CHAPTER I
INTRODUCTION

Proteins have many important functions in all biological processes such as enzymatic catalysis, transport and storage, coordinated motion, mechanical support, immune protection, generation and transmission of nerve impulses and control of growth and differentiation. Proteins are composed of one or more unbranced polypeptide chains which are, in turn, composed of amino acid residues joined together by *peptide bonds*. Each residue may be one of the 20 amino acid residues commonly found in proteins.

The work of Christian Anfinsen (1961) on ribonuclease showed that the information needed to specify complex three dimensional structure of a protein is contained in its amino acid sequence. Such three-dimensional structure and conformation dictate the function of the protein (Creighton, 1983; Blout *et al.*, 1960).

Works on polypeptides (Marqusee and Baldwin, 1987; Marqusee *et al.*, 1989) and small proteins (Ose & Kim, 1988; Roder *et al.*, 1988; Udgaonkar & Baldwin, 1988) suggested that a secondary structure can form independently and that the formation of some secondary structures may precede tertiary organization. Levitt and Chothia (1976) showed that there is a strong tendency for segments of secondary structure that are close together along the sequence to also be in close contact in the final three-dimensional (3D) structure. Such locally ordered regions, which are referred to here as *folding units*, associate to form the whole protein molecule or in the case of some of larger proteins, to form domains.

Most of the knowledge of protein structure came from the X-ray diffraction patterns of crystallized proteins or H-NMR spectrum (Williamson *et al.*, 1986; Bazzo *et al.*, 1988;

Marion & Wathrith, 1983). These methods can be very accurate, but many step are uncertain, complicated and time consuming (Qian & Sejnowski, 1988). Although developments in modem crystallography and NMR have clearly speeded up the process of tertiary structure determination, still they are very far from keeping pace with the protein primary structure output from DNA sequencing (Pascarella & Argos, 1992). Examination of three dimensional structures of proteins determined by x-ray diffraction and NMR has shown that the variety of folding patterns of proteins is significantly restricted (Chothia, 1992; Finkelstein & Ptitsyn, 1987). Protein sequence information, however, grew significantly faster than information on protein 3D structure, resulting in an enormous gap between the limited protein 3D structural information and the wealth of protein primary sequence data. In April 1991, a released of the Protein Data Bank (PDB) on 190 unique 3D protein structures was reported (Bernstein *et al.*, 1977), there were nearly 20,000 primary sequence entries from a release 17.0 of SWISSPROT sequence data bank (Bairoch & Boeckmann, 1991) (Fig. 1.1). In 1995, there were about 36,000 sequences but only 2,000 of them had known 3D structures (Rost and Sander, 1995).

Neural network methods have been used as an approach for prediction of 3D structures from amino acid sequences. Numerous pieces of work have been published during the past few years on the usability of these methods in protein structure research. Computational neural network is inspired from an architecture of neurons in the human brain (Alexander & Morton, 1990). In principle, simple neural networks consist of processing elements arranged in several layers and interconnected between such layers. Information and signals are transferred through these connections and processed by such processing elements or "neurons". The connections are numerically weighted. The weights are gradually changed and adapted in the "training phase" or " learning phase" until each pattern presented to the input layer of "neurons" is accurately projected onto the corresponding resulting pattern on the output layer. Threshold predefined values of the incoming signals have to accumulate in each respective processing element until a value is reached before an output signal is passed on to the connected "neurons" in the next layer. The methodological advantage with a neural network is its sensitivity to detect subtle patterns in the incoming data which may in some cases not be recognized by statistical or algorithmic methods (Eisenhaber *et al.*, 1995; Bohm, 1996)

Neural network technology has been applied in several ways for the study of proteins and nucleic acids (Hirst & Sternberg, 1992), for example;

Qian and Sejnowki (1988) presented the secondary structure prediction (in single protein sequence) using a neural network method. They achieved a success rate of 64.3% for a three-state model (α, β and coil). This level of accuracy, which was reproducibly obtained by other researchers (Holley & Karplus, 1989) is substantially better than the prediction accuracy from statistical approaches (Chou & Fasman, 1974; Lim, 1974, Garnier *et al.*, 1978) which is in the 50 to 56% range (Kabsch & Sander, 1983). The test set contained only proteins not homologous to those in the training set.

Muggleton *et al.* (1992) have used neural network in the prediction of all-helix domains with a learning set of 12 nonhomologous proteins. The accuracy was 81% for four different proteins. An approach for increasing rate accuracy in the prediction of α, β and turn simultaneously, a weak turn prediction would be discarded if the same segment has a strong prediction for an α-helix (McGregor *et al.*, 1989).

By creating profiles of aligned, homologous sequences, and training and testing neural networks on these rather than on individual proteins, Rost and Sander have obtained substantial improvements, with an average 3-state prediction accuracy of 72.5% on sequences not homologous to any in the training set (Rost & Sander, 1994). Although prediction accuracy may improve with the addition of more well resolved protein structures (Roman & Wodak, 1988), much of the inaccuracy in current secondary structure prediction method is believed to be due to the lack of consideration of long range interactions that arise from the unknown tertiary structure. This is a consequence of the fact that many sequences have alternative secondary structure possibilities (Kabsch & Sander, 1984; Argos, 1987; Holly & Karplus, 1991).

It has been found that basic information on protein tertiary structure such as the folding class can be helpful in improving the accuracy of secondary structure prediction (Taylor & Thornton, 1984; Kneller *et al.*, 1990; Presnell *et al.*, 1992). Four simple folding classes of protein have been defined by Levitt & Chothia (1976) :

All α has mainly α-helix secondary structure.

All β has mainly β-sheet secondary structure.

$\alpha+\beta$ has both α-helix and β - strand secondary structure segments that do not mix.

α/β has mixed or alternating segments of α-helical and β-strand secondary structure.

Several statistical methods were developed to predict whether a protein belongs to one of these classes. Kneller *et al.* found that prediction accuracy on proteins in the all α class was improved by 16% (from 63% to 79%) by using a neural network trained on similar proteins. The accuracy of β-class prediction improved by 6% (from 63% to 69%). And the accuracy on α/β did not improve and other classes were not examined (Kneller *et al* 1989).

Neural networks have yielded promising results in identifying specific tertiary folds with no experimental information besides the amino acid content and length (Duchak *et al.*, 1993). An accuracy of 87% was achieved at distinguishing protein of 4 specific fold : 4-helix bundles, barrels, nucleotide binding fold, and immunoglobulins. The folds that were tested are very different from each other in size, amino composition and helix and sheet contents.

The theme of this thesis relates to research and development of new methodologies and algorithms for the prediction of three dimensional structure of the proteins. New approaches for these predictions will focus on an information and properties of amino acid sequence , hydropathy, amino acid side chain properties, hydrophobicity and helical tendencies, as input vector applied to the computational neural network methods.

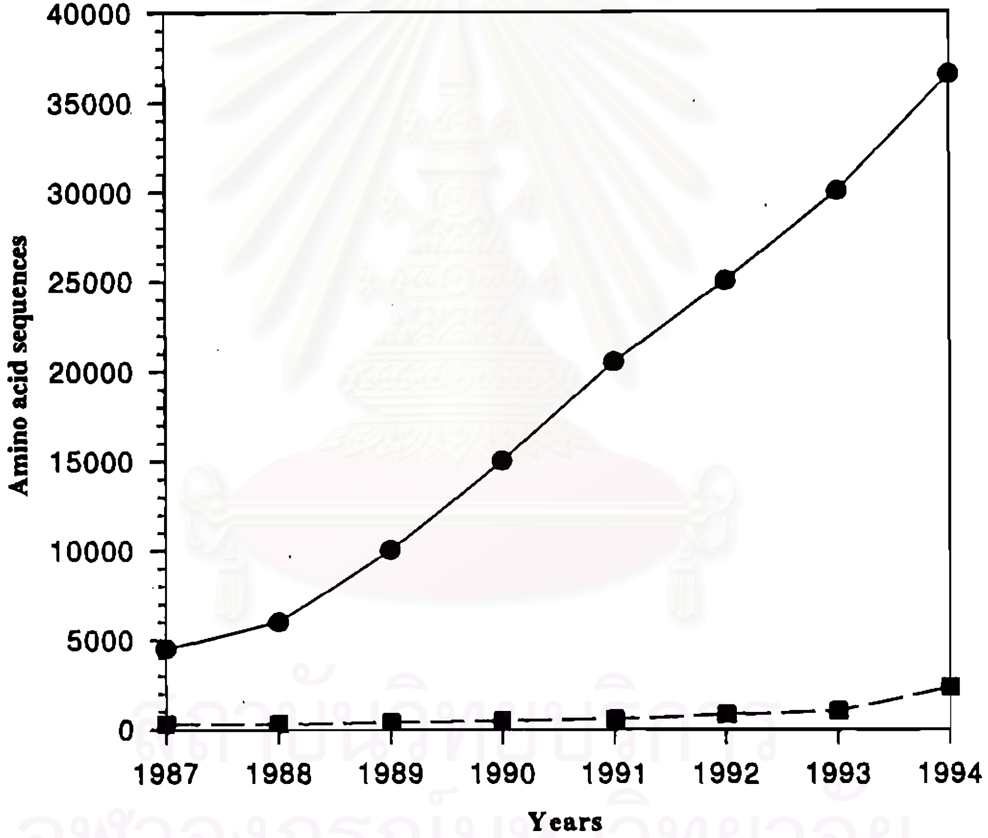

Figure 1.1 The gab between the number of protein sequences (•——•) and three dimensional structures (■---■). The graph shows the accumulation of protein amino acid sequences in SWISS-PROT (Bairoch and Boeckmann, 1993) and of protein tertiary structure in the Brookhaven Protein Databank (PDB; [Bernstein *et al.*, 1977; Abola *et al.*, 1987]) since 1987.