# LECTURE 09: PREDICTION OF THREE DIMENSIONAL STRUCTURES

## PREDICTION OF THREE DIMENSIONAL STRUCTURES FROM PRIMARY STRUCTURES

Proteins are one of the major biological macromolecules performing a variety functions such as enzymatic catalysis, transport, regulation of metabolism, nerve conduction, immune response etc. The three-dimensional structure of a protein is responsible for its function. In this an overview of the need for protein structure prediction, the different approaches available as of now and their applications and limitations will be discussed.

### Sequence-Structure Gap and the Need for Structure Prediction

With the advent of recombinant DNA technology it has become possible to determine the amino acid sequences of proteins quite rapidly. However, determining the three dimensional structure of proteins is a time consuming task and hence there exists a vast gap between the number of proteins of known amino acid sequence and that of known structures. This is called as the sequence-structure gap. As the knowledge of the 3-D structure of a protein is very essential to understand its function, it is imperative to develop techniques to predict the structure of a protein from its amino acid sequence.

### Basis for Structure Prediction:

The classic experiments carried out by C.B. Afinson in the 60's on the enzyme ribonuclease led to the conclusion that the information to specify the 3-D structure of a protein resides in its amino acid sequence. Within the cell a newly synthesized protein chain spontaneously folds into the compact globular structure to perform its function. Thus nature has an algorithm to fold proteins to their native structures. Efforts have been directed for the past four decades to discover nature's algorithm and computational methods have been developed to predict the structure of proteins from their sequences.

# PROTEIN ENGINEERING

## Approaches to Structure Prediction

Prediction of protein structures can be classified into two major categories viz.

(i) Prediction of secondary structure and

(ii) Prediction of tertiary (3-D) structure.

Prediction of secondary structure of proteins attempts to locate segments of the polypeptide chain adopting the α-helical or β-strand structure. Regions that are devoid of these regular secondary structural elements are considered to adopt coil conformation.

In tertiary structure prediction, one attempts to predict the three-dimensional structure of a protein or the native structure. While so far this has remained an elusive goal, different methods have been developed to press forward to the attainment of this goal.

## Secondary structure prediction

### What?

- Given a protein sequence (primary structure)

  GHWIATRGQLIREAYEDYRHFSSECPFIP

- 1 st step in prediction of protein structure.

  (C=Coils H=Alpha Helix E=Beta Strands)

  CEEEEECHHHHHHHHHHHCCCHHCCCCCC

- Technique concerned with determination of secondary structure of given polypeptide by locating the Coils Alpha Helix Beta Strands in polypeptide

# PROTEIN ENGINEERING

**Why?**

- secondary structure —tertiary structure prediction
- Protein function prediction
- Protein classification
- Predicting structural change
- detection and alignment of remote homology between proteins
- on detecting transmembrane regions, solvent-accessible residues, and other important features of molecules
- Detection of hydrophobic region and hydrophilic region

**Prediction methods**

**Chou-Fasman method**

• Based on the propensities of different amino acids to adopt different secondary structures

• Predictions are made using a rules-based approach to identify groups of amino acids with shared secondary structure propensities

**Garnier, Osguthorpe, Robson (GOR) method**

• Statistical method of secondary structure prediction based on information theory & Bayesian probability

**Multiple Sequence Alignment (MSA) methods**

• Performs secondary structure prediction on a multiple sequence alignment as opposed to a single protein sequence

**Neural network-based methods**

• Example: **P**rofile network from **Hei**d**e**lberg (PHD)

**Chou-Fasman method**,

**1. Alpha Helix Prediction:**

A. Nucleate a helix by scanning for groups of 6 residues with at least 4 helix formers (Hα and hα) and no more than 1 helix breaker (Bα and bα).

• Two Iα residues count as one helix former for nucleating a helix

B. Propagate predicted helix in both directions until reach a four residue window with average propensity (Pα) < 1.0

C. The average propensity (Pα) for a predicted helix must be Pα > 1.03 and Pα > Pβ

**2. Beta Strand Prediction:**

A. Nucleate a β-strand by scanning for groups of 5 residues with at least 3 strand formers (Hβ and hβ) and no more than 1 strand breaker (B$ and b$).

B. Propagate predicted β-strand in both directions until reach a four residue window with average propensity (Pβ) < 1.0

C. The average propensity (Pβ) for a predicted β-strand must be Pβ > 1.05 and Pβ > Pα

**3. Resolving conflicting predictions:**

(regions with both α-helix and β-strand assignment)

• If average Pα > average Pβ, then the region is alpha helix

• If average Pβ > average Pα, then the region is beta strand

§ Notes about Chou-Fasman algorithm:

• Later versions of the algorithm included predictions for turns

• The original algorithm contained additional rules about the location of certain residues (e.g., proline) in α-helices and β-strands

• More recent versions of the algorithm have used sequential tetrapeptide average propensities to predict secondary structure

• The propensity values have also been variously recalculated with larger protein data sets (original data sets based on 15 and 29 proteins)

§ Example of Chou-Fasman method:

Sequence: **MLNPKSYENAIQLGRCFTTHYA**

**alpha helix nucleation**

| M | L | N | P | K | S | Y | E | N | A | I | Q | L | G | R | C | F | T | T | H | Y | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| h | h | b | b | h | i | b | h | b | h | h | h | h | b | i | i | h | i | i | I | b | h |

• Has at least 4 helix formers
• Has no more than 1 helix breaker

• Note: Counts as 0.5 helix former

**propagating alpha helix**

Propagate helix in both directions until reach a four residue window with average propensity ($P_\alpha$) < 1.0

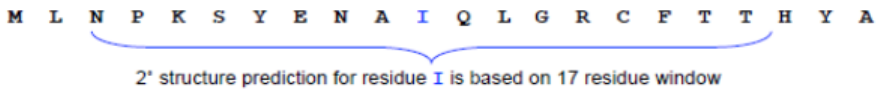
# PROTEIN ENGINEERING

## GOR (Garnier,Osguthorpe,Robson) Method

Key difference: Chou-Fasman uses individual amino acid propensities, while GOR incorporates information about neighboring amino acids to make prediction

A 20 x 17 matrix of directional information values for each secondary structure class was calculated from a database of known structures

These matrices are used to predict the secondary structure of the central (9th) residue in a 17 residue window:




The secondary structure class with highest information score over 17 residue window is selected as the prediction for the central residue of the window (e.g., I is predicted to be α-helix)

## Multiple sequence alignment method

A multiple sequence alignment arranges protein sequences into a rectangular array with the goal that residues in a given column are homologous (derived from a common ancestor), superposable (in a 3D structural alignment - α helix / β sheet) or play a common functional role (catalytic sites, nuclear localisation signal, protein-protein interaction sites,...). Uses BLAST to identify homologous protein sequence fragments in a protein structure database (PDB)

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTS-NIGS-ITVNWYQQLPG-
LRLS-CSVSGFIFSS-YAMYWVRQAPG
-LS-LTCTVSGTSFDDYYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFN-WYVDG-
A--TLVCTISDFYPGAVTVA-WKADS-
AALGCTVKDYFPEPVTVSWN--SG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

# PROTEIN ENGINEERING

Goal: try to have a maximum of identical/similar residues in a given column of the alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSVSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCTISDFYPGA--VTVAWKADS--
AALGCTVKDYFPEP--VTVSWNSG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

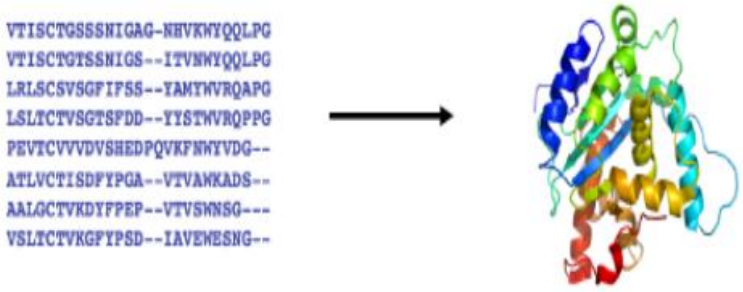
**Main Criteria for building a multiple sequence alignment**

| *Criterion* | *Meaning* |
|---|---|
| **Structure similarity** | **Amino acids that play the same role in each structure are in the same column.** Structure superposition programs are the only ones that use this criterion. |
| **Evolutionary similarity** | **Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column.** No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it. |
| **Functional similarity** | **Amino acids or nucleotides with the same function are in the same column.** No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually. |
| **Sequence similarity** | **Amino acids in the same column are those that yield an alignment with maximum similarity.** Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity. |

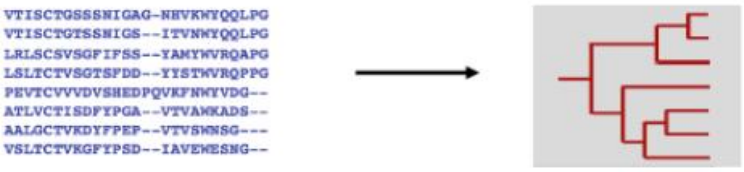
**PROTEIN ENGINEERING**

## What are the applications of multiple sequence alignment

§ Protein structure and function prediction




§ Phylogenetic inference




§ Detecting similarities between sequences (closely or distantly related) and conserved regions / motifs in sequences.

§ Detection of structural patterns (hydrophobicity/hydrophilicity, gaps etc), thus assisting improved prediction of secondary and tertiary structures and loops and variable regions.

§ Predict features of aligned sequences like conserved positions which may have structural or functional importance.

§ Computing consensus sequence.

§ Making patterns or profiles that can be further used to predict new sequences falling in a given family.

§ Deriving profiles or Hidden Markov Models that can be used to remove distant sequences (outliers) from protein families.

§ Inferring evolutionary trees / linkage.

## How is a multiple sequence alignment used?

```
chite     ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat     --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr     KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
unknown   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                ***. ::: .: ..  .    :  . .      *  .  *: *

chite     AATAKQNYIRALQEYERNGG-
wheat     ANKLKGEYNKAIAAYNKGESA
trybr     AEKDKERYKREM---------
unknown   AKDDRIRYDNEMKSWEEQMAE
          *   : .* . :
```

**Less Than 30 % id**
**BUT**
**Conserved where it matters!**

```
chite     ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat     --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr     KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
unknown   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                ***. ::: .: ..  .    :  . .      *  .  *: *

chite     AATAKQNYIRALQEYERNGG-
wheat     ANKLKGEYNKAIAAYNKGESA
trybr     AEKDKERYKREM---------
unknown   AKDDRIRYDNEMKSWEEQMAE
          *   : .* . :
```

**Conserved residues may be important for the function of the protein (catalytic site, etc).**

## How to score a multiple sequence alignment?

### The usual scoring method:

- assumes independance between the columns

$$S(m) = \sum_i S(m_i)$$

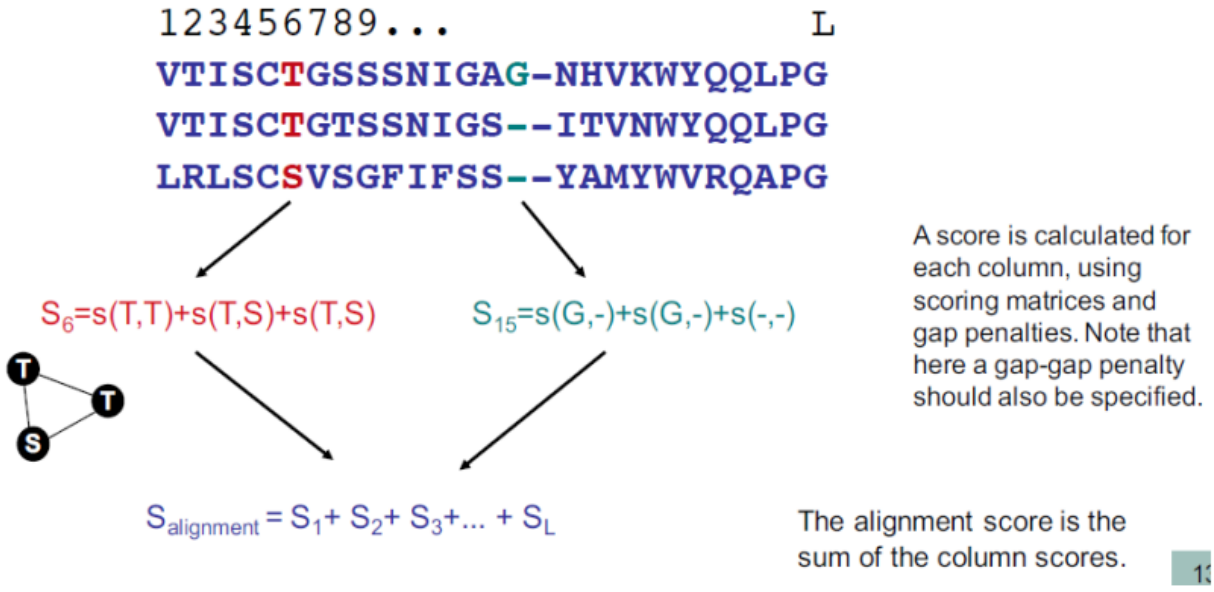
*S(m) = score of the whole alignment m*
*S(m$_i$) = score of column i in this alignmen*

- scores each column according a "sum-of-pairs" (SP) function using a substitution scoring matrix.

$$S(m_i) = \sum_{k<l} s\left(m_i^k, m_i^l\right)$$

*$m_i^k$ = residue in sequence k in column i*
*S(a,b) = score from a substitution matrix (PAM or BLOSUM for example)*

**Example**



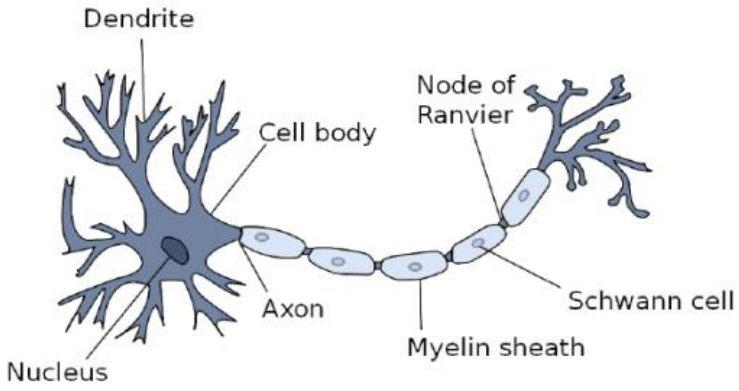

A score is calculated for each column, using scoring matrices and gap penalties. Note that here a gap-gap penalty should also be specified.

The alignment score is the sum of the column scores.

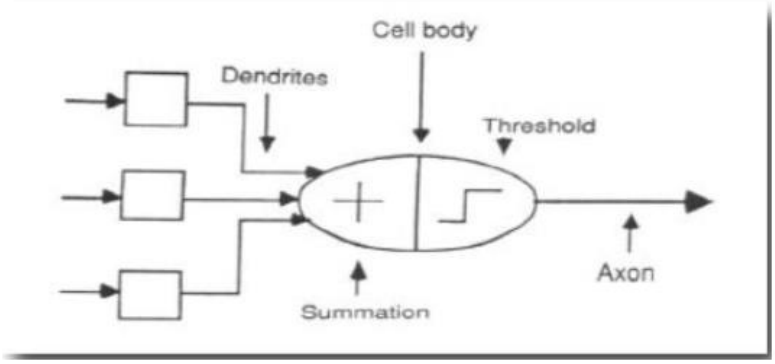
13

**Neural network secondary structure prediction methods**

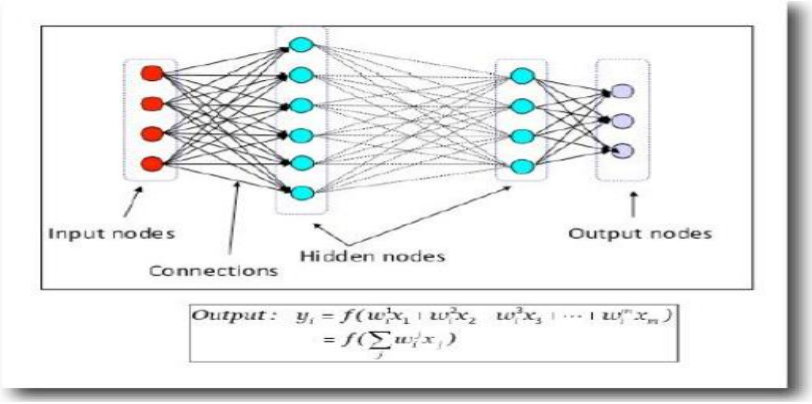
Artificial neural networks (ANN), with both statistical (linear regression and discriminant analysis) and artificial intelligence roots, are information processing units that that are modeled after the brain and its 100 billion neurons. In a neuron, the distal and proximal dendrites receive signals and communicate to the cell body, which in turn communicates with other neurons via its axon and its terminals.




Similarly, an ANN receives inputs (dendrites) that are processed with influence by weights to become outputs (axon).




The neurons or nodes interconnect with informational flows (unidirectional or bidirectional) at various weights or strengths. The simplest architecture is the perceptron, which consists of 2 layers (input and output layers) that are separated by a linear discrimination function (10). In a multi-layer perceptron (MLP) model, there are three layers: the input nodes, the hidden nodes layer, and the output nodes.



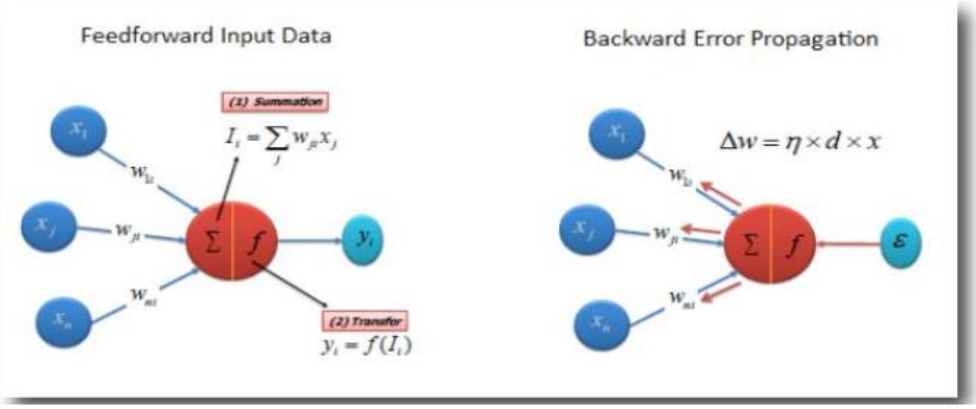
# PROTEIN ENGINEERING

Learning/ Training

In a feed-forward neural network architecture, a unit will receive input from several nodes or neurons belonging to another layer. These highly interconnected neurons therefore form an infrastructure (similar to the biological central nervous system) that is capable of learning by successfully perform pattern recognition and classification tasks. Training of the ANN is a process in which learning occurs from representative data and the knowledge is applied to the new situation.

This training or learning process occurs by arranging the algorithms so that the weights of the ANN are adjusted to lead to the final desired output. The learning in neural networks can be supervised (such as the multilayer perceptron that trained with sets of input data) or unsupervised (such as the Kohonen self-organizing maps which learn by finding patterns). Neural networks can also perform both regression and classification.

The ANN learning process consists of both a forward and a backward propagation process. The forward propagation process involves presenting data into the ANN whereas the important backward propagation algorithm determines the values of the weights for the nodes during a training phase. This latter process is accomplished by directing the errors for input values backwards so that corrections for the weights can be made to minimize the error of actual and desired output data. A recurrent neural network is a series of feed-forward neural networks sharing the same weights and is good for time series data. ANN can therefore extract patterns or detect trends from complicated and imprecise data sets.

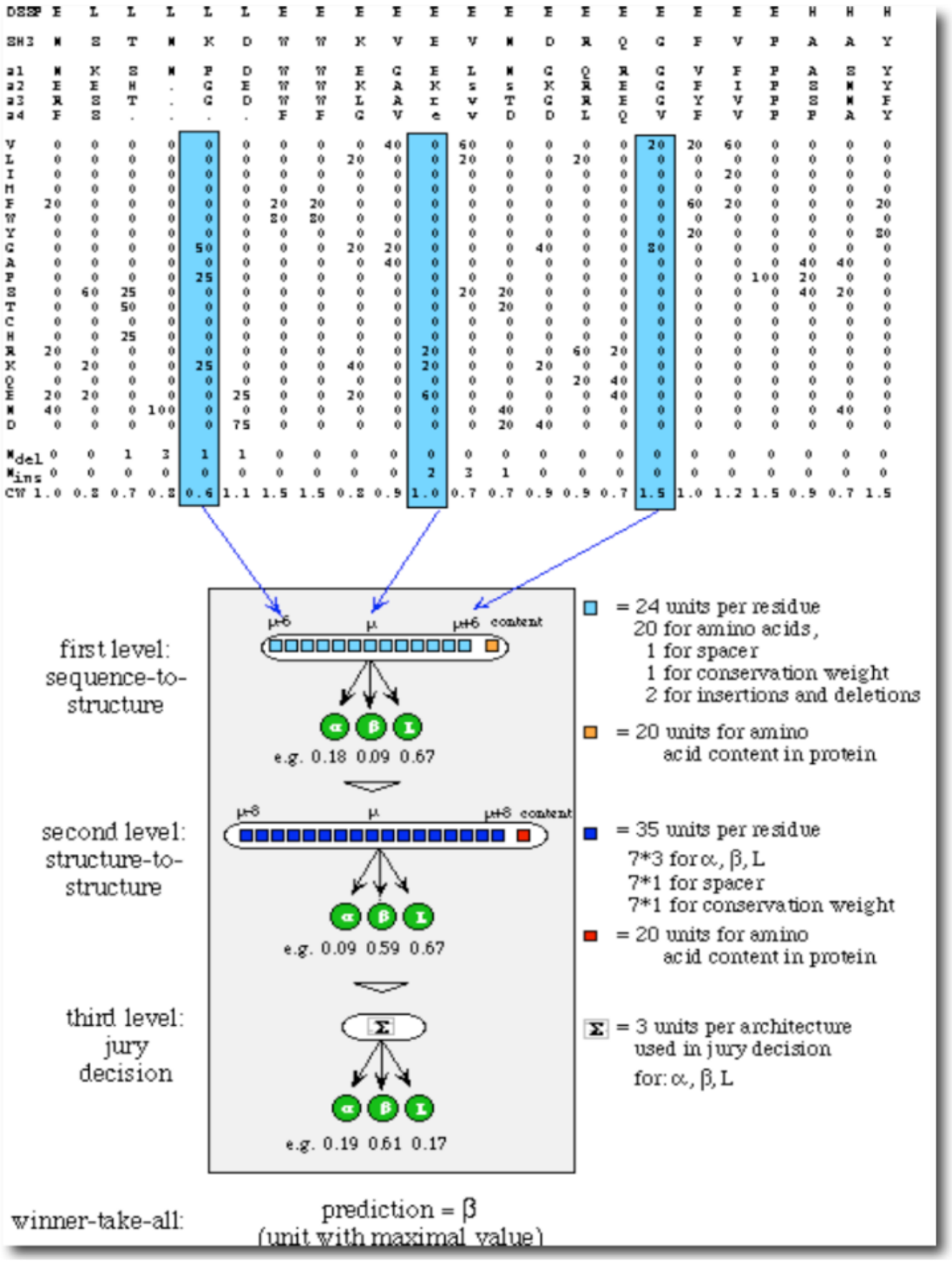



**Application of ANN to bioinformatics needs the following strategy:**

Extraction of features from molecular sequences to serve as training/prediction data; preprocessing that consists of feature selection and encoding into vectors of real numbers; neural network for training or prediction; postprocessing that consists of output encoding from the neural network; and finally the myriad of applications (such as sequence analysis, gene expression data analysis, or protein structure prediction).

In secondary structure prediction, neural network methods are trained using sequences with known secondary structure, and then asked to predict the secondary structure of proteins of unknown structure

§ Example: **P**rofile network from **H**ei**d**elberg (PHD) uses multiple sequence alignment with neural network methods to predict secondary structure




**Network architecture (PHD). A profile-based neural network system for protein secondary structure prediction. The multiple alignment is seen at the top with a profile of amino acid occurrences compiled. Then the alignment is fed into the neural network, which consists of 3 layers: 2 network layers and an additional layer for averaging over the independently trained networks**

# PROTEIN ENGINEERING

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) andbiotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

**Accuracy of Secondary Structure Prediction**

§ Prediction accuracy

- Accuracy is usually measured by Q3 (or Qindex) value
- For a single conformation state, i:

$$Q_i = \frac{\text{number of residues correctly predicted in state i}}{\text{number of residues observed in state i}} * 100\%$$

- where i is either helix, strand, or coil. For all three states:

$$Q_3 = \frac{\text{number of residues correctly predicted}}{\text{number of all residues}} * 100\%$$

§ Accuracy of prediction methods

- A random prediction has a Q3 value of ~ 33-38%
- Chou-Fasman method typically has a Q3 ~ 56-60%
- GOR method (depending upon version) has a Q3 ~ 60-65%

- MSA, neural network methods have Q3 ~70%