Lecture 11

SUPERVISED NEURAL NETWORKS FOR PROTEIN SEQUENCE ANALYSIS

Dr Lee Nung Kion
Faculty of Cognitive Sciences and Human Development
UNIMAS, http://www.unimas.my

Introduction

- Protein sequences are composed of 20 amino acids
- The twenty amino acid letters are: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
- Proteins are product of genes which have many functions in our body: antibodies, enzymes, structural (hairs, tendons etc) etc.

Introduction

- A **sequence motif** is a short amino pattern sequence in a protein sequence that has biological significance.
- For example:
- AATCKLMMVTVVWTTAGA

Underlined are motifs important for the function of this protein

- Proteins in the same functional domain will share a common motif

Introduction

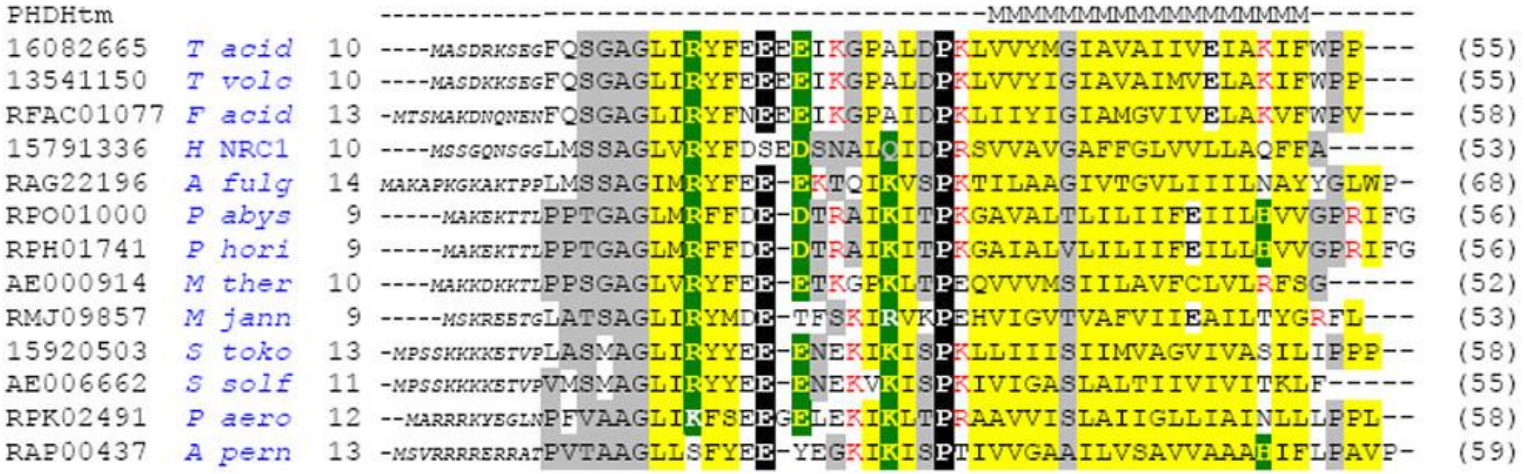
- A protein superfamily comprises set of protein sequences that are evolutionary and therefore functionally and structurally related
- Protein sequences in a family share some common motifs
- Two protein sequences are assigned to the same class if they have high homology in the sequence level (e.g., common motif).

First fact of biology

- “if two peptides stretches exhibit sufficient similarity at the sequence level, then they are likely to be biologically related”

Sequence alignment

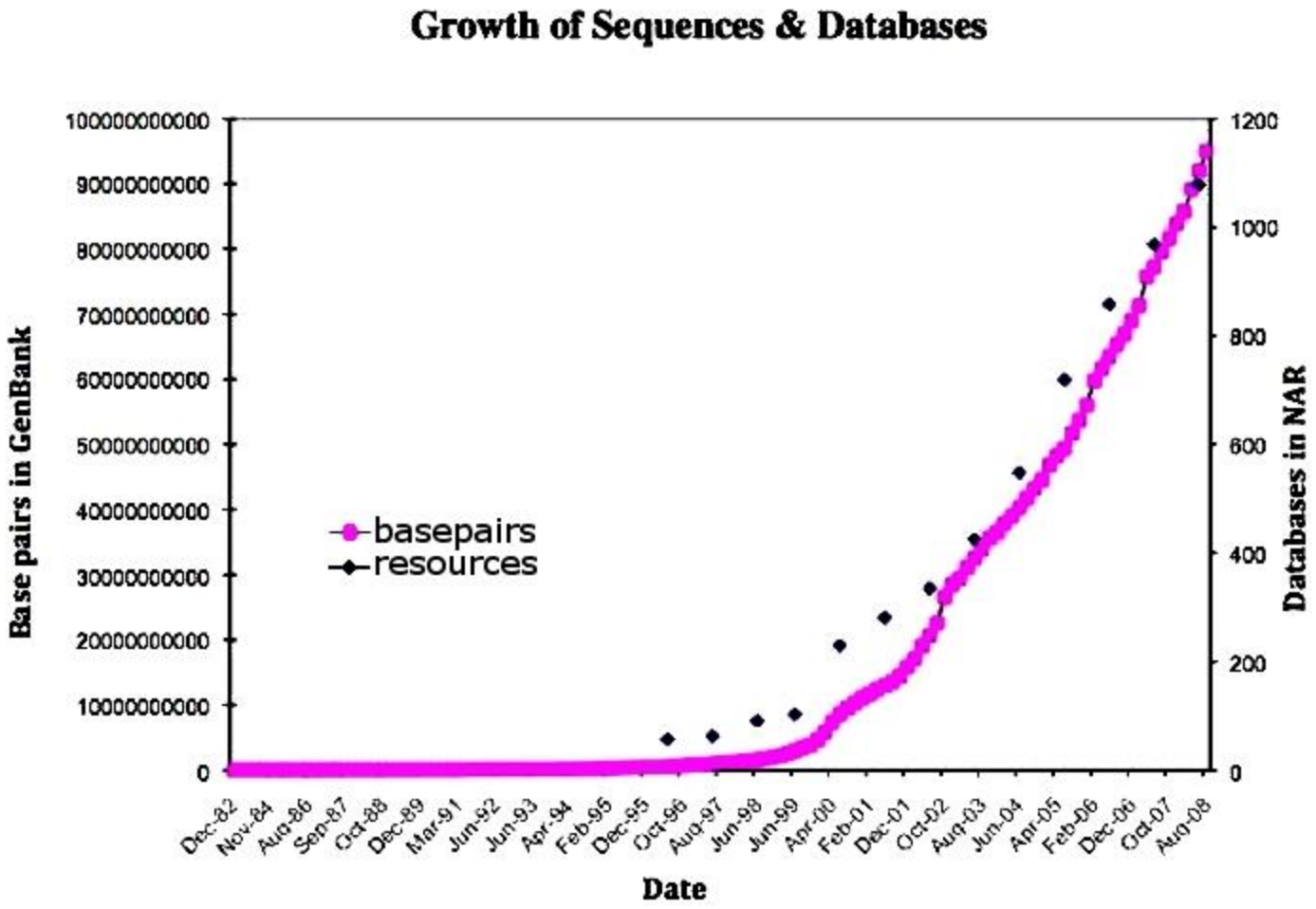
- The similarity between two protein sequences are commonly established through multiple alignment algorithm (e.g., BLAST)



Example of multiple sequence alignment to identify common motifs in protein sequences

Protein families

- Rapid grow in the number of protein sequences
- Searching one query sequence against all in all the databases is computationally expensive, need super-computer or weeks of computational time

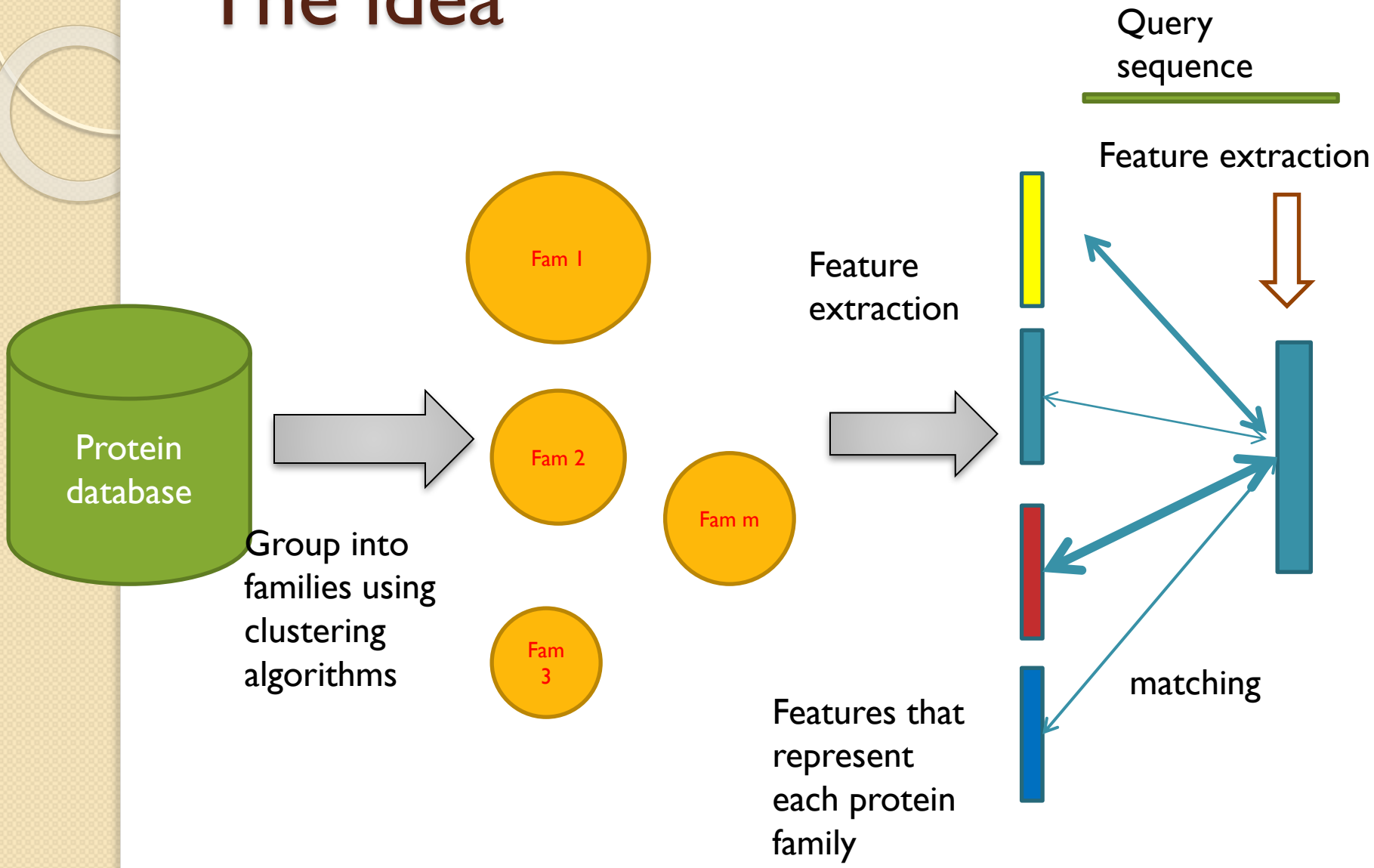


growth of sequence data in GenBank

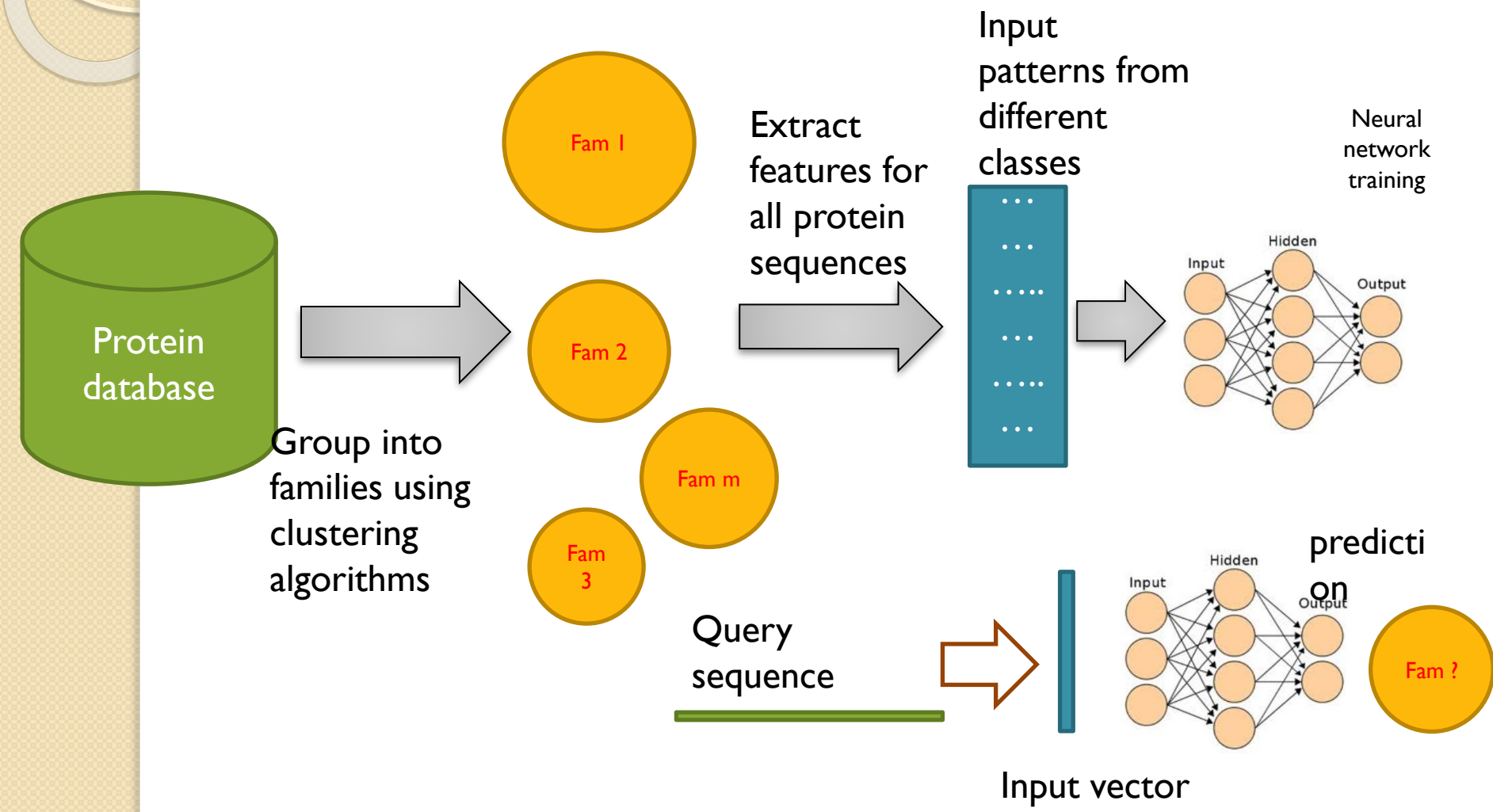
Solution

- Grouping protein into families based on the sequence level similarity can aid in:
  - Group analysis of sequences to identify common motifs;
  - Support rapid search of a protein sequence

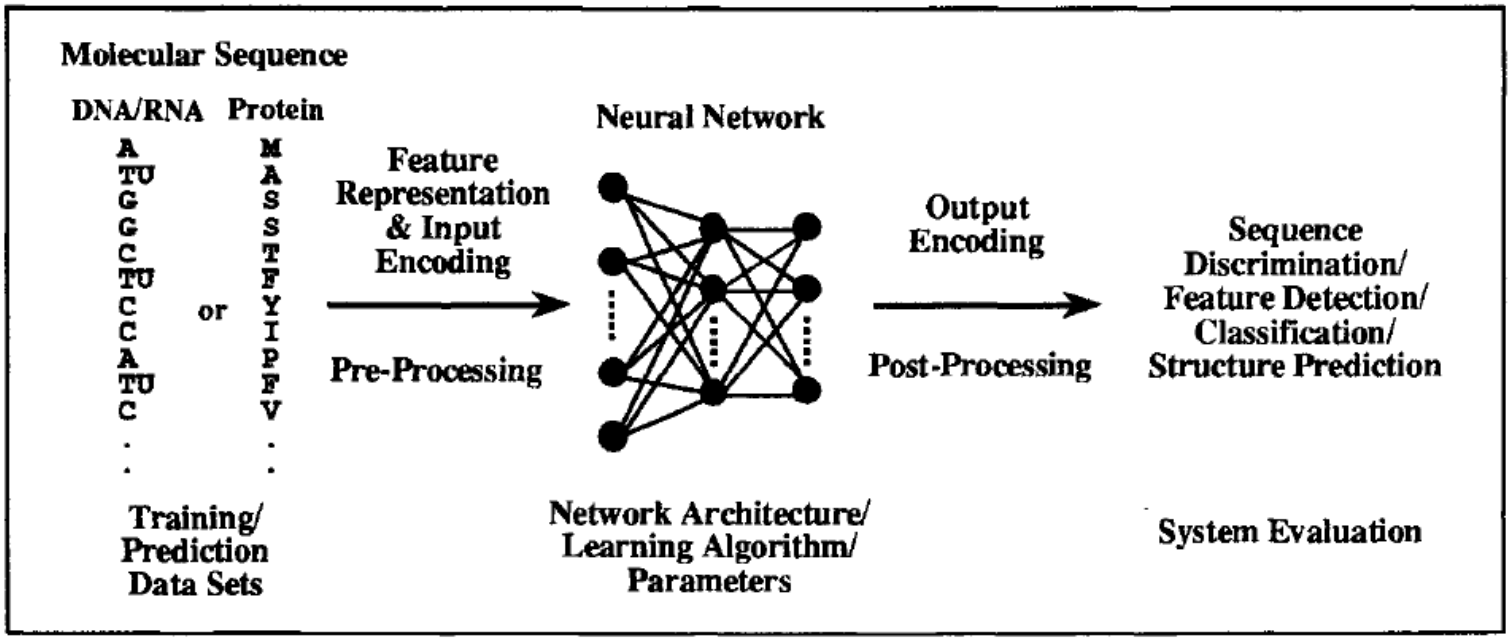
The idea


Query sequence
Feature extraction
Fam 1
Feature extraction
Protein database
Fam 2
Fam m
Group into families using clustering algorithms
Fam 3
matching
Features that represent each protein family

Neural network for protein family classification



Issues on application on NN for biological sequences analysis



Input encoding

- How to convert the amino acids into numerical values?
- Direct encoding:
  - Each amino acid letter is represented by a binary coding
  - Various binary representation can be obtained based on the properties of amino acids

Input encoding

- Each amino acid is converted into numerical value based on its properties

Table 6.1 *Physicochemical and structural properties of amino acid residues.*

Amino Acid Residues (3-, 1-letter code)	Chemical Property	Volume (A^3)	Mass (daltons)	HP Scale K&D	Surface Area	2D Structure Propensity α-helix	β-strand	Turn
Alanine (Ala, A)	Aliphatic	67	71.09	1.8	0.74	1.41	0.72	0.82
Arginine (Arg, R)	Basic	148	156.19	-4.5	0.64	1.21	0.84	0.90
Asparagine (Asn, N)	Amide	96	114.11	-3.5	0.63	0.76	0.48	1.34
Aspartic Acid (Asp, D)	Acidic	91	115.09	-3.5	0.62	0.99	0.39	1.24
Cysteine (Cys, C)	Reactive	86	103.15	2.5	0.91	0.66	1.40	0.54
Glutamine (Gln, Q)	Amide	114	128.14	-3.5	0.62	1.27	0.98	0.84
Glutamic Acid (Glu, E)	Acidic	109	129.12	-3.5	0.62	1.59	0.52	1.01
Glycine (Gly, G)	Small	48	57.05	-0.4	0.72	0.43	0.58	1.77
Histidine (His, H)	Aromatic	118	137.14	-3.2	0.78	1.05	0.80	0.81
Isoleucine (Ile, I)	Aliphatic	124	113.16	4.5	0.88	1.09	1.67	0.47
Leucine (Leu, L)	Aliphatic	124	113.16	3.8	0.85	1.34	1.22	0.57
Lysine (Lys, K)	Basic	135	128.17	-3.9	0.52	1.23	0.69	1.07
Methionine (Met, M)	Aliphatic	124	131.19	1.9	0.85	1.30	1.14	0.52
Phenylalanine (Phe, F)	Aromatic	135	147.18	2.8	0.88	1.16	1.33	0.59
Proline (Pro, P)	Cyclic Imino	90	97.12	-1.6	0.64	0.34	0.31	1.32
Serine (Ser, S)	Hydroxyl	73	87.08	-0.8	0.66	0.57	0.96	1.22
Threonine (Thr, T)	Hydroxyl	93	101.11	-0.7	0.70	0.76	1.17	0.90
Tryptophan (Trp, W)	Aromatic	163	186.21	-0.9	0.85	1.02	1.35	0.65
Tyrosine (Tyr, Y)	Aromatic	141	163.18	-1.3	0.76	0.74	1.45	0.76
Valine (Val, V)	Aliphatic	105	99.14	4.2	0.86	0.90	1.87	0.41

Input encoding

- Amino acids are converted to binary vectors or feature values

Encoding Method	Residue/ Window	Vector Size	Vector or Scalar Value*
Indicator Vector of AA Alphabet (BIN20)	Alanine	20	{1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}
Ind. Vector of AA Alph. & Spacer (BIN21)	Alanine	21	{1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}
Ind. Vector of Exchange Group (EG) Alph.	Alanine	6	{0, 0, 0, 1, 0, 0}
Ind. Vector of Hydrophobicity (HP) Alph.	Alanine	3	{0, 0, 1}
Fuzzy Vector of HP Alph.	Alanine	3	{0, 0.2, 0.8}
Feature Vector of K&D HP Scale	Alanine	1	{1.8}
PAM Substitution Vector	Alanine	20	Values varied
Evolutionary Vector	Alanine	20	Values varied
Ind. Vector of NA Alph. (Sparse) (BIN4)	Adenine	4	{1, 0, 0, 0}; others are 0100, 0010,0001
Ind. Vector of NA Alph. (Dense)	Adenine	2	{0, 0}; other are 01, 10, 11
Ind. Vector of NA Alph. (Ordinal)	Adenine	1	{1}; others are 2, 3, 4
Ind.Vector of AA Alph. (BIN20)	ANLAIDV	20 x 7	{1,19x 0;11x0,1,8x0; 9x0,...;17x0,1,0,0}
Ind. Vector of Exchange Group (EG) Alph.	ANLAIDV	6 x 7	{000100; 010000; 000010; ...; 000010}
Ind. Vector of Hydrophobicity (HP) Alph.	ANLAIDV	3 x 7	{001; 100; 001; 001; 001; 100; 001}
Ind. Vector of HP Alph. Pairs	ANLAIDV	9 x 6	{000000100;001000000;....; 001000000}
Feature Vector of K&D HP Scale	ANLAIDV	1 x 7	{1,8; -3.5; 3.8; 1.8; 4.5; -3.5; 4.1}
PAM Substitution Vector	ANLAIDV	20 x 7	Values varied
Evolutionary Vector	ANLAIDV	20 x 7	Values varied

(Wu & McLarty, 2000)

Direct encoding 1-of-20 (Bin20)

- Each amino acid is represented by 20 binary digits, with only one of them is 1, others zero.
- E.g.,
- A = [1, 0, 0, 0, 0, 0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0]
- D = [0, 1, 0, 0, 0, 0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0]
- …
- K = [0, 0, 0, 0, 0, 0, 0,0,0,0,0,0,0,0,0,0,0,0,0,1].
- A protein sequence becomes the concatenation of these binary vectors.

Exchange groups

- The twenty amino acids can be grouped according to various chemical/structural properties.

***Table 6.2** Alphabet sets for feature representation: some examples.*

Alphabet Name	Size	Features	Membership
AAIdentity	20	Sequence Identity	A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y
ExchangeGroup	6	Conservative Substitution	{HRK} {DENQ} {C} {STPAG} {MILV} {FYW}
ChargePolarity	4	Charge and Polarity	{HRK} {DE} {CTSGNQY} {APMLIVFW}
Hydrophobicity	3	Hydrophobicity	{DENQRK} {CSTPGHY} {AMILVFW}
Mass	3	Mass	{GASPVTC}{NDQEHILKM}{RFWY}
Structural	3	Surface Exposure	{DENQHRK} {CSTPAGWY} {MILVF}
2DPropensity	3	2D Structure Propensity	{AEQHKMLR} {CTIVFYW} {SGPDN}

Indirect encoding

- n-gram features
  - n-gram feature is the number of occurrences of a short protein sequence of length n in a protein sequence.
  - Definition:

    The n-gram features is a pair of values (v_i, c_i) where $v_i \in \Sigma$ depicts the i-th gram feature and c_i is the counts of this feature in a protein sequence for i=i… . $|\Sigma|^n$
  - For example: the 2-gram are:
  - (AA, AB, …, AY, BA, BB, …, BY, YA, …, YY).

N-gram feature example

- AGCCDDAGAGKDDV

AG – 3
GC- 1
CC -1
CD-1
DD-2
DA-1
GA-1
GK -1
KD -1
DV - 1

Indirect encoding

- Problems with n-gram features:
  - Large input dimension for n > 2

N-gram	# of inputs
2	400
3	8000
4	16000
5	3200000
6	64000000

  - Solution: feature selection

N-gram feature

- N-gram feature can also be reduced by using the amino acid exchange groups in Table 6.2.
- E.g.

***Table 6.2** Alphabet sets for feature representation: some examples.*

Alphabet Name	Size	Features	Membership
AAIdentity	20	Sequence Identity	A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y
ExchangeGroup	6	Conservative Substitution	{HRK} {DENQ} {C} {STPAG} {MILV} {FYW}
ChargePolarity	4	Charge and Polarity	{HRK} {DE} {CTSGNQY} {APMLIVFW}
Hydrophobicity	3	Hydrophobicity	{DENQRK} {CSTPGHY} {AMILVFW}
Mass	3	Mass	{GASPVTC}{NDQEHILKM}{RFWY}
Structural	3	Surface Exposure	{DENQHRK} {CSTPAGWY} {MILVF}
2DPropensity	3	2D Structure Propensity	{AEQHKMLR} {CTIVFYW} {SGPDN}

N-gram with exchange groups

- Using hydrophobicity
  - A={DENQRK}, B={CSTPGHY}, C={AMILVFW}

- **AGCCDDAGAGKDDV→ CBBBAACBCBAAAC**
- 2-grams are:
- CB-2, BB-2, BA-2, AA-2, AC-1, BC-1, AA-2
- # of features just 3^2

Conclusion

- Neural network is an effective and efficient option for protein family classification with proper feature representation and input encoding.