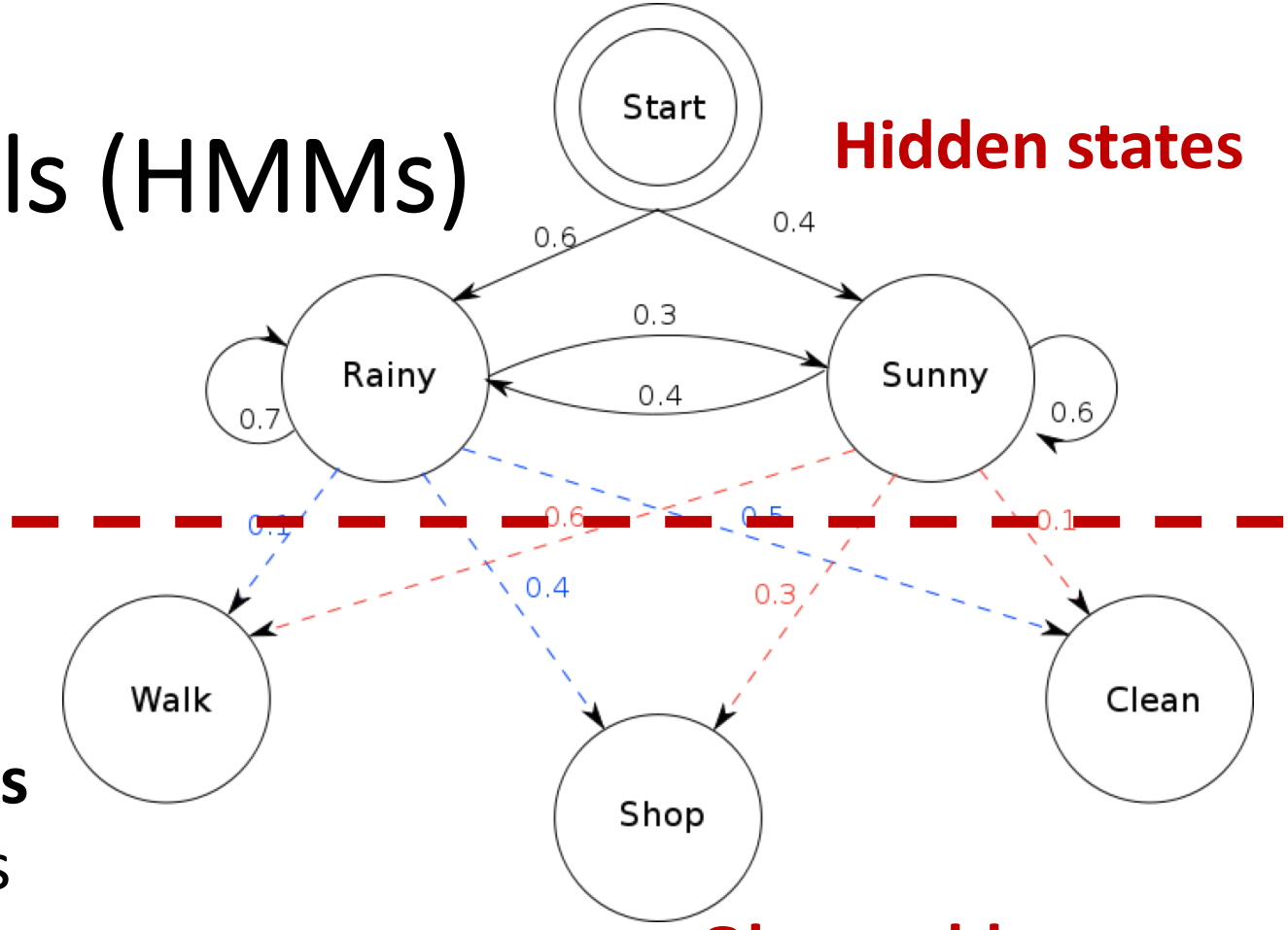
$$P(X) = P(x_L, x_{L-1}, \ldots, x_1) =$$
$$= P(x_L | x_{L-1}, \ldots, x_1) \cdot P(x_{L-1} | x_{L-2}, \ldots, x_1) \cdot \cdots \cdot P(x_2 | x_1) \cdot P(x_1) =$$
$$= P(x_L | x_{L-1}) \cdot P(x_{L-1} | x_{L-2}) \cdot \cdots \cdot P(x_2 | x_1) \cdot P(x_1) =$$
$$= \prod_{i=2}^{L} \boldsymbol{a_{x_i x_{i-1}}} \cdot \boldsymbol{P(x_1)}$$

Markov chains cont'd

- To get rid of $P(x_1)$,
  add technical **start** and **end** states
- Then $\boldsymbol{P(X) = \prod_{i=2}^{L} a_{x_i x_{i-1}}}$

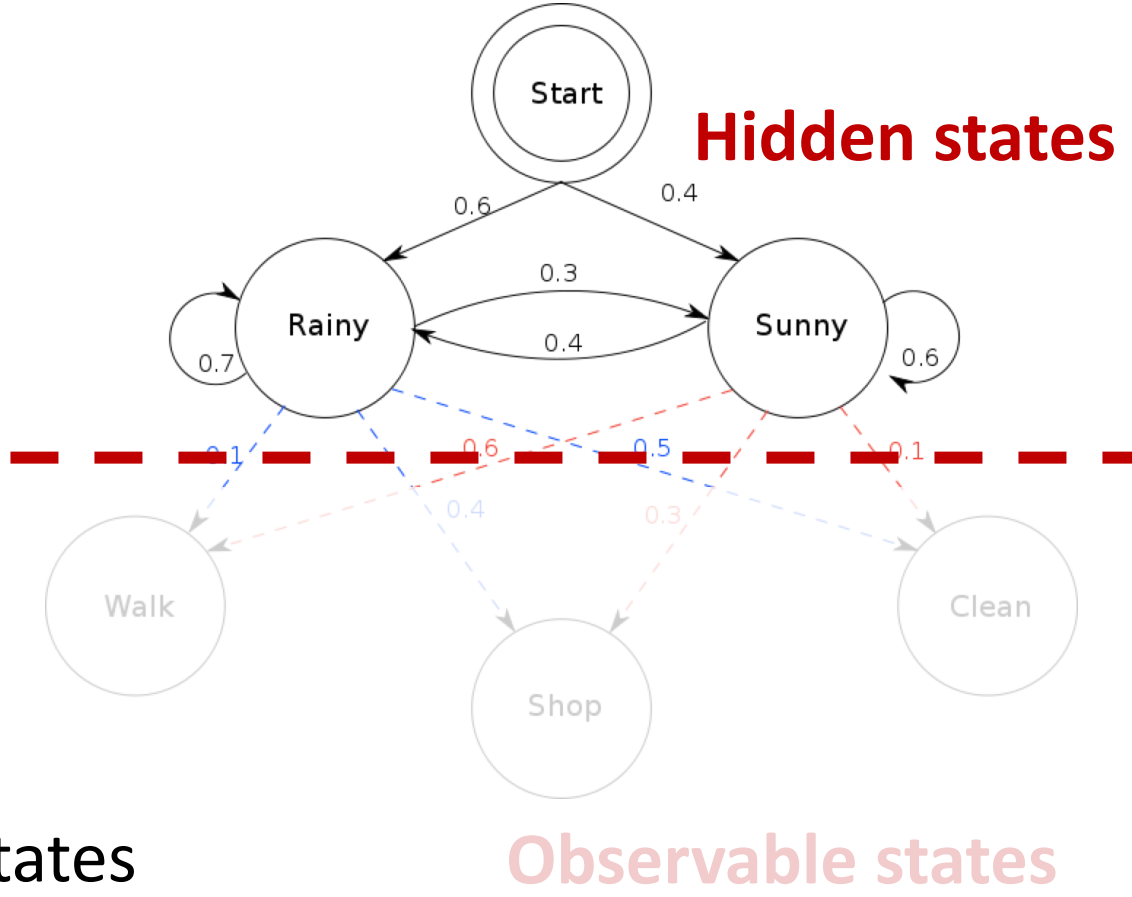


Hidden Markov models (HMMs)



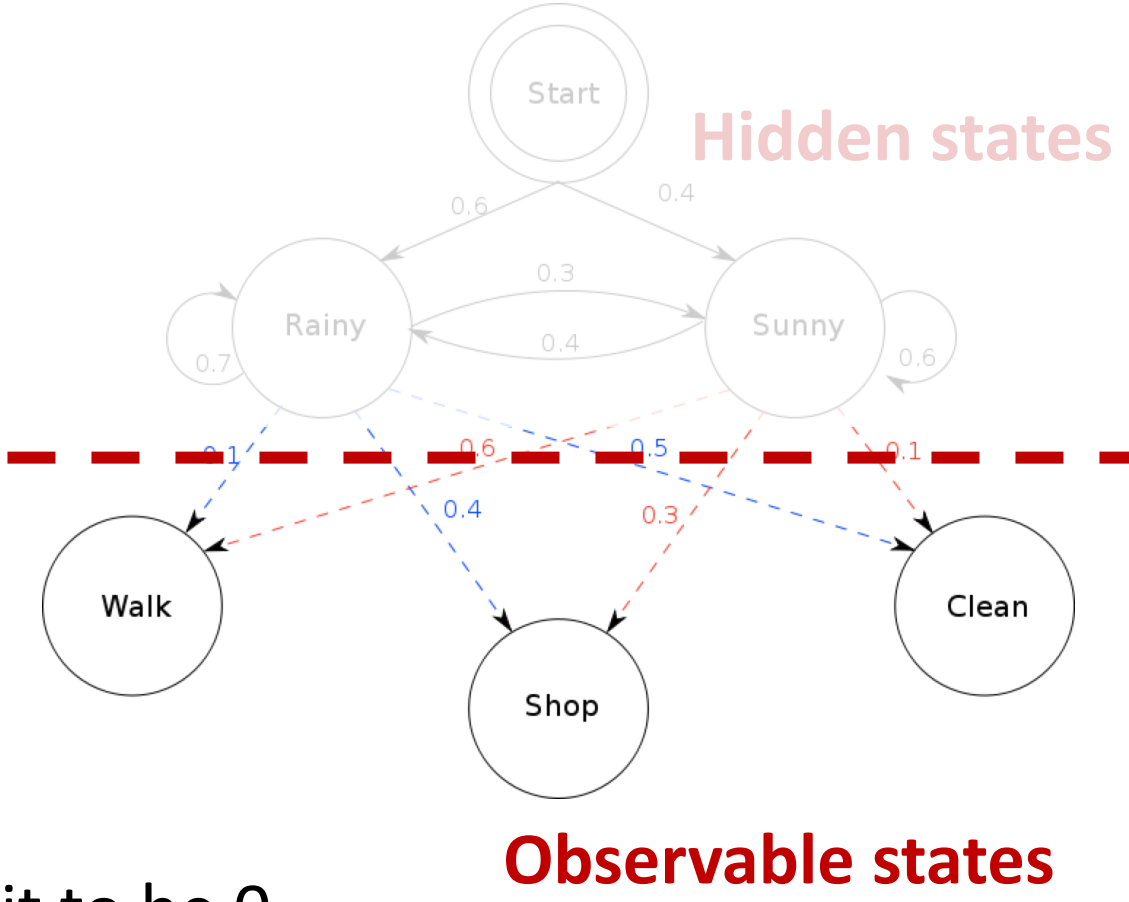
- The modelled system is assumed to be a **Markov process** over unobserved (**hidden**) states
- Hidden states emit observable states
- **Task**: given the sequence of observable states, **reconstruct the most probable sequence of the hidden states**

HMMs cont'd: parameters



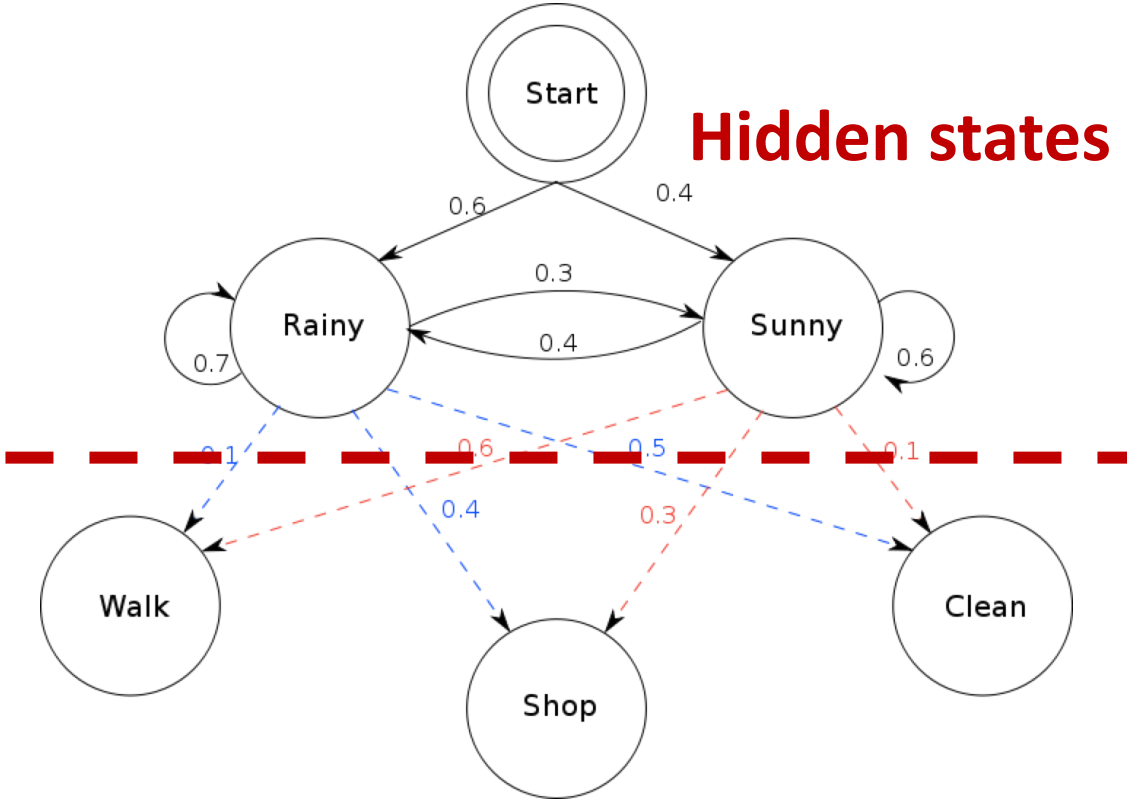
- **Hidden states (n)**:
  k transitions between the states =>
  $k \leq n \times n$ **transition probabilities**
  - assume non-existing transitions have a probability of 0, then n^2 transition probabilities between n states
- $n \times n$ matrix, $n \times (n-1)$ independent parameters
  - **Markov matrix**: stochastic, the state at time t depends only on the state at time $t-1$

HMMs cont'd: parameters



- **Observable states (m)**:
  n emitting states;
  m observable states =>
  $n \times m$ **emission probabilities**
  - if a state is not emitting,
    assume all emission probabilities of it to be 0
- Emission probabilities: $n \times m$ matrix, $n \times (m-1)$ independent parameters

HMMs cont'd: parameters



Hidden states (n) and observable states (m):

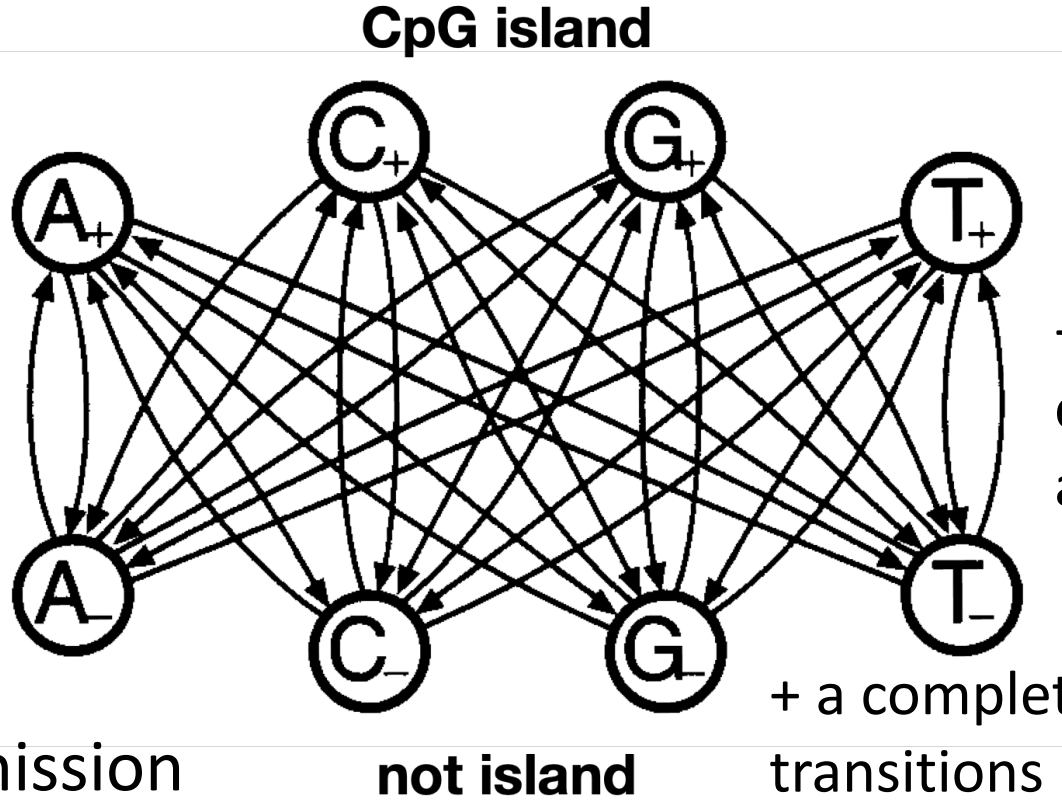
- Transition probabilities:
  $n \times n$ matrix, $n \times (n-1)$ parameters
- Emission probabilities:
  $n \times m$ matrix, $n \times (m-1)$ parameters
- => $n \times (n-1) + n \times (m-1) = \boldsymbol{n \times (m+n-2)}$ **independent parameters**

Modelling sequences with HMMs

- In a HMM, states **do not correspond** to observed characters
- States can **emit** observed characters (states)
- Example: in the genome of vertebrate, C's are avoided before G's (CpG context)
  - C's before G's get methylated, which leads to functional consequences (beyond the scope of this course)
  - in some regions, CpG context is needed for functional reasons — **CpG islands**
- Find CpG island in an annotated sequence using an HMM
- **=> hidden states: island / non-island; observable states: {A, T, G, C}**

Model topology: piece of creativity

- For CpG islands:



+ chances of going from one component to another relatively low

+ a complete set of transitions inside each component

- Transition and emission probabilities have to be estimated from annotated sequences

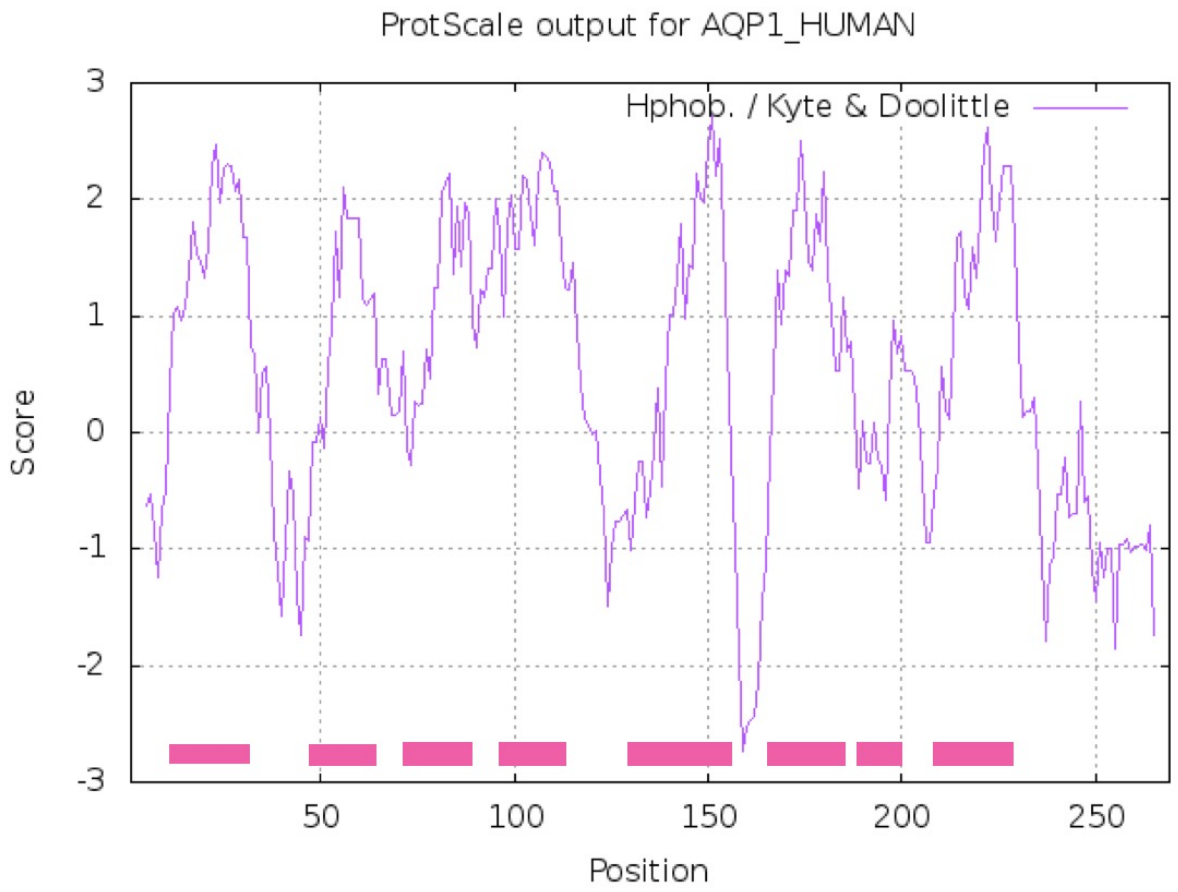
For details: See Durbin et al., “Biological Sequence Analysis”, Cambridge University Press, 1998

Model topology (a.k.a. HMM architecture)

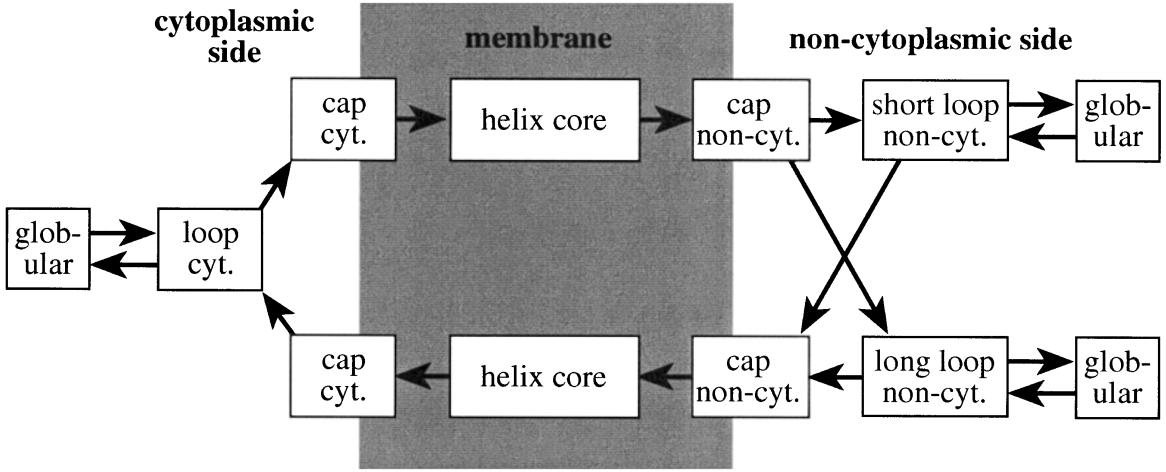
- How to model a sequence of a certain length?
- Suppose first 5 characters of a sequence are always **A**, second five are always **C**, then **A's** again, etc.?
- Suppose we have two sets of distributions for emission probabilities for different parts of the sequence (membrane vs. non-membrane regions)?

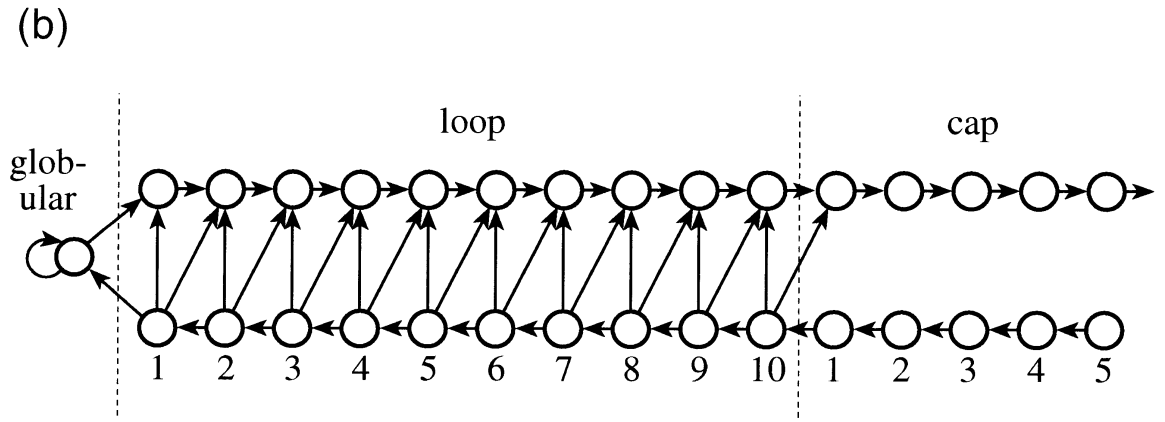
TMHMM (Krogh, Larsson, von Heijne, Sonnhammer, *J Mol Biol*, 2001)

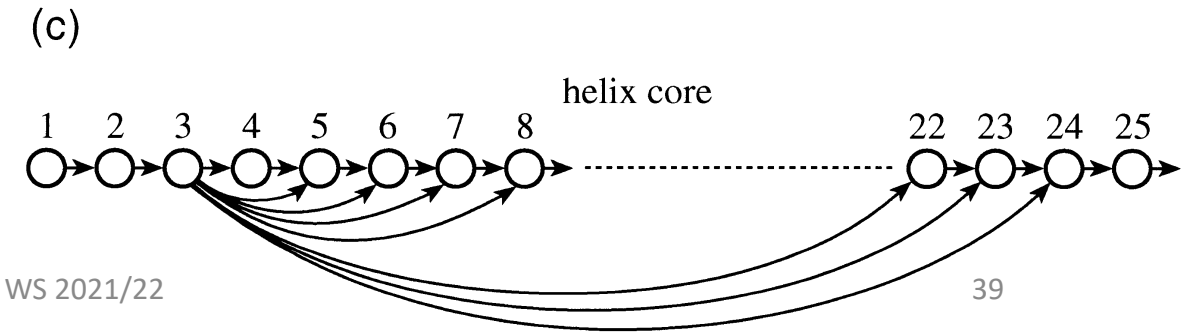
- Predict TM segments from protein sequences
- TM segments correspond to more hydrophobic regions
- Idea: predict TM segments using an HMM



HMM architecture



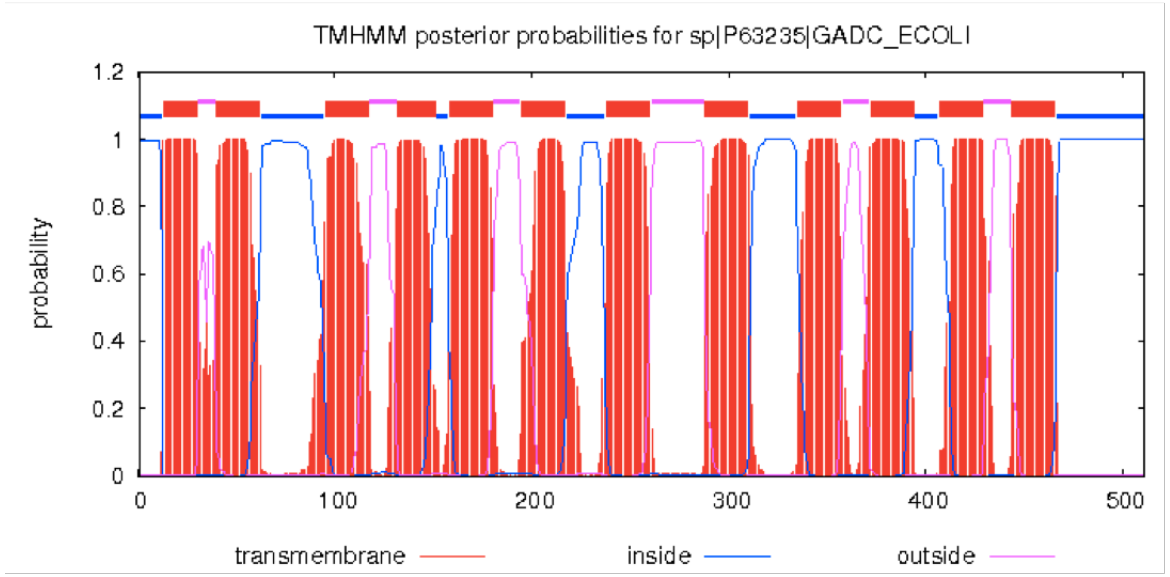


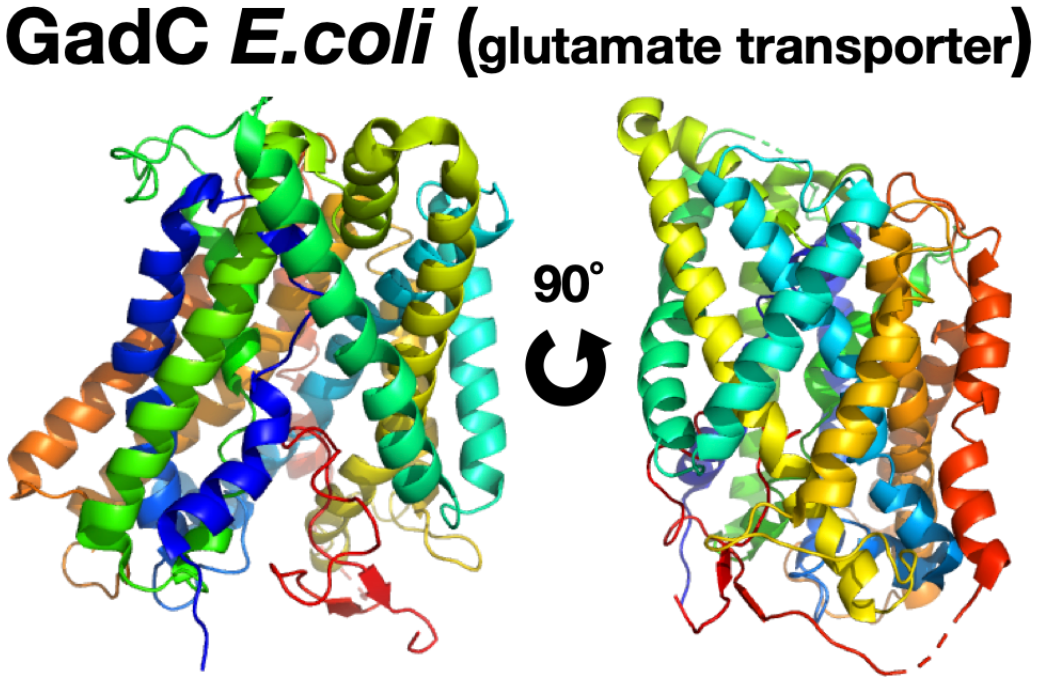


TMHMM performance

	Cross-validation		10-fold cross-validation Mean and std. dev.	
Number of proteins	160			
of which single-spanning:	52	32.50 %		
Correctly predicted topology:	124	77.50 %	120.2	1.3
Invertedly predicted topology:	11	6.88 %	10.5	0.9
Correctly predicted N-terminal:	141	88.12 %	138.0	1.3
Under-predictions:	16	10.00 %	18.4	1.4
of which single-spanning:	1	0.62 %	0.6	0.5
Over-predictions:	12	7.50 %	14.1	0.6
of which single-spanning:	7	4.38 %	7.0	0.2
Both over- and under-predictions:	3	1.88 %	3.60	0.8
of which single-spanning:	1	0.62 %	0.58	0.5
Total number of real helices:	696			
Number of over-predicted helices:	17	2.44 %	20.1	0.6
Number of under-predicted helices:	19	2.73 %	21.7	1.8
Number of shifted helix predictions:	0		0.33	0.5
Number of falsely merged helices:	0		0.50	0.6
Number of falsely split helices:	0		0	0

TMHMM performance: example

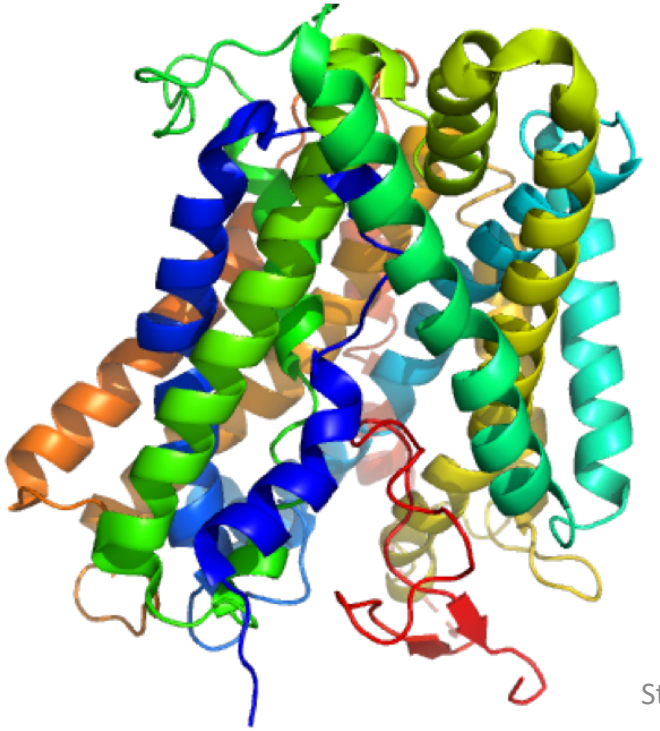


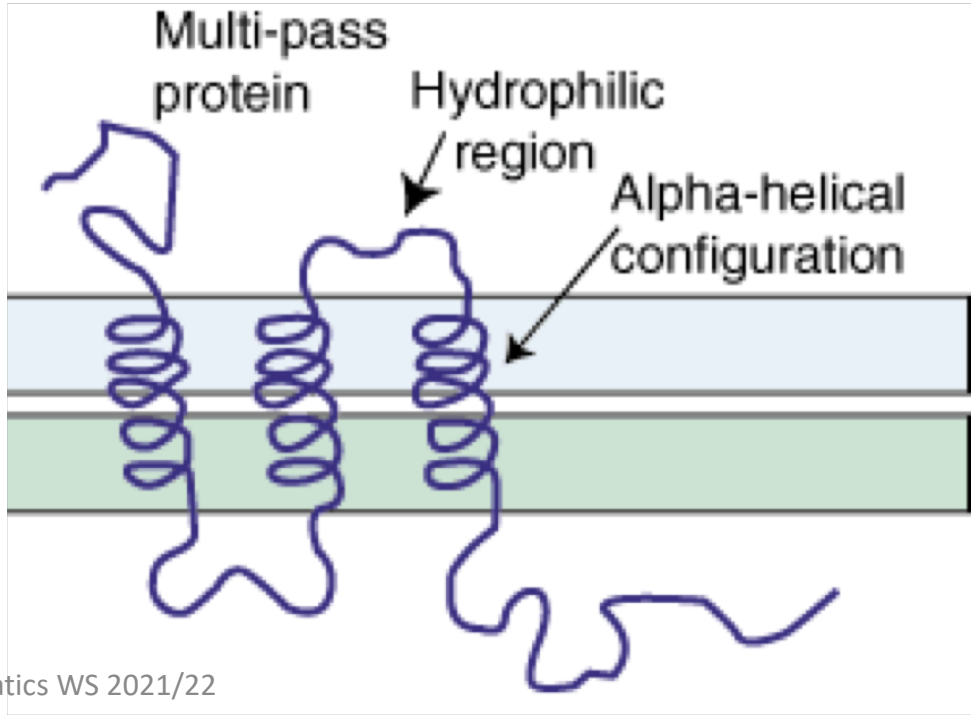


TMHMM: a caveat

- Can’t account for even slightly non-standard topologies

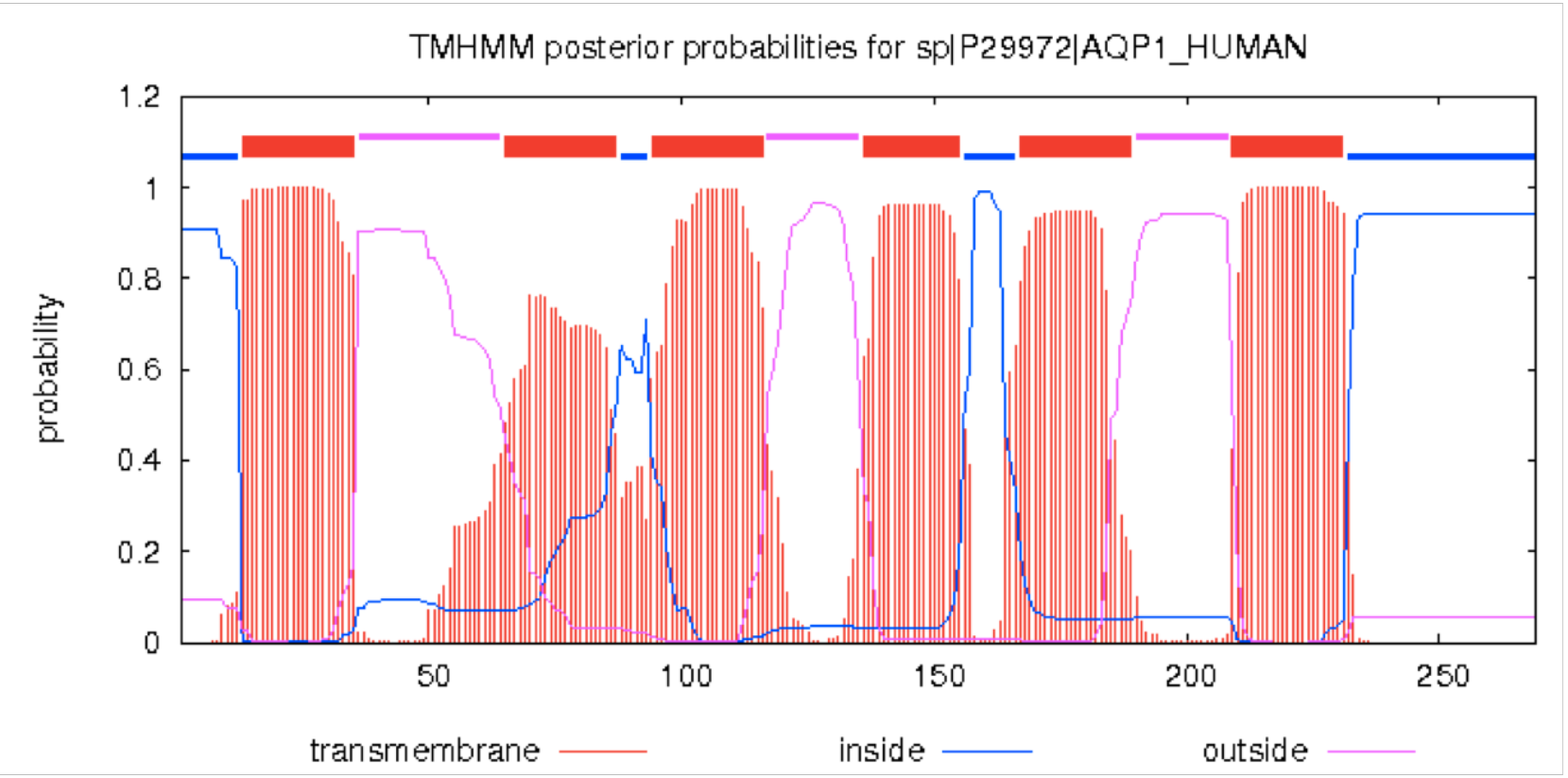
Standard topology of multi-pass α-helical proteins

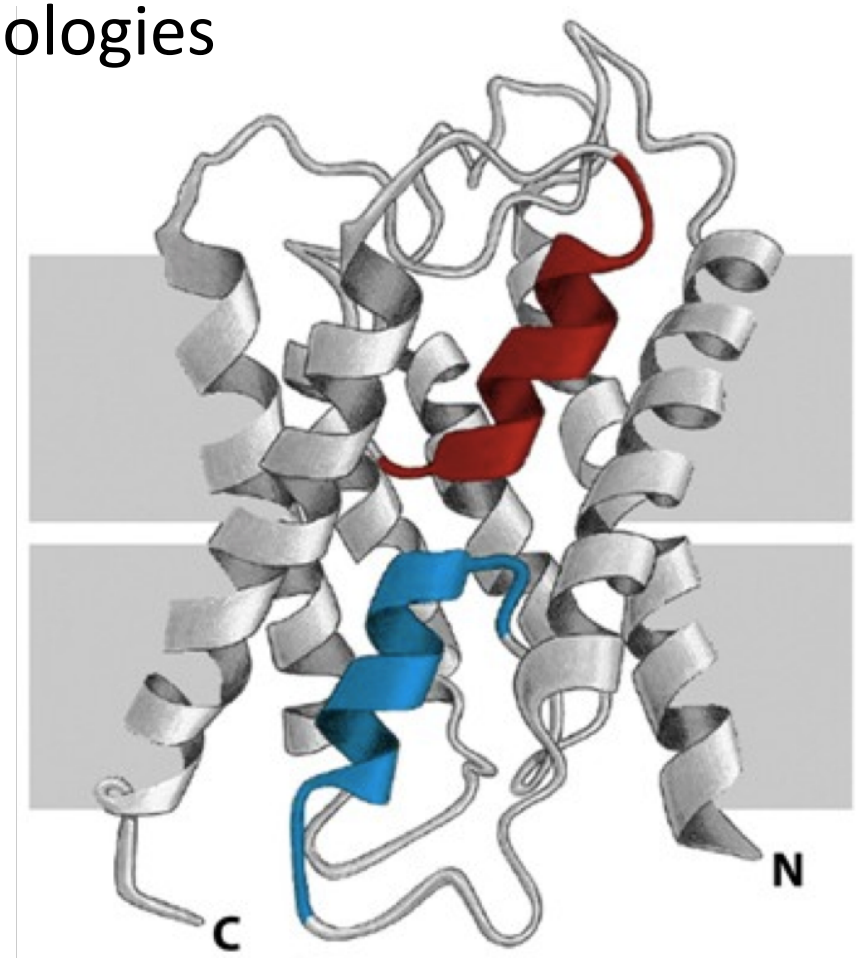




TMHMM: a caveat

- Can’t account for even slightly non-standard topologies





Prediction of secondary structure

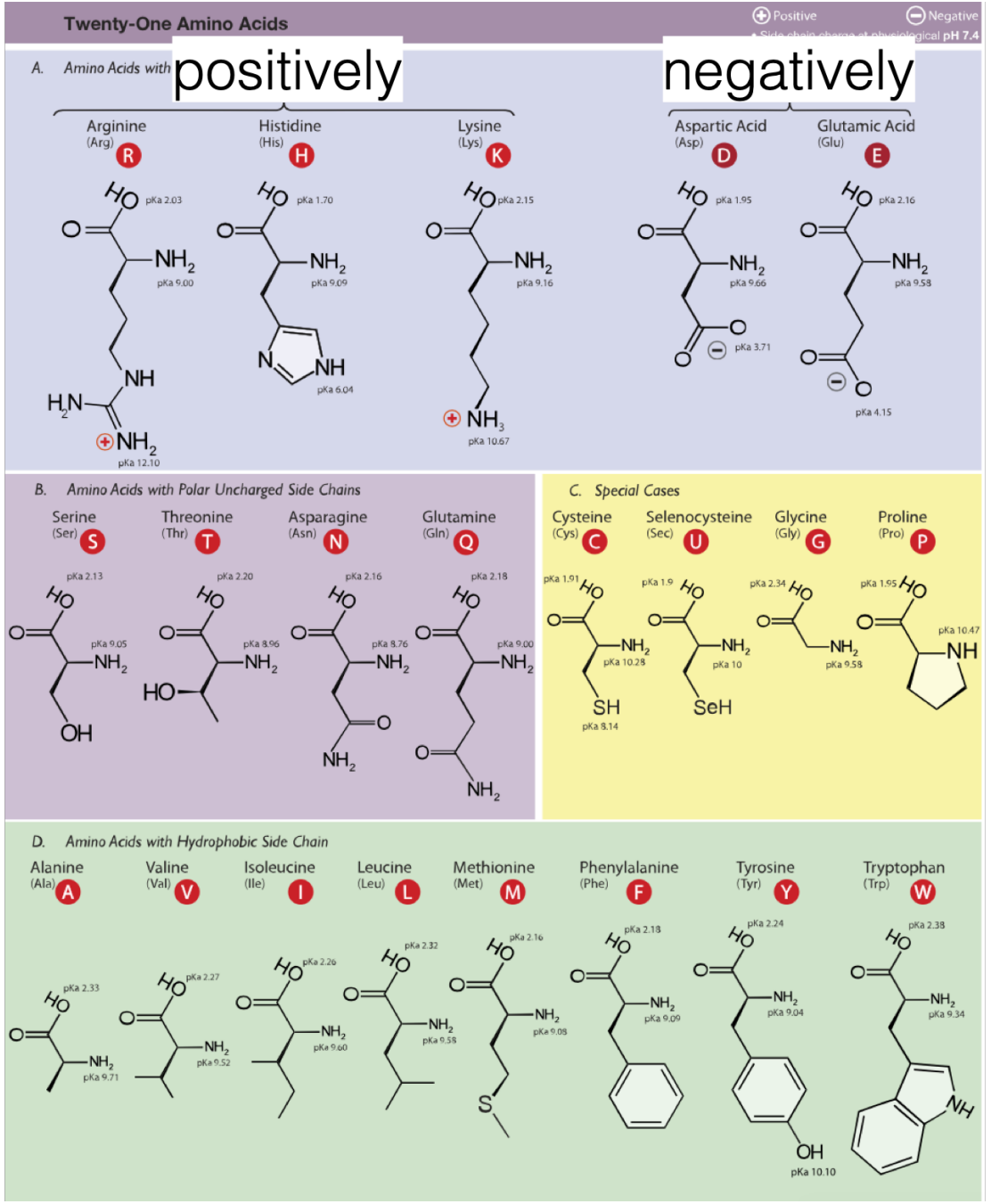
Amino acids have different chemical properties

charged

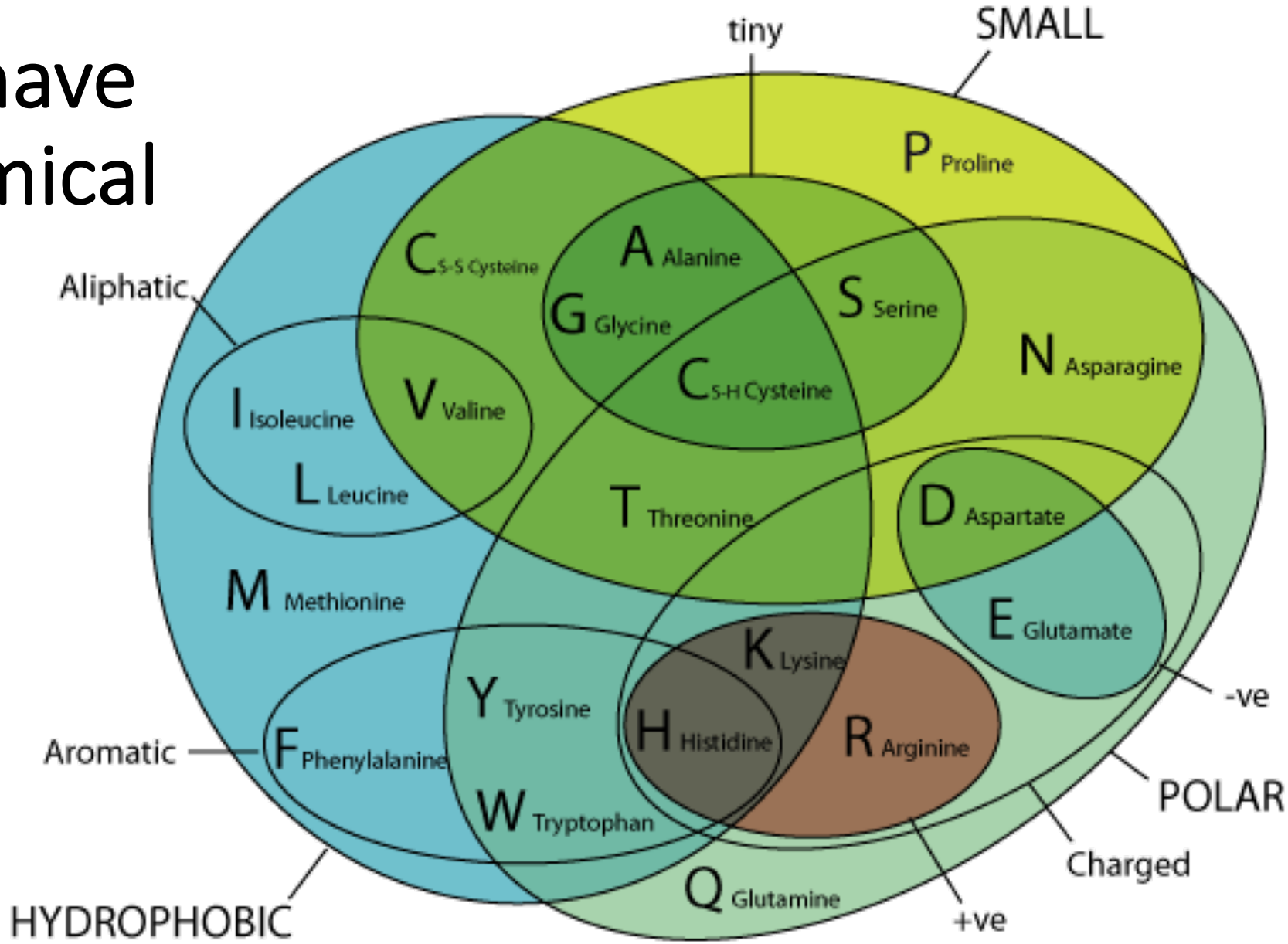
polar

hydrophobic

special

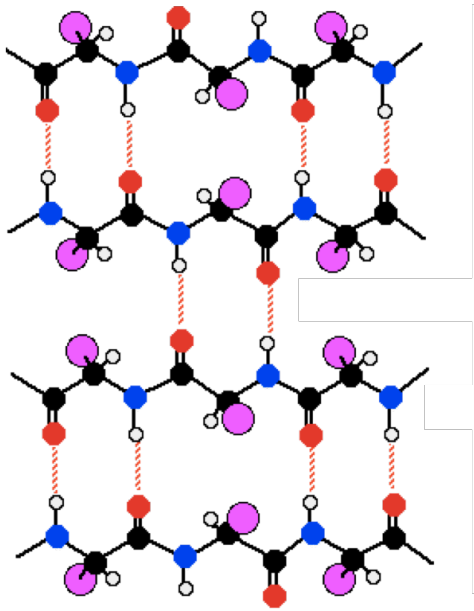


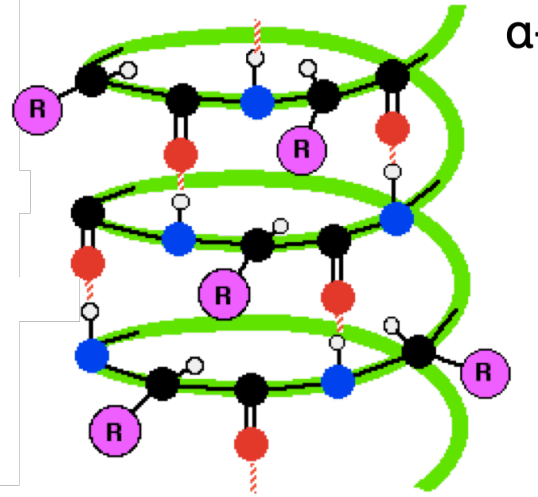
Amino acids have different chemical properties



Recap: most important secondary structure elements

β-sheet

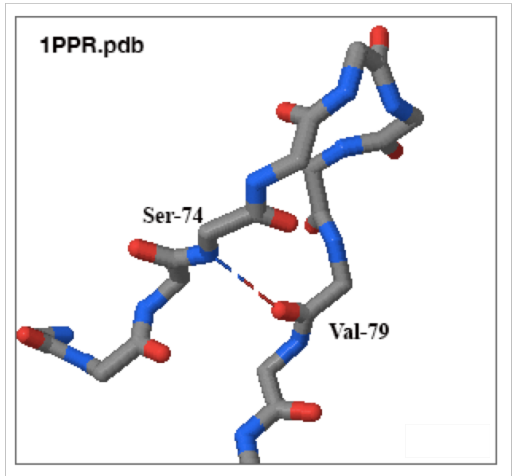


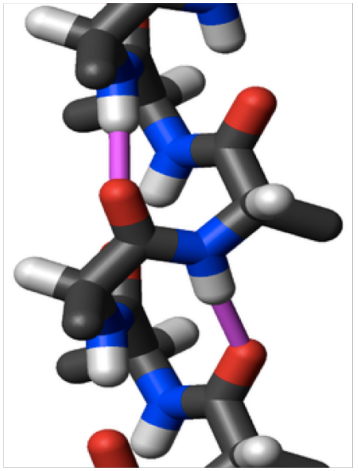


α-helix

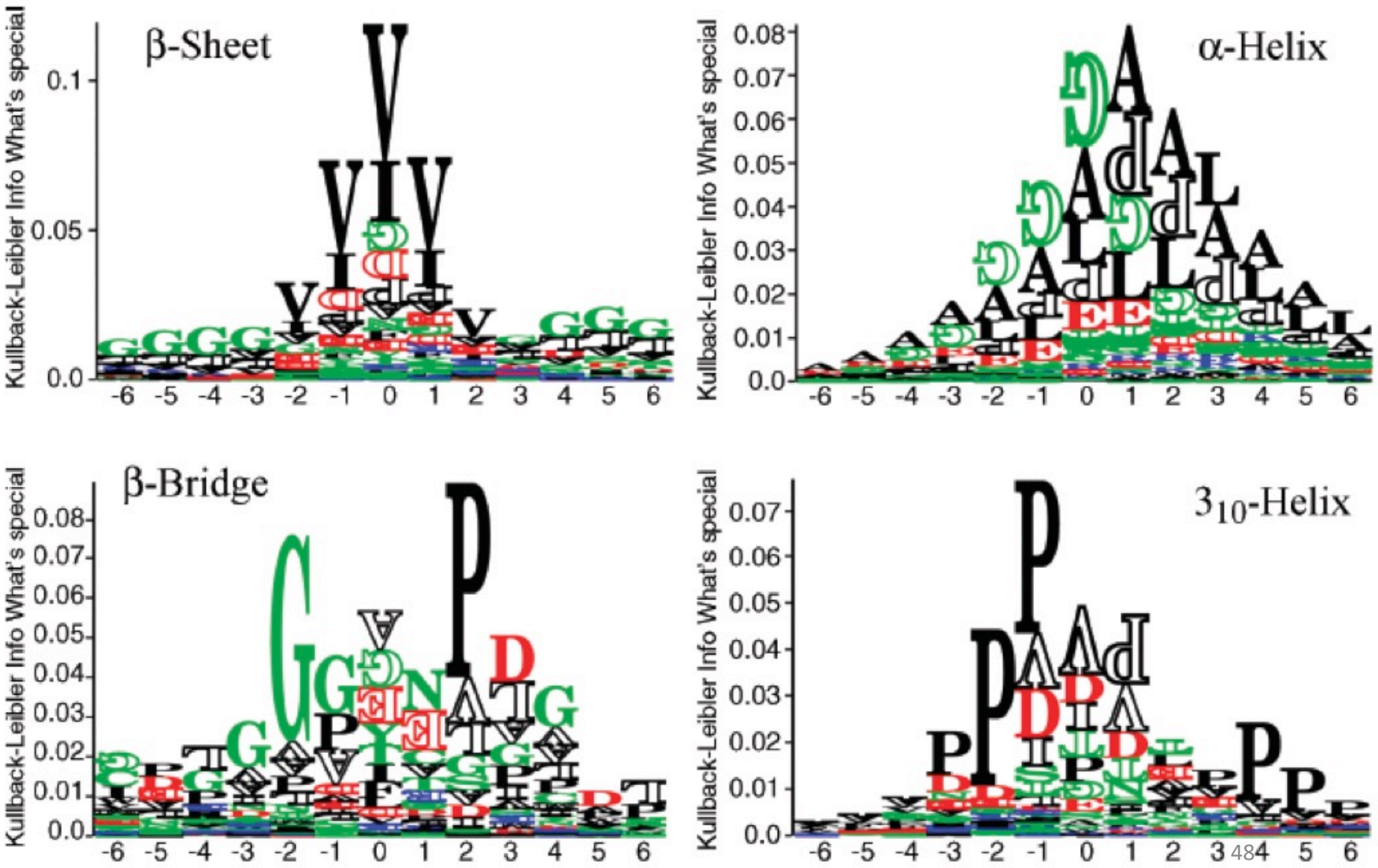
β-bridge: two β-strands connected by a single H-bond (rare)

3_{10}-helix: *i* + 3 → *i* H-bonding; 10-15% of all helices



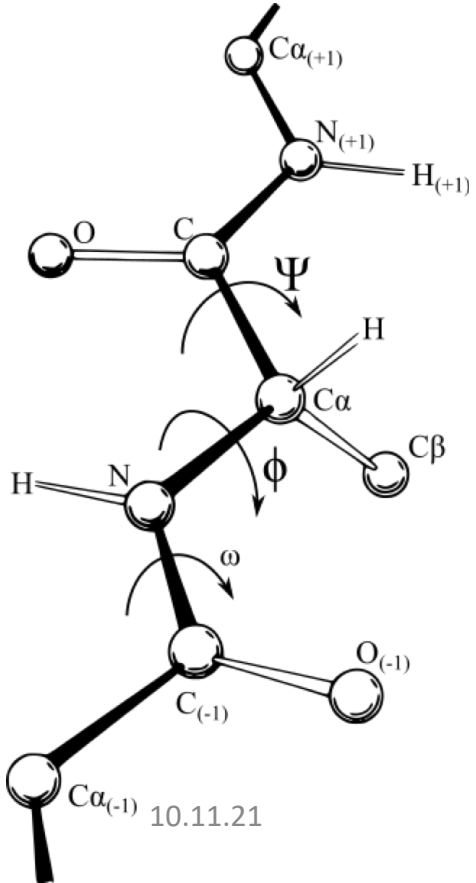


Amino acid preferences in secondary structure elements

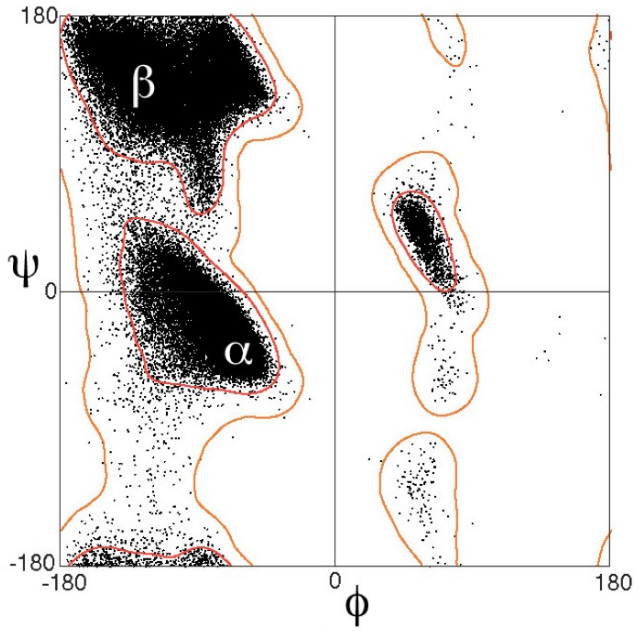


Ramachandran plot: unusual amino acids

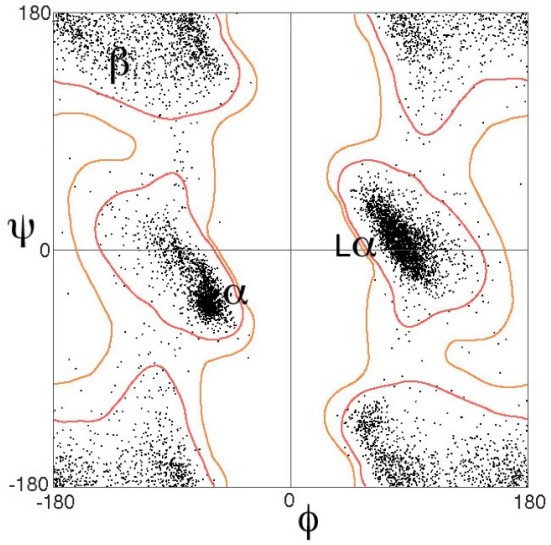
- Steric hindrance: the reason for disallowed regions in the Ramachandran plots



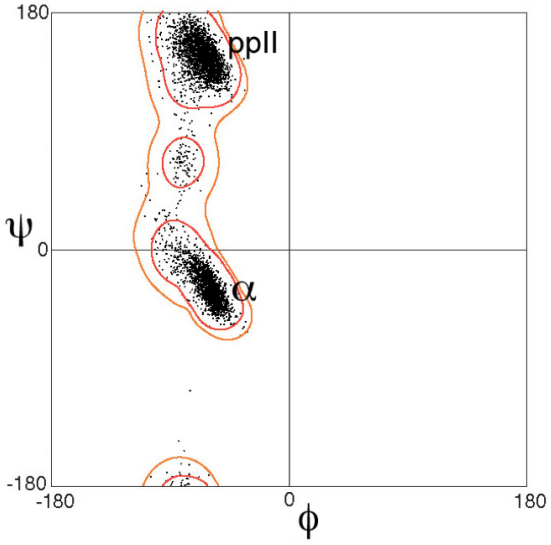
Almost all amino acids



Glycine



Proline



Secondary structure assignment: DSSP

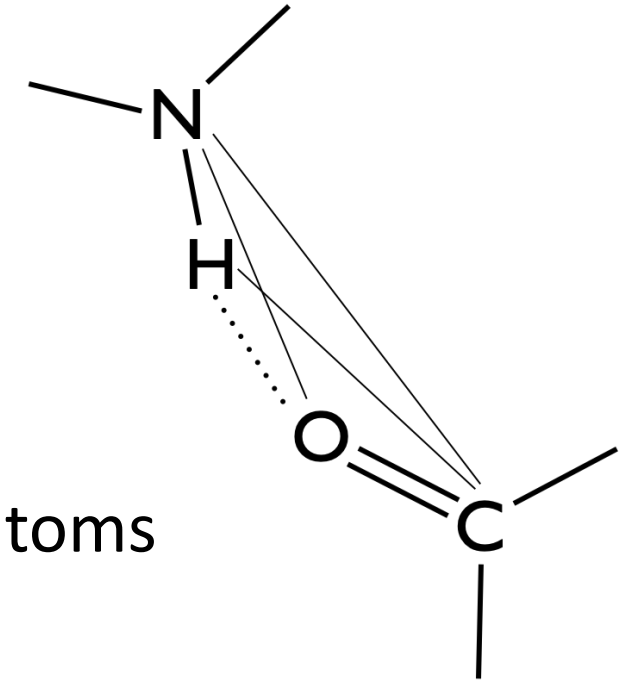
- **Assignment**: from known coordinates
- DSSP (Kabsch & Sander, 1983): H-bond energy from geometry => H-bonding pattern => secondary structure

$$E = 0.084\left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}\right) \cdot 332 \text{ kcal/mol}$$

0.084: product of partial charges

332: dimensionality factor

- **H-bond exists, if $E < -0.5$ kcal/mol**
- Position of H can be inferred from positions of other atoms (parallel to C=O, $r_{NH} = 1$)



DSSP cont'd

- H-bond energy from geometry => H-bonding pattern => secondary structure
- Define elementary H-bonding patterns:
  - **n-turn**: H-bond between residues (i, $i + n$), n = 3,4,5 => **helix**
  - **Bridge** between ($i - 1$, i, $i + 1$) and ($j - 1$, j, $j + 1$):
    - parallel (i, j): (H-bond($i - 1$, j) && H-bond(i, $j + 1$)) || (H-bond($j - 1$, i) && H-bond(j, $i + 1$))
    - anti-parallel (i, j): (H-bond(i, j) && H-bond(j, i)) || (H-bond($i - 1$, $j + 1$) && H-bond($j - 1$, $i + 1$))

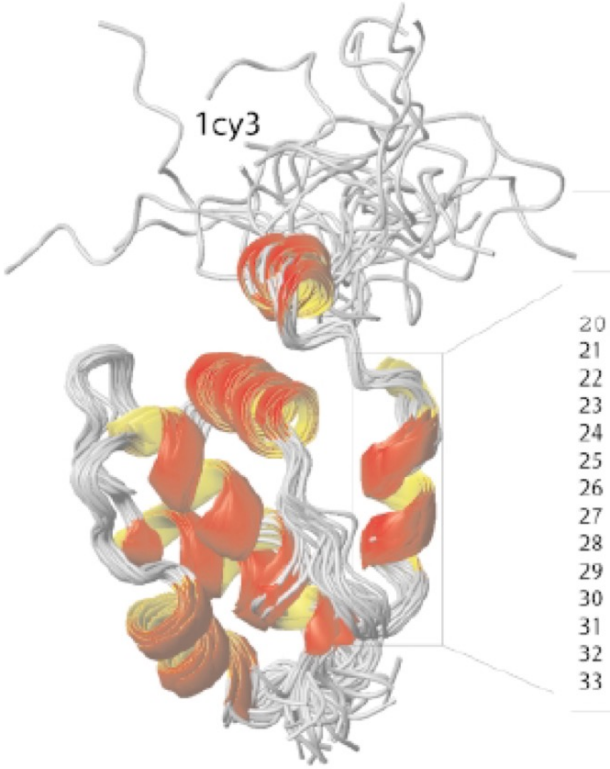
  => **beta-ladder** => **sheet**

Reminder: DSSP code

- H = alpha helix
- B = residue in isolated beta-bridge
- E = extended strand, participates in beta ladder
- G = 3-helix (3/10 helix)
- I = 5 helix (pi helix)
- T = hydrogen bonded turn
- S = bend
- C/- = no secondary structure (coil)

DSSPcont: Continuous secondary structure assignment (Carter et al., 2003)

- 9 different H-bond energy thresholds (−1.0 kcal/mol to −0.2 kcal/mol) weighted using NMR structures
- difference between assignment using single models vs. for the average model

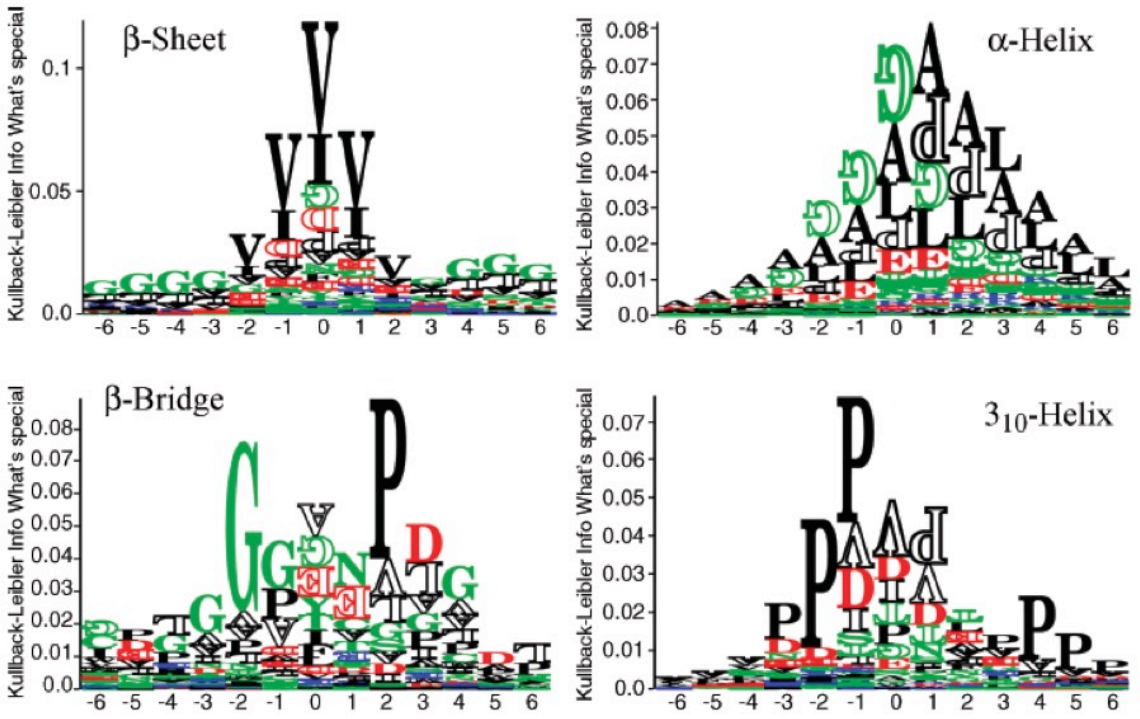


		(a) DSSP assignment for models 1-23	(b) DSSPcont assignment for model 1						
		1 5 10 15 20 23	G	H	I	T	E	B	S	L
20	Val	LLLLLLLLLLLLLLLLLLLLLLL	0	0	0	32	0	0	0	68
21	Ser	TLLHTLLTLLLLHLLLHLLTLLT	0	0	0	100	0	0	0	0
22	Glu	TTTHTTTTHTSSHTSTHSSTSTT	0	0	0	100	0	0	0	0
23	Glu	STTHHTTHHTSSHTHTHSSSHTH	0	32	0	21	0	0	47	0
24	Ser	HHHHHHHHHHHHHHHHHHHHHHH	0	99	0	1	0	0	0	0
25	Leu	HHHHHHHHHHHHHHHHHHHHHHH	0	99	0	1	0	0	0	0
26	Asn	HHHHHHHHHHHHHHHHHHHHHHH	0	99	0	1	0	0	0	0
27	Lys	HHHHHHHHHHHHHHHHHHHHHHH	1	99	0	0	0	0	0	0
28	Val	HTHTHHTHHTHHHTHHHHHTKHH	1	87	0	12	0	0	0	0
29	Arg	TTTTTTTTHTTTTTTTTTTTTTT	1	53	0	46	0	0	0	0
30	Asn	SSTSSSSSSTSSSSSTSTSSSTS	0	0	0	10	0	0	90	0
31	Arg	SSTSTSSSSTTSSTTTTTTSSTS	0	0	0	32	0	0	68	0
32	Glu	SSTSTLSLSTTSSTTTTTTLSLL	0	0	0	32	0	0	68	0
33	Glu	LLLLLLLLLLLLLLLLLLLLLLL	0	0	0	0	0	0	0	100

Prediction of secondary structure

- **NO 3D coordinates**

- Based on amino acid propensities to be in certain secondary structure elements
- Single amino acids or k-mers
- Information from one sequence or from multiple sequences can be used



CS detour: Intro to classification problems

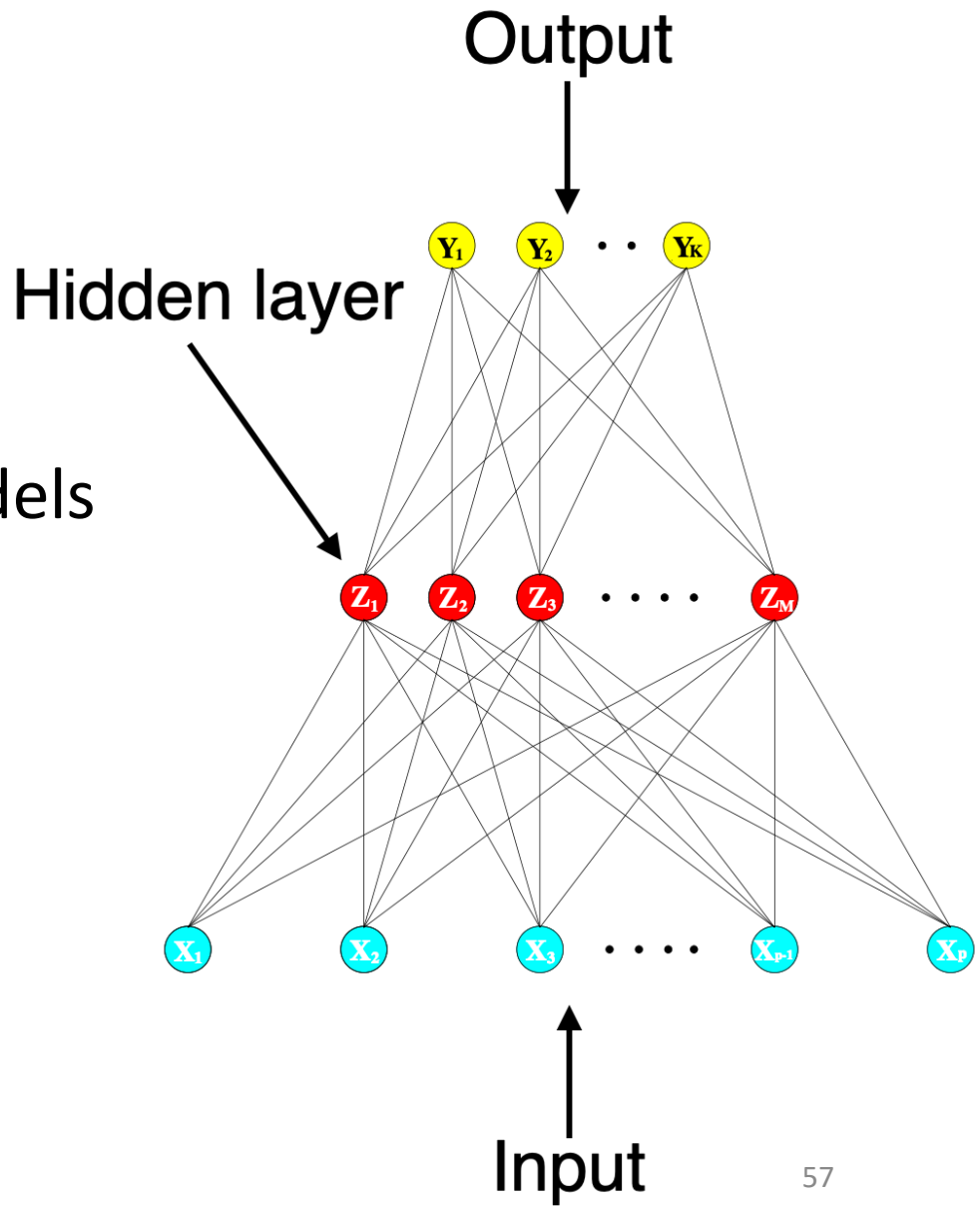
- Given a set of categories (e.g. α-helices, β-strands, turns, …, or DSSP categories), identify to which a new observation (e.g. residue) belongs
  - judging based on **features** of the new observation: measurable traits
- **Training set:** a set of observations with known features and categories (sequences with known 3D structures and assigned SSEs)
- **Test set:** new observations with only features known (protein sequences)

Regression problem (only for completeness)

- Estimate relationship between variables
  - given a set of **independent variables** (~features, inputs), estimate the value of the **dependent variable** (~output)
- **Training set:** a set of observations with known values of independent and dependent variables
- **Test set:** new observations with only values of independent variables known

Artificial neural networks

- Feed-forward
- Two-stage **regression** or **classification** models
- Input: numerical features
  - Feature vectors
- Output: class probabilities
  - For regression: only one output unit

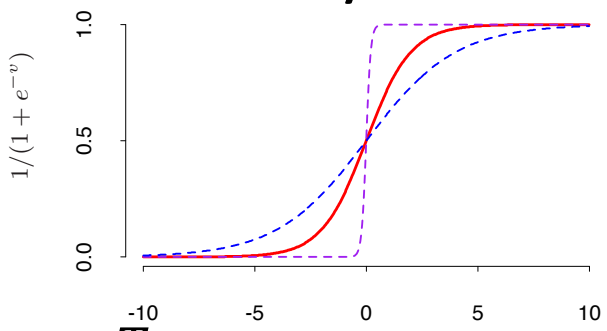


Artificial neural networks

- **Input features**: numerical vectors X
- Derived features (**hidden layer**): linear combination of the input features:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \dots, M$$

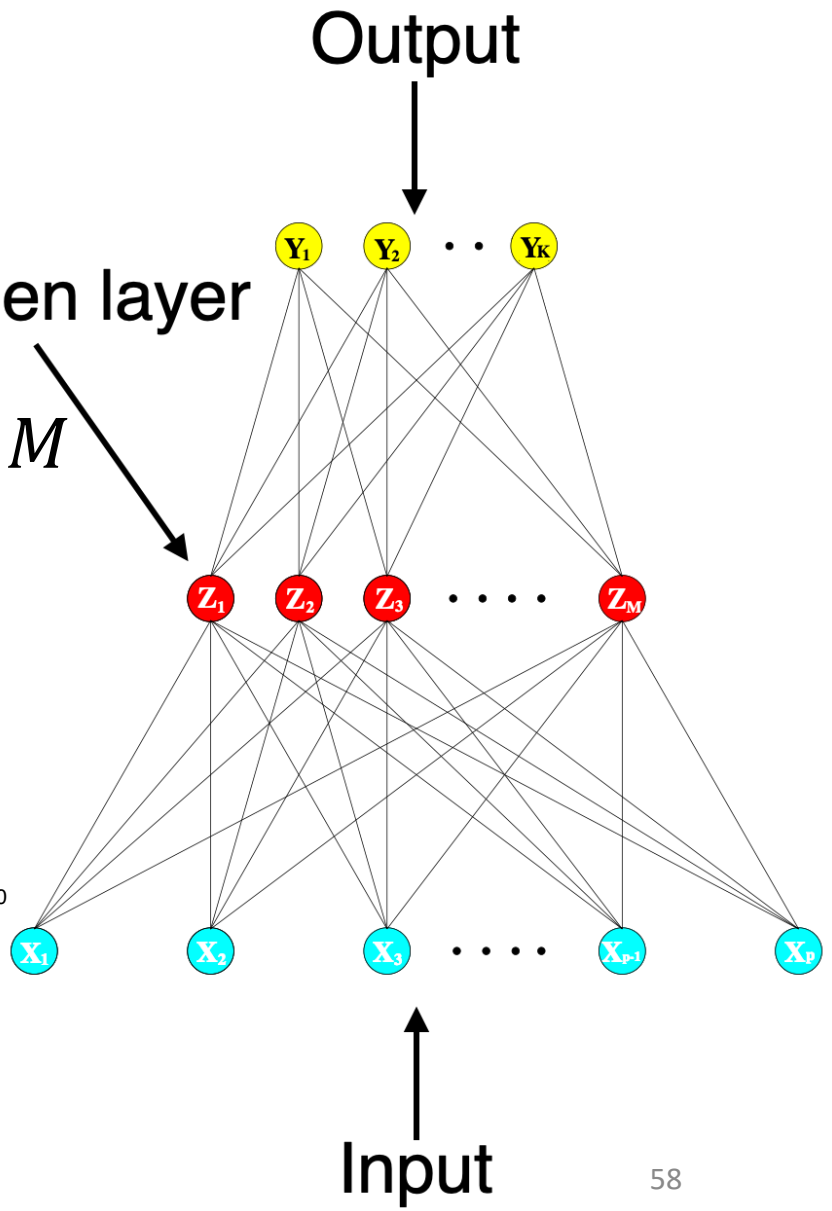
- σ: **activation functio**n, inserts non-linearity
  - Sigmoid: $\sigma(v) = \frac{1}{1+e^{-v}}$



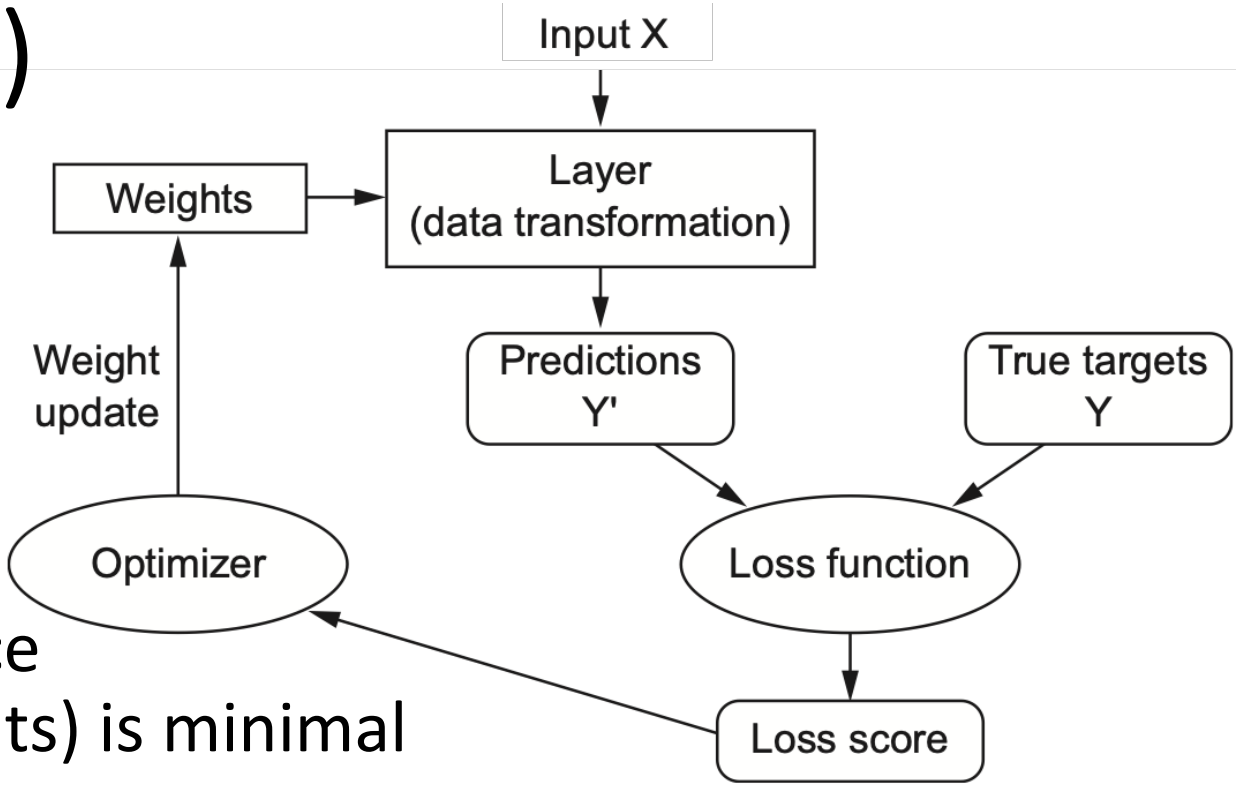
- **Target features** (K classes):

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, \dots, K$$

- Final normalization to turn T_k into class probabilities



Training a network (or any other machine-learning model, actually)

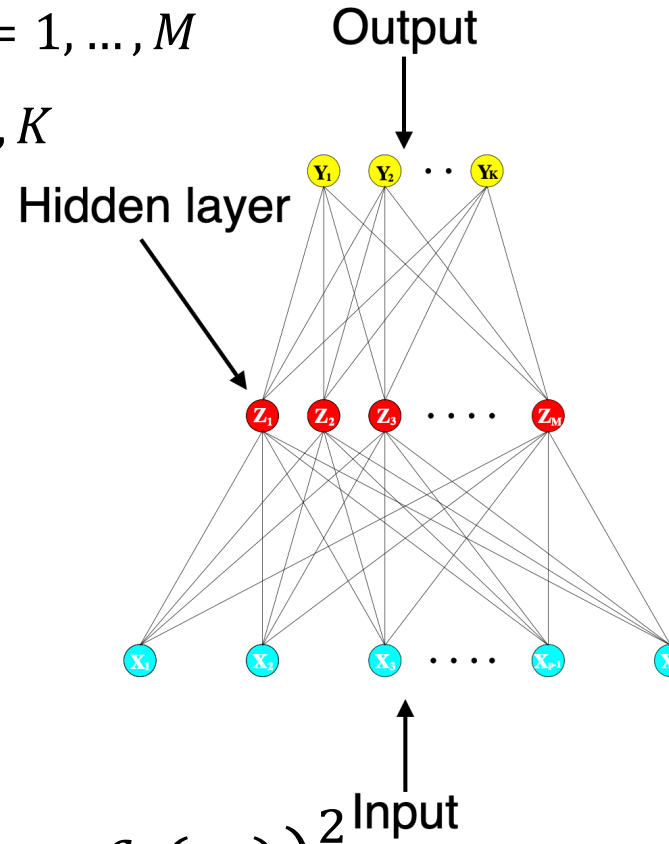


- N samples:
  - Input features: x_i
  - Known outputs: y_i
- **Training**: adjust parameters, such that an loss function (difference between true and computed outputs) is minimal
- ***K*-fold cross-validation**: split the data in *K* roughly equal parts, use *K* – 1 for training and 1 for testing
- **Leave-one-out cross-validation**: train on all samples but one

Fitting a network

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \dots, M$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, \dots, K$$



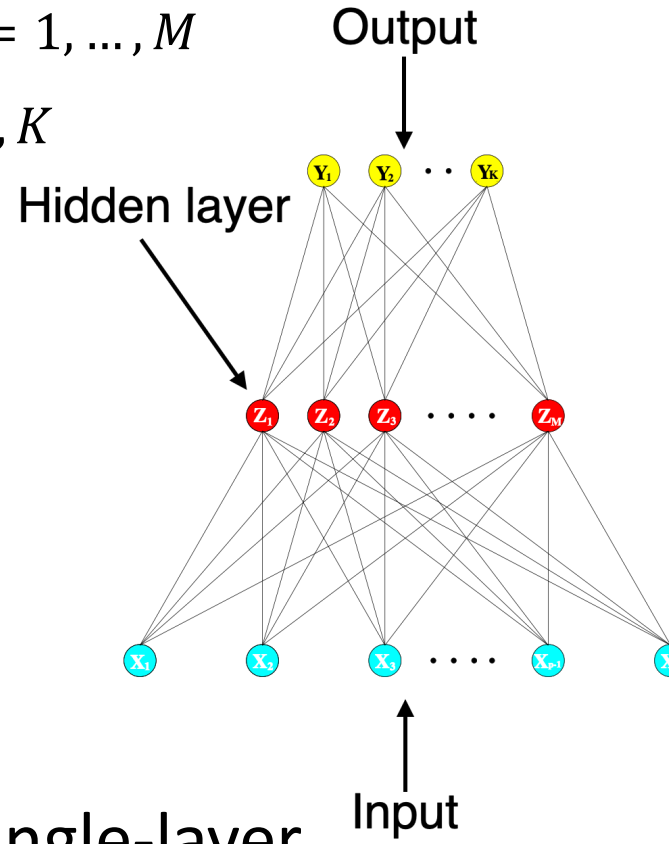
- Unknown parameters (**weights**):
  $\{\alpha_0, \alpha_m\}, m = 1, \dots, M$: $M(1 + p)$ weights,
  $\{\beta_0, \beta_k\}, k = 1, \dots, K$: $K(1 + M)$ weights,
- **Error function** (to be minimized):
  - For regression **sum-of-squares**: $R = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(y_{ik} - f_k(x_i)\right)^2$
  - For classification **cross-entropy**: $R = -\sum_{k=1}^{K} \sum_{i=1}^{N} y_{ik} \log f_k(x_i)$
- Start with random values close to 0 of weights, then minimize error by gradient descent

For details: See Hastie et al., “The Elements of Statistical Learning”, Springer, 2009

Detour: deep learning

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \dots, M$$

$$T_k = \beta_{0k} + \beta_k^T Z, k = 1, \dots, K$$



- >1 hidden layers
- Non-linear activation function => not equivalent to a single-layer network

Back to bioinformatics: Protein family profile

- Suppose we have collected a bunch of related proteins (e.g. with BLAST) and built an alignment (multiple sequence alignment methods will not be covered in this course)
  - Can we use it to find other related proteins?
  - Is this better (i.e. more sensitive) than BLASTing individually with the protein sequences?

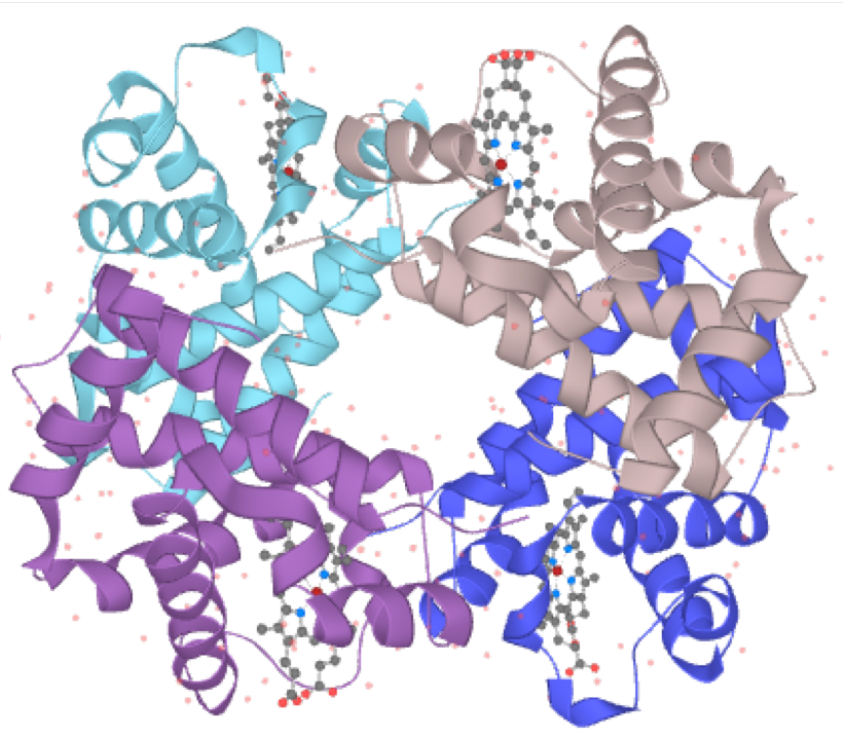
- YES — protein family **profile**

Conservation in MSA

```
Helix                 AAAAAAAAAAAAAAAA  BBBBBBBBBBBBBBBBB CCCCCCCCCCC
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP ----------LSADQISTVQASFDKVKG-----DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus             Ls....  v a W kv .  .    g . L.. f . P .   F  F

Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE          FFFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK---K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG---AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    ..  . v..Hg kv. a   a...l   d   . a l. l   H  .

Helix       FFGGGGGGGGGGGGGGGGGG    HHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR-------
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-------
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -QVDPQYFKVLAAVIADTVAAG----F----DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus   v.   f  l . ..  ....    f   . aa. k. .     l sky
```



Helices are better conserved than loops between them!

Protein family profile: idea

- Some protein regions are better conserved than other
  - Conserved sequence regions often correspond to conserved / structured 3D regions
- Some residues are particularly conserved
  - They often correspond to functionally important residues
- **Conserved regions and residues are more important**

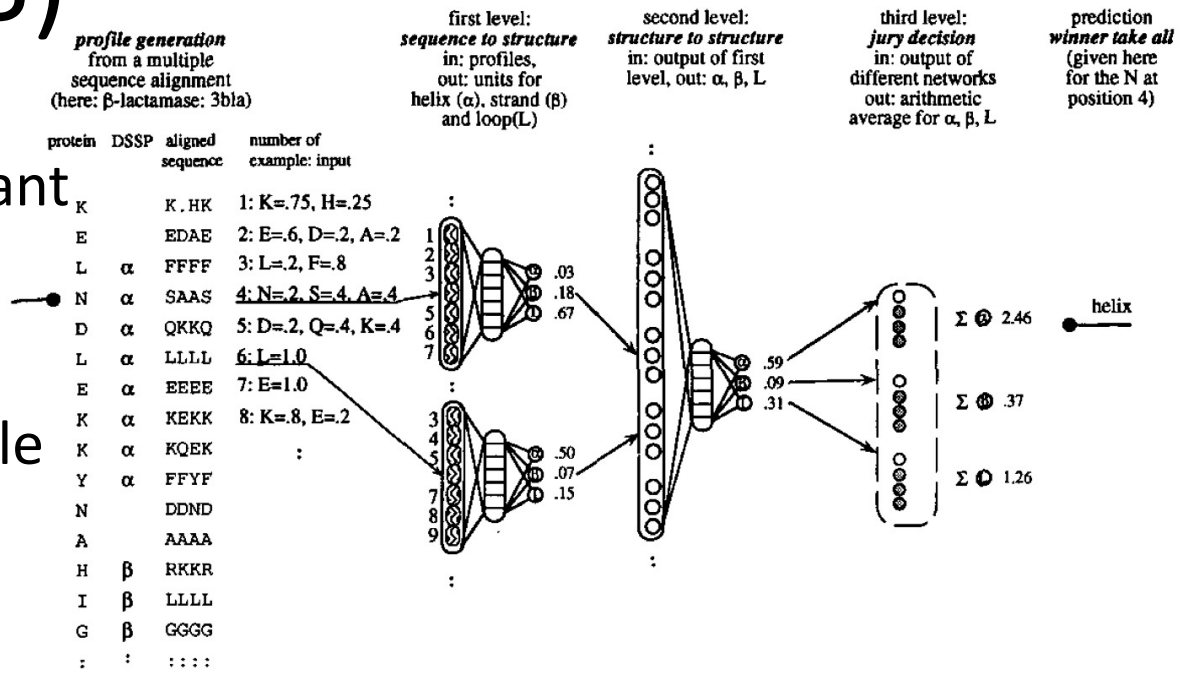
Calculating protein family profile

```
Helix                AAAAAAAAAAAAAAAA
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV--
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA
GLB3_CHITP ----------LSADQISTVQASFDKVKG
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAG
Consensus             Ls....  v a W kv .
```

- For each amino acid and each alignment column, count the number of occurrences of this amino acid in this column
- Divide by the number of sequences in the alignment
- **How to use it?**
  - For **sequence similarity search** (Lecture 6): can be naturally turned into an HMM
    - Calculate the probability of observing a certain amino acid at a certain position: emission probabilities $e(x_i) =$ observed frequencies (almost), normalized by background frequencies q_{x_i}
    - Probability of a sequence given a model: $\prod_{i=1,\dots,L} e_i(x_i)$
    - Transition probabilities: slightly more tricky
  - For **predicting secondary structure** (now)

PHD (Rost & Sander, 1993)

- Predicts 3 states: **helix**, **strand**, **loop**
- **Neural network** trained of **130** non-redundant proteins with known 3D structure
- 3 simple ANNs (1 hidden layer)
- **Input**: multiple sequence alignment -> profile
  - Amino acid frequencies
- 7-fold cross-validation
- Overlapping window moves down the sequence:
  - 1st level: independent prediction for central residue
  - 2nd level: combines output from overlapping windows
  - 3rd level: average over several networks
- Overall Q_3 accuracy: **70.8%**



Quality of prediction

- **Q_3 accuracy** = # residues correctly predicted / total # residues
  - Expected to be less than 100%, since SSE can only be identified with a certain uncertainty
  - Does not perfectly represent the quality of prediction

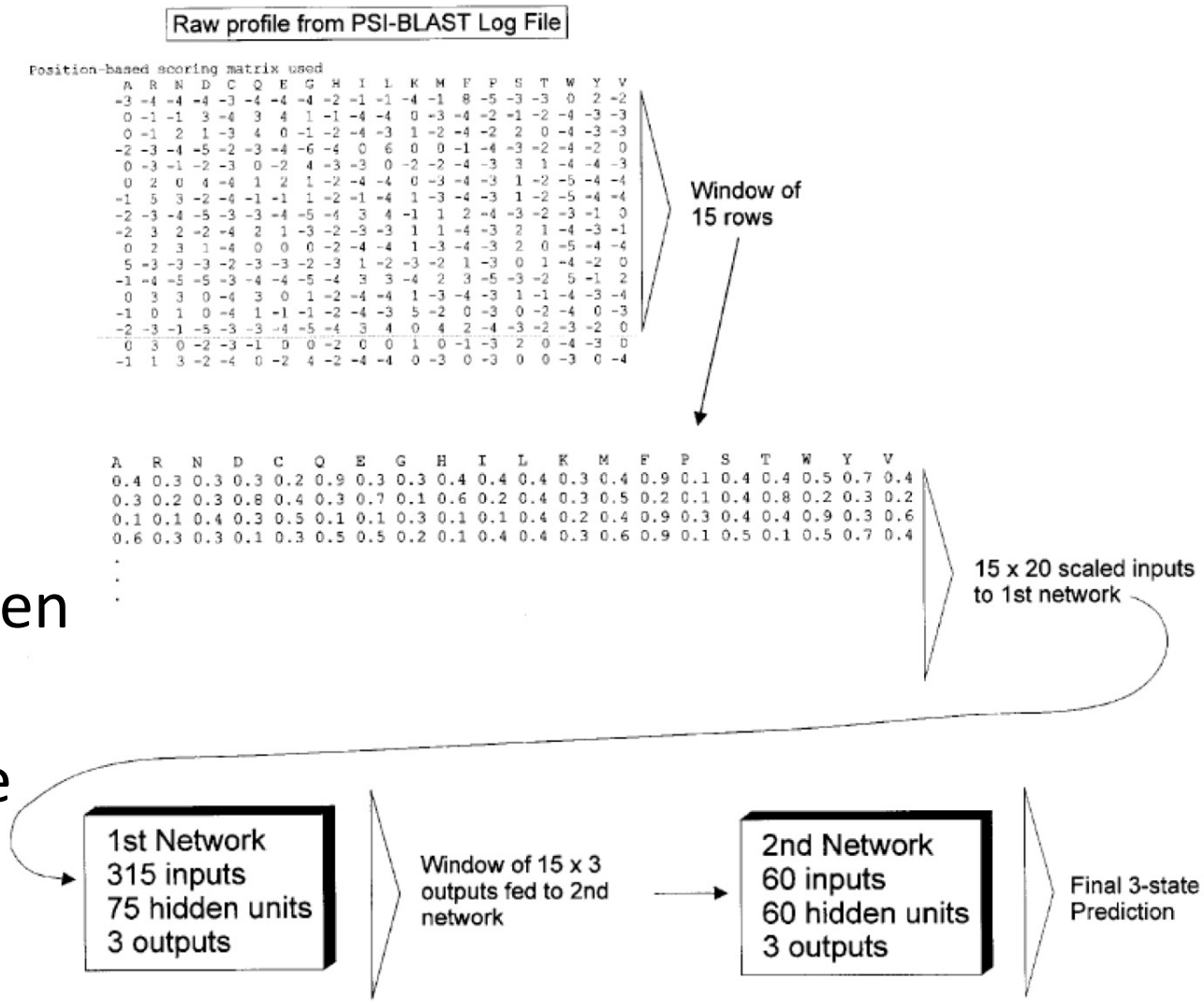
VLHQASGNSVILFGSDVTVPGATNAEQAR
HHHHHCCCCEEEECCCEEECCCCCHHHHH — actual SSEs
CHHHHCCCEEEECCCCCEEECCCHHHHHH — prediction 1, Q3 = 76%, useful
HHHHHCCCCHHHHCCCHHHCCCCCHHHHH — prediction 2, Q3 = 76%, bad

- **Sov** (segment overlap): $\frac{1}{N}\sum_s \frac{minov(s_1,s_2)+\delta}{maxov(s_1,s_2)} \cdot length(s_1)$

VLHQASGNSVILFGSDVTVPGATNAEQAR
HHHHHCCCCEEEECCCEEECCCCCHHHHH — actual SSEs
CHHHHCCCEEEECCCCCEEECCCHHHHHH — prediction
>|||< **— minov**
>|||||< **— maxov**

PSIPRED (Jones, 1999)

- PSI-BLAST profiles on a non-redundant set of sequences of globular proteins
- Neural network with a single hidden layer
- Second neural network to average over positions in the window
- 3-fold cross-validation
- Q_3 accuracy: **76.5%**; Sov_3: **73.5%**



Nowadays: deep learning

- Deep learning models can be trained to predict secondary structure
  - Usually the same 3 outputs: helix, strand, loop (coil)
- Not usually used on their own, but as a part of models to predict protein 3D structure *ab initio* (Lecture 11)

Summary and possible exam questions

- What is the residue pattern in the coiled-coil regions?
- What residues are preferred in transmembrane regions? Why?
- What is a Markov chain?
- What is a hidden Markov model?
- What are hidden and observable states, transmission and emission probabilities?
- How hidden Markov models are used to predict transmembrane regions in proteins?
- What amino acids are avoided in regular secondary structure elements, such as α-helices and β-sheets? Why?
- What is a protein family profile? How is it calculated?
- How is the quality of prediction of secondary structure measured?