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.

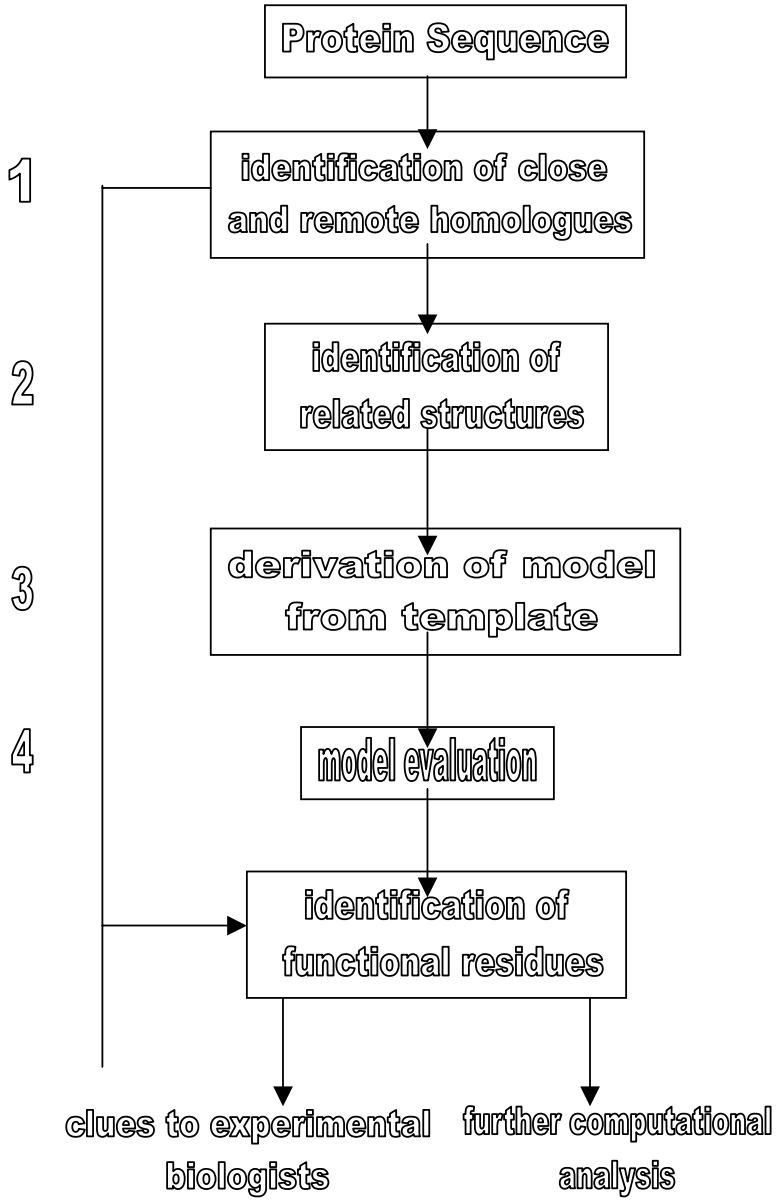

Figure 1.1 Flolw chart showing the approximate logic used to carry out the analysis project throughout this work. Every step is described in detail below.

Chapter 2

Overview of Protein Structure

2.1 Synthesis and Information Contents

It has been long recognized that life is based on morphological units known as **cells**. The formulation of this concept is generally attributed to an 1838 paper by Matthias Schleiden and Theodor Schwann, but its origins may be traced to the seventeenth century observations of early microscopists such as Robert Hooke. Most of the molecular constituents of living systems are composed of carbon atoms covalently joined with other carbon atoms and with hydrogen, oxygen, or nitrogen. The special bonding properties of carbon permit the formation of a great variety of molecules. Organic compounds of molecular weight (M_r) less than about 500, such as nucleotides, amino acids and monosaccharides, serve as monomeric subunits of nucleic acids, proteins and polysaccharides, respectively. The Deoxyribonucleic acids (DNA) are polymers of nucleotides Adenine (A), Guanine (G) Cytosine (C) and Thymine (T) while in Ribonucleic acids (RNA) Thymine is replaced by Uracil (U). DNA (and sometimes RNA) is the cell's master repository of genetic "**information**". The expression of the genetic information is a two-stage process. In the first stage, which is termed **transcription**, a DNA strand serves as a template for the synthesis of a complementary strand of RNA. In the second stage of genetic expression, which is known as **translation**, ribosomes enzymatically link together amino acids to form proteins. The order in which the amino acids are linked together is prescribed by RNA's sequence of bases, since proteins are self-assembling, the genetic information encoded by DNA serves, through the intermediacy of RNA, to specify protein structure and function. Proteins are

composed of 20 different kind of amino acids. The nucleotides from which nucleic acids are built and the amino acids from which proteins are built are identical in all leaving organisms. Consider following example. One can make a 8 unit word out of 26 letters of English alphabet, 4 different deoxyribonucletides and 20 different amino acids in 26^8 ($2.1x\ 10^{11}$), 4^8 (65,536) and 20^8 ($2.56x10^{10}$) ways, respectively. It is clear from the above example that such monomeric subsist in linear sequences can spell infinitely complex messages depends upon its length and as the information flows from DNA to protein the complexity increases rapidly (Voet and Voet, 1995).

Proteins are important as structural, functional and information career molecules. Talking biologically, proteins store and transport a variety of particles ranging from macromolecules to electrons. They guide the flow of electrons in the vital process of photosynthesis; as hormones, they transmit the information between specific cells and organs in complex organisms. Some proteins control the passage of molecules across the membranes that compartmentalize cells and organelles; proteins function in the immune systems of the complex organisms to defend against intruders; and proteins control gene expression by binding to the specific sequence of nucleic acids, thereby turning genes on and off. Proteins are the crucial components of muscles and other systems for converting chemical energy in to mechanical energy. They are also necessary for sight, hearing, and other senses. Many proteins are simply structural providing the filamentous architecture within cells and materials that are used in hair, nails, tendons, and bones of animals (Creighton, 1993).

2.2 Structural Hierarchy in Proteins

All proteins, in all species, regardless of the their function or biological activity, are polymers of the same set of 20 amino acids which are linked by covalent bonds. Amino acid sequences of the proteins can be deduced form the direct sequencing or from the DNA sequences of the related gene. Conceptually, protein structure can be considered at four levels.

Primary structure includes all the covalent bonds between amino acids and is normally defined by the peptide-bonded amino acids and location of disulfide bonds. The relative spatial arrangement of the linked amino acids is unspecified.

Secondary structure refers to regular, recurring arrangements in the space of adjacent amino acids in a polypeptide chain. There are a few common types of secondary structure, most prominent being the α helix and β conformation.

Tertiary structure refers to the spatial relationship among all amino acids in a polypeptide; it is the complete three-dimensional structure of the polypeptide. The boundary between secondary structure and tertiary structure is not always clear. Several different types of secondary structure are often found within the three-dimensional structure of a large protein. Proteins with several polypeptide chains have one more level of structure:

quaternary structure, which refers to the spatial relationship of the polypeptides, or subunits, within the protein. Understanding of protein structure, folding, and evolution has made it necessary to define two additional level between secondary structure and tertiary structure. A stable clustering of several elements of secondary structure is sometimes referred to as

supersecondary structure. The term is used to describe particularly stable arrangements that occur in many different proteins and sometimes many times in a single protein.

A somewhat higher level of structure is the **domain**. This refers to a compact region, including perhaps 40 to 400 amino acids, that is a distinct structural unit within a larger polypeptide chain. Many domains fold independently into thermodynamically stable structures. A large polypeptide chain can contain several domains that are readily distinguishable within overall structure. In some cases the individual domains have separate functions. However, important patterns exist at each of these levels of structure that provide clues to understanding the overall structure and function of large proteins.

2.3 Amino Acids and the Peptide Bond

Of the 20 amino acids usually found in proteins, 19 have the general structure as shown in Figure 2.1a-d.

a) b) c) d)

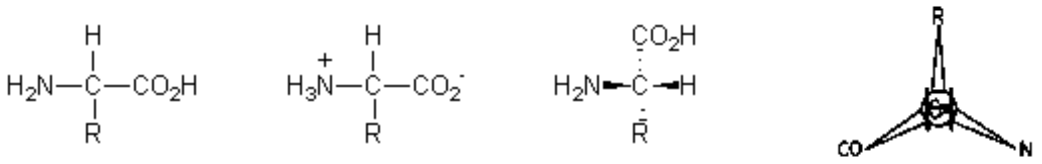

Figure 2.1: Different representations of general structure of amino acids. (a) Normal representation (b) Zwitter ionic structure (c) Geometric representation (d) The "CORN crib" for determining the handedness of an amino acid. Looking at the α carbon from the direction of hydrogen, the other substituents should read CO (carbonyls), R (side chain), and N (backbone NH) in clockwise order for a biologically appropriate L-amino acid.

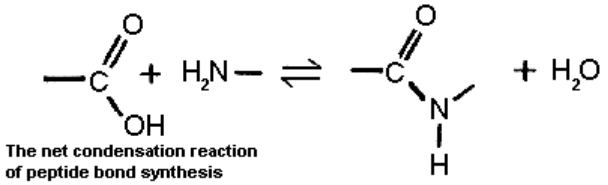

Figure 2.2 Showing net condensation reaction that results in formation of peptide bond. In the process a water molecule gets liberated.

The amino acids are linked in to proteins by the peptide bond, as in Figure 2.2, by the condensation of two amino acids. Generally, between 50 and 3000 such amino acids are linked in this way to form a typical linear polypeptide chain

2.4 Properties of Polypeptide Backbone and the Ramachandaran Plot

The backbone of the linear polypeptide chain consists of three atoms of each residue in the chain, the amide N_i, the C_i^{α}, and the carbonyl C_i', where *i* is the number of the residue, starting from the amino end of the chain.

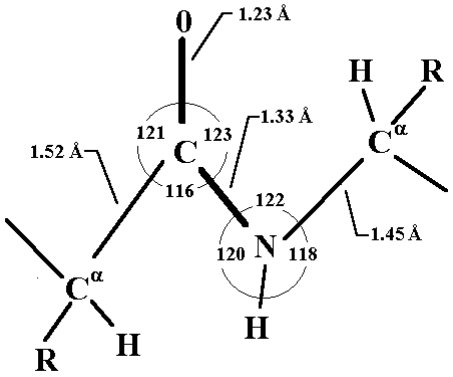

Figure 2.3: Showing the dimension of protein backbone derived from Ramachandran *et al*., 1974.

The dimension of the peptide group of a residue is given in Figure 2.3, have been derived from three-dimensional structure analysis of small peptides (Ramachandran *et al*., 1974). The presence of an asymmetric center at the C_α carbon atom, and only **L** amino acid residues, results in an inherent asymmetry of the polypeptide chain, that is important for spectral and conformational properties of polypeptides and proteins. The convention used to recognize correct L-amino acid handedness when dealing with physical models, stereo figures. Or molecular graphical displays: if one looks down on the α carbon from the direction of the hydrogen, other substituents should read "CO-R-N" in the clockwise order as shown in Figure 2.1d. In all the structures the central carbon or α carbon is bonded to an amino group, a carbonyl group, a hydrogen and an **R** group, that acts as

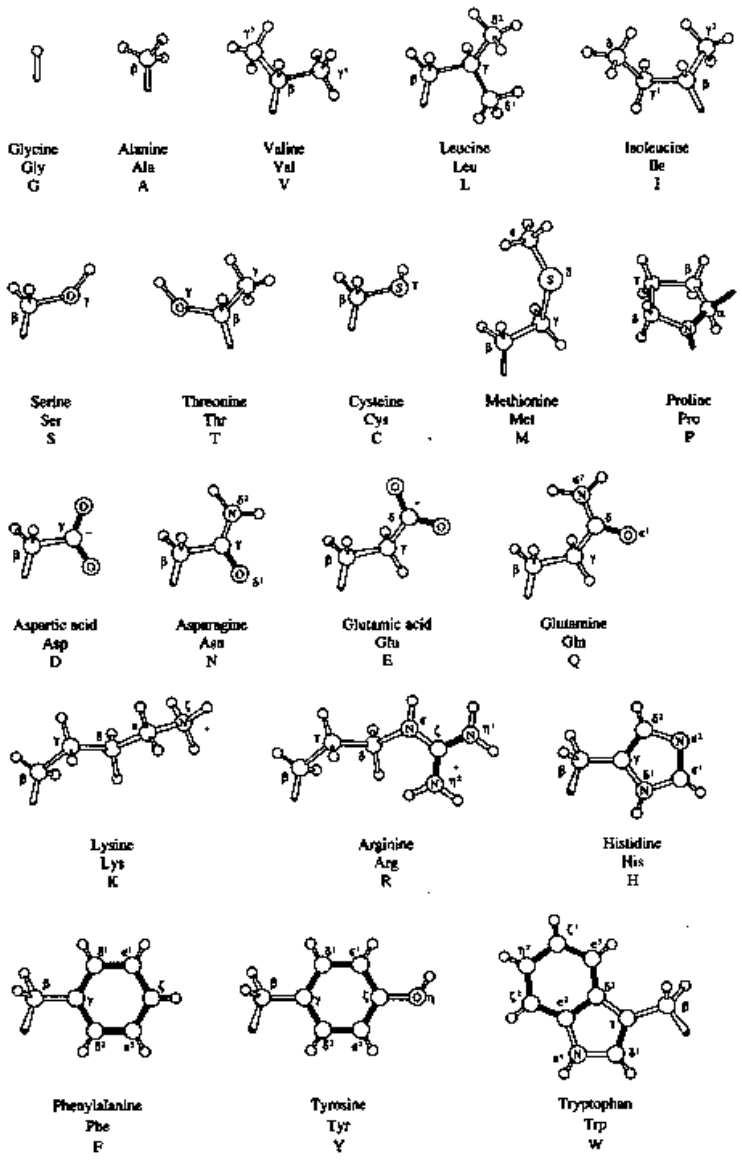

Figure 2.4 Side chains of the 20 amino acids that occur naturally in proteins. Small-unlabeled spheres are hydrogen atoms, and large unlabeled atoms are carbon atoms; other atoms are labeled. Double bonds are black, and partial double bonds are shaded. In the case of Pro, the bonds of the polypeptide backbone are included and are black. Below the name of the amino acid are the three-letter and the one-letter abbreviations commonly used. Note that isoleucine and threonine have asymmetric centers in the side chains, and only isomer illustrated is used biologically.

"side chain". The amino acids differ only in the chemical structures of the side chain **R**. The 20th natural amino acid, proline, is similar, but its side chain is bonded to the nitrogen atom to give the imino acid. Except in glycine, where the side chain is only a hydrogen atom, the central carbon atom is asymmetric and is always the **L** isomer. The side chain structure of each amino acid is shown in the Figure *2.4* with its full name, three- and one-letter codes. The central atom is designated as α, and the atoms of the side chains are commonly designated β,γ, δ, ε, and ζ, in order away from the α carbon

In principle, rotation could occur about any of the three bonds of each residue of the polypeptide backbone, but the peptide bond appears to have partial double-bonded character due to resonance. Consequently, the peptide bond length is only 1.33 Å, shorter than the usual C–N bond length of 1.45 Å, as in the C_α–N bond. It is however, longer than the value of 1.25 Å for the average C=N double bond. The peptide bond appears to have approximately 40% double-bonded character. As a result, rotation of this bond is restricted, and residues shown in Figure 2.3 have a strong tendency to be coplanar.

Resonance of the peptide bond tends to redistribute its electrons, and peptide backbone is correspondingly polar. The H and N atoms appear to have, respectively, positive and negative equivalent charges of 0.20 electron, where as C and O, respectively, have positive and negative equivalent charges of 0.42 electron. This gives the peptide bond a substantial permanent dipole moment of about 3.5 Debye units. The polypeptide backbone of the each residue contains one potent hydrogen bond donor, –NH–, and a hydrogen bond acceptor, carbonyl –CO–. This property is crucial for the polypeptide chain for three-dimensional architecture of proteins.

Two configurations of the planar peptide bonds are possible, one in which the C^α atoms are *trans*, and the other in which they are *cis* in conformation. The *trans* form is intrinsically favored energetically, probably owing to fewer repulsions between non-bonded atoms. If the residue that follows the peptide bond is Pro, how ever, its cyclic side chain diminishes the repulsions between atoms, and the intrinsic stability of the *cis* isomer is comparable to that of the *trans* isomer.

The above description indicates that the backbone of the protein is a linked sequence of rigid planar peptide groups. It is possible, therefore specify a polypeptide's backbone conformation by the dihedral angles (the angle formed between two planes) or rotation angles about C_α–N and C_α–C' bonds of each amino acid residues. A dihedral angle involves four successive atoms –A, B, C, and D–and three bonds joining them. If one look directly down the length of the central bond joining atoms B and C (the answer is the same as viewed from either end of this bond) and put the atom A at 12 on the clock face, then clock position of the far atom D reads out the dihedral angle for B–C bond. By convention, dihedral angles are assigned in the range of -180° to +180° with the clockwise direction being positive. The dihedral angle formed by C_i–N_i–$C_{\alpha i}$–C'_{i+1} is denoted as φ and generally referred as rotation angle of N_i–C_α bond (Figure 2.5). The dihedral angle formed by N_i–$C_{\alpha\ i}$–C'_{i+1}–N_{i+1} is denoted as ψ and generally referred as rotation angle of $C_{\alpha i}$–C'_{i+1} bond (Figure 2.5). The third dihedral angle is formed by $C_{\alpha\ i-1}$–C'_i–N_i–$C_{\alpha i}$ is denoted as ω or and generally referred as rotation angle of C'_i–N_i or the peptide bond (Figure 2.5).

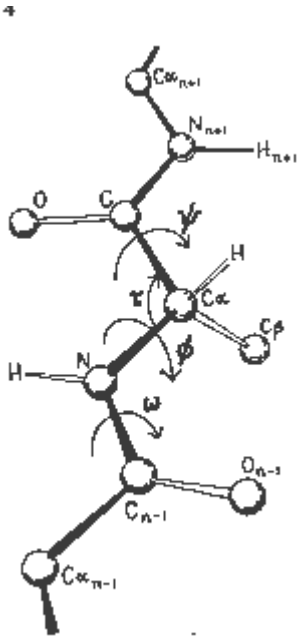

Figure 2.5 Nomenclature for the atoms of the polypeptide chain, the tetrahedral bond angle τ, and backbone dihedral angles φ, ψ, and ω.

Torisional angles of side chains are designated by χ_j, where *j* is the number of the bond counting outward from the C_α atom of the main chain. Assuming the ideality for the rest of the geometry, then three backbone dihedral angles per residue (φ,ψ, and ω) plus the side chain dihedral angles χ_j provides complete description of the local conformation. In practice, just φ and ψ suffice for the main chain, because the partial double bond character of the peptide bond keeps ω very close to flat. ω has a monomodel distribution with a mean of 180° and a small standard deviation of approximately 6°, which is the fully extended or *trans* conformation (Creighton, 1993). The curled up *cis* conformation of ω at or near 0° is observed about 10% of the time of proline and extremely rare for any other kind of amino acid. Hence, proline is the only exception in where ω distribution is bimodal.

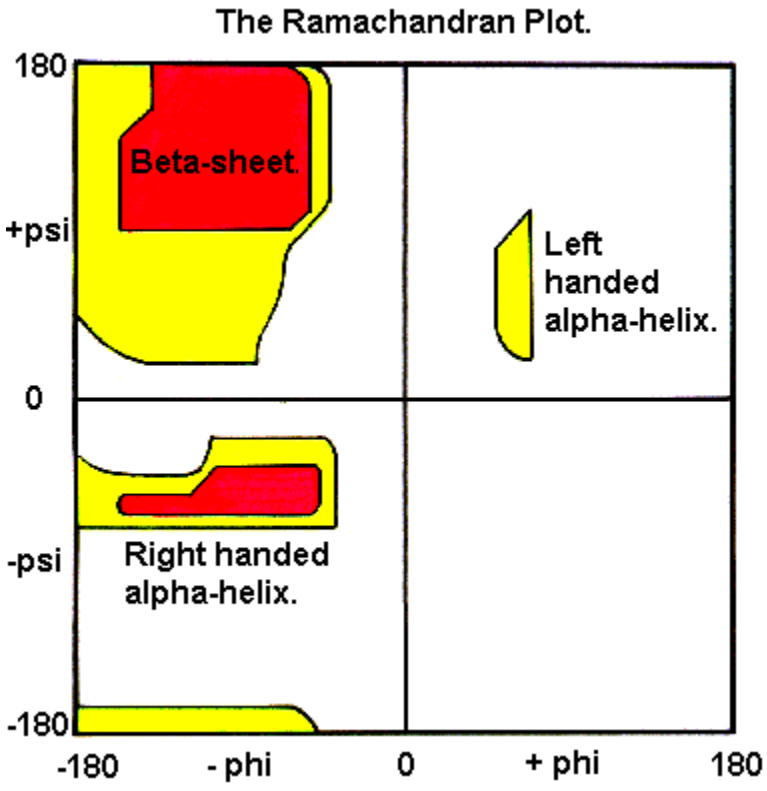

Figure 2.6 The positions of commonly found regular conformations of the proteins on a Ramachandran plot. The shown conformations are right and left-handed α helices and β sheet.

Since φ and ψ form a virtually complete description of the backbone conformation, a two dimensional plot of them is an important type of representation (Ramachandran and Sasiekharan, 1968). The plot is know as Ramachandran plot (Figure 2.6). Ramachandran plot can be used to illustrate properties of repeating conformations, single residues, or two successive residues and in general for studying the conformational properties. The regions of φ,ψ space, are generally named after the conformation the conformation that results, if they are repeated.

2.5 Definition of Secondary Structures

Main chain conformation can be classified in to secondary structures using Kabsch and Sander definitions (Kabsch and Sander, 1983). The distribution of φ,ψ pairs obtained from the protein structures in the protein structures in the protein databank shows six different picks. They correspond to right-handed α-helix (A), idealised β-strand (B), polyproline conformation (P), the ε region accessible primarily to Gly residues with positive φ angle (G), left-handed α-helix (L) and extended conformation. (E). These six peaks represent six different conformation states of the main chain of a particular residue.

The major conformations on the Ramachandran plot are the right-handed α helical cluster in the lower left near -60°, -40°; the broad region of extended β strands in the upper left quadrant (centered around -120°, 140°); and sparsely populated left-handed α-helical region in the upper right around +60°, +40° (Figure 2.6). Other regular conformations, like 3_{10}-helix (-49°, -26°), π-helix (-57°, -70°), polyproline1 (-83°, +158°), polyproline2 (-78°, +149°) and polyglycine2 (-80°, +150°) however do occur in proteins. The approximate mean values and standard deviations of the main chain dihedral angles in the classes are listed in Table1. Vacant areas in the Ramachandran plot (Figure 2.6) are the

conformations that place the atoms unfavorably close together within the dipeptide unit. The asymmetry of the plot results from the collisions of the C_β.

	Mean (°)		Standard deviation (°)		Residue per turn
	φ_i	ψ_i	$\sigma_i(\varphi)$	$\sigma_i(\psi)$	
A (α-helix, R)	-65	-41	15	15	3.6
B (β-strand)	-130	135	15	20	2.0
P (polyproline)	-65	140	15	15	3.0
G (Gly with +φ)	60	40	10	10	NA
L (α-helix, L)	90	-10	15	10	3.6
E (extended)	130	180	25	25	NA

Table 1 Showing the mean dihedral angles for defining a secondary structure. The table is modified from Sali *et al*., 1993.

2.6 Hydrogen Bonding

One of the more remarkable properties of the repetitive secondary structures observed in proteins is that the optimum φ, ψ values and the permissible range for good long-range H-bonding and steric fit are close to the optimum and range favorable for dipeptide conformations.

The dual hydrogen bonding capacity of the backbone peptide group is a persuasive influence on the protein structure. Although H-bonds are weak, non-covalent interactions, they are fairly directional and specific. Since each peptide can form a bond in both the

directions, the co-operative effect of a network of such interactions can hold the polypeptide together in a strong and specific network.

Hydrogen bond involves an electrostatic attraction, either between two actual or between dipoles and they also involve the sharing of a proton. The group on one side of the H-bond is the "donor" D (usually, in proteins, a nitrogen or a water but sometimes an OH), which has a hydrogen it can contribute to the bond. The other group is the "acceptor", A, with accessible pair of electrons (usually a CO or water, but sometimes an unprotonated N or the backside of an OH). The optimum distance for a strong H-bond is about 3Å between D and A or 2Å between H and A. Angular criterion is important for hydrogen bonding.

2.7 Secondary Structures or Repetitive Structures

Patterns of main-chain hydrogen bonding, combined with repeating values of φ, ψ angles define secondary structures in proteins. The β-structures involves repeating patterns of H-bonds between distant part of the backbone, whereas helices involve repeating patterns of local H-bonding.

2.7.1 Helices

Helices are predominant, recurring form of secondary structure. The number of residues (*i)* and atoms (*x*) per single turn defines each type of helix. Two of the first helices hypothesized by Pauling and Corey in 1951, to occur in proteins were the α-helix or 3.6_{13}-helix, where $i = 3.6$ and $x = 13$, and the γ-helix or 5.1_{17}-helix, where $i = 5.1$ and $x = 17$. In conjunction with α- and γ- helices of Pauling and Corey, Donohue hypothesized the 2.2_7-helix, the 3_{10}-helix, the 4.3_{14}-helix and the 4.4_{16}-helix (Donohue, 1953). Of these hypothesized structures α-helix and 3_{10}-helices are reported in numerous reported protein structures, with α-helix being most abundant. Of other helices hypothesized by Donohue

only π-helix has been reported and the occurrence is rare. The main features of the helical structures reported by known protein structures are as follows.

The right-handed α-helix (Figure 2.7b) is the best known and most easily recognized of the polypeptide regular structures, formed by repeated H-bonds between the CO of residue *i* and NH of residue *i+4*, with repeated φ, ψ values near -60º, -40º. Though the preferred values for φ and ψ angles differs with different analysis. The α-helices observed in actual protein structures are nearly always right-handed both because of the cumulative effect of a moderate energy difference for each residue and even more because each C_β would collide with the following turn of a left-handed α-helix.

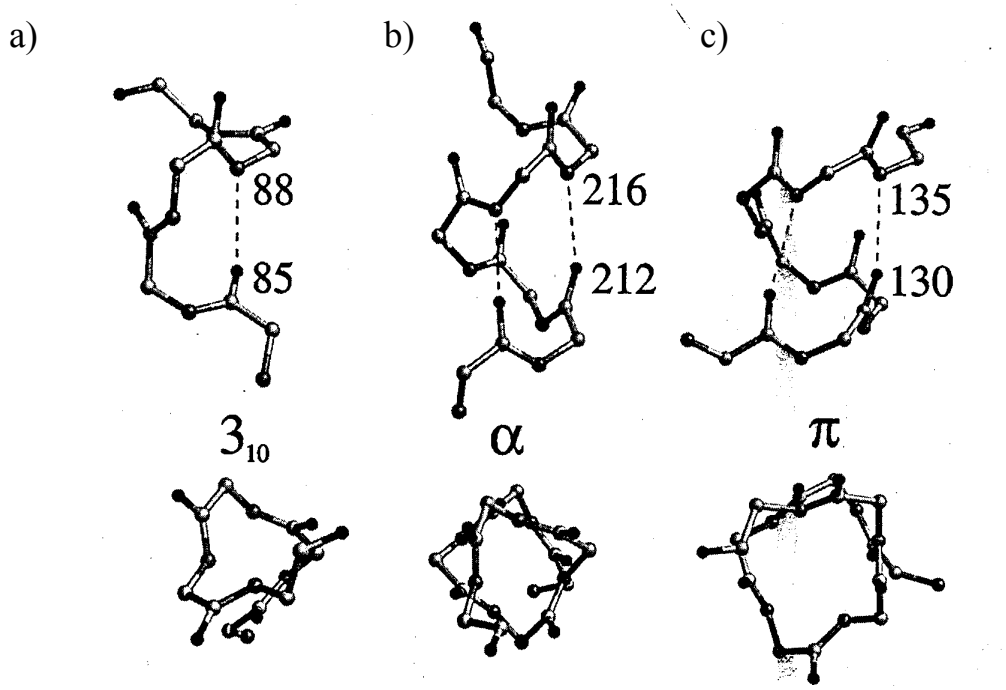

Figure 2.7: Helical structures witnessed within protein crystal structures. They are (a) 3_{10}-helix (b) α helix and (c) π helix. Figures on the top show the side view and that on the bottom shoes the top view of the helices. Residue numbering is random and shows the number of residues per turn in case of each helix. Hydrogen bonds between the backbone carbonyl oxygen and the backbone nitrogen are represented as dashed lines in the figures.

All atoms have been colored by type, where light gray = carbon, dark gray = nitrogen and black = oxygen.

The H-bonds in a α-helix are nearly parallel to the helix axis, with the CO all pointing towards the C-terminal end. Each peptide is tilted slightly, however, so that all the oxygen atoms point a bit outward. The β carbons do not extend radially out from the α carbons but make a clockwise pinwheel shape with Cβ nearly in the plane of the preceding peptide. The pitch, or repeat, of an ideal α helix is 3.6 residue per turn. For that pitch, the rise per residue along the helix axis is 1.5 Å, or 5.4 Å per turn. Real helices match this value quite well; however, a difference in average pitch of 5% (between, say, 3.5 and 3.7 residues per turn, which is well within the common range of variation) produces an offset of an entire residue by the end of a typical four or five turn helix. That 5% difference makes a trivial change in φ, ψ angles but has a substantial effect on side chain packing.

Variations on the helices come when the chain is either more tightly or more loosely coiled. Helices with hydrogen bonds to residues *i+3* and φ, ψ values near to -70°, -5° are designated as 3_{10} helix (Figure 2.7a). The 3_{10} helix is more tightly wound than α-helix and it has very distinctive triangular appearance in the end view. In the 3_{10} helix the α carbons on the successive turns are exactly in line with one another since there are an integral number of residues per turn; this makes the H-bond quite tilted relative to the helix axis. In contrast, the non-integral pitch of a α-helix lines up a CO on one turn with NH on the next to make parallel H-bonds, and α-carbons does not line up. The H-bond geometry and van der Waals interactions between successive turns are not quite as favorable in3_{10} helix, and long stretches are rare. The major importance of 3_{10} helix is that it very frequently forms the last turn at the C terminus of a α-helix and it is fourth most common structure found in proteins. Helices with hydrogen bond to *i+5* and φ, ψ values -57°, -70° are termed as π helices (Figure 2.7c). Their occurrence in structures is rare and till date only ten π helices are reported in the literature (Weaver, 2000). In each case the occurrence of the π helix was correlated with function. The conformation of π helix has

been postulated to be disfavored for three reasons: (1) the dihedral angles are unfavorable (Low and Greenville-Wells, 1953; Ramachandran and Sasiekharan, 1968); (2) the 1Å hole at the center is wide enough to create a loss of van der Waals interactions, but too narrow to accommodate the water molecule for compensation, and (3) four residues need to be correctly aligned to allow collinear *i* to *i+5* hydrogen bond (Rohl and Doig, 1996). In the case of α and 3_{10} helices all main-chain H-bonding groups within the body of the helix are satisfied by the secondary structure formation. Each end produces three unsatisfied groups that often H-bonds to solvent, especially the open carbonyls at the C terminus. Very frequently, one of the free NHs nears the N-terminus H-bonds to the side chain of N-cap residue.

Pro residues are not ideally suited for either α-helix or β-sheet conformations. Poly(Pro) forms other regular conformations known as poly(Pro) I and II. Proline residues are special in permitting both *cis* and *trans* peptide bonds, and the two forms of poly(Pro) differ in this respect. Poly(Pro) I contains all *cis* peptide bonds whereas form II has all *trans* (Sasiekharan, 1959; Creighton, 1993). The values of φ and ψ are very similar for both, but form I is a right-handed helix with 3.3 residues per turn, whereas form II is a left-handed helix with 3.0 residues per turn. The values of φ (-83° and -78° for forms I and II, respectively) are compatible with that dictated by cyclic Pro side chain. The values for ψ are (+158 and +149 for forms I and II, respectively) constrained by steric repulsions and very similar in both the cases. Gly residue, owing to the fact that it lack the side chain, have unique conformational flexibility, and poly(Gly) likewise forms two regular conformations, designated as I and II. The former has a β-sheet conformation; the later is a helix with three residues per turn like that of poly(Pro) I.

2.7.2 β-sheet

After the α-helix the second most regular and identifiable secondary structure is the extended β strand (Pauling and Corey, 1951) with φ, ψ values in the upper left quadrant of the plot near -120°, 140°. In the extended β strand, the polypeptide backbone is fully

extended and it has 2.0 residues per turn and a translation of 3.4 Å per residue. The backbone H-bonding groups are again completely satisfied within the body of β-sheet, but since the H-bonds go from one strand to another, β structure is inherently less local and modular than helices (Chou *et al.*, 1983). As a result, the primitive unit of β structure is not the individual β strand but the β strand pair, which can be hydrogen bonded in either parallel or anti-parallel arrangement with close to optimal geometry and dipole

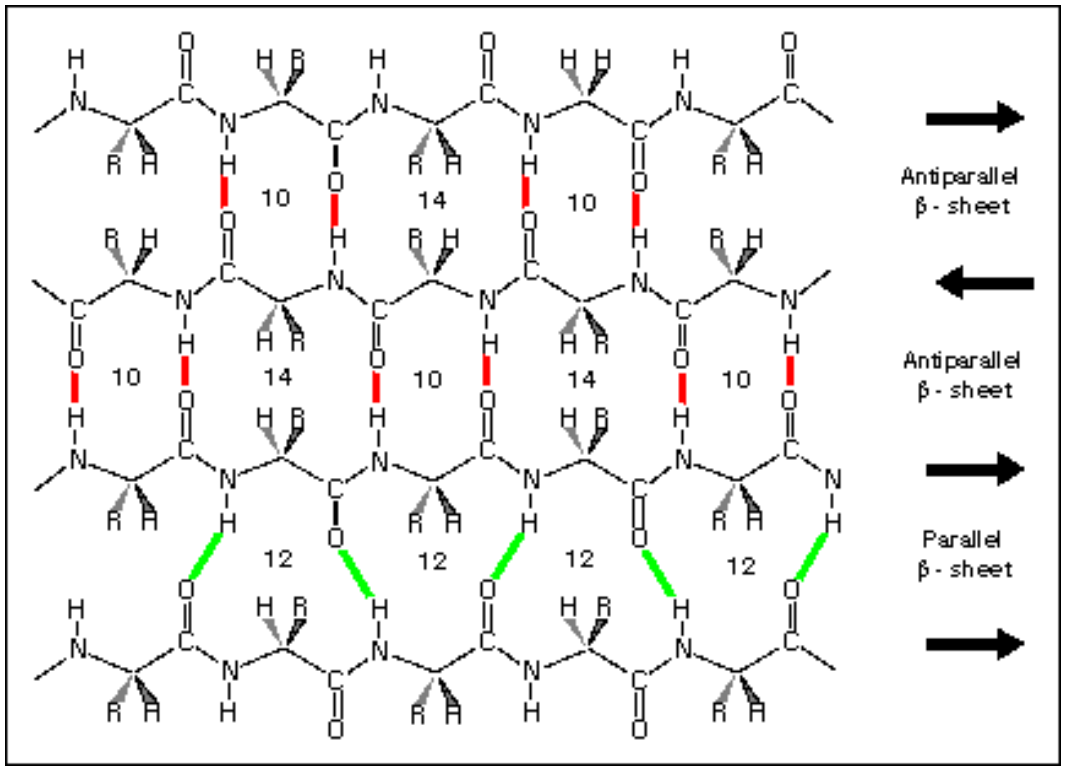

Figure 2.8 Showing the hydrogen bonding patterns in parallel and anti-parallel β- sheet structures. The direction of polypeptide backbone is marked with arrows and hydrogen bonds are shown with solid lines. The figure illustrates narrow and wide pairing of H-bonds and the side-chain alternation above and below the plane in anti-parallel β-sheet and evenly spaced but alternately slanting H-bonds in parallel β-sheet.

moments of the strands aligned favorably. Adjacent strands can be either parallel or anti-parallel (figure 2.8), and the stereochemistries of the strand in both the cases are slightly

different. For anti-parallel β sheet, the relationship between adjacent strands is a two-fold axis perpendicular to the sheet, with the H-bonds perpendicular to the strands and alternating between a closely spaced pair and a widely spaced pair. In parallel β sheet, the H-bonds are evenly spaced but alternatively slant forward and backward, and the relationship between adjacent strand is a translation. The side chains on β strands extend approximately perpendicular to the plane of H-bonding. Along the strand they alternate from one side to the other, but on adjacent strand they are in register. For anti-parallel β sheets typically one side is buried in the interior and the other side is exposed to solvent, so that the amino acid types tend to alternate hydrophobic and hydrophilic. Parallel sheets, on the other hand, are usually buried on both sides, so their central sequences are highly hydrophobic, and hydrophilics concentrate at the ends. For both types of β structure, edge strands can be much more hydrophilic than central strands (Fasman, 1989).

Distinguishing these characteristic patterns can be of some help in secondary structure prediction and is clearly important for working towards probable tertiary structures. The usefulness of this results from a strong tendency for β sheets to be either pure parallel or pure anti-parallel. Mixed sheets occur, but not at anything like random expectation. More efforts are usually needed for prediction of the sheets.

The most prevalent local disruption in a sheet is the β bulge (Richardson *et al*., 1978). A β bulge can be thought of as an insertion of an extra residue into one strand, so that between a pair of H-bonds there is one residue on the normal strand but two residues on the bulged strand. Bulges are common in anti-parallel β structure but rare in parallel β. Usually they are located between a close pair of H-bonds rather than a wide pair. The extra residue puts the hydrophobic-hydrophilic side-chain alternation out of register across the bulge, an effect that is sometimes recognizable in the sequence. To accommodate the surrounding H-bond pattern, usually one of the two bulge- strand residues stays close to normal β-conformation while the other is close either to α-helical conformation (a "classical" bulge) or close to left-handed 3_{10} conformation (a "G-1"

bulge). The single residue on the opposite strand is usually near polyproline conformation in order to match greatly accentuated right-handed twist produced by a β bulge. Bulges can mitigate the damage done by single residue insertion or deletions in β strands, at least when they occur near an end or an edge of the β sheet (Chan *et al*., 1993).

2.8 Non-repetitive Structure: Turns, Connections and Compact Loops

The secondary structures (described above) are one in which the φ,ψ angles repeats for each consecutive residues. Large portions of protein structure, however, are made up of well-ordered but nonrepeating conformations. These have often been referred to as "coil" or even "random coil", which unfortunately has connotation of disordered, mobile, unfolded chain. Nearly one third of the residues of globular proteins are involved in tight turns that reverse the direction of polypeptide chains at the surfaces of the molecules and make possible overall globular structure. Turns have also been implicated in molecular recognition (Rose *et al*., 1985) and in protein folding. Because of their prevalence, these reverse turns or loops are frequently classified as a third type of secondary structure.

Various types of reverse turns occur, involving different numbers of residues and depending upon which type of secondary structure they link. The best characterized are the β hairpins that link adjacent strands in antiparallel β-sheet. If only one residue is not involved in the H-bonding pattern of the sheet, there is a γ –turn, of which two types are possible. This very tight turn requires unfavorable geometry for the adjacent hydrogen bond of the β-sheet and unusual values of φ,ψ in the central residue of the turn. More common are β turns, in which two residues are not involved in the hydrogen bonding of the β-sheet; the two residues on either side of the non-hydrogen-bonded residues are included in the β turn, which, therefore, defined by four residues at the positions designated *i* to *i+3*. The existence of three ideal β turns, designated as types I, II and III,

was predicted by Venkatachalam (1968) on the basis of allowed polypeptide geometry with planar *trans* peptide bonds. Mirror images of backbone -but not the side chains, occur in variants I', II', III'. There have been a lot of efforts to classify the β turns and loops in general. Themean dihedral angles for γ turns and β-turn types are tabulated in Table 2. Loops have been analyzed and classified according to various structural properties and relationships, amog them main chain conformation, size, inter C_α distances, hydrogen bonding patterns, orientation, and type of secondary structure flanking the loop (Donate *et al*., 1996 and references within). A recent automated classification of conformational clusters and consensus sequences for the protein loops have been derived from a non-redundant data set by computational analysis (Oliva *et al*., 1997).

Turn type	Ramachandran nomenclature[a]	Mean dihedral angles[b] φ(*i*+1)	ψ(*i*+1)	φ(*i*+2)	ψ(*i*+2)
γ turn[c]					
Classical		70 to 85	-60 to -70		
Inverse		-70 to -85	60 to 70		
β turns					
I	$\alpha_R\alpha_R$	-64(-60)	-27(-30)	-90(-90)	-7(0)
I'	$\alpha_L\alpha_L$	55(60)	38(30)	78(90)	6(0)
II	$\beta_{\gamma L}$	-60(-60)	131(120)	84(80)	1(0)
II'	$\varepsilon\alpha_R$	60(60)	-126 (-120)	-91(-80)	1(0)
III[c]		-60	-30	-60	-30
III'[c]		60	30	60	30
IV		-61	10	-53	17
VIa1	$\beta\alpha_R$	-64 (-60)	142(120)	-93(-90)	5(0)
VIa2	$\beta\alpha_R$	-132 (-120)	139 (120)	-80(-60)	-10(0)
VIb	ββ	-135(-135)	131(135)	-76(-75)	157(160)

Table 2 Mean dihedral angles for γ turns and β-turn types derived from Crighton (1993) and Hutchinson *et al*., (1994).

[a] Ramachandran nomenclature for turn types as in Wilmot and Thornton (1990). The nomenclature describes the region of the Ramachandran plot occupied by residues i+1 and i+2 of the turn.

[b] The idealized φ, ψ values as determined by Lewis *et al*.(1973) are given in the parenthesis after the averaged values determined from dataset of Thornton *et al*., (1994).

[c] Values taken from Crighton T.E. (1983).

Loops have been classified into five types (α-α, β-β links, β-β hairpins, α-β and β-α) according to the secondary structures they embrace and in to total 56 classes (9 α-α, 11 β-β links, 14 β-β hairpins, 13 α-β and 9 β-α) were identified with consensus Ramachandran angles in the loops and consensus sequence patterns for each class. However, still the amino acid sequences of the loop region do not provide a fingerprint that can be used to identify the presence of a loop of a particular conformation anywhere in a protein sequence. However, if the position and nature of the neighboring β-strands and helices are known or suspected on the basis of the known three-dimensional structure of a homologue, or form a reliable secondary structure prediction, then the particular conformation of the connecting peptide may be identified by comparison with sequence templates or patterns that characterize loop classes. These can be useful in comparative modeling (Greer, 1980; Thornton *et al*., 1988; Sibanda *et al*, 1989; Overington *et al*., 1990; Topham *et al*., 1993) as well as in suggesting conformations of super-secondary motifs where a satisfactory structure prediction has been performed (Donate *et al*., 1996).

2.9 Amino acid Residues

The 20 different amino acids possess a variety of chemical properties. This variety is greatly enhanced when the various groups are combined in various sequences in a single molecule, which gives a protein properties far beyond those of simpler molecules. The chemical properties of a protein molecule are far more complex than the sum of the properties of its constituent amino acids but understanding side chain properties can be a

good beginning towards it. Side chains are divided and discussed briefly according its various properties in **Appendix A**. However, it should be mentioned that residues in biologically active proteins may have chemical and physical properties very different from those described. Amino acids can be classification as shown in the Venn diagram of Figure 2.9 (Taylor, 1986a,b). Thus we have seen that the amino acid properties and their preferences of staying in a particular kind of environment is the one that determine the protein secondary structures and may be the tertiary structure.

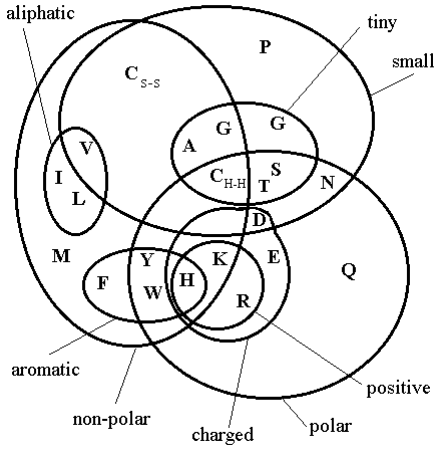

Figure 2.9 Venn diagram showing the classification of amino acids according to physical properties of their side chains (Taylor, 1986a,b).

Assemblies of a number of secondary structure elements, including the connecting loops, that have been observed often enough that they are becoming recognized as another level of structure, termed as super-secondary structures or motifs. These structures are a higher level of structure than secondary structure but does not constitute entire structural domains. However, description of super-secondary structures and structural domains is out of the scope of this thesis. For the description of recurring super-secondary structures please see Rossmann and Argos, (1981); Branden and Tooze., (1991) and Sowdhamini *et al*., (1992) etc. For description of protein structural domains and their classification

please see Sowdhamini *et al*., (1995); Sternberg *et al*., (1995); Sowdhamini *et al*., (1996) Orengo *et al*., (1997); and Holm and Sander, (1998).

Chapter 3

Biological sequence analysis

The steps outlined in the Figure1.1 are discussed in this chapter. Each step contains many sub steps, and there may have different approaches known to tackle the same problem. The methodologies that are used in this thesis for deriving results are described in details.

3.1 Identifying Close and Remote Homologues to the Query

Nature is a tinkerer and not an inventor. New sequences are adapted from pre-existing sequences rather than invented *de novo* (Jacob, 1977). This is very fortunate for computational sequence analysis, since discovery of sequence homology (recognition of significant similarity) to a known protein or family of proteins often provides the first clues about the function of a query sequence (Altschul *et al*., 1990). When the homologue is encountered the information about structure and/function can be transferred to query sequence by *homology*. Homologous proteins are defined as one that shares clear evolutionary relationship (or a common ancestor) with each other, while remote homologues are one in which the evolutionary relationships can not be detected at the first glance (e.g., using sequence similarity) due to divergent evolution. Following flowchart (Figure 3.1) summarizes ways of identification of functional and structural similarity.

Well-curated databases assume prime importance as sequence analysis is totally based upon the quality of the databases. An error in database may progress by repeated copying

of annotations between similar sequences. Some of the important publicly available sequence and structure databases are listed below.

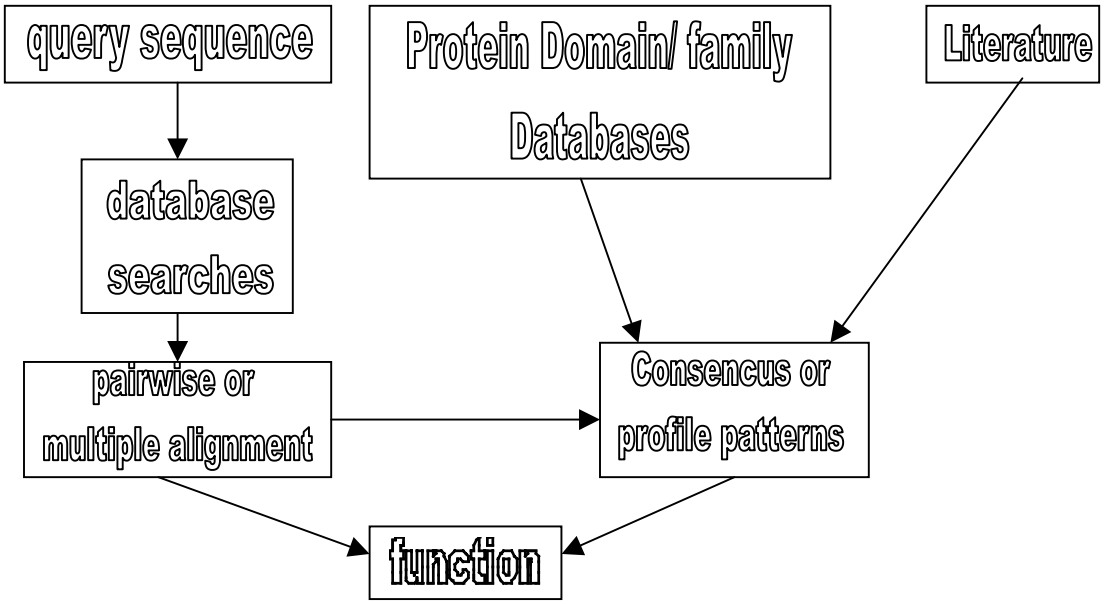

Figure 3.1 Identification of close and remote homolouges of query sequence by searching databases of deposited sequences and profiles. Plese see text for clear discussion.

3.1.1 Protein Sequence Databases

- SWISS-PROT http://www.expasy.ch/sprot
- TrEMBL http://www.expasy.ch/sprot
- PIR http://pir.georgetown.edu
- Entrz protein (NRDB) http://www.ncbi.nlm.nih.gov:80/entrez/
- OWL http://www.leeds.ac.uk/bmb/owl/owl.htm
- GenPept http://www.ncifcrf.gov/pub/genpept/

3.1.1.1 SWISS-PROT (Bairoch and Apweiler, 2000)

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria:

Annotation: In SWISS-PROT, as in most other sequence databases, two classes of data can be distinguished. First are the core data and the annotation. For each sequence entry the core data consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein). The annotation consists of the description of the following items: Function(s) of the protein, post-translational modification(s), domains, secondary structure, quaternary structure, similarities to other proteins, disease(s) associated with deficiencies in the protein, sequence conflicts and variants, etc. Systematic recourse both to publications other than those reporting the core data and to subject referees represents a unique and beneficial feature of SWISS- PROT.

Minimal redundancy: In SWISS-PROT all possible data are merged so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

Integration with other databases: Users of biomolecular databases are provided with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections. SWISS-PROT is currently cross-referenced with 30 different databases. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT.

3.1.1.2 TrEMBL (Bairoch and Apweiler, 2000)

It consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL Nucleotide Sequence Database, except the CDSs already included in SWISS-PROT.

3.1.1.3 PIR (Barkar *et al*., 2000)

The Protein Information Resource, is the most comprehensive and expertly annotated protein sequence database in the public domain, aiming to provide timely and high quality annotation and promote database interoperability. PIR employs rule-based and classification-driven procedures based on controlled vocabulary and standard nomenclature and include status tags to distinguish experimentally determined from predicted protein features. The database contains about 200000 non-redundant protein sequences, which are classified into families and superfamilies and their domains and motifs identified. Entries are extensively cross-referenced to other sequence, classification, genome, structure and activity databases. The PIR web site features search engines that use sequence similarity and database annotation to facilitate the analysis and functional identification of proteins.

3.1.1.4 Entrez (on NCBI)

It is a search and retrieval system have been compiled from sources, including SwissProt, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq.

3.1.2 Protein Structure Databases

- PDB http://rcsb.org/pdb
- NRL3D http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html
- MODBASE http://pipe.rockefeller.edu/modbase/index.shtml

3.1.3 Protein Family/ Domain Databases

- PFAM http://www.sanger.ac.uk/Pfam/
- PRODOM http://protein.toulouse.inra.fr/prodom.html
- PROSITE http://www.expasy.ch/sprot/prosite.html

- BLOCKS http://www.blocks.fhcrc.org/
- SMART http://smart.embl-heidelberg.de/
- DOMO http://www.infobiogen.fr/~gracy/domo
- PRINTS http://www.biochem.ucl.ac.uk/bsm/dbbbrowser/PRINTS/PRINTS.html
- PROFILESCAN http://www.isrec.isb-sib.ch/software/PFSCAN_form.html

3.1.3.1 PFAM (Bateman *et al*.., 2000)

Pfam is a database of protein domain families. Pfam contains curated multiple sequence alignments for each family, as well as profile hidden Markov models (profile HMMs) for finding these domains in new sequences. Pfam contains functional annotation, literature references and database links for each family. There are two multiple alignments for each Pfam family, the seed alignment that contains a relatively small number of representative members of the family and the full alignment that contains all members in the database that can be detected. All alignments use sequences taken from pfamseq, which is a non-redundant protein set composed of SWISS-PROT and TrEMBL. The profile HMM is built from the seed alignment using the HMMER package (Durbin *et al*., 1998), which is then used to search the pfamseq sequence database. All the matches found above the curated thresholds are aligned using the profile HMM to make the full alignment. The Pfam WWW servers can present the domain architecture of a protein graphically as 'beads on a string' with a color-coded and hyperlinked bead for each domain. For a fine-grained analysis of the evolution of domain architectures, a Java tool displays the graphical domain schematics of each sequence connected in an evolutionary tree.

3.1.3.2 PRODOM (Corpet *et al*., 2000)

The rapid growth of primary sequence databases makes it more and more difficult to comprehend the ever increasing diversity of known proteins. One major underlying difficulty is that many proteins exhibit a combinatorial arrangement of domains, which

makes it desirable to develop databases and tools to describe proteins at an intermediary level of structure, in terms of domain arrangements. The ProDom database was designed with this explicit purpose, with particular emphasis on the user interface. Domains are detected in an automatic process that uses sequence similarities between homologous domains of SWISS-PROT and TrEMBL (Bairoch and Apweiler, 2000) sequences using PSI-BLAST (Altschul *et al*., 1997). ProDom `domains' thus essentially reflect protein subsequences conserved in various proteins. To increase the number of these expert-validated families, the curated part of Pfam (Bateman *et al*., 2000) is used: the seed alignments of Pfam-A families were added to the list of 21 ProDom expert-validated multiple alignments and used to build new ProDom families with the PSI-BLAST program. An interactive graphical interface is available to allow for easy navigation between schematic domain arrangements, multiple alignments, phylogenetic trees, SWISS-PROT entries, PROSITE patterns (Hoffman *et al*., 1999), Pfam-A families and 3-D structures in the PDB (Bernstein *et al*., 1977; Berman *et al*., 2000). Alignments and trees can be reduced or developed to facilitate the analysis of sequence relationships within large domain families. New sequences can be searched against ProDom and aligned with existing domain families, and modeled on the basis of homologous domains in the PDB.

3.1.3.3 PROSITE (Hoffman *et al*., 1999)

PROSITE is a database of protein families and domains. It is based on the observation that, while there is a huge number of different proteins, most of them can be grouped, on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. It is apparent, when studying protein sequence families, that some regions have been better conserved than others during evolution. These regions are generally important for the function of a protein and/or for the maintenance of its three- dimensional structure. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated

proteins. A biologically significant patterns and profiles formulated in such a way that with appropriate computational tools it can help to determine to which known family of protein (if any) a new sequence belongs, or which known domain(s) it contains. PRINTS (Attwood *et al*., 1999) is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs: the database thus provides a useful adjunct to PROSITE.

3.1.4 For Text Based searches in Sequence Databases

- Entrez at NCBI http://www.ncbi.nlm.nih.gov/Entrz/
- SRS at EBI http://srs.ebi.ac.uk/
- WWW-QUERY http://pbil.univ-lyon1.fr/
- ExPASy http://www.expasy.ch/sprot/
- DBGET http://www.genome.ad.jp/

3.1.5 Sequence Alignment and Detection of Similarity

The concept of alignment is crucial for understanding the sequence searching procedures from known databases. The most basic sequence analysis task is to ask if two sequences are related and this is usually done by first aligning the sequences (or parts of them) and then deciding whether that alignment is more likely to have occurred because the sequences are related or just by chance. When sequences are compared, in essence, we are looking for the evidence that they have diverged from a common ancestor by a process of mutation and selection. We look for a series of individual characters or character patterns in the same order in the sequences. The key issues are: (1) what sorts of alignment should be considered; (2) the scoring system used to rank the alignment; (3) the algorithm used to find optimal (or good) scoring alignments; and (4) the statistical methods used to evaluate the significance of an alignment score. The basic mutational processes that are considered are *substitutions*, which change residues in sequences and

insertions and *deletions*, which add or remove residues. Insertions and deletions are together referred to as *gaps*. The total score assigned to an alignment will be a sum of terms for each aligned pair of residues, provisions for substitutions, plus terms for each gap.

The Scoring System

Since the proteins, in the course of evolution accommodates the substitutions and gaps. It is important to use appropriate substitution matrices while doing sequence alignments. The matrices are simply the prediction of tolerable amino acid changes that might occur to a sequence during the course of evolution. Two major families of matrices are available: (1) Point Accepted Mutation or PAM matrices (Dayhoff *et al*., 1983) and (2) Blocks Amino Acid Substitution Matrices (Henikoff and Henikoff, 1992). The matrices are discussed in detail while talking about statistics of sequence similarity scores. The probability of occurrence of a gap depends upon its length. Thus, when computing an alignment, penalties (P) associated with gaps are often estimated using a linear or "affine" model such as

$$P = \alpha + \beta\phi$$

Where, ϕ is the length of the gap, α the gap opening penalty, and β is the gap extension penalty. The gap opening penalties are higher than the gap extension penalties.

Alignment of Pairs of Sequences

There are two types of sequence alignment, global and local. In global alignment, an attempt is made to align the entire sequence, using as many characters as possible, up to both ends of each sequence. Sequences that are quite similar and approximately the same length are suitable for global alignment. In local alignment, stretches of sequence with the highest density of matches are aligned, thus generating one or more islands of matches or subalignments in the aligned sequences. Local alignments are more suitable for aligning sequences that are similar along some of their lengths but dissimilar in others, sequences that differ in lengths, or sequences that share a conserved region or

domain. It is also the most sensitive way to detect similarity when comparing two highly divergent sequences.

Alignment of pairs of sequences can be performed using: (1) Dot matrix analysis (Gibbs and McIntyre, 1970); (2) Dynamic programming algorithms for global (Needleman and Wunsch, 1970), and local alignment (Smith and Waterman, 1981); and (3) Word or k-tuple methods, as used by BLAST (Altschul *et al*., 1990).

Unless the sequences are known to be very much alike, the **dot matrix** method should be used first. This method displays any possible sequence alignments as diagonals on the matrix. Dot matrix analysis can readily reveal the presence of gaps and direct and inverted repeats that are more difficult to find by other methods.

The dynamic programming method, first used for global alignment of the sequences (Needleman and Wunsch, 1970) and subsequently for local alignment (Smith and Waterman, 1981) provides one or more alignments of the sequences. An alignment is generated by starting at the ends of the two sequences and attempting to match all possible pairs of characters between the sequences and following a scoring scheme (as described before; using substitution matrix and gap penalties) for matches, mismatches and gaps. This procedure generates a matrix of numbers that represents all possible alignments between two sequences. The highest set of sequential scores in the matrix defines an optimal alignment. The dynamic programming is guaranteed in a mathematical sense to provide the optimal or highest scoring alignment for user defined variables, including the substitution matrix and gap penalties.

3.1.5.1 Global Alignment: Needleman-Wunsch Algorithm:

As the name suggests Needleman and Wunsch suggested the first global alignment algorithm in 1970. A more efficient version of the algorithm was introduced by Gotoh in 1982. The later version is described here.

A matrix F indexed by i and j, is constructed. Where i and j are index for each sequence. The value F (i, j) is the score of the best alignment the initial segment $\chi_{1\ldots i}$ of χ up to χ_i and initial segment $\gamma_{1\ldots j}$ of γ up to γ_j. F(i, j) is built recursively. One can start by initializing F (0,0) = 0. Then proceed to fill the matrix from top left to bottom right (or from bottom right to top left). If F $(i-1, j-1)$, F $(i-1, j)$ and F $(i, j-1)$ are known, it is possible to calculate F (i, j). There are three possible ways that the best score F(i, j) of an alignment up to χ_i , γ_j could be obtained: χ_i could be aligned to γ_j in which case F$(i, j) =$ F$(i-1, j-1)$ + s (χ_i , γ_j); or χ_i is aligned to gap, in which case F$(i, j) =$ F$(i-1, j)$ – P; or γ_j is aligned to gap, in which case F$(i, j) =$ F$(i, j-1)$ –P. Here s (χ_i , γ_j) is the local score of the previous step and d is the gap penalty, which can be of the format described before. The best score up to (i, j) will be largest of the three points.

There fore, we have

$$
\begin{aligned}
F(i,j) = \max \{ & F(i-1, j-1) + s(\chi_i, \gamma_j), \\
& F(i-1, j) - P, \\
& F(i, j-1) - P \}
\end{aligned}
$$

The above equation is applied repeatedly to fill the matrix F(i, j) values, calculating the value in the bottom right-hand corner of each sequence to the top-left. As one fill in the F(i, j) values, the pointer is kept in each cell back to the cell from which its F(i, j) is derived. The boundary conditions are calculated as follows. Along the top row, where $j =$ 0, the values F$(i, j-1)$ and F$(i-1, j-1)$ are not defined, so the values F$(i, 0)$ must be handled specially. The value F$(i, 0)$ represent alignments of a prefix of χ to all the gaps in γ, so we can define F$(i, 0) = -i$P. Likewise down the left column F $(0, j) = -j$P. The value in the final cell of matrix, F (n, m), is by definition the best score for a alignment of $\chi_{1\ldots n}$ to $\gamma_{1\ldots m}$, which is the score of the best global alignment of χ and γ. To find the alignment itself, one should find the path of choices that lead to the final highest score using the pointers. The procedure for doing this is known as *traceback*. It works by building alignment in reverse.

3.1.5.2. Local Alignment: Smith-Waterman Algorithm:

Local alignment arises when say for example one is looking for the best alignment between subsequences of χ, γ. The highest scoring alignment of subsequences of χ and γ is called the best local alignment. The algorithm of local alignment is closely related to that described for global alignment. There are two differences. First, in each cell in the previous set of equation extra possibility is added, allowing F (i, j) to take the value 0 if all other options have value less than 0:

$$
\begin{aligned}
F(i,j) = \max \{ & 0, \\
& F(i-1, j-1) + s(\chi_i, \gamma_j), \\
& F(i-1, j) - P, \\
& F(i, j-1) - P \}
\end{aligned}
$$

Taking option 0 corresponds to starting a new alignment. As a result of it the boundary values of top row and left column will be 0 and not $-i$P and $-j$P respectively.

The second change is that the alignment can start anywhere in the matrix, so instead of taking the value in the bottom right corner, F (n, m), for the best score, one have to look for the highest value of F(i, j) over the whole matrix, and start *traceback* from there. The *traceback* ends when the cell with 0 value is encountered.

3.1.6 The Blast Algorithm For Searching Databases

Database Searching Programs

- BLAST http://www.ncbi.nlm.nih.gov/BLAST/
- PSI-BLAST http://www.ncbi.nlm.nih.gov/BLAST/
- FASTA3 http://www.ebi.ac.uk/fasta3/
- HMMER http://hmmer.wustal.edu/
- SAM http://www.cse.ucsc.edu/research/compbio/sam.html
- PFSEARCH http://www.isrec.isb-sib.ch/ftp-server/pftools/pft2.2/

- IMPALA http://bioinformatics.weizmann.ac.il/blocks/impala.html

Sequence searches algorithms like FASTA and BLAST use the word or K-tuple methods. They align two sequences very quickly, by first searching for identical short stretches sequences (called word or k-tuples) and then by joining words in to alignment by the dynamic programming method. These methods are fast and suitable for searching an entire database for the sequences that align best with the query sequence. The FASTA and BLAST methods are heuristic and use feedback to improve performance.

3.1.6.1 The Statistics of Sequence Similarity Scores

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone. In this context, "chance" can mean the comparison of (i) real but non-homologous sequences; (ii) real sequences that are shuffled to preserve compositional properties (Fitch, 1983; Lipman *et al*., 1984; Altschul, 1985) or (iii) sequences that are generated randomly based upon a DNA or protein sequence model. Analytic statistical results invariably use the last of these definitions of chance, while empirical results based on simulation and curve fitting may use any of the definitions.

3.1.6.2 The statistics of local sequence comparison (BLAST)

Statistics for the scores of local alignments, unlike those of global alignments, are well understood. This is particularly true for local alignments lacking gaps, which we will consider first. Such alignments were precisely those sought by the original Basic Local Alignment Search Tool (BLAST) database search programs (Altschul *et al*., 1990). A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences being compared. A modification of the Smith-Waterman (Smith and Waterman, 1981) or Sellers (Sellers, 1984) algorithms will find all segment pairs whose scores can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs. To analyze how high a score is likely to arise by chance, a model of random sequences is needed. For proteins, the simplest model chooses the

amino acid residues in a sequence independently, with specific background probabilities for the various residues. Additionally, the expected score for aligning a random pair of amino acid is required to be negative. Where this not the case, long alignments would tend to have high score independently of whether the segments aligned were related, and the statistical theory would break down.

Just as the sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution, the maximum of a large number of i.i.d. random variables tends to an extreme value distribution (Gumble, 1958). (We will elide the many technical points required to make this statement rigorous.) In studying optimal local sequence alignments, we are essentially dealing with the latter case (Karlin and Altschul, 1990; Dembo *et al*., 1994). In the limit of sufficiently large sequence lengths m and n, the statistics of HSP scores are characterized by two parameters, K and λ. Most simply, the expected number of HSPs with score at least S is given by the formula

$$E = K\, mn\, e^{-\lambda S} \qquad (1)$$

We call this the **E-value** for the score S. This formula makes eminently intuitive sense. Doubling the length of either sequence should double the number of HSPs attaining a given score. Also, for an HSP to attain the score 2x it must attain the score x twice in a row, so one expects E to decrease exponentially with score. The parameters K and λ can be thought of simply as natural scales for the search space size and the scoring system respectively.

3.1.6.3 Bit scores

Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters K and λ.

$$S' = \lambda S - \ln K / \ln 2 \qquad (2)$$

Using above equation one attains a "bit score" S', which has a standard set of units. The E-value corresponding to a given bit score is simply

$$E = mn\, 2^{-S'} \qquad (3)$$

Bit scores subsume the statistical essence of the scoring system employed, so that to calculate significance one needs to know in addition only the size of the search space.

3.1.6.4 P-values

The number of random HSPs with score ≥ S is described by a Poisson distribution (Karlin and Altschul, 1990; Dembo *et al*., 1994). This means that the probability of finding exactly a HSPs with score ≥S is given by

$$e^{-E} * E^{a} / a! \qquad (4)$$

where E is the E-value of S given by equation (1) above. Specifically the chance of finding zero HSPs with score ≥S is e^{-E}, so the probability of finding at least one such HSP is

$$P = 1 - e^{-E} \qquad (5)$$

This is the P-value associated with the score S. For example, if one expects to find three HSPs with score ≥ S, the probability of finding at least one is 0.95. The BLAST programs report E-value rather than P-values because it is easier to understand the difference between, for example, E-value of 5 and 10 than P-values of 0.993 and 0.99995. However, when E < 0.01, P-values and E-value are nearly identical.

3.1.6.5 Database searches

The E-value of equation (1) applies to the comparison of two proteins of lengths m and n. How does one assess the significance of an alignment that arises from the comparison of a protein of length m to a database containing many different proteins, of varying lengths? One view is that all proteins in the database are a priori equally likely to be related to the query. This implies that a low E-value for an alignment involving a short database sequence should carry the same weight as a low E-value for an alignment involving a long database sequence. To calculate a "database search" E-value, one simply multiplies the pairwise-comparison E-value by the number of sequences in the database.

An alternative view is that a query is *a priori* more likely to be related to a long than to a short sequence, because long sequences are often composed of multiple distinct domains. If we assume the *a priori* chance of relatedness is proportional to sequence length, then the pairwise E-value involving a database sequence of length n should be multiplied by N/n, where N is the total length of the database in residues. Examining equation (1), this can be accomplished by treating the database as a single long sequence of length N.
The BLAST programs (Smith *et al*., 1985; Collins *et al*., 1988; Altschul *et al*., 1990; Mott, 1992; Waterman and Vingron, 1994; Altschul and Gish, 1996; Altschul *et al*., 1997; Pearson, 1998) take this approach to calculating database E-value.

3.1.6.6 The Statistics of Gapped Alignment:

The statistics developed above have a solid theoretical foundation only for local alignments that are not permitted to have gaps. However, many computational experiments (Altschul and Gish, 1996; Altschul *et al*., 1997; and some analytic results (Arratia and Waterman, 1994) strongly suggest that the same theory applies as well to gapped alignments. For ungapped alignments, the statistical parameters can be calculated, using analytic formulas, from the substitution scores and the background residue frequencies of the sequences being compared. For gapped alignments, these parameters must be estimated from a large-scale comparison of "random" sequences. The BLAST programs achieve much of their speed by avoiding the calculation of optimal alignment scores for all but a handful of unrelated sequences. The must therefore rely upon a pre-estimation of the parameters lambda and K, for a selected set of substitution matrices and gap costs. This estimation could be done using real sequences, but has instead relied upon a random sequence model (Altschul and Gish, 1996), which appears to yield fairly accurate result (Pearson, 1998). The BLAST programs also correct for Edge effects (Altschul and Gish, 1996).

3.1.6.7 The choice of substitution scores

The results a local alignment program produces depend strongly upon the scores it uses. No single scoring scheme is best for all purposes, and an understanding of the basic theory of local alignment scores can improve the sensitivity of one's sequence analyses. A large number of different amino acid substitution scores, based upon a variety of rationales, have been described (Dayhoff *et al*., 1978; Altschul, 1991; Gonnet *et al*., 1992; Henikoff and Henikoff, 1992). However the scores of any substitution matrix with negative expected score can be written uniquely in the form

$$S_{ij} = (\ln q_{ij} / p_i p_j) \setminus \lambda \qquad (6)$$

Where, the q_{ij}, called target frequencies, are positive numbers that sum to 1, the p_i are background frequencies for the various residues, and λ is a positive constant (Karlin and Altschul, 1990; Altschul, 1991). The λ here is identical to the λ of equation (1). Multiplying all the scores in a substitution matrix by a positive constant does not change their essence: an alignment that was optimal using the original scores remains optimal. Such multiplication alters the parameter lambda but not the target frequencies $q_{ij.}$ Thus, up to a constant scaling factor, every substitution matrix is uniquely determined by its target frequencies. These frequencies have a special significance (Karlin and Altschul, 1990; Altschul, 1991): A given class of alignments is best distinguished from chance by the substitution matrix whose target frequencies characterize the class. The most direct way to construct appropriate substitution matrices for local sequence comparison is to estimate target and background frequencies, and calculate the corresponding log-odds scores of formula (6). These frequencies in general can not be derived from first principles, and their estimation requires empirical input.

3.1.6.8 The PAM and BLOSUM amino acid substitution matrices

While all substitution matrices are implicitly of log-odds form, the first explicit construction using formula (6) was by Dayhoff and coworkers (Dayhoff *et al*., 1978; Schwartz *et al*., 1978). From a study of observed residue replacements in closely related proteins, they constructed the PAM (point accepted mutation) model of molecular evolution. An alternative approach to estimating target frequencies, and the corresponding log-odds matrices, has been advanced by Henikoff and Henikoff (Henikoff

and Henikoff, 1992). They examine multiple alignments of distantly related protein regions directly, rather than extrapolate from closely related sequences. An advantage of this approach is that it cleaves closer to observation; a disadvantage is that it yields no evolutionary model. A number of tests (Pearson, 1995; Henikoff and Henikoff, 1993) suggest that the BLOSUM matrices (Blocks Substitution Matrix derived using BLOCKS database) produced by this method generally are superior to the PAM matrices for detecting biological relationships. BLOSUM62 is default matrix for blast searches.

3.1.6.9 Gap scores and Low Complexity Regions

The theoretical development concerning the optimality of matrices constructed using equation (6) unfortunately is invalid as soon as gaps and associated gap scores are introduced, and no more general theory is available to take its place. However, if the gap scores employed are sufficiently large, one can expect that the optimal substitution scores for a given application will not change substantially. In practice, the same substitution scores have been applied fruitfully to local alignments both with and without gaps. Appropriate gap scores have been selected over the years by trial and error (Pearson, 1995), and most alignment programs will have a default set of gap scores to go with a default set of substitution scores. No clear theoretical guidance can be given, but "affine gap scores" (Gotoh, 1982; Fitch and Smith, 1983; Altschul and Erickson, 1986) with a large penalty for opening a gap and a much smaller one for extending it, have generally proved among the most effective. The BLAST programs employ the SEG algorithm (Wootton and Federhen, 1993) to filter low complexity regions from proteins before executing a database search.

3.1.7 Database Searching with PSI-BLAST

Many functionally and evolutionarily important protein similarities are recognizable only through comparison of three-dimensional structures (Holm and Sander, 1997; Brenner *et al.*, 1998). When such structures are not available, patterns of conservation identified from the alignment of related sequences can aid the recognition of distant similarities.

There is a large literature on the definition and construction of these patterns, which have been variously called motifs, profiles, position-specific score matrices, and Hidden Markov Models (Gribskov, 1987; Staden, 1988; Tatusov *et al*., 1994; Altschul and Gish, 1996; Altschul *et al*., 1997; Durbin *et al*., 1998). In essence, for each position in the derived pattern, every amino acid is assigned a score. If a residue is highly conserved at a particular position, that residue is assigned a high positive score, and others are assigned high negative scores. At weakly conserved positions, all residues receive scores near zero. Position-specific scores can also be assigned to potential insertions and deletions (Gribskov *et al*., 1987; Altschul *et al*., 1997; Durbin *et al*., 1998). The power of profile methods can be further enhanced through iteration of the search procedure (Gribskov, 1992; Tatusov, 1994; Yi and Lander, 1994; Altschul *et al*., 1997). After a profile is run against a database, new similar sequences can be detected. A new multiple alignment, which includes these sequences, can be constructed, a new profile abstracted, and a new database search performed. The procedure can be iterated as often as desired or until the search converges, when no new statistically significant sequences are detected.

3.1.7.1 The design of PSI-BLAST

Iterated profile search methods have led to biologically important observations but, for many years, were quite slow and generally did not provide precise means for evaluating the significance of their results. This limited their utility for systematic mining of the protein databases. The principal design goals in developing the Position-Specific Iterated BLAST (PSI-BLAST) program (Altschul *et al*., 1997) were speed, simplicity and automatic operation. The procedure PSI-BLAST uses can be summarized in five steps: (1) PSI-BLAST takes as an input a single protein sequence and compares it to a protein database, using the gapped BLAST program (Altschul *et al*., 1997). (2) The program constructs a multiple alignment, and then a profile, from any significant local alignments found. The original query sequence serves as a template for the multiple alignment and profile, whose lengths are identical to that of the query. Different numbers of sequences can be aligned in different template positions. (3) The profile is compared to the protein database, again seeking local alignments. After a few minor modifications, the BLAST

algorithm (Altschul *et al*., 1997; Altschul *et al*., 1990) can be used for this directly. (4) PSI-BLAST estimates the statistical significance of the local alignments found. Because profile substitution scores are constructed to a fixed scale (Karlin and Altschul, 1990), and gap scores remain independent of position, the statistical theory and parameters for gapped BLAST alignments (Altschu and Gish, 1994) remain applicable to profile alignments (Altschul *et al*., 1997). (5) Finally, PSI-BLAST iterates, by returning to step (2), an arbitrary number of times or until convergence. Profile-alignment statistics allow PSI-BLAST to proceed as a natural extension of BLAST; the results produced in iterative search steps are comparable to those produced from the first pass. Unlike most profile-based search methods, PSI-BLAST runs as one program, starting with a single protein sequence, and the intermediate steps of multiple alignment and profile construction are invisible to the user.

3.1.7.2 Estimation of statistical parameters for local alignment scores

As discussed previously, computation experiments strongly suggest that the optimal gapped local alignment scores produced by the Smith-Waterman algorithm (Smith and Waterman, 1981) and approximated by FASTA (Pearson and Lipman, 1988) or Gapped BLAST (Waterman and Vingron, 1994; Altschul and Gish, 1996) follow an extreme value distribution (Gumble, 1958). Specifically, the probability that the optimal score S from the comparison of unrelated proteins is at least x is given by the equation,

$$P(S \geq X) = 1 - \exp(-K\, mn\, e^{-\lambda x}) \qquad (1)$$

Where, K and λ are statistical parameters dependent upon the scoring system and the background amino acid frequencies of the sequences being compared. BLAST estimates parameters beforehand for specific scoring schemes by comparing many random sequences generated using a standard protein amino acid composition (Robinson and Robinson, 1991). For example, using BLOSUM-62 amino acid substitution scores (Henikoff and Henikoff, 1992), and affine gap costs (Fitch and Smith, 1983; Altschul and Erickson, 1986; Myers and Miller, 1988) in which a gap of length k is assigned a score of -(10 + k), 10,000 pairs of length-1000 random protein sequences were generated, and Smith-Waterman algorithm was used to calculate 10,000 optimal local alignment scores.

From these scores, λ was estimated at 0.252 and K at 0.035 by the method of maximum-likelihood (Lawless, 1982). In general, given M samples from an extreme value distribution, the ratio of the maximum-likelihood estimate of lambda to its actual value is approximately normally distributed, with mean 1.0 and standard deviation 0.78/sqrt(M) (Lawless, 1982). Thus the standard error for our estimate of λ is about 0.002, or less than 1%. The chi-squared goodness-of-fit test for these data, with 34 degrees of freedom, is 25.6, which is lower than would be expected to occur by chance 87% of the time even were the theory precisely valid.

3.1.7.3 Generalization to PSI-BLAST alignment scores

In order for PSI-BLAST to iterate automatically, it needs to be able to generate accurate estimates of the statistical significance of the alignments it produces. Unfortunately, there is no analytic theory with which to estimate the statistical significance of a gapped local alignment of a profile and a simple sequence. One hope is that if amino acid scores within each column of a PSI-BLAST profile can be constructed to the same scale (Karlin and Altschul, 1990; Altschul, 1991) i.e. with the same ungapped λ, as those for a standard amino acid substitution matrix, and then use the same position-independent gap costs, the same gapped λ may result. To review, for ungapped local alignments, any substitution matrix takes the form

$$S_{ij} = (\ln q_{ij} / p_i p_j) \lambda_u \qquad (2)$$

Where, the q_{ij} are the target frequencies for aligned pairs of amino acids, the p_i are background frequencies, and the subscript for λ indicates it is the statistical parameter for ungapped local alignments scale (Karlin and Altschul, 1990; Altschul, 1991). For a PSI-BLAST profile (Altschul *et al*., 1997), each column has its own unique set of amino acid target frequencies q_i. Following (2), the amino acid scores for this column may be constructed to the same scale by using the formula

$$S_i = (\ln q_i / p_i) / \lambda_u \qquad (3)$$

The hope is that, given a specific set of gap costs, the gapped λ for the PSI-BLAST profile will be the same as the gapped λ for the standard substitution matrix, which may be calculated in advance.

3.1.8 Multiple Alignment using CLUSTAL

CLUSTAL has been written and subsequently improved during the span of last ten years (Higgins and Sharp, 1988; Thompson *et al*., 1994a; Higgins *et al*., 1996). CLUSTAL performs a global multiple alignment using following steps: (1) Perform pairwise alignment of all the sequences; (2) use the alignment scores to produce the phylogenetic tree (see later); and (3) align the sequences sequentially, guided by the phylogenetic relationships indicated by the tree. Thus, the most closely related sequences are aligned first, and then additional sequences and groups of sequences are added, guided by the initial alignments to produce a multiple sequence alignment. The quality of the alignments produced in such way is excellent, as judged by the ability to correctly align corresponding domains from sequences of known secondary or tertiary structure. The initial alignments used to produce the guide tree may be obtained by a fast k-tuple or pattern finding approach similar to BLAST that is useful for many sequences, or a slower, dynamic programming method may be used. An enhanced dynamic programming alignment algorithm (Myers and Miller, 1988) is used to obtain optimal alignment scores. For producing a phylogenetic tree, genetic distances between the sequences are required. The genetic distance is the number of mismatched positions in an alignment divided by the total number of matched positions (positions opposite to gaps are not scored).

The recent version is CLUSTALW (Thompson *et al*., 1994) with the W standing for "weighing" represent the ability of the program to provide weights to sequence and program parameters. The sensitivity of the commonly used progressive multiple sequence alignment (CLUSTALV) method has been greatly improved for the alignment of divergent protein sequences using following steps. Firstly, individual weights are

assigned to each sequence in a partial alignment in order to downweight near-duplicate sequences and upweight the most divergent ones. Secondly, amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. Fourthly, positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions.

The CLUSTALX (Thompson *et al*., 1997) is graphic interface to CLUSTALW. CLUSTALX is new windows interface for the widely used progressive multiple sequence alignment program CLUSTALW. It is easy to use, providing an integrated system for performing multiple sequence and profile alignments and analyzing the results. CLUSTALX displays the sequence alignment in a window on the screen. A versatile sequence coloring scheme allows the user to highlight conserved features in the alignment. Pull-down menus provide all the options required for traditional multiple sequence and profile alignment. New features include: the ability to cut-and-paste sequences to change the order of the alignment, selection of a subset of the sequences to be realigned, and selection of a sub-range of the alignment to be realigned and inserted back into the original alignment. Alignment quality analysis can be performed and low-scoring segments or exceptional residues can be highlighted. Quality analysis and realignment of selected residue ranges provide the user with a powerful tool to improve and refine difficult alignments and to trap errors in input sequences.

3.1.9 Literature

Searching for literature can be of prime importance for a computational biologist. It is equally important for biologists working in all areas of research to stay acquinted with the latest development in the field. The Literature can be searched over the web in PubMed. PubMed is a project developed by the National Center for Biotechnology Information (NCBI). It has been developed in conjunction with publishers of biomedical literature as

a search tool for accessing literature citations and linking to full-text journals at Web sites of participating publishers. The PubMed is available at NCBI web site at http://www.ncbi.nlm.nih.gov/.

3.1.10 Uses of Patterns

Patterns, searched using family alignment databases or multiple sequence alignments, are used to describe the residues that are conserved in a set of sequences. Discovering patterns conserved in a protein family can help in the understanding between sequence, structure and function of the protein under study. When a conserved pattern is discovered, one should analyze how likely it is that pattern has been discovered by chance. The less likely this is, the more likely the pattern is to describe functionally or structurally conserved residues.

If one finds a pattern that not only is conserved in the family, but also is unique to the family, i.e., no (or few) sequences outside the family matches the pattern, then pattern can be used to identify new members of the family. The PROSITE database (Hoffman *et al*., 1999) of protein sites and families illustrates this. The patterns in PROSITE can be used not only for finding out structurally and functionally important residues but also for classification purposes for removing false family members.

3.2 Identification of related structures

PSI-BLAST (Altschul *et al*., 1997) or its relatives has been the best sequence (or homology) searching. Probabilistic or Bayesian models also have been applied (e.g., hidden markov models; Durbin *et al*, 1998) for detection of remote homologues. If structure level similarity in terms of PDB hit(s) is suggested by sequence searching methods, one can straight forward transfer information by homology or can set stage for homology modeling (step 3) for more refined function prediction. But in case that the sequence searches doesn't arrive at any useful hits once can resolve for secondary structure prediction or fold recognition methods for identifying the related structures in fold library.

3.2.1 Secondary Structure Prediction

3.2.1.1 History and General Comments

In one of the earliest studies involved in the analysis of helix content in proteins by optical rotatory dispersion, Szent -Gyorgyi and Cohen (1957) showed that proteins with high proline content also exhibit less helicity. Thus, this established the idea of proline as, in some sense, a helix breaker. Cook in 1967 has given some early rules for helix formation, using then available structures and chemical properties of residues. Some of them are (1) Ala, Val and Leu are the helix formers and they tend to occur in the middle of helix. (2) The size of the side chain of a helix-forming residue is important. (3) Residues Asp, Asn and phe are helix breaking. (4) Asp, Glu, and Thr favor N-termini of α-helical region. (5) Lys, His and Arg prefers the C-termini of α-helical regions.

As observations the above rules were good and that started the search for more sophisticated rules. The x-ray determined structures of 15 proteins were examined by Chou and Fasman (1974a) and the number of occurrence of a given amino acid in the α helix, β sheet and coil was tabulated. From this, the conformational parameters (propensities) for each amino acid within a protein, its occurrence in a given type of secondary structure, and the fraction of residues occurring in that type of structure. The residue preferences found by Chou and Fasman has been quite accurate and has been discussed before while discussing about properties of amino acid side chains. Having computed the propensities Chou and Fasman derived the rules for secondary structure prediction. This rules, when applied then resulted in 70-80% predictive accuracy. However now that accuracy is predicted to be around 50% only. This was the first attempt to apply statistical methods for secondary structure prediction. With this Chou and Fasman has unknowingly set a trend to do a three-state prediction for a given sequence. The GORIII method (Garnier *et al*., 1978; Gibrat *et al*., 1987) is a representative of the methods based not only on single residue propensities but also on statistically significant pairwise residue interactions. The preference (information content) I of a residue with sequence number j and residue type Rj for a secondary structure type Z ∈ {helix, sheet, coil} is approximated as

$$I(S_j = Z; R_{j-8}, \ldots, R_{j+8}) = \sum I(S_j\, Z; R_{j+m} / R_j)$$, where Σ runs from -8 to 8.

in a sequence environment of eight residues on either side of a central one. The information I carried by the amino acid pair (R_{j+m} / R_j) on the occurrence of the event Z (adoption of a specific secondary structure state) is defined as

$$I(S_j = Z; R_{j+m} / R_j) = \log \left[P\left(Z / (R_{j+m} / R_j)\right) / P(Z) \right]$$

Where, P denotes the conditional probability. The enormous amount of parameters (3 structural states × 20 amino acid types × 20 amino acid types × 17 sequence positions) is estimated from a set of 68 non-redundant protein crystallographic structures. The prediction accuracy achieved was about 63% then (Gibrat *et al*., 1987; Garnier and Levin, 1991). A further improvement of 2.5 to 6.5% (Biou *et al*., 1988) was obtained by combining GORIII method with two other prediction schemes. First based on hydrophobicity patterns that are observed in regular secondary structures (bit pattern

method, and second using structural similarity between short, sequentially homologous peptides (Levin and Garnier, 1988). It is important to note that the predictive power of methods relying on only sequentially local structure information is limited by about 65% (Gibrat *et al*., 1991). A further increase requires the consideration of tertiary interactions.

3.2.1.2 Importance of Evolutionary information

One of the most successful applications of the multiple sequence alignment has been to improve the accuracy of secondary structure prediction. This has been first used by Zvelebil *et al*., (1987) and subsequently used by Levin *et al*., (1993); Rost and Sander (1993); Salamov *et al*., (1995); Cuff *et al*., (1999) and others for reaching an overall three state prediction accuracy more than 70%. It is around 9% more than single sequence based methods. Some of the methods use multi neuron neural networks and jury of neural network to give three-state prediction.

It is well known that the structure is more conserved than sequences (Chothia and Lesk, 1986; Pastore and Lesk, 1990). What we see in alignment of native proteins is a record of the evolution. If proteins share more than 30% identity most likely they share same fold (Chothia and Lesk, 1986). Of course, not any two residues can be exchanged. On the contrary, the pattern of residue substitutions within one structure family contains specific information about the structure. Gaps in multiple alignments occur more often in loop regions than in regular secondary structure elements such as helix and strand (Pascarella and Argos, 1992). This implies that the number of gaps at a particular position carries information about secondary structure: the more gaps found in a region, the more likely it is a loop region (provided the alignment is correct).

Although secondary structure alone can is generally of limited use, it is nonetheless helpful to be able to refer to a reliable secondary-structure prediction to predict the tertiary structure by fold recognition or motif searches and secondary structure based threading. The following structural clues can sometimes be obtained through inspection of predicted secondary structural elements:

- The structural class of target proteins may be ascertained (all α, all β, or α-β)
- Structural repeats can be detected. By identifying a repeating sequence of secondary structures, it is sometimes possible to identify repeated domains in the target proteins.
- The sequence of secondary structural elements can be compared to the folds matched by fold recognition. For the fold-recognition methods, which do not use predicted secondary structure, this "second opinion" is of great value in determining the degree of confidence to assign to the prediction.

Online servers available for Secondary Structure Prediction

- GOR IV http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html
- PREDATOR http://www.embl-heidelberg.de/cgi/predator_serv.pl
- PHDsec http://cubic.bioc.columbia.edu/predictprotein/
- JPRED http://jura.ebi.ac.uk/
- NNPREDICT http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
- PSIPRED http://insulin.brunel.ac.uk/psiform.html

3.2.1.3 Secondary Structure Prediction Using Tertiary Interactions

PREDATOR (Frishman and Argos, 1997) is a secondary structure prediction program. It takes as input a single protein sequence to be predicted and can optimally use a set of unaligned sequences as additional information to predict the query sequence. The principal step in the procedure involves generation of seven secondary structural propensities for input sequences and the related sequences. Three propensities are based on long-range interactions involving potential hydrogen bonding resides in antiparallel (P_1) and parallel (P_2) β strands as well as α-helices (P_3). Three further propensities for helix (P_4), strand (P_5) and coil (P_6) rely on the similarity of the sequence segments to be predicted with those of known conformation (nearest neighbor approach; Zhang *et al*., 1992). Finally a statistically based turn propensity (P_7) is used over a four-residue window (Hutchinson and Thornton, 1994). The mean prediction accuracy of

PREDATOR is 68% for a single sequence and 75% for a set of related sequences. PREDATOR does not use multiple sequence alignment. Instead, it relies on careful pairwise local alignments of the sequences in the set with the query sequence to be predicted.

3.2.1.4 Prediction with Neural Networks

A neural network mimics the architecture of brain neurons. Since 1958, when psychologist Frank Rosenblatt proposed the "Perceptron," a pattern recognition device with learning capabilities, the hierarchical neural network has been the most widely studied form of network structure. A hierarchical neural network is one that links multiple neurons together hierarchically. The special characteristic of this type of network is its simple dynamics. That is, when a signal is input into the input layer, it is propagated to the next layer by the interconnections between the neurons. Simple processing is performed on this signal by the neurons of the receiving layer prior to its being propagated on to the next layer. This process is repeated until the signal reaches the output layer completing the processing process for that signal. The manner in which the various neurons in the intermediary (hidden) layers process the input signal will determine the kind of output signal it becomes (how it is transformed). As you can see, then, hierarchical network dynamics are determined by the weight and threshold parameters of each of their units. If input signals can be transformed to the proper output signals by adjusting these values (parameters), then hierarchical networks can be used effectively to perform information processing.

Since it is difficult to accurately determine multiple parameter values, a learning method is employed. This involves creating a network that randomly determines parameter values. This network is then used to carry out input-to-output transformations for actual problems. The correct final parameters are obtained by properly modifying the parameters in accordance with the errors that the network makes in the process. Error back-propagation learning method has played a major role in the recent neural network computing boom. The back-propagation paradigm has been tested in numerous applications including bond rating, mortgage application evaluation, protein structure determination, backgammon playing, and handwritten digit recognition.

Qian and Sejnowski (1988) presented a neural network method for prediction of secondary structures for single protein sequences using supervised learning method and back-propagation. They trained a standard network with 13 input groups, with 21 units/group using 106 protein structures and different window lengths of 1-21 residues. They achieved a success rate of 64.3% for three-state prediction. This is substantially better than the prediction from statistical methods described before. This however, opened a way for next generation secondary structure prediction methods, as described below.

3.2.1.5 Prediction with Neural Networks and Multiple Alignments:

3.2.1.5.1 PHD Secondary Structure Prediction Method:

PHD is made of three individual prediction methods that use evolutionary information as input to predict secondary structure (PHDsec; Rost and Sander, 1993a,b; 1994a), relative solvent accessibility (PHDacc; 1994b) and transmembrane helices (PHDhtm; Rost *et al*., 1995). Presently it is available on predict protein server. The method consists of following steps.

Generating Multiple Alignment

First step in a PHD prediction is to search for remote homologues from PRODOM domain database using SAM-T98 (Karplus *et al*., 1997). The pairwise profile-based alignment is generated using the program MaxHom (Sander and Schneider, 1991).

Multiple level of Computation

The PHD methods process the input information on multiple levels. The first level is a feed-forward neural network with three layers of units (input, hidden, and output). Input to this first level sequence-to-structure network consists of two contributions: one from

the local sequence that is, taken from a window of 13 adjacent residues and another from global sequence. The global information contents for example can be percentage of each amino acid in protein or length of protein etc. Output of the first level network is the 1D structural state for the residue at the center of the output window. For PHDsec and PHDhtm the second level is a structure-to-structure network. The second level structure-to-structure network introduces a correlation between adjacent residues. It is important that the neural network get trained by balanced data for improved prediction of less populated states (e.g., strand) but this is associated with less accurate prediction of more populated states (e.g., loops). Consequently, the overall accuracy is lower for balanced training than for the unbalanced training. To find a compromise between this, a third and final jury decision is performed (effectively a compromise between over- and under prediction). This jury is a simple arithmetic average over, typically, four differently trained networks: all combination of first and second level networks with balanced and unbalanced training, and with balanced and unbalanced training of second level network. The final prediction is assigned to the unit with maximum output value.

Final Filtering

For secondary structure prediction (PHDsec) the filter affects only drastic and unrealistic predictions. Only filter used for predicting transmembrane helices (PHDhtm) is crucial for performance. Predicted transmembrane helices, which are too long, are either split or shortened. Predicted transmembrane helices, which are too short are either elongated or deleted. All decisions are based on the strength of the prediction and length of the transmembrane helix predicted. PHD predicts secondary structure at more than 72% accuracy and transmembrane helices are predicted with accuracy of more than 95%.

3.2.1.5.2 Secondary Structure Prediction using JPRED:

JPRED is a consensus prediction method (Cuff *et al*., 1998) It applies combination of various methods and returns consensus prediction which improves the average three state accuracy of prediction by 1% that to PHD. The server simplifies the use of current

prediction algorithms and allows conservation patterns important to structure and function to be identified. The server accepts two input types, a family of aligned protein sequences or a single protein sequence. If a single sequence is submitted, an automatic process creates a multiple sequence alignment, prior to prediction (Cuff & Barton, 1998). Six different prediction methods: DSC (King & Sternberg, 1996), PHD (Rost & Sander, 1993), NNSSP (Salamov & Solovyev, 1995), PREDATOR (Frishman & Argos, 1997), ZPRED (Zvelebil *et al*., 1987) and MULPRED (Barton, 1988, unpublished) are then run, and the results from each method are combined into a simple file format.

The NNSSP, DSC, PREDATOR, MULPRED, ZPRED and PHD methods were chosen as representatives of current state of the art secondary structure prediction methods, that exploit the evolutionary information from multiple sequences. Each derives its prediction using a different heuristic, based upon nearest neighbors (NNSSP), jury decision neural networks (PHD), linear discrimination (DSC), consensus single sequence method combination (MULPRED), hydrogen bonding propensities (PREDATOR), or conservation number weighted prediction (ZPRED).

The predictions and corresponding sequence alignment are rendered in colored HTML, Java (Clamp *et al*., 1998) and Postscript. The predictions are colored and aligned with their corresponding family of sequences. Physico-chemical properties, solvent accessibility, prediction reliability and conservation number values (Zvelebil *et al*., 1987) for each amino acid are included in the output. The original ASCII text data from each of the prediction methods can also be downloaded. For example, BLAST results, MSF and HSSP format alignments, pair comparison files and so on.

3.2.1.6 Transmembrane Region Prediction

Onlione Servers for Transmembrane Region Detection

- DAS http://www.sbc.su.se/~miklos/DAS/
- HMMTOP http://www.enzim.hu/hmmtop/submit.html

- PHDhtm http://dodo.cpmc.columbia.edu/predictprotein/
- SOSUI http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
- TMAP http://www.mbb.ki.se/tmap/
- TMHMM http://www.cbs.dtu.dk/services/TMHMM-1.0/
- TMpred http://www.ch.embnet.org/software/TMPRED_form.html
- TopPred2 http://www.sbc.su.se/~erikw/toppred2/

3.2.1.6 .1 Using TOPRED

TOPPRED (von Heijne, 1992) is strategy for predicting the topology of bacterial inner membrane proteins and it is proposed on the basis of hydrophobicity analysis, automatic generation of a set of possible topologies and ranking of these according to the positive-inside rule. It is shown that positively charged residues in short loop guide the orientation of helices by preventing translocation across membranes (von Heijne, 1994). It applies two empirical hydrophobicity cutoffs to the output of a sliding trapezoid window in order to compile certain and putative transmembrane helices. The combination of the putative helices that produces strongest enrichment of positively charged residues on the cytoplasmic side is selected as best prediction.

3.2.1.6 .2 Using TMPRED

The TMpred (Hofmann and Stoffel, 1993) program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occuring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. TMbase is mainly based on SwissProt, but contains informations from other sources as well. All data is stored in different tables, suited for use with any relational database management system. These tables are distributed as ASCII files.

3.2.1.6 .3 Using HMMTOP

HMMTOP (Tusnády and Simon, 1998) is based on the hypothesis that the localization of the transmembrane segments and the topology are determined by the difference in the amino acid distributions in various structural parts of these proteins rather than by specific amino acid compositions of these parts. Five structural parts were defined in membrane proteins: membrane helix (H), inside and outside helix tail (i and o), inside and outside loop (I and O). Topology is determined by partitioning amino acid sequence in a way that product of the relative frequencies of amino acids in these structural parts along the sequence should be maximal. This task can be solved by the hidden Markov model (HMM), in which biological constraints can be taken into account by the architecture of HMM using the Baum-Welch algorithm. The structural parts, which are described above, correspond to the five states used by the model. With use of this HMM architecture a state sequence (i.e. a prediction) can be generated as follows: first a state is chosen according to the initial state probabilities. Every following state is chosen according to the transition probabilities of the present state. The aim is to maximize the product of these probabilities and the emission symbol probabilities along the given sequence. The method has been a successful demonstration of HMM in secondary structure prediction.

3.2.1.6 .4 Using TMHMM

TMHMM (Sonnhammer *et al*., 1998) is based on a hidden Markov model (HMM) with an architecture that corresponds closely to the biological system. The model is cyclic with 7 types of states for helix core, helix caps on either side, loop on the cytoplasmic side, two loops for the non-cytoplasmic side, and a globular domain state in the middle of each loop. The two loop paths on the non-cytoplasmic side are used to model short and long loops separately, which corresponds biologically to the two known different membrane insertion mechanisms. The close mapping between the biological and computational states allows us to infer which parts of the model architecture are important to capture the information that encodes the membrane topology, and to gain a better understanding of the mechanisms and constraints involved. Models were estimated both by maximum likelihood and a discriminative method, and a method for reassignment of the membrane

helix boundaries was developed. In a cross-validated test on single sequences, our TMHMM correctly predicts the entire topology for 77% of the sequences in a standard dataset of 83 proteins with known topology. The same accuracy was achieved on a larger dataset of 160 proteins. These results compare favorably with existing methods. The TMHMM method is very similar to HMMTOP and uses the same algorithm for training the internal parameter of markov model.

3.2.1.6 .5 Using SOSUI

SOSUI (Hirokawa *et al*., 1998) is a system for discrimination of membrane proteins together with soluble ones and the prediction of transmembrane helices. One important assumption SOSUI system makes is that, a primary transmembrane helix is stabilized by a combination of amphiphilic side chains at helix ends as well as high hydrophobicity in the central region. The system uses four paramamters in form of four indices. A hydropathy index (Kyte and Doolittle, 1982), an amphphphilicity index, an index of amino acid charges and length of each sequence. The SOSUI output contains (i) the type of protein; (ii) the region of transmembrane helices; (iii) a graph of the hydropathy plot; and (iv) helix wheel diagram for all transmembrane helices.

3.2.1.7 Perscan: a method for predicting 3D models of transmembrane helices

The structure prediction of integral membrane proteins is a difficult task. However since the membranes are essentially 2 dimensional, they provide a powerful constraint upon arrangement of the elements that cross them. Therefore structure prediction of α– helical membrane proteins can often be viewd as a two dimensional problem for which four pieces of information are required: (1) The region of the sequences that form transmembrane helices (2) the basic topology of transmembrane domain; (3) The side of each helix that faces the helix bundle. (4) The relative depth that each helix is inserted into membrane.

Perscan is a collection of programs that attempts to address some of the requirements in order to get information about system under study. Perscan (V7.0) is a collection of 13 FORTRAN programs that detect and display periodicity in protein sequences or structures. These are 2 'PROF' programs, 5 'PER' and 5 'SCAN' programs and one utility program called SELHEL. The PROF programs are more traditional method for searching transmembrane helices. Perscan use Fourier transform methods in order to identify periodicity of hydrophobic and hydrophilic residues in sequence and sequence alignments to identify amphipathic helices (Eisenberg *et al*., 1984; Cornette *et al*., 1987). The periodicity of conserved/variable residues can be used to predict the presence of helix (Komiya *et al*., 1988). The third method uses different between substitution patterns described for soluble (Overington *et al*., 1990; Overington *et al*., 1992) and membrane proteins (Donnelly *et al*., 1993). These environment-specific substitution tables can also be used to assign a value that quantifies the extent to which each position in a sequence alignment is buried. The periodicity in such values can be used to assign values to predict the presence of α-helix and also allows the buried face of each helix to be identified.

The SCAN programs (SCANHYD, SCANVAR, SCANCON, SCANMUT and SCANACC) are designed to look for sequences in complete sequence alignments or structures, whereas the PER programs (PERHYD, PERVAR, PERCON, PERMUT and PERACC) carry out a more detailed analysis of a single putative helical region. The five identifiers (HYD, VAR, CON, MUT and ACC) indicate the different properties for which helical periodicity is searched (i.e., hydrophobicity, variability, conservation, substitution-patterns and solvent accessibility). This information is then used to predict the point at which the helix makes contact with the aqueous environment at the borders of bilayer (Donnelly *et al*., 1993;Donnelly and Codgell, 1993).

PERSCAN is also useful as secondary structure prediction method. The results of PERSCAN including the number of helices, variable and constant faces of a helix, buried faces, hydrophobic moments combined with helix-wheel diagram provided by it, can be used to build a model of the system under study.

3.2.2 Tertiary Structure Prediction (or Fold Recognition)

Lists of threading servers

- 123D http://www-lmmb.ncifcrf.gov/~nicka/123D.html
- 3D-PSSM http://www.bmm.icnet.uk/~3dpssm
- Honig lab http://honiglab.cpmc.columbia.edu/
- Libra I http://www.ddbj.nig.ac.jp/htmls/E-mail/libra/libra/libra.html
- NCBI http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/threading.html
- Profit http://lore.came.sbg.ac.at/home.html
- Threader2 http://insulin.brunel.ac.uk/threader/threader.html
- TOPITS http://www.embl-heidelberg.de/predictprotein/help05.html
- UCLA-DOE http://www.doe-mbi.ucla.edu/people/frsvr/srsvr.html
- GenThreader http://insulin.brunel.ac.uk/psiform.html

The term "threading was first coined in 1992 by Jones *et al*. (Jones *et al*., 1992), but the field has grown considerably with many different methods being proposed. The idea behind threading comes from the fact that a large percentage of proteins adopt a limited number of folds (Orengo *et al*., 1994).

Description of the methods is out of scope of this thesis. How ever, the most important methods so far had been the 1-D-3-D profiles (Bowie *et al*., 1991), threading (Jones *et al*., 1992), using secondary structure predictions (Rost, 1997), combining sequence similarity with threading as implimented in Gen-THREADER (Jones, 1999), and using structural profiles (3D-PSSM; Kelly *et al*., 2000).

3.3 Derivation of Model from Template(s)

3.3.1 History and General Comments

Comparative modeling uses experimentally determined protein structures to predict conformation of other proteins with similar amino acid sequences. This is possible because a small change in the sequence usually results in a small change in structure (Lesk and Chothia, 1986; Hubbard and Blundell, 1987). The accuracy of protein models obtained by comparative modeling compares favorably with the model calculated by other theoretical models. The comparative method produces models with an r.m.s error as low as 1Å for the sequences that have sufficiently similar homologues with known 3D structures (Topham *et al*., 1991); in contrast, physical prediction methods and combinatorial modeling calculates structures with r.m.s. error of approximately 3.5Å for small proteins (Cohen and Kuntz, 1989; Wilson and Doniach, 1989). On the other hand, comparative modeling is not as accurate as X-ray crystallography and NMR, which can determine protein structures with an r.m.s. error of approximately 0.3 and 0.5Å, respectively (Clore and Gronenborn, 1991). It is also restricted to sequences with closely related proteins with known structures. It has been estimated that approximately one third of all knows sequences are related to at least one protein of known structure (Rost and Sander, 1996). With approximately 0.7 million sequences known, comparative modeling had been applied to 2,43,410 domains in known sequences (Sanchez *et al*., 2000). This is an order of magnitude more proteins than experimentally determined protein structures (~15,600). Furthermore, the usefulness of comparative modeling is steadily increasing because the number of different structural folds that protein can adopt is limited (Chothia, 1992), and because the number of experimentally determined new structures is increasong exponentially (Holm and Sander, 1996). Due to Structural Genomics Initiative

in less than 10years, atleast one example of most structural folds will be known, making comparative modeling applicable to most globular domains in most protein sequences (Sali *et al*., 1998).

Early modeling studies frequently relied on the construction of wire or plastic models and only later incorporated interactive computer graphics. The first models produced from homologous proteins were constructed by taking the existing coordinates of a single known structure and then altering those side chains that were not identical in the protein to be modeled. Browne and co-workers (1969) published the first model, they modeled bovine α-lactalbumin on the three dimensional structure of hen egg-white lysozyme. For reviews on the history and development of homology modeling please see Johnson *et al*., 1994; Sanchez and Sali, 1997 and Sanchez and Sali, 2000 etc.

3.3.2 Modeling Procedure

Modeling procedures can be envisaged as two steps. The first step is to solve the inverse folding problem: to define all those sequences that can adopt a particular fold (step1 and step 2 of this thesis; figure 1.1). It involves projecting restraints from a three-dimensional structure onto a one-dimensional sequence. The second step is to use the sequence with the knowledge that the protein belongs to a family of known fold to construct a model.

The modeling techniques used for comparative modeling generally falls into two classes: (a) assembly of rigid fragments and (b) use of distance geometry to construct the models that are in best agreement with the distance constraints. Both the approaches have been used while working towards this thesis. The packages used are COMPOSER (Sutcliffe *et al*., 1987a,b; a part of SYBYL suite) and MODELER (Sali and Blundell, 1993) respectively. The flow chart of the methodology used by COMPOSER and MODELER is given in figure 3.1 and figure 3.2 respectively. The step obviously important to both the methods is defining the topologically equivalent parts using the superposition of homologous structures and other structural properties. COMPOSER uses it to derive the structural framework (Structurally Conserved Regions or SCRs; Sutcliffe *et al*., 1987a)

for the model, while MODELER uses it to derive the spatial restraints for the model (Sali *et al*., 1993. The rules for comparative modeling are also derived from the database of homologous structures (Sali and Overington, 1994). Several methods are available for defining topological equivalence of residues. Most of them use superposition of the structures. However proteins can be compared at residue, secondary structure, supersecondary structure, motif or domain levels also. The features that can be used for the comparison at residue and segment levels of two structures is summarized table (derived from Sali and Blundell, 1990).

Comparison at Residue level
Properties: Identity, Residue type properties, Local conformation, Distance from gravity centre, Side- chain orientation, Main-chain orientation, Solvent accessibility, Position in space *Relations:* Hydrogen bond, Distance to one or more nearest neighbors, Disulfide bond, Ionic bond, Hydrophobic cluster
Comparison at Segment level
Properties: Secondary structure type, Amphipathicity, Improper-dihedral angle, Distance form gravity centre, Orientation relative to gravity centre, Solvent accessibility, Position in space, Orientation in space *Relations:* Distance to one or more nearest neighbors, Relative orientation of two or more segments

Table 3.1 Showing different levels at which two protein structures can be compared to derive topological equivalence

The methods in this thesis for superposition and generating structure-based alignments are MNYFIT (Sutcliffe *et al*., 1987), COMPARER. (Sali and Blundell, 1990) and STAMP (Russell and Barton, 1992). COMPOSER uses MNYFIT for generating the structural framework while the structure-based alignment for MODELER input can be prepared using either method.

3.3.2.1 MNYFIT

MNYFIT (Sutcliffe *et al*., 1987) works by method of unweighed least square fitting (Hermans and Ferro, 1971; McLachlan 1979, 1982; Sutcliffe *et al*., 1987a) choosing one of the structures at random to the framework and fit all the others to it pairwise. The process is iterative and, it does the fitting till an r.m.s of 10^{-5}Å is reached. In the second step, atomic positions are weighed while doing least square fit as to reflect how representative it is of the set of topologically equivalent positions. The third step generates a framework that is close to the specific structure to be modeled.

3.3.2.2 STAMP

STAMP (Russell and Barton, 1992) is designed with specific purpose of generating multiple sequence alignment from tertiary structure comparison. It provides not only multiple alignments and the corresponding 'best-fit' superpositions, but also a systematic and reproducible method for assessing the quality of such alignments. It also provides a method for protein 3D-structure database scanning.

STAMP uses Rossman and Argos equation (Rossman *et al*., 1975) for expressing the probability of equivalence of residue structural equivalence. STAMP then uses Smith Waterman dynamic programming algorithm (Smith and Waterman, 1981; Sankoff and Kruskal, 1983; Barton, 1994) for fast determination of best path through a matrix containing a numerical measure of the pairwise similarity of each position in one sequence to each position in another sequence. Within STAMP, these similarity values correspond to the modified values of Rossman and Argos equation. From this a set of equivalent C_α positions are obtained. These are used to obtain a best fit transformation

and r.m.s. deviation by a least square method (Kabsch, 1978; McLachlan, 1979). This transformation is applied to yield two new sets of coordinates for which the entire procedure is repeated in iterative fashion until the two sets of coordinates, and the corresponding alignment, converge on a single solution.

3.3.2.3 COMPARER

COMPARER (Sali and Blundell, 1990) attempts to define topological equivalences in protein structures by comparing properties of protein structures at various levels. Residue and segment properties that COMPARER takes in to are: residue local fold, residue type properties, residue distance from molecular gravity centre, side-chain orientation relative to molecular gravity centre, side-chain orientation relative to main-chain, main-chain orientation relative to molecular gravity centre, side-chain solvent accessibility, main-chain solvent accessibility, hydrogen bonding relationship, residue identity, residue position in space, φ, ψ dihedral angle and main chain directions. A normalized difference of a certain feature between residues from the pair of proteins is computed. A scaling factor is defined that determines the relative importance of a feature used for comparison. From this a weighted sum is calculated. Relationships were weighed using simulated annealing methods. Once the dissimilarity matrices are computed. Best pairwise or multiple alignment is searched using dynamic programming approach described before (Needleman and Wunsch, 1970; Sankoff and Kruskal).

COMPARER alignments are more useful in terms of modeling by spatial restraints since it gives the topologically equivalent residues using hierarchical definition of structure and used in MODELER (Sali and Blundell, 1993).

Description of Modeling Programs

3.3.2.4 COMPOSER

As mentioned before COMPOSER (Sutcliffe *et al*., 1987a,b) is an automated approach of comparative modeling based on assembly of rigid fragments. It is available as a part of SYBYL module of TRIPOS Inc. The flow chart of the COMPOSER methodology is as shown in figure 3.1. As described before for homologous structures are used to derive the structural framework or SCRs using MNYFIT. Modeling of gaps or Structurally Variable Regions (SVRs) involves search for fragments of suitable length and end-to-end

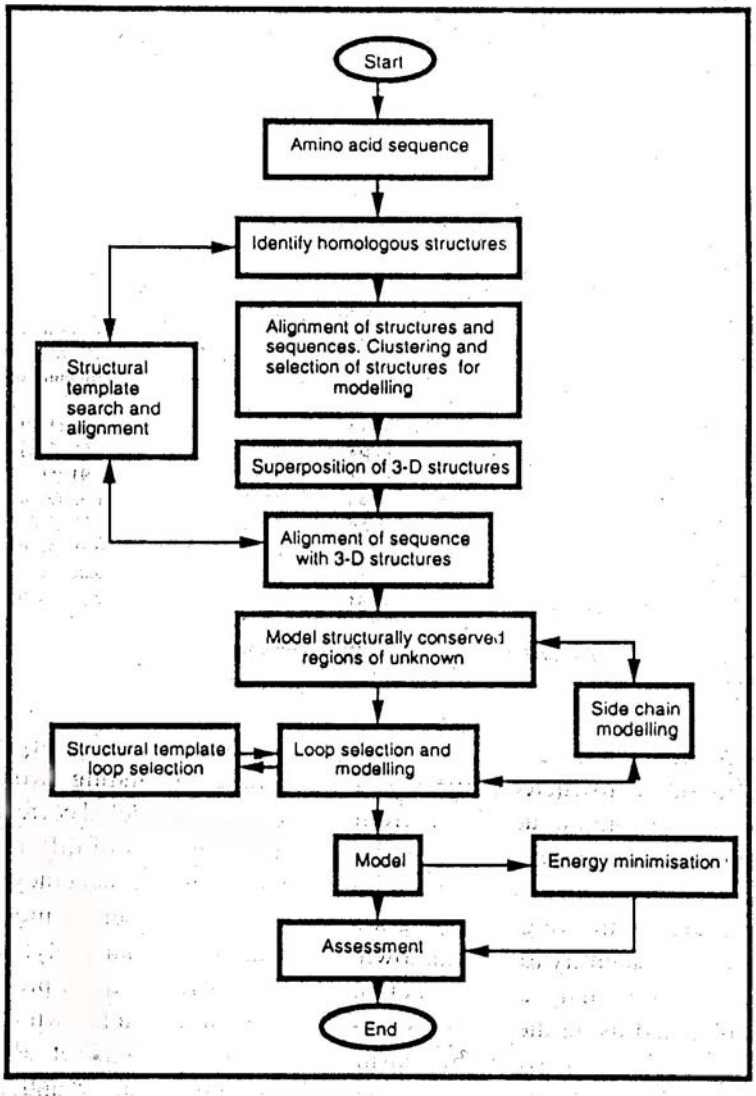

Figure 3.1 Showing the flowchart of methodology implimented in COMPOSER homology modeling program.

distances with a check that the modeled loop does not clash with the rest of the proteins. The identified region is usually fitted to anchor regions (the ends of the intervening regions in the model that are mainly the helices and strands). The selection of the correct conformation can be improved by considering the r.m.s. difference in the anchor regions and sequence similarity between the identified segment and one to be modeled.

Candidate loops can also be ranked by using structural templates (Topham *et al*., 1993). The templates reflect amino acid substitutions that are compatible with the local structural environment for each amino acid defined in terms of main chain conformation, solvent accessibility, hydrogen bonding, disulfide bonding, and *cis*-peptide conformation (Overington *et al*., 1990; 1992). The side chains are modeled depending on the orientation of the side chains in the equivalent positions in the known homologues or based on a large number of rules derived for their preferred conformations in various secondary structures (Sutcliffe *et al*., 1987b). Other techniques, including energy minimization and localized molecular dynamics can then be applied to the model.

3.3.2.5 MODELLER

MODELLER is an implementation of an automated approach to comparative protein structure modeling by satisfaction of spatial restraints extrapolated from homologous 3D-structures to the sequences to be modeled (Sali and Blundell, 1993, Sali *et al*., 1995). The modeling procedure begins with an alignment of the sequence to be modeled (target) with related known structures (templates). This alignment is usually the input to the program. The output is a 3D model for the target sequence containing all main chain and side chain non-hydrogen atoms.

First, many distance and dihedral angle restraints on the sequence are calculated from its alignment with template 3D structures. The form of these restraints was obtained from a statistical analysis of the relationship between many pairs of homologous structures. This analysis relied on the database of 105 family alignments that included 146 known structures (Sali and Overington, 1994). By scanning the database, tables quantifying various correlations were obtained, such as correlations between two equivalent C_α–C_α distances, or between equivalent main chain dihedral angles from two related proteins. These relationships were expressed as conditional probability density functions (pdf's) and can be used directly as spatial restraints. For example, probabilities for different values of the main chain dihedral angles are calculated from the type of a residue

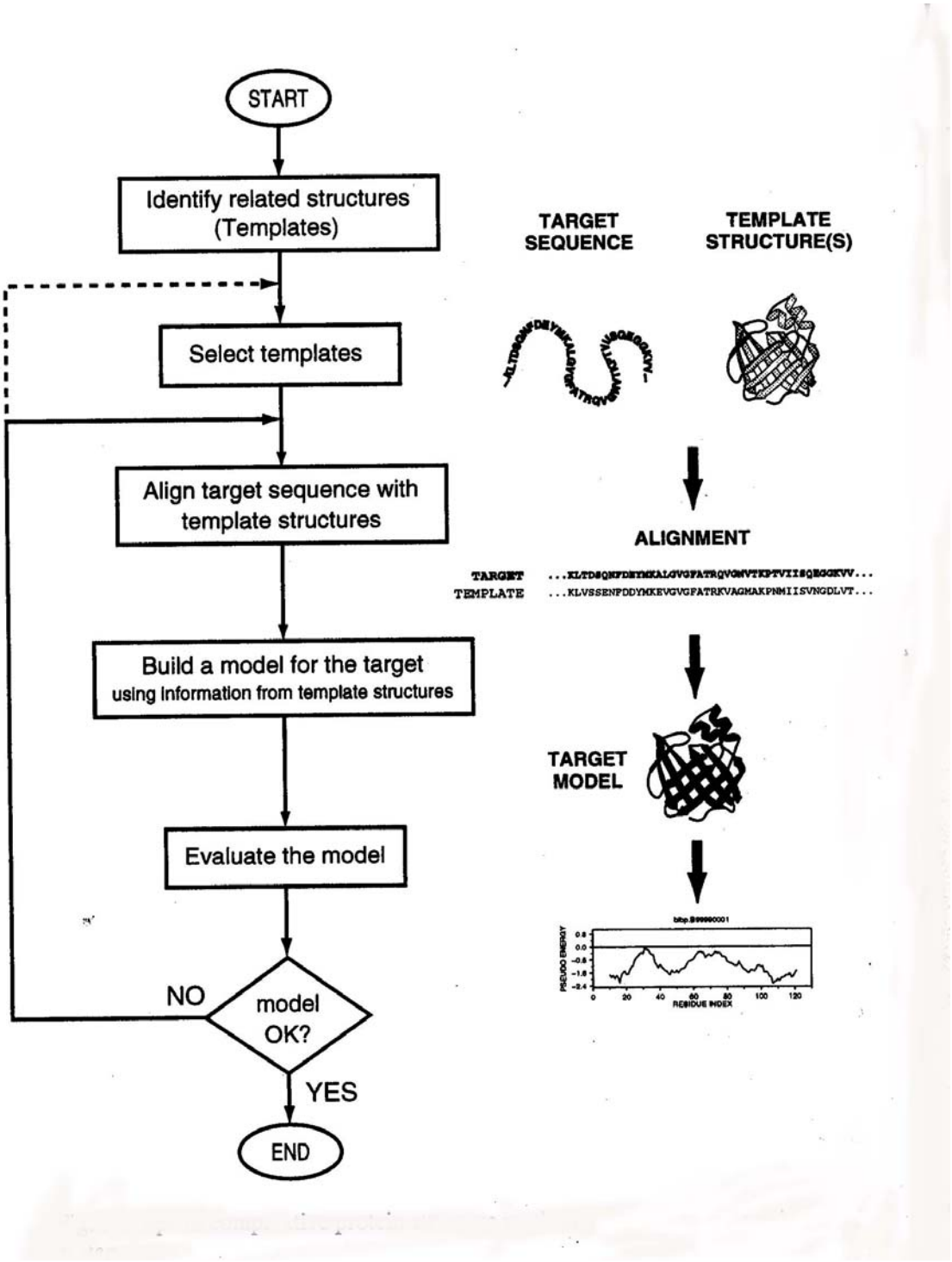

Figure 3.2 Flowchart showing methodology implimented in homology program MODELLER.

considered, form the main chain conformation of an equivalent residue, and form the sequence similarity between the two proteins. Another example is the pdf for a certain C_α–C_α distance given equivalent distances in two related protein structures. An important feature of the method is that the spatial restraints are obtained empirically, from the database of protein structure alignments. Next, the spatial restraints and CHARMM energy terms enforcing proper stereochemistry are combined in to an objective function. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method (Braun and Go, 1985) employing methods of conjugate gradients and molecular dynamics with simulated annealing.

Several slightly different models can be calculated by varying the initial structure. The variability among these models can be used to estimate the errors in the corresponding regions of the fold. MODELLER evaluates the model internally. The internal self-consistency check is that the model has to satisfy most restraints used to calculate it, especially the stereochemical restraints. If some restraints are grossly violated in all models it is likely that the alignment in the corresponding region is incorrect. The restraint violations are reported at the end of the log file.

3.4 Model Evaluation

Evaluation of the 3D model is an essential step that can be performed at different levels of structural organization, namely, to identify (1) the correctness of the overall fold, (2) detect errors over more localized regions, and (3) check stereochemical parameters like bond lengths, bond angles, and hydrogen bond geometry.

Model Evaluation programs and sites

- PROCHECK www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
- WHATCHECK www.sander.embl-heidelberg.de/whatcheck
- PROSAII www.came.sbg.ac.at
- PROCYON www.horus.com/sippl/
- BIOTECH biotech.embl-ebi.ac.uk:8400/
- VERIFY3D www.doe-mbi.ucla.edu/verify3d.html
- ERRAT www.doe-mbi.ucla.edu/errat_server.html

It is recommended to evaluate the model obtained by homology modeling for errors. Various programs that are developed for checking the quality of protein structures are also used for checking quality of models derived from homology modeling.

3.4.1 PROCHECK

PROCHECK (Laskowski *et al*., 1993) makes use of properties originally derived from a set of 119 non-homologous protein crystal structures at a resolution of 2.0 Å or higher and having an R-factor no greater than 20% (Morris *et al*., 1992). It checks the stereochemistry using C_α chirality, Percentage of residues found (more than 90%) in the core region of Ramachandran plot, torsion angles for secondary structures and χ_1, χ_2, χ_3 torsional angles etc. It also calculates the main chain hydrogen bond energy. The output is a series of postscript files. The most important file is the one that gives the Ramachandran plot, which has been discussed extensively before.

Chapter4

Study of Structure and Principles of Ion Channeling: Analysis of Ip_3R and RyR Sequences towards Ca^{2+} Channeling

4.1 Summary

Inositol 1,4,5-triphosphate receptors (Ip_3Rs) and ryanodine receptors (RyR) act as cationic channels transporting calcium ions from the endoplasmic reticulum to cytosol (Berridge and Irvine, 1989) by forming tetramers and are proteins localized to the Endoplasmic Reticulum (ER). Despite the absence of classical calcium-binding motifs, calcium channeling occurs at the transmembrane domain. Putative calcium binding motifs have been investigated in these sequences. Prediction methods indicate the presence of six transmembrane helices in the C-terminal domain, one of the three domains conserved between Ip_3R and RyR receptors. Recently, the crystal structure of tetrameric K^+ channel (Doyle *et al*., 1998) revealed that two transmembrane helices, an additional pore helix and a selectivity filter are responsible for selective ion K^+ channeling. The last three TM helices of Ip_3R and RyR are particularly well-conserved and analogous pore helix and selectivity filter motif is found in these sequences. Three-dimensional structural model for permeation pathway of the channel tetramer is generated by extrapolating the distant structural similarity to the K^+ channels.

4.2 Signal Transduction Pathways

The release of intracellular Ca^{2+} is an intermediate step in many cellular signaling processes (Berridge and Irvine, 1989; Tsein and Tsein, 1990). In vertebrates, two classes of proteins,

the Inositol 1,4,5-triphosphate receptor (Ip_3R) and the ryanodine receptor (RyR), act as channels for the release of intracellular Ca^{2+}. Ip_3R causes release of intracellular Ca^{2+} in response to Ip3 which is generated during signaling mechanisms that involves the activation of phospholipase C (Majerus *et al*., 1985). This signal transduction pathway is used in processes as diverse as the response to hormones, growth factors and neurotransmitters (Berridge and Irvine, 1984), as well as various sensory systems such as olfaction (Reed, 1992), gustation (Hwang *et al*., 1990) and vision (Payne *et al*., 1988; for a recent review see Patel *et al*., 1999). Ip_3R pathway must also function in the central brain, the tissue from which it was initially purified and cloned (Furuichi *et al*., 1989; Mignery *et al*., 1990).

Ryanodine receptor function is best understood in vertebrate skeletal muscle. It is required for the intracellular Ca^{2+} release that occurs prior to muscle contraction, in response to nerve impulses delivered to the muscle plasma membrane (Caterrall, 1991). The other two RyR isoforms are often referred to as the 'heart' and 'brain' forms, but the actual cell and tissue distribution of the isoforms is more complex than is suggested by this nomenclature (For reviews see Coronado *et al*., 1994; Meissner, 1994; Striggow and Ehrlich, 1996). Functional studies have shown that the channel may be regulated by various endogenous effector molecules including Ca^{2+}, ATP, cADP ribose and calmodulin, depending upon the isoforms. In addition, both Ip3R and RyR have been postulated to function during Ca^{2+}-induced Ca^{2+}-release in neuronal and non-neuronal tissues requiring Ca^{2+} oscillations (Tsein and Tsein, 1990). The presence of these intracellular Ca^{2+} channels in such diverse tissues indicates that they are likely to be involved in many different cellular functions. Ip_3R and RyR are thought to occur as homotetramers. Their monomers are of length ~3000 and ~5000 amino acids respectively (Mignery *et al.*, 1989; Serysheva *et al.*, 1995; Galvan *et al*., 1999).

4.3 Materials and Methods

The sequences of ip3r_rat, rynr_human and ip3r_drome have been extracted from the SWISSPROT protein sequence database (Appel *et al*., 1994). Blast searches were made in PRODOM database (Altschul *et al*., 1999). Sequences have been aligned using ClustalW multiple alignment program (Thompson *et al*., 1994). Several methods, both for secondary

structure prediction and membrane spanning region prediction, were used: PHD (Rost *et al*., 1995), PREDATOR (Frishman and Argos, 1997), JPRED (Cuff *et al*., 1998) were used to obtain secondary structure prediction. For the prediction of membrane spanning regions, PERSCAN (is a general purpose method; Donnelly *et al*., 1994), PHD (Rost *et al*., 1995), HMMTOP (Tusnády *et al*., 1998), TMHMM (Sonnhammer et al., 1998), TMPRED (Hofmann and Stoffel, 1993), SOSUI (Hirokawa *et al*., 1998) and TOPPRED II (Claros *et al*., 1994) were used. The comparative modeling program, COMPOSER (Sutcliffe *et al*., 1987; Blundell *et al*., 1988; Srinivasan and Blundell, 1993) was used to derive the three-dimensional structure of the last two TM helices of Ip_3R and RyRs. The tetramer co-ordinates were obtained by means of rigid-body superposition from the K^+ channel tetramer co-ordinates using the program SUPER (B.S. Neela, personal communication). The protomers were moved systematically away from the pore axis using SCHELAX (Chou *et al.*, *1984*; Sowdhamini *et al*., 1992) by 1.5 Å to suit the reported dimensions of Ca^{2+} channels.

4.4 Results and Discussions

4.4.1 Calcium-binding Sites on Primary Sequence

Ip_3R and RyR are poorly selective and high conductance Ca^{2+} channel with estimated permeability ratio (divalent/monovalent) of both the receptors is nearly six (Bezprozvanny and Ehrlich, 1994; Tinker and Williams, 1992). Calcium is known to be a regulator of both the receptor channels. Both the properties demand existence of calcium binding motifs on the channel structure. However no classical calcium binding motifs are reported for both the receptor channels. Analysis of individual domains suggested by PRODOM (Corpet *et al*., 1999) has been carried out for this purpose. The domain arrangement of Ip_3R and RyR, with putative helix transmembrane helix positions (see later), as suggested by PRODOM is as shown in Figure 4.1 a, b.

Table 1: Putative calcium binding sites in inositol triphosphate $Insp_3R$ and RyR

Amino acid From @	Amino acid to@	Linear Sequence	Comment
97	107	**DLEKKQNETEN**	
228*	255	**D**NK**D**DILKGG**D**VVRLFHAEQ**E**KFLTC *D*E	$Found conserved in domain 1922
317*#	381	**E**V**D**PDF*EEE*CL**E**FQPSVDP*D*QDASRSR LRNAQ**E**KMVYSLVSVP**E**GN*D*ISSIF*E*L **D**PTTLRGG*D*S L	$Found conserved in domain 1922
378*#	450	*D*SLVPRNSYVRLRHLCTNTWVHSTNI PI*D*K**EE**EK PVMLKIGTSPLK**ED**KEAFAIVPVSPAE VR*D*L*D*FAN**D**AS	$Found conserved in domain 1922
528*	544	*D*CG**D**GPMLRL**EE**LG**D**Q	$Found conserved in domain 1922
660*#	733	TNA**D**ILIETKLVLSRFEFEGVSTGENAL EAGED**EEE**VWLFWRDSNK**E**IRSKSV RELAQ**D**AKEGQKE**D**R DILSYY	Found at boundary of domain 1922
741#	849	ARMCL**D**RQYLAINEISGQLDV**D**LILRC MS**DE**NLPY**D/D**R**D**PQ**E**QVTPVKYARL WS**E**IPSEIAID**D**Y**D**SSGTSKDEIKERFA QTMEFV**EE**YLRDVVC	
994#	1059	LCIFKR**E**FD**E**SNSQSS**E**TSSGNSSQEGPS NVPGAL**DF**EHIEEQA**E**GIFGGSEENTP LDL**D**DHGGRT	
1107	1121	Q**D**V**D**NYKQIKQ**D**L**D**Q	
1140	1157	DEPMDGASGENEHKKTEE	Unstructured charged loop
1347#	1426	**D**RASFQTLIQMMRSER**D**RM**DE**NSPL	

		MYHIHLV**E** LLAVCT**E**GKNVYTEIKCNSLLPL**DD**IV RVVTH**ED** CIP**E**VKIAYINFL	
1685	1719	**D**RGYG**E**KQISI**D**ES**E**NA**E**LPQAP**E**A**E** NST**E**Q**E**L**E**P	
2124*#	2146	IKKAYMQGEVEFEDGENGEDGAA	Found at boundary of domain 2036 and replaced by two EF- hands in RyR, unstructured loop
2178	2186	QVDGDEALE	Unstructured charged loop
2463#	2528	K**DD**FILEV**D**RLPNETAVPETG**E**SLAND FLYSDVCRVETG**E**NCTSPAPK E**E**LLPAEETEQDK**E**HTC**E**	Part of luminal loop, domain 1555, Replaced by a charged region in RyR
2589*	2604	**D**TFA**D**LRS**E**KQKK***E***E	Found conserved in domain 1555

@ corresponds to ip3r_mouse residue numbering

* stretch of residues are found conserved in both Insp3R and RyR.

stretch of residues are reported to bind calcium (Sienaert *et al* ., 1996,1997).

$ - Domain 1922 is N-terminal, which is reported to be ligand binding domain in Insp3R (Miyawaki *et al*., 1991).

/ indicates gap in the sequence.

Amino acids in bold letters indicate the conserved charged residues, when both families are compared. Conservation only in $Insp_3R$ is shown in italics.

The domain numbering is as follows: Domain 1922 corresponds to N-terminal residues 180-650. Domain 2036 corresponds to middle region of residues 1963-2131 and domain 1555 corresponds to C-terminal region of residues 2382-2674 (numbering according to ip3r_mouse).

PRODOM records the N-terminal domain (domain id PD001922) of around 550 amino acids with ip3r_mouse-numbering 143-671, and rynr_human-numbering 180-650 to be similar. Interestingly enough, the N-terminal domain in case of Ip3R is shown to be the ligand binding domain (Mignery and Sudhof, 1990; Miyawaki *et al*., 1991). Furthermore, a middle domain of 168 amino acids (domain id PD002036; ip3r_mouse-numbering 1963-2131 and

rynr_human- numbering 3751-4123) shares high sequence similarity. The C-terminal transmembrane domain is divided into more than one domain according to PRODOM and a region of around 300 amino acids (domain id PD001555) ip3r_mouse-numbering 2382-2674 and rynr_human-numbering 4612-5032 shares relatively high sequence similarity (36% sequence identity).

12 Ip_3R sequences and 13 RyR sequences were chosen and aligned at the membrane-traversing transmembrane (TM) domain. The multiple alignment of Ip_3R and RyR sequences show the presence of several conserved negatively charged residues (Table 1) which could act as Ca^{2+} binding sites. While studying Ca^{2+}regulation of Ip_3R receptor at the molecular level and the structural determinants of Ca^{2+} binding, Sienaert and co-workers (Sienaert *et al*, 1996; 1997) had identified 8 linear sites which are shown to bind both calcium and ruthenium red (see Table1). Out of 8 sites, 3 are in regions where the two classes of receptors share high sequence identity. The regulatory calcium binding sites are therefore novel conserved motifs. Two EF-hand Ca^{2+} binding domains have been identified in Lobster skeletal muscle RyR, (Xiong *et al*., 1998) at positions (numbering according to rynr_human) 4070-4130, which are at the boundary of the middle domain which is conserved between Ip_3R and RyR receptors. Ip_3R, however, does not contain an equivalent EF-hand motif, but is replaced by an aspartate-glutamate rich region (2124-2146 of ip3r_mouse) which is shown to bind Ca^{2+} (Sienaert *et al*., 1997). Conversely, a region from ip3r_mouse (amino acids 2463-2528) which is the part of C-terminal domain is shown to bind Ca^{2+} but corresponding region in rynr_human is replaced by highly aspartate and glutamate rich region. Thus, the elements that are involved in binding calcium ions on primary structure are conserved and indicate the similar mode of regulation by Ca^{2+}.

4.4.2 Lessons from K+ channel structure

Recently structure of tetrameric K^+ channel (Doyle *et al*., 1998) from *S. Lividens* was reported, revealing many mysteries about the channel structures that had kept physiologists wondering for many decades. Apart from two membrane spanning helices, the loop region connecting the two helices (P-loop) forms the selectivity filter. The amino terminal region of

the P loop is also α-helical (which is termed as pore helix), slanting towards the pore axis from outside. The pore helix is followed by a signature sequence - Five amino acids in this zone, corresponding to VGYGD, form the lining of the selectivity filter orienting their main chain carbonyls towards the pore axis and their side chains outward thus stabilising the right ions of desired pore size. Sequence alignments from various K^+ channels, both inward and outward rectifiers, shows that most of the residues of pore helix and signature sequence are conserved (Doyle *et al*, 1998; MacKinnon *et al*., 1998; Armstrong, 1998), suggesting that the architecture of the channels is similar irrespective of the direction of ion transfer. Moreover, two membrane spanning helices, pore helix and selectivity filter per monomer would be the minimal requirement and sufficient for forming the functional channel tetramer.

4.4.3 Secondary Structure Prediction Studies on C-terminal Region

Prediction studies were carried out using methods that use both single sequence and multiple alignment on the sequences of one Ip_3R and one RyR, to map the putative transmembrane region on both the receptors. Various transmembrane region prediction methods available on SWISSPROT server (www.expasy.ch) were employed. The results from various methods with the predicted positions of the transmembrane helices are shown in Figure 4.2 for ip3r_mouse sequence. It is interesting to note that the helix marked as "pore helix" is predicted as a membrane-spanning region by three transmembrane region prediction methods while others miss it. However, it is predicted as a helix by all secondary structure prediction methods. Thus, confirming its existence as a helix. The existence of pore helix was confirmed also by applying these methods to the KcsA sequence, where all membrane region prediction methods miss the pore helix. The helix-wheel diagram is shown in Figure 4.3 for the region predicted to contain the sixth TM helix of ip3r_mouse by PERSCAN (Donnelly *et al.*, 1994). It is clear from the prediction studies reported that Ip_3R contains a topology of six membrane spanning helices. Prediction analysis was also performed for the C-terminal domain of ryanodine receptors. PHD TMpred, a method that employs multiple sequence alignments, suggests six membrane spanning helices and a pore helix for RyR, a topology analogous to that suggested for Ip_3R. All the other membrane region prediction methods predict different number of membrane spanning regions, but for the last two membrane spanning helices and

pore helix the results are identical to that for Inp_3R.

The pore helix is predicted in the loop region between the putative fifth and sixth membrane-spanning helices of the receptors, which is known to be analogous to P-loop of voltage-activated Ca^{2+}, Na^{+}, and K^{+} channels (Mignery and Sudhof, 1993). It is also implicated to be the pore-forming segment (Balshaw *et al*., 1999). Figure 4.4 shows the multiple sequence alignment of the region containing putative last three helices of both the receptors where the highest sequence similarity extends to a further 100 amino acids towards the C-terminus (36% I.D.) The predicted helix positions and certain conserved amino acid positions are indicated. This observed similarity is also in agreement with deletion studies on Ip_3R which demonstrates that the deletion of the first four TM helices of recombinant Ip_3R forms functional calcium channels and mutants lacking the last two helices do not form detectable channels (Ramos-Franco *et al*., 1999). The results of secondary structure prediction, inspection of sequence alignment and the deletion studies (Ramos-Franco *et al*., 1999) strongly suggests that the pore forming regions for both Ip_3R and RyR are similar and conserved.

4.4.4 Structural Paramemters for Calcium channel

From the above discussion and sequence alignment shown in Figure 4.4, it is clear that the conserved C-terminal region also contains the predicted pore helix, which has a length of 10 amino-acid residues. Following the pore helix, a motif, GXRXGGGXGD (starting from 4820 of RyRs and 2540 of Ip_3Rs) is found to be highly conserved, in all known Ip_3Rs and RyR. Mutations of glycine to alanine in this signature sequence in RyR, at first, fourth and sixth positions disrupt the calcium release from the channel (Zhao *et al*., 1999). Also isoleucine to threonine mutation of RyR1 (see Figure 4.4) decreases the threshold of Ca^{2+} requiring to initiate opening of wild type channel and resulted in a reduced release of Ca^{2+} from internal stores (Balshaw *et al*., 1999; Lynch *et al*., 1999). These data suggest that this conserved region constitutes channel conduction pathway or the central pore lining of this receptor (Zhao *et al*., 1999) reaffirming that the same topology is present in the channel forming

region as in the KcsA K^+ channel, *viz.* fifth helix, pore helix, pore-lining region and sixth helix. It is anticipated in earlier studies that Ca^{2+} channels have pores that are related architecturally to K^+ channels (Roux and MacKinnon 1999; Doyle *et al*., 1998).

Owing to the difference in mechanism of cation conduction, it is obvious that the structural parameters are different for the RyR Ca^{2+} channels than K^+ channels. Ryanodine receptors are reported to have a pore of diameter of ~6-7 Å (McCleskey and Almers, 1985; Tinker and Williams, 1993; Serysheva *et al*., 1999). The length of selectivity filter region is found to be 10.4 Å (Tinker and Williams, 1993; 1995), which is in good agreement with the KcsA selectivity filter length of 12 Å. Before the structure of K^+ channel was determined the experimental value of selectivity filter of K^+ selective channel was reported as 10 Å (Miller, 1982). Blocking studies with the impermeant charged derivative of triethyl amine reveal that this narrowing occurs over first 10-20% of the voltage drop when crossing from the lumen of SR to the cytoplasm showing that the narrow region (selectivity filter) occurs at the luminal mouth of the channel.

4.4.5 Building the Structure of Permeation Pathway

The three-dimensional structure of RyR human TM domain was derived using KcsA structure as the template and by employing the COMPOSER homology modeling program (Sutcliffe *et al*., 1987; Blundell *et al*., 1988; Srinivasan and Blundell, 1993). The length of KcsA sequence and that of RyR C-terminal regions that contains the pore forming region are similar, but both the sequences shares very low sequence similarity (8% ID).

The transmembrane helices, pore helix and selectivity filter region are taken as SCR (structurally conserved regions; Figure 4.5) and the resulting structure is energy minimized with a fixed backbone conformation. The tetramer positions of the calcium channel are generated from the K^+ channel tetramer by a structure superposition program called SUPER (Neela, B.S., personal communication). The pore diameter of RyR is 6Å, wider than that of K^+ channel by 3Å as mentioned above. Therefore, in the tetramer of the TM domain, each

monomer was moved 1.5Å away from the pore axis symmetrically, to suit the reported structural parameters. Interprotomer interactions before and after the change in pore dimensions were measured and no major destabilization was found due to the slight enlargement in pore diameter.

The tetramer model has a pore diameter of roughly 6 Å and selectivity filter length around 11 Å, (Figure 4.6) in correspondence with functioning calcium channels. This model satisfies most of the properties of calcium channels both used by binding model and continuum model (Hille B, 1992; Nooner and Eisenberg, 1998; Doyle *et al*., 1998). Figure 4.7 shows the ribbon diagram of the tetramer model of ryr_human derived by such comparative modeling studies. The presence of leucines and other hydrophobic residues in two adjacent protomers at the protomer interface might account for the stability of the tetramer. The model is in agreement with present theory of calcium permeation through large pores, which have large diameters than their preferred ions. The calcium is concentrated by the negatively charged residues, which are concentrated at the mouth of the pore (Figure 4.8a) passing through the selectivity filter region composed of the conserved motif, GGGIGD, which occurs at the luminal mouth of the channel. It can be stabilized by dipole moments of the pore helices and also water molecules present in the middle of the pore (as shown by structure of K^+ channel) and then it passes through the remainder of the pore. This narrow region is relatively short which is consistent with the large conductance of the channel (Latorre and Miller, 1983). The hydrophobic membrane spanning helices form the hydrophobic walls (Figure 4.8b). The pore helices, which are pointing towards the central axis of the pore provide the stabilization to the ions inside the pore by its dipole moments and also holds the amino acids of the selectivity filter region firmly at their position. (Figure 4.9)

4.5 Conclusions

To conclude, this chapter reports the regions of Ip_3R and RyR, which share high similarity and its importance for Ca^{2+} binding and channel regulation, are identified. High degree of partial sequence similarity between the two receptors suggests that the elements involved in calcium channel formation, regulation and selectivity are highly similar and conserved during

evolution.

It is well-known that all of the known Na^+, Ca^{2+} and K^+ channels are made of tetramers of either four internal repeats each containing six membrane spanning helices or four protomers each having six membrane spanning helices (see for example, Hille, 1992). Else some channels are tetramer of two transmembrane spanning α-helices.

On the basis of structural principles exemplified by the KcsA K^+ channel structure (Doyle *et al*., 1998), the first atomic level structure of a calcium channel has been proposed as a multi ion-single file pore. It is in agreement with existing structural and theoretical studies, which provides clues to the permeation pathway located in the linear sequence and how calcium ions might pass through it. The novel finding of this work is identification of pore helix in Ip_3R or RyR. The above analysis also confirms that the cationic channel proteins belong to a broad superfamily with highly conserved structures. It will be interesting to compare the four internal repeats of the Na^+ channels for similarities in secondary structural features.

4.6 References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller,W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Res*., **25**, 3389-3402.

Appel, R.D., Bairoch, A. and Hochstrasser, D.F. (1994). A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.,* **19**, 258-260.

Armstrong, C. (1998). The vision of the pore. *Science,* **280**, 56-57.

Balshaw, D., Gao, L. and Meissner, G. (1999). Luminal loop of the ryanodine receptor: a pore-forming segment? *Proc. Natl. Acad. Sci. U S A*., **30**, 3345-3347.

Berridge, M.J. and Irvine, R.F. (1984). *Nature,* 312, 315-320.

Berridge, M. J. and Irvine, R.F. (1989). Inositol phosphates and cell signalling. *Nature,* **341**, 197-205.

Bezprozvanny, I. and Ehrlich, B.E. (1994). Inositol (1,4,5)-trisphosphate (InsP3)-gated Ca channels from cerebellum: conduction properties for divalent cations and regulation by intraluminal calcium. *J Gen Physiol.,* **104**, 821-856.

Blundell, T. L., Carney, D., Gardner, S., Hayes ,F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibdana, B.L. and Sutcliffe, M. (1988). Knowledge-based protein modelling and design *Eur. J. Biochem*., **172**, 513-520.

Catterall, W.A. (1991). Excitation-contraction coupling in vertebrate skeletal muscle: a tale of two calcium channels. *Cell,* **64**, 871-874.

Chou, K.C., Nemethy,G. and Scheraga,H.A. (1984). Energetic approach to the packing of alpha-helices. 2. General treatment of nonequivalent and nonregular helices. *J. Amer. Chem. Soc*., **106**, 3161-3170.

Claros, M.G. and von Heijne,G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci*., **10**, 685-6.

Coronado, R., Morrissette, J., Sukhareva, M. and Vaughan, D.M. (1994). Structure and function of ryanodine receptors. *Am. J. Physiol*., **266**, C1485-C1504.

Corpet, F., Gouzy, J. and Kahn, D. (1999). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*., **27**, 263-267.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, A.M. and Barton, G.J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892-893.

Donnelly, D., Overington, J.P. and Blundell, T.L. (1994). The prediction and orientation of alpha-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules. *Protein Engng*., **7**, 645-653.

Doyle, D.A., Cabral, J.M., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L, Chait, B.T. and MacKinnon, R.(1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science,* **280**, 69-77.

Evans, S.V. (1993). SETOR: Hardware lighted three-demensional solid model representations of macromolecules, J. Mol. Graphics, **11**:134-138.

Frishman, D. and Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329-335.

Furuichi, T., Yoshikawa, S., Miyawaki, A., Wada, K., Maeda, N. and Mikoshiba, K. (1989). Primary structure and functional expression of the inositol 1,4,5-trisphosphate-binding protein P400. *Nature*, **342**, 32-38.

Galvan, D.L., Borrego-Diaz, E., Perez, P.J. and Mignery, G.A. (1999). Subunit oligomerization, and topology of the inositol 1,4, 5-trisphosphate receptor. *J. Biol. Chem*., **274**, 29483-29492.

Hille, B. (1992). In Ionic channels of excitable membranes. (Hille, B.,ed), pp. 250-254, Sinauer Associates,Inc., Mass.

Hirokawa, T., Boon-Chieng, S. and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378-379.

Hofmann, K. and Stoffel, W. (1993). TMbase: A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler,* **347**,166.

Hwang, P.M., Verma, A., Bredt, D.S. and Snyder, S.H. (1990). Localization of phosphatidylinositol signaling components in rat taste cells: role in bitter taste transduction. *Proc. Natl. Acad. Sci. USA,* **87**, 7395-7399.

Kraulis, P.J. (1991). MOLSCRIPT: A program to produce both Detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946-950.

Latorre, R. and Miller, C. (1983). Conduction and selectivity in potassium channels. *J. Membr. Biol*., **71**, 11-30.

Lynch, P. J., Tong, J., Lehane, M., Mallet, A., Giblin, L., Heffron, J., Vaughan, P., Zafra, G., MacLennan, D.H. and McCarthy, T.V. (1999). A mutation in the transmembrane/lumenal domain of the ryanodine receptor is associated with abnormal Ca^{2+} release channel function and severe central core disease. *Proc. Natl. Acad. Sci. USA*, **96**, 4164-4169.
MacKinnon., R, Cohen, S.L., Kuo, A., Lee, A. and Chait, B.T. (1998). Structural conservation in prokaryotic and eukaryotic potassium channels. *Science,* **280**, 106-109.

Majerus, P.W., Wilson, D.B., Connolly, T.M., Bross, T.E. and Meufeld, E.J. (1985). Phosphoinositide turnover provides a link in stimulus-response coupling. *Trends Biochem. Sci*., **10**, 168-171.

McCleskey, E.W. and Almers, W. (1985). The Ca channel in skeletal muscle is a large pore. *Proc. Natl. Acad. Sci. USA*, **82**, 7149-7153.

Meissner, G. (1994). Ryanodine receptor/Ca2+ release channels and their regulation by endogenous effectors. In *Annu. Rev. Physiol*., pp. 485-508. Annual reviews Inc.

Michikawa, T., Hamanaka, H., Otsu, H., Yamamoto, A., Miyawaki, A., Furuichi, T., Tashiro, Y. and Mikoshiba, K. (1994). Transmembrane topology and sites of N-glycosylation of inositol 1,4,5-trisphosphate receptor. *J. Biol. Chem.,* **269**, 9149-9198.

Mignery, G.A., Sudhof, T.C., Takei, K. and DeCamilli, P. (1989). Putative receptor for inositol 1,4,5-Trisphosphate similar to ryanodine receptor. *Nature,* **342**, 192-195.

Mignery, G.A. and Sudhof, T.C. (1990). The ligand-binding site and transduction mechanism in the inositol-1,4,5-trisphosphate receptor. *EMBO J*., **9**, 3893-3898.

Mignery, G.A., Newton, C.I., Archer III, B.T. and Sudhof, T.C. (1990). Structure and expression of the rat inositol 1,4,5-trisphosphate receptor. *J. Biol. Chem*., **265**, 12769-12685.

Mignery, G.A. and Sudhof, T.C. (1993). Molecular analysis of Inositol 1,4,5 triphosphate receptor. *Methods Neurosci.,* 18, 247-265

Miller, C. (1982). Coupling of water and ion fluxes in a K+-selective channel of sarcoplasmic reticulum. *Biophys. J.,* 38, 227-30.

Miyawaki, A., Furuichi, T., Ryou.,Y., Yoshikawa, S., Nagakawa,T., Saitoh, T. and Mikoshiba, K. (1991). Structure-function relationships of the mouse inositol 1,4,5-trisphosphate receptor. *Proc. Natl. Acad. Sci. USA*, 88, 4911-4915.

Nicholls, A., Bharadwaj, R. and Honig, B. (1993). GRASP--Graphical representation and analysis of surface properties. *Biophys. J*., 64, 166-170.

Nonner, W. and Eisenberg, B. (1998) Ion Permeation and Glutamate Residues Linked by Poisson-Nernst-Planck Theory in L-type Calcium Channels *Biophys. J.,* **75**, 1287-1305.

Patel, S., Joseph, S.K. and Thomas, A.P. (1999). Molecular properties of inositol 1,4,5-trisphosphate receptors. *Cell Calcium,* **25**, 247-264.

Payne, R., Walz, B., Levy, S. and Fein, A. (1988). The localization of calcium release by inositol trisphosphate in Limulus photoreceptors and its control by negative feedback. *Phil. Trans. R. Soc. Lond*., *B* **320**, 359-379.

Ramos-Franco, J., Galvan, D., Mignery,.G.A. and Fill, M. (1999). Location of the permeation pathway in the recombinant type 1 inositol 1,4,5-trisphosphate receptor. *J. Gen. Physiol.* **114**, 243-250.

Reed, RR. (1992). Signaling pathways in odorant detection. *Neuron,* **8**, 205-209.

Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995). Prediction of helical trans-membrane segments at 95% accuracy. *Prot Sci.,* 4, 521-533.

Roux,B. and MacKinnon,R. (1999). *Science,* 285, 100-102.

Serysheva, II., Orlova, EV., Chiu, W., Sherman, M.B., Hamilton, S.L. and van Heel, M. (1995). Electron cryomicroscopy and angular reconstitution used to visualize the skeletal muscle calcium elease channel. *Nat. Struct. Biol*., **2**, 18-24.

Serysheva, I.I., Schatz, M., van Heel,M., Chiu, W. and Hamilton, S.L. (1999). Structure of the Skeletal Muscle Calcium Release Channel Activared with Ca^{2+} and AMP-PCP. *Biophy. J.,* **77**, 1936-1944.

Sienaert, I., De Smedt, H., Parys, J.B., Missiaen, L., Valingen, S., Simpa, H. and Casteels, R. (1996). Characterization of a cytosolic and a luminal Ca^{2+} binding site in the type I inositol 1,4,5-trisphosphate receptor *J. Biol. Chem*., **271**, 27005-27012.

Sienaert, I., Missiaen, L., De Smedt, H., Parys, J.B., Simpa, H. and Casteels, R. (1997). Molecular and functional evidence for multiple Ca^{2+}-binding domains in the type 1 inositol

1,4,5-trisphosphate receptor *J Biol Chem*., **272**, 25899-25906.

Sonnhammer,E., von Heijne,G. and Krogh,A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc of sixth I S M B*, (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. and Sensen, C., eds), pp. 175-182, AAAI Press, Menlo Park, CA.

Sowdhamini,R., Srinivasan,N., Ramakrishnan,C. and Balaram,P. (1992). Orthogonal beta beta motifs in proteins. *J. Mol. Biol.*, 223, 845-851.

Srinivasan, N. and Blundell, T.L. (1993). An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Engng.,* **6**, 501-512.

Striggow, F. and Ehrlich, B.E. (1996). Ligand-Gated Calcium Channels inside and Out. *Curr. Op. Cell Biol*., **8**, 490-495.

Sutcliffe, M.J., Hayes, F.R. and Blundell, T.L. (1987). Three-Dimensional Frameworks Derived From The Simultaneous Superposition Of Multiple Structures *Protein Engng*., **1**, 377-384.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673-80.

Tinker,A. and Williams,A.J. (1992). Divalent cation conduction in the ryanodine receptor channel of sheep cardiac muscle sarcoplasmic reticulum. *J. Gen. Physiol.,* **100**, 479-93.

Tinker,A. and Williams,A.J. (1993). Probing the structure of the conduction pathway of the sheep cardiac sarcoplasmic reticulum calcium-release channel with permeant and impermeant organic cations. *J. Gen. Physiol*., **102**, 1107-1129.

Tinker, A. and Williams, A.J. (1995). Measuring the length of the pore of the sheep cardiac sarcoplasmic reticulum calcium-release channel using related trimethylammonium ions as molecular calipers. *Biophys. J.,* **68**, 111-120.

Tusnády, G.E. and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol*., **283**, 489-506.

Tsein, R.W. and Tsein, R.Y. (1990). Calcium channels, stores, and oscillations. *In Annual review of Cell Biology*, **6**. (ed. G. E. Palade). pp. 715-760. California: Annual Reviews, Inc.

Wagenknecht,T. and Radermacher,M. (1997). Ryanodine receptors: structure and macromolecular interactions. *Curr. Op. Struc. Biol.,* **7**, 258-265.

Zhao, M., Li, P., Li, X., Zhang, L., Winkfein, R.J. and Chen, S.R. (1999). Molecular identification of the ryanodine receptor pore-forming segment. *J. Biol. Chem.,* **274**, 25971-25974.

Figure Legends

Figure 4.1 Doamin structure of Ip_3R (4.1a) and RyR (4.1b) primary sequences as provided by PRODOM domain database (Corpet *et al*., 2000). It shows the domain level similarity shared by both the receptor channels and hence points to tertiary structure similarities.

Figure 4.2 Consensus secondary structure predicton of transmembrane (TM) helices in the C-terminal domain of Inositol triphosphate receptor (Ip_3R) sequence. Various methods used for predicting TM helix positions are mentioned (see text on Materials and Methods for details). Most methods identify six TM helices while a few of them predict an additional

shorter helix before the last TM helix.

Figure 4.3: Helix-wheel projection of the sixth TM helix using PERSCAN method (Donnelly *et al*., 1994). The helix positions are predicted by the periodicity and pattern in the occurrence of hydrophobic residues.

a) The occurrence of amino acids along the predicted TM helix is shown schematically, where helix is shown as a cylinder.

b) shows the distribution of conserved residues on the predicted TM helix projected down the helix axis. Several hydrophobic residues are distributed around the putative TM helix.

c) Fourier transform of this periodicity corresponds to an angle of 100° consistent with the prediction of an amphipathic membrane spanning α-helix.

Figure 4.4 Multiple sequence alignment of various Ip_3Rs and ryanodine receptors (RyRs) corresponding to the region of the C-terminal, transmembrane (TM) domain that has the highest sequence conservation across the two families (RYNR_PIG: ryanodine receptor, *Sus scrofa* skeletal muscle; RYNR_HUMAN: ryanodine receptor, *Homo sapiens* skeletal muscle; RYNR_RABIT: ryanodine receptor, *Oryctolagus cuniculus* skeletal muscle; O13054_EEEEE: ryanodine receptor ryr1 isoform, *Makaira nigricans*; Q91313_RANCA: alpha-ryanodine binding protein, *Rana catesbeiana*; Q15413_HUMAN: ryanodine receptor 3, *Homo sapiens* brain; Q95201_MUSVI: ryanodine receptor type 3, *Mustela vison*. Q91319_RANCA: beta-ryanodine binding protein, *Rana catesbeiana*; Q90985_CHICK: ryanodine receptor type 3, *Gallus gallus*; RYNC_RABIT: ryanodine receptor, *Oryctolagus cuniculus* cardiac muscle; Q92736_HUMAN: ryanodine receptor 2, *Homo sapiens* cardiac muscle; Q24500_DROME: ryanodine receptor, *Drosophila melanogaster*; P91905_CAEEL: ryanodine receptor, *Caenorhabditis elegans*; IP3R_DROME: inositol 1,4,5-trisphosphate-binding protein receptor, *Drosophila melanogaster*; O77089_PANAR: inositol 1,4,5-trisphosphate receptor, *Panulirus argus*; Q14643_HUMAN: human type 1 inositol 1,4,5-trisphosphate receptor, *Homo sapiens*; Q14460_HUMAN: inositol 1,4,5-trisphosphate receptor type 1, *Homo sapiens*; IP3R_RAT: inositol 1,4,5-trisphosphate-binding protein type 1 receptor, *Rattus norvegicus*; Q91908_XENLA: inositol 1,4,5-triphosphate receptor, *Xenopus laevis*; IP3R_MOUSE: inositol 1,4,5-trisphosphate-binding protein type 1 receptor, *Mus musculus*; IP3S_HUMAN: inositol 1,4,5-trisphosphate-binding protein type 2 receptor, *Homo sapiens*; IP3S_RAT: inositol 1,4,5-trisphosphate-binding protein type 2 receptor, *Rattus norvegicus*; Q14649_HUMAN: type 3 inositol 1,4,5-trisphosphate receptor, *Homo sapiens*; Q63269_RAT: inositol triphosphate receptor - subtype 3, *Rattus norvegicus*; O61193_CAEEL:E f33d4.2a protein, *Caenorhabditis elegans*).

The predicted TM helices 4, 5 and 6 are marked. The positions of the predicted functional

motifs, the pore helix and selectivity filter, are also indicated. Analogous motifs are shown to form the cationic pathway in K^+ channels (Doyle *et al*., 1998).

Figure 4.5 The alignment shows the Structurally Conserved Regions (SCRs) used by COMPOSER (Sutcliffe *et al*., 1987) for generating the monomer model of permeation pathway in RyR on the basis of 1bl8 structure.

Figure 4.6 A portion of three-dimensional model of the permeation pathway in RyR showing the structural parameters. The figure was prepared using SETOR (Evans, 1993)

Figure 4.7 Three-dimensional model of the last two predicted TM helices of human ryanodine receptor. This corresponds to the region that is most conserved between ryanodine receptors (RyRs) and inositol triphosphate receptors (Ip_3Rs). Owing to the similarity between Ca^{2+} channels and K^+ channels, the model has been built by extrapolating from the K^+ channel structure (Doyle *et al.*, 1998). Ca^{2+} ions pass through the pore helix and the selectivity filter. Two transmembrane helices are shown to be important for channeling activity. Ribbon diagram of the Ca^{2+} channel tetramer shown using MOLSCRIPT (Kraulis, 1991). Several hydrophobic residues (shown for two adjacent protomers) line the protomer interface along the TM helices and serve to stabilise the tetramer.

Figure 4.8 Electrostatic potential representation of the Ca^{2+} channel tetramer using GRASP (Nicholls *et al.*, 1993). Acidic residues are indicated by red patches and blue patches indicate basic residues.
a) the pore view (b) view down the tetramer pore helix axis.
A broad red patch at the mouth of the channel shown in this three-dimensional model of the tetramer might explain how Ca^{2+} ions are attracted towards the channel**.**

Figure 4.9 Ribbon diagramme showing three-dimensional model of the last two predicted TM helices of human ryanodine receptor. The pore helices and the selectivity filter regions are marked. The figure was preparewd using SETOR (Evans, 1993).

Chapter 5

Structural Determinants of Binding and Specificity in Transforming Growth Factor-Receptor Interactions.

5.1 Summary

The transforming growth factor (TGFβ) families of proteins are cytokines that occur as large number of homologous proteins. Three major subgroups of these proteins have been realized, the TGFβs, the activins/inhibins and the BMPs, with marked specificities for their receptors. Although structural information is available for some members of TGFβ family of ligands and receptors, very little is known about the way these growth factors interact with the extracellular domains of their cell surface receptors, especially receptor type2. The elements that are determinants of binding and specificity of the ligands are also poorly understood. The structure of the extracellular domain of the receptor is a three-finger fold similar to some of the toxin structures. Amino acid exchanges between multiply aligned homologous sequences of type2 receptors point to residues at the surface, specifically, finger1, as determinant of ligand specificity and complex formation. The 'knuckle' epitope of ligands is predicted as the surface that interacts with the type2 receptor. The residues on strands β2, β3, β7, β8 and the loop region joining β2-β3 and β7-β8 of the ligands are identified as determinants of binding and specificity. These results are independently supported by docking studies of the type2-receptor to the ligand dimer-type1-receptor complex.

5.2 Introduction

5.2.1 The system

The transforming growth factor β (TGFβ) family comprises a large number of structurally related polypeptide growth factors, each capable of regulating a fascinating array of cellular processes including cell proliferation, lineage determination, differentiation, motility, adhesion and death. Expressed in complex temporal and tissue-specific patterns, TGFβ and related factors play a prominent role in development, homeostasis, and repair of virtually all tissues in organisms (Massagué, 1998). For example, the founding member of the TGFβ1 family was identified as a regulator of mesenchymal growth and, separately, as an antimitogen in epithelial cells (Massagué, 1990; Roberts and Sporn, 1993). Activins were identified as endocrine regulators of pituitary function and, as inducers of mesoderm in frogs (Kingsley, 1994; Gaddy-Kurten *et al*., 1995). Bone morphogenic proteins (BMPs) were identified as bone repair factors and, independently, as dorsalizing agents in Drosophila (Hogan, 1996; Mehler *et al*., 1997). Nearly thirty members of the TGFβ family have been described in human and many orthologs are known in mouse, Xenopus and other vertebrates (Massagué, 1998; Hogan, 1996). Four are present in *Caenorhabditis elegans* (Padgett et al., 1998) and seven in *Drosophila melanogaster* (Raftery *et al*., 1999).

The family is divided into two general branches: the BMP/GDF (growth and differentiation factor) and TGFβ/Activin/Nodal branches, whose members have diverse, albeit often complementary effects. Additional members such as inhibin-α act as ligand antagonists. Some family members are expressed in a few cell types or for limited periods of time during development, whereas others are widespread during embryogenesis and in adult tissues. AMH/MIS (Anti-Müllerian hormone or Müllerian inhibiting substance) and GDF8/myostatin are examples of the former; TGFβ1 and BMP4 are of the latter (Massagué *et al*., 2000). TGFβ1-3 are ~70% conserved among themselves, while BMPs are ~60% identical among each other. TGFβs and BMPs share

~30% identity, while inhibinβB shares 30%, 40% and ~27% sequence identity with BMP7, BMP2 and TGFβs, respectively. *Dpp* (*decapentaplegic* protein of *Drosophila melanogaster*) shares 56, 72, 40 and ~31% sequence identity with BMP7, BMP2, inhibinβB and TGFβs, respectively. Glial cell line-derived neurotrophic factor (GDNF) and its subfamily members, undergo similar modes of dimerization as TGFβs, but share very low sequence similarities (~14%) with members of TGFβ family. GDNF subfamily, therefore, can be considered as a member of the broader 'cystine-knot' superfamily, which includes nerve growth factor and platelet-derived growth factor, that have similar protomer structures but display different modes of dimerization and share ~15% sequence similarity with TGFβ family (Saarma, 2000; Sowdhamini *et al*., 1998).

5.2.2 Signal Transduction Pathway

Members of TGFβ family of growth factors are synthesized as larger precursor molecules with an amino-terminal signal sequence and a pro-domain of varying size. These precursor proteins are usually cleaved at a dibasic or RXXR site to release a mature carboxy-terminal segment of 110-140 amino acids (Massagué, 1998; Murray-rust *et al*., 1993; Barr, 1991) and are biologically active as dimers. Members of TGFβ family regulate gene expression by bringing together two types of receptor serine/threonine kinases (Massagué, 1998), collectively known as TGFβ receptor family. Unlike other members of the TGFβ family, GDNF family ligands activate intracellular signaling cascades via the receptor tyrosine kinase Ret (Heldin *et al*., 1997). TGFβ receptor family is divided into two subfamilies: type1 receptors and type2 receptors, on the basis of their structural and functional properties. Table 5.1 summarizes various TGFβ family ligands and their receptors identified biologically. Two general modes of ligand binding have been observed: One mode involves direct ligand binding (The biologically active form of TGFβ ligands is dimer of two monomers. Henceforth, when ligand binding to receptor is discussed, it is assumed that the receptor(s) is interacting with ligand dimer) to ectodomain of the type2 receptor and subsequent interaction of this complex with the type1 receptor. Type1 receptor, in effect, becomes recruited to the complex, which is

characteristic of TGFβ and activin receptors. The second mode of binding is typical of BMP receptors and is cooperative, involving type1 and type2 receptor ectodomains that bind ligand with high affinity when expressed together but with low affinity when expressed separately (Massagué, 1998). In the mechanisms described above, type2 receptors bind to ligand dimers, subsequently (or simultaneously) recruit type1 receptors and finally phosphorylate type1 receptors at GS domain and thus activate them in transducing the signal to the nucleus via SMAD proteins. (Please see reviews by Massagué, 1998; Massagué *et al*., 2000; Miyazono *et al*., 2000; Ducy and Karsenty, 2000; Zimmerman and Padgett, 2000; Massagué, 2000; for the list of TGFβ family members, their activities and detailed signaling mechanism.)

5.2.3 Description of Structure of Ligands

The TGFβ isoforms show remarkable structural homology between each other, including seven absolutely conserved cysteine residues that form three intrachain disulfide bonds and one interchain disulfide bond. TGFβ and activins/inhibins contain an extra disulfide bridge at the N-terminus of the molecule. The structures of TGFβ2 (Daopin *et al*., 1992; Schlunegger and Grutter, 1993), TGFβ3 (Mittal *et al*., 1996), BMP7/OP1 (Griffith *et al*., 1996), BMP2 (Scheufler *et al*., 1999) and GDNF (Eigenbrot and Gerber, 1997) were determined by X-ray crystallography while a model of TGFβ1 (Hinck *et al*., 1996) was calculated from NMR restraints. The monomer is a thin, elongated and slightly curved molecule resembling an open left hand. As shown in Figure5.1a, each monomer is folded into nine β-strands (β1-β9) and a long α-helix (discussed below as α3). The fold can be described as a hand with the thumb as the N-terminus and the extended sheets as fingertips representing β2-β3 and β7-β8 loops. Accordingly, the convex surfaces of the fingers correspond to the knuckles and the helix region to the 'wrist'. The residues exposed on the convex surface involving β2, β3, β7 and β8 strands and loops joining them define the 'knuckle' epitope (Kirsch *et al*., 2000). All known ligand sequences contain seven invariant Cys residues, numbered as C2, C4 and C5-C9; many of them contain an extra pair of Cys residues numbered as C1 and C3. The structurally conserved

region of the fold is described as a '*cystine-knot*' since cysteines C4, C5, C8 and C9 participate in an eight-membered macrocycle wide enough for the last cystine bridge (formed by C3 and C7) to pass through. The cysteines that form N-terminal disulfide bridge in TGFβ are absent in other family members. Since the proteins of this family lack the hydrophobic core, the rigid cystine-knot scaffold is necessary for structural integrity. Further stabilization is achieved by dimerization that creates a hydrophobic core between the protomers. In most cases, such dimerization events are accompanied by the formation of a disulfide bridge connecting the two protomers at the C6 position.

5.2.4 The Structure of Receptors

The type1 and type2 receptors are glycoproteins of approximately 55kDa and 70kDa, respectively, with core polypeptides of 500 to 570 amino acids including the signal sequence. Each receptor contains an extracellular or ectodomain, a short membrane spanning helix and a cytoplasmic serine/threonine kinase domain (Mathews and Vale, 1991; ten Dijke *et al*., 1993; Lin *et al*., 1992; Attisano *et al*., 1992; Ebner *et al*., 1993). The type1 receptors have a higher level of sequence similarity than type2 receptor, particularly in the kinase domain (Massagué, 1998). Crystal structure of the extracellular domain of activin type2A receptor (AtR2-ECD) has been determined (Greenwald *et al*., 1999). The fold of AtR2-ECD comprises of three antiparallel sheets formed by seven β-strands (Figure 5.1b). The molecule has both concave and convex surfaces arising from a curvature in the first β-sheet (β1-β2). AtR2-ECD adopts a *three-finger toxin* fold*,* also observed in several toxins, which is characterized by a common pattern of eight cysteines, forming a conserved scaffold of four disulfide bridges. The three fingers refer to three pairs of strands (β1-β2, β3-β4, β5-β6) which all point roughly to the same direction (Figure 5.1b). AtR2 and cardiotoxin have the same disulfide pattern (C1-C3, C2-C4, C5-C8 and C9-C10), with the exception of an additional disulfide in AtR2 (C6-C7). Among the type2 receptors, there is some variability in the occurrence of the cysteines. The majority of the extra cysteines in other receptors are clustered in finger1, which constitutes the least conserved region in terms of both sequence and length. TβR2

has four additional cysteines in finger1, but lacks the two cysteines that constitute the C5-C8 disulfide bond in AtR2. Punt has two extra cysteines in finger1.

Crystal structure of human BMP2 ligand in complex with two high affinity receptor1A extracellular domains ($BR1A_{ec}$) has been reported recently which provide important information on TGF-receptor interactions at the molecular level (Kirsch *et al*., 2000). In this structure, two molecules of type1 receptor are bound to the ligand dimer at the 'wrist epitope' region of the ligand (further details please see Results and Discussion) by mainly hydrophobic surfaces of both the molecules. In addition, this report also confirms that both type1 and type2 receptor extracellular domains share the same fold, especially at the central β-sheet, despite poor sequence identity. Differences at loop regions and insertions of non-core secondary structures are evident; for example, a helix involved in primary interactions with the ligand in type1 receptor is absent in type2 receptors (Kirsch *et al*., 2000). Instead, an additional disulfide bridge, unique to type2 receptors, links the equivalent loop region at the convex surface to the central β-sheet. This suggests that the two types of receptors have different modes of binding to the ligand at the atomic level.

5.2.5 Previous Studies and Present Approach

Only a limited number of functionally important residues have been identified in TGFβ and related growth factors for binding to type2 receptor. The influence of segment deletions, residue replacements and isoform chimeras on the binding affinity of TGFβs of their type2 receptor (TβR2) were studied, highlighting the importance of C-terminal residues 83-112 of TGFβ1-3 (Qian *et al*., 1996). Structure-function analysis of activinβA molecule is reported and two amino acids involved in the binding of the activin molecule to its type2 receptor were identified as important for binding: Asp27 and Lys102, on the 'knuckle epitope' (Wuytens *et al*., 1999). Gray and coworkers have performed alanine scanning mutagenesis experiments on AtR2-ECD and identified a cluster of hydrophobic residues ('hydrophobic triad'), Phe42, Trp60 and Phe83, as critical for binding to activins/inhibins (Gray *et al*., 2000). It is known that type2 receptors form a heteromeric complex with the ligand, but exactly how many receptor molecules interact with the

ligand is not known (Massagué, 1998). It is apparent from Table 5.1 that TGFβ ligands can only bind to TβR1 and TβR2 but no such specificity is observed in the case of BMPs and activins. AtR1 binds to activins/inhibins, BMP7 and MIS/AMH; AtR2 binds to activins, BMP7 and GDF5 (Massagué, 1998). This report suggests that the determinants of ligand binding to receptors may be conserved within the TGFβ subfamily, the determinants of specificity are different between TGFβ and activin/BMP subfamilies, while activins and BMPs have similar residues that determine the specificity. Activin receptors bind to activins/inhibins, BMPs, MIS and GDF5. Despite a remarkable structural similarity, no such binding is observed for TGFβ ligands.

In order to determine the functionally important residues, we have compared the sequence distribution within the three fingers of the receptor, the nature of charge distribution of ligands and employ the Evolutionary Trace (ET), first applied by Litcharge *et al* on SH2 and SH3 domains method (Litcharge *et al*., 1996), to identify potential binding-site residues as targets for mutagenesis in TGFβ family of receptors. The five available structures of TGFβ ligands (3TGFβs, and 2BMPs) and three-dimensional models of *dpp* and inhibinβB derived by comparative modeling, have been analyzed for the differences in the distribution of polar and hydrophobic residues on the surface of the molecules, especially at the conserved residues (Innis et al., 2000) important for binding to type2 receptor. The extracellular domain of type2-receptor was docked to the ligand-dimer-type1 receptor complex. On the basis of previous mutagenesis studies and the results of our analysis, the 'knuckle' epitope is identified as a site of interaction with type2 receptor. Since the ligand molecules contain two symmetric knuckle epitopes, two receptors can bind to ligand dimer forming a tetrameric complex.

5.3 Materials and Methods:

5.3.1 Sequence alignment and clustering of receptor2 sequences

23 members of TGFβ receptor2 family were identified by PSIBLAST (Altschul *et al*., 1997) search using AtR2A ectodomain as query sequence against the Swissprot Databank (Bairoch and Apweiler, 1996) and used for evolutionary analysis. The ectodomains of the sequences were aligned using CLUSTALX (V 1.8; Thompson *et al*., 1997) and manually edited ensuring that gaps were not inserted into areas of known (or predicted) secondary structures. A PHYLIP (V 3.5) distance matrix based on sequence dissimilarity indices was generated and input into KITSCH clustering package to build a rooted phylogenetic tree (Felsenstein, 1985).

5.3.2 Evolutionary Trace analysis of receptor sequences

An evolutionary trace is generated by comparing consensus sequences for a group of proteins which originate from a common node in a phylogenetic tree and are characterized by a common Evolutionary Time Cut-off (ETC), and classifying each residue as one of the three types: absolutely conserved, class-specific and neutral. Here 'class-specific' denotes residues occupying a strictly conserved location in the sequence alignment, but differing in the nature of their conservation between various subgroups. When structural and functional residues of a protein family are not characterized, target residues can be chosen for mutagenesis. This can also be mapped on to known protein structures to identify clusters of important amino acids on the surface of the protein.

The ET analysis (Litcharge *et al*., 1996) was performed using TraceSuite (Innis et al., 2000). First, the phylogenetic tree was split along the evolutionary time into five evenly distributed partitions: P01 to P05 in order of increasing ETC. For each partition, a trace procedure was completed automatically in three steps: (1) Protein connected by a common node with evolutionary time greater than the given ETC were clustered together.

(2) A consensus sequence was generated for each group to distinguish between conserved and non-conserved positions. (3) A trace was generated by comparing the consensus sequences of receptors. Residues were classified into three types: absolutely conserved, class-specific and neutral. All the receptor sequences considered for the initial alignment were used for ET analysis. Punt sequence was not included since it is a lone element in the evolutionary tree and may bias the results.

5.3.3 Comparative modeling and visualization

Mature carboxy terminus peptides of *dpp* of *Drosophila melanogaster* and inhibinβB of *Homo sapiens* were taken from Swissprot databank (Bairoch and Apweiler, 1996). They were multiply aligned using CLUSTALX (V 1.8; Thompson et al., 1997) to other family members of TGFβ family. MODELLER (V 4.0; Sali and Blundell, 1993) was used to build three-dimensional models of both the proteins. BMP2 (PDB code 3bmp) was used as a template for modelling *dpp*; BMP2 (PDB code 3bmp) and TGFβ3 (PDB code 1tgj) were used as templates for modeling inhibinβB. MODELLER constructs a minimized 3D model(s) of a protein by the satisfaction of spatial restraints extracted from the template PDB (Bernstein *et al*., 1977) files. 20 models of the query sequence in each case were generated. The final models were chosen on the basis of lowest energy and least violation of structural restraints. The models with violated backbone CO and backbone NH restraints are not considered. Stereochemistry and geometry of the models were assessed using PROCHECK (V 3.4.4; Laskowaski *et al*., 1993) ensuring that the models have more than 85% residues in the core region of Ramachandran plot. The models were energy minimized using MAXIMIN2 option in SYBYL (Tripos Association, Inc., V6.5) using TRIPOS force field. For every run of energy minimization, 20 cycles of Simplex method and a further 50 cycles of Powell algorithm were employed. The resultant models have no short contacts or bad geometry. The dimer coordinates were generated using a superposition program called SUPER (Neela, B., personal communication). The punt (type2 receptor for *dpp* molecule) receptor ectodomain was also modeled following the same procedure with AtR2-ECD crystal structure (PDB code 1bte) as template. The resultant models and crystal structures were viewed by RASMOL (V 2.6b2; Sayle and

Milner-white, 1995) and solvent accessible surfaces and electrostatic potentials were calculated and displayed using GRASP (V 1.1; Nicholls *et al*., 1993). Structure-based sequence alignment of TGFβ ligands was compiled using the program COMPARER (V 2.0; Sali and Blundell, 1990) and structure-annotated using JOY (V 4.0; Overington *et al*., 1993; Mizuguchi *et al*., 1998).

5.3.4 Docking studies on ligand-receptor type2 receptor interactions

The Global Range Molecular Matching (GRAMM, V 1.03) methodology (Katchalski-Katzir *et al*., 1992; Vakser, 1995; Vakser, 1996) is an empirical approach to smoothing the intermolecular energy function by changing the range of the atom-atom potentials. The technique allows to locate the area of the global minimum of intermolecular energy for structures of different accuracy. The quality of the prediction depends on the accuracy of the structures. Thus, the docking of high-resolution structures with small conformational changes yields an accurate prediction, while the docking of ultra-low-resolution structures will give only the gross features of the complex. To predict the structure of a complex, it requires only the atomic coordinates of the two molecules (no information about the binding sites is needed). The program performs an exhaustive 6-dimensional search through the relative translations and rotations of the molecules.

The X-ray structures of activin type2 receptor (PDB code 1bte; solved at 1.5 Å resolution) and complex of BMP ligand dimer with its type1 receptors (PDB code 1es7; solved at 2.90 Å resolution) were docked using GRAMM program (Katchalski-Katzir *et al*., 1992; Vakser, 1995; Vakser, 1996) with a generic, hydrophobic mode and a grid step of 2.1Å. 1000 different models were generated to study every probable way of ligand-receptor interactions. The models were examined for maximal hydrophobic interactions and total interactions between the 1es7 and 1bte structures using the distance cut-off value derived from known cytokine-receptor crystal structures.

5.4 Results and Discussion

5.4.1 Analysis of receptor type2 sequences

Aligned non-redundant sequences of receptor type2, as shown in Figure 5.2, contain seven sequences of AtR2A, seven sequences of AtR2B, four BR2 sequences, four TβR2 sequences, and a punt receptor sequence from *Drosophila melanogaster*. Sequences of subfamilies show high conservation among themselves, but across the subfamily there is hardly any conservation apart from the cysteines. Phe42, Trp45, Gly58 and Asn92 are characteristic of a three-finger toxin fold and are largely conserved. However, Phe42 is substituted by Tyr in AtR2B, BR2 and punt but replaced by Val in TβR2; Gly58, which is conserved in AtR2B, BR2 and punt, is absent in TβR2. In general, the average sequence identity is around 25%. Punt receptor shares 28-30% identity with BR2 and AtR2B, 22-23% identity with AtR2A and ~15% identity with TβR2. Trp45 and Asn92 are absolutely conserved amongst all the type2 receptor subtypes considered. Evolutionary tree was generated using PHYLIP3.5 package (Figure 5.3; Felenstein, 1985). As expected, AtR and BR sequences are more similar and TβRs stand by their own as a separate cluster.

5.4.2 Analysis of residues in fingers

Finger1 contains loops of similar length that may be important for specificity in binding to the ligand since these loop regions display maximal sequence variation also confirmed by evolutionary trace method (discussed later). Two negatively charged residues at the tip of finger1, Glu19 and Asp21 (AtR2A numbering) are replaced by Asn and Leu in BR2 and Ser and Cys in TβR2. Finger1 of punt contains an extra disulfide bridge, while that of TβR2 contains two extra disulfide bridges (see Figure 5.1). This confirms previous modeling and scanning-deletion mutagenesis studies Guimond *et al*., 1999), which show that residues in finger1 (residues 58-60 and 63-65 of TβR2), facing the concave surface are important to bind TGF. Finger2 contains very few residues in each receptor sequence.

However, punt, TβR2 and BR2 receptors have relatively longer finger2 region: two-residue insertion in the case of TβR2 and punt and a one-residue insertion in BR2. Residues 74-79 of finger3 are exposed on the concave surface; AtR2A has two positive and two negative charges in this loop, while AtR2B is predominately negative. BR2 is polar and TβR2 is predominantly positive in this region, while punt contains one positive and one negatively charged residue.

It is reported that mutant receptors, containing deletions corresponding to loop regions of finger1, β2-β3 loop and finger2, do not bind the ligand (Guimond *et al*., 1999). However, mutant receptors containing deletion at finger3, loop region before β1, β4-β5 loop and after β7 do bind the ligand with similar affinities as the wild type receptors (Guimond *et al*., 1999). Deletion of the loop region corresponding to finger2, owing to the fact that finger2 is short, might cause structural changes to the receptor rendering inability to bind the ligand. Thus, finger2 may or may not be important for binding. The highly variable finger1 is not only a potential binding interface, but also the second most exposed, conserved hydrophobic surface (as observed in the crystal structure of AtR2), which is present at the convex side of the molecule. Finger1 is a good candidate to provide both hydrophobic docking surface and to act as primary determinants of interaction and binding specificity (Greenwald *et al*., 1999).

5.4.3 Evolutionary Trace of receptor2 sequences

The output of TRACESUITE program (by Innis *et al;* employing ET method; Litcharge *et al*., 1996) on the extracellular domain of TGF type2 receptors is shown in Figure 5.4. Analysis of the mapped traces for partitions P01 to P05 reveal clusters of potentially important residues on both concave and convex surfaces of the receptor structures. The residues defined by the 'hydrophobic triad' are located at the concave surface (Gray *et al*., 2000). In partition P01, apart from the structurally invariant cysteines, Trp45, Val55 and Asn92 are absolutely conserved among all receptor types considered in the ET analysis. The conserved Val55 lies on β4 and it is in the vicinity of finger2. Lys replaces Val55 in punt sequence. Other residues identified in partition P01 are Thr8, Glu10, Asn15, Glu19,

Glu29, Gly33, Ala43, Asn47, Asp62, Asp63, Val81, Glu93 and Phe95. Gly33 and Ala43 are buried in the core and may have a structural role and Glu19 is on the β1-β2 loop (finger1). Thr8, Glu10, Asn15, Asn47, Glu93 and Phe95 do not face the concave surface but are solvent accessible with no identified function. The trace residues facing the concave surface are Glu29 (on β2), Asp62 and Asp63 (on β4-β5 loop), Val81 (on β5-β6 loop; finger3) and Phe83 (on β6; finger3). No class-specific residues were identified at P02.

Phe83 is in the 'hydrophobic triad' identified by alanine scanning mutagenesis to be important for ligand binding (Gray *et al*., 2000). However, single mutations of Phe13, Phe14, Glu29 and Asp62 do not alter binding specificity for activins and inhibins (Gray *et al*., 2000). ET method does not identify Phe13 (exposed on concave side) and Phe14 (exposed in convex side) as trace residues, which implies that these residues are probably involved in non-specific binding. The method, however, identifies Glu29 (at the end of β1) and Asp62 (β4-β5 loop), which face away from the three fingers (Figure 5.1b). Glu29 is replaced by Ser in both TβR2 and BR2 and by Thr in punt; Asp62 is replaced by Gly in BR2 and by Tyr in TβR2 while the corresponding residue in punt is deleted. This suggests that Glu29 and Asp62 might be playing a functional role in other subfamilies not tested so far by mutagenesis experiments.

In the crystal structure of AtR2A (Greenwald *et al*., 1999), Thr44 (AtR2A numbering) identified as a conserved residue at partition P03 by ET analysis, is in the middle of a solvent-exposed hydrophobic surface, created by Ala16, Phe42, Val55, Trp60, Ile64, Val81 and Phe83 (to recall that three of these define the 'hydrophobic triad' important in ligand binding). Except Ala16 all others are 'trace' residues. Ala16 is considered in the analysis as it is solvent exposed hydrophobic residue and it lies in loop region of finger1. We refer to the 'hydrophobic triad' as the residues defining the 'principal' hydrophobic patch which can be further extended to include Ala16, Thr44, Val55, Leu61, Ile64 and Val81 termed as the 'surrounding' hydrophobic patch. In BR2 and punt, Thr44 is replaced by Leu in BR2 and a Val in TβR2. To note that position Lys56, spatially proximate to this extended hydrophobic patch and conserved in AtR, BR and punt when mutated to

Ala does not display drastic change in binding (Gray *et al*., 2000). Lys56 has not been identified as a trace residue by our present ET analysis.

5.4.4 Structure based analysis of TGFβ ligands and identification of determinants of binding and specificity

Large exposed hydrophobic patches on a protein surface often form part of a binding surface (Young *et al*., 1994). In the human growth hormone-receptor complex, a few hydrophobic residues at the interface contribute most to the free energy of interaction (Clackson and Wells, 1995). The recently solved crystal structure of the complex of BMP2-BR1A ectodomain (Kirsch *et al*., 2000), exemplifying TGF-TGF type1 receptor interactions, also demonstrates the same theme. Phe85 of $BR1A_{ec}$ helix α1 fits into a hydrophobic pocket of the ligand where it interacts with Trp28 and Trp31 of BMP2, among other residues. In the crystal structure of free BMP2, this pocket accommodates a 2-methylpentane-2, 4-diol molecule from the buffer solution, and a dioxane in the case of TGFβ3 (Mittal *et al*., 1996; Scheufler *et al*., 1999; Kirsch *et al*., 2000). Ile62, Val63, and Leu66 of BMP2 provide an almost exclusively hydrophobic surface, which together with Asn59, form the site of interaction with Phe85 of the receptor molecule (Kirsch *et al*., 2000). In addition, Phe60, Met78 and Ile99 of BR1A are central to the ligand-binding interface (Kirsch *et al*., 2000). The residues correspond to Asn59, Ile62, Val63, and Leu66 (BMP2) in case of TGFβ ligands (Innis *et al*., 2000) and the residues corresponds to Phe85, Phe60, Met78 and Ile99 (BR1A) in case of receptor1 sequences were identified as trace residues. In order to identify the determinants of binding and specificity for TGF-TGF type2 receptors, the following approaches were taken:

5.4.5 Structure of TGF growth factors and analysis of TGF-like sequences

The structures for TGFβ1-β3 (Daopin *et al*., 1992; Schlunegger *et al*., 1993; Mittal *et al*., 1996; Hinck *et al*., 1996), BMP7 (Griffith *et al*., 1996) and BMP2 (Scheufler *et al*., 1999) when superposed in the best fit, display an overall root mean square deviation of less than

1.1 A°. However, there are clear differences in some structural elements between TGFβs and BMPs; N-terminus is not visible in the crystal structure of BMP2 (Scheufler *et al*., 1999) and BMP7 (Griffith *et al*., 1996). In contrast, TGFβ1-3 exhibits a short N-terminal α-helix (α1), that is anchored to the protein core by an additional disulfide bridge (Daopin *et al*., 1992; Schlunegger *et al*., 1993; Mittal *et al*., 1996; Hinck *et al*., 1996). Moreover, BMP2 and BMP7 do not contain the short helix α2 observed after the second β-strand in TGFβs and is replaced by a tighter non-helical turn. This feature is conserved among known BMPs, GDFs, activins and other subfamilies. However, BMP2 and BMP7 structures show a unique conformation at the loop preceding α3: a longer loop with a three-residue insertion (a short β-strand in BMP2).

5.4.6 Analysis of surface residues of ligand molecules for difference in charge distribution

Figure 5.5 shows the GRASP surface representation (Nicholls *et al*., 1993) of structures of TGFβ1-3, BMP2, BMP7 and models of inhibinβB and *dpp*. Large hydrophobic areas are concentrated especially on the wrist and knuckle epitope regions of the ligand dimers. It is clear from the figure that the charge distribution is different between TGFβ isoforms and activin subfamily of proteins, especially in the knuckle epitope, in the loops of β2-β3 and β7-β8 strands. These regions contain high negative charge in case of BMPs, inhibins and *Dpp*, while they are positively charged in TGFβs (the conservation is confirmed using multiple sequence alignment of ligands). Unlike BMPs, *dpp* is in general polar at β7-β8 loop, at the knuckle epitope, where Asp93, Glu95 and Lys96 of BMP2 are replaced by Asn, Gln and Thr in *dpp*, respectively. This difference in charge distribution together with the structural differences discussed before can be instrumental in giving rise to specificity while binding to receptors. In addition, all the structures have positive charge at N-terminus (conserved positively charged residue, two residues after C2) which accounts for their heparin binding (Ruppert et al., 1996). However in BMPs, the N-terminus might fold back to shield this charge (as observed in case of inhibinβB model) and TGFβs are less positive than BMPs in this region.

5.4.7 Evolutionary Trace of Ligands and identification of residues implicated in binding and specificity:

ET method was applied to multiply aligned sequences of TGFβ superfamily of ligands and trace residues were identified by Innis and coworkers (Innis *et al*., 2000). Trp28 and Trp31 of BMP2, which have primary interactions with Phe85 of BR1Aec (please see above) are absolutely conserved in the TGFβ family alignment and are identified as 'trace' residues (Innis *et al*., 2000). Mutation of Trp31 to alanine significantly decreases the stability of the BMP2-BR1A$_{ec}$ complex (Kirsch *et al*., 2000). Interestingly, neither Trp28 nor Trp31 are conserved in the distant relatives like GDNF.

Here we will discuss those trace residues which occur on the knuckle epitope (important for receptor binding), are topologically equivalent (using COMPARER; Sali and Blundell, 1990) and display similar characteristics (identified by JOY; Overington et al., 1993; Mizuguchi et al., 1998) (Figure 5.6). Two interesting clusters of residues are identified: first at alignment positions 35, 36, 37, 92, 93, 94, 96, 98, 104, 105 and 106 (Figure 5.6; residues forming the 'knuckle epitope'). In case of BMP2 (PDB code 3bmp), these residues are Val33, Ala34, Pro35, Ala86, Ile87, Ser88, Leu90, Leu92, Val98, Val99 and Leu100. A second small cluster of residues at alignment positions 17, 18, 43 and 115 (Arg16, His17, Phe41 and Glu109 according to BMP2 numbering) is rather surprising. However, it should be noted that the N-terminus of the BMP family of ligands, that are not seen in the crystal structure could fold back in this region attributing a structural than functional role to these residues. The trace residues at the 'knuckle epitope' are divided into two classes (see Table 5.3 a;b): First, the residues, which are implicated for ligand binding and second, residues implicated for binding as well as specificity, most of which are subfamily specific and class-specific residues. The first cluster can be the preferred site of interaction with hydrophobic clusters on type2 receptor identified by mutagenesis studies and also in this study. Residue Pro35 is absolutely conserved in all the ligand sequences considered for our analysis but is absent in GDNF. Val33 and Ala34 are replaced by charged residues in TGFβ isoforms, while Leu100, is replaced by a charged residue in inhibins; other residues in cluster1 are conserved substitutions. In cluster2,

Arg16 is replaced by a hydrophobic residue in TGFβs. Alignment position 18 (Figure 5.6; His17 of BMP2) is occupied by a positively charged residue in all the known sequences. Phe41, adopting unusual φ/ψ angles (67°, 178°) in the Ramachandran plot (Ramachandran *et al*., 1963; Ramachandran and Sasiekharan, 1968), although not in the dimer interface, forms interchain contact at the backbone carbonyl with the neighboring subunit in bmp2 and its side chain is solvent exposed (Scheufler *et al*., 1999) on 'knuckle epitope'. This residue is present where the β-strands, β2, β5, β6 and β9 are arranged close enough to form a short segment of four-stranded antiparallel β-sheet as evident in the crystal structure (Scheufler *et al*., 1999). This arrangement is also observed in all known proteins in TGFβ superfamily. As shown in Table 5.3, Glu109 (BMP2 numbering) is specific for BMP2 and inhibin subfamilies, which is a positively charged residue (Arg/Lys) in TGFβs, BMP7 and a valine in *dpp*. Such differences at 'trace' residues point to class-specific electrostatic distribution and receptor specificity. Figure 5.7 shows the 'trace residues' mapped on the surface of ligand structure. The residues mapped are on 'knuckle epitope'.

In the light of the above results and the crystal structure of BMP2 with its type1 receptor, it is plausible to propose that finger1 (with loop region of β2-β3) of type2 receptor interacts with the 'knuckle epitope' of the ligand; if C-terminus of both type1 and type2 receptors need to point roughly in the same direction and the type2 receptor interacts with the ligand at its concave surface (Greenwald *et al*., 1999; Kirsch *et al*., 2000; Guimond *et al*., 1999) with the 'hydrophobic triad' residues. Finger3 loop region of the type2 receptor is not playing any important role in binding. Some of residues identified by ET method in 'surrounding patch' for receptor type2 are also reported in the deletion studies of core region (53-55, 83-85, 98-100 and 143-145; TβR2 numbering; See table 5.2) of TβR2 (Guimond *et al*., 1999). Thus ET method can be used both to rationalize the result of the mutagenesis studies and also to predict the targets for the mutagenesis.

5.4.8 Docking studies of type2-receptor to the ligand-dimer-type1 complex

The nature of interactions at the ligand-type1 receptor binding site were primarily hydrophobic (Kirsch *et al*., 2000). The above analyses on trace residues of solvent-exposed hydrophobics and the overall similarities between type1 and type2 receptors suggest similar hydrophobic interactions in TGF ligand-type2 receptor binding (Figure 5.7a,b). 1000 GRAMM (Katchalski-Katzir *et al*., 1992; Vakser, 1995; Vakser, 1996) models of type2 receptor interacting with the ligand dimer complexed with two type1 receptors were generated. The models were examined for maximal hydrophobic interactions, (defined as interaction between hydrophobic residues of both the structures) and total interactions (defined as interaction between all residues of both the structures) using the C_α distance cutoff of 12 Å for 'interacting' pairs. Figure 5.7c shows the distribution of the number of models with different number of hydrophobic interactions between ligand-type1 receptor complex and receptor type2. Models with appreciable number of hydrophobic interactions at the 'principal patch' were specifically examined after including the 'surrounding hydrophobic patch' residues (inset to Figure 5.7c). Interestingly, models with the highest number of hydrophobic interactions (Figure8c) and total interactions (as shown in Figure5.7d) with key residues of type2 receptor closely correspond to ET-results, suggesting a theme involving knuckle epitope at the ligand as the receptor-binding site.

5.5 Conclusions

In this chapter, the clusters of residues has been identified, which lie on the knuckle epitope of ligand molecules and the concave surface of type2 receptor molecules, which may play an important role in complex formation. These clusters are hydrophobic patches surrounded by charged residues on the surface of molecules. Finger1 and a part of finger2 of the type 2 receptor, with the central hydrophobic patch, interact with the 'knuckle epitope' of the ligand (mainly convex side on β2-β3, β7-β8 and the loop regions joining them) as it provides the large conserved hydrophobic surface for docking. The β2-β3

loop region may be interacting with the smaller cluster identified by ET bearing good agreement with GRAMM docking studies. While each type1 receptor interacts simultaneously with both the ligand protomers at the 'wrist' epitope (Kirsch *et al*., 2000), we predict that the type2 receptor interactions are with one protomer each at the 'knuckle' epitope. These predictions are supported by deletion studies on ligands (Hinck *et al*., 1996; Qian *et al*., 1996; Gray *et al*., 2000), deletion and mutagenesis studies on receptor type2 sequences (Gray *et al*., 2000; Guimond *et al*., 1999) and orientation of receptor type1 in the crystal structure in complex with BMP2 (Kirsch *et al*., 2000). The amino acids that emerge from the ET method as important for function can be targets for future mutagenesis studies. It will also be interesting to prepare TGF chimeras of loop region between β2-β3 loop and β7-β8 loop since difference in charge distribution in this region may contribute to specificity in identification of receptors. Various tools such as the study of evolutionary trees, conserved residues of the aligned sequences, spatial positions of interesting residues, charge distribution on their three-dimensional fold and docking studies have been employed to provide structural explanations for ligand-receptor specificity which have general value in the area of protein-protein interactions.

Table5.1 Abbreviations include: TGFβ (transforming growth factor β), BMP (bone morphogenic protein), Dpp (decapentaplegic), GDF (growth and differentiation factor), MIS/AMH (Mullerian inhibiting substance/anti mullerian hormone) TβR (transforming growth factor receptor) AtR (activin receptor) and BR (bone morphogenic protein receptor). This table was compiled according to Massague [1], with a few corrections. It should be noted that there is no species specificity observed in ligand-receptor interaction.

Sequential Binding		
Ligand	**Type 2 receptor**	**Type1 receptor**
TGFβ	TβR2	ALK1, ALK2?, TβR1, ALK7
Activins	AtR2, AtR2B	AtR1 , AtR1B
BMP7	AtR2, AtR2B	AtR1
GDF5	AtR2, AtR2B	AtR1 , AtR1B
MIS/AMH	AMHR	AtR1?
Co-operative Binding		
Ligand	**Type 2 receptor**	**Type1 receptor**
BMPs	BR2	BR1A ,BR1B
Dpp	Punt	Thick veins (tkv) Saxophone (sax)
GDF5		AtR1, BR1B

Table 5.2: Residues of receptor type2 important for interaction with ligands identified using ET method.

	Prin. Patch			Surrounding patch					
Position	91	111	146	58	93	106	116	119	139
AtR2A	F42	W60	F83	A16	T44	V55	L61	I64	V81
AtR2B	Y42	W60	F84	A16	S44	V55	L61	F64	V81
BR2	Y41	W59	F89	P13	L43	V54	I62	P65	I82
TβR2	V85	H102	M135	V56	V87	V100	T108	G110	G130
Punt	Y37	F57	F81	E10	L39	K52	T58	M60	G77

Table 5.3:

a) Specific residues implicated for high affinity binding

Position	21	23	25	37	43	94	96	98
Tgfβ3	Y21	D23	R25	P36	N42	T87	L89	Y91
Tgfβ2	Y21	D23	K25	P36	N42	T87	L89	Y91
Tgfβ1	Y21	D23	R25	P36	N37	P87	V89	Y91
BMP2	Y20	D22	S24	P35	F41	S88	L90	L92
BMP7	Y44	S46	R48	P59	Y65	S113	L115	F117
Dpp	Y9	D11	S13	P24	Y30	A78	L80	L82
IHB β	F17	D19	R21	P32	N39	S89	L91	F93

b) Subfamily specific residues implicated for receptor specificity and binding

Position	17	18	32	35	36	92	93	104
Tgfβ3	V17	R18	K31	H34	E35	P85	L86	P96
Tgfβ2	L17	R18	K31	H34	E35	P85	L86	P96
Tgfβ1	V17	R18	K31	H34	E35	P85	L86	P96
BMP2	R16	H17	D30	V33	A34	A86	I87	V98
BMP7	K40	H41	D54	I57	A58	A111	I112	V123
Dpp	R5	H6	D19	V22	A23	S76	V77	V88
IHB β	R13	Q14	D27	I30	A31	T87	M88	I99

Position	105	106	107	108	115
Tgfβ3	K97	V98	E99	Q100	K107
Tgfβ2	K97	I98	E99	Q100	K107
Tgfβ1	K97	V98	E99	Q100	R107
BMP2	V99	L100	K101	N102	E109
BMP7	I124	L125	K126	K127	R134
Dpp	V89	L90	K91	N92	V99
IHB β	V100	K101	R102	D103	E110

Figure Legends

Figure 5.1: Ribbon representation of (a) BMP2 ligand (3bmp.pdb) and (b) AtR2-ECD (PDB code: 1bte). Secondary structures are labeled. The knuckle and wrist epitopes are marked on TGF; fingers are marked on the receptor structure. Figure is prepared using MOLSCRIPT (Kraulis, 1991).

Figure 5.2: Multiple alignment of 23 sequences of various type2 receptor ecto-domain with secondary structure and fingers marked.

ActR2A_MOUSE*, ActR2A_RAT, ActR2A_HUMAN, ActR2A_BOVIN, ActR2A_SHEEP, ActR2A_GALLUS, ActR2A_XENLA: activin receptor type2A from mouse, rat, human, bovin, sheep, chicken and xenopus.

ActR2B_MOUSE, ActR2B_RAT, ActR2B_HUMAN, ActR2B_BOVIN, ActR2B_GALLUS, ActR2B_ZEBRAFISH, ActR2B_GOLDFISH: activin receptor type2B from mouse, rat, human, bovin, chicken, zebrafish and goldfish.

BMPR2_HUMAN, BMPR2_MOUSE, BMPR2_GALLUS, BMPR2_XENLA: bone morphogenic protein receptor type2 from human, mouse, chicken and xenopus.

TGR2_HUMAN, TGR2_PIG, TGR2_MOUSE, TGR2_RAT: transforming growth factor receptor type2 from human, pig, mouse, rat and fruitfly.

PUNT_DROSO: homologue of activin type2 receptor in fruit fly.

* ActR2A_MOUSE (PDB code 1bte) is activin type2 receptor ecto-domain sequence from mouse, with known crystal structure (Greenwald *et al*., 1999),. The sequence is showed in structure based annotation using JOY (Overington *et al*., 1993; Mizuguchi *et al*., 1998). Please refer to legend of Figure6 for the JOY key.

Figure 5.3: Dendrogram containing 23 TGFβ family of receptor type2 ecto-domain on the basis of their sequence dissimilarity using PHYLIP3.5 .(Felsenstein, J, 1985) The sequences are as described in legend of Figure2.

Figure 5.4: Evolutionary Trace of type2 receptor sequences (excluding punt) for partitions P01 to P05, aligned with the amino acid sequences of AtR2A_mouse (activin receptor2A from mouse), AtR2B_human (activin receptor2B from human), BR2_human (BMP receptor2 from human) and TβR2 (TGFβ type2 receptor from mouse). * indicates

the residues important for mutagenesis. n indicates solvent buried residues as shown in the crystal structure of AtR2-ECD.(Greenwald *et al*., 1999).

Figure 5.5: Electrostatic potential representation of the known and modeled structures of TGFβ family of ligands using GRASP (Nicholls, 1993). Acidic residues are indicated by red surface patches and blue patches indicate basic residues. The structures are (a) bone morphogenic protein 2, (b) inhibinβB, (c) bone morphogenic protein 7, (d) decapentaplegic protein (e) transforming growth factor β1, (f) transforming growth factor β2 and (g) transforming growth factor β3. The charged residues are marked and numberings are according to their structural positions given in PDB files. In case of *dpp* the hydrophobic residues are marked.

Figure 5.6: Structure based sequence alignment of the TGFβ family using COMPARER (Sali and Blundell, 1990) and compiled using JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998). Solvent-accessible and solvent-inaccessible residues are shown in upper case and lower case, respectively. Residues in positive phi are indicated in italics; residues with *cis* peptide in the backbone or disulfide bonds are indicated by the presence of breve (e.g. š) or cedilla (e.g. ç), respectively. Hydrogen bonds formed to the side chains, main chain amides and main chain carbonyls of the other residues are indicated by the presence of tilde on top, boldface or underline respectively. The secondary structures are marked and numbered.

Figure 5.7: Predicted mode of interaction between tranforming growth factor and type2 receptor.

a) Structure of bmp dimer in GRASP surface representation (Nicholls, 1993). Key residues identified by evolutionary trace (ET) method (Litcharge *et al*., 1996), probably important for binding (prinicipal patch) are denoted in cyan. Additional residues (surrounding patch), also identified by ET method are shown in magenta.

b) Same as (a) but for the extracellular domain of activin type2 receptor. Regions corresponding to finger1, finger2 and finger3 (also see Figure 1) are shown in yellow, green and violet arrows. Proposed complimentary areas of interaction in the ligand dimer

are marked in similar colors in (a).

c) Distribution of the total number of hydrophobic contacts for 1000 models of the interaction between type2 receptor (PDB code 1bte) and ligand-dimer complexed with two type1 receptor molecules (PDB code 1es7). Hydrophobic contacts are measured between the two molecules as the number of C^{α}-C^{α} distances of hydrophobic residues of 1es7 within 12Å from key residues on the type2 receptor. Key residues at the type2 receptor have been identified by ET method (Litcharge *et al*., 1996; also listed in Table 5.2) and also by alanine-scanning mutagenesis (Gray *et al*., 2000) to be important for binding. Inset to Figure 7c) Models with high hydrophobic contacts (7 or more) at the 'principal patch' of the type2 receptor are examined for additional hydrophobic interactions including the 'surrounding hydrophobic patch'.

d) Ribbon representation of one of the models of the interaction between type2-receptor and type1-receptor-bound ligand dimer. This model, suggested by GRAMM, has high number of hydrophobic residues at the predicted binding site (Figure 7c). Activin type2 receptors are shown in cyan, bmp dimer in magenta and the two type1 receptors are shown in grey. The knuckle epitope of the ligand dimer and finger1 and finger3 of the type2-receptor (shown in yellow and violet arrows) are the key regions predicted to form the binding interface. This picture has been prepared using SETOR (Evans, 1993).

5.6 References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*. **25**: 3389-3402.

Attisano, L., Wrana, J.L., Cheifetz, S., and Massague, J. (1992). Novel activin receptors: distinct genes and alternative mRNA splicing generate a repertoire of serine/threonine kinase receptors. *Cell*. **10** 68:97-108.

Bairoch, A. and Apwiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucl. Acids Res*. **24**: 21-25.

Barr, P.J. (1991). Mammalian subtilisins: the long-sought dibasic processing endoproteases. *Cell*. **66**:1-3.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*. **112**: 535-542.

Clackson, T. and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* **267**: 383-386.

Daopin, S., Piez, K.A., Ogawa, Y. and Daevis, D.R. (1992). Crystal structure of transforming growth factor-β2: an unusual fold for the superfamily. *Science* **257**: 369-374.

Ducy, P., and Karsenty, G. (2000). The family of bone morphogenetic proteins. *Kidney Int* **57**: 2207-14.

Ebner, R., Chen, R.H., Lawler, S., Zioncheck, T., and Derynck, R. (1993). Determination of type I receptor specificity by the type II receptors for TGF-β or activin. *Science*. **262**: 900-2.

Eigenbrot, C., and Gerber, N. (1997). X-ray structure of glial cell-derived neurotrophic factor at 1.9 A resolution and implications for receptor binding. *Nat. Struct. Biol.* **4**: 435-8.

Evans, S. V. (1993). SETOR--hardware-lighted 3-dimensional solid model representation of macromolecules. *J. Mol. Graph*. **11**: 134-138.

Felsenstein, J. (1985). Confidence-limits on phylogenies - An approach using the bootstrap. *Evolution* **39**: 783-791.

Gaddy-Kurten, D., Tsuchida, K, and Vale W. (1995). Activins and the receptor serine kinase superfamily. *Recent Prog Horm Res* **50**:109-29.

Gray, P.C., Greenwald, J., Blount, A.L., Kunitake, K.S., Donaldson, C.J., Choe, S. and Vale, W. (2000). Identification of a binding site on the type II activin receptor for activin and inhibin. *J.Biol.Chem*. **275**: 3206-3212.

Greenwald, J., Fischer, W.H., Vale, W.W. and Choe S. (1999). Three-finger toxin fold for the extracellular ligand-binding domain of the type II activin receptor serine kinase. *Nat*. *Struct. Biol*. **6**: 18-22.

Griffith, D.L., Keck, P.C., Sampath, T.K., Rueger, D.C. and Carlson, W.D. (1996). Three-dimensional structure of recombinant human osteogenic protein 1: structural paradigm for the transforming growth factor β superfamily. *Proc. Natl. Acad. Sci*. USA **93**: 878-883.

Guimond, A., Sulea, T., Pepin, M.C.O. and Connor-McCourt, M.D. (1999). Mapping of putative binding sites on the ectodomain of the type II TGF-β receptor by scanning-deletion mutagenesis and knowledge-based modeling. *FEBS Letters* **456**:79-84.

Heldin, C.H., Miyazono, K., and ten Dijke, P. (1997). TGF-β signalling from cell membrane to nucleus through SMAD proteins. *Nature* **390**:465-71.

Hinck, A.P., Archer, S.J., Qian, S.W., Roberts, A.B., Sporn, M.B., Weatherbee, J.A., Tsang, M.L.S., Lucas, R., Zhang, B.L., Wenker, J. and Torchia, D.A. (1996). Transforming growth factorβ1: three-dimensional structure in solution and comparison with the X-ray structure of transforming growth factor β2. *Biochemistry* **35**: 8517-8534.

Hogan, B.L.M. (1996). Bone morphogenetic proteins: multifunctional regulators of vertebrate development. *Genes Dev*. **10**:1580-94.

Innis, C.A., Shi, J., and Blundell, T.L. (2000). Evolutionary trace analysis of TGF-β and related growth factors: implications for site-directed mutagenesis. *Protein Eng* **13**:839-47.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* **89**:2195-2199.

Kingsley, D.M. (1994). The TGF-β superfamily: new members, new receptors, and new genetic tests of function in different organisms. *Genes Dev* **8**:133-46.

Kirsch, T., Sebald, W. and Dreyer, M.K. (2000). Crystal structure of the BMP-2-BRIA ectodomain complex. *Nat. Struct. Biol*. **7**:492-496.

Kraulis, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24: 946-950.

Laskowaski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst*. **26**: 283-291.

Lin, H.Y., Wang, X.F., Ng-Eaton, E., Weinberg, R.A., and Lodish, H.F. (1992). Expression cloning of the TGF-β type II receptor, a functional transmembrane serine/threonine kinase. *Cell*. **68**:775-85.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. An evolutionary trace method defines binding surfaces common to protein families. J.Mol.Biol. 1996; 257: 342-58.

Massagué, J. (1990). The transforming growth factor-β family. *Annu. Rev. Cell Biol*. **6**:597-641.

Massagué, J. (1998). TGF-β signal transduction. *Annu. Rev. Biochem*. **67**: 753-91.

Massagué, J., Blain, S.W., and Lo, R.S. (2000). TGFβ signaling in growth control, cancer, and heritable disorders. *Cell* **103**:295-309.

Massagué, J. (2000) How cells read TGF-β signals. *Nat Rev Mol Cell Biol* **1**:169-78.

Mathews, L.S., and Vale, W.W. (1991). Expression cloning of an activin receptor, a predicted transmembrane serine kinase. *Cell*. **65**:973-82.

Mehler, M.F., Mabie, P.C., Zhang, D., and Kessler, J.A. (1997). Bone morphogenetic proteins in the nervous system. *Trends Neurosci*., **20**:309-17.

Mittal, P.R.E., Pristle, J.P., Cox, D.A., McMaster, G., Cerletti, N. and Grutter, M.G. (1996). The crystal structure of TGF-β3 and comparison to TGF-β2: implications for receptor binding. *Prot. Sci*. **5**: 1261-71.

Miyazono, K. (1993). Activin receptor-like kinases: a novel subclass of cell-surface receptors with predicted serine/threonine kinase activity. *Oncogene*. **8**:2879-87.

Miyazono, K., ten Dijke, P., and Heldin, C.H. (2000). TGF-β signaling by Smad proteins. *Adv Immunol* **75**: 115-57.

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. and Overington, J.P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**: 617-623.

Murray-Rust, J., McDonald, N.Q., Blundell, T.L., Hosang, M., Oefner, C., Winkler, F., and Bradshaw, R.A. (1993). Topological similarities in TGF-β2, PDGF-BB and NGF define a superfamily of polypeptide growth factors. *Structure* **1**:153-9.

Nicholls, A., Sharp, K.A. and Honig, B. (1993). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Biophys. J*. **64**: 166-170.

Overington, J.P., Zhu, Z.Y., Sali, A., Johnson, M.S., Sowdhamini, R., Louie, G.V. and Blundell, T.L. (1993). Molecular recognition in protein families: A database of aligned three-dimensional structures of related proteins. *Biochem. Soc. Trans*. 21: 597-604.

Padgett, R.W., Das, P., and Krishna, S. (1998). TGF-β signaling, Smads, and tumor suppressors. *Bioessays*, **20**:382-90.

Qian, S.W., Burmester, J.K., Tsang, M.L.S., Weatherbee, J.A., Hinck, A.P., Ohlsen, D.J., Sporn, M.B., and Roberts, A.B. (1996). Binding affinity of transforming growth factor-β for its type II receptor is determined by the C-terminal region of the molecule.

J.Biol.Chem. **271**: 30656-662.

Raftery, L.A., and Sutherland, D.J. (1999). TGF-β family signal transduction in Drosophila development: from Mad to Smads. *Dev Biol*, **210**:251-68.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J.Mol.Biol*., **7**: 95-99

Ramachandran, G.N. and Sasisekharan,V. (1968). Conformation of polypeptides and proteins. *Adv. Protein. Chem*. **23**: 283-438.

Roberts, A.B., and Sporn, M.B.(1993). Physiological actions and clinical applications of transforming growth factor-β (TGF-β). *Growth Factors* **8**: 1-9.

Ruppert, R., Hoffmann, E. and Sebald,W. (1996). Human bone morphogenetic protein 2 contains a heparin-binding site which modifies its biological activity. *Eur. J. Biochem*. **237**: 295-302.

Saarma, M. (2000). GDNF - a stranger in the TGF-β superfamily? *Eur J Biochem* **267**: 6968-71.

Sali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J.Mol.Biol*. **234**: 799-815.

Sali, A. and Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol*. **212**: 403-428.

Sayle, A.R. and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. *T.I.B.S*. **20**: 374-376.

Scheufler, C., Sebald, W. and Hulsmeyer, M. (1999). Crystal structure of human bone morphogenetic protein-2 at 2.7 A resolution. *J. Mol. Biol*. **287**: 103-115.

Sowdhamini, R., Burke, D.F., Huang, J.F., Mizuguchi, K., Nagarajaram, H.A., Srinivasan, N., Steward, R.E., and Blundell, T.L. (1998). CAMPASS: a database of structurally aligned protein superfamilies. *Structure*. **6** :1087-94.

ten Dijke, P., Ichijo, H., Franzen, P., Schulz, P., Saras, J., Toyoshima, H., Heldin, C.H., Schlunegger, M. P. and Grutter, M.G. J. (1993). Refined crystal structure of human transforming growth factor β2 at 1.95A resolution. *J Mol. Biol*. **231**: 445-458.

Thomson, J.D., Gibson, T.J., Planwniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **25**: 4876-82.

Vakser, I.A. (1995). Protein docking for low-resolution structures. *Protein Eng*. **8**:371-377.

Vakser, I.A. (1996). Long-distance potentials: An approach to the multiple-minima problem in ligand-receptor interaction. *Protein Eng*. **9**:37-41.

Wuytens, G., Verschueren, K., de Winter, J.P., Gajendren, N., Beek, L., Devos, K., Bosman, F., de Waele, P., Andries, M., van den Eijnden-van Raaij, a.J.M., Smith, J.C., and Huylebroeck, D. (1999). Identification of two amino acids in activin A that are important for biological activity and binding to the activin type II receptors. *J.Biol.Chem.* **274**: 9821-9827.

Young, L., Jernigan, R.L. and Covell, D.G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Prot. Sci*. **3**:717-729.

Zimmerman, C.M., and Padgett, R.W. (2000). Transforming growth factorβ signaling

mediators and modulators. *Gene* **249**:17-30.

Chapter 6

Analysis of Masquerade: Case Study for Finding Function of Serine Protease Domains in Modular Proteins Involved in Patterning and Immune Response

6.1 Abstract

We describe here the structure-function relationship of the eukaryotic serine proteases, the most comprehensively studied enzyme family. A method to identify functional information from structures alone is described by analysis of structural properties of redundant data set of five sub families and a non-redundant data set of 43 protease structures derived from PDB using position specific information. Sequence information from SwissProt sequence database is used subsequently to derive 'consensus' functional residues. The analysis is then used to find the functional role of modular proteins containing functional and non-functional serine protease domains involved in early development, patterning and immune response with emphasis of masquerade sequence and its reported homologues. Masquerade is a protein reported for cell adhesion and contains C-terminal non-functional serine protease like domain. We have identified five chitin-binding motifs in N-terminal cysteine-knot domain in masquerade and related proteins. We propose the mechanism of binding and subsequent cleavage for the proteins having dual role in patterning and immune responses. Role of masquerade serine protease domain in early immune response is proposed. We also report two chitin-binding motifs for Drosophila *GRAAL* gene product identified from analysis of Drosophila genome and propose its role in patterning and immune response.

6.2 Introduction

6.2.1 The system

This chapter describes structure-function relationships in one of the largest and most comprehensively studied of all enzyme families, the serine proteases with trypsin fold and its application to other modular proteins containing this domain. It is also called the chymotrypsin fold, as it was the first member of the family to be crystallised. These enzymes function using catalytic triad made of His57, Asp102 and Ser195 (Lesk, 1981; Fersht, 1984). A similar catalytic triad has been observed in other proteases such as subtilisin (Kraut, 1971) and carboxypeptidases (Liao *et al*., 1992), as well as lipases (Dodson *et al*., 1992), but theses molecules have different structures. The crystal structure of α-chymotrypsin was reported in early days of crystallography in 1967 (Mathews *et al*., 1967). The 2Å-resolution structure followed soon (Birktoft and Blow, 1972). The structural information has grown enormously since then. Extensive investigations have elucidated the mechanism of enzymatic catalysis, folding pattern, inhibition, activation, and substrate specificity and have described the evolutionary variation in the family (Stroud, 1974; Neurath, 1975; Lesk and Fordham 1996).

Mammalian serine proteases (also called proteinases) participate in numerous physiological processes (Barret, 1977,1994; Horl, 1989; Bond, 1991; Twining *et al*., 1994); the best known are digestion, blood clotting (Davie *et al*., 1991), fertilisation (Baba *et al*., 1989), development (Gurwitz and Cunningham 1998), complement activation in the immune response in vertebrates (Reid *et al*., 1986; Goldberger *et al*., 1987) and insects (Kwon *et al*., 2000; Paskewitz *et al*., 1999). Their roles have also been suggested in signal transduction systems (Smirnova *et al*., 2001; Pendurthi *et al*., 2000). In several disease states, including emphysema (Watorek *et al*., 1988), tumor metastasis (Henderson *et al*., 1992), and arthritis (Froelich *et al*., 1993), the levels of proteases or inhibitors are elevated or out of balance. Proteins containing serine proteases domain are often modular (or mosaic) in nature. They sometimes contain multiple copies of different domains. Various proteins having protease domain are also known to contain Kringle (e.g

plasminogen and apolipoprotein); Sushi (e.g. complement factor B and Limulus clotting factor); Apple (plasma kallikrein and coagulation factor XI); Growth factor and Ca^{2+} binding (coagulation factor VII and coagulation factor IX); and Finger (coagulation factor XII and t-plasminogen activator) domains (Barrett, 1994). Proteins like Sb-Sbd (Appel *et al*., 1992), snake (Delotto and Spierer, 1986) easter (Chasan and Anderson, 1989) and Limulus contain five repeats of disulfide motifs in N-terminus and serine proteases in C-terminus. These proteins are reported in taking part in patterning during development (Murugasu-Oei *et al*., 1995). There are reports of mosaic proteins having domains homologues to serine proteases that have mutations in their catalytic triad (Donate *et al*., 1994; Murugasu-Oei *et al*., 1995). Masquerade is a member of this category.

6.2.2 Description of Structure

The 1.55 release of **S**tructural **C**lassification **O**f **P**roteins (SCOP) database (Murzin *et al*., 1995) describes proteins having trypsin fold under all β class. It describes the fold as containing internal gene duplication with two closed barrels having 6 strands in each barrel and shear number of 8 as shown in Figure1. There are 3 invariant disulfide bridges (C42-C58, C168-C182 and C191-C220 for all family members (chymotrypsin numberings are used throughout in text). The other disulfide bridges differ. For example, trypsin of higher organisms has disulfide bridge formed by C128 and C132, which is, absent in thrombin. On the other hand thrombin has a disulfide bridge between C1 and C122, absent in trypsin. The superfamily of trypsin fold serine protease is divided in to four broad families. They are eukaryotic protease, prokaryotic protease, viral protease and viral cysteine protease of trypsin fold with respectively forty-eight, nine, four and three family members with different functional specificity. For example trypsin, chymotrypsin, thrombin, elastase, collegenase, coagulation factors etc. are classified as eukaryotic serine proteases family members. It is evident that a wealth of structural information is available for this particular domain. More than one structure is available for certain subfamilies as shown by SCOP records. For example crystal structures of trypsin (ogen) from various species like cow, pig, rat, human, atlantic salmon and mold

are available. Further more there are 125 structures deposited alone for bovine (cow) trypsin, 21 structures from rat and 13 structures from pig. The reason being that each structure talks about different mutants or protease complexes with different chemical or protein inhibitors and how that might give important insights about the catalytic capacity or specificity of the proteases.

There has been a previous report of analysis of serine proteases of chymotrypsin family by Lesk and Fordham (Lesk and Fordham, 1996) describing in detail about basic structure, spatial relationship between the domains, mechanism of catalysis, packing of residues in individual domains, domain-domain interface, specificity pocket and similarity and divergence of proteases. Structural basis of substrate specificity (in terms of residues in S1-S4 positions and role of surface loop) and divergence has also been reviewed (Perona and Craik, 1995; 1997). There are no serious attempts to review the knowledge about the entire family afterwards best to our knowledge. Also, the data set used by Lesk and Fordham contained 13 structures only while the non-redundant data set used here is more than 3 times (43 structures).

This work has specifically focused on residues that renders specificity to the proteases or involved in structural changes that might be involved in inhibition or activation of proteases, giving a broader definition to functional residues. We will also discuss about the geometry of catalytic pocket and structural factors that renders them. It also demonstrates how biologically important information can be derived from data sets of 'redundant' (Please see Materials and Methods), and 'non-redundant' structures of eukaryotic trypsin fold serine proteases deposited in PDB (Berman *et al*., 2000), and sequences of Swissprot database (Bairoch and Apweiler, 2000) with available tools for structure and sequence analysis. Position specific properties of structures are used previously for detecting overall structure similarities and attempting fold prediction (Jones *et al*., 1992; Jones, 1999; Kelley *et al*., 2000) or homology detection (Gribskov *et al*., 1987; Karplus *et al*., 1998; Shi *et al*., 2001) or verification (Luthy R *et al*., 1992) but not function directly. This work reports position specific properties of loop regions spatially proximate to catalytic triad of serine proteases and adjoining secondary

structures for predicting the functional residues. It also demonstrates increasing importance of well-curated and publicly available databases for biologists. We have also discussed some interesting evolutionary relationships with respect to SCOP classification at sub-family level.

The analysis is then applied to find out functional role of 'masquerade' and related proteins. Masquerade is a non-functional serine protease from *Drosophila melanogaster* reported to be functioning in the process of somatic muscle attachment during larval stages. Its homologues from other species like crayfish *Pacifastacus leniusculus* (Huang *et al*., 2000) and insect *Holotrichia diomphalia* larvae (Kwon *et al*., 2000) are also reported recently. Masquerade sequence is used to fish its homologues in Drosophila genome database. Chitin binding motifs has been identified over primary sequence of masquerade and other proteins. We suggest role of masquerade in both development and immune response in Drosophila. The putative function for GRAAL gene product and its mosquito homologue Sp22D (Danielli *et al*., 2000) were also examined. We have extensively worked with eukaryotic tryspsin fold serine proteases (with one exception of *Streptomyces griseus* trypsin) for reasons of ample availability of structures and sequences and also its immediate functional importance to disease biology. Hence forth, they will be referred to as 'serine proteases' only.

6.3 Materials and Methods

6.3.1 The Redundant Data Set

Using masquerade sequence (Swissprot accession number Q24019) a PSI-BLAST (Altschual *et al*., 1997) search ('NCBI gi' option ON) is made on Protein Databank (PDB) database (Berman *et al*., 2000) at RCSB web site. The resultant output essentially listed all the eukaryotic trypsin fold serine proteases in the database. All serine protease entries have been downloaded from structure explorer pages of PDB (http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=query id). Here query id is four-letter unique pdb code assigned to each unique structure. The key words- Title, Compound,

Source, Primary Citation, Resolution, R-Value, Polymer chains and HET Groups were extracted from the pages. For all the entries keywords Compound and HET groups were analysed to seek for 'redundant structures' having different protein inhibitors and/or chemical or artificial inhibitors bound to them as described using Compound or HET Groups key words. The structures might also contain important mutations and can be from different species. The keywords Resolution and R-Value were used to choose high-resolution structure and also to prevent cases of 'tie'. Only high-resolution structures are used for the analysis. The resultant data set contained 23 structures of trypsin, 13 structures of thrombin, 7 structures of plasminogen activators and 6 structures of elastase and coagulation factors each. These families are selected also since there is plentiful information about structure-function relationships of these sub-families available in literature. The PDB files of resultant data set of structures were processed to remove all other protein chains, HET Group and water entries.

6.3.2 The Non-redundant Data Set

The PSI-BLAST (Altschul *et al*., 1997) output was also used for generating a non-redundant data set of structures with no two members sharing more than 95% identity as follows. The sequences reported in PSI-BLAST were extracted using the 'gi' identifiers from ENTREZ protein database using batch download option. An in-house program is written that discards one of the sequence (structure) of the pair with low quality (using keywords Resolution and R-Value) iteratively, using the pairwise identity from CLUSTALX output file (version 1.8; Jeanmougin *et al*., 1998). The resultant dataset has 43 structures. The data set includes the *Streptomyces griseus* trypsin as a protein under investigation, which is a prokaryotic serine protease. We will discuss the reasons for it in the discussion part.

6.3.3 Visulization and Analysis

The serine protease structures were viewed using RASMOL (Bernstein, 2000; Sayle and Milner-White, 1995). The loop regions spatially proximate to catalytic triad (please see

introduction) were chosen and marked (Figure1). Loop nomenclature is adopted as described by Peisach and co-workers (Peisach *et al*., 1999). Loops are named from L1 to L15 (Figure 2) and loops L3, L5, L7, L9, L11, L12 and L14 with adjoining secondary structures are chosen for further examination.

6.3.3.1 Analysis Using Redundant Data Set

As mentioned before this set of proteins contains the proteins/chemical inhibitors or HET groups bound to proteases. A program is written to find out the interacting residues of proteins with its inhibitors using distance criterion of 4Å. The results were tabulated and plotted as shown in Figure 2.

6.3.3.2 Analysis Using Non-Redundant Data Set

The program H-BOND is used to calculate hydrogen bonding (Overington J., unpublished). H-BOND calculates hydrogen bonding of main chain to main chain (i.e. those responsible for secondary structure), side chain to main chain carbonyl, side chain to main chain amide and side chain to side chain (hetero-atoms). The unprocessed PDB files were used while running H-BOND. For solvent accessibility calculations the PDB files were processed and all non-protease chains and HETATM entries were removed. The calculation was done using PSA that implements algorithm of Lee and Richards (Lee and Richards, 1971). The 7% relative cut-off is applied (Hubbard and Blundell, 1987). Secondary structure and main chain conformation were calculated using SSTRUC (Smith D, Unpublished) that uses Kasbach and Sander definitions; Kasbach and Sander, 1983) It is also a part of the PROCHECK (Laskowski *et al*., 1993) suit of programs. The programs HBOND, PSA and SSTRUC are part of sequence-structure representation and analysis program JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998). The structure-based alignment of all 43 structures as shown in Figure 6.3 is prepared using STAMP (Russell and Barton, 1992) and manually edited to remove local misalignments. The alignment is annotated using JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998) to compare the structural properties. Perl scripts are written to process the JOY ‘tem’ output

file to extract information about accessibility, hydrogen bonding, secondary structure, Ooi number for the segments under examination (loop region under investigation and adjoining secondary structures for each protein). The results for each property are plotted for loop regions under examination. The solvent accessibility results are the most interesting ones and reported here in Figure 6.4. The alignment was used to generate evolutionary tree using Neighbour-Joining method (Saitou and Nei, 1987; Figure 6.5) provided by CLUSTALX8.1 (Jeanmougin *et al*., 1998) and also using PHYLIP3.5 package (Felsenstein, 1985) that uses KITSCH algorithm. The protease domain of sequences of all five sub families were extracted from Swissprot database (Bairoch and Apweiler, 2000) using their function as keyword. Structure based alignment is used to guide the profile alignment of sequences using CLUSTALX8.1 profile alignment mode where the sequences are added to structural alignments. The resultant alignment was manually edited (Figure 6.6) and examined for conservation of functionally important residues among sub-families. Sequences of masquerade like proteins from different species were obtained from NCBI web site using Entrez search and retrieval system (http://www.ncbi.nlm.nih.gov:80/Entrez/) and aligned as mentioned above (with other proteins involved in patterning).

A BLASTP (Altschul *et al*., 1997) search is made using masquerade as query sequence on Drosophila genome and hits were examined using disulfide bridge conservation as a criterion (see results) to identify its homologues.

6.4 Results

6.4.1 Finding the Functional Residues

The interacting residues of structures in redundant data set is listed and their occurrence is converted in to frequencies by dividing their occurrence to number of the structures used for the analysis. They are then plotted according to their positions as bars as shown in Figure 6.2. Surprisingly all resides that are reported as utilised in binding inhibitors or HET groups falls on loop regions and a very few on neighbouring secondary structures.

Highest number of structures analysed (23 structures) are of trypsins, but the variation reported is larger in case of thrombin (13 structures). The reported residues on a linear sequence starts from residue number 34 (reported as interacting residue once and thrice respectively; chymotrypsin numberings used throughout the 'Results' and 'Discussion' sections), but for trypsin last residue reported in binding is 228 and for thrombin it is 245 (reported once each). For elastase, co-agulation factors and plasminogen activators the first and last residues reported interacting are 35, 41, 36 and 224, 228, 217 respectively (each reported once). While deciding for functional residues, it is important to put a cut-off value on data interpreted from frequency of binding. After studying the behaviour of all plots a cut off value of 0.4 was used to discriminate between commonly used residues and very specific residues for each protease or a false positive (see discussion). As it is evident from graphs corresponding to loop 9 where no peaks cross the cut off of 0.4. All the reported residues where checked for published reports of functionally mutations for all 5 types of proteases under study. A frequency value of 0.5 is decided as 'interacting' value by studying behaviour exhibited by His 57 residue of thrombin. Cases where the cut off values fall between 0.35 and 0.5 (as exhibited by behaviour of loop 14 for thrombin) were solved by studying the behaviour of analogues segments from other proteases under study and literature. This problem may arise because of the segment movements in reported structures.

The residues that come as interacting residue from loop3 are residues 40, 41 and 42 of trypsin and residue 41 of elastase. However, residues 38 and 40 of thrombin fall under investigating range. Catalytic His 57 (Loop 5) comes out as predominant residue for all of the proteases except co-agulation factors. Residues 60A-60F of thrombin are insertion relative to chymotrypsin (and other proteases). These residues are not reported as interacting residues in proteases under investigation and may be a unique insertion for thrombin. Residues 97 (investigating range), 99 of trypsin, 97A, 98, 99 of thrombin, 97, 98, 99 of co-agulation factors and 99 of elastase and plasminogen activators comes as interacting residues. The residues in loop 9 are well below cut off value and they seem to be examples of false positives. Loop11 contributes residue 174 in thrombin, co-agulation factors and plasminogen activators but in case of trypsin residue 175 is reported. Loop12

lines the S_i-S_i' residues and it is the most important loop for catalysis. It harbours Ser195. The residues 189 to 195 are reported as binding residues (except for residue 189 for elastase) in all sub families. The most important were results for loop 14. The residues 213-217, 219, 220 and 226 are reported as interacting residues in trypsin and plasminogen activators. Thrombin structures lack report of residue 216, coagulation factors lack report of reside 214 and elastase structures lack report of residues 213 and 214 but they have addition of residue 218 reported as interacting residue.

The alignment was analysed for conservation of interacting residues among sub-families. Interacting residues were mapped on to loop residues as shown in Figure 6.6 for trypsin and thrombin. The variation is mapped for regions in loop3, loop5, loop7, loop 12 and loop 14 on the alignment containing structures and sequences of all five sub-families as shown in the figure 6.6. Residues 40-42, 57, 60, 60A-F (for thrombin), 99, 174, 189-195 and 213-218 and 226 are identified as residues rendering functional specificity to respective proteases. Hence, the residues needed for specificity resides in loop regions and barrel structures are simply providing the scaffold needed. The experimental evidences supporting the above mentioned residues are discussed later. It should be noted that the evidences are not available for all residues for all five sub-types under examination.

6.4.2 Solvent Accessibilities and Hydrogen Bonding of Functional Regions

Mean side chain accessibilities of 43 proteases at each position of functional loops are plotted. The error bars are the variation observed from the mean value. The gap positions of each protease were assigned an arbitrary value of 100. Therefore positions showing accessibility values near 100 with very low error bars are actually gap regions. The side chains of residues in beta stands β2, β4, β7, β10, and β11 are buried in side the barrel structures. Catalytic His57 is adjacent to β4, catalytic Asp102 is adjacent to β7 and catalytic Ser195 is harboured by loop region of β10 and β11. The hydrogen bonding properties (main chain to main chain, side chain to main chain carbonyl and side chain to main chain amide H-bonding) of each loop region under examination were also plotted

for each positions for each of 43 proteases. The H-bonding is found to be conserved (data not shown) as evident from the JOY alignment of structures shown in Figure 6.3. These results suggest the catalytic triad residues to be very rigidly held by the adjacent secondary structures to maintain the geometry of catalytic triad. It is also evident from Figures 6.3 and 6.4 that side chains of catalytic His57 and Ser195 are only moderately solvent accessible (20-40%), but these residues are coming out as interacting residues in analysis with redundant data set. These results also support the fact that the catalytic triad is rigidly held in all proteases. The side chain of Asp102 is reported not at all solvent accessible. This fact is well known from previous studies and is also supported by the fact that Asp102 is not reported as interacting residues in analysis with redundant data set. The region of residues 215 to 226 shows highest variation in solvent accessibilities at each position with the fact that we are investigating family properties. This region is highly conserved among all known serine proteases. This points to the fact that it is the most highly mobile region among the structures under investigation and can be of functional importance.

6.4.3 Phylogeny and SCOP

We have used here NJ method (Saitou and Nei, 1987; Figure 6.5) provided by CLUSTALX8.1 package (Jeanmougin *et al*., 1998) to generate and visualise (NJview) the evolutionary tree of 43 structures of non-redundant database. As expected proteases with similar function are clustered together in the evolutionary tree. Here we discuss those examples where the proteases come together forming single node in a cluster but are assigned to different sub-family by SCOP (Murzin *et al*., 1995). The occurrence of the pair of proteins at a single node is also confirmed using PHYLIP3.5 (Felsenstein, 1985) that uses KITSCH algorithm. Trypsin from *Fusarium oxyparium* (1try-; the last character of PDB ID in each case shows the chain identifier) and *Streptomyce*s *griseus* (prokaryotic protease; 1sgt-) forms a pair different from rest of the trypsins. Human leukocyte elastase (1ppfe) gets clustered different from other elastases (1qnja, 1elt, and 1brup) but it pairs up with myeloblastin (1fuja). Tissue type plasminogen activators 1a5ha and 1a5ia groups together with other plasminogen activators (1ejna, 1ddja) with

different functional specificity but they are also assigned to different sub-families by SCOP (Murzin *et al*., 1995). Chymotrypsinogen C (1pytd) forms a pair with porcine pancreatic elastase (1brup) and cluster together with other elastase structures showing that it is related to elastase sub-family but it is grouped with chymotrypsin(ogen) by SCOP (Murzin *et al*., 1995).

1try- is grouped with other trypsin(ogen) in SCOP database (Murzin *et al*., 1995) and 1sgt- is grouped with prokaryotic proteases. These two proteins are similar in many respects. 1try- shares respectively 42% and 38% identity with 1sgt- and 5ptp- (bovine trypsin) but 5ptp- shares only 30% similarity with 1sgt-. However, R.M.S deviations (a measure of structural similarity) between 1try- and 1sgt-, 1try- and 5ptp-, and 1sgt- and 5ptp- are 1.16Å, 1.21Å and 1.21Å respectively. This shows that the backbones of all structures are similar. As shown in the alignment of Figure6.6 1sgt- and 1try- (with trypsin of lower organisms) also lacks disulfide bridges by C22-C157 and C128-C232 reported for trypsins from higher organisms. However, 1sgt- shares very less similarity with other prokaryotic serine proteases (for example ~20% with 1hpga, a glutamic-acid specific protease).

1ppfe is grouped with 1qnja, 1elt- and 1brup under elastase sub-family by SCOP (Murzin *et al*., 1995). While, 1fuja is defined as a lone member of myeloblastin sub-family. 1ppfe and 1fuja shares 55% sequence identity and are equally similar to other members of elastase sub-family (~36%). The R.M.S. deviation of the pair of 1ppfe and 1fuja is 0.68 Å, a virtually identical backbone. While that of 1ppfe and 1fuja with other members of elastase sub-family is ~1.19 Å and ~1.26 Å. These numbers are suggestive of the fact that they are sub-family members (for example, see the R.M.S. deviation reported for 1try- and 5ptp-). The alignment in Figure6.6 shows that the functional residues are very similar for 1fuja and 1ppfe but different than other elastases. This protein is also known to degrade elastin, fibronectin, laminin, vitronectin, and collagen types 1, 3, and 4 (Swissprot entry P24158 at www.expasy.ch). It is also known that genes for human neutrophile elastase (1ppfe), myeloblastin (1fuja) or proteinase3 (PR3) and azurocidin are organized as single genetic locus and are homologues (Zimmer *et al*., 1992).

Chymotrypsinogen C (1pytd) which also known as caldecrin or elastase4 shares 63% identity with porcine pancreatic elastase (1brup) but only 30% with leukocyte elastase (1ppfe). It also shares 41% identity with α-chymotrypsinogen. The R.M.S deviations of 1pytd-1brup, 1pytd-1ppfe and 1pytd-4cha are 1.02 Å, 1.37 Å and 1.29Å respectively. This suggests that 1pytd is equally related to both the subfamilies. Indeed 1pytd is reported to have characteristics of both elastase and chymotrypsin sub-families. As described it is sequentially more related to elastase sub-family but it shares disulfide bridge pattern and catalytic specificity of chmotrypsins (Gomis-Ruth *et al*., 1995). Caldecrin are known to be expressed in pancreas but TPCK doesn't inhibit them, suggesting that it is different than chymotrypsins (Yoshino-Yasuda *et al*., 1998). The enzyme comission numbers for elastase, caldecrin and chymotrypsin sub-families are 3.4.21.36, 3.4.21.2 and 3.4.21.1 respectively suggesting difference in the sub-families. Thus 1pytd should be assigned a different sub-family than both elastases and chymotrypsins.

Tissue type plasminogen activators 1a5ha and 1a5ia share 78% identity and the R.M.S. deviation for this pair is 0.78 Å. They are assigned the same Enzyme Commission number (EC 3.4.21.68) suggesting that the catalytic specificity and other features are same. There has been studies in past reporting that above 70% sequence similarities the function can be reliably transferred for sequences under consideration (Devos and Valencia, 2000; Wilson *et al*., 2000). Hence, this pair should also be considered as part of same sub-family.

6.4.4 Finding Function for Masquerade and others

Masquerade is a secreted molecule encoded by Drosophila *masquerade* (*mas*) gene. It is 1047 amino acid long and reported to contain an N-terminal domain containing five disulfide knotted motifs and C-terminal serine protease like domain (Murugasu-Oei *et al*., 1995).

BLASTP search (Altschul *et al*., 1997) of PDB database (Berman *et al*., 2000) and NR database at NCBI (www.ncbi.nlm.nih.gov) reports trypsins (~ 35% identity with bovine trypsin 2tld-) and thrombins (~ 32% identity with thrombin 1ucyk) as the closest homologues of masquerade. It is well known that masquerade is a non-functional serine protease due to mutation of catalytic serine residue to glycine at position 195 (Murugasu-Oei *et al*., 1995). The other proteins containing five repeats of disulfide motifs in arthropod serine proteases includes Sb-Sbd (Appel *et al*., 1992), snake (Delotto and Spierer, 1986) easter (Chasan and Anderson, 1989) and Limulus (Muta *et al*., 1990). Drosophila *GRAAL* or *Tequila* gene product and its *Anopheles gambiae* homologue Sp22D, which is associated with hemocytes and hemolymph, are also reported to have cysteine rich N-terminal domains (Danielli *et al*., 2000) and a functional C-terminal serine protease domain. In addition, both *GRAAL* gene product and Sp22D is shown to have poly-threonine stretches also common to masquerade. Sp22D is shown to be a chitin binding (adhesive) protein and it is suggested to serve as sentinel to detect exposed chitin, and then trigger appropriate physiological, developmental or immune response as chitin is also found on the surface of invading agents (Danielli *et al*., 2000). In drosophila the pathways involving toll ligand is implicated both in patterning drosophila embryo and generating early immune response and it involve serine proteases like Nudel, Gastrulation defective (GD), Snake and Easter (Lemosy *et al*., 2001; Levashina *et al*., 1999). Furthermore, the defensive prophenoloxidase cascade in *Manuuca sexta* is reported to be initiated by proteolytic processing (Jiang *et al*., 1998). The prophenoloxidase activating factor-1 (PPF1) in hemolymph of *Holotrichia dimphalia* larve (Lee *et al*., 1998) and *Anopheles gambiae* (Paskewitz *et al*., 1999) is shown to be a homologue of drosophila easter protease. Recently drosophila masquerade like proteins were reported from coleopteran *Holotrichia dimphalia* and (gi 10697070) *Tenebrio molitor* larvae (gi 10697178) and shown to be necessary for prophenoloxidase activity (Kwon *et al*., 2000). Analysis of sequences shows that they contain only one N-terminal disulfide knot and serine protease domain catalytic serine mutated to glycine. A cell adhesion protein containing multiple disulfide-knotted motif (total 7) and serine protease domain catalytic serine mutated to glycine is reported in crayfish *Pacifastacus lenisculus*.

Serine proteases like domain of masquerade sequences from drosophila, crayfish and coleopteran larves were aligned with the 43 proteins of non-redundant database using profile alignment mode of CLUSTALX8.1 (Jeanmougin *et al*., 1998). The resultant evolutionary tree has shown the sequences to be equally related to trypsins and thrombins (not shown). Noticeably the proteins contain disulfide bridge C1-C122 like thrombins which is absent in trypsins and C136-C201 like trypsins which is absent in thrombins. A BLASTP search (Altschul *et al*., 1997) is made on drosophila genome database at NCBI web site (ncbi.nlm.nih.gov) using serine protease like domain of masquerade as a query sequence (total 226 reported hits). The hits were analysed and grouped as masquerade homologue using disulfide bridge as criterion till the hit annotated having different function was found (first 34 hits). The gene product CG4998 is found to be among top hits with catalytic serine mutated to glycine. We hypothesise a masquerade like function for this gene product. The other hits with serine as catalytic residues are gene products CG5390, CG8586, CG8738, Tequila (GRAAL), CG13318, CG6639, CG2105, CG3117, CG14990, Nudel and CG18557. The GRAAL gene product, its mosquito homologue (Sp22D) and Nudel gene product were taken for further examination.

An alignment of Nudel, Snake, Easter (drosophila proteases involved in patterning), 5ptp-, 1c1uh, *GRAAL*, Sp22D and masquerade like proteins from drosophila, crayfish and coleopteran larvae serine protease like domain is shown in the Figure 6.7. The functional residues identified before are marked in the alignment. It is clear from the figure that the functional residues are mutated to random in all four masquerades like proteins with catalytic serine mutated to glycine. Hence, catalytic serine is not the only mutated functional residue. Thus suggesting that they can not function as a competitive antagonist of serine proteases.

Five cysteine rich repeats are identified by careful analysis of masquerade N-terminal domain. Similar repeats were also found in other masquerade like proteins (not shown). An alignment of the repeats with ‘chitin binding motif’ reported by Suetake *et al* (2000) shows each repeat contained one chitin binding motif as shown in Figure 6.8. Chitin

binding motifs of GRAAL and Sp22D as identified are also shown in the figure 6.8. Thus showing masquerade having a chitin binding or very similar activity.

6.5 Discussion

We have identified positions 40, 41 and 42 (loop3), 57, 60, and 60A-F (loop5), 97 and 99 (loop7), 174 (loop11), 189-195 (loop12) and 213-219 and 226 (loop14) as residues rendering specificity and catalysis to serine proteases. Catalytic Asp102 was not reported as it is completely buried and doesn't interact directly with substrate or inhibitors. The barrel structures are shown only as supporting structures for loops to carry out a particular function. The allosteric sites were not identified by this study since all the HETATM (mostly ions) entries other than specific inhibitors were removed from the PDB files. We than searched for literature reports of structure function relationships for the proteases.

Indeed, there has been a wealth of information of mutagenesis studies in chosen sub-families of serine proteases, while others are uncharacterised in terms of function (Maxwell *et al*., 1999). New members of the family are constantly being added due to genome sequencing efforts (for example, 29 out of 34 top hits (from Drosophila genome database) examined in this work are not annotated). Hence, searching for "consensus" positions providing functional specificity is important for fast characterisation of known sequences. For reports correspond to catalytic triad and description of catalytic mechanism we recommend analysis by Lesk and Fordham, a review by Perona and Craik and literature cited within (Lesk and Fordham, 1996; Perona and Craik, 1995).

The diversity of substrate specificity among the chymotrypsin like proteases rests upon small differences in structure of the substrate-binding cleft composed of two juxtaposed β-barrel domains, with catalytic residues bridging the barrels (Figure1; Kraut, 1977; Steitz and Shulman, 1982; Bazan and fletterick, 1990). Position 189, located at the base of S1 pocket, is highly conserved as an Asp in enzymes with trypsin like specificity. It is found as Ser or other small amino acid in chymotrypsin and elastase like enzymes.

Position 190 extends into the base of the pocket as well as plays an additional role to modulate specificity profile. Amino acids at positions 216 and 226 are usually Gly in both trypsin and chymotrypsin-like enzymes; larger amino acid side chains at positions partially or fully block access of larger substrate side chains to the base of the pocket (Craik *et al*., 1985; Wilke *et al*., 1991). Accordingly elastases possess larger, usually non-polar residues at this positions, providing a platform for interaction with small hydrophobic substrates (Perona and Craik, 1995). Consequently, C-terminal sequence is postulated to encode function for serine proteases (Stroud, 1974; Maxwell *et al*., 1999). A view supported by the fact in protease structures as the C-terminal end is approached, the surface area containing the substrate increases sharply. The residue 192 is shown to be important for blood coagulation and fibrinolytic systems but not tissue type plasminogen activators and have different roles in other sub-families (Zhang *et al*., 1999). The residues 189-220 in C-terminal sequences were found to account for >95% of the area around the specificity pocket S1 and catalytic His57 and >70% of the area of around specificity sites S2 and S3 (Maxwell *et al*., 1999). Role of residue 172 for trypsin substrate specificity is also known (Hedstrom *et al*., 1994).

But this view is not entirely correct as N-terminal residues are identified as the functional residues. Role of residues 60B-F (loop5), Trp96 and effect of charge reversal of Arg93, 97 and 99 (loop7) for thrombin has been reported (DiBella and Scharaga, 1998; He *et al*., 1997). Glu39 of thrombin is reported to play part in P3 specificity (Le Bonniec *et al*., 1991). Role of residue 99 of factor Xa, activated protein C and thrombin is shown to exhibit P2 specificity (Rezaie, 1997). Role of loop5 (loop60) is reported to be instrumental in S1′ specificity for trypsin (Kurth *et al*., 1997). But these reports has been scattered, information is derived from largely biochemical analysis and discussed about role of particular residue for only particular sub-family. We have identified these residues on the basis of structural and sequence analysis and hypothesise the role of the consensus residues in rendering specificity in all the characterised and yet to be characterised sub-families.

Loop12 and Loop14 are longest loops in the serine proteases that take part in the catalysis and substrate specificity. Loop 12 and loop14 are also highly conserved among all the proteases. Interestingly the movement of the loop12 is shown to be restricted by the secondary structures from both the sides. Here it should be noted that the loop is glycine rich and is reported in limited amount of conformational change forming the oxyanion hole for catalysis by mediating changing in hydrogen bonding state of residue 194 (Lesk and Fordham, 1996; Peisch *et al*., 1999). Loop14 however on the contrast shows the highest variability in the solvent accessibility analysis. This simply suggests that this loop is highly mobile in all 43 structures used for the analysis. According to the expectations, loop14 of proenzyme domain of plasminogen is reported to have an entirely different loop conformation than active conformation of trypsin and chymotrypsingoen showing W214 side chain blocking the S1 specificity pocket in a foot in mouth mechanism of inactivation (Peisach *et al*., 1999). Residues 214-220 of chymotrypsinogen that makes up the opposite of S1 subsite is reported to narrow down active-site pocket. Such change is not reported for trypsinogens and plasmins (Parry *et al*., 1998). Thus analysis solely based on structural properties is capable of identifying the functional sites in proteins. Though it requires large amount of structural information which we hope to be common in future with high number of redundant structures deposition and current rate of growth of PDB database (http://www.rcsb.org/pdb/holdings.html).

Homology of *Streptomyces griseus* trypsin (1sgt-) with other eukaryotic proteases has been reported and attributed to gene transfer from eukaryote to the bacterium as early as in 1970 (Hartely, 1970). It is grouped differently as the proteases are grouped on the basis of the source unlike other proteins in SCOP database (Murzin *et al*., 1995). We propose that 1sgt- should be grouped with other eukaryotic trypsins. The tissue type plasminogen activators 1a5ha and 1a5ia shows identical functional residues identified by our analysis and other similarities described above and hence should be grouped under single sub-family. The neutrophile elastase (1ppfe) and myeloblastin (1fuja) shows conserved functional residues among the pair but different than other elastases at positions 41, 60, 189, 190, 191 and 220 leading to the suggestions that these pair should be classified as differently than other elastases. We also propose that chymotrypsinogen C to be

considered in a different sub-family than both elastase and chymotrypsin sub-families. These reported differences with the SCOP classification however points to an interesting question of definition of function. There is no clear measure for functional similarity. The definition of function itself is often vague. For example, all the proteins under consideration in this paper serve function of a protease and cleave the scissile bond. But they have been divided or grouped as 'different' on the basis of the substrate specificity or catalytic mechanism. It is also evident form the case of chymotrypsinogen C that functional similarity can not be inferred always with confidence on the basis of sequence similarities. Hence, such differences should be considered subjective and waiting for clearer structure-function relationships.

At last we apply our analysis of serine protease domain and hits identified from the drosophila genome to a cell adhesion protein masquerade and its homologues reported functioning in adhesion as well as immune responses. The closest functional serine protease domain from Drosophila genome shares 34% identity with that of masquerade. As shown previously all the functional residues of masquerade and its homologues have been mutated randomly. Thus, suggested role of masquerade acting, as antagonist of a seine protease domain is very unlikely. Instead we have identified 5 chitin binding (adhesive) motifs in masquerade N-terminal cysteine-knotted repeats. Such motifs are also identified for GRAAL gene product and Sp22D protein of Anopheles in this study. As mentioned before many proteases simultaneously involved in patterning and immune response possess such motifs. We propose that these proteins carry out their both the roles with common mechanism of adhesion (to chitin of invading pathogen or a tissue for degradation) and subsequent protease activity. We strongly attribute this role to GRAAL gene product and Sp22D but also in general all the proteins involved in patterning and immune response (like snake, easter etc). The role of Sp22D during early immune response is demonstrated (Danielli *et al*., 2000). In addition to this we propose the involvement of GRALL and Sp22D in development. We suggest that the C-terminal domain of masquerade like proteins are acting as prophenoloxidase response factor or serving a signalling role during the early development reported for HGF/SF like proteins with non-functional serine protease domain (Thery *et al*., 1995). This suggestion is

strengthen by the fact that total loss of function of *mas* gene is embryonic lethal (Murugasu-Oei *et al*., 1995) and masquerade for its homologues with prophenyloxidase activity contain only one improper cysteine-knot motif (unpublished results).

6.6 Conclusions

In this study the consensus residues involved in rendering in substrate specificity in eukaryotic serine proteases has been identified by analysis of redundant and non-redundant data set of structures and sequence information. We have predicted functional sites on the basis of structural properties like solvent accessibility and hydrogen bonding (not shown) analysis of non-redundant data set and showed that the secondary structures adjacent to catalytic triad residues are immobile and maintain rigid geometry required for efficient catalysis. The results were applied to proteins 'masquerading' its real function. Chitin binding motifs have been identified in masquerade, GRAAL and Sp22D and multiple roles of during adhesion process, immune response and development has been suggested for the proteins.

6.7 References

Andersen, N. H., Cao, B., Rodriguez-Romero, A., and Arreguin, B. (1993). Hevein: NMR assignment and assessment of solution-state folding for the agglutinin-toxin motif. *Biochemistry* **32**, 1407-22.

Appel, L. F., Prout, M., Abu-Shumays, R., Hammonds, A., Garbe, J. C., Fristrom, D., and Fristrom, J. (1993). The Drosophila Stubble-stubbloid gene encodes an apparent transmembrane serine protease required for epithelial morphogenesis. *Proc Natl Acad Sci U S A*, **90**, 4937-41

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-402.

Baba, T., Watanabe, K., Kashiwabara, S., Arai, Y. (1989). Primary structure of human proacrosin deduced from its cDNA sequence. *FEBS Lett* **44**, 296-300.

Baker, B. M., and Murphy, K. P. (1997). Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase. *J Mol Biol*, **268**, 557-69.

Barrett, A. J. (1977). Editor of Proteinases in Mammalian Cells and Tissues. North-holland, amesterdam.

Barrett, A. J. (1994). Editor of Proteolytic Enzymes: Serine and Cysteine Peptidases. Methods Enzymol, 244, Academic Press, New York.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res*. **28**, 263-6.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, **28**, 45-8.

Bazan, J. F., and Fletterick, R. J. (1990). Structural and catalytic models of trypsin like viral proteases. *Semin Virol*. **1**, 311-22.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*,**7**, 957-9.

Bernstein, H. J. (2000). Recent changes to RasMol, recombining the variants. *Trends Biochem Sci*, **25**, 453-5.

Birktoft, J. J., and Blow, D. M. (1972). Structure of crystalline -chymotrypsin. V. The atomic structure of tosyl--chymotrypsin at 2 A resolution. *J Mol Biol.* **68**, 187-240.

Bond, J. S. (1991). Plasma-membrane proteases-introductory remarks. *Biomed. Biochem. Acta*, 50, 775-780.

Boggon, T. J., Shan, W. S., Santagata, S., Myers, S. C., and Shapiro, L. (1999). Implication of tubby proteins as transcription factors by structure-based functional analysis.

Chasan, R, and Anderson, K. V. (1989). The role of easter, an apparent serine protease, in organizing the dorsal-ventral pattern of the Drosophila embryo. *Cell*, **56**, 391-400.

Chothia, C., and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**, 705-8.

Craik, C. S., Largman, C., Fletcher, T., Roczniak, S., Barr, P. J., Fletterick, R., and Rutter, W. J. (1985). Redesigning trypsin: alteration of substrate specificity. *Science*, **228**, 291-7.

Danielli, A., Loukeris, T. G., Lagueux, M., Muller, H. M., Richman, A., and Kafatos, F. C. (2000). A modular chitin-binding protease associated with hemocytes and hemolymph in the mosquito Anopheles gambiae. *Proc Natl Acad Sci U S A*, **97**, 7136-41.

DeLotto, R., and Spierer, P. (1986). A gene required for the specification of dorsal-ventral pattern in Drosophila appears to encode a serine protease. Nature, 323, 688-92.

Devos, D., and Valencia, A. (2000). Practical limits of function prediction. *Proteins*, **41**, 98-107.

DiBella, E. E., Scheraga, H. A. (1998). Thrombin specificity: further evidence for the importance of the beta-insertion loop and Trp96. Implications of the hydrophobic interaction between Trp96 and Pro60B Pro60C for the activity of thrombin. *J Protein Chem*, **17**, 197-208.

Dodson , G. G., Lawson, D. M., and Winkler, F. K. (1992). Structure and evolutionary relationships in the lipase mechanism and activation. *Faraday Discuss*. **93**, 95-105.

Donate, L. E., Gherardi, E., Srinivasan, N., Sowdhamini, R., Aparicio, S., and Blundell, T. L. (1994). Molecular evolution and domain structure of plasminogen-related growth factors (HGF/SF and HGF1/MSP). *Protein Sci.* **3,** 2378-94.

Felsenstein, J. (1985). *Evolution*, **39**, 583.

Fersht, A. (1984). *Enzyme Structure and Mechanism*, 2nd edit., W.H. Freeman, San Fransisco, CA.

Froelich, C. J., Zhang, X., Turbov, J., Hudig, D., Winkler, U., Hanna, W. L. (1993). Human granzyme B degrades aggrecan proteoglycan in matrix synthesized by chondrocytes. *J Immunol*, **151**,7161-71.

Goldberger, G., Bruns, G. A., Rits, M., Edge, M. D., and Kwiatkowski, D. J. (1987). Human complement factor I: analysis of cDNA-derived primary structure and assignment of its gene to chromosome 4. *J Biol Chem*, **262**, 10065-71.

Gomis-Ruth, F. X., Gomez, M., Bode, W., Huber, R., and Aviles, F. X. (1995). The three-dimensional structure of the native ternary complex of bovine pancreatic procarboxypeptidase A with proproteinase E and chymotrypsinogen C. *EMBO J*, **14**, 4387-94.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**, 4355-8.

Gurwitz, D., Cunningham, D.D. (1988). Thrombin modulates and reverses neuroblastoma neurite outgrowth. *Proc Natl Acad Sci U S A*, **85**, 3440-4.

Hartley, B. S. (1970). Homologies in serine proteinases. *Philos Trans R Soc Lond B Biol Sci*, **257**, 77-87.

He, X., Ye, J., Esmon, C. T., Rezaie, A. R. (1997). Influence of Arginines 93, 97, and 101 of thrombin to its functional specificity. Biochemistry 1997, 36, 8969-76.

Hedstrom, L., Szilagyi, L., and Rutter, W. J. (1992) Converting trypsin to chymotrypsin: the role of surface loops. *Science*, **255**, 1249-53.

Hedstrom, L., Perona, J. J., and Rutter, W. J. (1994). Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry*, **33**, 8757-63.

Henderson, B. R., Tansey, W. P., Phillips, S. M., Ramshaw, I. A., Kefford, R.F. (1992). Transcriptional and posttranscriptional activation of urokinase plasminogen activator gene expression in metastatic tumor cells. *Cancer Res*, **52**, 2489-96.

Horl, W. H. (1989). Proteinases: potential role in health and disease. In *Design of Enzyme Inhibitors as Drugs* (Sandler, M & Smith, H. J., eds), pp.573-581, Oxford University Press, Oxford.

Huang, T. S., Wang, H., Lee, S.Y., Johansson, M. W., Soderhall, K., and Cerenius, L. (2000). A cell adhesion protein from the crayfish Pacifastacus leniusculus, a serine proteinase homologue similar to Drosophila masquerade. *J Biol Chem*, **275**, 9996-10001.

Hung, S. H., and Hedstrom, L. (1998). Converting trypsin to elastase: substitution of the S1 site and adjacent loops reconstitutes esterase specificity but not amidase activity. *Protein Eng*, **11**, 669-73.

Hubbard, T. J., and Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling.

Hubbard, S. J., Eisenmenger, F., and Thornton, J. M. (1994). Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci*, **3**, 757-68.

Iengar, P., and Ramakrishnan, C. (1999). Knowledge-based modeling of the serine protease triad into non-proteases. *Protein Eng*, **12**, 649-56.

Janin, J., and Chothia, C. (1976). Stability and specificity of protein-protein interactions: the case of the trypsin-trypsin inhibitor complexes. *J Mol Biol*, **100**, 197-211.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem Sci*, **23**, 403-5.

Jiang, H., Wang, Y., and Kanost, M. R. (1998). Pro-phenol oxidase activating proteinase from an insect, Manduca sexta: a bacteria-inducible protein similar to Drosophila easter. *Proc Natl Acad Sci U S A*, **95**, 12220-5.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-9.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, **287**, 797-815.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-637.

Karplus, K, Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. (1997). Predicting protein structure using hidden Markov models. *Proteins*, Suppl **1**, 134-9.

Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol*, **299**, 499-520.

Kraut, J. (1971). Subtilisin: X-ray structure. In *The Enzymes* (Boyer, P. D., ed.), vol. 3, pp. 547-560, Academic Press, New York and London.

Kraut, J. (1977). Serine proteases: structure and mechanism of catalysis. *Annu Rev Biochem*, **46**, 331-58.

Krem, M. M., Rose, T., Di Cera, E. (1999). The C-terminal sequence encodes function in serine proteases. *J Biol Chem*, 274, 28063-6.

Kurth, T., Ullmann, D., Jakubke, H. D., and Hedstrom, L. (1997). Converting trypsin to chymotrypsin: structural determinants of S1' specificity. *Biochemistry*, **36**, 10098-104.

Kwon, T. H., Kim, M. S., Choi, H. W., Joo, C. H., Cho, M. Y., Lee, B. L. (2000). A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, Holotrichia diomphalia larvae. *Eur J Biochem*, **267**,6188-96.

Laskowski, R. A., Moss, D. S., and Thornton, J. M. (1993). Main-chain bond lengths and bond angles in protein structures. *J Mol Biol*, **231**, 1049-67.

Levashina, E. A., Langley, E., Green, C., Gubb, D., Ashburner, M., Hoffmann, J. A., and Reichhart, J. M. (1999). Constitutive activation of toll-mediated antifungal defense in serpin-deficient Drosophila. *Science*, **285**, 1917-9.

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, **55**, 379-400.

LeMosy, E. K., Tan, Y. Q., and Hashimoto, C. (2001). Activation of a protease cascade involved in patterning the Drosophila embryo. *Proc Natl Acad Sci U S A*, 98, 5055-60.

Lesk, A.M. (1981). *Introduction to Physical Chemistrey*, sections 18-10, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Lesk, A. M., and Fordham, W. D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *J Mol Biol* **258**, 501-37.

Liao, D. I., Breddam, K., Sweet, R. M., Bullock, T., and Remington, S. J. (1992). Refined atomic model of wheat serine carboxypeptidase II at 2.2-A resolution. *Biochemistry*, **31**, 9796-812.

Luthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-5.

Matthews B. W., Sigler, P. B., Henderson, R., and Blow, D. M. (1967).Three-dimensional structure of tosyl-alpha-chymotrypsin. *Nature*, **21**, 4652-6.

McLachlan, A. D., and Shotton, D. M. (1971). Structural similarities between alpha-lytic protease of Myxobacter 495 and elastase. *Nat New Biol*, **229**,202-5.

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617-23.

Murugasu-Oei, B., Rodrigues, V., Yang, X., and Chia, W. (1995). Masquerade: a novel secreted serine protease-like molecule is required for somatic muscle attachment in the Drosophila embryo. *Genes Dev.* **9**, 139-54.

Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.*, **247**, 536-40.

Muta, T., Hashimoto, R., Miyata, T., Nishimura, H., Toh, Y., and Iwanaga, S. (1990). Proclotting enzyme from horseshoe crab hemocytes. cDNA cloning, disulfide locations, and subcellular localization. *J Biol Chem.* 265, 426-33.

Neurath, H. (1985). Proteolytic enzymes, past and present. *Fed Proc.* **44**, 2907-13.

Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc R Soc Lond B Biol Sci*., **241**, 132-45.

Parry, M. A., Fernandez-Catalan, C., Bergner, A., Huber, R., Hopfner, K. P., Schlott, B., Guhrs, K. H., and Bode, W. (1998). The ternary microplasmin-staphylokinase-microplasmin complex is a proteinase-cofactor-substrate complex in action. *Nat Struct Biol*, **5**, 917-23.

Paskewitz, S. M., Reese-Stardy, S., Gorman, M. J. (1999). An easter-like serine protease from Anopheles gambiae exhibits changes in transcript abundance following immune challenge. *Insect Mol Biol*, **8**,329-37.

Peisach, E., Wang, J., de los Santos, T., Reich, E., and Ringe, D. (1999). Crystal structure of the proenzyme domain of plasminogen. *Biochemistry*, **38**, 11180-8.

Pendurthi, U. R., Allen, K. E., Ezban, M., Rao, L. V. (2000). Factor VIIa and thrombin induce the expression of Cyr61 and connective tissue growth factor, extracellular matrix signaling proteins that could act as possible downstream mediators in factor VIIa x tissue factor-induced signal transduction. *J Biol Chem* **275**, 14632-41.

Perona, J. J., and Craik, C. S. (1995). Structural basis of substrate specificity in the serine proteases. *Protein Sci*, **4**, 337-60.

Perona, J. J., and Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J Biol Chem*, **272**, 29987-90.

Reid, K. B. M., Bentley, D. R., Campbell, R. D., Chung, L. D., Sim, R. B. Kristensen, T. and Tack, B. F. (1986). Complement system proteins which interact with C3B or C4B. *Immunol. Today*, **7**, 230-234.

Russell, R. B., and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309-23.

Smirnova, I. V., Citron, B. A., Arnold, P. M., Festoff, B. W. (2001). Neuroprotective signal transduction in model motor neurons exposed to thrombin: G-protein modulation effects on neurite outgrowth, Ca(2+) mobilization, and apoptosis. *J Neurobiol*, **48**,87-100.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779-815.

Sali, A., and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**, 403-28.

Sayle, R. A., and Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, **20**, 374.

Saitou, N, and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-25.

Shapiro, L., and Harris, T. (2000). Finding function through structural genomics. *Curr Opin Biotechnol*, **11**, 31-5.

Shi, J, Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**, 243-57.

Srinivasan, N., and Blundell, T. L. (1993). An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng*, **6**, 501-12.

Steitz, T. A., and Shulman, R. G. (1982). Crystallographic and NMR studies of the serine proteases. *Annu Rev Biophys Bioeng*, **11**, 419-44.

Stroud, R. M. (1974). A family of protein-cutting proteins. *Sci Am*. **231**, 74-88.
Twining, S. S. (1994). Regulation of proteolytic activity in tissues. *Crit Rev Biochem Mol Biol*, **29**, 315-83.

Suetake, T., Tsuda, S., Kawabata, S., Miura, K., Iwanaga, S., Hikichi, K., Nitta, K., and Kawano, K. (2000). Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. *J Biol Chem*, **275**, 17929-32.

Thery, C., Sharpe, M. J., Batley, S. J., Stern, C. D., and Gherardi, E. (1995). Expression of HGF/SF, HGF1/MSP, and c-met suggests new functions during early chick development. *Dev Genet*, **17**, 90-101.

Tsiang, M., Paborsky, L. R., Li, W. X., Jain, A. K., Mao, C. T., Dunn, K. E., Lee, D.W., Matsumura, S.Y., Matteucci, M. D., Coutre, S. E., Leung, L. L., and Gibbs, C. S. (1996). Protein engineering thrombin for optimal specificity and potency of anticoagulant activity in vivo. *Biochemistry*, **35**, 16449-57.

Watorek, W., Farley, D., Salvensen, G., and Travis, J. (1988). Nwutrophil elastase and CatepsinG: structure, function and biological control. *Advn. Expt. Med. Biol.* **240**, 23-31.

Wilke, M. E., Higaki, J. N., Craik, C. S., Fletterick, R. J. (1991). Crystallographic analysis of trypsin-G226A. A specificity pocket mutant of rat trypsin with altered binding and catalysis. *J Mol Biol*, **219**, 525-32.

Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, **297**, 233-49.

Yang, F., Gustafson, K. R., Boyd, M. R., and Wlodawer, A. (1998). Crystal structure of Escherichia coli HdeA. *Nat Struct Biol*, **5**, 763-4.

Zarembinski, T. I., Hung, L. W., Mueller-Dieckmann, H. J., Kim, K. K., Yokota, H., Kim, R., and Kim, S. H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc Natl Acad Sci U S A*, **95**, 15189-93.

Zhang, Y. L., Hervio, L., Strandberg, L., Madison, E. L. (1999). Distinct contributions of residue 192 to the specificity of coagulation and fibrinolytic serine proteases. *J Biol Chem*, **274**, 7153-6.

Zimmer, M., Medcalf, R. L., Fink, T. M., Mattmann, C., Lichter, P., and Jenne, D. E. (1992). Three human elastase-like genes coordinately expressed in the myelomonocyte lineage are organized as a single genetic locus on 19pter. *Proc Natl Acad Sci U S A*, **89**, 8215-9.

Figure Legends

Figure 6.1.

The Figure shows typical trypsin fold serine protease domain (shown bovine trypsin; PDB ID 5ptp-). The catalytic residues and loop regions spatially proximate to catalytic residues are marked. The figure was prepared using Setor (Evans, 1993).

Figure 6.2.

The Figure shows the binding site analysis as reported. The binding residues were defined as those having atoms less than 4Å apart. The reasons for considering the redundant structures can be 1) different inhibitors 2) different mutations 3) speciation and 4) Literature 5) deletions etc. Catalytic residues are marked.

Proteins used for Analysis are as follows:

Trypsin Structures

3tgj	RattusNorvegicus	BPTI
1bra	pig (D189G, G226D)	BENZAMIDINE
1anb	Rattus Rattus (S 214 E)	BENZAMIDINE
1and	Rattus Rattus (R 96 H)	BENZAMIDINE
1bit	Salmon (different crystal)	BENZAMIDINE
1c9p	Pig	Bdellastasin
2sta	Salmon	Squash Seed Inhibitor
1zzz	Bovine	C9H18N4O2 and C5H11N1O2
1ezs	Rattus Norvegicus	Ecotin Mutant
1f0t	Bovine	Rpr131247
1f2s	bovine beta trypsin	Mcti-A
1g3b	bovine beta trypsin	Meta-Amidino Schiff Base
1ldt	pig	Leech-Derived Tryptase Inhibitor
1ntp		C3H8O3P1 (MIP)

1ql9	X99Rt	Factor Xa Specific Inhibitor
1slw	Rat (N143H, E151H)	Ecotin (Nickel bound)
1slx	Rat (N143H, E151H)	Ecotin (Zinc bound)
1taw	Rat	Appi
1tgs		Porcine Pancreatic Secretory Inhibitor
1fy8	trypsinogen delta 16, 17	Bpti
1a0l		Appa
2trm	(D 102 N) at PH 7	BENZAMIDINE

Thrombin

1b7x	Thrombin Y225I	D-Phe-Pro-Arg-Chloromethylketone
1bhx	Human	Sdz (C19H28N6O4S2)
1thp	Human Y225P	D-Phe-Pro-Arg-Chloromethylketone
1bth	Bovine	BPTI
1d6w	Human	Decapeptide Inhibitor
1d9i	Human	Hirugen
1dm4	Human S195A	Fibrinopeptide A
1qhr	Human	TYS (C9H11N1O6S1)
1doj	Human	Rwj-51438
1e0f	Human	Haemadin
1eoj	Human	CPI and TIH
1vr1	Human	Plasminogen Activator Inhibitor-1
2thf	Human Y225F	D-Phe-Pro-Arg-Chloromethylketone

Elastase

1b0f	Human	Mdl 101, 146
1hne	Human	MSACK
1ppf	Human	turkey ovomucoid inhibitor

1ppg	Human	chloromethyl keton inhibitor
1bru	Pig	Gr143783
1qr3	Pig	Fr90127

Coagulation Facotor

1ezq	Human	Rpr128515
1fax	Human	DX9
1fjs	Human	Zk-807834
1kig	Bovine	Anticoagulant Peptide
1xka	Human	Fx-2212A
1pfx	Pig	D-Phe-Pro-Arg

Plasminogen Activators

1a5h	Human	Bis-Benzamidine
1a5i	Vampire Bat	Egr-Cmk
1bda	Human	Dansyl-Egr-Cmk
1bqy	Snake Venom	Chloromethylketone Inhibitor
1c5w	Human	ESI and FLC
1ejn	Human	Phenylguanidine
1lmw	Human	DEOXY-METHYL-ARGININE

Figure 6.3

The alignment of structures was built using STAMP (Russell and Barton, 1992) and annotated using JOY (Overington *et al*., 1990; Mizuguchi *et al*., 1998). Secondary structures and loop regions are as marked. Loop nomenclature was adopted from Peisach *et al*., (1999).

The proteins in the alignments are as follows.
1a0j Trypsin ; 1a0l Beta Tryptase; 1a5h Two-Chain Tissue Plasminogen Activator; 1a5i Saliva Plasminogen Activator; 1a7s Heparin Binding Protein; 1agj Epidermolytic Toxin A; 1ao5 Glandular Kallikrein-13; 1aut Activated Protein C; 1azz Collagenase; 1bio Complement Factor D; 1bqy Plasminogen Activator; 1bru Pancreatic Elastase; 1c1u Alpha Thrombin; 1cgh Cathepsin G; 1ddj Plasminogen Catalytic Domain; 1dpo Anionic Trypsin; 1dva Coagulation Factor VIIa; 1ejn Urokinase Plasminogen Activator; 1ekb Enteropeptidase; 1elt Native Pancreatic Elastase; 1fon Procarboxypeptidase; 1fuj Myeloblastin (PR3); 1fxy Coagulation Factor Xa-Trypsin Chimera; 1hcg Blood Coagulation Factor Xa; 1kig Bovine Factor Xa; 1klt Chymase; 1npm Neuropsin; 1pfx Factor Ixa; 1ppf Leukocyte Elastase; 1pyt Chymotrypsinogen C; 1qnj Pancreatic Elastase; 1qqu Beta Trypsin; 1rfn Coagulation Factor Ixa; 1sgf Nerve Growth Factor; 1sgt Streptomyces griseus trypsin; 1ton Tonin; 1trn Trypsin 1; 1try Fusarium-Oxysporum Trypsin; 2hlc Collagenase; 2pka Pancreatic Kallikrein A; 2sga Streptomyces griseus protease A; 2tbs Trypsin; 3rp2 Mast Cell Protease II; 4cha Alpha Chymotrypsin; 5ptp Beta Trypsin

Figure 6.4

Solvent accessibilities of 43 proteases structures listed in legend of Figure 6.3 were calculated using PSA (Lee and Richards, 1971) after removing all nonprotease entries from the PDB files. The mean accessibility (Y-axis) for each alignment position (X-axis) is shown as a solid bar and the root mean square deviation is shown as error bar.

Secondary structures are marked. The gap regions were assigned an arbitary accessibility of 100. The analysis was done using loop regions spatially proximate to catalytic triad.

Figure 6.5

Figure5 displays a tree calculated by CLUSTALX8.1 (Jeanmougin and Thompson, 1998) using Neighbour-Joining method (Saitou and Nei, 1987) from alignment of serine proteases shown in Figure3. The protein codes are as described in legend of Figure 6.3. Branch lengths are proportional to sequence divergence and can be measured relative to bar shown (top right). Branch labels record the stability of the branches over 1000 bootstrap replicates.

Figure6 6

The multiple alignment was prepared using profile mode of CLUSTALX8.1 (Jeanmougin and Thompson, 1998), where annotated sequences from the Swissprot database (Bairoch and Apweiler, 2000) are added to the structure based alignment prepared using STAMP (Russell and Barton, 1992). The residues identified as “binding” residues (see results) are marked with boxes. It should be noted that the marked residues are either absolutely conserved or sub-family specific, hence assumed to render specificity to the proteases.

The swissprot accession numbers (for structures please see legend of Figure 6.3) and short description for the sequences are as follows.
TRY2_MOUSE sp|P07146| Trypsin II Anionic Precursor; TRY1_CANFA sp|P06871| Trypsinogen Cationic Precursor; TRY1_CHICK sp|Q90627| Trypsin I-P1 Precursor; TRY1_XENLA sp|P19799| Trypsin Precursor; TRY1_GADMO sp|P16049| Trypsin I Precursor; TRY1_SALSA sp|P35031| TRypsin I Precursor; TRYP_SQUAC sp|P00764| Trypsin Precursor; 1TRY-_FUSOX Trypsin; 1SGT-_STRGR Streptomyces Griseus Trypsin; TRYA_DROER sp|P54624| Trypsin Alpha Precursor; TRYA_DROME sp|P04814| Trypsin Alpha Precursor; TRY4_LUCCU sp|P35044| Trypsin Alpha-4 Precursor; TRYE_DROER sp|P54627| Trypsin Epsilon Precursor; TRYE_DROME

sp|P35005| Trypsin Epsilon Precursor; TRYT_DROER sp|P54628| Trypsin Theta Precursor; TRYT_DROME sp|P42278| Trypsin Theta Precursor; TRYP_SARBU sp|P51588| Trypsin Precursor; TRYI_DROME sp|P52905| Trypsin Iota Precursor; TRYU_DROER sp|P54629| Trypsin Eta Precursor; TRYU_DROME sp|P42279| Trypsin Eta Precursor; TRYZ_DROER sp|P54630| Trypsin Zeta Precursor; TRYZ_DROME sp|P42280| Trypsin Zeta Precursor; TRY1_ANOGA sp|P35035| Trypsin 1 Precursor; TRY3_AEDAE sp|P29786| Trypsin 3A1 Precursor; TRYP_SIMVI sp|P35048| Trypsin Precursor; TRYA_MANSE sp|P35045| Trypsin Alkaline A Precursor; TRYP_CHOFU sp|P35042| Trypsin CFT-1 Precursor; THRB_MOUSE sp|P19221| ProThrombin Precursor; THRB_RAT sp|P18292| ProThrombin Precursor; THRB_BOVIN sp|P00735| ProThrombin Precursor; PLMN_BOVIN sp|P06868| Plasminogen Precursor; PLMN_SHEEP sp|P81286| Plasminogen; PLMN_PIG sp|P06867| Plasminogen; PLMN_MACMU sp|P12545| Plasminogen Precursor; PLMN_CANFA sp|P80009| Plasminogen; PLMN_MOUSE sp|P20918| Plasminogen Precursor; PLMN_ERIEU sp|Q29485| Plasminogen Precursor; PLMN_HORSE sp|P80010| Plasminogen; TPA_BOVIN sp|Q28198| Tissue-Type Plasminogen Activator Precursor; TPA_MOUSE sp|P11214| Tissue-Type Plasminogen Activator Precursor; TPA_RAT sp|P19637| Tissue-Type Plasminogen Activator Precursor; UROK_PAPCY sp|P16227| Urokinase-Type Plasminogen Activator Precursor; UROK_PIG sp|P04185| Urokinase-Type Plasminogen Activator Precursor; UROK_BOVIN sp|Q05589| Urokinase-Type Plasminogen Activator Precursor; UROK_MOUSE sp|P06869| Urokinase-Type Plasminogen Activator Precursor; UROK_RAT sp|P29598| Urokinase-Type Plasminogen Activator Precursor; UROK_CHICK sp|P15120| Urokinase-Type Plasminogen Activator Precursor; HCGA_HUMAN Blood Coagulation Factor Xa; KIGH_BOVINE Factor Xa; A10_RABIT sp|O19045| Coagulation Factor X Precursor; A10_CHICK sp|P25155| Coagulation Factor X Precursor; A10_TROCA sp|P81428| Coagulation Factor X; DVAH_HUMAN Coagulation Factor VIIa; A7_MOUSE sp|P70375| Coagulation Factor VII Precursor; FA7_RABIT sp|P98139| Coagulation Factor VII Precursor; FA7_BOVIN sp|P22457| Coagulation Factor VII; FA9_BOVIN sp|P00741| Coagulation Factor IX; FA9_SHEEP sp|P16291| Coagulation Factor IX; FA9_RAT sp|P16296| Coagulation Factor IX; EL1_BOVIN sp|Q28153| Elastase 1 Precursor; EL1_RAT sp|P00773| Elastase

1 Precursor; EL2_MOUSE sp|P05208| Elastase 2 Precursor; EL2_RAT sp|P00774| Elastase 2 Precursor; EL2_BOVIN sp|Q29461| Elastase 2 Precursor;

Figure 6.7

The figure shows alignment of serine protease like domain of masquerade like sequences with other proteins involved in patterning in Drosophila and bovine trypsin (5ptp-) and human thrombin (1c1uh). The residues conserved in masquerade like sequences are boxed. The functional residues identified are marked as star. Secondary structures shown according to bovine trypsin (5ptp-). The loop regions are as marked. Loop nomenclature was adopted from Peisach *et al*., (1999).

Figure 6.8

The chitin-binding motifs identified from masquerade and GRAAL gene product of Drosophila and Sp22D protein of Anophilis are shown with those identified from plants and other invertebrates. The residue numbers are shown and repeats are indicated by english uppercase letters (A to E). Proposed chitin-binding residues are boxed and conserved cysteines are marked with star.

Invertebrates are as follows: T. tridentatus tachycitin (Tachycitin), Anopheles gambiae chitinase (Ag-chit), Penaeus japonica chitinase 1 (Pj-chit1), Chelonus sp. chitinase (Ch-chit), 44-kDa glycoprotein from Lucilia cuprina (Peritrophin-44), Trichoplusia ni intestinal mucin (Tn-IM), five repeats of Drsophila masquerade (masquerade A, -B, -C, -D and -E), two repeats of Drosophila Graal gene product (Graal A, -B) and two repeats in Anophilis Sp22D protein (Sp22D A, and -B).

Plants are as follows: hevein from rubber tree (Hevein), Amaranthu caudatus antimicrobial protein, 2 (Ac-AMP2) and four homologous domains of wheat germ agglutinin (WGA A, -B, -C, and -D). Alignments of Proteins other than Masquerade, Graal and Sp22D are taken from Sutake et al., (2000). NMR structures of tachycitin (Sutake et al., 2000) and hevein (Anderson et al., 1993) are known.