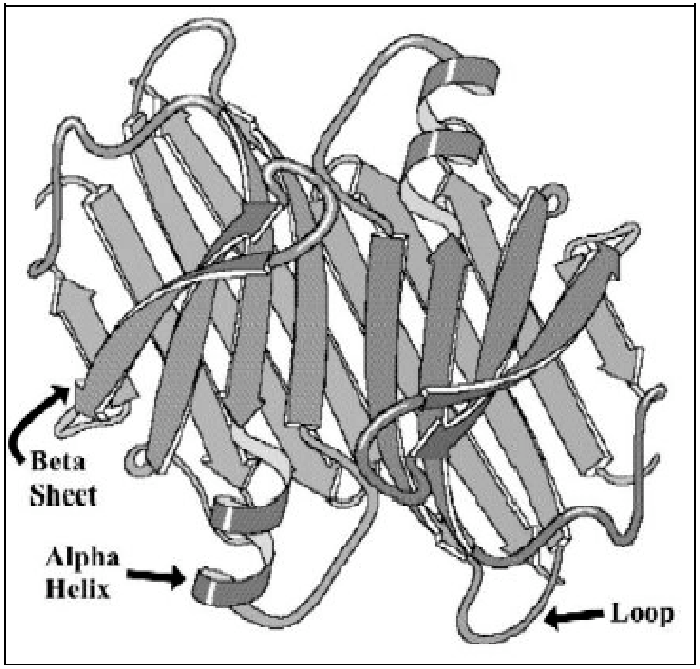
Protein Structure Prediction Using Neural Networks

Martha Mercaldi
Kasia Wilamowska

Literature Review
December 16, 2003

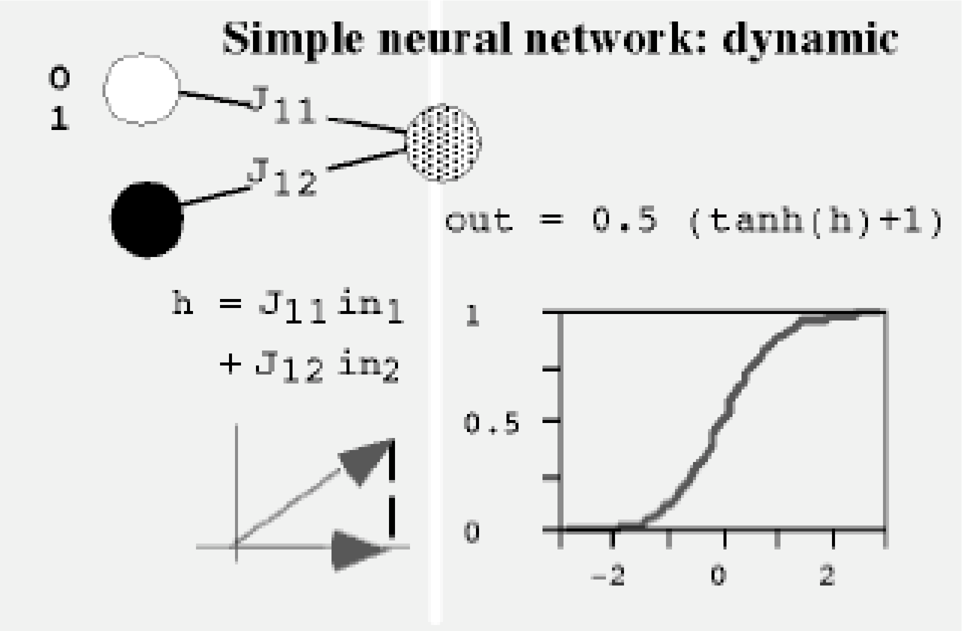
The Protein Folding Problem

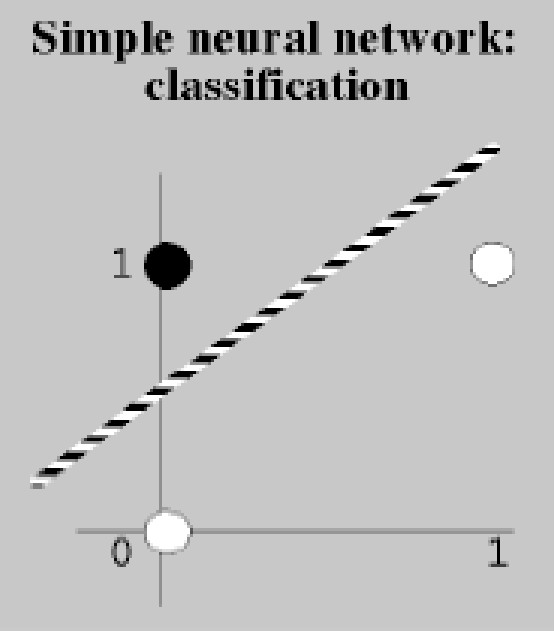
DNA Sequence	AGG**AAA**AGC**AGA**ATT**ACT**AAT**TAC**CCT								
	AGG	AAA	AGC	AGA	ATT	ACT	AAT	TAC	CCT
Amino Acid Sequence	R	K	S	R	I	T	N	Y	P
	RKSRITNYP								



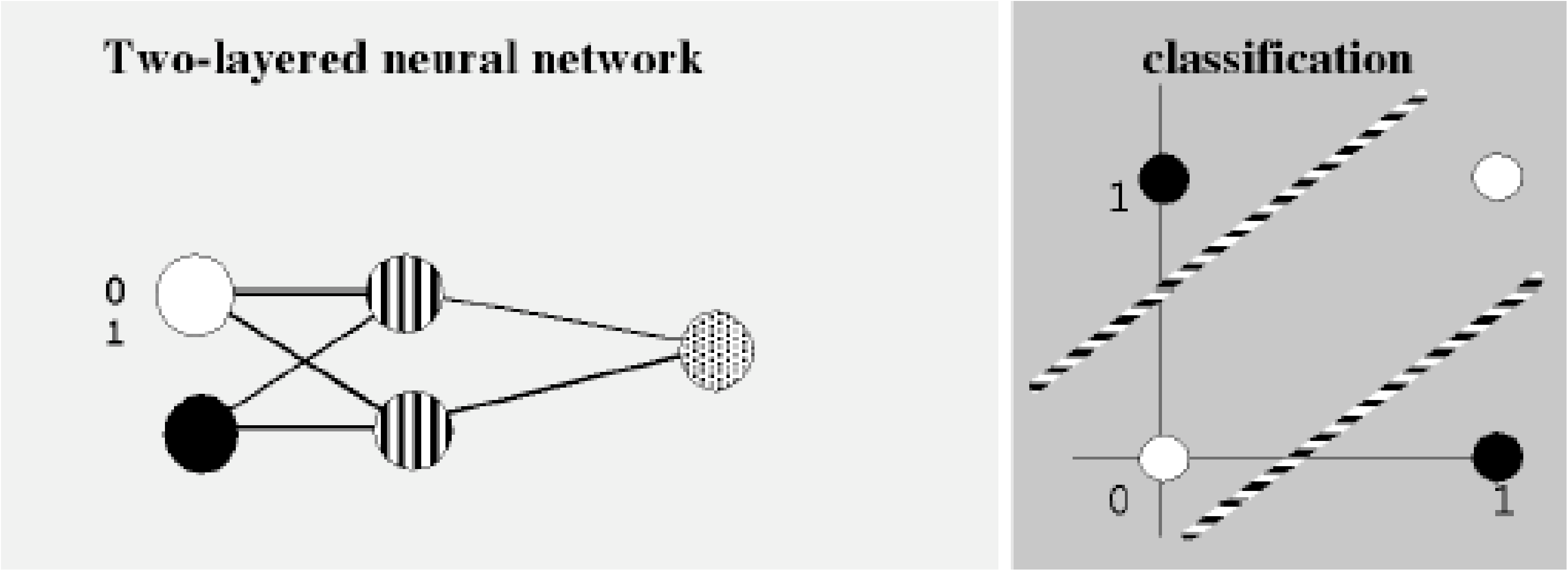
Evolution of Neural Networks

- Neural networks originally designed to approximate connections between neurons in the brain


Simple neural network: dynamic
0
1
J_{11}
J_{12}
out = 0.5 (tanh(h)+1)
$h = J_{11}\,in_1 + J_{12}\,in_2$
1
0.5
0
-2
0
2



Evolution of Neural Networks



Why use Neural Nets for Protein Folding?

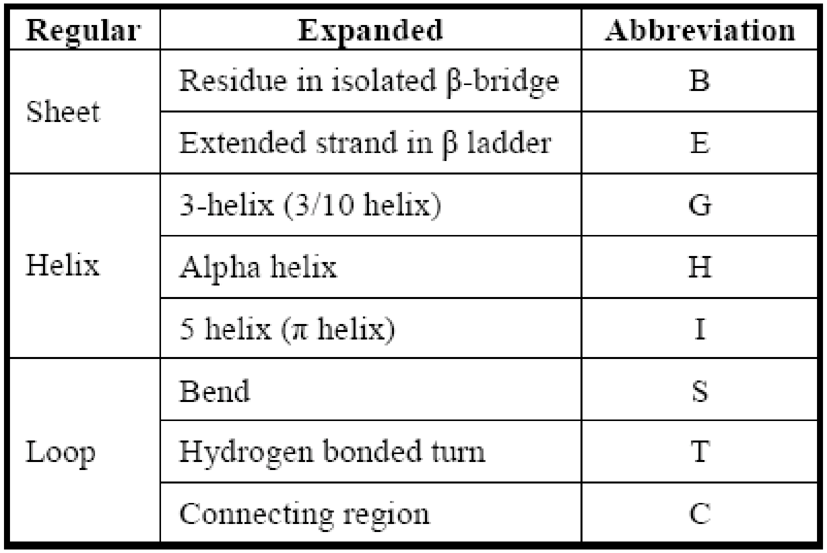
- Successful applications in:
  - Secondary structure prediction
  - Solvent access
- No “inherent shortcoming” yet found
- Can incorporate evolutionary information via multiple alignments
- Detect previous misclassifications

Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

- Purpose
  - Using neural nets, effectively predict the secondary structure of proteins.
- Current best for secondary structure prediction is SSpro8 with accuracy in the range of 62-63%

Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

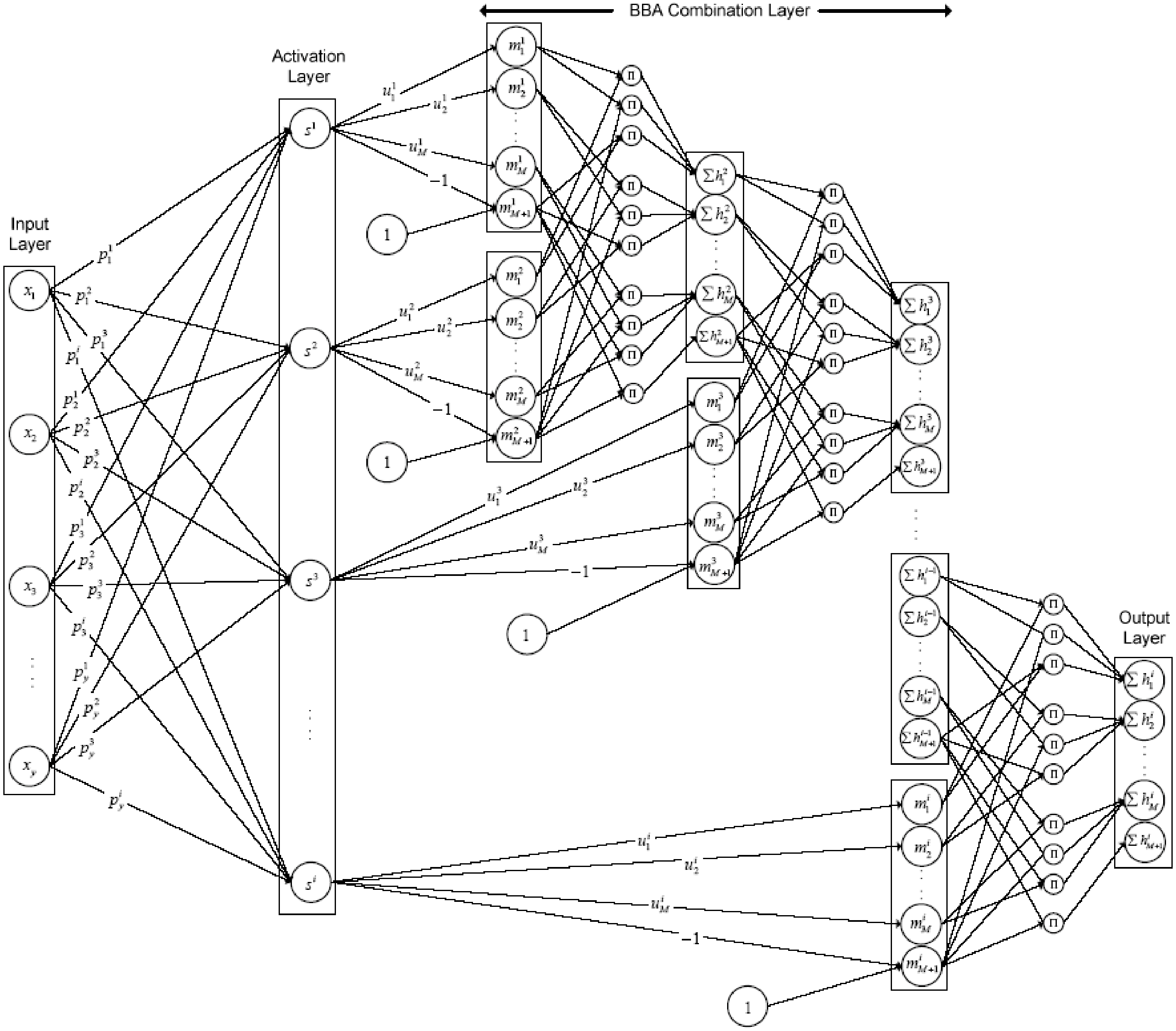
- Input to the system
  - Can choose to use DNA or amino acid sequences
  - SSpro8 uses amino acid sequences
  - The authors' system, UTMPred, uses DNA
- Output - forms consisting of alpha helices, beta sheets and loops expanded to eight structure forms



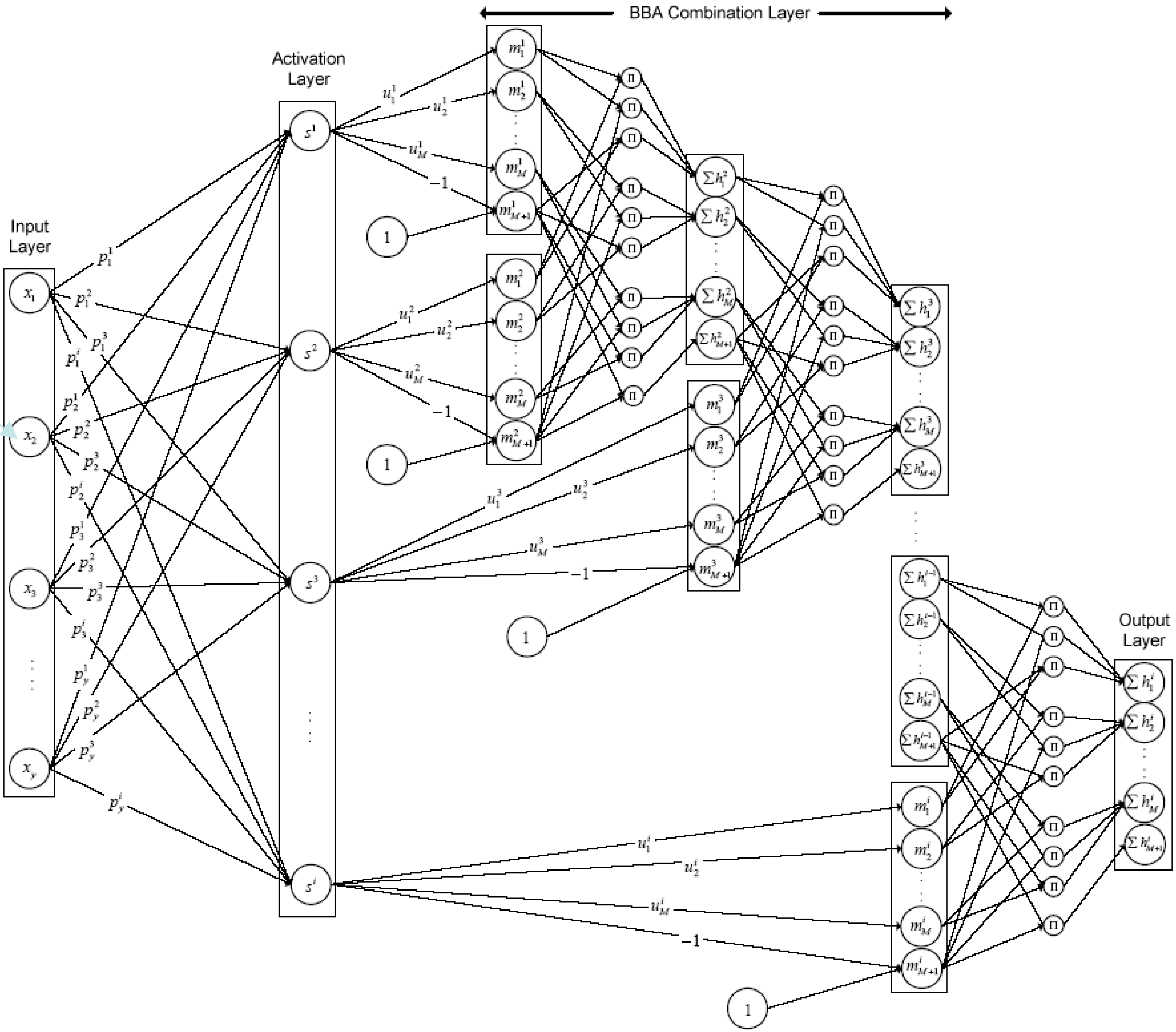
Regular	Expanded	Abbreviation
Sheet	Residue in isolated β-bridge	B
	Extended strand in β ladder	E
Helix	3-helix (3/10 helix)	G
	Alpha helix	H
	5 helix (π helix)	I
Loop	Bend	S
	Hydrogen bonded turn	T
	Connecting region	C

Protein Secondary Structure Forms

Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

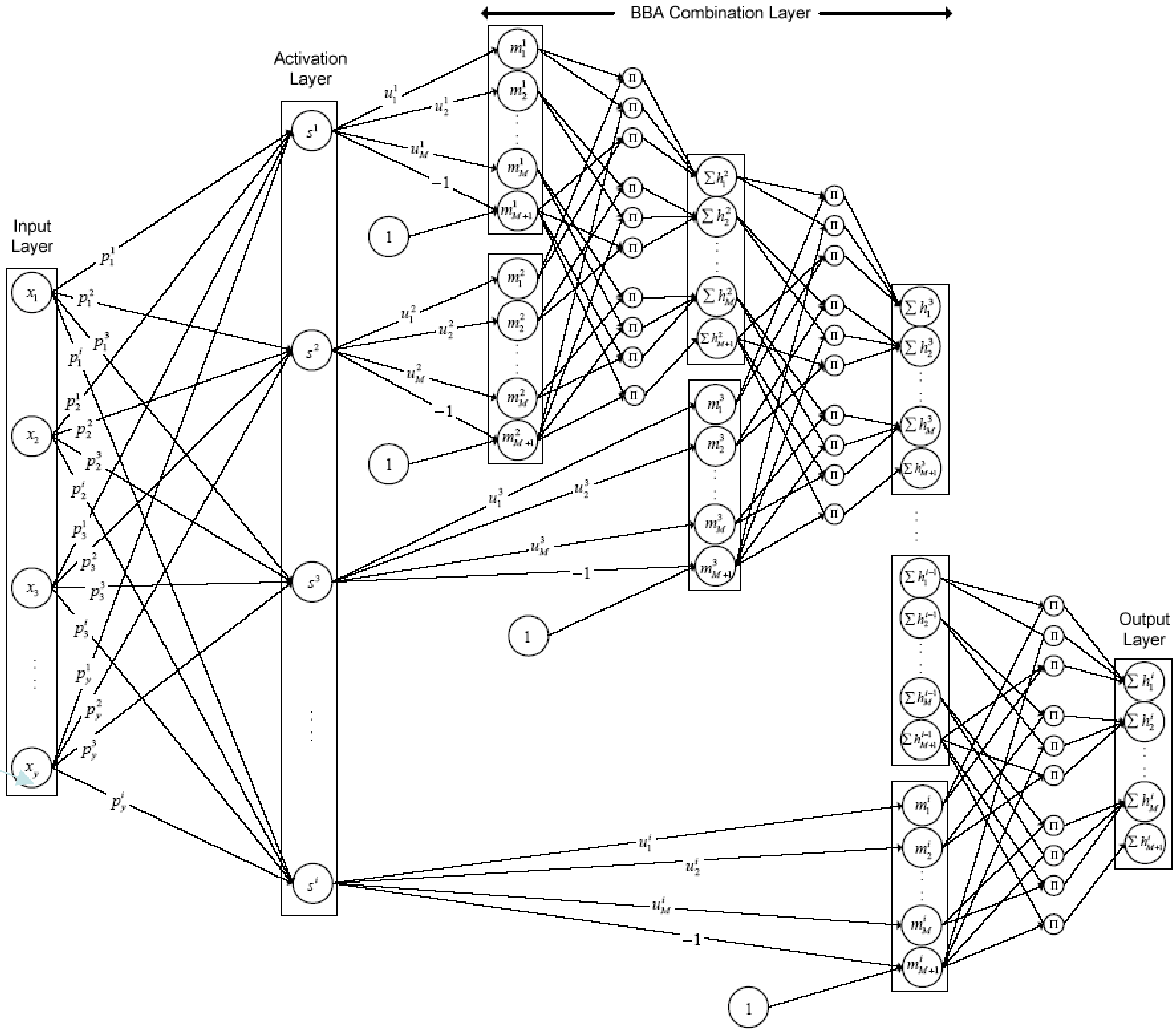


Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network



x - input to the network

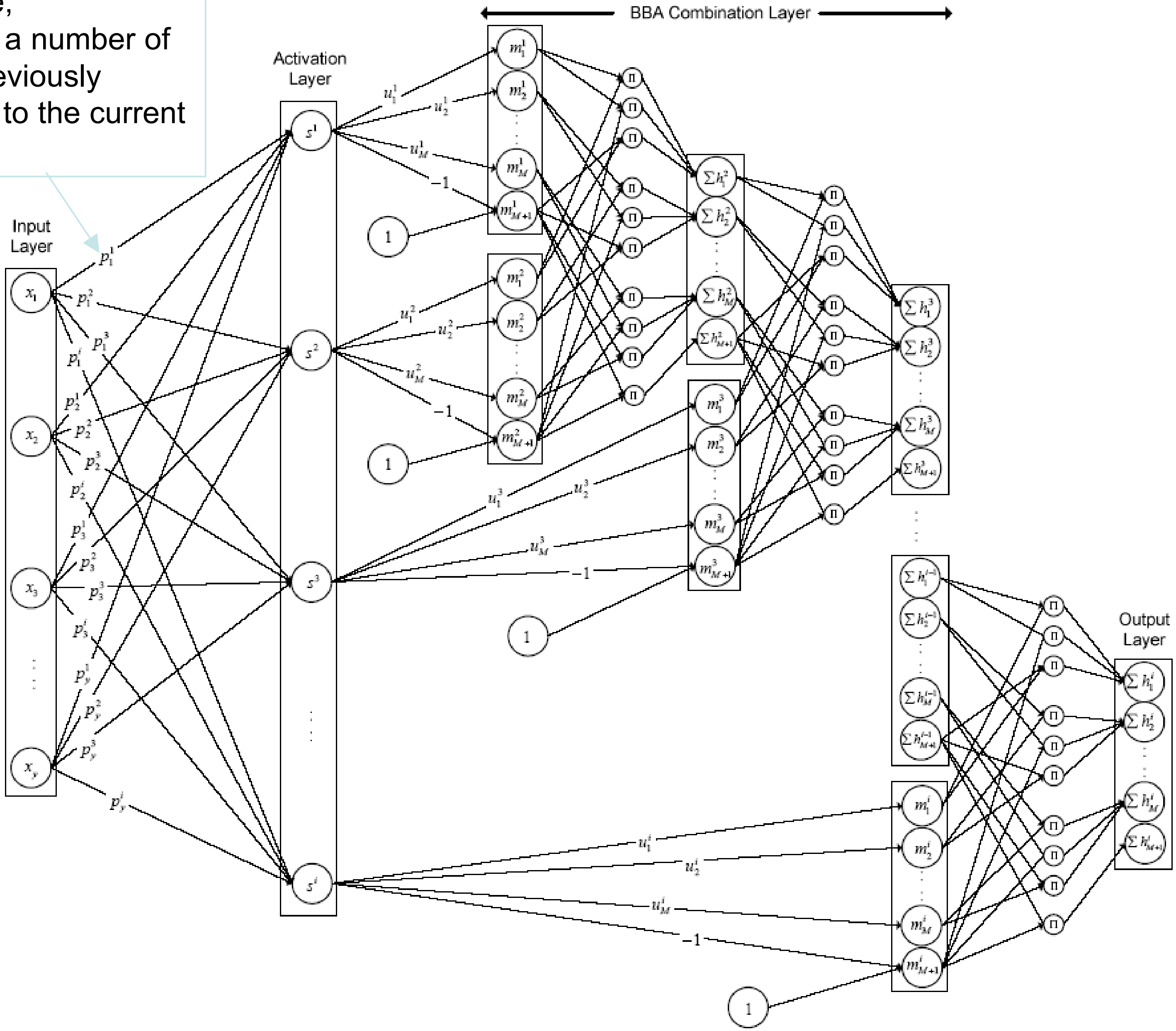
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network



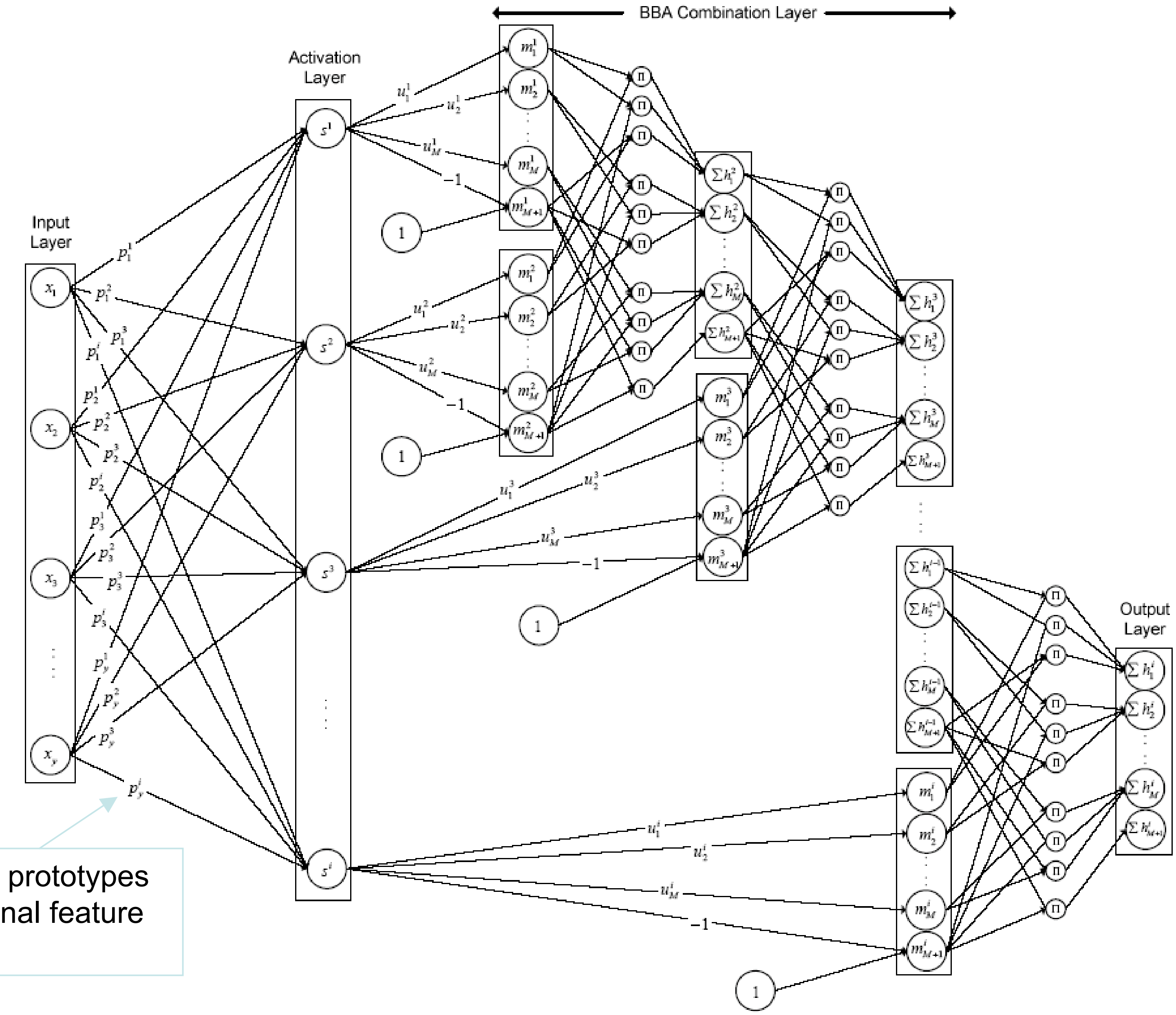
y – number of input nodes

Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

p – prototype, representing a number of ***k*** nearest previously trained input to the current tested input



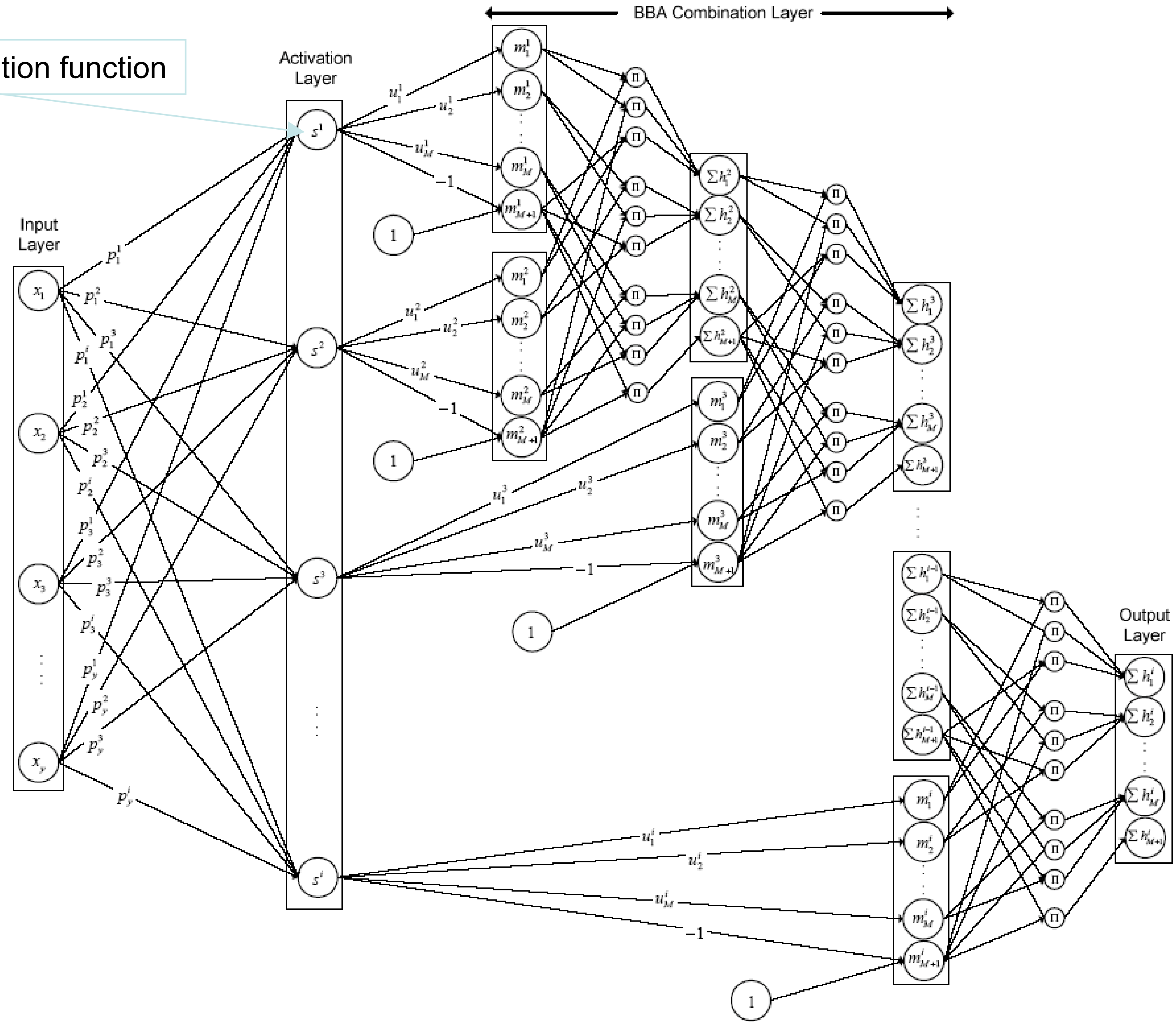
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network



i – number of prototypes in y-dimensional feature space

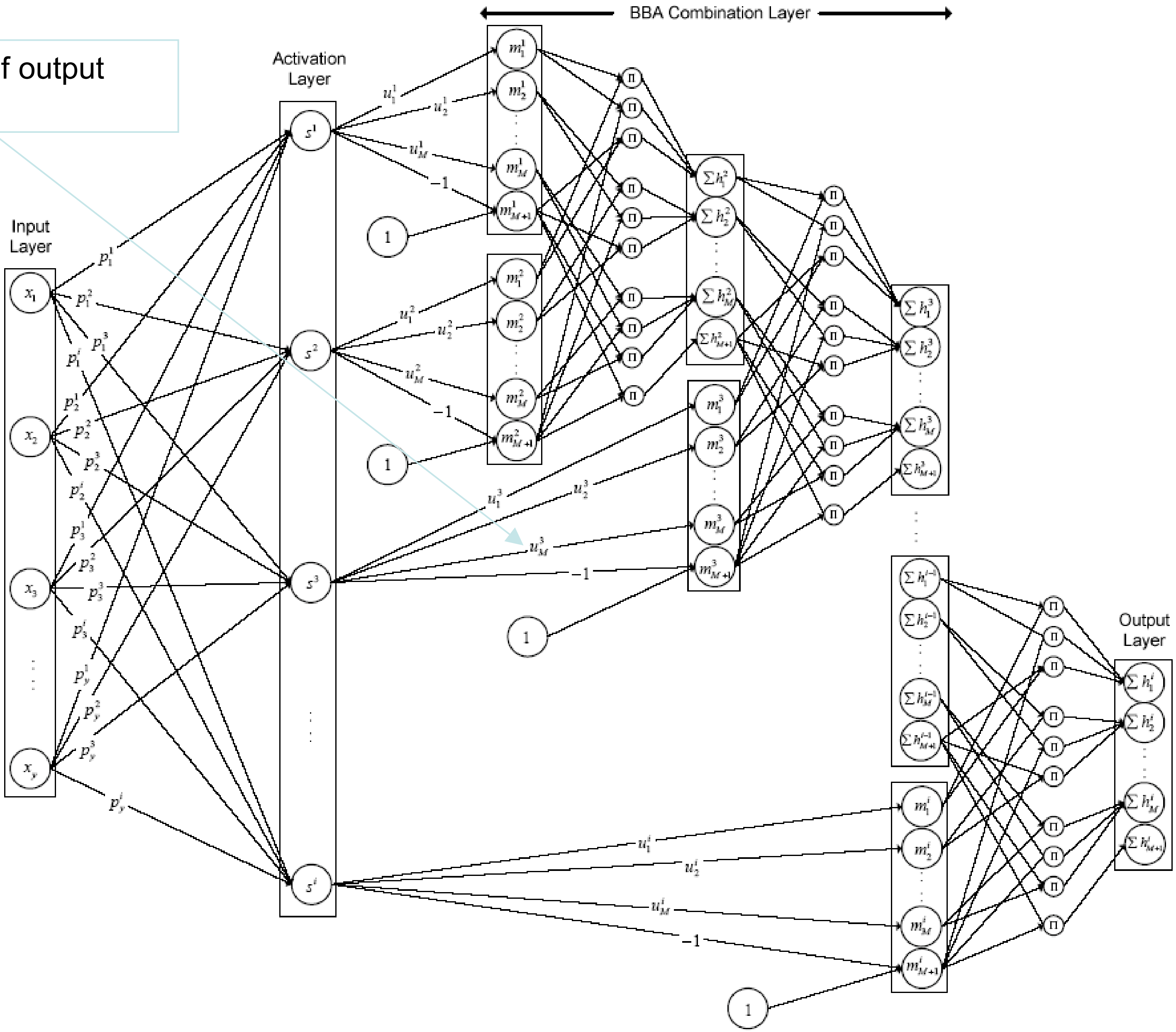
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

s – the activation function



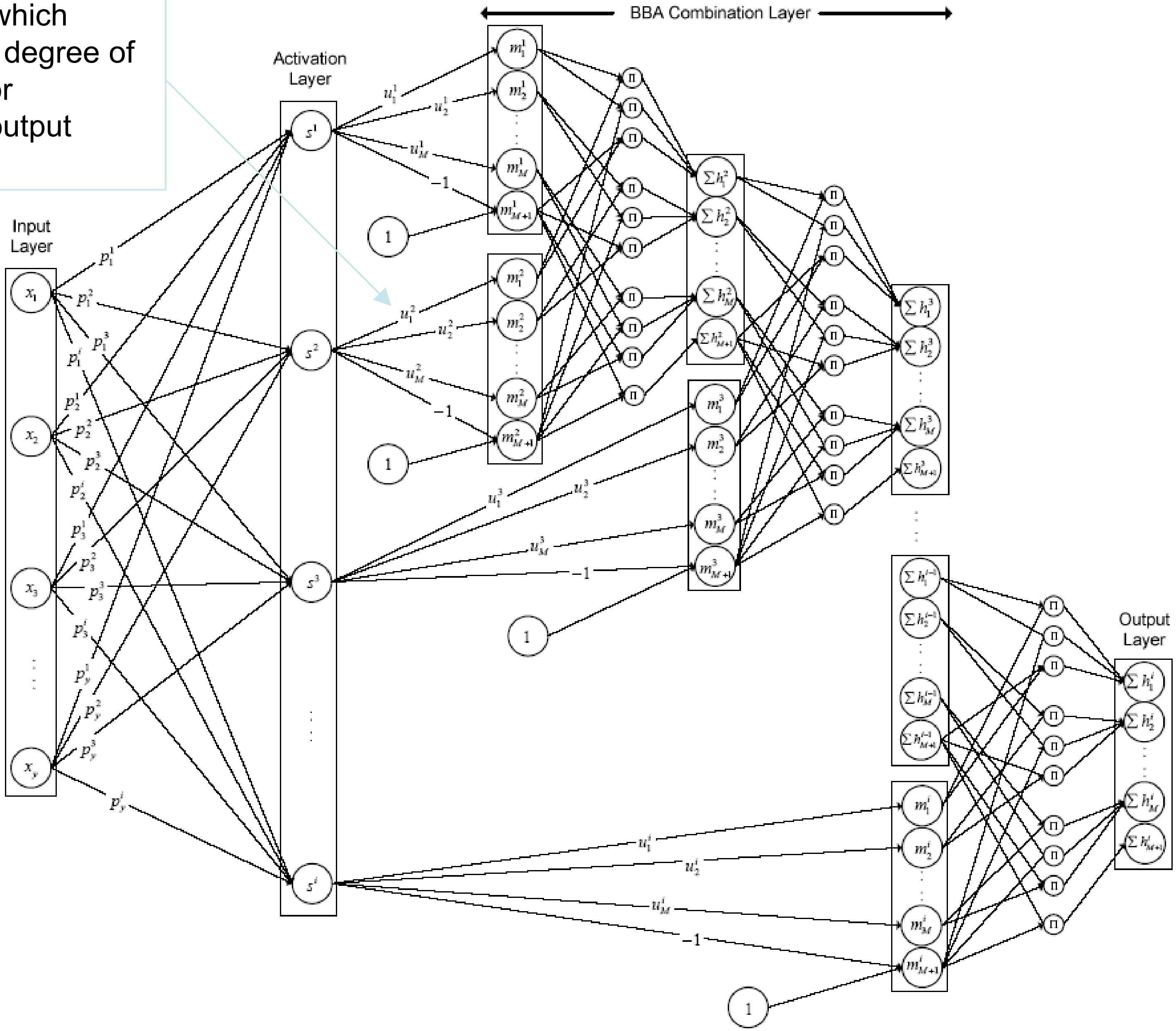
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

M – number of output classes


BBA Combination Layer
Activation Layer
Input Layer
Output Layer
x_1
x_2
x_3
x_y
p_1^1 p_1^2 p_1^3 p_1^i
p_2^1 p_2^2 p_2^3 p_2^i
p_3^1 p_3^2 p_3^3 p_3^i
p_y^1 p_y^2 p_y^3 p_y^i
s^1
s^2
s^3
s^i
u_1^1 u_2^1 u_M^1 -1
u_1^2 u_2^2 u_M^2 -1
u_1^3 u_2^3 u_M^3 -1
u_1^i u_2^i u_M^i -1
1
m_1^1 m_2^1 m_M^1 m_{M+1}^1
m_1^2 m_2^2 m_M^2 m_{M+1}^2
m_1^3 m_2^3 m_M^3 m_{M+1}^3
m_1^i m_2^i m_M^i m_{M+1}^i
Π
$\sum h_1^2$ $\sum h_2^2$ $\sum h_M^2$ $\sum h_{M+1}^2$
$\sum h_1^3$ $\sum h_2^3$ $\sum h_M^3$ $\sum h_{M+1}^3$
$\sum h_1^{i-1}$ $\sum h_2^{i-1}$ $\sum h_M^{i-1}$ $\sum h_{M+1}^{i-1}$
$\sum h_1^i$ $\sum h_2^i$ $\sum h_M^i$ $\sum h_{M+1}^i$

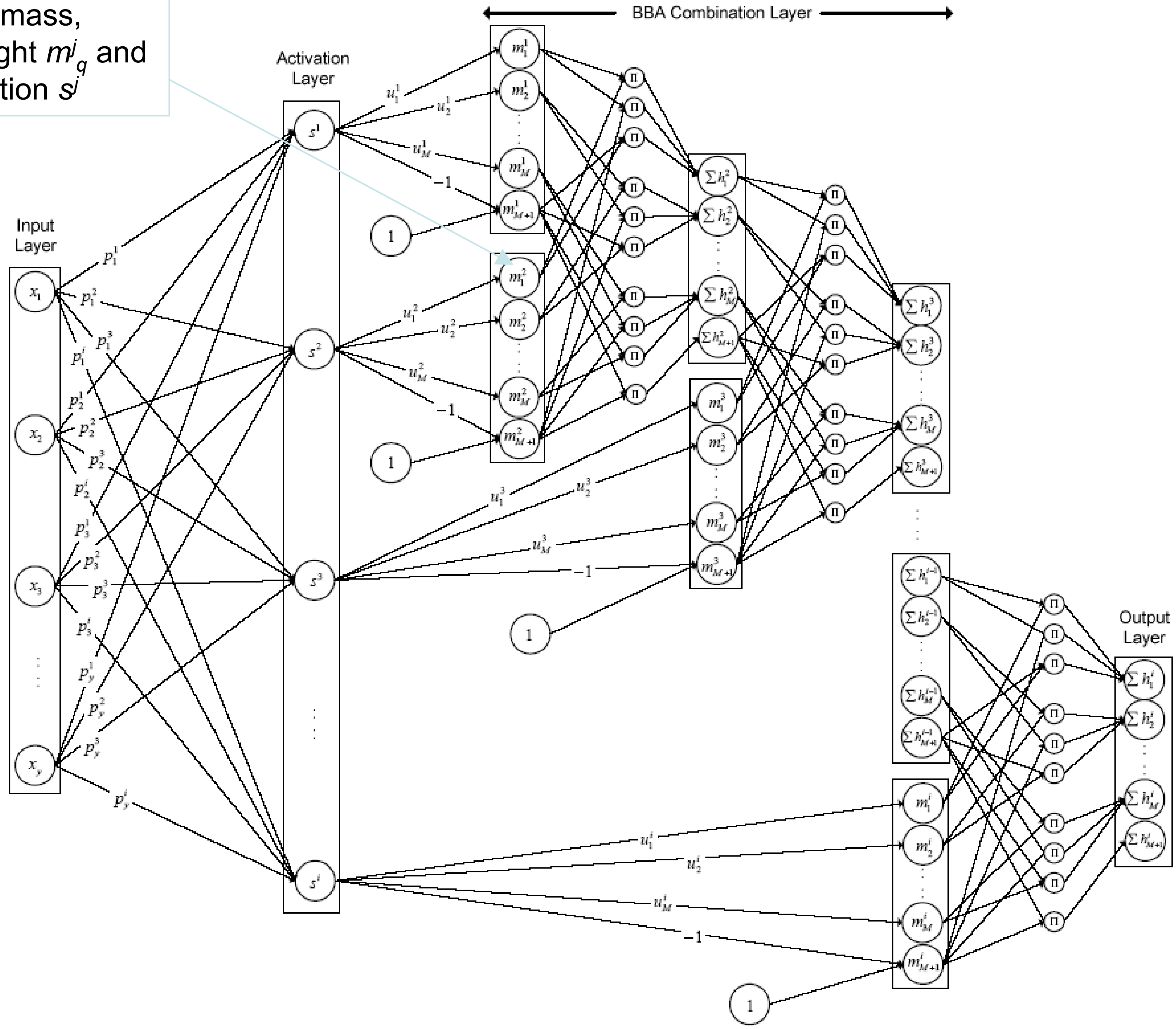
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

u^j_q – a weight which represents the degree of membership for prototype j to output class ***q***



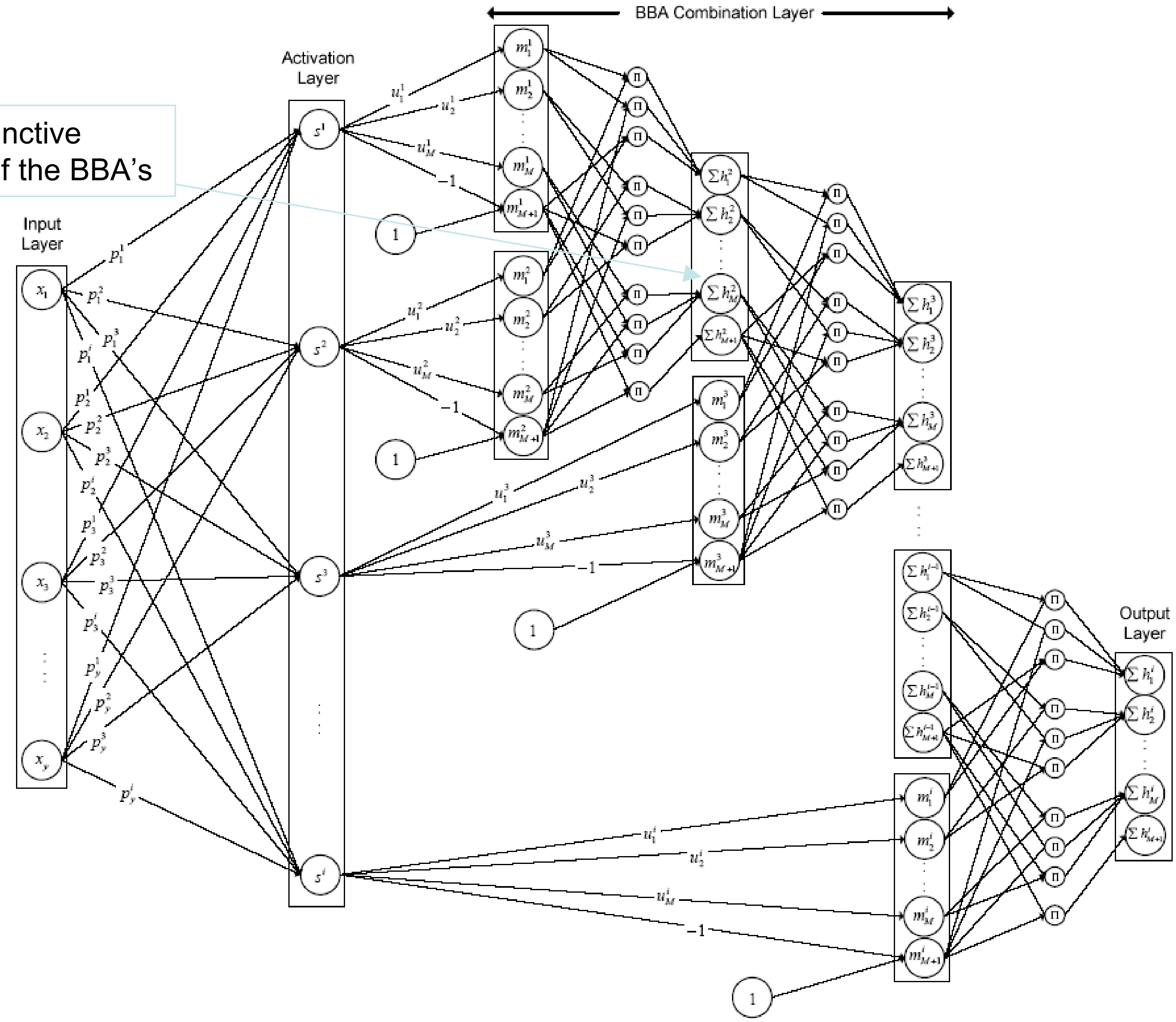
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

m^j_q – the BAA mass, product of weight m^j_q and activation function s^j



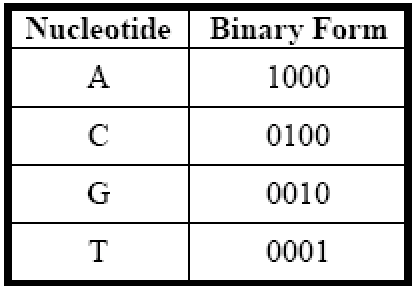
Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

h^j_q – the conjunctive combination of the BBA's



Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

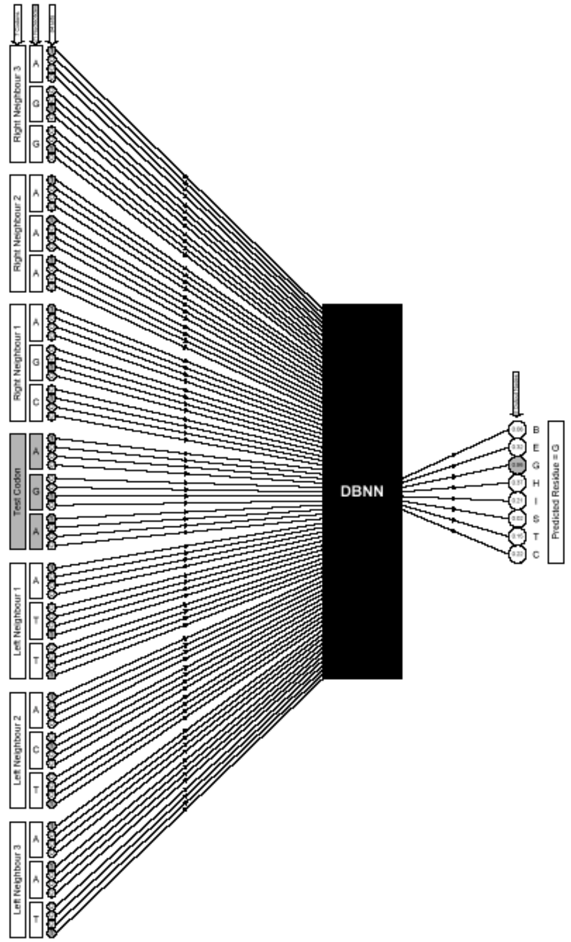
- The DBNN input are DNA sequences converted to binary format prior to use
- The sequences are:
  - 88 *Escheichia coli* proteins
  - 25 yeast *Saccharomyces cerevisiae* proteins
  - 166 mammalian proteins (80 of which are human)



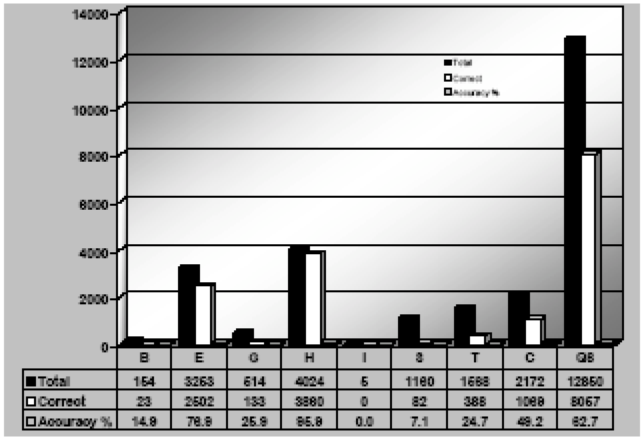
Nucleotide	Binary Form
A	1000
C	0100
G	0010
T	0001

Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network

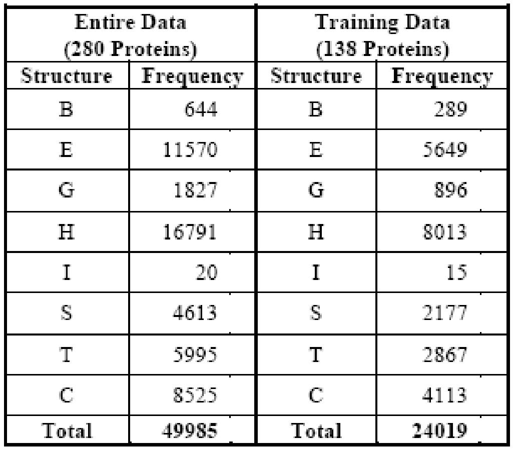
- The input window size for UTMPred is set to 7 codons, which results in 84 input nodes and 8 output notes which represent the expanded structural forms.



Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network



	B	E	G	H	I	S	T	C	Q8
■ Total	154	3263	614	4024	5	1160	1668	2172	12860
□ Correct	23	2602	133	3860	0	82	388	1069	8067
□ Accuracy %	14.9	79.9	25.9	95.9	0.0	7.1	24.7	49.2	62.7



Entire Data (280 Proteins)		Training Data (138 Proteins)	
Structure	Frequency	Structure	Frequency
B	644	B	289
E	11570	E	5649
G	1827	G	896
H	16791	H	8013
I	20	I	15
S	4613	S	2177
T	5995	T	2867
C	8525	C	4113
Total	**49985**	**Total**	**24019**

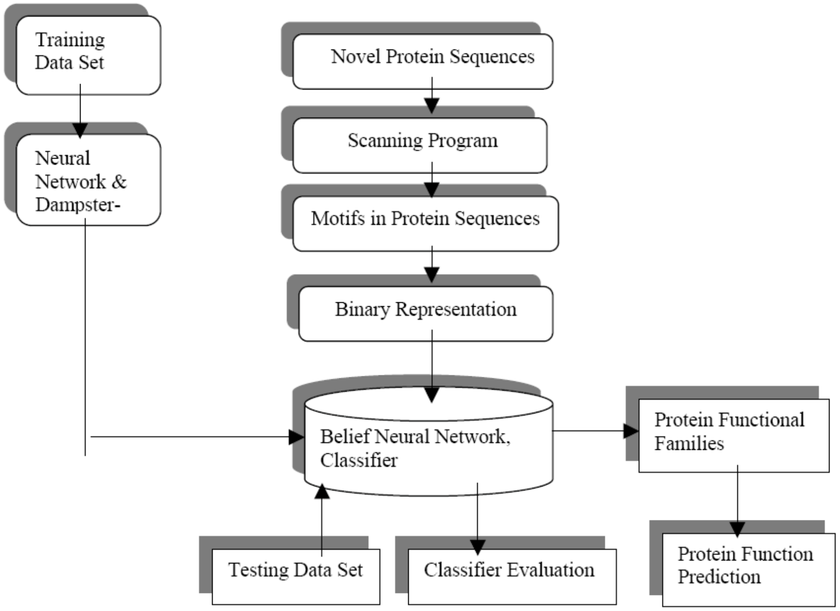
- UTMPred used 200 prototypes and after the training was completed, the system was able to predict H and E forms with accuracy above 75%. At the same time, the system had difficulty predicting form I, due to a small amount of data in the training samples.

Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory

- Purpose
  - Using neural networks, efficiently predict protein function
- Using databases such as Prosite, Pfam, and Prints, either query the databases for motifs within a protein in question, or query for an absence or presence of arbitrary combinations of motifs.

Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory

- Given a training set, induce a classifier able to assign novel protein sequences to one of the protein families represented in the training set
- Once trained, the classifier will be able to predict novel proteins into specific functional families based on its knowledge of the training set



Training Data Set

Neural Network & Dampster-

Novel Protein Sequences

Scanning Program

Motifs in Protein Sequences

Binary Representation

Belief Neural Network, Classifier

Protein Functional Families

Testing Data Set

Classifier Evaluation

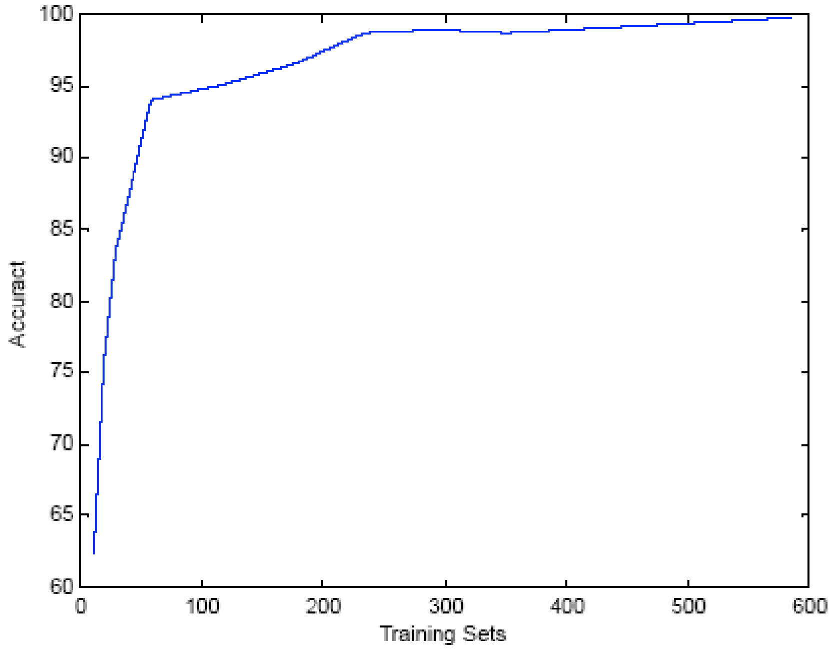
Protein Function Prediction

Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory

- Input data
  - From the Prosite database containing over 1100 entries. Each entry describes a function shared by some proteins. In the experiment one Prosite documentation entry corresponded to a protein class, and each protein class could, in turn, be characterized by one or more motif patterns/profiles. Only motifs considered significant matches by profileScan were chosen.
- DBNN was used as the classifier.

Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory

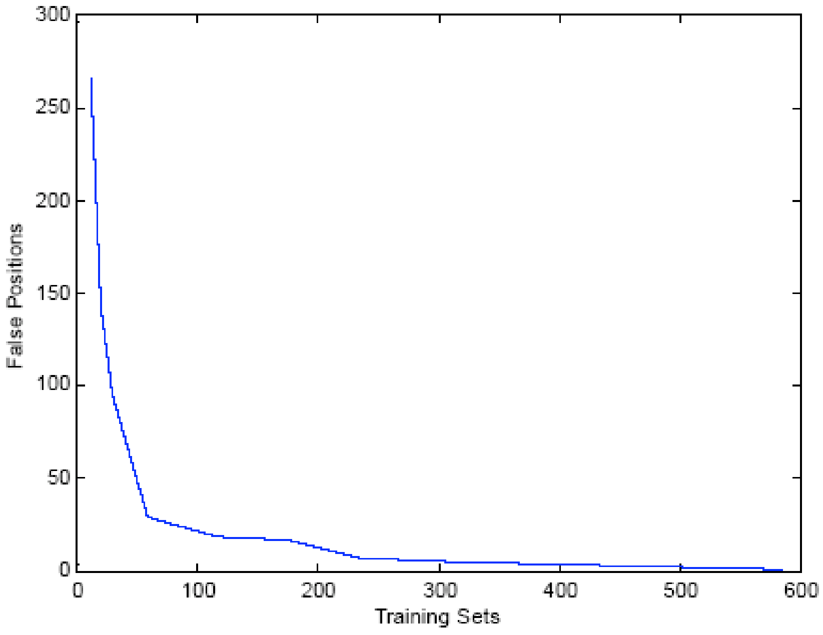
- 585 proteins belonging to one of ten classes were used, out of which subsets of varying size were picked randomly to become the training set.
- Once the DBNN was trained, all 585 proteins were used as the test set to determine accuracy



With only 10% of the total training samples, DBNN could be constructed to classify proteins with a 95% accuracy.

Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory

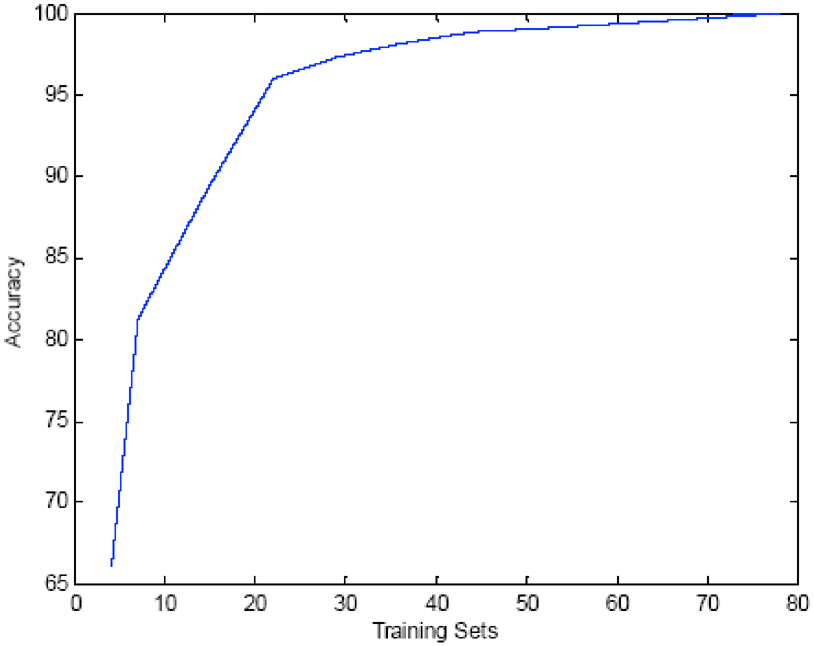
- The number of false positives generated by DBNN were significantly lower than those resulting from a Prosite search.
- As the size of the data set approaches 100%, the false positives discovered by DBNN approaches zero.



The number of false positives resulting from the use of the DBNN trained using training sets of different sizes.

Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory

- A second data set of 73 protein sequences drawn from five classes were used to build a DBNN classifier
- Using the DBNN classifier built by random sized datasets, the output exceeded 96% accuracy when the training set was greater or equal to 22
- Once the input contained more than 80% (58 or more sequences) of the dataset, all sequences were correctly predicted



Result of classifying proteins containing common motifs

Future Work

- Ultimate solution to “protein folding” will probably be a hybrid
- Neural networks likely to be included due to their successful application to related problems
  - Secondary structure
  - Solvent access
  - Distance between residues in final structure
  - Protein interface recognition
- In addition, neural nets can combine knowledge from multiple sources

Bibliography

- B. Rost. “Neural networks for protein structure prediction: hype or hit?” Artificial intelligence and heuristic methods for bioinformatics (2003): 34-50.
- S.N.V. Arjunan, S. Deris, R.M. Illias. “Protein Secondary Structure Prediction Based on Denoeux Belief Neural Network.” ICAIET Proceedings (2002): 554-560.
- N.M.Zaki, S. Deris, S.N.V. Arjunan. “Assignment of Protein Sequence to Functional Family Using Neural Network and Dempster-Shafer Theory” Journal of Theoretics 5-1 (2003).

Background information

- S.N.V Arkimam, S. Deris, R.M.Illias. “Prediction of Protein Secondary Structure” Jurnal Teknologi 35(C) (2001): 81-90.
- T.Wessels, C.W. Omlin.”Refining Hidden Markov Models with Recurrent Neural Networks”.