Chapter 1

INTRODUCTION
Tuberculosis (TB) is one of the leading infectious diseases and affects roughly one-third of the entire world population (Sander et al., 2004). *Mycobacterium tuberculosis* is the etiological agent of the disease, which most commonly affects the lungs. In healthy people, infection with the pathogen often causes no symptoms, since the person's immune system acts to "wall off" the bacteria. There is a 10% lifetime chance that a latent TB infection will progress to TB disease. The pathogenicity of *M. tuberculosis* results from its potential to multiply to high parasite burdens and its unique capability to modulate the host immune system. There were 5.7 million TB case notifications in 2008. In view of the annual death toll of approximately 2 million individuals due to infection by *M. tuberculosis*, the emergence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) strains of *M. tuberculosis* poses a major threat to human health worldwide (http://www.who.int) TB occurred, of which 0.45-0.62 million deaths were in HIV-positive people globally (WHO, 2009). There were an estimated 0.5 million cases of multidrug resistant TB (MDR-TB) in 2007 (WHO, 2009).

Treatment of this disease is further complicated by the ability of *M. tuberculosis* to persist in the lungs of infected individuals for decades by switching to a dormant or latent phase (Bloom et al., 1999) which also induces tolerance to current antibiotics (Wayne, 1994 & Wallis et al., 1999). About one-third of the world's populations are infected with persistent mycobacterium, providing an enormous potential reservoir for further spread of this disease. Dormancy has been associated with nonreplicating or very slow growth of *M. tuberculosis* that resides in granulomas, a heterogeneous assembly of macrophages, in the lungs of infected individuals. It is generally assumed that the microenvironment in the granulomas is characterized by hypoxia, nutrient starvation, presence of reactive oxygen and nitrogen species (Wayne, 1994; Zhang, 2004 & Fenton et al., 1996)).

The availability of complete genome sequence has enhanced our capability to understand molecular basis of disease and understand the complex host-pathogen interactions. The sequencing of its genome revealed that it is comprised of 4,411,529 base pairs which code for ~4000 genes. *M. tuberculosis* differs radically from other bacteria in that a significant portion of its coding capacity is devoted to the production of enzymes involved in lipogenesis and lipolysis (Cole et al., 1998). Analysis of the sequences have resulted in attributing precise functions to 40% of the predicted proteins, some information is available for 44% and the remaining 16% might account for specific Mycobacterial functions (Cole et
A re-annotation suggests that the genome contains 3995 genes coding for proteins (ORFs) and 50 RNA genes (45 tRNA, 3 rRNA and 2 stable rRNA) (Cole et al., 1998 & Camus et al., 2002).

The complete genome information available on public databases such as Tuberculist (http://genolist.pasteur.fr/TubercuList/) provides access to full range of biosynthetic pathway responsible for the lifecycle and pathogenesis of the bacteria. Coordinated efforts have been launched by several TB Structural Genomics Consortiums (http://www.doe-mbi.ucla.edu/TB) to elucidate the crystal structures of proteins from the pathogen. These efforts are expected to greatly contribute to the identification of novel drug targets and the development of new inhibitors with therapeutic potential aided by rational methods. Structure genomics effort will also have a significance impact in identification of function and characterization of putative proteins.

In the initial part of the present introductory chapter, the reader will be introduced to the haloacid dehalogenase (HAD) superfamily of enzymes. Subsequently the focus will be on the sub-family, phosphoserine phosphatases, which are a main concern of the thesis. A brief introduction to structure based rational techniques is also given here as Chapter 5 deals with the same.

1.1 HAD super family of enzymes

The HAD superfamily, named after the archetypal enzyme haloacid dehalogenase. Current understanding of the haloacid dehalogenase (HAD) superfamily derives largely from studies of prokaryotes, although interest in Eukaryotic HAD proteins appears to be increasing. With the exception of the epoxide hydrolases and the L-2 haloacid dehalogenases, most HAD proteins function as magnesium-dependent phosphohydrolases (Collet et al., 1998). Members of the phospho transferase group include the cation-transporting membrane-bound P-type ATPases, phosphoserine phosphatases, phosphonatases (e.g. phosphonoacetaldehyde hydrolase), phosphomutases (e.g. β-phosphoglucomutase) and protein tyrosine phosphatases.
1.2 Structural variation in the HAD superfamily

1.2.1 Core structure of the HAD superfamily

To provide the basic context for a structure–function analysis of the HAD superfamily it is necessary to first define its essential structural core, and compare it to other structurally related folds. The core catalytic domain of the HAD superfamily contains a three layered α/β sandwich comprised of repeating β-α units which adopt the topology typical of the Rossmannoid class of α/β folds. The central sheet is parallel and is typically comprised of 5 and more strands in a 54123 order (Fig. 1.1). These strands are numbered as S1–S5 and so on. The HAD fold is differentiated from other related Rossmannoid folds by two key structural motifs (Fig. 1.1). First, immediately after strand S1 is a unique, approximately of five to six residues that forms a phosphoaspartyl-intermediate during hydrolysis (Morais et al, 2000, Collet et al, 1997 & Seal et al, 1987). Second, acidic amino acid of this motif functions as a general acid-base in phosphatases and phosphomutases. In some cases it binds while in many cases it protonates the substrate in first step of catalysis and in second step it deprotonates the nucleophile (Lahiri et al, 2002). The rate of aspartyl phosphate hydrolysis is lower in the case of ATPases which is due occurrence of threonine at first position of this motif, which may allow for the time lag necessary for the consequent conformational change. The presence of alanine instead of second aspartate depicts its unique role played by the enamine intermediate as a general acid-base hydrolysis in aspartyl phosphate hydrolysis by phosphonatases.

Threonine or serine was conserved at the end of the second Motif which is corresponds to the strand 2. The Motif III is associated with the conserved Lysine positioned at the end of helix just upstream the strand 4. Motif III and Motif II are important for providing stability to the reaction intermediates of the catalytic reaction. The Lysine present in Motif III is reminiscent of the basic residue generally termed as arginine finger, which stabilize the negative charge on reaction intermediates in most of the phosphohydrolases. So, we can say that they play a similar role also in HAD hydrolases. An analysis of the available structures shows that the lysine in motif III may occur in either of two structural contexts in different HAD hydrolases. In the P-type ATPases, acid phosphatases, phosphoserine phosphatases and the Cof hydrolases the lysine is incorporated into the helix immediately preceding strand 4. However, in all other HAD hydrolases it emerges from the loop immediately prior to the helix. On account of this difference in the secondary structure
context of the lysine, motif III is poorly conserved relative to the other motifs. The poor local
conservation beyond the functionally critical basic residue is also comparable to the regions
bearing the arginine finger in the AAA+ ATPases. Motif IV is present on strand 4 and the
conserved residues are located at the end of this motif. These terminal acidic residues of
motif IV typically exhibit one of three basic signatures: DD, GDxxxD, or GDxxxxD (where x
is any amino acid). These acidic residues along with those in motif I are necessary for
coordinating the Mg ion in the active site (Morais et al., 2000, Baker et al., 1998, Peisach et
Rinaldo-Matthis et al., 2002). All these four Motifs arrange themselves in such a manner to
form a single binding cleft at the C-terminal end of the central strand which typically form the
active site of the HAD superfamily. The binding pocket is partially covered by the β-hairpin
flap present after the first strand (Fig. 1.2). There are some cases, which have additional
covering between two strands of the flap or just after strand. These coverings provide extensive
shielding for catalytic activity. These coverings are also important for providing specificity or
other auxiliary catalytic functions and play very crucial role in the catalytic reactions of HAD
hydrolases. (Baker et al., 1998; Olsen et al., 1988 & Kurihara et al., 1995). These inserts are
termed as caps.

1.2.2 Comparative study of HAD superfamily with other Rossmannoid folds

The folds are classified on the basis of the arrangement of the structural constituents
viz helix and strands. This arrangement of structural constituents is called topology of the
protein. The topology is typical for the particular superfamily and also for family. The
topology of the central β-sheet of the HAD fold makes it a typical representative of the
Rossmannoid class of three-layered α/β sandwich folds. The location of the active site is
centered at the C-terminal end of the central sheet. More specifically, the location of the
substrate binding site in the HAD fold is shared by other Rossmannoid fold enzymes which is
present in the loop between strand 1 and the downstream α-helix, and a second active site
residue positioned immediately downstream of the strand occurring after the crossover in the
β-sheet, i.e. strand 4 (Fig1.1).
Fig 1.1 Rossmannoid domains.
Topology diagrams of domains representative of the major divisions of Rossmann like folds with catalytic acidic residues. Two subtypes of the HAD domain (P-type ATPase HAD and BcbF HAD) that show significant modifications to the classic HAD domain are also shown. Strands are shown as arrows with the arrowhead on the C-terminal end and are labeled from S1 to S6 in the classic HAD. The HAD domain of BcbF is an obligate dimer, and strands from the two dimers are differentiated as A and B (e.g. S1A and S1B). The first strand containing the catalytic D residue is colored in yellow; other core strands conserved across all members of the domain are in blue; non-conserved elements that may have been absent from the ancestral state of a domain are in gray. The HAD C1 cap insertion point is represented as a bright green line and the C2 cap insertion point is represented as an orange line. The pink loop in the HAD domain represents the conserved squiggle.
There are two major divisions in the Rossmannoid fold:

(1) The nucleotide binding domains

These domains are characterized by the presence of the nucleotide-binding loop between strand 1 and the helix positioned downstream to it. This is the classic Rossmann fold named after the Rossmann who first acknowledged it. This group includes many large monophyletic assemblages of proteins. Some of which are given here – the classic Rossmann NAD/FAD-dependent dehydrogenases, the S-AdoMet-binding methyltransferases (Martin et al, 2002), the Sir2-like deacetylase (Zhao et al, 2003), the GTPase FtsZ (Lowe et al, 1998), the HUP superclass (class I tRNA synthetases, HIGH nucleotidyltransferases, photolyase and electron transport flavoprotein). Most of the members of this domain are characterized by the presence of specific signatures in the nucleotide-binding loop.

(2) Phosphohydrolases or divalent cation chelating domain:

This is also a huge division, which is characterized by the presence of a conserved acidic residue in the loop between the strand 1, and the helix just after it. The HAD superfamily enzymes come under this division. The DxD motif is found in this loop. There are several other superfamilies having the same configuration come under this division. Some of these superfamilies are the DHH domain phosphoesterases (e.g. the DNAse involved in repair and recombination, RecJ) (Aravind et al, 1998), the receiver or CheY domain of the two component signaling system (Grebe et al, 1999 & Koretke et al, 2000), the TOPOIM domain, which is the shared catalytic domain of the topoisomerases and DnaG-type primases(Aravind et al, 1998), the classical histone deacetylases/arginases(Finnin et al, 1999), and the von Willebrandt factor A (vWA) domain(Whittaker et al, 2002). The second acidic residue is present at the end of the strand present after the first helix, which occurs after the crossover of the sheet to the opposite side. This acidic residue is also shared by almost all the members of this division (Fig 1.1). Like the HAD domains, the receiver domain forms an aspartyl phosphate intermediate, which receives a phosphate from a histidinyl-phosphate on the histidine kinase. Because of the mechanistic similarity, the receiver domain has previously been claimed to be a member of the HAD fold (Ridder et al, 1999 & Meng et al, 2004). However, a careful examination of the active site organization and sheet topology of the
receiver domains shows that it does not share any of the other specific features conserved throughout the HAD superfamily beyond the phosphorylated aspartate and other generic features of the acidic active-site-containing division of Rossmannoid folds (Fig 1.1). The DHH phosphoesterases contain a DxD signature while the histone deacetylases/argininases contain a DxH signature at the end of strand 1, which help in chelating the metal ion. Except this, these enzymes also contain their own characteristic motifs. These enzymes do not form any phosphoaspartyl intermediate during the course of the catalytic reaction mechanism. The first acidic residue at the end of the strand 1 is always a glutamate in TOPRIM domains of primases and topoisomerases. This glutamate acts as a general acid or base in the hydrolysis of phosphoester bond or polynucleotide transfer. In the case of PIN/5'-3' domains, a catalytic Mg$^{2+}$ activates a water molecule for nucleophilic attack. The acidic residues on the first strand and strand chelate this Mg$^{2+}$ immediately after its cleft (Clissold et al., 2000). First aspartate is part of metal binding motif (DxSxS) called as MIDAS metal binding motif in vWA domain (Whittaker et al., 2002), which is critical for chelation of metal by these domains. It becomes clear from above discussion that different superfamilies of this division of Rossmannoid folds, despite the similarity of the position of acidic residue and metal chelation sites, have different catalytic mechanisms. Like HAD superfamily, TOPRIM, P5N/5'-3' nuclease domains, histone deacetylase/argininase and DHH superfamilies also have large to moderate inserts within the Rossmannoid domain. The presence of these caps suggests that there is some control for the access to active site.

1.2.3 The core domain of the HAD superfamily and its variations

The core Rossmannoid fold of the HAD superfamily is generally conserved in all members except the insertion of the cap modules. The central sheet, however, often shows lateral modifications corresponding to the both ends of the sheet. The ancestral condition of the HAD appears to have been the five-stranded central sheet (Fig 1.1). A major division of HAD superfamily is found to have a additional C-terminal β-α unit after the strand 5 – helix unit (strand6) due to which sandwich get extended (at the left side of the sheet in). Similarly, there are some inserts of additional strands on the right side of the sheet. These inserts arrange themselves parallel to each other in same plane as the core strands and hence extend the sheet. The simplest of these is a β-hairpin, which folds backs and extends the central sheet.
Chapter 1

Hairpin is the defining feature of a large clade within the HAD superfamily. This clade includes the phosphosmannomutases, mannosyl-3-phosphoglycerate phosphatases, sucrose phosphate phosphatases, Cof-type phosphatases and trehalose phosphatases. In P-type ATPases, there is a second independent insert in the right side of the sheet, which form an additional $\alpha$-$\beta$ unit just after strand 3 (Fig 1.1). This additional strand is accommodated in between the strand 2 and strand 3. This is the defining character of P-type ATPases. There is dramatic modification in proteobacterial BcbF family of phosphatases (Fig 1.1). In this case, the helix immediately after the conserved lysine in third motif is replaced by a loop. This loop displaces the strand 4 away from the core sheet and place it in anti-parallel configuration, where it arrange itself with remaining strand (S1- S3) of second monomer in parallel configuration. Thus, the strand 4 appears to be swapped between the two monomers and two identical active sites are formed by a combination of two monomers, one monomer supplying motifs I, II and III and the other monomer-supplying motif IV associated with the swapped strand. This dramatic modification appears to have evolved rather recently through a relatively simple process.

The additional insertions found in ancestral Rossmannoid fold are known as caps. The different caps found in HAD superfamily is discussed below:

1.2.4 Classification of Caps of the HAD superfamily

The most notable inserts seen in the HAD superfamily are the caps which can be classified in three categories:

1. $C_0$ caps: This is the structurally simplest representatives of the HAD superfamily. These caps have only small inserts in either of the two points of cap insertion.

2. $C_1$ caps: are defined as inserts occurring in the middle of the $\beta$-hairpin of the flap motif, and fold into a structural unit distinct from the core domain.

3. $C_2$ caps: are defined as inserts occurring in the linker immediately after strand S3 (Fig 1.1).

Though most members of the HAD superfamily possess either a $C_1$ cap or $C_2$ cap but there are few cases which possess both $C_1$ and $C_2$ caps. The $C_0$ is the simplest state, which has not any additional inserts in the $C_2$ position. These are frequently observed caps in HAD superfamily such as deoxy-D-mannose-octulosonate 8-phosphate (KDP 8-P) phosphatase.
(Fig.1. 2). In the case of polynucleotide phosphatases, there is slightly longer inserts are seen, which have a long loop separating the two β-strands of the flap. In the case of the CTD phosphatases and MDP-1 like phosphatases, this basic condition is elaborated further, with the addition of a strand between the two sheets forming the β-hairpin; resulting in a cap in the form of three-stranded sheet.

The classical C1 caps belong to two distinct structural classes, the α-helical C1 caps and the cap with the unique α+β fold seen in the P-type ATPases (Fig 1.2). The most basic α-helical cap in the form of bi-helical α-hairpin is observed in the acid phosphatase (Fig 1.2). The tetra helical bundle, a form of C1 cap, is found in most of the HAD domains with a cap in this position. This is a next level of complexity in C1 cap. It can be further sub classified into three subcategories on the basis of structural properties and conserved interactions.

- The first subclass, represented by β- phosphoglucomutases and deoxyribonucleotidases, has conserved contacts between the descending arm of the cap domain and the second helix of the Rossmannoid core.
- The second subclass seen in haloacid dehalogenases and their close relatives has conserved contacts involving the loop between the second and third helices of the cap and the linker between strand S3 and the core helix downstream of it.
- The third subclass, typified by the Phosphoserine phosphatase family, shows contacts in the region between the third and fourth α-helices of the C1 cap and a smaller C2 cap that is unique to this family.

The caps in P-type ATPases are formed by the internal duplication of a simple α + β unit and a core sheet formed by a three-stranded β-sheet (Fig 1.2). This indicates that they possibly developed from a single ancestral unit, which in turn could have developed from a precursor similar to Co caps of the MDP1 and CTD phosphatases via the addition of a small α-helical hairpin to the three-stranded sheet. Subsequent duplication of this unit appears to have generated the C1 cap seen in extant P-type ATPases (Fig 1.2). All the known α-helical caps can be conservatively pictured as an evolutionary series of α-helical bundles of increasing complexity emerging through serial duplication from a basic bihelical precursor, along with rapid sequence divergence and reorganization of the helical packing (Fig 1.2).

The C2 caps are found in CoF-like phosphatases and the NagD-like phosphatases and its relatives. These caps can be further subdivided in two groups depending on CoF-like
Fig. 1.2 Topology diagrams of selected C0 and C1 cap HAD domains.
Strands are shown as blue arrows with the arrowhead on the C-terminal side and red coils represent helices. Central to the diagram is the ancestral strand-strand C0 cap. Arrows refer to the likely evolutionary progression leading to the diversification of C1 caps. P-type ATPase cap is colored to denote a possible duplication event. The first unit of the cap is colored in yellow and the second is colored in green. Arrows refer to the likely evolutionary progression leading to the diversification of caps. Broken arrows reflect two possible progression scenarios.
Chapter 1

phosphatases and NagD-like phosphatases (Fig 1.2). Except these two major groups, there is a small, unique C2 cap found in the histidinol phosphatase family.

There is a remarkable diversification of C2 cap through addition of secondary structure elements to the basic unit with a three-stranded anti-parallel β-sheet in Cof-type phosphatases (Fig 1.2). Trehalose 6 phosphatase orthologs shows a more complex form of C2 caps where a strand is added to the sheet at the N terminus. In YedP, there are entire β-α units, instead of a single strand, added to the N terminus of the ancestral unit. In the Tm0651 from Thermotoga maritima (Shin et al, 2003), this trend is further-extended via the addition of three α-β units to the ancestral unit. The elaboration of this cap was observed in YwpJ from Bacillus subtilis by the duplication of a helix in one of the α-β units. Thus, it appears that the C2 caps of the Cof-type phosphatases evolved through a process of serial addition of simple secondary structure units, most probably through duplications limited to the N-terminal region of the cap.

The C2 cap of the NagD-like phosphatases might be arises from simple duplication of the core domain. It includes an α/ β domain with a core four-stranded parallel β-sheet, with an additional N-terminal antiparallel strand. This configuration is different from from all other known domains. However, at the sequence level there is no significant similarity with the core domain. This group of C2 caps also contains a unique β-hairpin inserted after the third strand. An examination of the sequence of the C2 caps of the histidinol phosphatase family reveals a conserved CxHx(6-13) CxC signature (where x is any amino acid). This suggests that this C2 cap is stabilized through the chelation of a divalent metal ion, and is likely to assume a simple flap-like structure (Fig 1.2).

Several members of the HAD superfamily simultaneously possess both C1 and C2 caps, both of which may be similarly sized, or one of them may be the dominant cap. Small β-hairpin C2 caps are seen in the phosphoserine phosphatase and the pyrimidine 5-nucleotidase families, which also contain helical C1 caps (Fig 1.2). In an archaeal sub-family of the phosphoserine phosphatases, typified by the protein AF1437, a small C2 cap assuming the form of a tri-helical bundle is seen, suggesting that there have been multiple independent innovations of such smaller C2 caps. In all these families the C1 cap is clearly the dominant cap with the C2 cap.
1.2.5 Importance of caps in catalytic mechanism of the HAD superfamily

There are many studies on the mechanism of HAD superfamily enzymes. All enzymes having C1 cap are likely to follow the same catalytic cycle (Allen et al, 2004, Morais et al, 2000, Lahiri et al, 2002, Wang et al, 2001). The enzyme in the "open" configuration allows the substrate (typically a phosphoester) to enter the active site. Once the substrate is bound the enzyme assumes the "closed" configuration and the Mg$^{2+}$ in the active site interacts with the negatively charged phosphate, preparing it for nucleophilic attack by the first conserved aspartate of Motif I, which is situated at the end of strand 1. Due to this an acyl phosphate intermediate is formed. As the acyl intermediate is formed, the enzyme enters the open configuration again. This open state of the enzyme allows solvent to enter in bulk quantity in the active site and a water is deprotonated by the second aspartate of strand one; hydrolyzing the acyl phosphate intermediate and returning the enzyme to the native state (Lahiri et al, 2002). There is a slight variation in haloacid dehalogenases, which release a halide ion along with the formation of a regular ester linkage (Ridder et al, 1999). In phosphonatases, the initial stage is different to that of classical reaction cycle while in sugar phosphate mutases, the final stage is different. (Morais et al, 2000, Lahiri et al, 2003) but the core phosphoryl transfer mechanism remains the same.

Key aspects of the HAD catalytic mechanism that emerged from these studies are:

1. The alternation between open and close states
2. A preliminary reaction favored by solvent exclusion and a subsequent step favored by extensive solvent contact.

The squiggle and flap of the core domain are the structural features, which are responsible for this process. The squiggle, close to a helical structure, has ability to wound tightly or loosely alternatively. This differential function of squiggle in turn induces a movement in the flap immediately juxtaposed to the active site and alternatively results in the closed and open states. These squiggle and flap are conserved across the HAD superfamily and though the essential structural features of this superfamily. The emergence of cap modules makes the solvent exclusion and acyl phosphate formation more efficient. Besides helping in catalytic mechanism, the emergence of caps provided them a mean for substrate recognition. The recognition of substrate is provided by several additional interactions, which are not possible in ancestral active site (Baker et al, 1998, Kurihara et al, 1995). The simplest
structures add the cap to the flap motif itself, so as to completely seal the active site in the closed state. Thus, the flap region was a hotspot for the insertion of the various C1 caps, which appears to suggest intense natural selection for efficient solvent exclusion. Though C2 caps do not have the mobility like the C1 caps but squiggle-flap elements exhibits drastic movements similar to C1 caps. As a result there would be an open state in which the substrate, solvent and leaving group can be exchanged with the active site cavity and a closed state where the flap occludes the cavity formed by the C2 cap completely and excludes the solvent. In most cases where both C1 and C2 caps are present such as the Phosphoserine phosphatase family, the C1 cap is responsible for closing the active site while that of C2 cap for occlusion. Co caps form crater like structures associated with active sites (e.g. MDP-1 and the CTD phosphatase families). These crater-like accesses to the active site of the C0 cap enzymes are unlikely to completely occlude the solvent, but their substrates are large molecules (proteins and Oligonucleotide), which may block the rest of the active site from solvent while being bound to it.

1.3 Classification of the HAD Super family

This is a huge superfamily, which, on the structural basis, can be divided into following families:

- **Probable Phosphatase YrbI**: The insertion subdomain is a beta-hairpin involved into tetramerisation
- **Beta-phosphoglucomutase like**: The insertion subdomain is a 4-helical bundle.
- **Phosphonoacetaldehyde hydrolase like**: the insertion subdomain is a 4-helical bundle.
- **Phosphoserine phosphatase**: Here still the insertion subdomain is a 4-helical bundle.
- **Homoserine Kinase**: the insertion subdomain is a rudiment 4-helical bundle.
- **Acid Phosphatase**: the insertion subdomain is a helical hairpin.
- **Phosphatase domain of polynucleotide kinase**: This is unique in HAD superfamily because it have no insertion subdomain.
- **Cof family**: contains an alpha+beta subdomain inserted into a new site after strand 3.
- **NagD-like**: It consists of two segment-swapped domains of this fold; this results in the insertion of a circularly permuted domain after strand 3, analogously to the Cof family.
Trehalose phosphatase: contains an insert alpha+beta subdomain; similar overall fold to the Cof family.

Magnesium dependent phosphatase-1, Mdp1: the insertion subdomain is a 3-stranded beta-sheet different from the NIF family.

Histidinol phosphatase: there is not any insertion subdomain but is unique.

MtnX-like: The insertion subdomain is rudiment 4-helical bundle.

5'-nucleotidase like: the insertion domain consists of 3-helical bundle and a pseudo beta-barrel; contains extra C-terminal long alpha hairpin subdomain.

Meta-cation ATPase, catalytic domain P: Interrupted by a large insertion domain N.

P-type ATPase: The insertion domain is 4-helix bundle in duplicate.

As we saw above HAD is huge superfamily. The Phosphoserine phosphatase is discussed in detail because we mainly concerned with it.

1.3.1 Phosphoserine phosphatase SerB (EC 3.1.3.3)

The systematic name of this enzyme class is O-phosphoserine phosphohydrolase. Hydrolases are the enzymes that catalyze the hydrolysis of various bonds. Some of these enzymes pose problems because they have a very wide specificity. While the systematic name always includes 'hydrolase', the common name is, in most cases, formed by the name of the substrate with the suffix -ase. It is understood that the name of the substrate with this suffix, and no other indicator, means a hydrolytic enzyme. Phosphoserine phosphatase is a hydrolase working on the phosphoric monoester bond. The monomer of the enzyme consists of two domains: a core $\alpha/\beta$ domain and a four-helix-bundle domain. The core $\alpha/\beta$ structure has a parallel $\beta$-pleated sheet with 7 strands ordered as $S_3S_2S_1S_5S_6S_7$. This domain resembles an NAD(P) binding Rossman fold (Rossmann et. al, 1974) The importance of phosphoserine phosphatase is underscored by its presence in mammalian brains and in membranes in lower organisms.

The enzyme formally catalyzes the chemical reaction:

$$O\text{-phospho-L (or D)-serine + H}_2\text{O} \rightleftharpoons L \text{(or D)-serine + phosphate}$$
which is the last step in the biosynthesis of L-serine from carbohydrates.

The 3 substrates of this enzyme are O-phospho-L-serine, O-phospho-D-serine, and H$_2$O, while its 3 products are L-serine, D-serine, and phosphate. Among other possible functions, the enzyme participates in glycine, serine and threonine metabolism.

A brief reaction mechanism cycle of phosphoserine phosphatases is given below:

![Reaction Mechanism](image)

**Fig. 1.3 General Scheme of the reaction cycle of Phosphoserine phosphatase (Wang et al, 2002).**

- (II) L-Phosphoserine binds to the active site, presenting its phosphate group to Asp185.
- (III) Transition state with nucleophilic attack of Asp185.
- (IV) Covalent phospho-aspartyl enzyme intermediate.
- (V) Transition state with a nucleophilic attack of a water molecule causing the dephosphorylation of Asp185.
- (VI) Phosphate noncovalently bound in the active site with Mg$^{2+}$. 

16
Some functions of this enzyme are summarised below:

1.3.1.1 Phosphoserine phosphatase in the mammalian brain

L-Serine has for several years been recognized as having a role in cellular proliferation, being a precursor for nucleotide synthesis (Snell 1984). However, after the first report of patients with serine deficiency disorders it became clear that L-serine and L-serine-derived metabolites are important for brain development and function. L-serine has been shown to be a requisite growth factor for neurons (Savoca et al., 1995; Furuya et al., 2000), while D-serine is a co-agonist at the N-methyl-D-aspartate subtype of glutamate-gated ion channels (Mothet et al., 2000; Shleper et al., 2005), and may also act as a ligand at the GluRδ2 glutamate-like receptor (Naur et al., 2007). In mammalian cells L-serine is synthesized via the “phosphorylated pathway” which encompasses three enzymatic steps. In the first step, 3-phosphoglycerate derived from glycolysis is metabolized to phosphohydroxypyruvate by 3-phosphoglycerate dehydrogenase (de Koning et al., 2003). In the second step phosphohydroxypyruvate is converted into L-serine-O-phosphate (L-SOP), also known as L-phosphoserine, by the vitamin B6-dependent enzyme phosphoserine aminotransferase (PSAT; Hester et al., 1999). L-SOP is then dephosphorylated by Phosphoserine phosphatase (PSP; Collet et al., 1999) to produce L-serine, which is subsequently metabolized to glycine by the enzyme serine hydroxymethyltransferase, or isomerized to D-serine by serine racemase.

1.3.1.2 Phosphoserine phosphatase in neural stem cell niche

Neural stem cells are multipotent cells of the embryonic and adult brain that have the potential to produce multiple lineages of mature neural cells. Neural stem cells are found in the embryonic brain, as well as in the ventricular zone and in the subventricular zone of the lateral ventricle and in the subgranular zone of the dentate gyrus in the adult brain. L-Serine is an amino acid necessary for protein and nucleotide synthesis, as well as potentially playing a role directly in regulating neuronal differentiation. D-Serine, Glycine, and L-phosphoserine have all been shown to bind as cofactors or ligands to various extracellular receptors. Phosphoserine phosphatase is expressed in the neural stem cell niche and may regulate proliferation of neural stem cells (Nakano et al., 2007).
1.3.1.3 Phosphoserine phosphatase in pathogenesis

The role of Phosphoserine is well understood in the central nervous system, stem cell niche and nuclear proliferation but nothing known about its involvement in pathogenesis. The study of the Porphyromonas gingivalis Phosphoserine phosphatase establishes its role in the invasion of gingival epithelial cells (Tribble et al, 2006). They used biochemical analysis to confirm this enzyme as a member of the HAD family of phosphatases. Study of allelic exchange mutants in antibiotic protection assays indicated that the SerB653, but not other HAD phosphatases or metabolic enzymes, was required for maximum invasion efficiency. Furthermore, both internalization and intracellular survival were affected by loss of this enzyme. Screens of SerB653 interactions with gingival epithelial cell extracts showed that the phosphatase may exert its effect on invasion through components of membrane vesicular transport systems, including GAPDH and microtubules. SerB653 is a previously uncharacterized HAD family phosphatase that is exploited by a prokaryote to facilitate an intracellular lifestyle (Tribble et al, 2006).

1.3.1.4 Phosphoserine phosphatase in Cytoskeleton remodeling

In 2008, Lamont establishes the involvement of Phosphoserine phosphatase in cytoskeleton remodeling. Phosphoserine phosphatase is a multifunctional enzyme, it expresses during the invasion and also involve in modulation of cytoskeleton in P. gingivalis. In addition to a role in intracellular invasion through control of microtubule dynamics, secreted SerB acts on epithelial surfaces to impact actin cytoskeletal structure and cytokine production. These properties are associated with broadly based regulation of the epithelial cell transcriptome.

1.3.1.5 Phosphoserine phosphatase in CNS

The metabolite of L-SOP, L-serine is principally synthesized in glial cells (Savoca et al, 1995; Verleysdonk et al, 2000). However, Hampson (2009) found that PSAT and PSP are expressed in both neurons and in glia. The lack of induction of PSAT and PSP after kainic acid induced seizures and gliosis indicated a primarily neuronal localization in the hippocampus. A sufficient supply of L-serine is crucial for the normal development of the
CNS. This is dramatically illustrated in children with mutations that cripple the enzymatic activity of 3-phosphoglycerate dehydrogenase, PSAT, or PSP, where severe neurological defects become apparent within several months after birth (de Koning et al, 2003), and also in 3-phosphoglycerate dehydrogenase knockout mice, which die in utero (Yoshida et al, 2004). PSP is expressed more ubiquitously in the immature rat brain compared to the adult CNS (Hampson et al, 2009). Cerebellum displayed a convergence in Purkinje neurons where PSP displayed relatively high expression.

1.4 Catalytic mechanism of Phosphoserine phosphatase

A comparison of the probable reaction mechanism of Phosphoserine phosphatase, P-type ATPases with that of HAD indicates notable similarities. Probing of the active site of HAD with hydroxylamine suggests that the reaction proceeds by esterification of first D of motif I followed by ester hydrolysis releasing the alkanoic acid (Liu et al, 1997). This invariant aspartate probably performs a similar role in the other members of the superfamily. Consistent with this, this residue is the autophosphorylation site in P-type ATPases (Asano et al, 1996). The aspartate carboxyl oxygen could initiate a nucleophilic attack on the electron-rich γ-phosphate bond, resulting in phosphoester formation. The phosphoester is subsequently hydrolyzed and the free energy of hydrolysis is utilized for ion transport. The other phosphatases of the HAD superfamily, for example phosphoserine phosphatase, are likely to have the same reaction mechanism, and consistent with this, P-type ATPases behave like typical phosphatases in their use of acetyl phosphate as a substrate. The role of the other two motifs in catalysis is less clear, although evidence of their direct involvement is available. Based on its spatial position it is proposed that the conserved lysine of motif III is involved in the stabilization of the excess of negative charge in the substrate /reaction intermediates. The two conserved aspartates in motif III might form a system to direct water for the hydrolysis of the acyl-phosphate intermediate.

In summary, Phosphoserine phosphatases have a wide range of functions. In mammals, one of the functions is in the development of the brain while in lower organisms like *P. gingivalis* it helps in invasion of epithelial cells. Their roles in other organisms like *M. tuberculosis* needs to be studied in detail.
1.5 Structure Based Inhibitor Design

Drug discovery and development is an intense, lengthy and an interdisciplinary process. Drug discovery is mostly portrayed as a linear, consecutive process that starts with target and lead discovery, followed by lead optimization and pre-clinical *in vitro* and *in vivo* studies to determine whether such compounds satisfy a number of pre-set criteria for initiating clinical development. For the pharmaceutical industry, the number of years to bring a drug from discovery to market is approximately 12-14 years and costing up to $1.2 - $1.4 billion dollars. Traditionally, drugs were discovered by synthesizing compounds in a time-consuming multi-step processes against a battery of *in vivo* biological screens and further investigating the promising candidates for their pharmacokinetic properties, metabolism and potential toxicity. Such a development process has resulted in high attrition rates with failures attributed to poor pharmacokinetics (39%), lack of efficacy (30%), animal toxicity (11%), adverse effects in humans (10%) and various commercial and miscellaneous factors. Today, the process of drug discovery has been revolutionized with the advent of genomics, proteomics, bioinformatics and efficient technologies like, combinatorial chemistry, high throughput screening (HTS), virtual screening, *de novo* design, *in vitro*, *in silico* ADMET screening and structure-based drug design.

There are two major types of drug design.

- Ligand-based drug design.
- Structure-based drug design.

Ligand-based drug design (or indirect drug design) relies on knowledge of other molecules that bind to the biological target of interest. These other molecules may be used to derive a pharmacophor, which defines the minimum necessary structural characteristics a molecule must possess in order to bind to the target (Robertson *et al*, 2007). In other words, a model of the biological target may be built based on the knowledge of what binds to it and this model in turn may be used to design new molecular entities that interact with the target.

Structure-based drug design (or direct drug design) relies on knowledge of the three dimensional structure of the biological target obtained through methods such as x-ray crystallography or NMR spectroscopy (Rajamani *et al*, 2007). If an experimental structure of a target is not available, it may be possible to create a homology model of the target based on
the experimental structure of a related protein. Current methods for structure-based drug design can be divided roughly into two categories.

1.5.1 Ligand based drug design

This involves “finding” ligands for a given receptor, which is usually referred as database searching. A large number of potential ligand molecules are screened to find those fitting the binding pocket of the receptor. The key advantage of database searching is that it saves synthetic effort to obtain new lead compounds.

![Flowchart](image)

**Fig 1.4** Schematic representation of strategy for Ligand based drug design.

1.5.2 Structure-based drug design

This is about “building” ligands, which is usually referred as receptor-based drug design. Here, ligand molecules are built up within the constraints of the binding pocket by assembling small pieces in a stepwise manner. These pieces can be either individual atoms or
molecular fragments. The key advantage of such a method is that novel structures, not contained in any database, can be suggested. These techniques are raising much excitement to the drug design community.

1.5.3 Scoring Functions

There are some scoring functions, which are used in drug design and molecular modeling. Scoring functions are fast approximate mathematical methods used to predict the strength of the non-covalent interaction (also referred to as binding affinity) between two molecules after they have been docked (Jain et al., 2006). There are three general classes of scoring functions:

1.5.3.1 Force Field

The affinities between ligand and receptor are estimated by summing up van der waals, electrostatic interactions, intramolecular energies and desolvation energies of ligand and protein.

![Fig 1.5 Schematic representation of strategy for Structure based drug design.](image)
1.5.3.2 Empirical

This is based on counting of intermolecular interactions between two molecules or by calculating the change in solvent accessible surface area before and after binding. These interaction includes hydrophobic - hydrophobic, hydrophobic - hydrophilic, hydrophobic unfavorable, hydrophilic - hydrophilic and number of rotatable bonds (Böhm HJ et. al, 1998).

1.5.3.3 Knowledge based

This is based on statistical observations intermolecular close contacts in large structural databases such as Cambridge structural database and Protein-Data Bank.

A general-purpose empirical function in order to define binding energy was given by Böhm was-

$$\Delta G_{binding} = -RT \ln K_d$$

(Eq. 1.1)

Where,

$$K_d = \frac{[\text{Receptor}][\text{Acceptor}]}{[\text{Complex}]}$$

$$\Delta G_{binding} = \Delta G_{motion} + \Delta G_{interaction} + \Delta G_{desolvation} + \Delta G_{configuration}$$

1.5.4 Utility of structure-guided methods

There are more than 40 compounds were discovered with the aid of structure-guided methods that have entered clinical trials. These drugs and drug candidates are directed against two-dozen different molecular targets, in a wide range of therapeutic areas, although over half of the compounds are used for oncology or viral infections. There may even be additional compounds that we have overlooked in our survey of information in the public domain. Some of these compounds have become approved and marketed drugs. Captopril was the first drug whose discovery relied upon the explicit use of X-ray structural information as a guide for small molecule design. This occurred over two decades ago. Although the template used for discovery of captopril was a homology model (based on the X-ray structure of bovine carboxypeptidase A), the recognition of the 'heuristic' value of such a model was a landmark. A similar conceptual model of acetylcholinesterase (based on the experimental X-ray
structure of chymotrypsin) was used for the development of the agent zanapezil. Two of the drugs (captopril and dorzolamide) have now become generic products. The largest fraction of approved drugs developed by structure-guided methods is the set of HIV protease inhibitors (Clercq et al, 2007), which were pivotal during the late 1990s in converting HIV infection from a death sentence into a chronic disease. Obviously, structure guided inhibitor design has become an invaluable tool in the accelerated discovery of new therapeutics.