Master Course

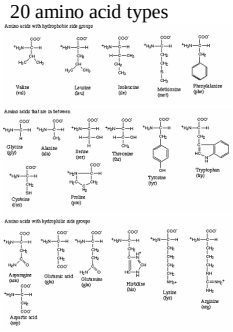
DNA/Protein Structure-function Analysis and Prediction

Lecture 7

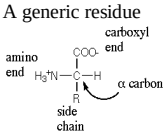
Protein Secondary Structure Prediction

Protein primary structure

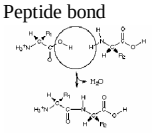
20 amino acid types



A generic residue



Peptide bond



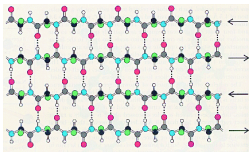
SARS Protein From Staphylococcus Aureus

```
  1 MKYNNHDKIR DFIIIEAYMF RFKKKVKPEV DMTIKEFILL TYLFHQQENT LPFKKIVSDL
 61 CYKQSDLVQH IKVLVKHSYI SKVRSKIDER NTYISISEEQ REKIAERVTL FDQIIKQFNL
121 ADQSESQMIP KDSKEFLNLM MYTMYFKNII KKHLTLSFVE FTILAIITSQ NKNIVLLKDL
181 IETIHHKYPQ TVRALNNLKK QGYLIKERST EDERKILIHM DDAQQDHAEQ LLAQVNQLLA
241 DKDHLHLVFE
```

Protein secondary structure

Alpha-helix

Beta strands/sheet



SARS Protein From Staphylococcus Aureus

```
  1 MKYNNHDKIR DFIIIEAYMF RFKKKVKPEV DMTIKEFILL TYLFHQQENT
        SHHH HHHHHHHHHH HHHHHHTTT  SS HHHHHHH HHHHS S SE
 51 LPFKKIVSDL CYKQSDLVQH IKVLVKHSYI SKVRSKIDER NTYISISEEQ
    EEHHHHHHHS SS GGGTHHH HHHHHHTTS  EEEE SSSTT EEEE HHH
101 REKIAERVTL FDQIIKQFNL ADQSESQMIP KDSKEFLNLM MYTMYFKNII
    HHHHHHHHHH HHHHHHHHHH HTT SS S   SHHHHHHHH  HHHHHHHHHH
151 KKHLTLSFVE FTILAIITSQ NKNIVLLKDL IETIHHKYPQ TVRALNNLKK
    HHH SS HHH HHHHHHHHTT TT EEHHHH  HHHSSS HHH HHHHHHHHHH
201 QGYLIKERST EDERKILIHM DDAQQDHAEQ LLAQVNQLLA DKDHLHLVFE
    HTSSEEEE S SSTT EEEE  HHHHHHHHH  HHHHHHHHTS SS TT SS
```

Secondary Structure

- An easier question – what is the secondary structure when the 3D structure is known?

DSSP

- **DSSP** (Dictionary of Secondary Structure of a Protein) – ***assigns*** **secondary structure** to proteins which have a crystal (x-ray) or NMR (Nuclear Magnetic Resonance) structure

DSSP uses hydrogen-bonding structure to assign Secondary Structure Elements (SSEs). The method is strict but consistent (as opposed to expert assignments in PDB

H = alpha helix
B = beta bridge (isolated residue)
E = extended beta strand
G = 3-turn (3/10) helix
I = 5-turn (π) helix
T = hydrogen bonded turn
S = bend

A more challenging task:
Predicting secondary structure from primary sequence alone

Improvements in the 1990's

- Conservation in MSA
- Smarter algorithms (e.g. HMM, neural networks).

Accuracy

- Accuracy of prediction seems to hit a ceiling of 70-80% accuracy
  - Long-range interactions are not included
  - Beta-strand prediction is difficult

Method	Accuracy
Chou & Fasman	50%
Adding the MSA	69%
MSA+ sophisticated computations	70-80%

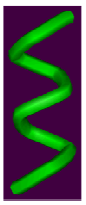
Secondary Structure Method Improvements

- 'Sliding window' approach
- Most alpha helices are ~12 residues long
  Most beta strands are ~6 residues long
- Look at all windows of size 6/12
- Calculate a score for each window.
- If >threshold
  → predict this is an alpha helix/beta sheet
  → otherwise this is coil

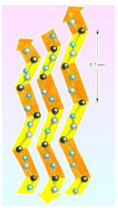
TGTAGPOLKCHIQWMLPLKK

Secondary Structure

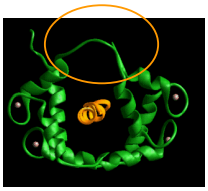
- Reminder-
  secondary structure is usually divided into three categories:



Alpha helix



Beta strand (sheet)



Anything else – turn/loop

Protein secondary structure

Why bother predicting them?

SS Information can be used for downstream analysis:

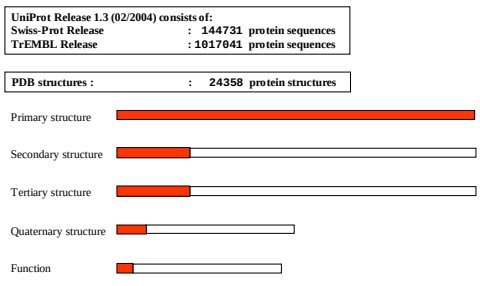
• Framework model of protein folding, collapse secondary structures

• Fold prediction by comparing to database of known structures

• Can be used as information to predict function

• Can also be used to help align sequences (e.g. SS-Praline)

Why predict when you can have the real thing?



Mind the (sequence-structure) gap!

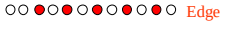
What we need to do

1) Train a method on a *diverse* set of proteins of known structure
2) Test the method on a test set separate from our training set
3) Assess our results in a useful way against a standard of truth
4) Compare to already existing methods using the same assessment

Some key features

ALPHA-HELIX: Hydrophobic-hydrophilic residue periodicity patterns

BETA-STRAND: Edge and buried strands, hydrophobic-hydrophilic residue periodicity patterns

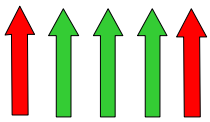


OTHER: Loop regions contain a high proportion of small polar residues like alanine, glycine, serine and threonine.

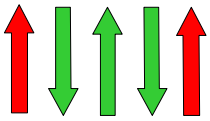
The abundance of glycine is due to its flexibility and proline for entropic reasons relating to the observed rigidity in its kinking the main-chain.

As proline residues kink the main-chain in an incompatible way for helices and strands, they are normally not observed in these two structures (breakers), although they can occur in the N-terminal two positions of α-helices.

Burried and Edge strands



Parallel β-sheet



Anti-parallel β-sheet

History (1)

Using computers in predicting protein secondary has its onset 30 ago (Nagano (**1973**) *J. Mol. Biol.*, *75*, 401) on single sequences.

The accuracy of the computational methods devised early-on was in the range 50-56% (Q3). The highest accuracy was achieved by Lim with a Q3 of 56% (Lim, V. I. (**1974**) *J. Mol. Biol., 88*, 857). The most widely used method was that of Chou-Fasman (Chou, P. Y. , Fasman, G. D. (**1974**) *Biochemistry*, *13*, 211).

Random prediction would yield about 40% (Q3) correctness given the observed distribution of the three states H, E and C in globular proteins (with generally about 30% helix, 20% strand and 50% coil).

History (2)

Nagano 1973 – Interactions of residues in a window of ≤6. The interactions were linearly combined to calculate interacting residue propensities for each SSE type (H, E or C) over 95 crystallographically determined protein tertiary structures.

Lim 1974 – Predictions are based on a set of complicated stereochemical prediction rules for α–helices and β–sheets based on their observed frequencies in globular proteins.

Chou-Fasman 1974 - Predictions are based on differences in residue type composition for three states of secondary structure: α–helix, β–strand and turn (i.e., neither α–helix nor β–strand). Neighbouring residues were checked for helices and strands and predicted types were selected according to the higher scoring preference and extended as long as unobserved residues were not detected (e.g. proline) and the scores remained high.

Chou and Fasman (1974)

The propensity of an amino acid to be part of a certain secondary structure (e.g. – Proline has a **low** propensity of being in an alpha helix or beta sheet → **breaker**)

Name	P(a)	P(b)	P(turn)
Alanine	142	83	66
Arginine	98	93	95
Aspartic Acid	101	54	146
Asparagine	67	89	156
Cysteine	70	119	119
Glutamic Acid	151	037	74
Glutamine	111	110	98
Glycine	57	75	156
Histidine	100	87	95
Isoleucine	108	160	47
Leucine	121	130	59
Lysine	114	74	101
Methionine	145	105	60
Phenylalanine	113	138	60
Proline	57	55	152
Serine	77	75	143
Threonine	83	119	96
Tryptophan	108	137	96
Tyrosine	69	147	114
Valine	106	170	50

Chou-Fasman prediction

- Look for a series of >4 amino acids which all have (for instance) alpha helix values >100
- Extend (…)
- Accept as alpha helix if
  average alpha score > average beta score

	Ala	Pro	Tyr	Phe	Phe	Lys	Lys	His	Val	Ala	Thr
α	142	57	69	113	113	114	114	100	106	142	83
β	83	55	147	138	138	74	74	87	170	83	119

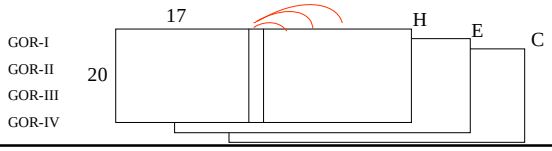
Chou and Fasman (1974)

- Success rate of 50%

GOR: the older standard

The GOR method (version IV) was reported by the authors to perform single sequence prediction accuracy with an accuracy of 64.4% as assessed through *jackknife* testing over a database of 267 proteins with known structure. (Garnier, J. G., Gibrat, J.-F., , Robson, B. (**1996**) In: *Methods in Enzymology* (Doolittle, R. F., Ed.) Vol. 266, pp. 540-53.)

The GOR method relies on the frequencies observed in the database for residues in a 17- residue window (i.e. eight residues N-terminal and eight C-terminal of the central window position) for each of the three structural states.

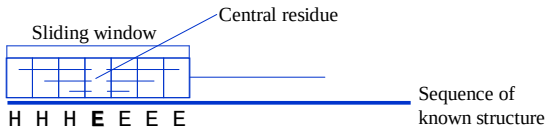


How do secondary structure prediction methods work?

•They often use a window approach to include a local stretch of amino acids around a considered sequence position in predicting the secondary structure state of that position

•The next slides provide basic explanations of the window approach (for the GOR method as an example) and two basic techniques to train a method and predict SSEs: *k-nearest neighbour* and *neural nets*

Sliding window



A constant window of *n* residues long slides along sequence

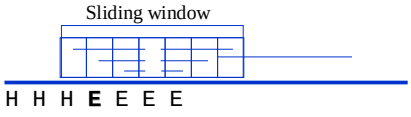
•The frequencies of the residues in the window are converted to probabilities of observing a SS type
•The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types
•At each position, the highest probability (helix, strand or coil) is taken.

Sliding window

A constant window of *n* residues long slides along sequence

•The frequencies of the residues in the window are converted to probabilities of observing a SS type
•The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types
•At each position, the highest probability (helix, strand or coil) is taken.

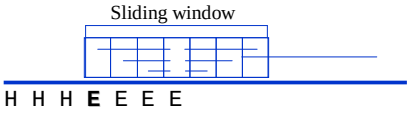
Sliding window



A constant window of *n* residues long slides along sequence

- The frequencies of the residues in the window are converted to probabilities of observing a SS type
- The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types
- At each position, the highest probability (helix, strand or coil) is taken.

Sliding window

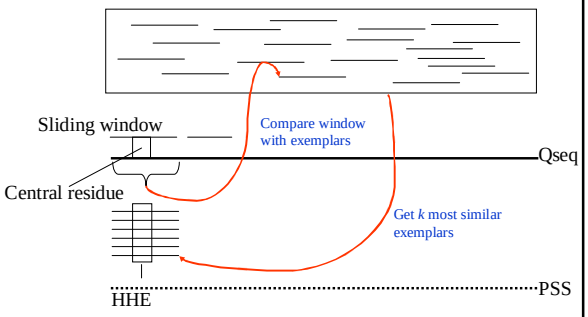


A constant window of *n* residues long slides along sequence

- The frequencies of the residues in the window are converted to probabilities of observing a SS type
- The GOR method uses three 17*20 windows for predicting helix, strand and coil; where 17 is the window length and 20 the number of a.a. types
- At each position, the highest probability (helix, strand or coil) is taken.

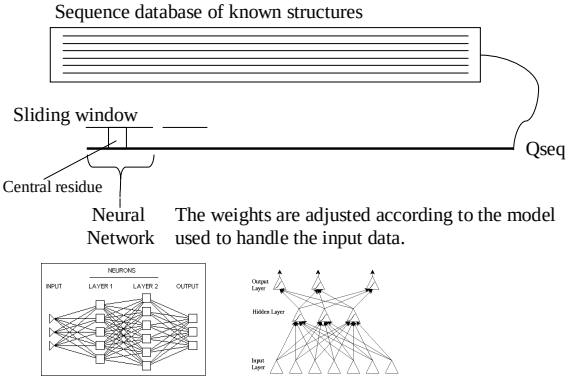
K-nearest neighbour

Sequence fragments from database of known structures (exemplars)



Neural nets

Sequence database of known structures



Neural nets

Training an NN:

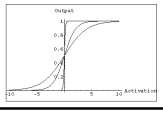
Forward pass:

the outputs are calculated and the error at the output units calculated.

Backward pass:

The output unit error is used to alter weights on the output units. Then the error at the hidden nodes is calculated (by *back-propagating* the error at the output units through the weights), and the weights on the hidden nodes altered using these values.

For each data pair to be learned a forward pass and backwards pass is performed. This is repeated over and over again until the error is at a low enough level (or we give up).



$Y = 1 / (1+ \exp(-k.(\Sigma W_{in} * X_{in}))$, where W_{in} is weight and X_{in} is input

The graph shows the output for *k=0.5, 1, and 10*, as the activation varies from -10 to 10.

Example of widely used neural net method: PHD, PHDpsi, PROFsec

The three above names refer to the same basic technique and come from the same laboratory (Rost's lab at Columbia, NYC)

Three neural networks:

1) A 13 residue window slides over the multiple alignment and produces 3-state raw secondary structure predictions.
2) A 17-residue window filters the output of network 1. The output of the second network then comprises for each alignment position three adjusted state probabilities. This post-processing step for the raw predictions of the first network is aimed at correcting unfeasible predictions and would, for example, change (HHH**E**EHH) into (HHH**HH**HH).
3) A network for a so-called jury decision over a set of independently trained networks 1 and 2 (extra predictions to correct for training biases).
4) The predictions obtained by the jury network undergo a final simple filtering step to delete predicted helices of one or two residues and changing those into coil.

Multiple Sequence Alignments are the superior input to a secondary structure prediction method

Multiple sequence alignment: three or more sequences that are aligned so that overall the greatest number of similar characters are matched in the same column of the alignment.

Enables detection of:

- Regions of high mutation rates over evolutionary time.
- Evolutionary conservation.
- Regions or domains that are critical to functionality.
- Sequence changes that cause a change in functionality.

Modern SS prediction methods all use Multiple Sequence Alignments (compared to single sequence prediction >10% better)

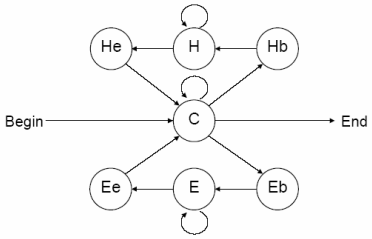
Rules of thumb when looking at a multiple alignment (MA)

- Hydrophobic residues are internal
- Gly (Thr, Ser) in loops
- MA: hydrophobic block -> internal β-strand
- MA: alternating (1-1) hydrophobic/hydrophilic => edge β-strand
- MA: alternating 2-2 (or 3-1) periodicity => α-helix
- MA: gaps in loops
- MA: Conserved column => functional? => active site

Rules of thumb when looking at a multiple alignment (MA)

- Active site residues are together in 3D structure
- MA: 'inconsistent' alignment columns and alignment match errors!
- Helices often cover up core of strands
- Helices less extended than strands => more residues to cross protein
- β-α-β motif is right-handed in >95% of cases (with parallel strands)
- Secondary structures have local anomalies, e.g. β-bulges

How to optimise?
Differentiate along SSEs – The Yaspin method (Lin et al., 2005)



Helices and strands are dissected in (begin, middle, end) sections. The Yaspin method then tries to regognise these sections.

Lin K., Simossis V.A., Taylor W.R. and Heringa J. (2005) A simple and fast secondary structure prediction algorithm using hidden neural networks. Bioinformatics. 21(2):152-9.

How to optimise?
Capture long-range interactions (Important for β-strand prediction)

- Predator (Frishman and Argos, 1995)
  - side-chains show subtle patterns in cross-strand contacts

- SSPro (Polastri et al., 2002) – uses bidirectional recurrent neural networks
  - One basic sliding window is used, with two more windows that slight in from opposite sites at each basic window position. This way all-possible long-range interactions are checked.

A stepwise hierarchy

1) Sequence database searching
   - PSI-BLAST, SAM-T2K

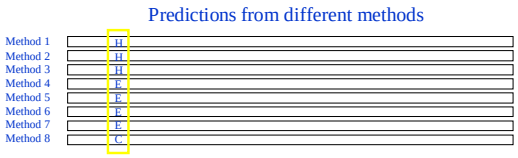
These basically are local alignment techniques to collect homologous sequences from a database so a multiple alignment containing the query sequence can be made

2) Multiple sequence alignment of selected sequences
   - PSSMs, HMM models, MSAs

3) Secondary structure prediction of query sequences based on the generated MSAs
   - Single methods: PHD, PROFsec, PSIPred, SSPro, JNET (consensus method), YASPIN
   - consensus methods using best methods
     - Issue is accuracy and correlation of methods used in consensus

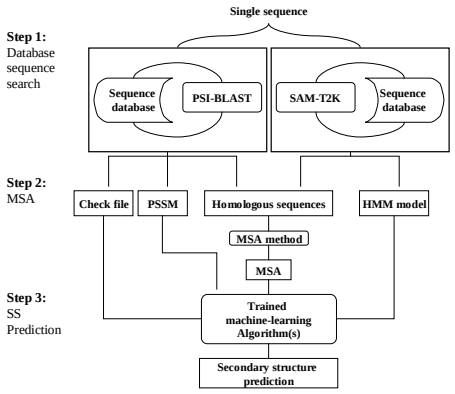
Single vs. Consensus predictions

The current standard ~1% better on average



Listen most to the most accurate method?
Accuracy and correlation of methods used in consensus

The current picture



Jackknife test

A **jackknife test** is a test scenario for prediction methods that need to be tuned using a ***training* database**.

In its simplest form:

For a database containing *N* sequences with known tertiary (and hence secondary) structure, a prediction is made for one test sequence after training the method on a training database containing the *N*-1 remaining sequences (one-at-a-time jackknife testing).

A complete jackknife test involves *N* such predictions, after which for all sequences a prediction is made.

If *N* is large enough, meaningful statistics can be derived from the observed performance. For example, the mean prediction accuracy and associated standard deviation give a good indication of the sustained performance of the method tested.

If the jackknife test is computationally too expensive, the database can be split in larger groups, which are then jackknifed. The latter is called **Cross-validation**

Cross validation

To save on computation time relative to the Jackknife, the database is split up in a number of non-overlapping sub-databases.

For example, with 10-fold cross-validation, the database is divided into 10 equally (or near equally) sized groups. One group is then taken out of the database as a test set, the method trained on the remaining nine groups, after which predictions are made for the sequences in the test group and the predictions assessed.

The amount of training required is now only 10% of what would be needed with jackknife testing.

Standards of truth

What is a standard of truth?

- a structurally derived secondary structure assignment (using a 3D structure from the PDB)

Why do we need one?

- it dictates how accurate our prediction is

How do we get it?

- methods use hydrogen-bonding patterns along the main-chain or knowledge-based approaches to assign the Secondary Structure Elements (SSEs) in experimentally solved tertiary structures.

Some examples of programs that assign secondary structures in 3D structures

1) DSSP (Kabsch and Sander, 1983) – most popular
2) STRIDE (Frishman and Argos, 1995)
3) DEFINE (Richards and Kundrot, 1988)

Annotation:

Helix: 3/10-helix (G), α-helix (H), π-helix (I)

Strand: β-strand (E), β-bulge (B)

Turn: H-bonded turn (T), bend (S)

Rest: Coil (“ “)

Assessing a prediction

How do we decide how good a prediction is?

1. Q_n : the number of correctly predicted *n* SSE states over the total number of predicted states
   $Q3 = [(PH + PE + PC)/N] \times 100\%$
2. Segment OVerlap (SOV): the number of correctly predicted *n* SSE states over the total number of predictions with higher penalties for core segment regions (Zemla *et al*, 1999)

Assessing a prediction

How do we decide how good a prediction is?

3. Matthews Correlation Coefficients (MCC): the number of correctly predicted *n* SSE states over the total number of predictions taking into account how many prediction errors were made for each state:

$$C_S = \frac{(P_S \times N_S) - (\tilde{P}_S \times \tilde{N}_S)}{\sqrt{(P_S + \tilde{P}_S) \times (P_S + \tilde{N}_S) \times (N_S + \tilde{P}_S) \times (N_S + \tilde{N}_S)}},$$

$\tilde{P}$ = false positive

Some Servers

- PSI-pred uses PSI-BLAST profiles
- JPRED Consensus prediction
- PHD home page – all-in-one prediction, includes secondary structure
- nnPredict – uses neural networks
- BMERC PSA Server
- IBIVU YASPIN server
- BMC launcher – choose your prediction program