Freie Universität Berlin


VL Algorithmische BioInformatik (19710)

WS2013/2014

Woche 8 - Mittwoch

Tim Conrad
AG Medical Bioinformatics
Institut für Mathematik & Informatik, Freie Universität Berlin

- Vorlesung nächsten Montag:
  - Vertretung durch Prof. Reinert.
  - Thema: „Repeat Resolution in Genomics Assembly“
- Praktikum
  - Programm muss nach dem SVN Check-out direkt auf einem Fachbereichsrechner lauffähig sein
  - Das gilt auch für Ruby/Python Projekte
  - Im Ausnahmefall (!) kann mit der Tutorin abgesprochen werden, das manuelle Schritte (z.B. Download von Bibliotheken) notwendig sind
  - Bericht muss vollständig und „vernünftig“ formatiert sein (insbesondere bzgl. Umlaute, Tabellen etc.)

Vorlesungsthemen

Part 1: Background Basics (4)

1. The Nucleic Acid World
2. Protein Structure
3. Dealing with Databases

Part 2: Sequence Alignments (3)

4. Producing and Analyzing Sequence Alignments
5. Pairwise Sequence Alignment and Database Searching
6. Patterns, Profiles, and Multiple Alignments

Part 3: Evolutionary Processes (3)

7. Recovering Evolutionary History
8. Building Phylogenetic Trees

Part 4: Genome Characteristics (4)

9. Revealing Genome Features
10. Gene Detection and Genome Annotation

Part 5: Secondary Structures (4)

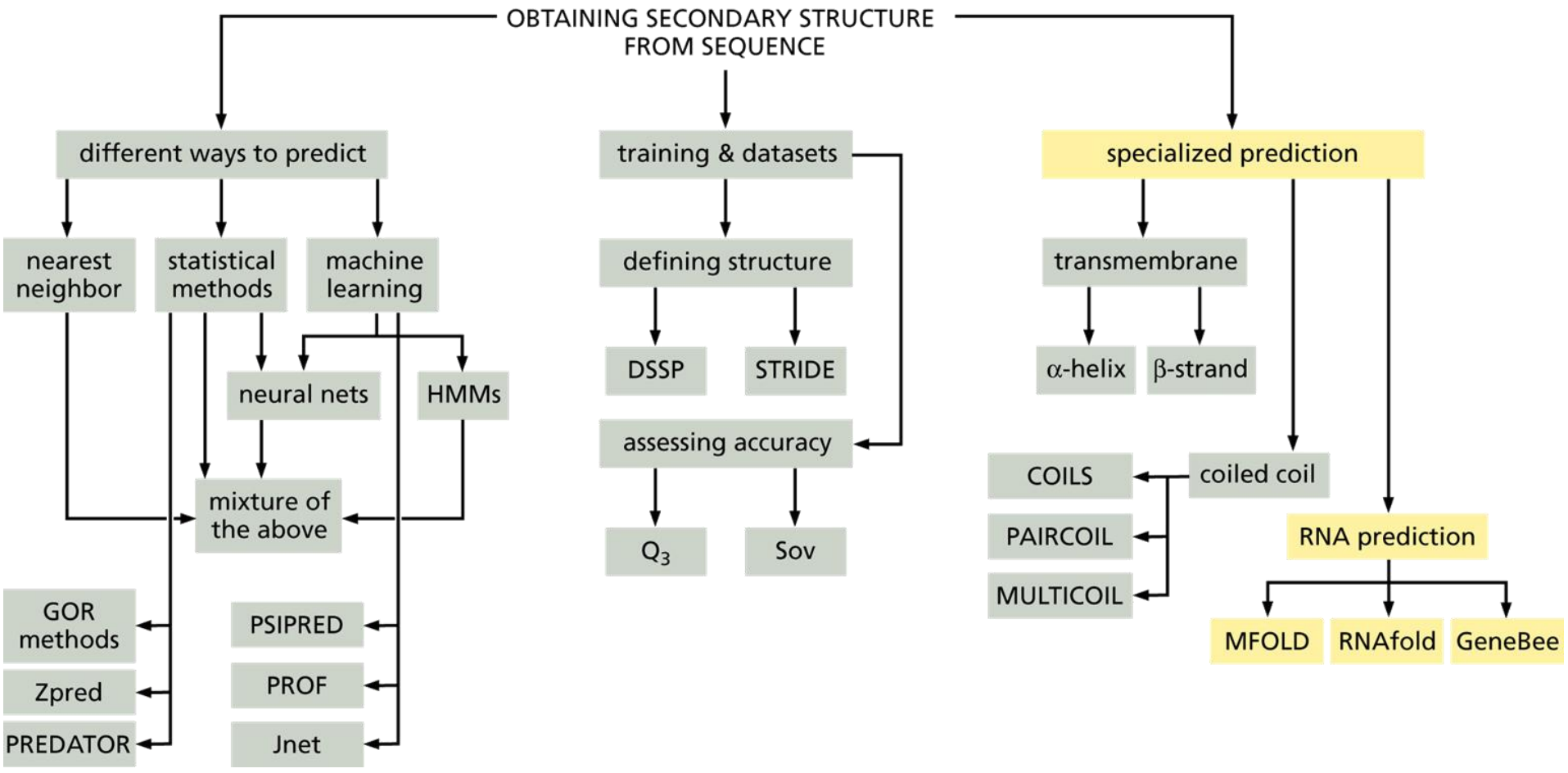
11. Obtaining Secondary Structure from Sequence
12. Predicting Secondary Structures

Part 6: Tertiary Structures (4)

13. Modeling Protein Structure
14. Analyzing Structure-Function Relationships

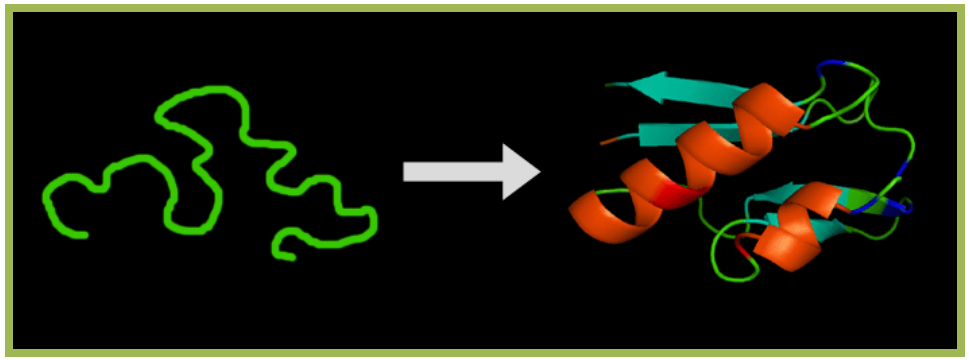
Part 7: Cells and Organisms (8)

15. Proteome and Gene Expression Analysis
16. Clustering Methods and Statistics
17. Systems Biology

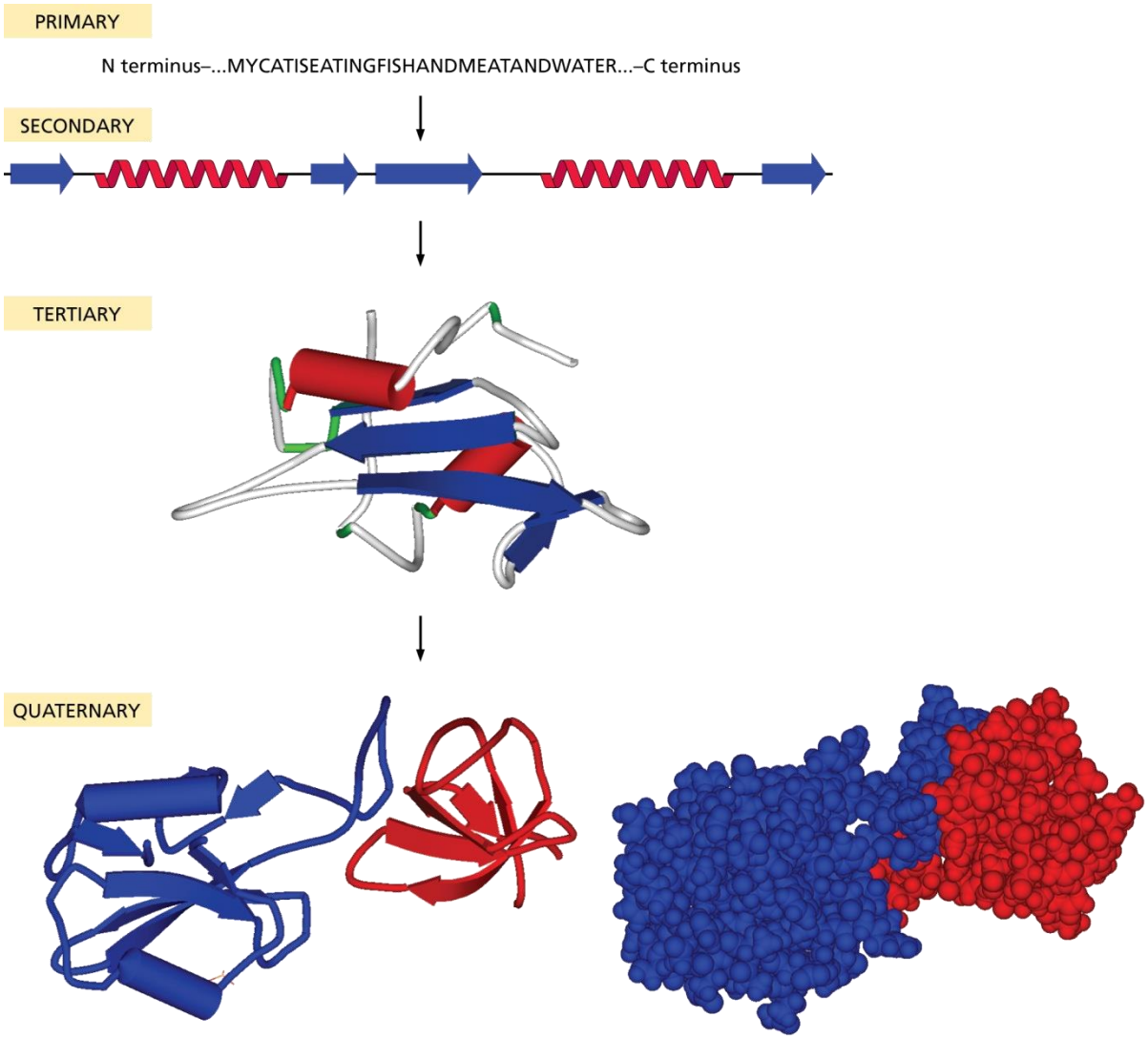


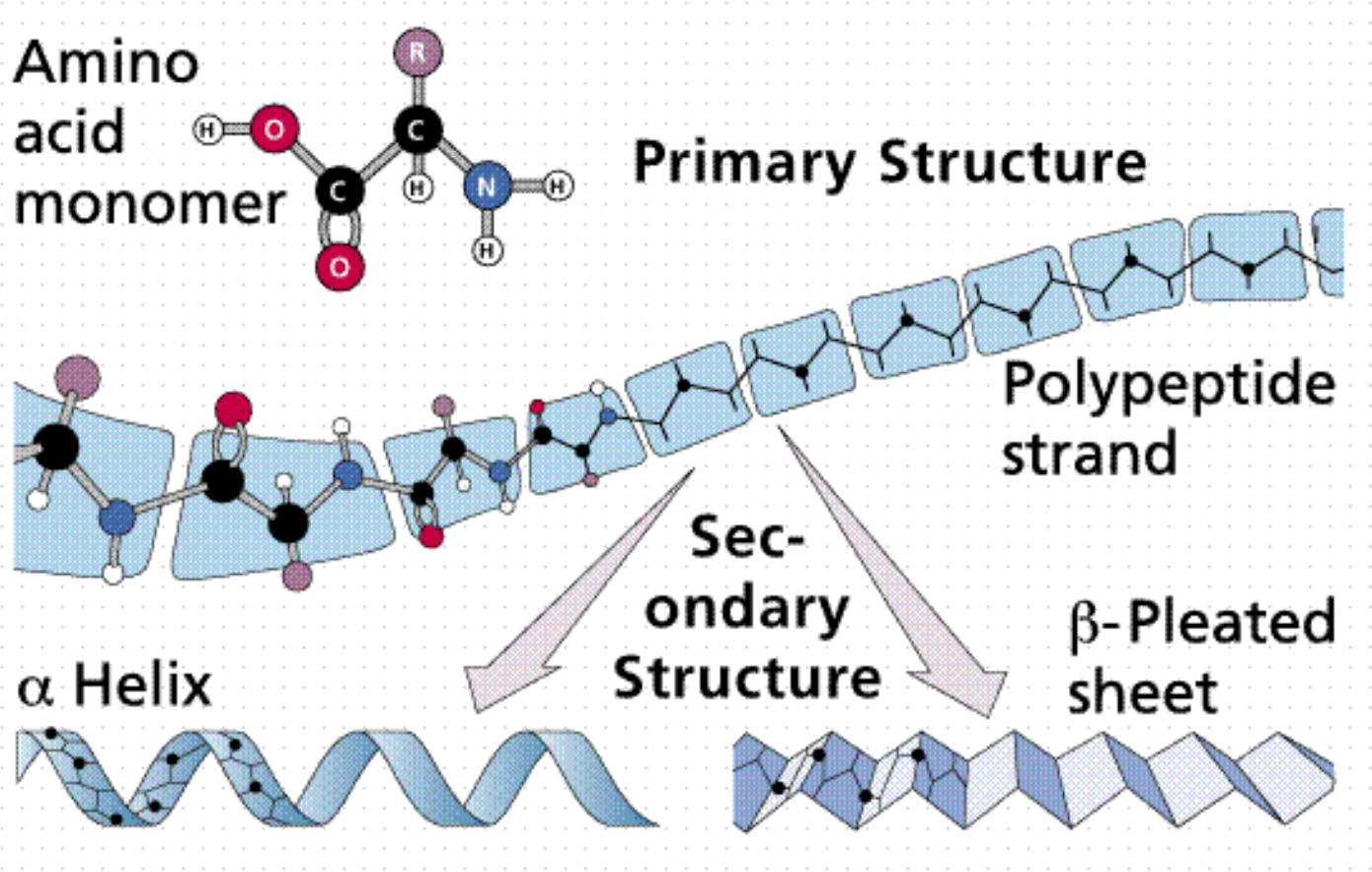
Today: 11.3, 12.2, 12.3

- How do proteins do so much?
  - Proteins FOLD spontaneously
  - Assume a characteristic 3D SHAPE
  - Shape depends on particular Amino Acid Sequence
  - Shape gives SPECIFIC function



Proteins are linear polymers that fold up by themselves…mostly.





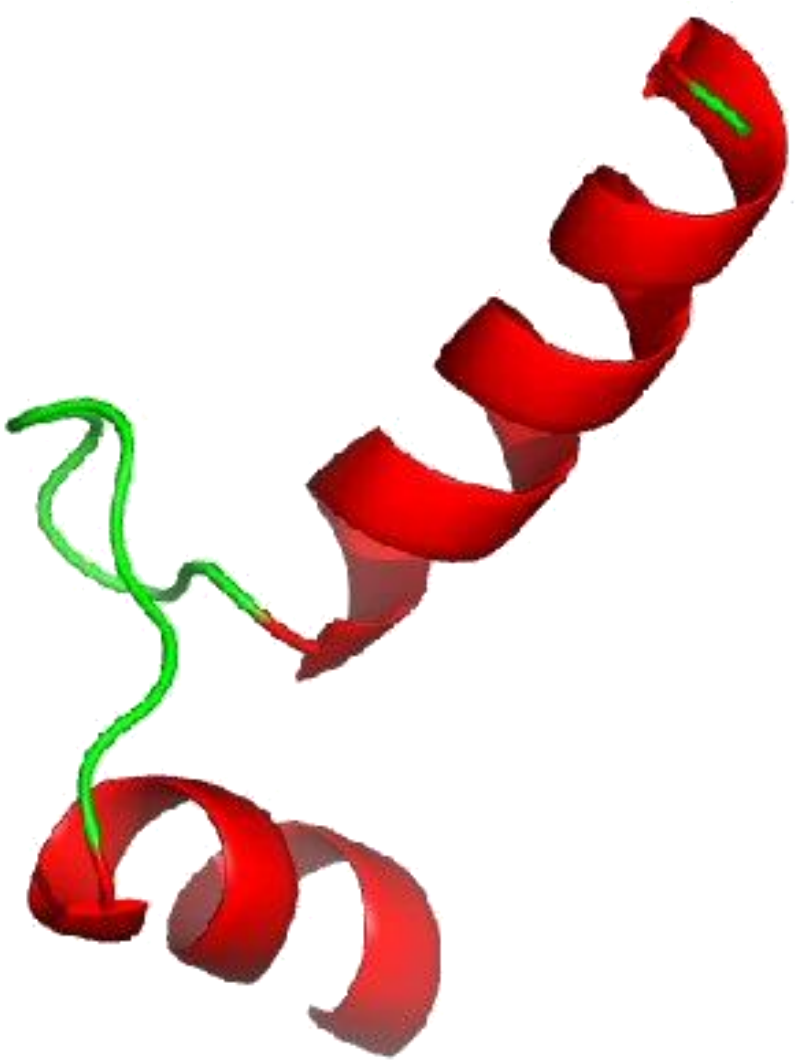
Secondary structure = spatial arrangement of amino-acid residues that are adjacent in the primary structure

- One of the first fields to emerge in bioinformatics (~1967)
- Grew from a simple observation that certain amino acids or combinations of amino acids seemed to prefer to be in certain secondary structures
- Subject of hundreds of papers and dozens of books, many methods…

2° Structure Prediction

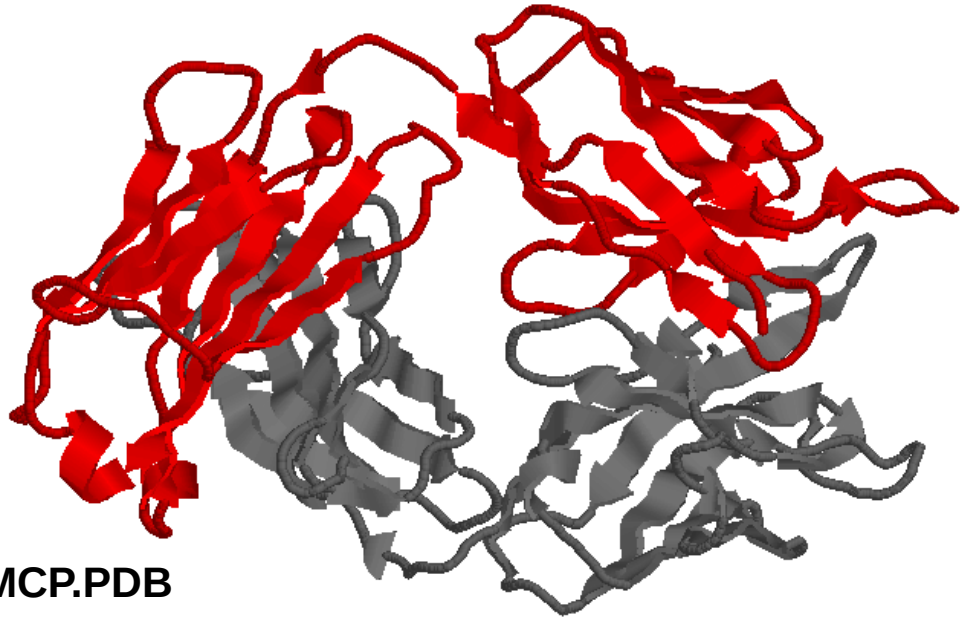
- Statistical (Chou-Fasman, GOR)
- Homology or Nearest Neighbor (Levin)
- Physico-Chemical (Lim, Eisenberg)
- Pattern Matching (Cohen, Rooman)
- Neural Nets (Qian & Sejnowski, Karplus)
- Evolutionary Methods (Barton, Niemann)
- Combined Approaches (Rost, Levin, Argos)

Helix-loop-helix



Large proteins are composed of compact, semi-independent units - domains.

Reason:
Modularity
Folding efficiency

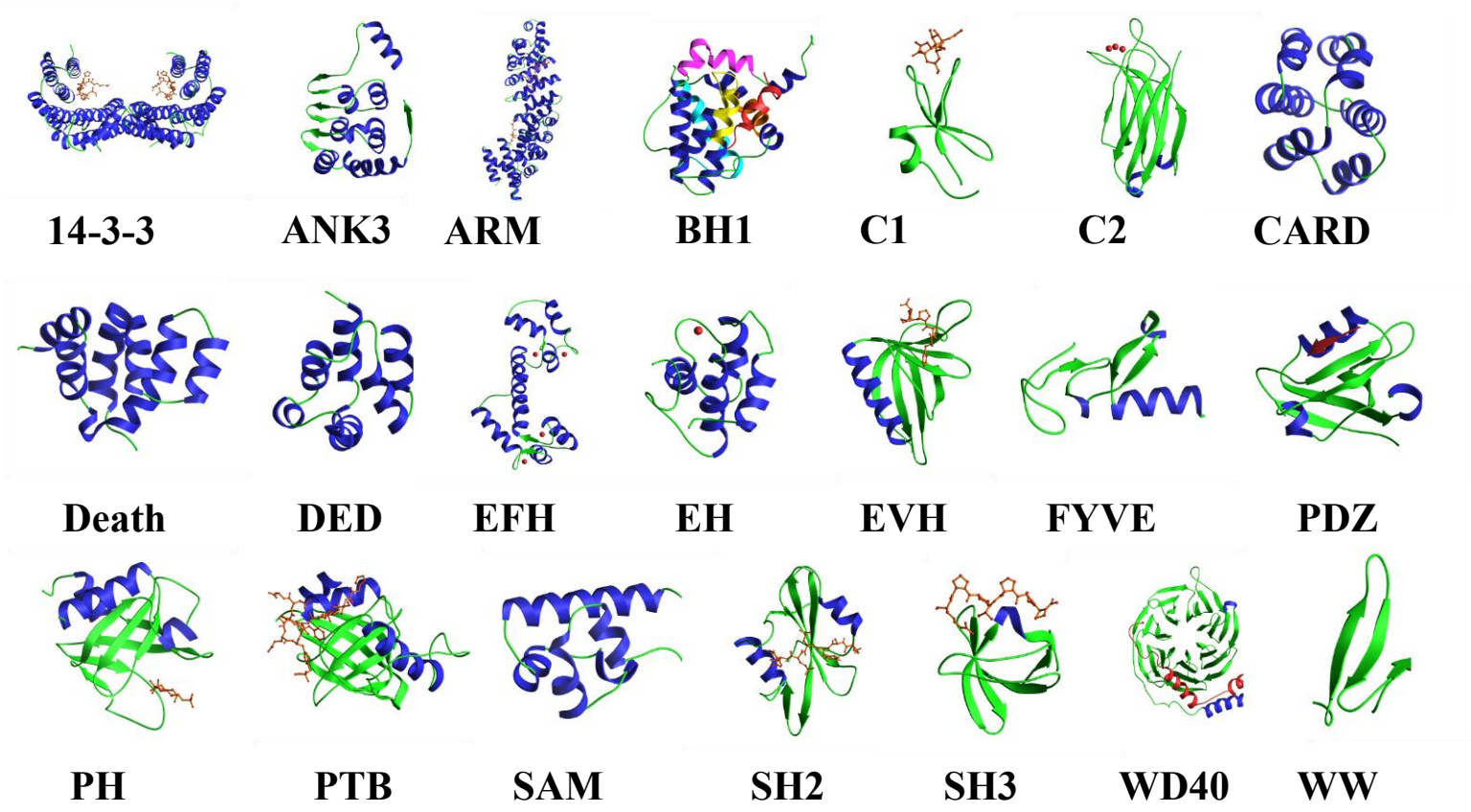


2MCP.PDB

From Domains to Structure

What's the goal again?

Protein Domains – an alphabet of functional modules

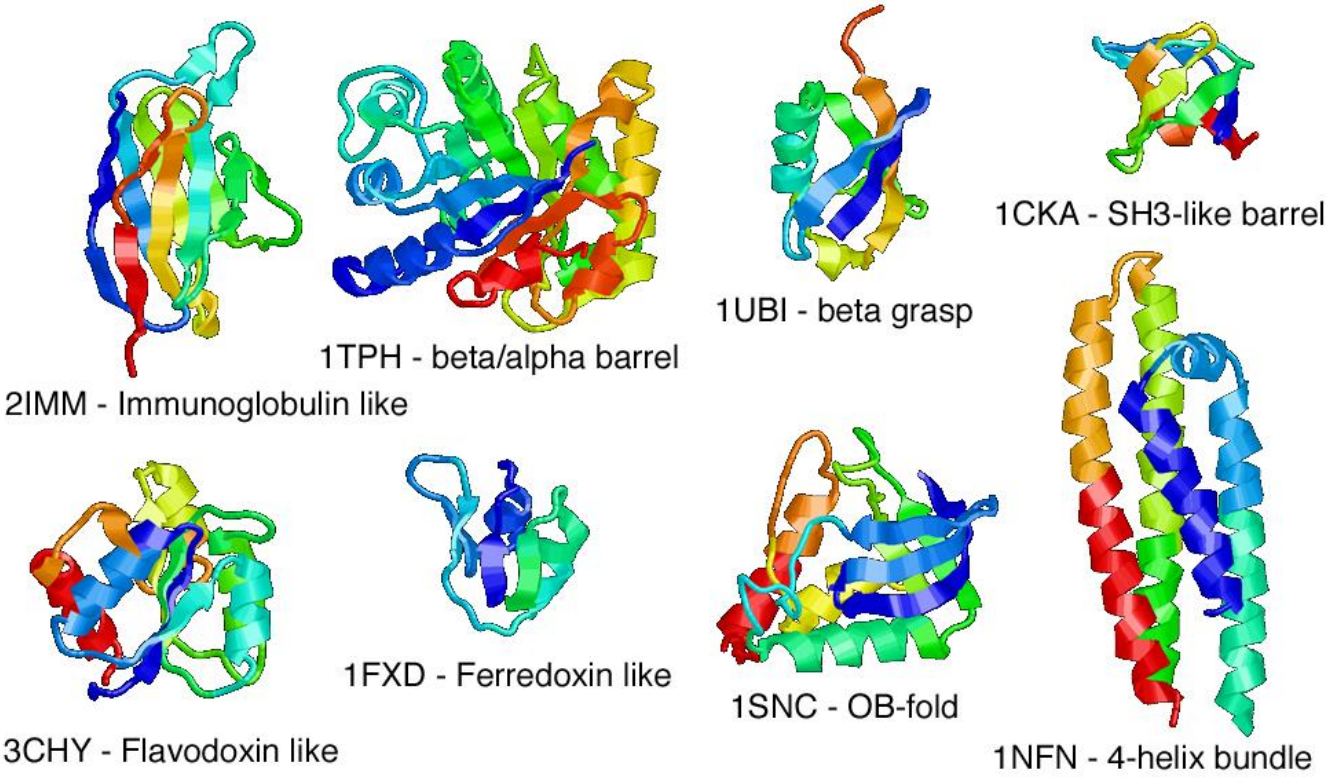


SCOP

- The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- Created by manual inspection and aided by automated methods
- Consists of four hierarchical categories:
  - Class, Fold, Superfamily and Family.

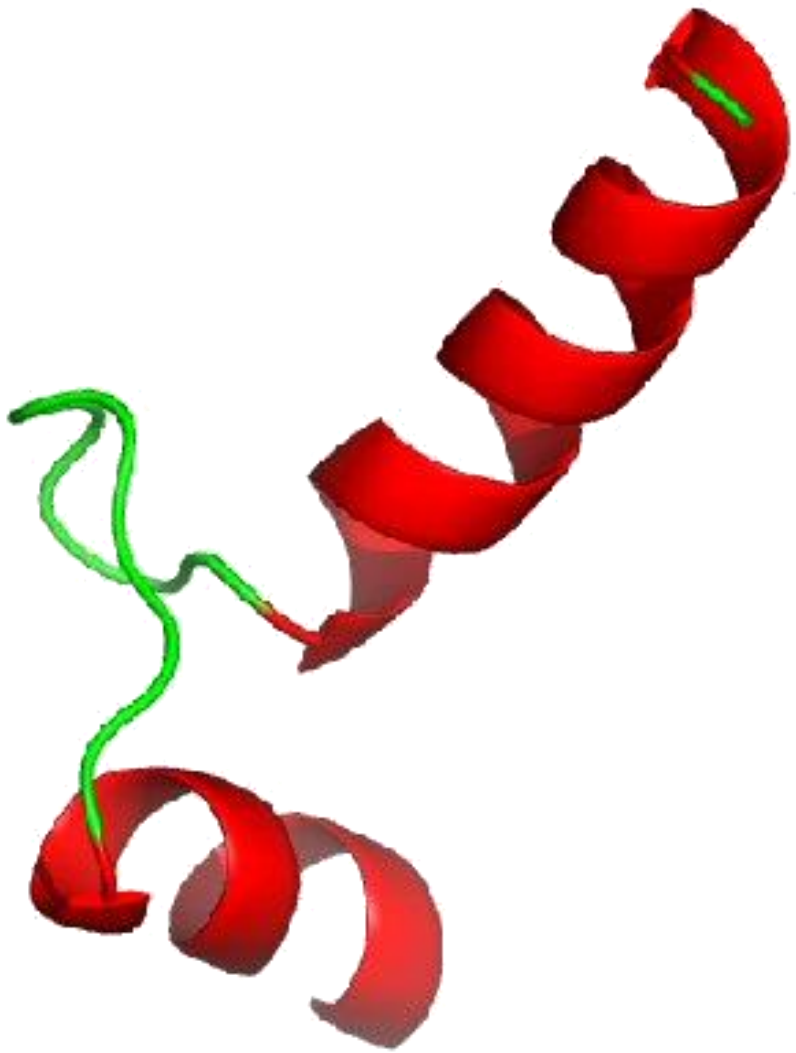
- http://scop.mrc-lmb.cam.ac.uk/scop

Structural **classification**

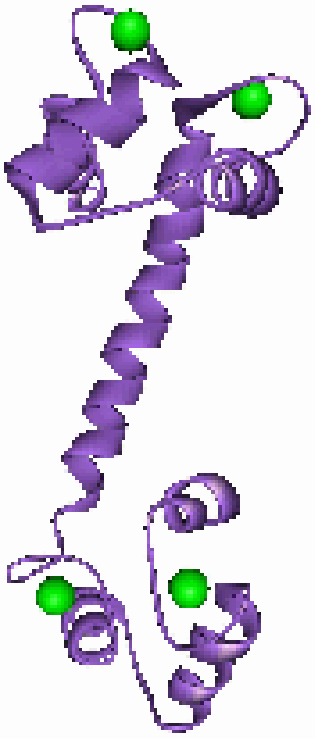


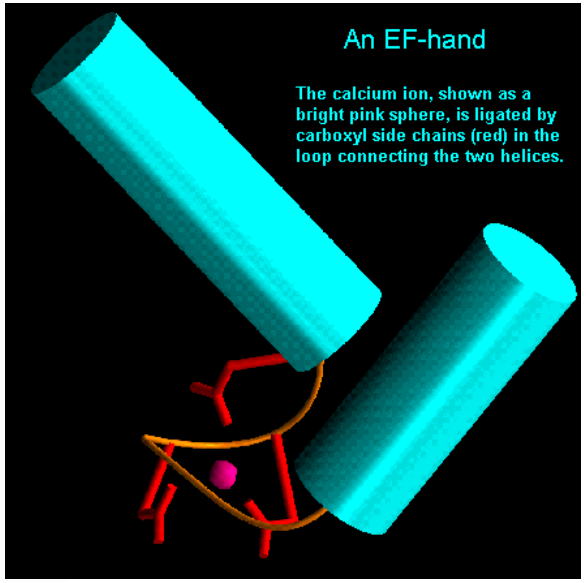
The eight most frequent SCOP superfolds

Helix-loop-helix



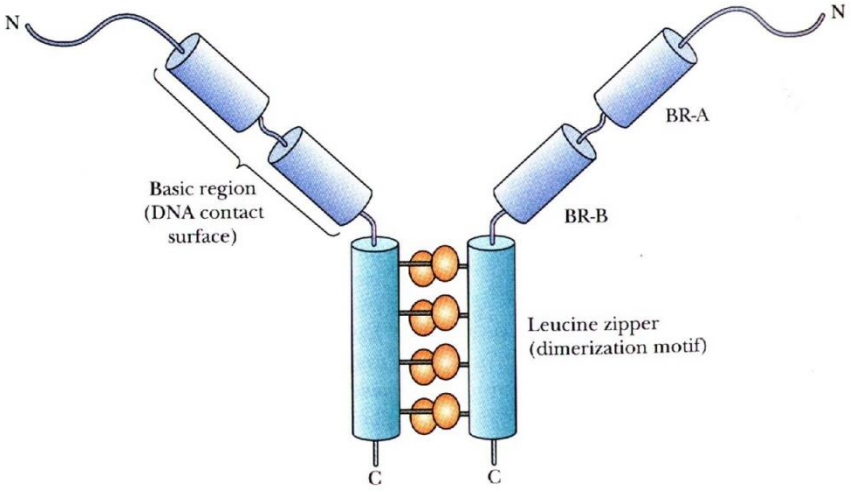
EF Fold

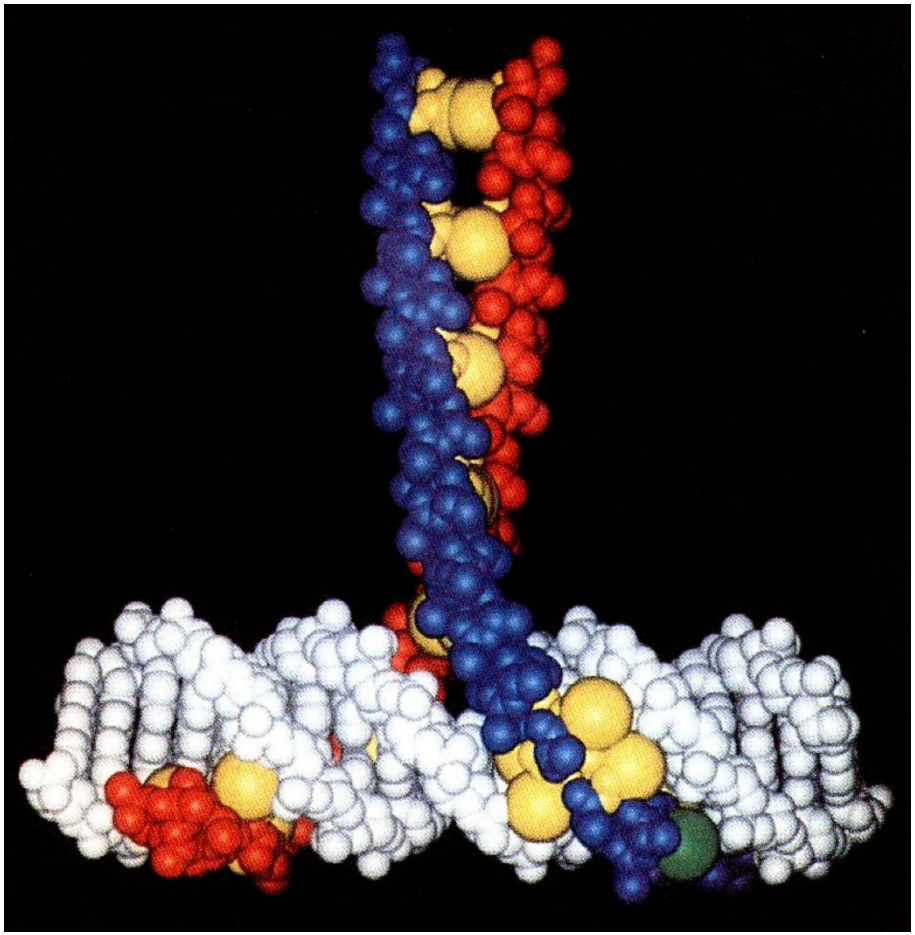




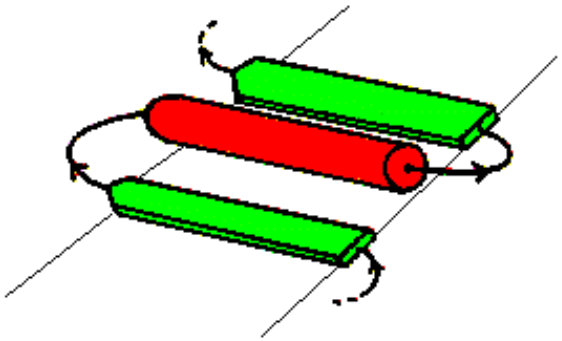
Found in Calcium binding proteins such as Calmodulin

Leucine Zipper





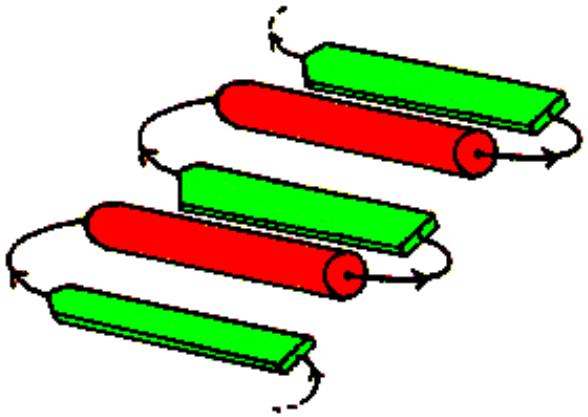
Rossman Fold

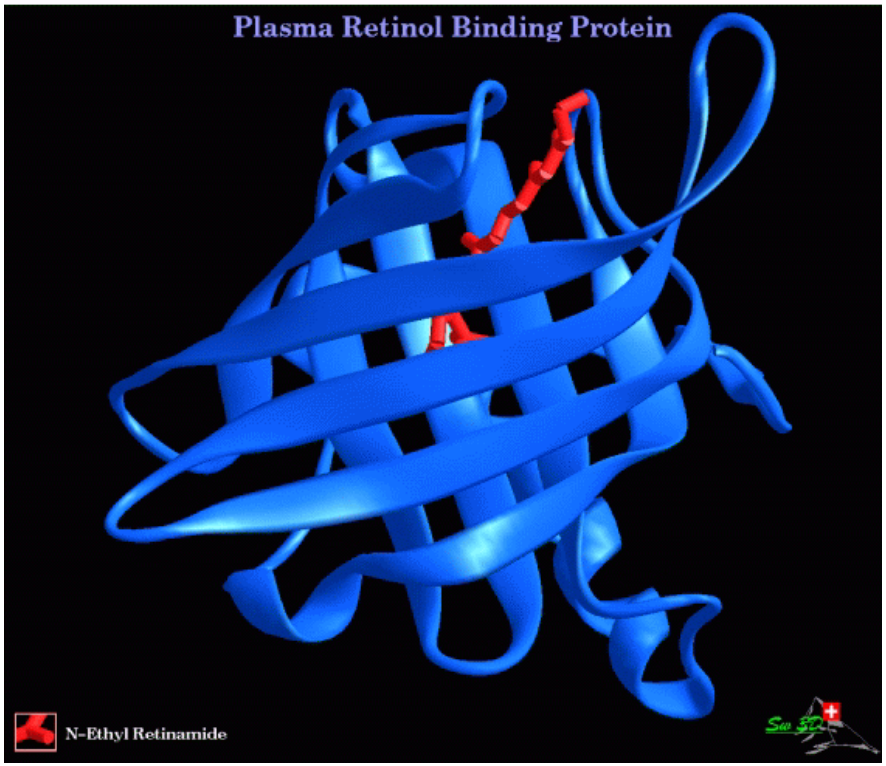


The right–handed beta–alpha–beta unit. The helix lies above the plane of the strands.

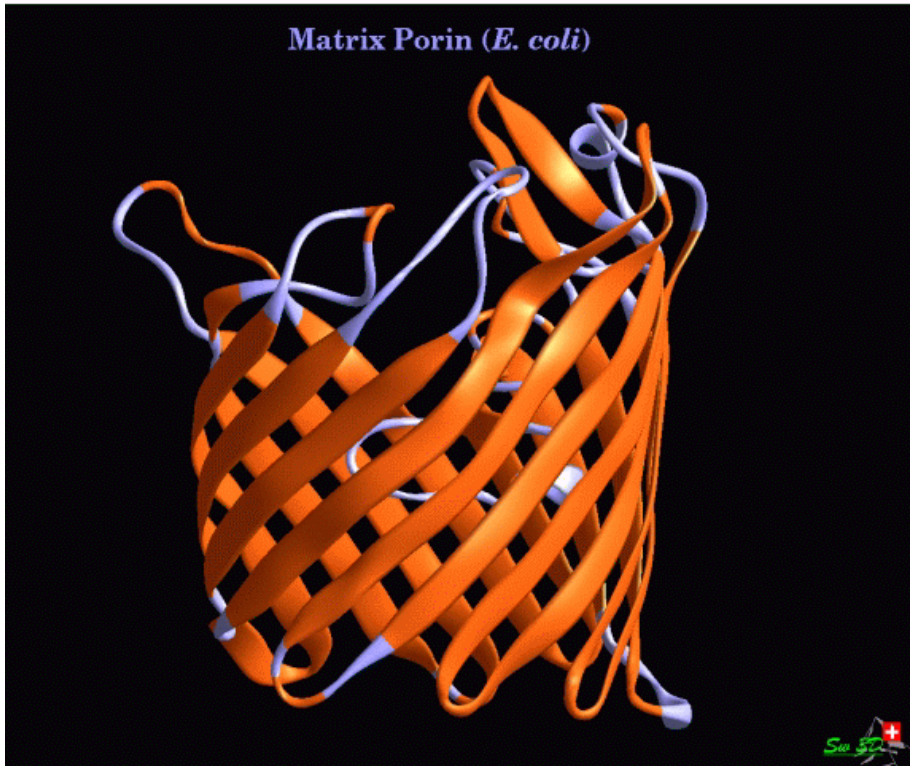
- The beta-alpha-beta-alpha-beta subunit
- Often present in nucleotide-binding proteins

The Rossman fold



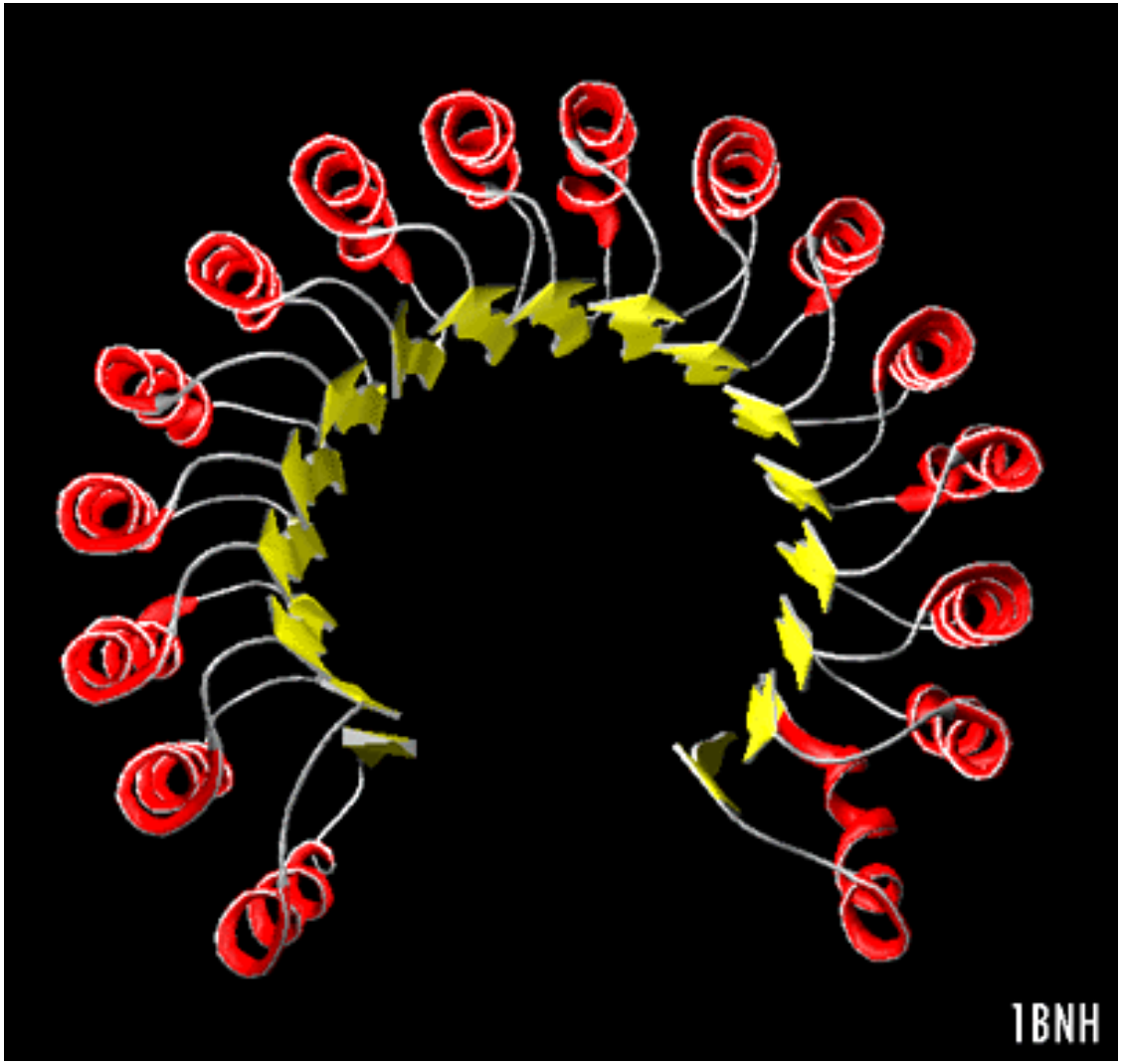


β sandwich

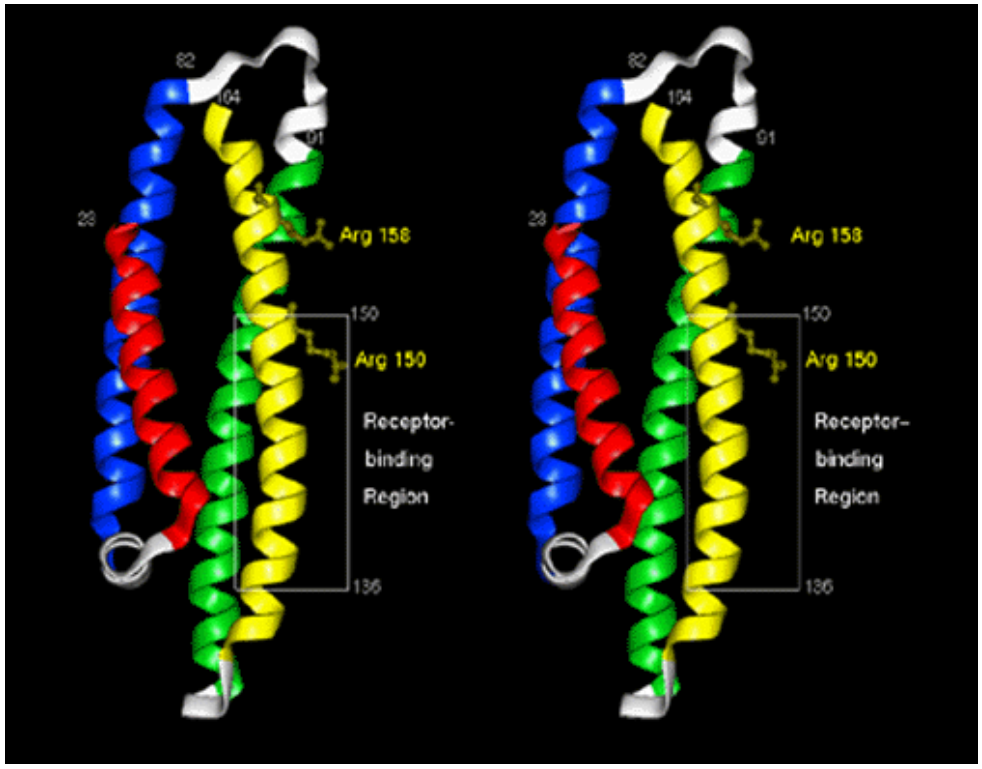


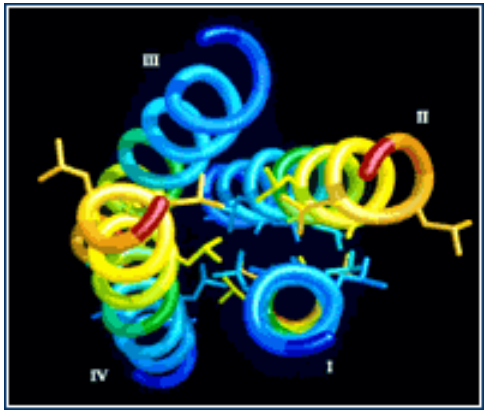
β barrel

α/β horseshoe



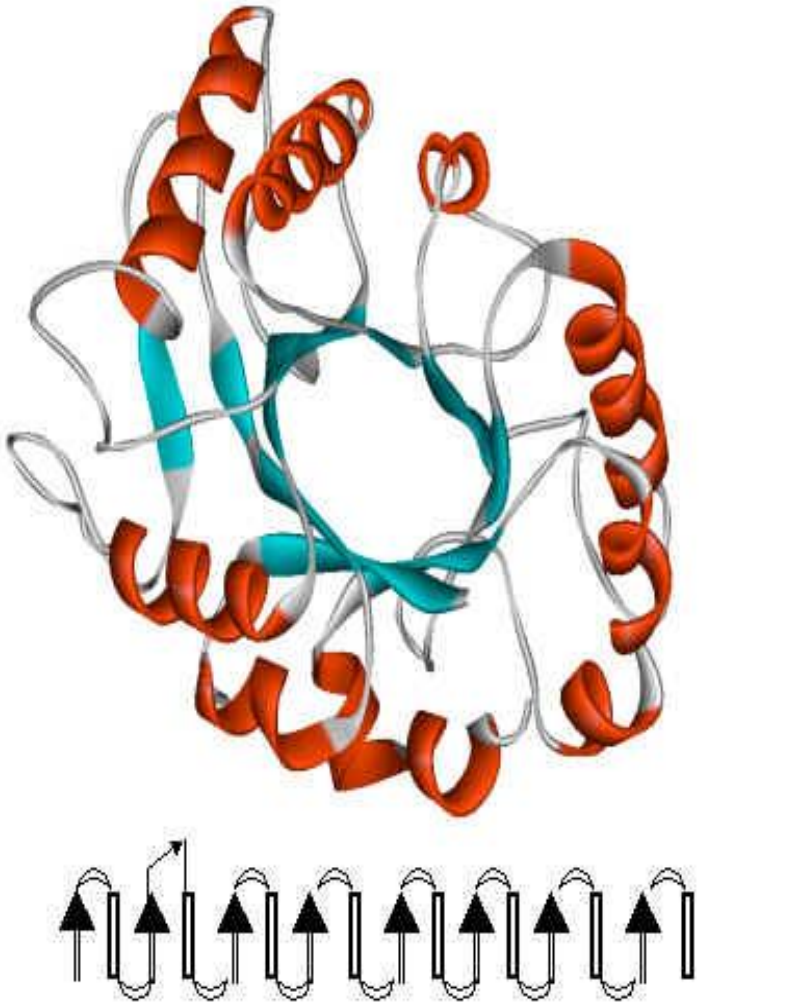
Four helix bundle





- 24 amino acid peptide with a hydrophobic surface
- Assembles into 4 helix bundle through hydrophobic regions
- Maintains solubility of membrane proteins

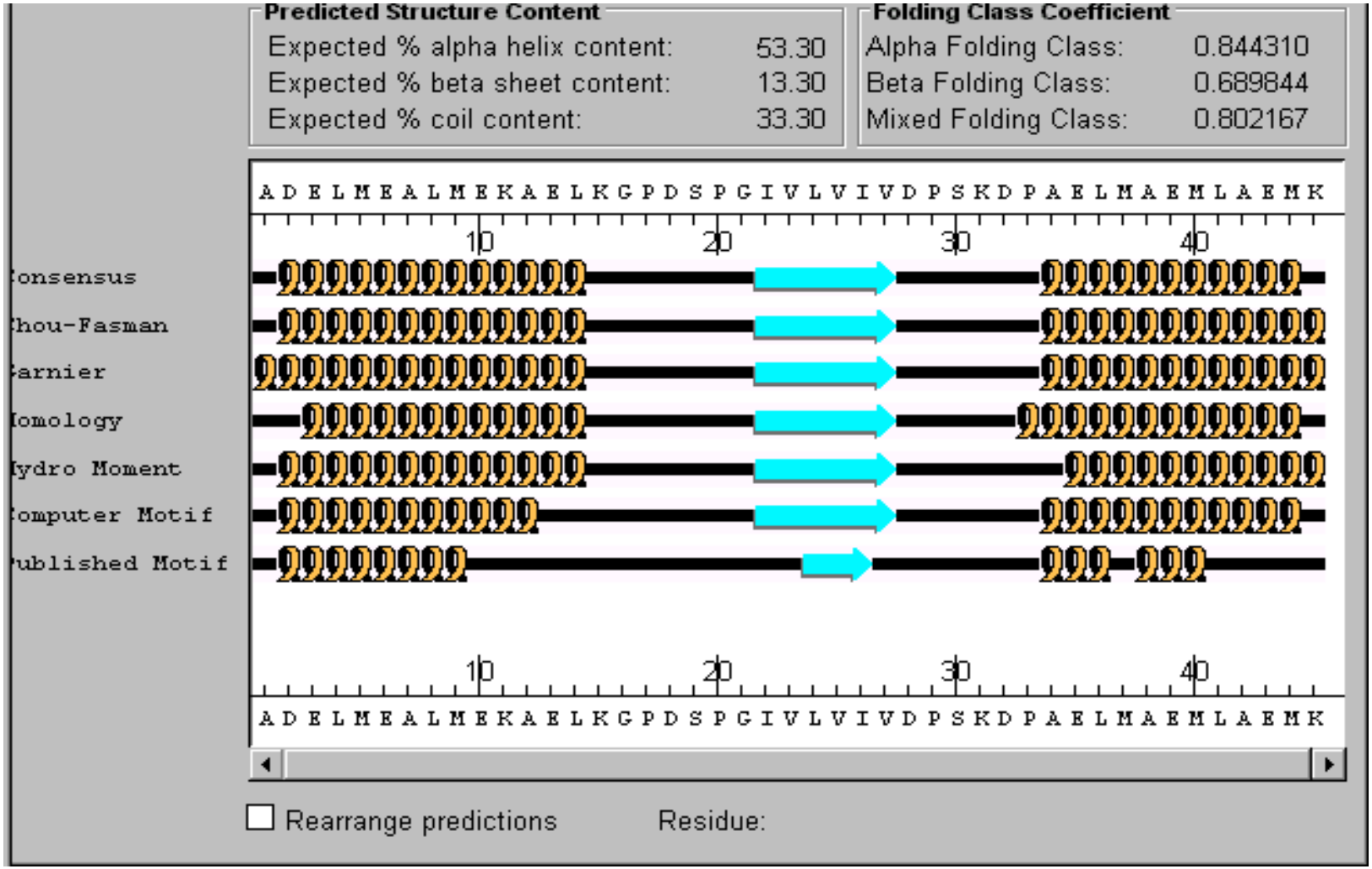
TIM Barrel



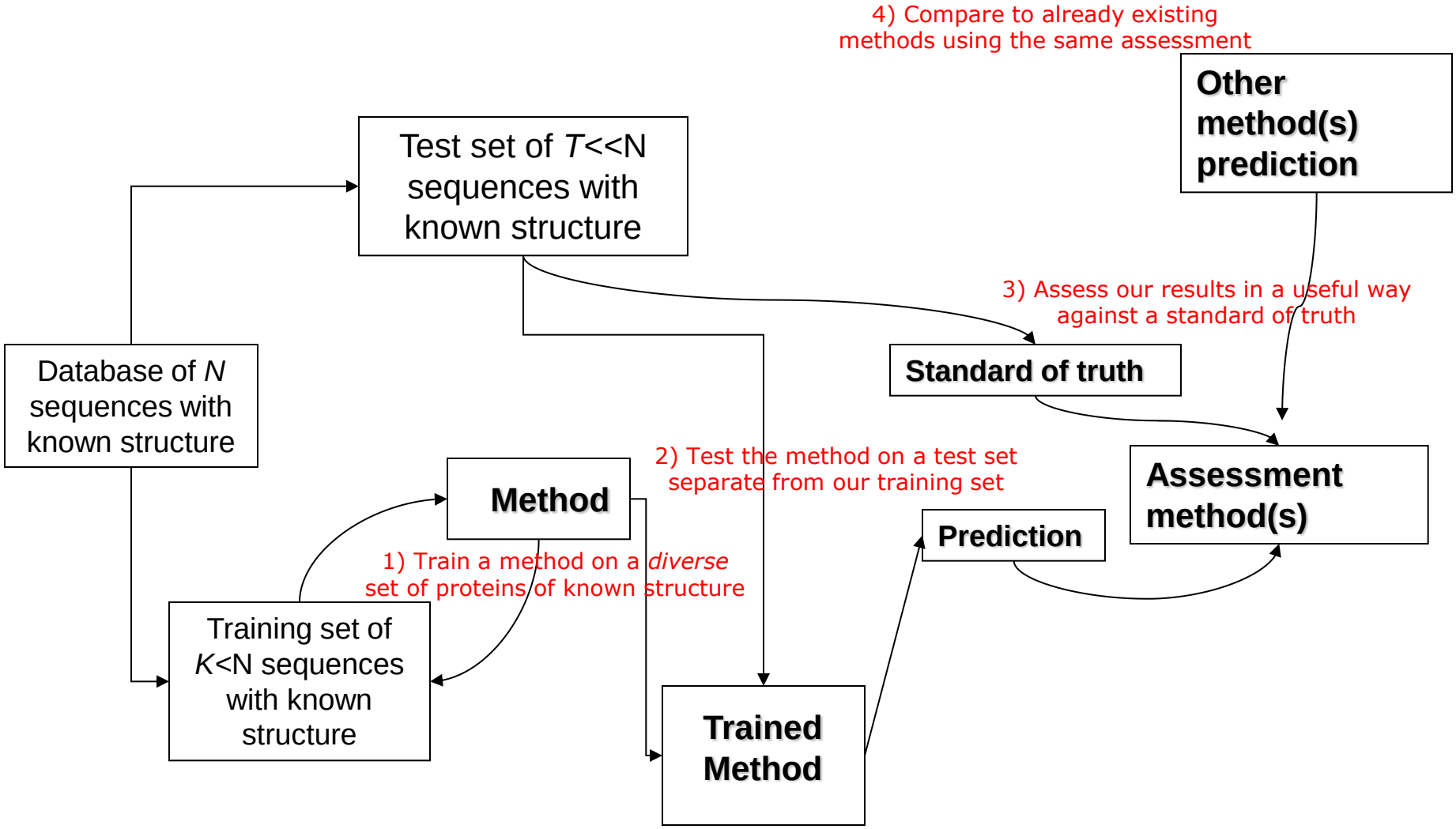
Topology diagram of Hevamine - one of the TIM barrel structures

How do we start?

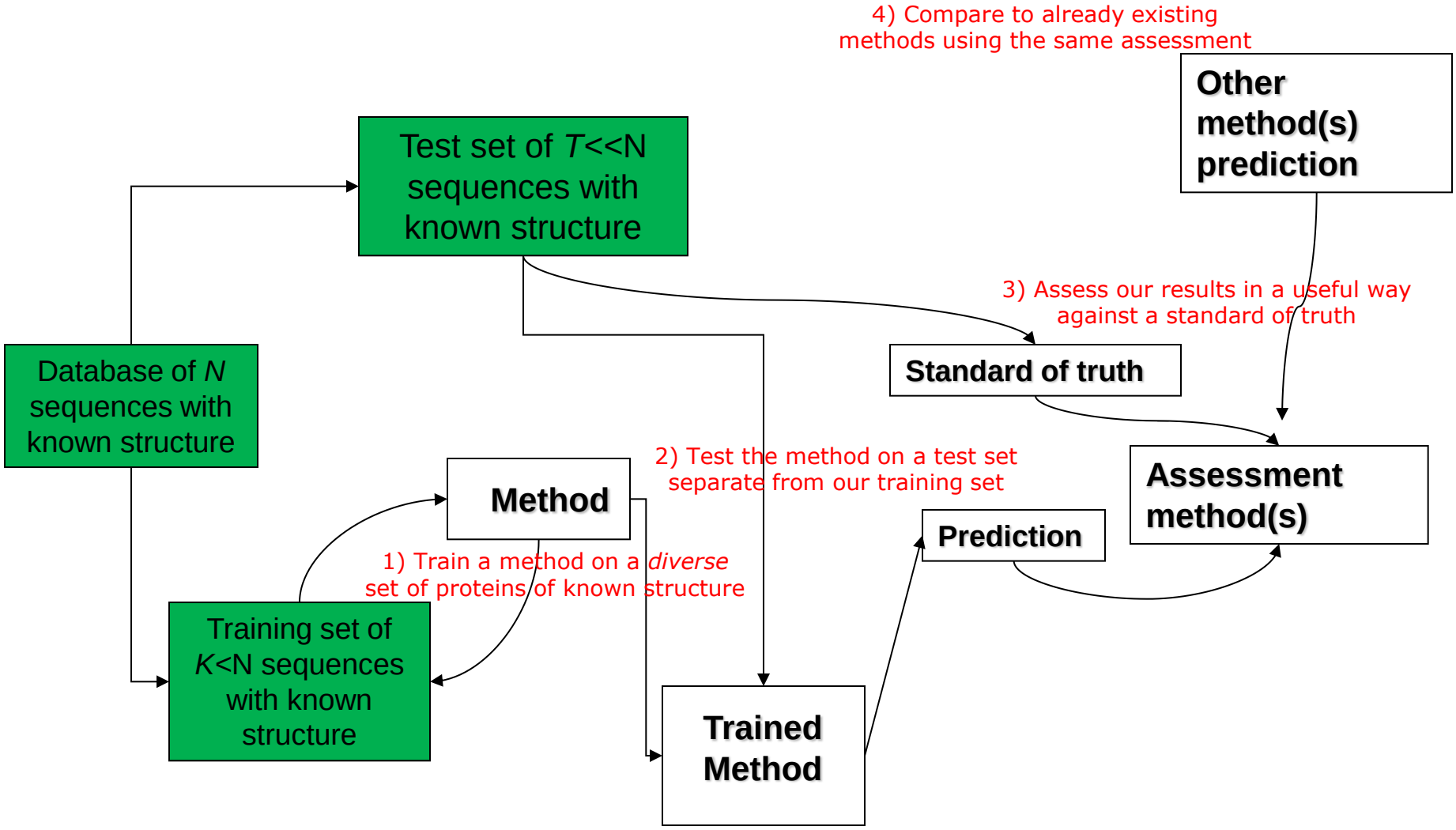
Secondary Structure Prediction



How to develop a method



How to develop a method



First, get some examples to study…

- We need some examples of proteins with known secondary structure to try and formulate a prediction approach…

This what we want lots of…

- Primary sequences labeled with the secondary structure of the residue's environment.

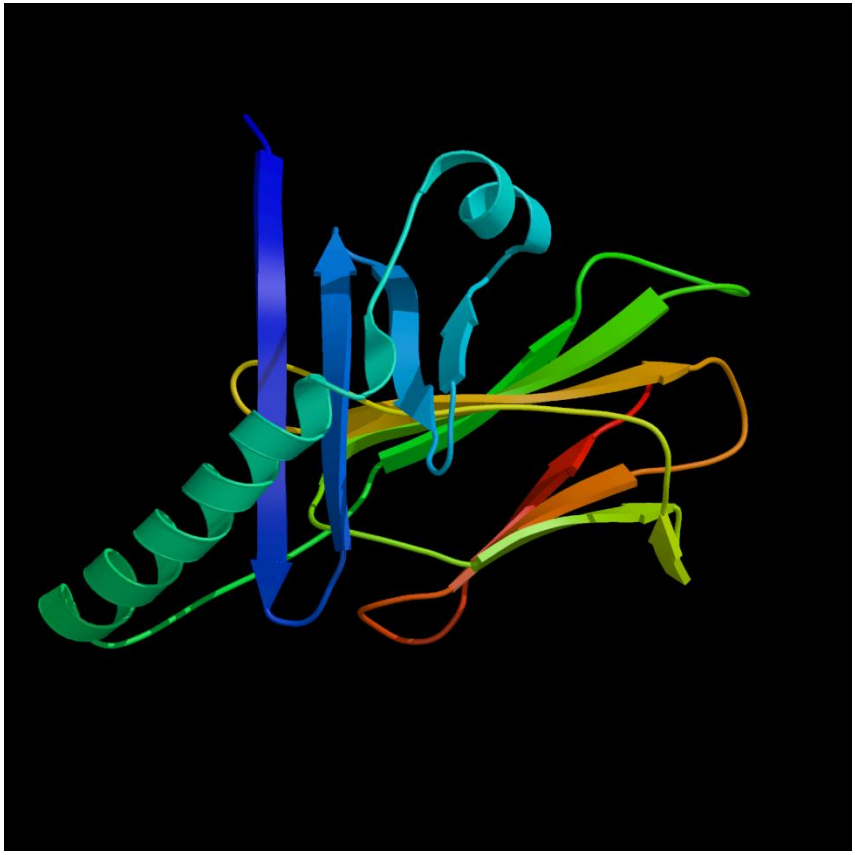
H=Alpha Helix, E=Beta strand, C=Coil/other

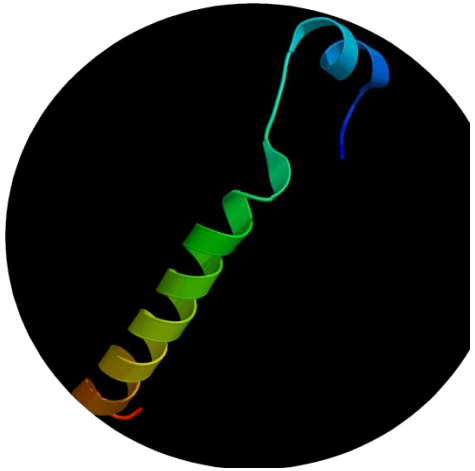
Start with some proteins of known structure

- Get some good X-ray or NMR models of proteins.
- Since we know their tertiary structures, certainly we can assign each residue in each protein a secondary state.
- Or can we?

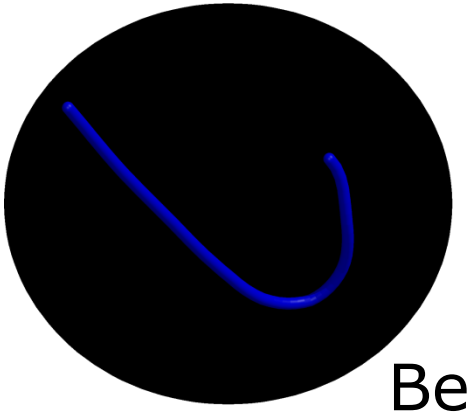
Secondary Structure Elements

ß-
strand





Helix



Bend

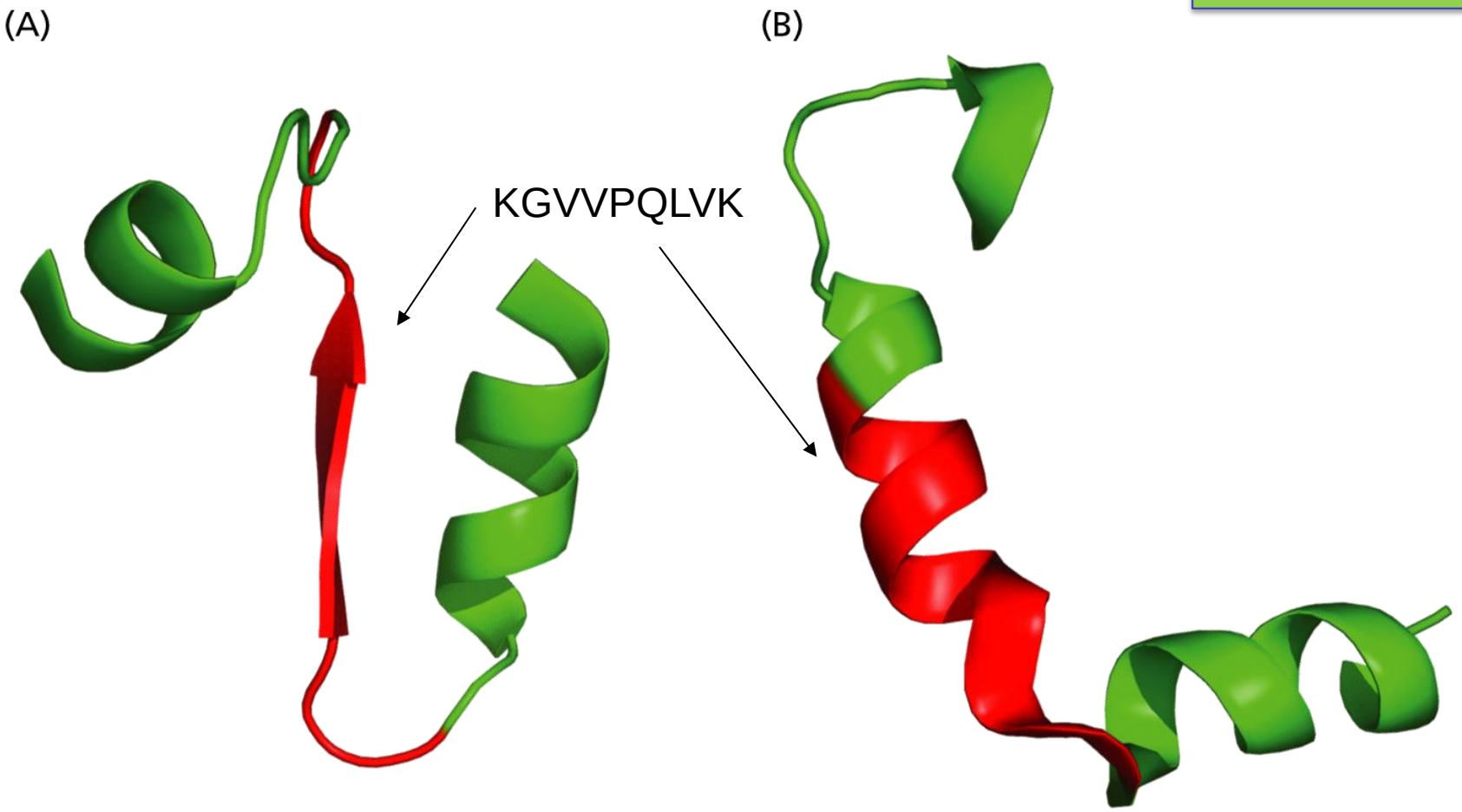
Turn

Is even that trivial?

- Is it even trivial to label the secondary state of each residue if we know the tertiary structure?
  - Where does a helix begin/end?
  - Is that a beta sheet or not?
  - …
- If the residue-state assignments are subjective, we're doomed!

Same sequence – different structure

Buch 12.12 (p.480)



DSSP to the rescue!

- In 1983 Kabsch and Sander introduced DSSP (Dictionary of Protein Secondary Structure)
- It automated the assignment of secondary structure from tertiary structure to make it less arbitrary.

ABSTRACT

For a successful analysis of the relation between amino acid sequence and protein structure, an unambiguous and physically meaningful definition of secondary structure is essential. We have developed a set of simple and physically motivated criteria for secondary structure, programmed as a pattern-recognition process of hydrogen-bonded and geometrical features extracted from x-ray coordinates. Cooperative secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge." Repeating turns are "helices," repeating bridges are "ladders," connected ladders are "sheets." Geometric structure is defined in terms of the concepts torsion and curvature of differential geometry. Local chain "chirality" is the torsional handedness of four consecutive C^{α} positions and is positive for right-handed helices and negative for ideal twisted β-sheets. Curved pieces are defined as "bends." Solvent "exposure" is given as the number of water molecules in possible contact with a residue. The end result is a compilation of the primary structure, including SS bonds, secondary structure, and

Kabsch and Sander. Dictionary of Secondary Structure in Proteins: pattern recognition of hydrogen-bonded and geometrical features. Biopolymer 22: 2571-2637 (1983)

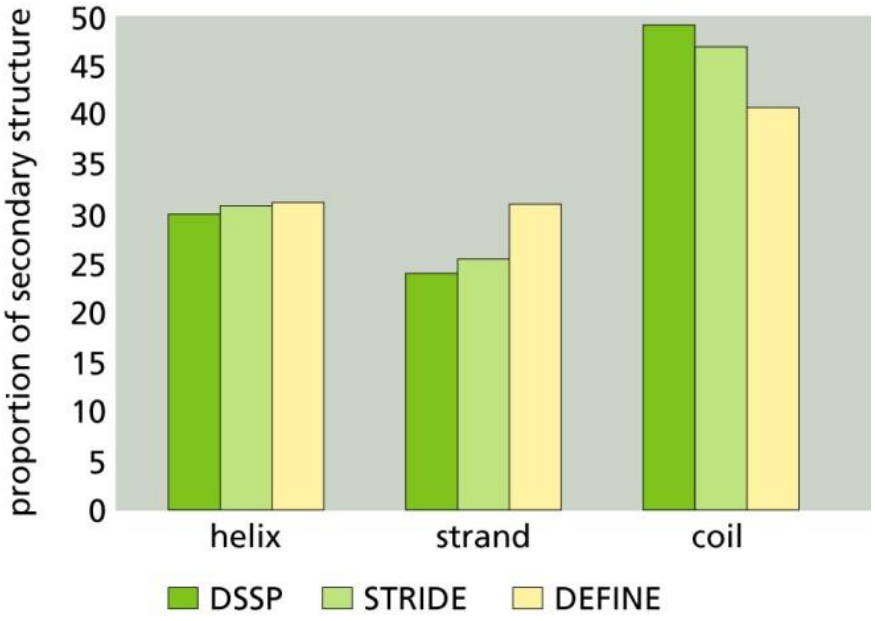
There are several ways to define protein secondary structures

- Defining features
  - Dihedral angles
  - Hydrogen bonds
  - Geometry
- Assigned manually by crystallographers or
- Automatic
  - **DSSP (Kabsch & Sander,1983)**
  - DEFINE (Richards & Kundrot, 1988)
  - STRIDE (Frishman & Argos, 1995)

Frischman and Argos. Knowledge-based secondary structure assignments. Proteins, 23:566-571 (1995) (STRIDE)

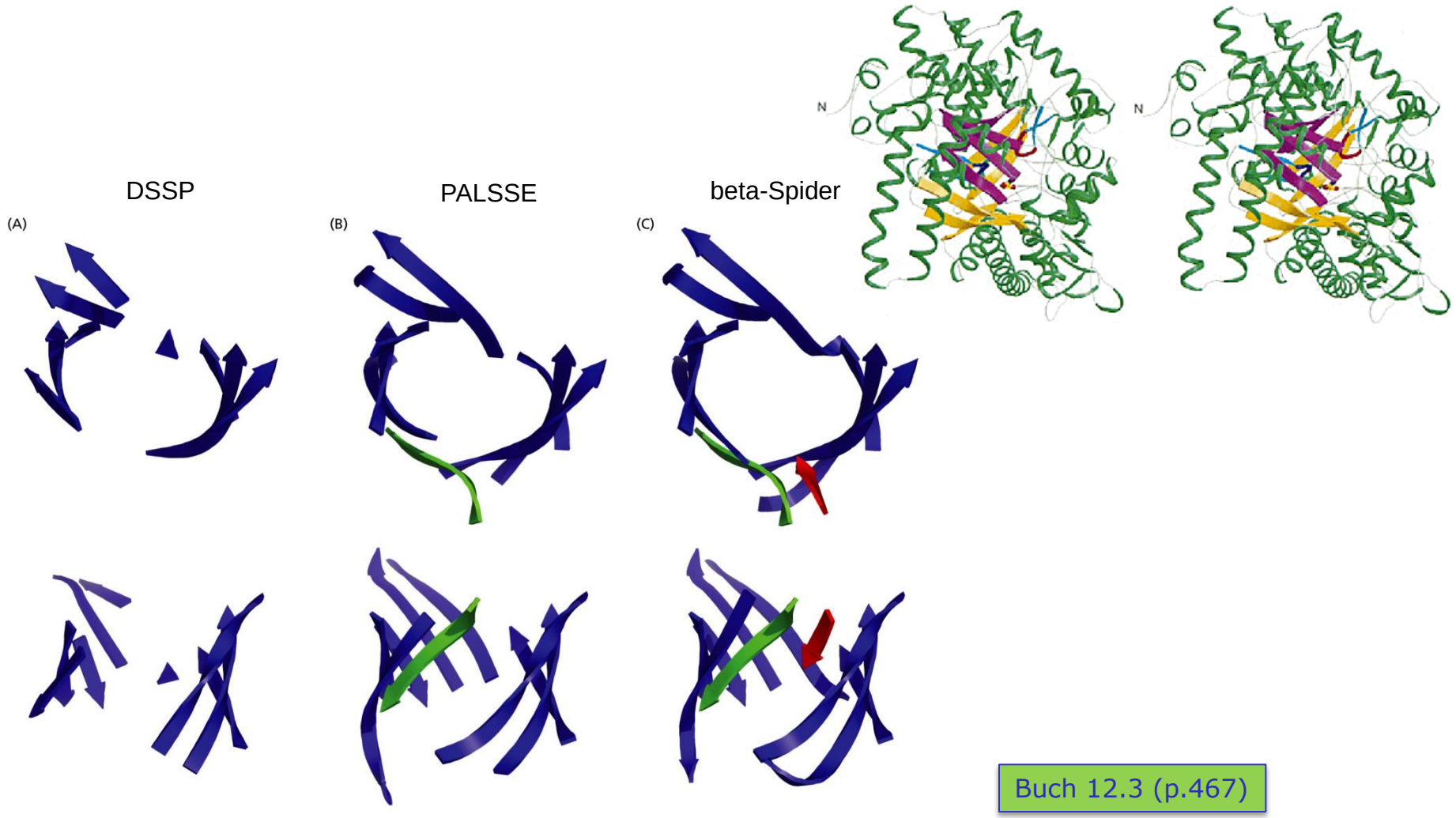
Freie Universität Berlin

- STRIDE and DEFINE are two other automatic "secondary-from-tertiary" programs.
- They agree (mostly) with DSSP.
- Moral: even when we know the tertiary structure, the "prediction" of secondary structure is hard!



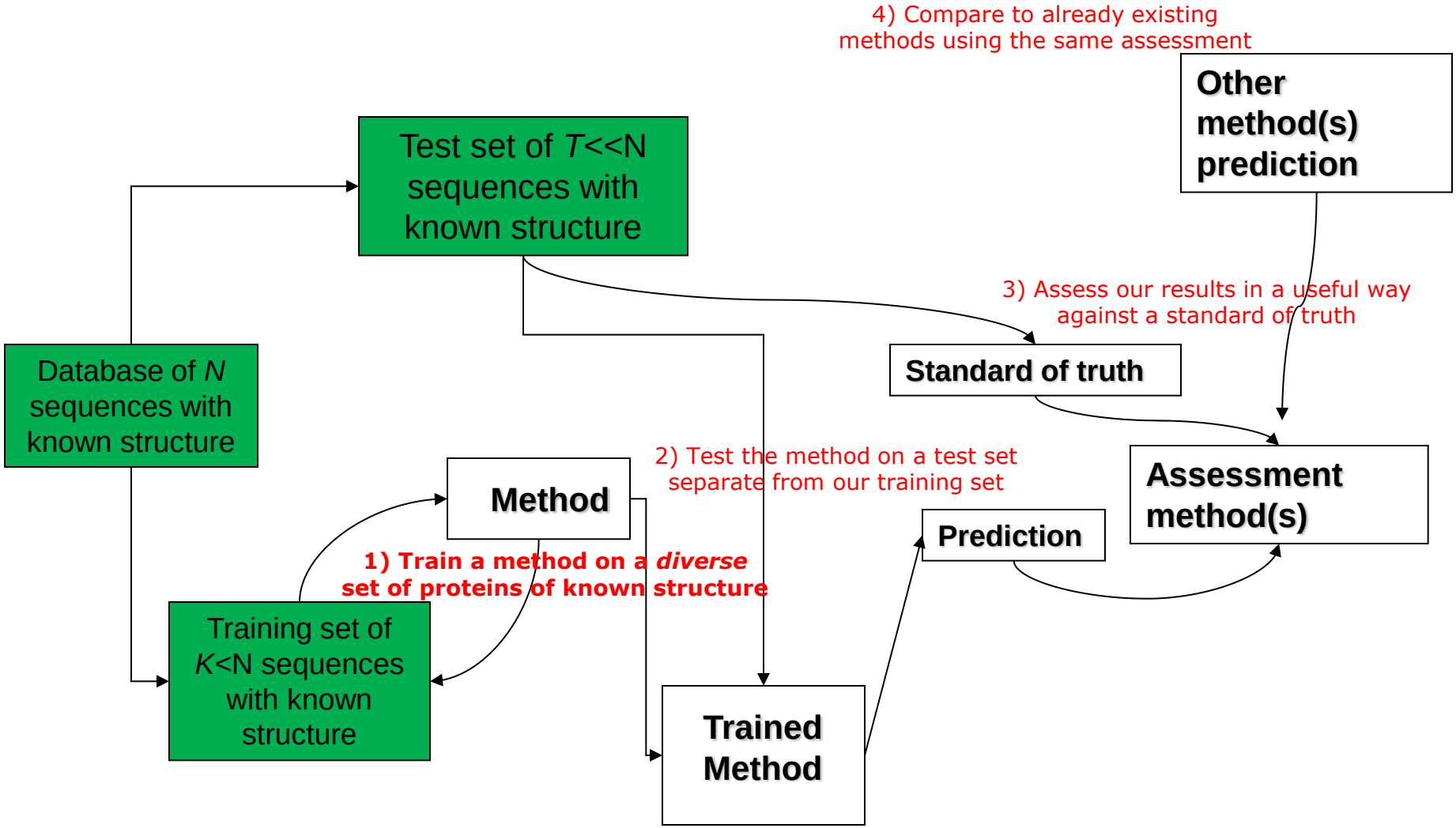
Buch 11.4 (p.417)

Automatic assignment is difficult…

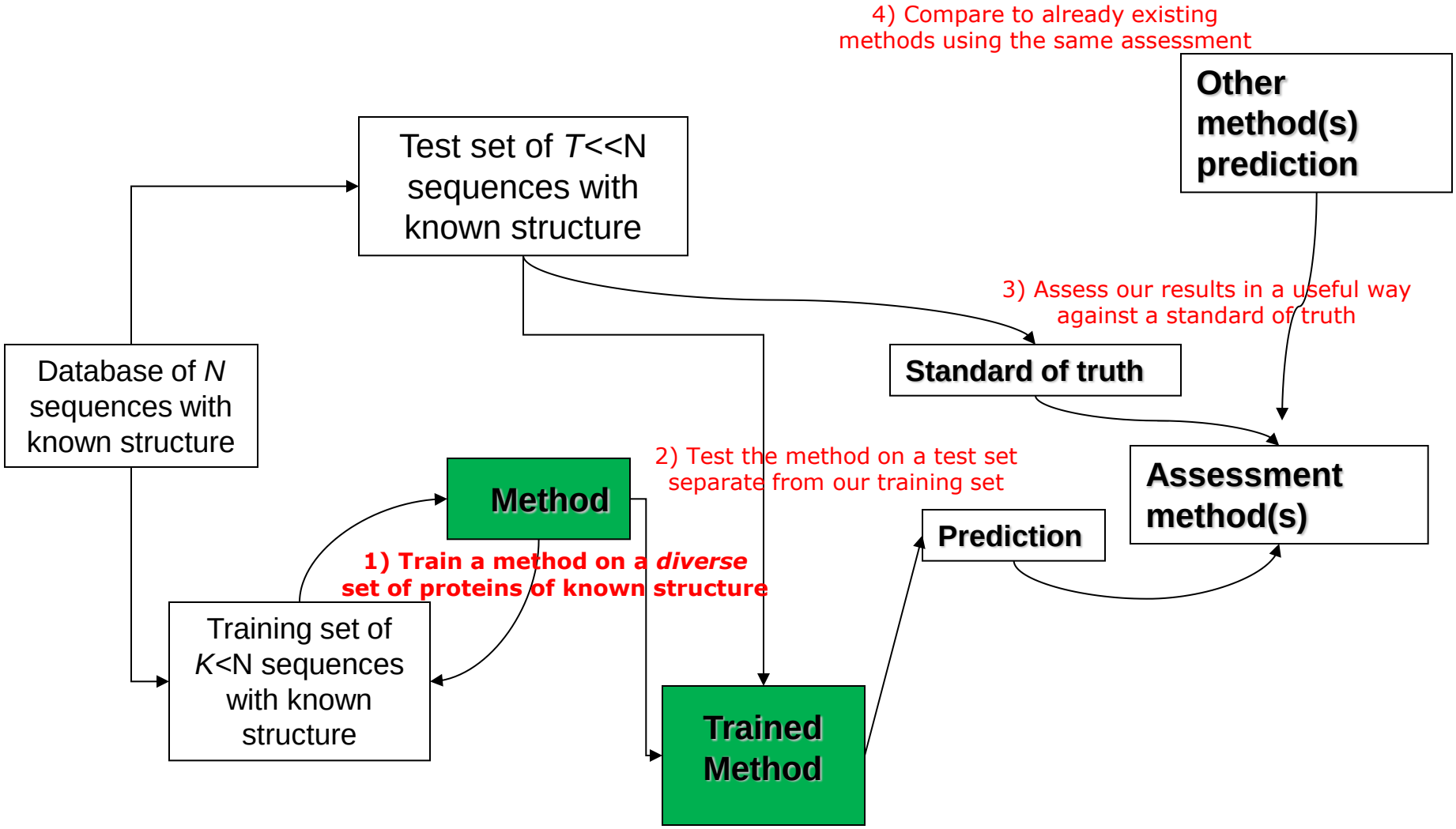


Buch 12.3 (p.467)

How to develop a method



How to develop a method



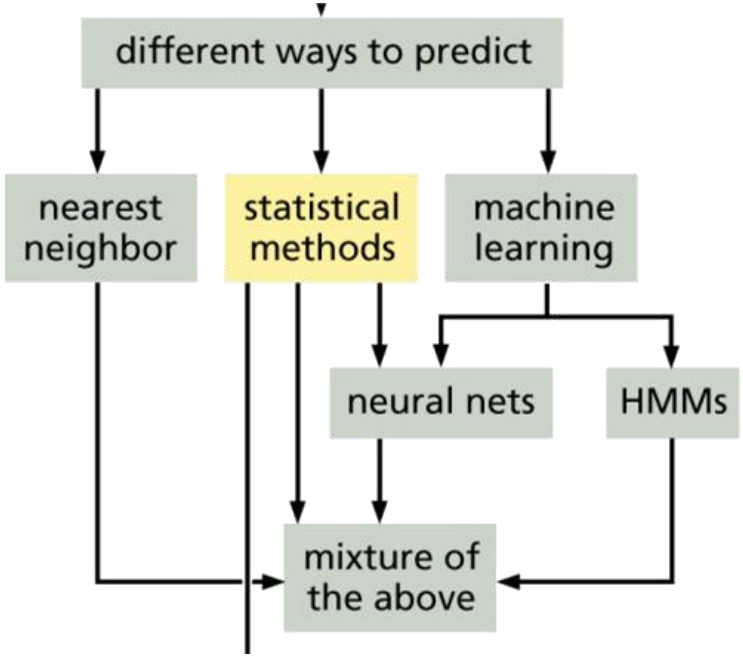
OK, now what?

- What can we learn from a set of proteins with each residue labeled as having a particular secondary structure state?
- How can we incorporate that knowledge into an automatic primary-to-secondary structure predictor?
- We need to decide for a method and teach it how to do predictions!

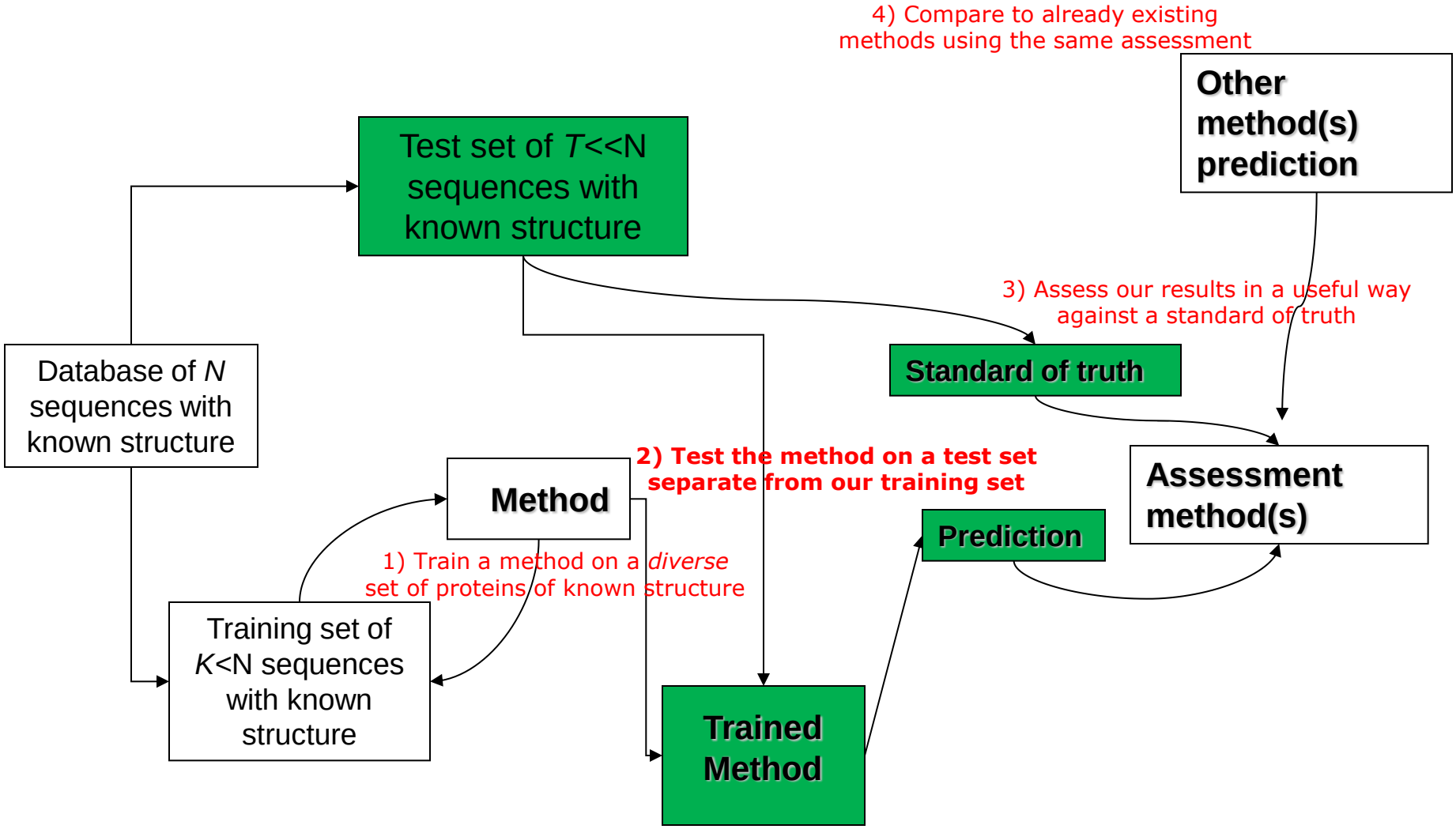
Classification Models: Different Classifiers

- Decision Trees
- Neural Networks
- Bayesian Networks
- Genetic Algorithms

Most of the best classifiers for secondary structure prediction are based on the Neural Network model.



How to develop a method



Test and Training Sets

- The golden rule of machine learning:
  - Don’t test and train on the same data!
- Why not?

Generalization

- We want to know how well a model will generalize to data it has never "seen".
- If we test (measure accuracy) on the same data we trained on:
  - We overestimate the generalization accuracy
  - We will tend to over-fit the training data (by adjusting the model design to fit it)

- **Predictive accuracy depends heavily on a choice of the test and training data.**

- There are many methods of choosing test and and training sets and hence evaluating the predictive accuracy. This is a separate field of research.

Cross-validation and hold-out sets

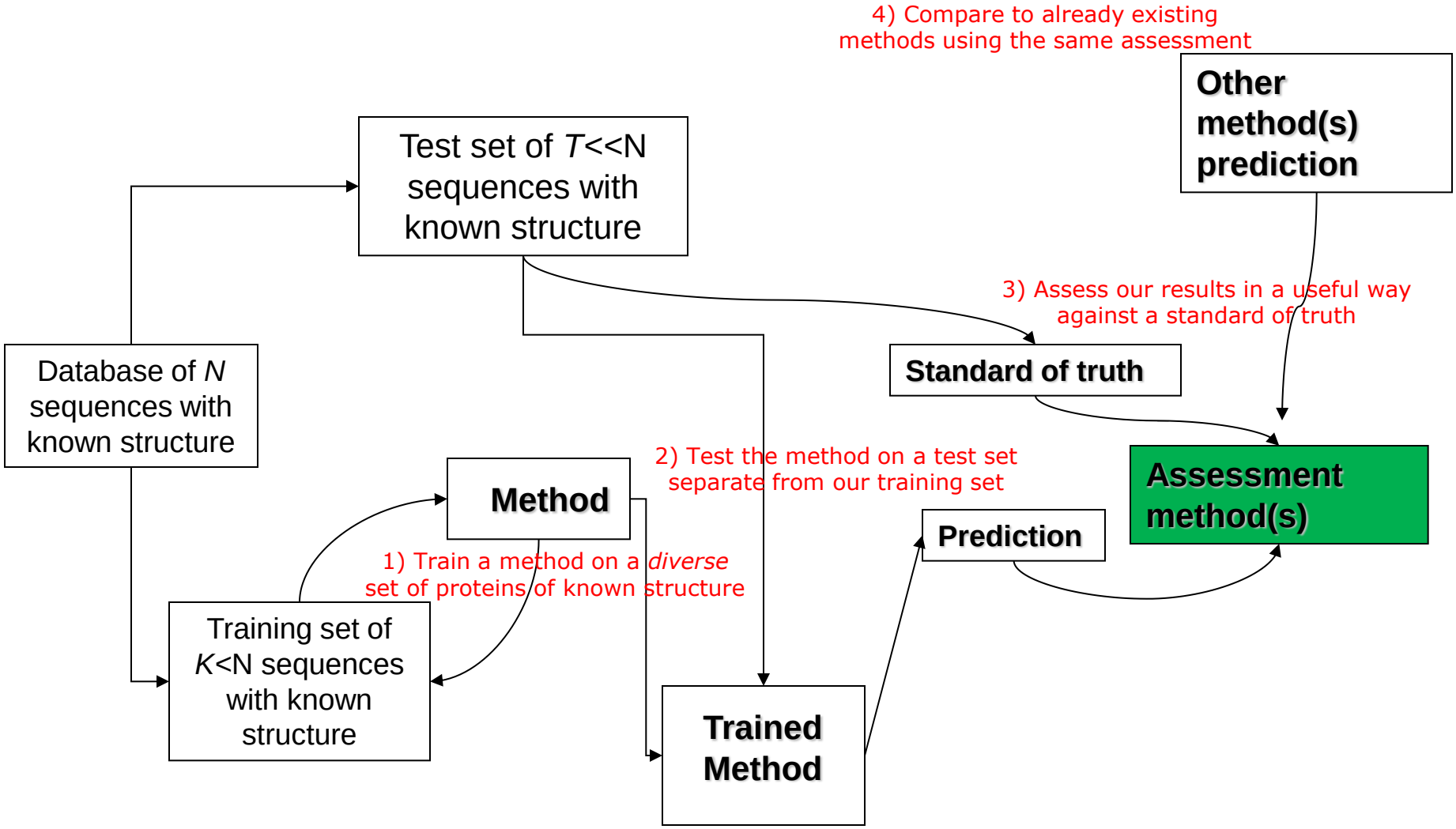
- The safest way to avoid biasing our results is with a "hold-out" set.
  - Lock some our data in a safe until we are all done designing and training our models.
  - Use the "held-out" data to measure the accuracy of our final model(s).

- Cross-validation
  - Split the data into n groups.
  - Train on n-1, test on 1.
  - Report average on the testing groups.

- The main methods of predictive accuracy evaluations are:
  - Re-substitution (N ; N)
  - Holdout (2N/3 ; N/3)
  - x-fold cross-validation (N-N/x ; N/x)
  - Leave-one-out (N-1 ; 1),

  where N is the number of instances in the dataset

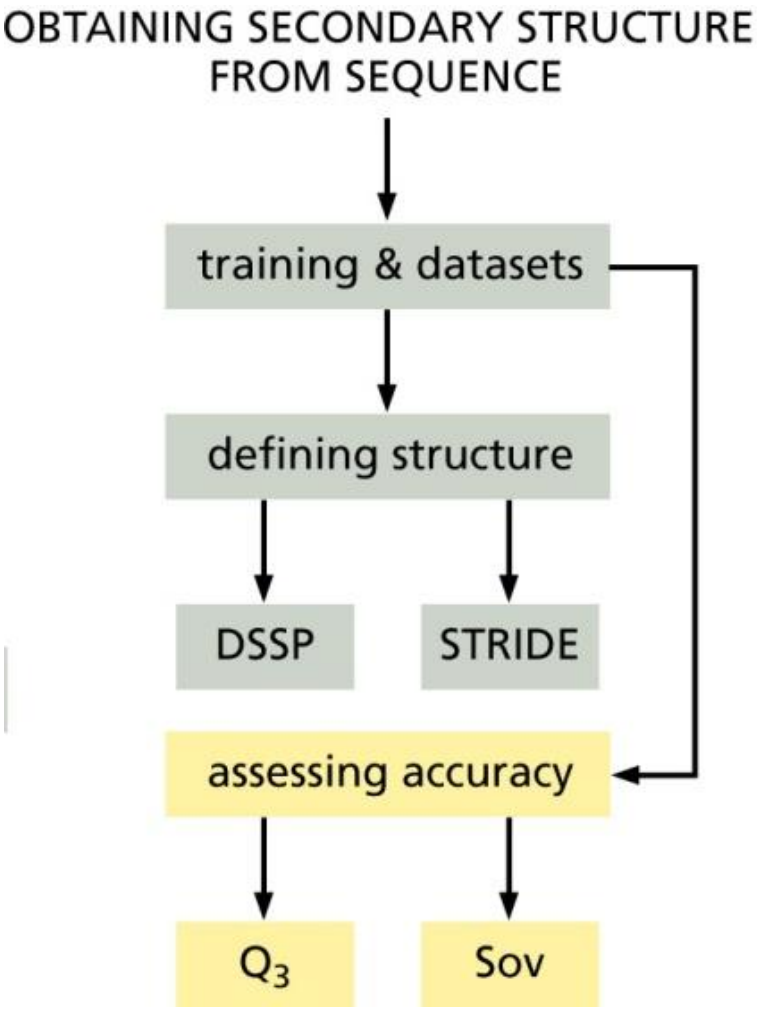
How to develop a method



Prediction Accuracy Evaluation

Q3 & SOV

Assessing the Accuracy of Prediction Programs



Secondary Structure Evaluation

- Q3 score
  - standard method in evaluating performance, 3 states (H,C,B) evaluated like a multiple choice exam with 3 choices. Same as % correct
- SOV (segment overlap score)
  - more useful measure of how segments overlap and how much overlap exists

Qindex: (Qhelix, Qstrand, Qcoil, Q3)

Percentage of residues correctly predicted as a-helix, b-strand, coil, or for all 3 conformations

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

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

- Secondary structure assignment in real proteins is uncertain to about 10%; Therefore, a "perfect" prediction would have Q3=90%.
- Even a random assignment of structure can achieve a high score Q3 for random prediction is 33% (*Holley & Karpus 1991*)

Buch 11.5 (p.418)

- Problems with per-residue accuracy measurement
  - Proteins with same 3D folding differ by 12% in Secondary Structure
  - This means maximum performance of Q_3 should be ~88%
  - End of segments might vary for proteins with same 3D structure (so their classification is less relevant to determining protein structure)
- SOV accounts for the following
  - Type and position of secondary structure segments rather than per-residue
  - Variation of the residues at the end of the segments

The Sov value measures the prediction accuracy for whole elements

SOV: segment overlap

- More useful to predict the correct number, type and order of secondary structure elements.
- If SOV is high, it will be easier to classify the protein into the correct fold.
- More complicated to compute.

```
x-ray: CCCEEEECCHHHHHHHHHCCCCEEEEECCCCCCCCCCCHHHHHHCCCC
pred:  CCCEEECCCCHHHHHHHHHCCCEECCCECCCCCCCCCCCHCHCHCHHCC
```

Buch 11.7 (p.419)

SOV

$$Sov = \frac{1}{N} \sum_{s} \frac{\text{minov}\ (s_1; s_2) + \delta}{\text{maxov}\ (s_1; s_2)} \times len(s_1),$$

$$Sov^{\delta} = \frac{1}{N} \sum_{s} \frac{\min\{\text{end}(s_1); \text{end}(s_2)\} - \max\{\text{beg}(s_1); \text{beg}(s_2)\} + 1 + \delta}{\max\{\text{end}(s_1); \text{end}(s_2)\} - \min\{\text{beg}(s_1); \text{beg}(s_2)\} + 1} * len(s_1)$$

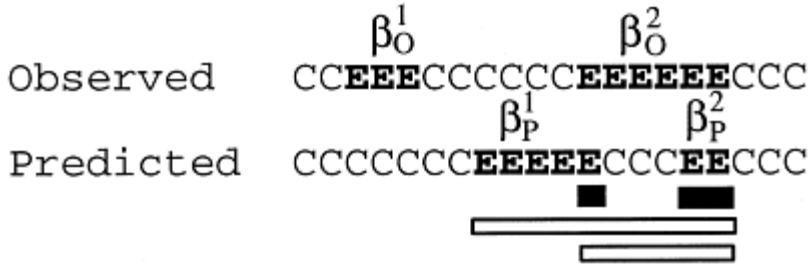


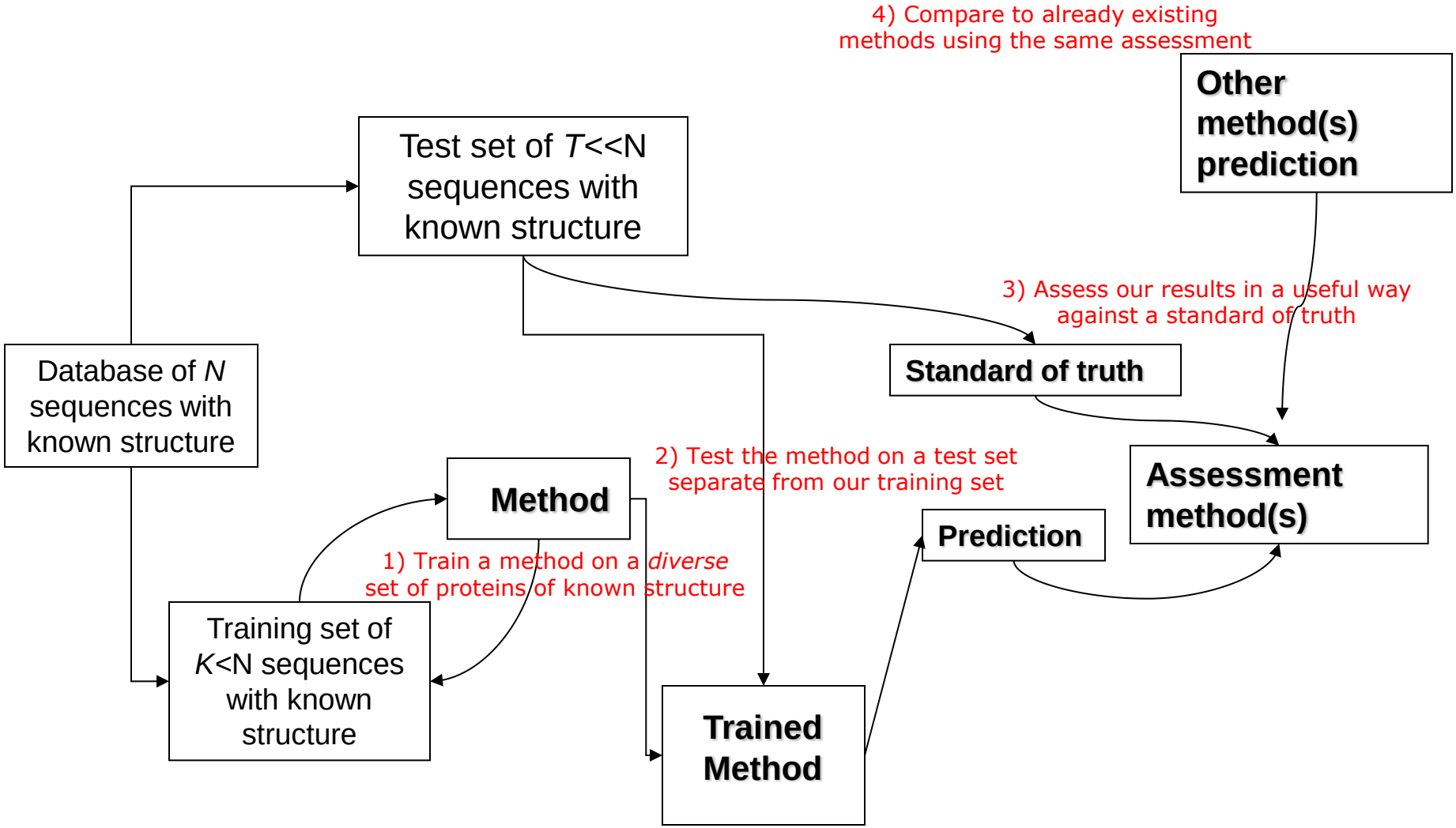
Illustration of a Sov(E) calculation. Black bars corresponds to minov and white bars to maxov in the overlapping segment pairs from observed and predicted structures.

Rost, B., Sander, C., Schneider, R.: Redefining the goals of protein secondary structure prediction. Journal of Molecular Biology 235, 13–26 (1994)

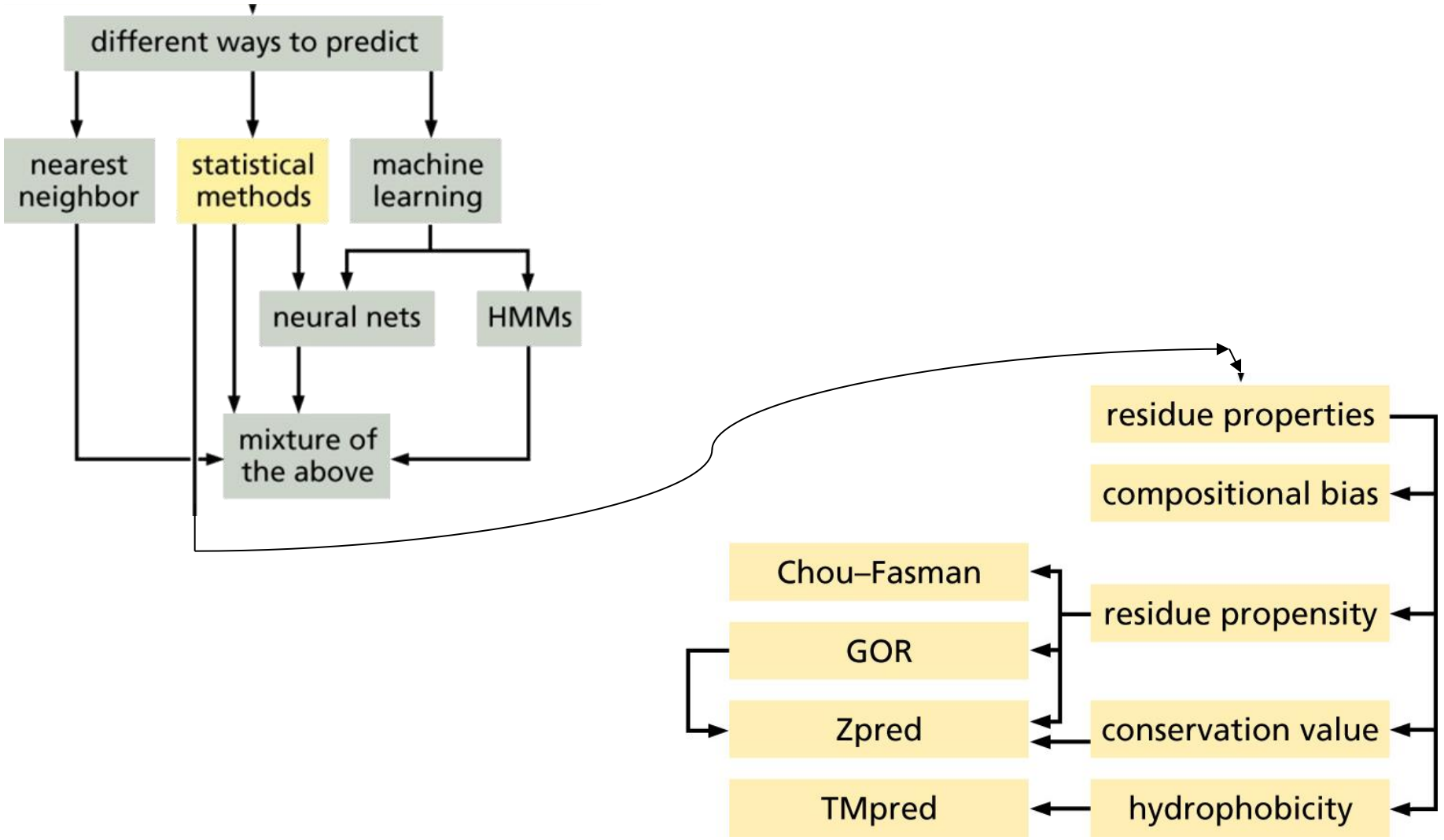
Prediction Algorithms

Nearest Neightbor, Statistical Methods

How to develop a method



Statistical and Knowledge-Based Methods



Types of Prediction Methods

1. Statistical methods are based on rules that give the probability that a residue will form part of a particular secondary structure

   1.2. Nearest-neighbor methods are statistical methods that incorporate additional information about protein structure (e.g. physicochemical properties)

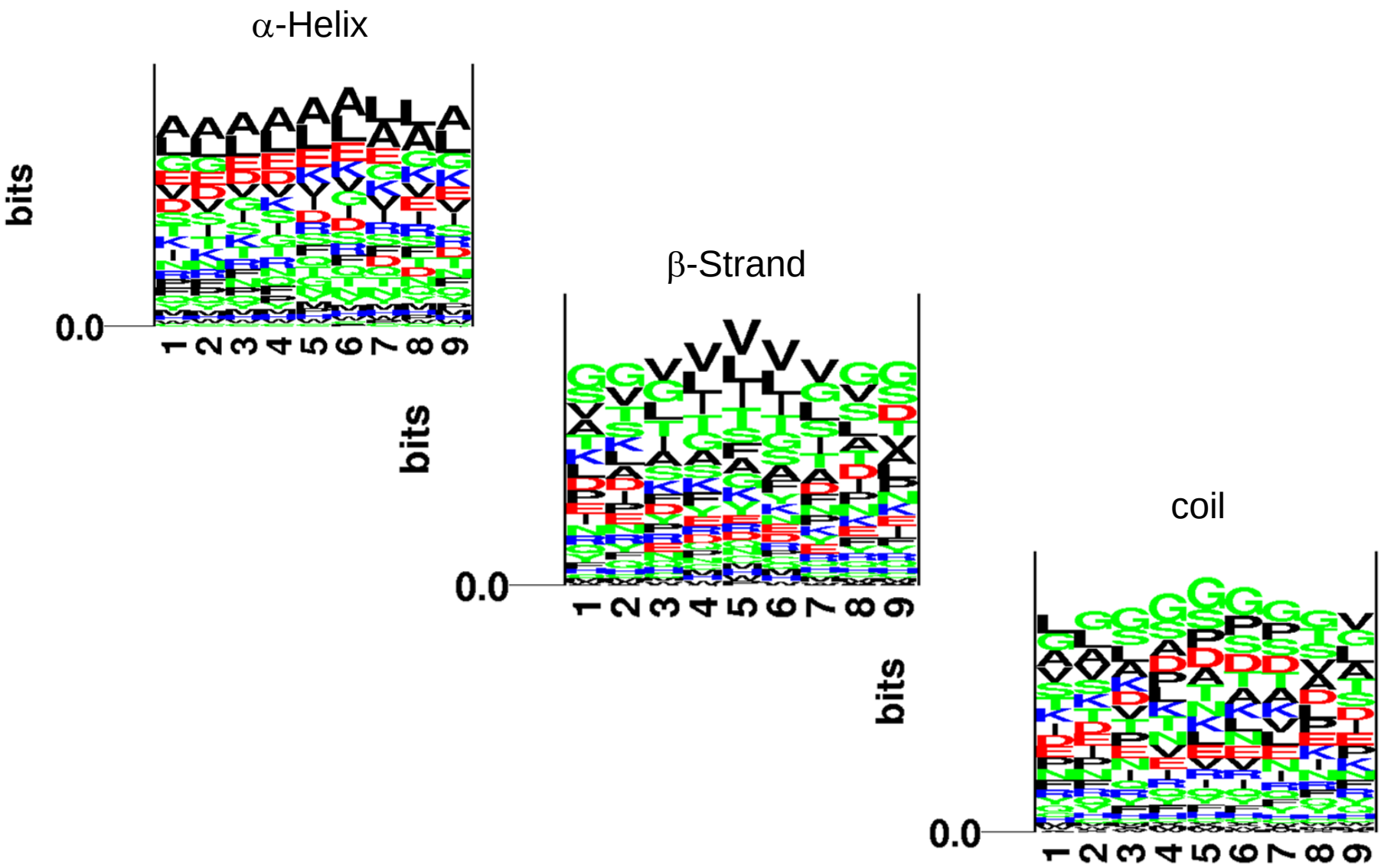
Why limit ourselves to single residues?

- Certain sequences of residues may occur frequently in a given secondary structure so find out:
  - What short "strings of residues" are common within or at the boundaries of secondary structures?
- The "nearest neighbor" idea compares a window of residues in the query protein to the database of labeled proteins.
- The conformations of the central residues in each of the closest matches can be used to create a prediction feature.

Early methods for Secondary Structure Prediction

- *Chou and Fasman*
  - (Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974)
- *GOR*
  - Garnier, Osguthorpe and Robson.
  - Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.
  - J. Mol. Biol., 120:97-120, 1978

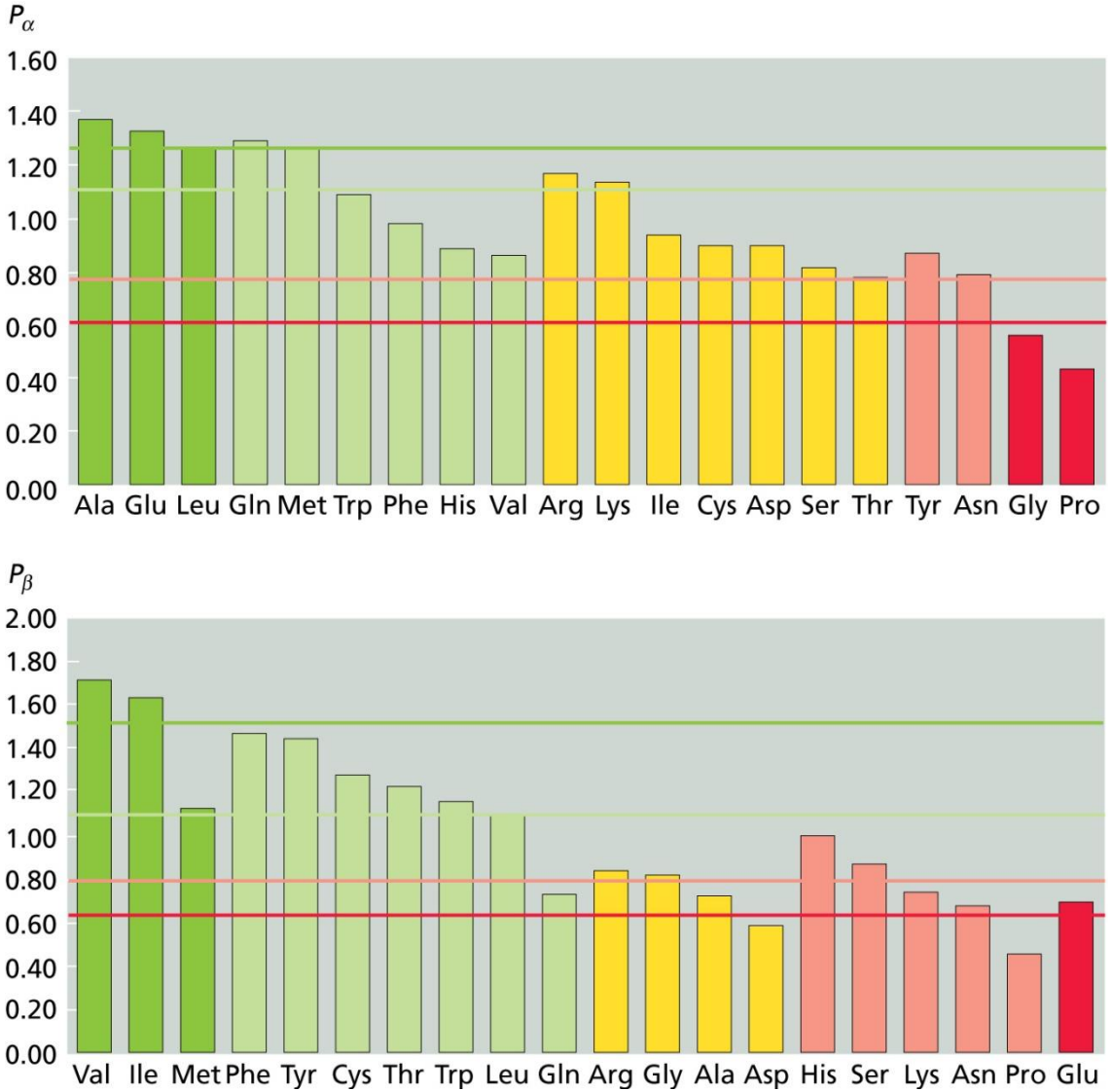
Amino acid preferences



Chou-Fasman Statistics

Chou & Fasman Secondary Structure Propensity of the Amino Acids							
	Pα	Pβ	Pc		Pα	Pβ	Pc
A	1.42	0.83	0.75	M	1.45	1.05	0.5
C	0.7	1.19	1.11	N	0.67	0.89	1.44
D	1.01	0.54	1.45	P	0.57	0.55	1.88
E	1.51	0.37	1.12	Q	1.11	1.1	0.79
F	1.13	1.38	0.49	R	0.98	0.93	1.09
G	0.57	0.75	1.68	S	0.77	0.75	1.48
H	1	0.87	1.13	T	0.83	1.19	0.98
I	1.08	1.6	0.32	V	1.06	1.7	0.24
K	1.16	0.74	1.1	W	1.08	1.37	0.45
L	1.21	1.3	0.49	Y	0.69	1.47	0.84

Chou-Fasman Statistics



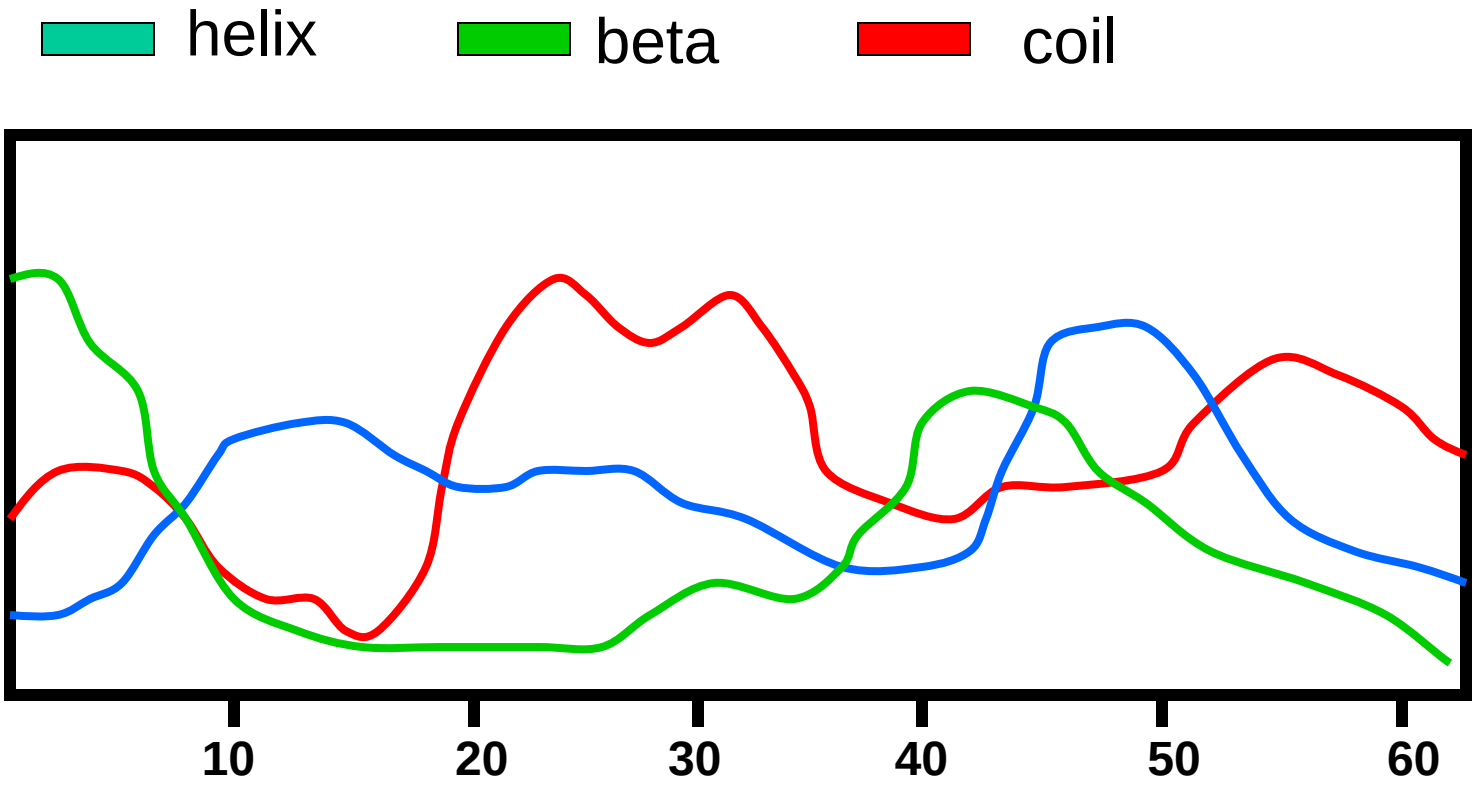
- Strong Former
- Former
- Indifferent
- Breaker
- Strong Breaker

Buch 12.8 (p.474)

Simplified C-F Algorithm

1. Select a window of 3-5 residues
2. Calculate average P_α over this window and assign that value to the central residue
3. Repeat the calculation for P_β and P_c
4. Slide the window down one residue and repeat until sequence is complete
5. Analyze resulting "plot" and assign secondary structure (H, B, C) for each residue to highest value.

Simplified C-F Algorithm



- Predicting helices:
  1. find nucleation site: 4 out of 6 contiguous residues with P(α)>1
  2. extension: extend helix in both directions until a set of 4 contiguous residues has an average P(α) < 1 (breaker)
  3. if average P(α) over whole region is >1, it is predicted to be helical

- Predicting strands:
  1. find nucleation site: 3 out of 5 contiguous residues with P(β)>1
  2. extension: extend strand in both directions until a set of 4 contiguous residues has an average P(β) < 1 (breaker)
  3. if average P(β) over whole region is >1, it is predicted to be a strand

Chou and Fasman

- *Position-specific parameters for turn:*
  - Each position has distinct amino acid preferences.
- Examples:
  - At position 2, Pro is highly preferred; Trp is disfavored
  - At position 3, Asp, Asn and Gly are preferred
  - At position 4, Trp, Gly and Cys preferred

	f(i)	f(i+1)	f(i+2)	f(i+3)
Ala	0.060	0.076	0.035	0.058
Arg	0.070	0.106	0.099	0.085
Asp	0.147	0.110	0.179	0.081
Asn	0.161	0.083	0.191	0.091
Cys	0.149	0.050	0.117	0.128
Glu	0.056	0.060	0.077	0.064
Gln	0.074	0.098	0.037	0.098
Gly	0.102	0.085	0.190	0.152
His	0.140	0.047	0.093	0.054
Ile	0.043	0.034	0.013	0.056
Leu	0.061	0.025	0.036	0.070
Lys	0.055	0.115	0.072	0.095
Met	0.068	0.082	0.014	0.055
Phe	0.059	0.041	0.065	0.065
Pro	0.102	0.301	0.034	0.068
Ser	0.120	0.139	0.125	0.106
Thr	0.086	0.108	0.065	0.079
Trp	0.077	0.013	0.064	0.167
Tyr	0.082	0.065	0.114	0.125
Val	0.062	0.048	0.028	0.053

- Predicting turns:
  - for each tetrapeptide starting at residue i, compute:
    - P_{Turn} (average propensity over all 4 residues)
    - F = f(i)*f(i+1)*f(i+2)*f(i+3)
  - tetrapeptide is considered a turn, if:

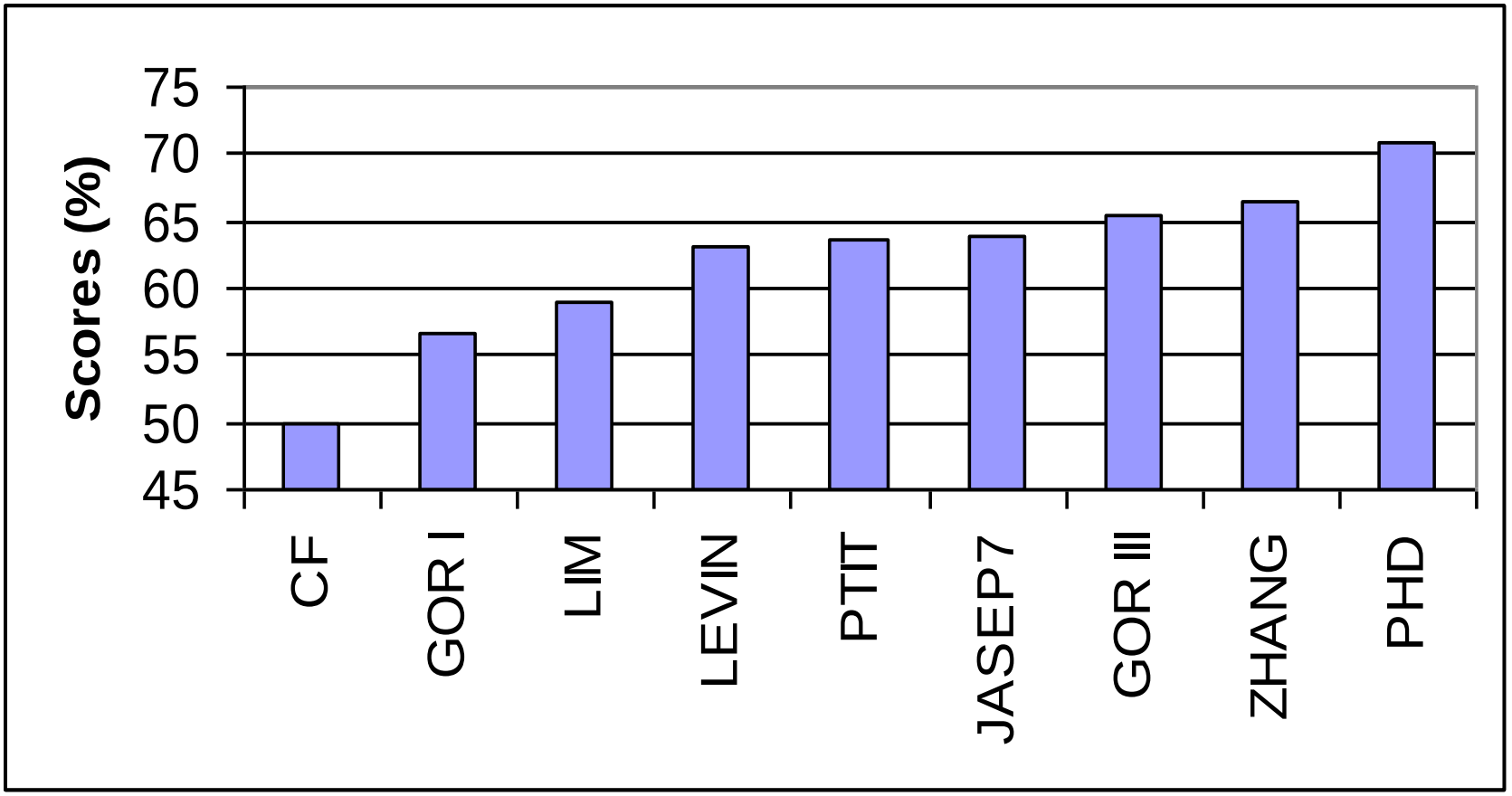
$P_{Turn} > Pa$ and
$P_{Turn} > Pb$ and
$P_{Turn} > 1$ and
$F > 0.000075$

Chou and Fasman online:
http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

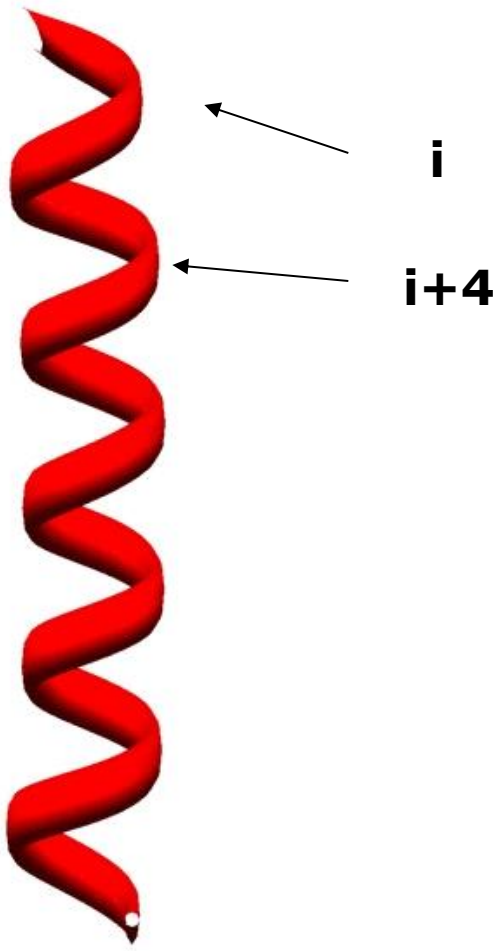
Limitations of Chou-Fasman

- Does not take into account
  - long range information (>3 residues away)
  - structure class
- Does not include
  - related sequences or alignments in prediction process
- Only about 55% accurate

Prediction Performance

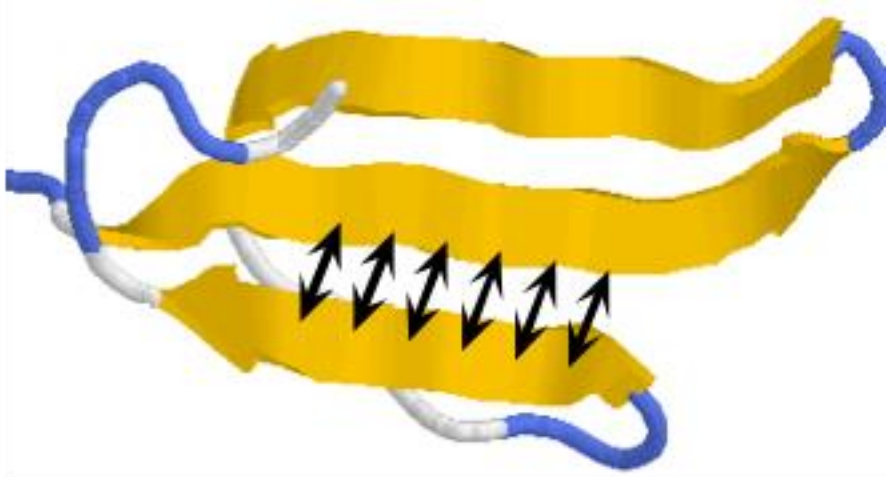


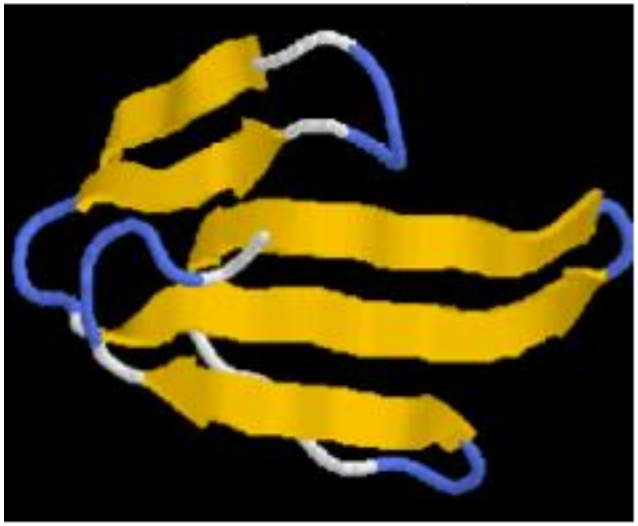
CF works good for predicting Helix formations (is local)



THYROID hormone receptor (2nll)

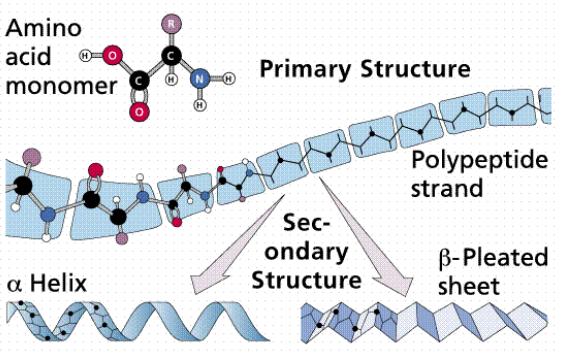
CF works not good for predicting β-sheet formation - is NOT local

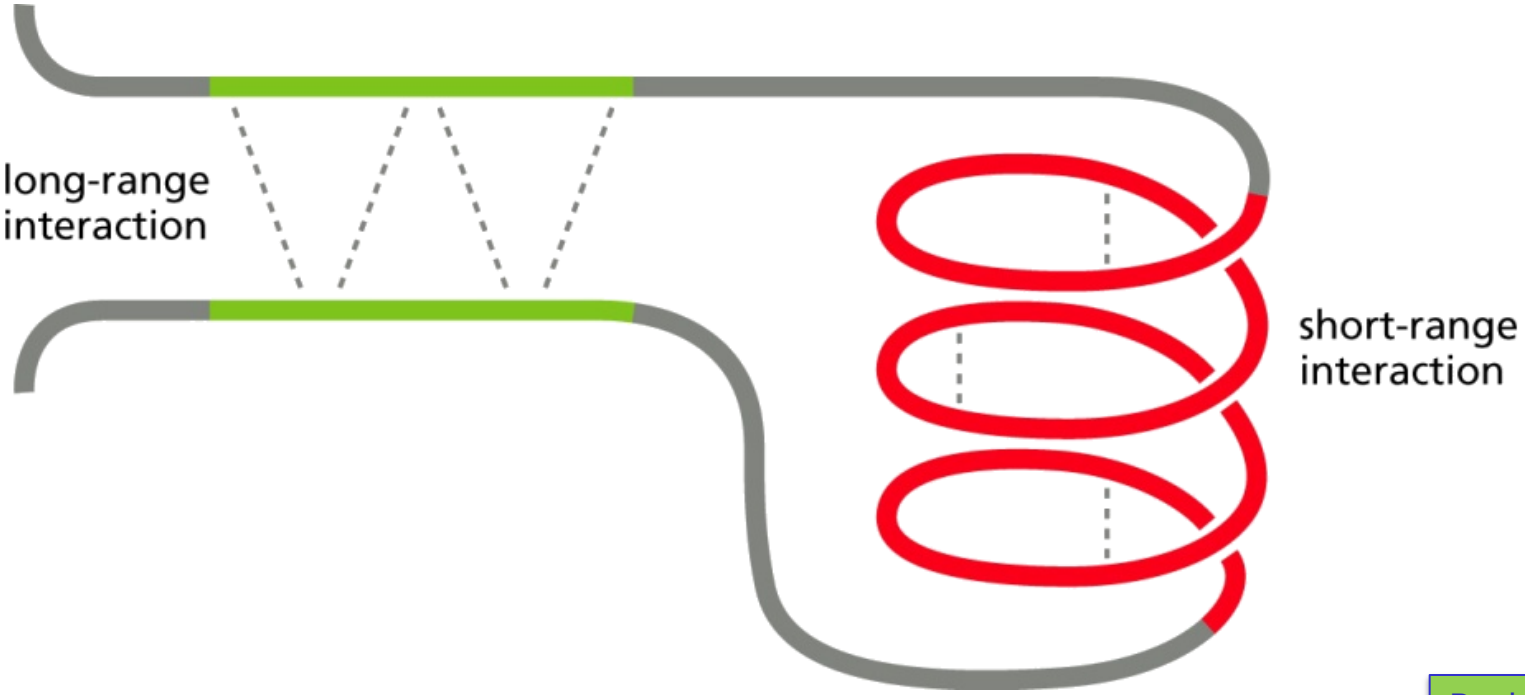




Erabutoxin β (3ebx)

CF Main issues: only local information





Buch 11.19 (p.429)

Early methods for Secondary Structure Prediction

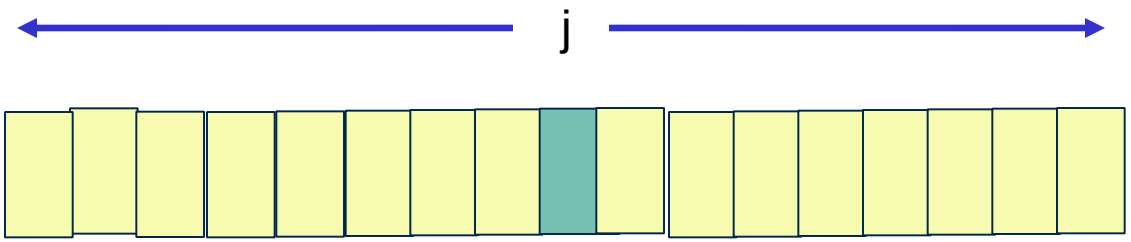
- *Chou and Fasman*
  - (Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974)
- *GOR*
  - 2nd generation method
  - Garnier, Osguthorpe and Robson.
  - Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.
  - J. Mol. Biol., 120:97-120, 1978

The GOR method

- Developed by Garnier, Osguthorpe & Robson
- Build on Chou-Fasman *Pij* values
- Evaluate each residue PLUS adjacent 8 N-terminal and 8 carboxyl-terminal residues
- Sliding window of 17 residues
- Improved detectino of β-strand regions (but still underpredicted)
- GOR method accuracy Q3 = ~64%

The GOR method

- Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.

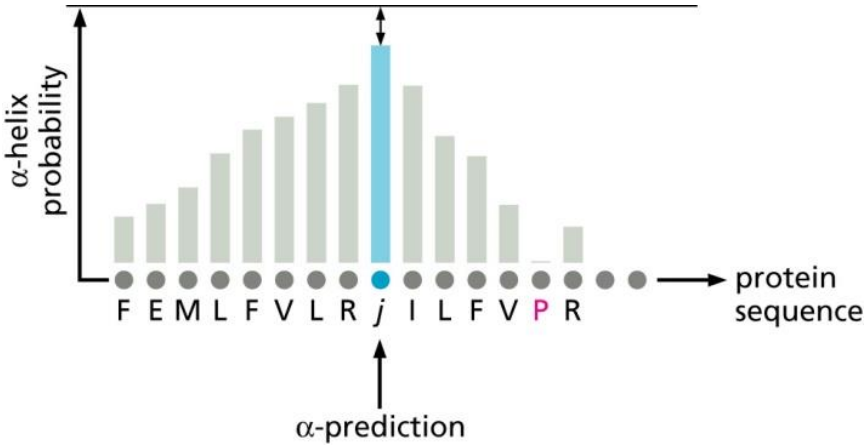


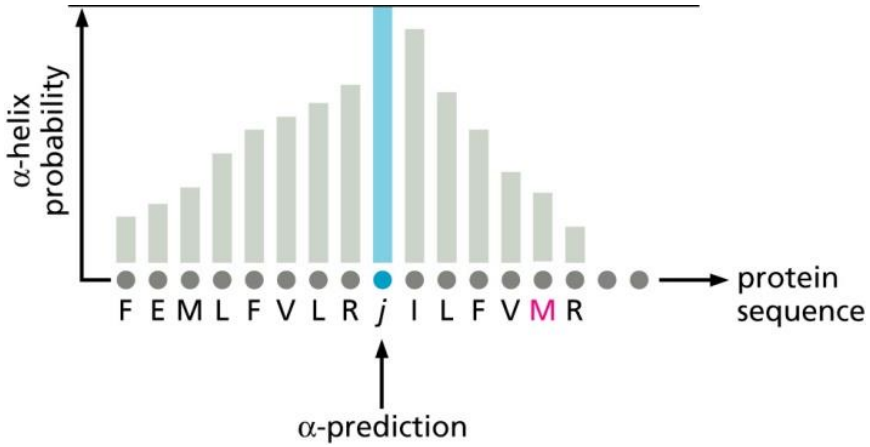
- A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries.
- Build similar tables for strands and turns.

- *GOR simplification:*
  - The predicted state of AAj is calculated as the sum of the position-dependent propensities of all residues around AAj.
  - Based on assumption that each amino acid individually influences the propensity of the central residue to adopt a particular secondary structure. Each flanking position evaluated independently like a PSSM.

- GOR (IV) can be used at : http://abs.cit.nih.gov/gor/

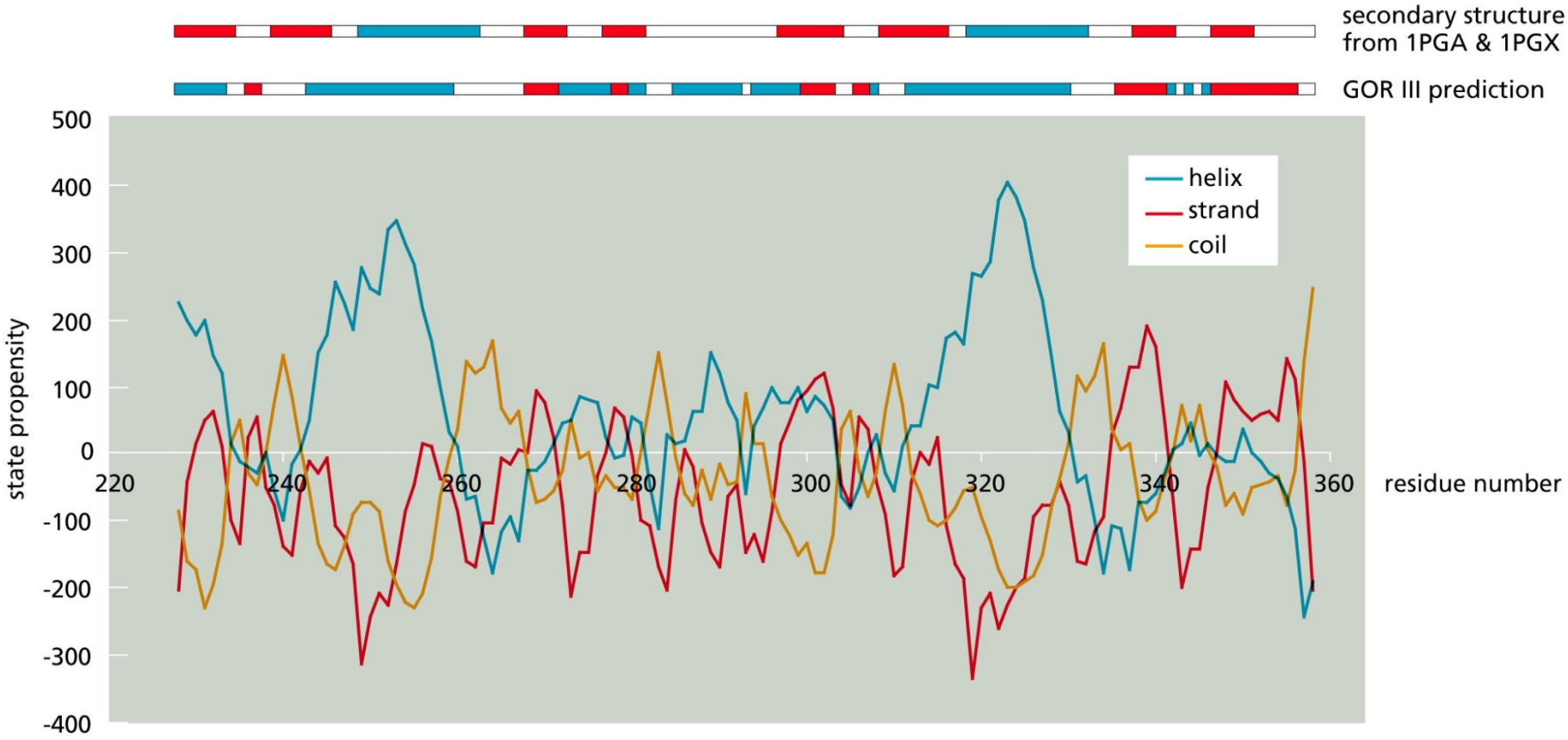
GOR: Influence of the neighborhood





Buch 11.13 (p.425)

Example results for GORIII



Buch 12.14 (p.484)

A small excurse…

Bayes formula

$$X : \{x_i\}, \quad Y : \{y_j\},$$

Count of $(x_i, y_j) = N_{ij},$

$$N_{i\bullet} = \sum_j N_{ij}, \quad N_{\bullet j} = \sum_i N_{ij}, \quad N = \sum_{i,j} N_{ij} = \sum_i N_{i\bullet} = \sum_j N_{\bullet j},$$

$$P_{\bullet j} = N_{\bullet j}/N, \quad \text{Prob}(x_i|y_j) \equiv P(i|j) = N_{ij}/N_{\bullet j},$$

$$P_{ij} = N_{ij}/N = (N_{ij}/N_{\bullet j})(N_{\bullet j}/N) = P(i|j)P_{\bullet j}.$$

Generally, P(x, y) = P(x|y)P(y),

$$P(x|y) = \frac{P(x,y)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}.$$

Protein sequence A, $\{a_i\}$, i=1,2,…,n
Secondary structure sequence S, $\{s_i\}$, i=1,2,…,n

1. Simple Chou-Fasman approach

$$\frac{P(S|A)}{P(S)} = \frac{P(A,S)}{P(A)P(S)} \approx \prod_{i=1}^{n} \frac{P(a_i,s_i)}{P(s_i)P(a_i)} \propto \prod_{i=1}^{n} \frac{P(a_i|s_i)}{P(a_i)},$$

Chou-Fasman's propensity of amino acid to conformational state

$$\frac{P(a_i,s_i)}{P(s_i)P(a_i)}$$

+ independence approximation

One residue, one state

GOR window version

$$\frac{P(a_i, s_i)}{P(s_i)P(a_i)} \longrightarrow \frac{P(W_i|s_i)}{P(W_i)} \equiv \frac{P(a_{i-8}\cdots a_{i-1}a_i a_{i+1}\cdots a_{i+8}|s_i)}{P(a_{i-8}\cdots a_{i-1}a_i a_{i+1}\cdots a_{i+8})}$$

$$\sim \frac{P(a_{i-8}\cdots a_{i-1}a_i a_{i+1}\cdots a_{i+8}|s_i)}{P(a_{i-8}\cdots a_{i-1}a_i a_{i+1}\cdots a_{i+8}|\bar{s}_i)}$$

Conditional Independency

$$\approx \frac{P(a_{i-8}|s_i)\cdots P(a_{i-1}|s_i)P(a_i|s_i)P(a_{i+1}|s_i)\cdots P(a_{i+8}|s_i)}{P(a_{i-8})\cdots P(a_{i-1})P(a_i)P(a_{i+1})\cdots P(a_{i+8})}$$

Weight matrix (20x17)x3 P(W|s)

3. Improved GOR (20x20x16x3, to include pair correlation)

$$\frac{P(W_i|s_i)}{P(W_i)} = \frac{P(W_i'|a_i, s_i)P(a_i|s_i)}{P(W_i'|a_i)P(a_i)}$$

$$\approx \frac{P(a_{i-8}|a_i, s_i)\cdots P(a_{i-1}|a_i, s_i)P(a_{i+1}|a_i, s_i)\cdots P(a_{i+8}|a_i, s_i)}{P(a_{i-8}|a_i)\cdots P(a_{i-1}|a_i)P(a_{i+1}|a_i)\cdots P(a_{i+8}|a_i)}\frac{P(a_i|s_i)}{P(a_i)}$$

Width determined by mutual information.

Directional Information

$$\text{LOD} = \ln \frac{\text{OBS}}{\text{EXP}}$$

helix

strand

coil

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-12	-15	-12	-12	-17	-13	-25	-24	-32	-35	-32	-29	-24	-20	-12	-5	-6
c	36	26	41	50	45	31	29	19	7	5	27	29	38	48	41	45	59
d	-8	-10	-13	-8	-13	-10	12	25	50	43	39	27	7	-7	-4	-9	-5
e	-3	-11	-10	-11	-10	-7	-5	-23	-26	-23	-2	5	-1	-3	3	-5	-9
f	22	25	28	25	21	9	-23	-34	-49	-40	-29	-12	9	20	13	18	13
g	-3	-8	-18	-17	-7	2	26	68	97	58	19	-2	-18	-14	-18	-11	-11
h	15	9	-4	-7	8	-2	12	8	8	5	-4	1	-3	-5	-5	-10	-9
i	7	12	19	14	7	1	-21	-42	-66	-55	-26	-14	14	18	4	2	1
k	-12	-7	-10	-9	-1	5	11	5	0	9	5	-8	-20	-15	-7	-10	-12
l	2	8	11	11	11	2	-23	-42	-65	-63	-52	-39	-15	-11	-10	-6	0
m	11	14	4	3	-9	-16	-33	-52	-62	-77	-71	-54	-32	-7	3	9	9
n	-2	-8	-11	1	8	12	32	51	61	31	18	6	-6	-8	-4	2	2
p	4	8	4	-1	5	15	39	76	120	159	98	59	32	17	11	3	0
q	-1	-11	-12	-15	-17	-4	5	-5	-13	1	1	2	-2	-5	-1	-9	-20
r	-4	-9	-8	-10	-10	-13	-18	-16	-14	-9	-14	-16	-14	-11	-5	-3	-2
s	-3	-4	-4	-4	4	11	22	26	41	31	20	13	3	5	4	8	11
t	-5	-5	-5	-4	-7	-5	0	2	15	21	29	30	19	7	3	-4	-5
v	3	17	20	20	8	-2	-26	-46	-68	-51	-20	3	25	24	23	15	11
w	5	9	28	28	12	-16	-32	-46	-53	-38	-20	5	13	30	9	2	16
y	10	7	12	7	6	3	7	-1	-31	-14	-11	11	13	1	3	12	15

Table 3. *Directional informational parameters: I(Sj = x:x': Rj + m) for residue position versus residue type for α-helices*[a]

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	19	21	22	24	34	36	44	47	60	60	53	50	44	40	31	23	24
c	-47	-45	-44	-47	-44	-36	-44	-55	-56	-58	-54	-55	-58	-58	-59	-53	-66
d	14	15	14	15	17	21	15	17	-7	-11	-31	-42	-28	-12	-8	1	-5
e	14	16	15	20	26	27	34	52	62	57	32	15	19	12	6	7	9
f	-19	-14	-10	-4	-2	-1	6	-1	10	10	12	12	-4	-5	2	0	2
g	5	2	1	-5	-22	-30	-50	-70	-92	-52	-28	-21	-13	-17	-8	-6	-6
h	-22	-20	-9	-10	-19	-10	-14	-7	-11	-4	0	-3	-2	2	6	11	12
i	7	7	0	0	1	1	2	-5	1	2	1	7	-6	-3	10	8	6
k	-2	-1	-1	-1	-6	-9	-6	5	17	17	21	27	35	33	21	22	23
l	0	-1	0	6	9	16	30	33	45	47	51	53	37	32	30	25	18
m	4	3	15	23	30	30	39	36	45	54	57	53	44	29	30	14	1
n	2	3	2	-5	-9	-10	-16	-17	-31	-16	-17	-16	-9	-8	-9	-10	-5
p	-12	-15	-14	-19	-23	-25	-30	-48	-82	-195	-145	-104	-67	-49	-43	-33	-17
q	-4	3	7	4	13	8	10	24	35	32	31	21	18	18	9	8	6
r	5	3	6	13	7	13	19	27	34	32	36	41	33	29	23	21	18
s	-10	-7	-10	-10	-16	-17	-25	-21	-39	-35	-39	-41	-32	-35	-34	-35	-33
t	1	-1	-6	-8	-6	-11	-16	-25	-48	-47	-48	-46	-34	-31	-34	-26	-24
v	-5	-12	-13	-14	-13	-19	-17	-20	-15	-22	-22	-20	-26	-19	-15	-10	-5
w	0	-4	-12	-19	-7	14	16	12	18	17	12	8	1	-6	1	3	-13
y	-22	-19	-17	-20	-16	-21	-30	-32	-8	-10	-4	-12	-17	-9	-10	-14	-15

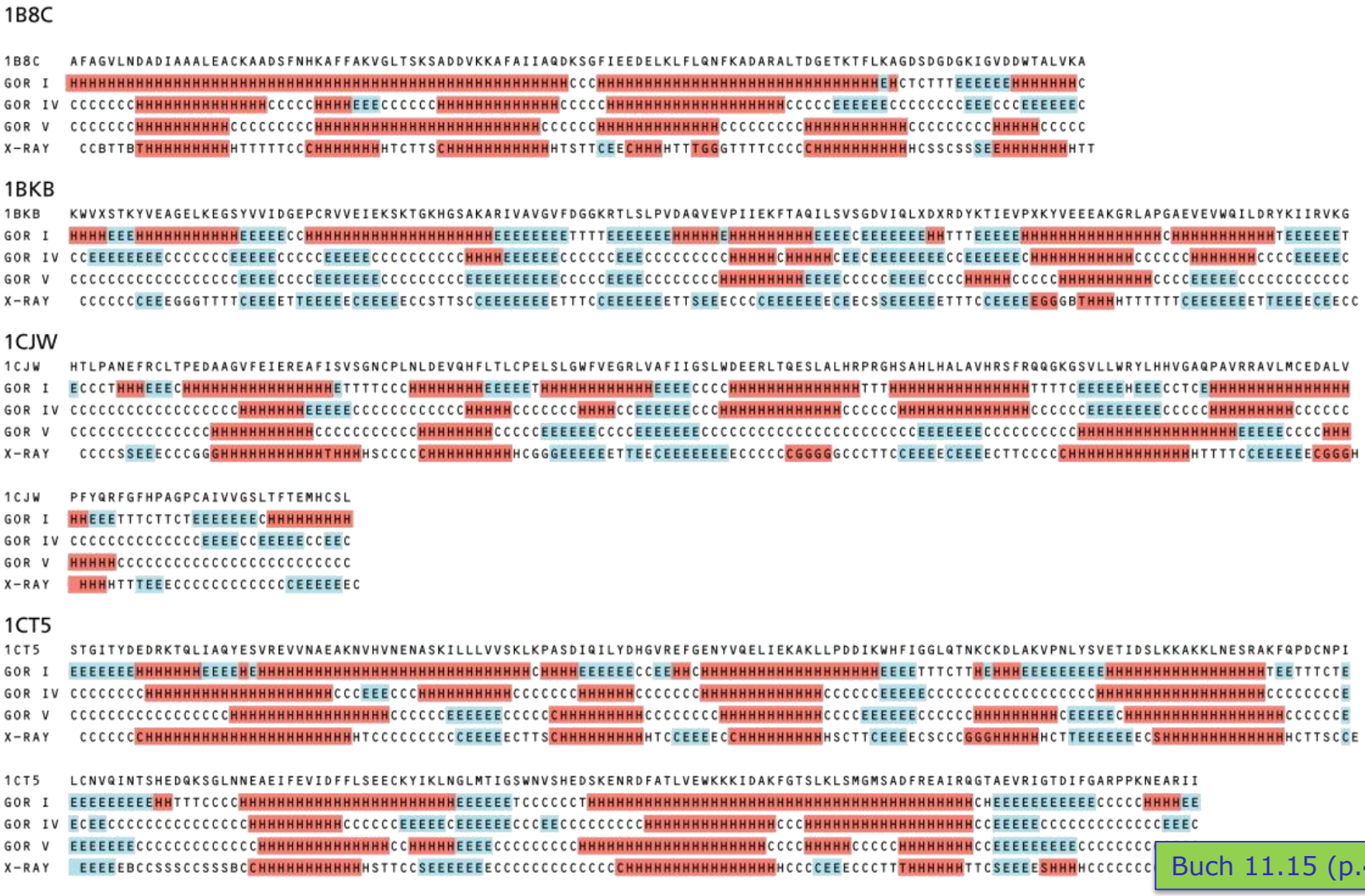
[a]Note that the convention used is the reverse of that adopted by (Garnier et al., 1978), for example the first entry for alanine at position j-8 is the amount of information that an alanine residue eight positions toward the N terminus has for predicting an α-helix

Table 4. *Directional informational parameters for residue position versus residue type for β-strands*

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-8	-7	-13	-17	-23	-33	-26	-32	-43	-37	-30	-30	-26	-27	-26	-25	-25
c	3	13	-9	-20	-15	-3	9	33	47	51	21	19	9	-5	7	-5	-14
d	-7	-5	0	-9	-4	-14	-42	-73	-83	-59	-21	10	22	24	16	11	13
e	-14	-5	-5	-11	-21	-27	-45	-44	-57	-54	-46	-29	-25	-12	-12	-2	0
f	-9	-20	-32	-34	-30	-12	24	44	49	39	24	2	-9	-23	-24	-29	-23
g	-3	9	24	29	34	30	18	-23	-48	-27	6	27	39	38	33	23	23
h	6	11	17	22	12	16	0	-2	3	-2	5	3	8	4	-1	1	-3
i	-21	-30	-31	-21	-12	-3	26	58	76	64	33	11	-14	-24	-20	-14	-11
k	20	12	15	14	8	4	-8	-14	-25	-40	-39	-27	-20	-24	-20	-15	-15
l	-2	-10	-18	-27	-30	-27	-6	15	27	21	2	-19	-31	-29	-28	-26	-25
m	-22	-26	-29	-40	-31	-17	-7	23	24	28	17	2	-15	-31	-53	-36	-16
n	1	8	14	5	0	-6	-30	-65	-62	-28	-6	11	18	21	16	10	3
p	9	7	12	24	20	8	-22	-65	-108	-64	-8	17	25	30	32	31	21
q	6	12	8	16	8	-5	-22	-27	-30	-52	-49	-34	-22	-17	-9	2	20
r	0	8	3	-3	5	2	1	-14	-26	-32	-30	-35	-27	-26	-25	-25	-21
s	16	14	17	19	14	5	-3	-13	-15	-4	15	27	32	32	31	28	21
t	6	8	14	15	16	21	19	25	31	22	13	9	12	25	34	34	34
v	1	-11	-15	-11	4	25	51	75	91	81	49	19	-6	-12	-16	-11	-11
w	-8	-8	-28	-19	-9	5	23	44	45	30	13	-18	-22	-40	-15	-7	-9
y	13	13	4	14	12	20	24	37	48	31	20	-1	2	11	7	0	-4

Credits: King & Sternberg, 1996

Evolution of GOR: I -> V



Buch 11.15 (p.426)

Mehr Informationen im Internet unter

medicalbioinformatics.de/teaching

Vielen Dank!

Tim Conrad

AG Medical Bioinformatics

www.medicalbioinformatics.de