Part I

Introduction

Chapter 1

1.1 The need for Biological Sequence Analysis

Despite great improvements to the basic techniques of X-ray crystallography, rate-limiting step in structure determination remains the expression, purification and crystallization of target proteins. NMR techniques offer some scope for some of theses difficulties, but they are still limited with respect to the size of the proteins that can be routinely tackled (Jones, 2000). Therefore, it has not been possible to study large proteins or protein complexes in molecular details by means of routine techniques identifying individual domains and ascribing distinct functions to each.

In the past few years, the technology of sequencing has developed to stage at which the sequencing of a complete genome can be contemplated as a practical and routine possibility. The complete sequences of more than 55 genomes have been published and at least 100 more are known to be nearing completion. These projects produce large amount of sequence data lacking experimental determination of structure and biological function of predicted gene products (Kriventseva *et al*., 2001). Predicting structural and functional features from primary sequence is becoming increasingly important for many reasons. The current publicly available sequence database contains 705,144 sequences (NR database as on july 2001), while the protein databank (Bernstein *et al*., 1977; Berman *et al*., 2000) contains only 15,531 structures (PDB as on July 2001; www.rcsb.org/pdb) of which only ~2300 are 'non-redundant'. These structures belong to about 600 fold families (e.g. SCOP release 1.53; Murzin *et al*., 1995), where a fold level similarity implies conservation of over all structure. Thus, it is hoped that theoretical methods may help 'fill the gaps' in fold space.

1.2 The Structure/function paradigm

It is now well known that protein structure is much more highly conserved than protein sequence. Homologous proteins resemble each other in sequence, three-dimensional structure and usually function (Rossman and Argos, 1977; Chothia, 1984; Overington *et al*., 1990; Sowdhamini *et al*., 1998). Divergent evolution has also led to the existence of superfamilies with very low sequence identities, but very similar topologies and often related functions (Sowdhamini *et al.*, 1998). Chothia and Lesk (1986, 1987) found that structural divergence, when expressed in terms of RMS

separation of matching C_α atoms, was an exponential function of sequence divergence expressed in terms of the fraction of residues that differed between sequences. It has been shown that the fold can be transferred reliably from a protein whose structure is known to an uncharacterized sequence, when the identity between them is >20% (Devos and Valencia, 2000). It is also evident that the tertiary structure of a protein creates the means by which it functions (Jones, 2000). Precise function is not conserved below 30-40% sequence identity, but functional class is conserved for sequence identities as low as 20-25% (Wilson *et al*., 2000). These observations emphasize that determining the three-dimensional structure of a protein is a prerequisite for understanding of function. It may give clues not apparent from sequence, about distant relatives that share a catalytic mechanism or recognize same ligand for sequences sharing identities as low as 20%. In short, it can give a way to find out details of function of a protein and ease further biochemical characterization of the protein (Johnson *et al*., 1994; Jones, 2000). The protein sequences seeking attention can be divided into three categories: (1) Proteins with a known function, but no apparent relationship to protein of known structure and (2) Proteins with a known function and a distant relationship to proteins of known structure and (3) Hypothetical proteins.

It is proposed that large number of proteins come from no more than 1000 superfamilies and number of folds expected are even less (Orengo *et al*., 1994; Brenner *et al*., 1997). Also the observation that the probability of finding a gene product having an entirely new fold is less than 30% (Orengo *et al*., 1997), gives a hope of **gaining knowledge about structure (and therefore function) of a major portion of sequences deposited in sequence databases by means of sequence analysis.** This assumes, of course that the fold has already been associated with a known function. Fortunately, the vast majority of proteins with known 3D structures belong to well-characterized families for which a lot of biochemical knowledge has been collected.

In this thesis three examples of **sequence analysis** have been presented to demonstrate its power in reaching to helpful results. The goal reached is same in each case viz., starting from sequence to function, via structure. In each case, the protein domains involved, indulge in different signal transduction pathways, which gives this

work additional importance. The methods used in each case can be summarized using following flow chart (Figure 1.1). Each step of the flow chart has been discussed at length in the following portions of introduction after describing basics of protein structure. In the basics of protein structures (Chapter 2) the first the properties of peptide bond, dihedral angles and Ramachandran plot are discussed. Secondary structures and other regular conformations are defined on the basis of H-bonding patterns and positions they occupy in Ramachandran plot. Chapter 3 describes in length about the methods and databases used for this work while discuss about others in short. It also provides links to various sequence analysis services available on world wide web.

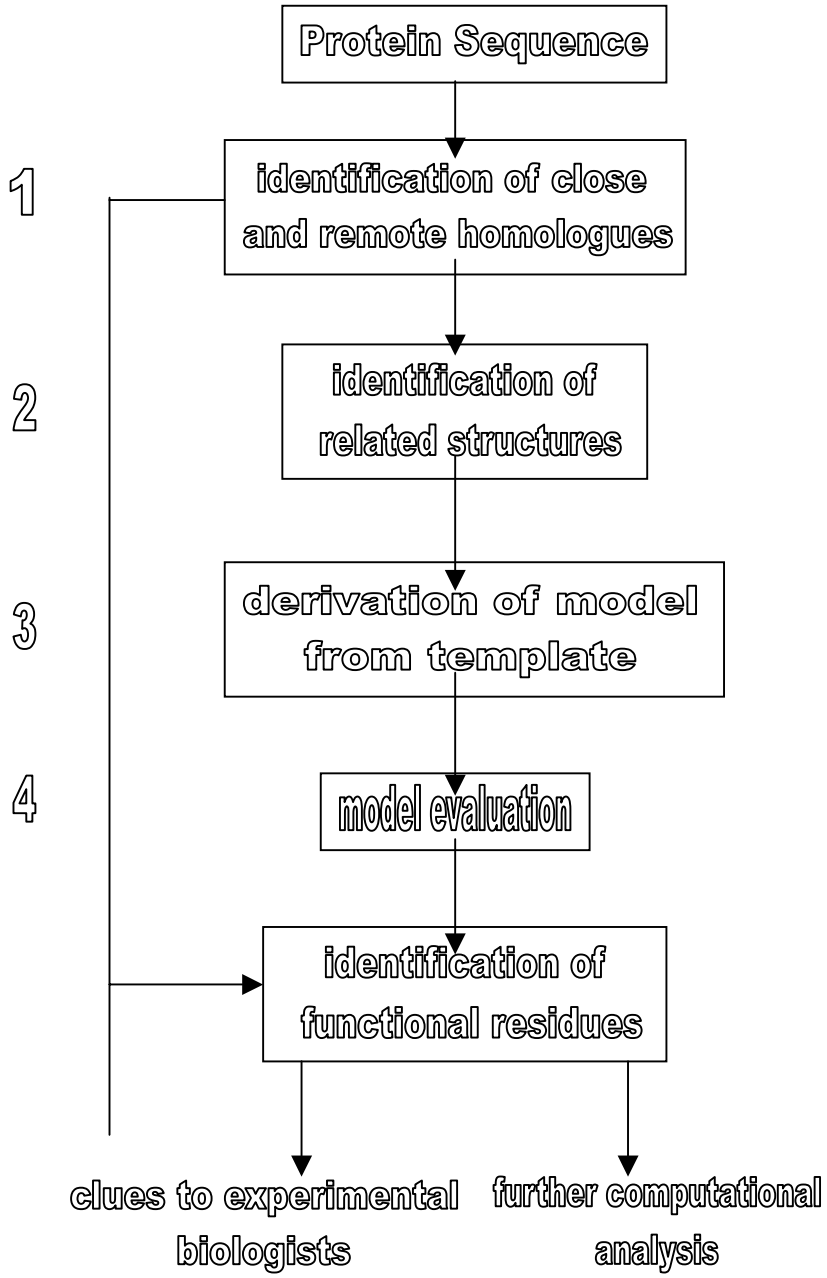

Figure 1.1 Flolw chart showing the approximate logic used to carry out the analysis project throughout this work. Every step is described in detail below.