# Structure Prediction

Homayoun Valafar

Department of Computer Science and Engineering, USC

# Demonstration of AVOGADRO

# SCOP

- Structural Classification of Proteins (SCOP) utilizes four levels of hierarchical structural classification
  - Class: general "structural architecture" of the domain
  - Fold: similar arrangement of regular secondary structures but without evidence of evolutionary relationship
    - No significant sequence of functional similarity
  - Superfamily: proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable
    - No sequence similarity but functional similarity
  - Family: proteins that are clearly evolutionarily related (sequence identity > 30%)

# Structure Prediction Tools

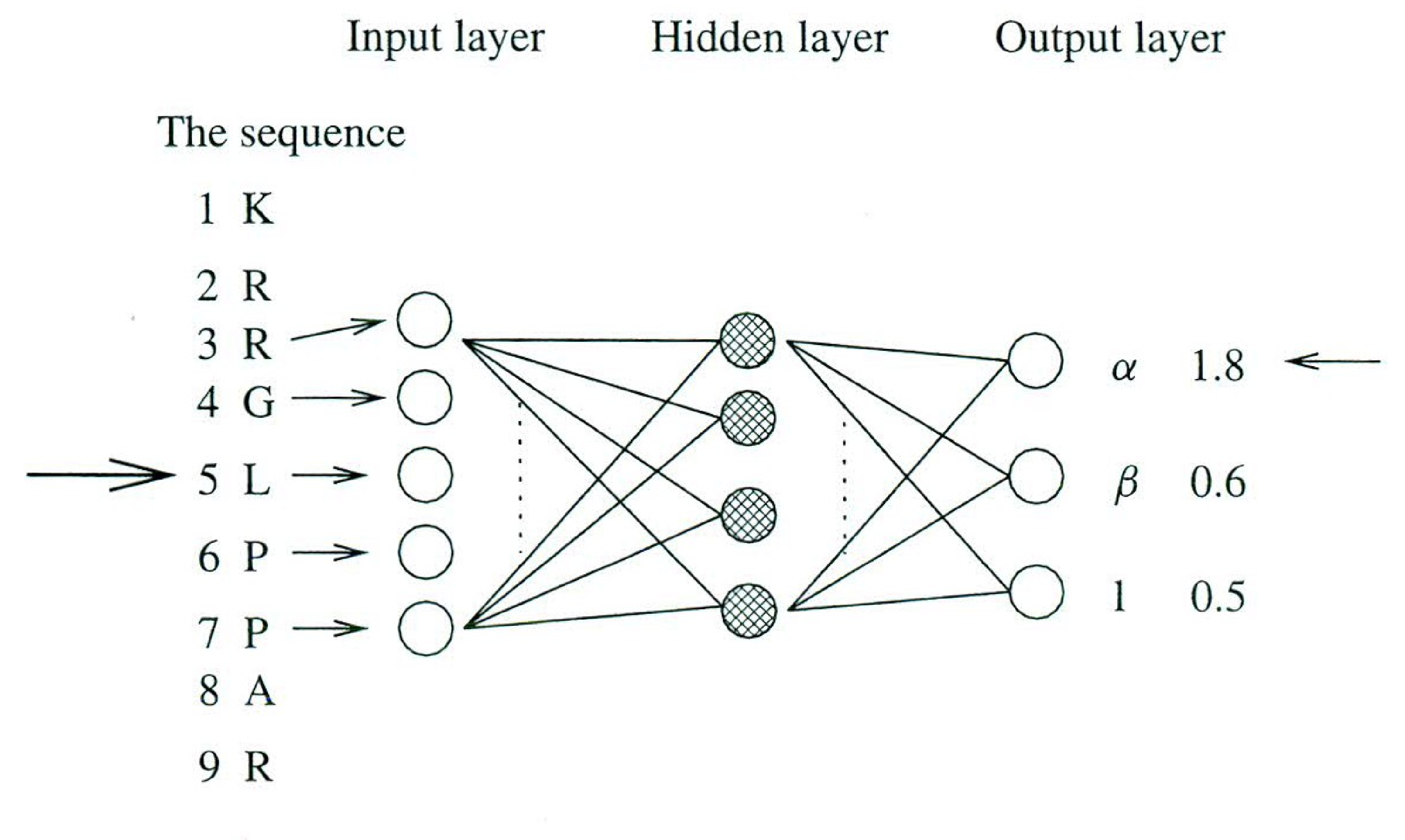
- Structure prediction tools aim to provide structural information at various levels
  - Secondary structure predictors
    - PhD
    - PredictProtein (PHDsec, PHDacc, …)
    - HNN (Hierarchical Neural Network method)
    - Jpred
  - Tertiary structure predictors
    - Modeller
    - Prospect
    - Rosetta
    - I-TASSER

# Secondary Structure Prediction Tools

- These programs accept a primary sequence and produce a predicted secondary structure per residue
- The most commonly known and used is PHD
- The results are usually in the form of a sequence of letters
  - H stands for Helical
  - E stands for Extended
  - L or T stand for Loop or Turn

```
AKYVCKICGYIYDEDAGDPDNGVSPGTKFEEIPDDWVCPICGAPKSEFEK
--EEE----EEE--------------------------------HHHH--
```

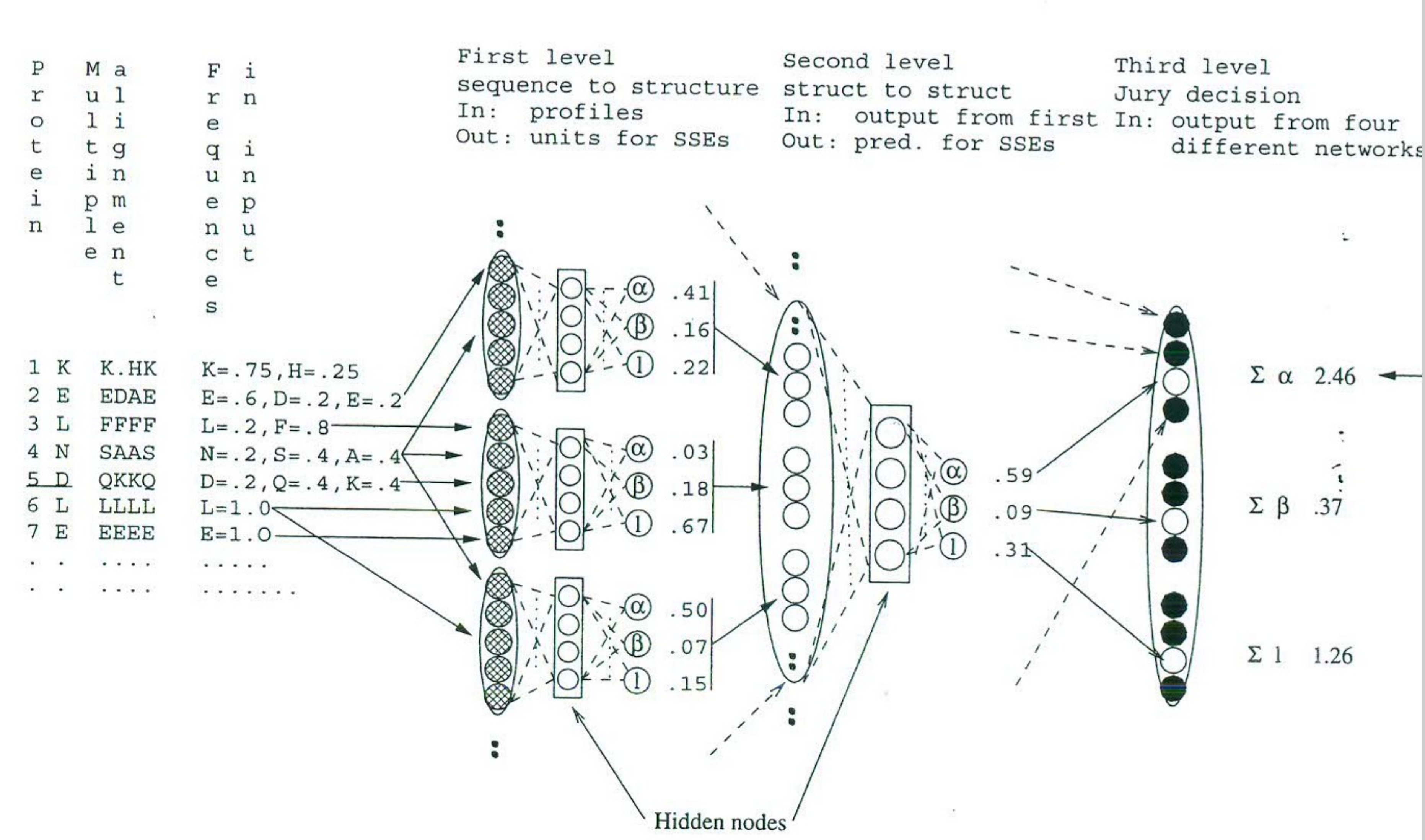
# Neural Network Based SS Prediction Tool



# PHD

- Sequence alignment is performed against the library of all proteins with known structures
- A multiple sequence alignment is generated
- Results of sequence alignment are presented multiple ANN
- ANNs analyze the structure in a sliding window
- Results from different ANNs are combined and SSE are predicted
- Refer to pages 298-301 of the handout

# PHD



| Protein | Multiple alignment | Frequencies in input |
|---|---|---|
| 1 K | K.HK | K=.75,H=.25 |
| 2 E | EDAE | E=.6,D=.2,E=.2 |
| 3 L | FFFF | L=.2,F=.8 |
| 4 N | SAAS | N=.2,S=.4,A=.4 |
| 5 D | QKKQ | D=.2,Q=.4,K=.4 |
| 6 L | LLLL | L=1.0 |
| 7 E | EEEE | E=1.O |
| . . | .... | ..... |
| . . | .... | ....... |

First level
sequence to structure
In: profiles
Out: units for SSEs

Second level
struct to struct
In: output from first
Out: pred. for SSEs

Third level
Jury decision
In: output from four
different networks

# SS Prediction Tool Demo

# Tertiary Structure Prediction Tools

- Over the past few years, a number different TSPT have been presented
- These tools present three distinct TSP strategies that are presented in the historical order and order of sophistication
  - Homology modeling tools
    - SWISS-MODEL
  - Threading tools
    - Threader
    - Prospect
  - Ab initio tools
    - Rosetta
    - I-TASSER

# Homology Modeling Tools

- HMT leverage strong sequence identity (seq identity > 30%)
- The unknown protein needs to exhibit more than 30% seq identity to another protein with known structure
- Structure of the homologue is used as a template to fold the unknown protein
- Can be applied to proteins in the same “Family”
- If a family of proteins is not structurally characterized, then HMT do not apply

# Limitations of Homology Modeling

- Homology modeling tools fail when
  - The target protein does not have a homologue in the PDB
  - Two structures can be very similar in structure but different in sequence
  - Homstrad database
  - Example: Peptidase M17

# Threading Tools

- Apply to proteins in the same SuperFamily
- There maybe low sequence identity but evolutionary relationship is evident
- Threading strategy of the following four steps:
  1. A database of structure templates (CATH, SCOP, FSSP)
  2. A scoring function to evaluate fitness of a sequence to structure
  3. Threading alignment to align the target sequence with each of the structure templates
  4. Select the most optimal threading alignment

# Limitations of Threading Tools

- Threading tools fails when the target protein is dissimilar in sequence and structure to what is in the PDB
- Can we still utilize the information in the PDB to construct a structure for a protein based on structures in Fold or Class level?

# Ab Initio Protein Structure Prediction Tools

- This class to tools may be successful with even little sequence similarity
- They may rely on a number of other tools such as SS prediction
- They may seed the structure based on small conserved sequences (13 residues long)
- General force terms are deployed to further refine the structure
    - VDW collision is not presented at the atomic level but at the residue level