Structural studies of three cell signaling proteins

Crystal structures of EphB1 kinase domain, PTPA, and YegS

M. Amin Bakali H.

PhD thesis
Department of biochemistry and biophysics
Stockholm University, Sweden 2007
Doctoral Thesis 2007
Department of Biochemistry and Biophysics
The Arrhenius Laboratories for Natural Sciences
Stockholm University
SE-106 91 Stockholm
Sweden

All previously published papers are reprinted
with permission from the publisher.

© 2007 M-Amin Bakali H.
ISBN 91-7155-364-9, pp. 1-74
To my family especially my parents

A voyage through time and the PN-lab
Kinases and phosphatases are key regulatory proteins in the cell. The disruption of their activities leads ultimately to the abolishment of the homeostasis of the cell, and is frequently correlated with cancer. EphB1 is a member of the largest family of receptor tyrosine kinases. It is associated with neurogenesis, angiogenesis, and cancer. The cytosolic part of the human EphB1 receptor is composed of two domains. Successful generation of soluble constructs, using a novel random construct screening approach, led to the structure determination of the kinase domain of this receptor. The native structure and the complex structure with an ATP analogue revealed novel features in the regulation of the Eph family of kinases.

The structure of PTPA, an activator of protein phosphatase 2 A, a tumor suppressor and a key phosphatase in the cell was solved. The structure revealed a novel fold containing a conserved cleft predicted to be involved in interaction with PP2A.

Finally, the structure of YegS, an *Escherichia coli* protein annotated as a putative diacylglycerol kinase, has been determined. Beside the elucidation of its atomic structure, a phosphatidylglycerol (PG) kinase activity, never seen before, has been assigned to YegS based on biochemical studies. The YegS structure shows resemblance to the fold previously seen in NAD kinases. The structure also revealed the existence of a novel metal site that could potentially play a regulatory role. The YegS structure has important implications for understanding related proteins in pathogenic organisms and is the first homologue of a human lipid kinase for which the structure has been elucidated.
Table of content

Abstract 4
Table of content 5
List of publications 7
Additional publications 8
Abbreviations 9
General introduction 10
DNA, genomics and "omics", to structure hunters 12
Structural genomics 13
Protein production 14
Protein production; now and then 14
Protein production problems 15
Suggested solutions for reluctant targets 15
An appealing approach 16
Screening for soluble proteins 17
Solubility screen by FiDo 18
Deletion library screening for soluble proteins (Paper I) 20
Dephosphorylation of EphB1 construct 23
Protein kinases, RTK, and Eph family 25
The catalytic domain 25
Receptor tyrosine kinases 26
Eph receptors, the largest family of RTK 27
Eph receptors and cancer 28
Unique features of the Eph receptor 28
Previous structural work on the Eph family 29
Motivation of the Eph study 29
Human EphB1 kinase domain structure (Paper II) 30
The overall structure of the kinase domain of EphB1 30
The conformational change of Tyr740 31
The αC helix 31
The activation loop 32
More on the activation mechanism 34
PTPA a regulator of PP2A 35
PP2A 35
PTPA 37
PTPA structure (Paper III) 37
List of publications

The following publications are the basis for this thesis, and will be referred to by their Roman numerals in the text.

I. Rosemarie K.C. Knaust, M. Amin Bakali H., and Pär Nordlund
Expression and phosphorylation studies of the cytosolic domain of the EphB1 receptor.
*Manuscript*

II. M. Amin Bakali H., Rosemarie K.C. Knaust, and Pär Nordlund
Crystal structure of the kinase domain of the human EphB1.
*Manuscript*

III. Magnusdottir A, Stenmark P, Flodin S, Nyman T, Hammarstrom M, Ehn M., Bakali H MA, Berglund H, and Nordlund P.

IV. Bakali M. A., Nordlund P, and Hallberg BM.
Expression, purification, crystallization and preliminary diffraction studies of the mammalian DAG kinase homologue YegS from Escherichia coli.

V. M. Amin Bakali H., Maria Dolores Herman, Kenneth A. Johnson, Amélie A. Kelly, Åke Wieslander, B. Martin Hallberg, and Pär Nordlund
Crystal structure of YegS a homologue to the mammalian diacylglycerol kinases reveals a novel regulatory metal binding site. *
Submitted

*) Presented the content of this paper at the SBnet conference 2006 in the hot stuff session.
Additional publications

1. Martin Högbom, Ulrika B. Ericsson, Robert Lam, M. Amin Bakali H., Ekaterina Kuznetsova, Pär Nordlund and Deborah B. Zamble.

C1-Tetrahydrofolate synthase stabilizes HIF-1a by disrupting the interaction with the von Hippel Lindau tumor suppressor gene product.
Manuscript.
## Abbreviation

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP-PNP</td>
<td>Adenosine 5’-(β,γ-imido) triphosphate</td>
</tr>
<tr>
<td>CAT</td>
<td>Chloramphenicol acyltransferase protein</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>C-terminus</td>
<td>Carboxy terminus</td>
</tr>
<tr>
<td>DAG</td>
<td>Diacylglycerol</td>
</tr>
<tr>
<td>DGK</td>
<td>Diacylglycerol kinase</td>
</tr>
<tr>
<td>Eph</td>
<td>Erythropoitin producing hepatocellular</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
</tr>
<tr>
<td>His-tag</td>
<td>Polyhistidine tag</td>
</tr>
<tr>
<td>HTP</td>
<td>High throughput</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>N-terminus</td>
<td>Amino terminus</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl-β-D-thiogalactopyranoside</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>PG</td>
<td>Phosphatidylglycerol</td>
</tr>
<tr>
<td>PGP</td>
<td>Phosphatidylglycerol phosphate</td>
</tr>
<tr>
<td>PTPA</td>
<td>Protein phosphatase two A phosphatase activator</td>
</tr>
<tr>
<td>SAM</td>
<td>Sterile Alpha motif</td>
</tr>
<tr>
<td>SGs</td>
<td>Structural genomics</td>
</tr>
<tr>
<td>RTK</td>
<td>Receptor tyrosine kinase</td>
</tr>
</tbody>
</table>
General introduction

Before diving into the details contained within this thesis I would like to start by placing things in a wider context. To start with, a cell is the fundamental unit of life, and it is the smallest functional unit in all living organisms. Cells differ in size, shape, and the role they play. The human body for instance is composed of roughly $10^{14}$ cells and approximately ten times that number of bacterial cells inhabit different parts of the human body such as the gastrointestinal tract and skin. If we stick to the human body as the basis of our discussion, we can as an illustrative example take the difference between a liver cell and an insulin producing beta cell. What dictates the differences between these two different cell types? The two cells have the potential to encode exactly the same set of proteins, since they have exactly the same gene sequences, and thereby the potential to produce exactly the same set of proteins. However it is the timing and the amount of a certain protein or set of proteins that are produced and degraded that make the difference. This is most pronounced in the red blood cell that is involved in oxygen transport. Red blood cells contain approximately 80% hemoglobin in the cytosol. Proteins are involved in virtually every process in the cell whether it is synthesizing the building blocks of the cell, maintaining its integrity or degrading it. Of the different classes of proteins, enzymes are an interesting group. Enzymes catalyze, i.e. speed up, chemical reactions in a controlled and efficient way, so that the reaction can occur under physiological conditions. Other groups of proteins play a structural role such as proteins in the cytoskeleton, others are involved in immune response, cell cycle etc. In short there is no process in the cell that proteins are not part off either directly or indirectly.

Proteins are biological macromolecules made up of a set of 20 basic building blocks called amino acids. The amino acids are arranged in a specific linear order called the primary sequence of the protein. The content and the order of the amino acids in a certain protein are defined by the gene that encodes the protein. The primary sequence is ultimately what dictates the three-dimensional structure of the protein. The three-dimensional structure of a macromolecule is valuable for understanding its function; the most famous example of this principle must be the structure of the DNA double helix [1]. The structure led James Watson and Francis Crick to express in a powerful sentence: “it has not escaped our notice that the specific pairing
we have postulated immediately suggests a possible copying mechanism for the genetic material”. Structural information of a protein is of tremendous value not only for understanding its function but also in e.g. drug development. The knowledge about the structure of a protein can be used as a tool in designing an inhibitor to cure a specific disease. This process, structure-based drug design [2] is useful for a wide range of potential therapeutics such as the antibacterial, antiviral, and anticancer research. Structure-based drug design is used to combat the AIDS epidemic, by targeting HIV-1 reverse transcriptase [3], and to improved HIV-1 protease inhibitors to circumvent drug resistance are good examples [4]. Determination of protein three-dimensional structures can be done by different methods such as X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy (EM). X-ray crystallography is a frequently used method; however, it requires the protein to be in a crystalline form. One of the first steps in the process is the production of large amount of a highly pure protein. A significant part of the work presented here deals with the protein production process. In an attempt to overcome the rate-limiting step in the production of suitable soluble proteins, our laboratory has undertaken method development. Two methods were developed which have been important for my projects. The first made screening for soluble proteins manageable by employing filtration instead of centrifugation to separate soluble proteins from inclusion bodies. The solubility screen method was suited for high throughput (HTP) screening. The second contribution is the deletion screening method. The method includes generation of protein constructs with varying lengths and different solubility profiles. The first structure reported here was produced by combining the two methods, and thus proved the usefulness of these approaches.

The structural studies in the present work aims at deriving structure-function understanding of the protein studied. The proteins are of biomedical importance when they are either potential human drug target proteins, or related to medically relevant human proteins with no reported structure. The first two structures reported here, EphB1 and PTPA are of the first kind, while YegS was chosen because it is homologous to the human diacylglycerol kinases. An introduction to the field and to the different proteins investigated here will be given along with detailed discussion of the insights contributed by this work.
DNA, genomics and "Omics" to structure hunters

Realizing that the DNA is the genetic material [5], and later the elucidation of the DNA double helix structure [1] led to a huge advancement of science. Later on the central dogma was proposed [6]. The essence of the central dogma is that DNA is responsible for coding for a protein and that the information is transferred by mediator molecules named messenger RNA (mRNA). Many more interesting discoveries were to follow as an inevitable consequence of this insight. The DNA replication machinery was revealed, the genetic code was cracked and numerous exciting events were to follow. In parallel to the increased basic understanding of how the cell functions, a series of technological advancements were made in parallel such as protein [7, 8] and DNA sequencing [7, 8], and the polymerase chain reaction (PCR) [9] to mention just a few. The knowledge about the central role played by DNA, and the availability of sequencing tools led to the proposition of sequencing whole genomes. The first genome to be sequenced was that of the 5375 nucleotide bacteriophage phi X174 [10]. This eventually gave birth to the field of genomics meaning the study of the DNA sequence in whole genomes. The genome of "Haemophilus influenzae" was reported 1995 as the first bacterial genome [11]. Two years later the genome of the yeast "Saccharomyces cerevisiae" was completed representing the first eukaryote to have its genome sequenced [12]. The success in sequencing genomes eventually culminated in sequencing the human genome [13-15]. In 1987 a journal named Genomics was inaugurated by an editorial titled “Genomics: A new Discipline, a new name, a new Journal” celebrating the new era of genome exploration. However, its should be mentioned that at the outset genomics as a discipline, and the enterprise that aimed to sequence the human genome were heavily criticized [16]. Currently, there is complete sequence information for more than 800 genomes and the ongoing world wide sequencing will result in an exponential increase. Today many scientific fields include “omics” in their names such as proteomics, and oncogenomics. A new field named with an “omics” is the exciting field of structural genomics (SGs) [17]. Inspired by the success of the genomic projects, and shifting focus from sequences to the three-dimensional structure of the proteins they encode instead. World wide efforts to solve the structures of every possible protein were initiated, and in September 2000 the Protein Structure Initiative (PSI) [18] was launched providing financial support and setting strategies to advance the progress. To illustrate the progress, and the maturation that structural genomics has
reached, many laboratories are now targeting not only soluble proteins, but also consider the notorious membrane proteins. The structural genomics on soluble proteins are now entering into “production mode” and projects on the challenging membrane proteins are starting to report progress [19, 20]. The framework for the thesis you are now reading is the structural genomics project initiated at Stockholm University and later at the Karolinska Institute. Arthur Kornberg in his autobiography described scientists as vitamin hunters, enzyme hunters, and gene hunters to describe the different epochs in biological science [21]. In accordance with Kornberg’s analogy I am tempted to extend this to describe the emerging era of structural genomics as the one of structure hunters.

**Structural genomics**

Structural genomics projects were launched with the ambition to establish technologies, that allow structures to be solved at a minimal cost. Eventually the hope was to deliver structures of representative members of most structural families as well as allow key biomedical drug target families and pathways to be structurally characterized en masse [17, 22]. At present, a number of structure production centers have been established around the world and it is clear that structural genomics will give important contributions to both basic and applied biomedical research [23]. The determination of structures will help to annotate newly identified genes by structure-based assignment of molecular function [24, 25] and some structures will be highly valuable for rational drug design and other “traditional” projects. Solving a crystal structure has been a tedious and costly process. The cost for a protein structure determination in current structural genomics initiatives is estimated to range from $50,000 to $200,000 per novel structure, not including membrane proteins [26]. The cost of determining a structure in a traditional structural biology project is estimated to be significantly higher than that. Streamlined parallel and partially automated methods have emerged as an important component that improve speed, and minimize costs for structure determination [27]. Such high-throughput (HTP) methods are currently being developed and improved in structural genomics initiatives all over the world. Robust strategies for parallel cloning, automated nano-crystallization and imaging have reached a high level of sophistication, and are now integrated into the gene-to-structure process of most structural genomics centers. The availability of cDNA libraries from different organisms at a minimal cost and the different
HTP cloning methods such as LIC (ligation independent cloning) and GATEWAY™ cloning [28] are major contributors. Producing soluble proteins in sufficient amounts and quality for successful crystallization is however still the major bottle-neck, and the need for screening different constructs and conditions for expression and purification is widely acknowledged [29]. In well-established structural genomics projects, two-thirds of the investments are into protein production and protein chemistry.

In the initial phase of structural genomics, strategies were set up to allow large numbers of proteins to be processed and allow the most accessible to be structurally characterized, i.e. “the low hanging fruits”. This strategy has proved itself to be especially useful for prokaryotic proteins. A recent assessment of the impact of structural genomics projects states that the cost of solving a structure in the most efficient SGs center in the US has dropped to one quarter of the estimated cost of solving it in the traditional approach [30]. Furthermore the future direction of SGs will be focused on challenging proteins, such as membrane proteins, large eukaryotic proteins and complexes. Therefore considerable resources are planned for specialized centers aimed at developing technologies for the high-throughput solution of challenging targets [30].

**Protein production**

**Protein production; now and then**

In the early days of structural biology, the proteins to be crystallized were obtained from a source rich in a certain protein. For instance myoglobin, the first protein that had its three-dimensional structure elucidated by x-ray crystallography, was obtained from the red dark tissue of diving animals such as whales [31]. Isolating proteins directly from tissues has been a common practice in biochemistry, but had great limitations, since it requires a high abundance of the protein in a certain cell type or tissue/organism. Alternatively, for proteins of lower abundance, large quantities of the cell or tissue/organism were required, but the purification of the protein remained challenging. For example 1000 pig brains were needed for the isolation of 200 μg porcine neuropeptide Y2 receptor [32]. The majority of proteins are in fact only present at very low levels in the cell, making their isolation in useful amounts impractical. Traditionally, protein crystallographers have not been very concerned with protein production when they obtained purified protein.
samples from collaborating biochemists. Recombinant DNA technology [33] allowed proteins to be overexpressed in recombinant systems, which revolutionized biochemistry and related fields. With the development of recombinant DNA technology it became possible to introduce a certain gene into a fast growing organism such as *E. coli* and “force” it to produce the protein of interest in substantial amounts. The human hormone somatostatin was the first heterologously expressed protein produced in *E. coli* [34]. The production of human recombinant insulin is another important example of the success of this approach [35]. The method, beside being cost effective, had the advantage of producing an insulin that is chemically identical to its naturally produced counter part [36, 37], and therefore did not result in the immune response experienced by some patients against porcine or bovine insulin.

Today there are multiple choices of strains, expression vectors, promoters, and different purification tags such as polyhistidine tags (His-tag), that one can employ as tools for protein production and purification. However overexpressing a heterologous protein is often not without challenges.

**Protein production problems**

In a systematic study of the *Thermotoga maritima* proteins, 29% could be produced in a soluble form, which allowed facile and generic scale up of protein for crystallization studies [38]. However, eukaryotic proteins are expected to be considerably more challenging. When 2078 randomly selected *Caenorhabditis elegans* full-length genes were expressed in *E. coli* only 11% yielded soluble proteins [39]. In a similar study, where approximately half of the ORFeome of *C. elegans* i.e. 10,167 ORFs were overexpressed in *E. coli*, it was demonstrated that 47.7% of this unfiltered and unbiased large gene collection could be expressed. On the other hand only about 15% of the ORFs resulted in soluble protein [40]. Upon overexpression of a protein in a host the metabolic burden increases, which is manifested in a decrease in cell growth rates, an increase in cell lysis, and the induction of stress proteins such as proteases and chaperones [41]. The failure to overexpress heterologous proteins is often associated with folding problems, that may lead to the formation of inclusion bodies or degradation by the host organism [42].

**Suggested solutions for reluctant targets**

Expression vectors are equipped with different promoters such as the *lac*, arabinose, T5 and T7 promoters [43-46]. These promoters differ in strength
dictating the level of expression of the gene they control. The use of different promoters has been exploited to improve the yield of soluble proteins. Additional regulation is also required to eliminate the background levels of expression, by for instance, supplying the cells with an additional plasmid encoding a repressor gene. This is especially valuable in cases where expression of the exogeneous protein has a toxic effect on the host cell. Studies on the effects of different promoters, and host strains for the high level expression have been carried out [47]. Further expression regulation can be used in cases where the protein is stabilized by complex formation with other proteins. In these situations, three solutions have been suggested: bicistronic system, dual vector systems and two promoter systems [48-50].

Numerous other strategies are currently employed to overcome the problem of low soluble protein yield. Heat shock induction has been used to increase yield [51], and lowering the expression temperature is another popular approach [52, 53]. It has also been shown that a 6-fold increase in expression of active protein could be achieved by simply fine tuning the inducer concentration [54, 55]. Therefore optimizing the inducer concentration can be a rewarding approach. Furthermore, in the case of the production of interleukin-1β, screening for temperature, media, and point of induction has also been beneficial [56]. Screening cultivation conditions such as pH, temperature, and other additives is often a useful and a highly recommended approach to increase the efficiency of protein production. Vincentelli et al. [57], made use of an iterative procedure to increase the success rate towards reluctant targets by screening for combinations of seven strains and three different temperatures. An integrated system for cultivation has been developed in our laboratory [58] with the ability to screen multiple parameters, and additionally to increase the yield by increasing cell density.

**An appealing approach**

Another approach for increasing the success rate is to instead of changing parameters mentioned earlier change the protein itself i.e. using the protein itself as a variable in expression and crystallization experiments. There are suggestions that random mutagenesis could also be efficient to improve solubility [59]. However, the simultaneous processing of homologues, for instance from lower eukaryotes or prokaryotes, may be more rewarding, since structures of variants of proteins are satisfactory when the function of the protein generally is preserved. A successful example is the work by Locher et
al. on ABC transporters [60]. In order to increase the likelihood of success, they subcloned the genes for 28 distinct ABC transporters originating from different organisms. They also varied the location of the His-tag. Together this led to the successful determination of E. coli BtuCD structure. A similar strategy was also applied in the structure determination of the E. coli mechanosensitive channel MscL [61]. This should be the initial method of choice for challenging targets. In a recent report, and also with a membrane protein this approach has been successfully applied [62]. It should be mentioned that as a concept, the homologue screening as an approach is not at all a new innovation. It was originally exploited by Kendrew already in 1954 in his original selection of myoglobin [63].

Screening for soluble proteins

Whatever problems one may encounter upon trying to overexpress a protein, and regardless of what strategy is chosen to overcome them, it is useful to have a “HTP” solubility screening method to monitor the expression progress. Different laboratories have suggested diverse approaches for this task. A fusion solubility reporter is a widely used method, where a reporter domain with an easily measured characteristic is fused to the desired protein domain via a flexible linker. If the protein studied is soluble, the reporter domain remains functional and therefore its activity can be detected. On the other hand, if the protein studied is insoluble the function of the reporter domain will be affected. The use of green fluorescent protein (GFP) as a solubility reporter is a representative of this approach [64]. GFP is suitable as a reporter protein [65] because of its relatively small size, non toxicity, autoflorescence, and noninvasive in situ detection. Several reports have described the successful use of this approach. For instance it has been effectively used as an indicator to monitor membrane protein overexpression in E. coli [66], and in a global topology analysis of the E. coli inner membrane proteome [67]. There are numerous examples where GFP has been used to increase the amounts of soluble protein. One is the use of GFP in combination with directed evolution, which led to a 5 fold increase in soluble expression of TEV protease [68]. Splitting a reporter protein is yet another innovative approach and is a modified version of the fusion solubility reporter method. The underlying principle here is that separate pieces of the reporter protein are not active, while the activity is restored when the two pieces are associated. This
structural complementation requires that the protein of interest is fused to a small piece of the reporter protein, and if the protein of interest is soluble then the complementation is allowed, and the signal can consequently be detected. Reporters such as β-galactosidase (β-gal) [69], and GFP [70] have been used. However the drawback of this approach, as reported for the β-gal, is that the fused fragment was found to render soluble proteins insoluble [71]. Use of an antibiotic resistance gene as a reporter is a different and inventive approach as a tool for screening for soluble proteins. It relies on the use of antibiotics as a selective marker. The fused protein, if not soluble, will inactivate the resistance gene and so only cells expressing soluble fused protein will grow in media containing high antibiotic concentration, therefore allowing large libraries to be screened. This basic idea has been investigated and it has been demonstrated that the survival of E. coli positively correlates with the solubility of the fused test proteins with the chloramphenicol acyltransferase (CAT) protein [72]. The drawback of this method is that CAT functions as a trimer, and therefore can affect the solubility of the fusion protein.

A common feature of the methods outlined above is that the fusion partner has to be cleaved off before the target protein can be used in biochemical and structural studies, because of the interference of the fusion protein. This removal may affect the solubility of the target protein, since fusion partners can have a chaperone-like effect [73-75]. Another complication is that it is not always possible to cleave off the fusion partner, since there are cases where the cleavage site is not available for the protease. Due to these considerations, and because of ease of purification, the His-tag has become the tag of choice. In a survey of crystallographic studies it was found that almost 60% of the proteins produced were found to include a His-tag [76]. In summary, fusion partners are useful, but should be avoided whenever possible, because in addition to the reasons discussed above, they add extra steps to the protein production platform, which is undesirable in the context of HTP.

**Solubility screening by FiDo**

Antibodies are widely used in biochemistry. Their high specificity and ease of detection have made them invaluable tools. Immunological methods are appropriate as reporters of solubility in SGs projects since most proteins are expressed with purification tags, and antibodies directed against purification tags are available. Our laboratory has contributed with a method to help cutting down the cost of small scale screening for soluble proteins by
eliminating the tedious step of centrifugation. The method proved that small-scale lysis on a filter plate followed by a filtration step is an adequate method for separating soluble proteins from inclusion bodies [77]. The following is a short description of the procedure. A 96 well filter plate replaces the centrifugation step to separate the soluble fraction from inclusion bodies, and is used to separate soluble from insoluble protein. In our validation experiments, the insoluble form of the cytosolic part of the human EphB1 was used as a negative control to demonstrate that the filter could separate soluble proteins from inclusion bodies, and as a positive control the soluble SAM domain was used. Cells are harvested by applying a vacuum, and lysed on the membrane. Soluble proteins are obtained in the flowthrough, whereas inclusion bodies are retained in the filter. Screening for total expression can be done by using a denaturing buffer directly after harvesting. Protein samples are analyzed by dot blot using antibodies directed against the tag. The method was named FiDo, because it includes Filtration, and Dot blot. Besides the benefit of removing the tedious and the labor-intensive process of centrifugation, and the potential to handle large number of samples, another pivotal advantage is the feasibility of this method for automation. The screening method has now been adapted by others [78]. Most notable is that it was utilized for expression of soluble proteins in a cell free system [79]. The method is schematically presented in Figure 1. A recent method development for screening for soluble proteins has been developed in our laboratory, and it is based on detection of soluble protein at the colony level. This approach employs the same basic principle as FiDo. Upon lysis of colonies on a filter membrane, soluble proteins are allowed to diffuse to a layer of nitrocellulose [80]. The method is therefore called the Colony Filtration (CoFi) blot.
Figure 1: Flow chart for FiDo procedure. 96 different proteins or protein variants are grown in a 96 deep well plate. Cells are transferred to two filtration plates and are harvested by vacuum. The collected cells are lysed either by a native lysis buffer or a denaturing lysis buffer for soluble and total protein analysis respectively. The flowthrough is analyzed by dot blot.

Deletion library screening for soluble proteins (Paper I)

It is widely recognized that protein production is particularly challenging in the case of human proteins with multiple domains. Dyson et al. tested 30 different human proteins, in a multi-parameter approach, using 14 different expression vectors in *E. coli* [81]. The main conclusion from the study was that there is a strong correlation between successful soluble expression and molecular mass of the protein. Proteins with an average molecular mass of 22.8 kDa did not require fusion to a solubility-enhancing protein for soluble expression. Truncation of multidomain proteins, i.e. expressing each domain *per se* is a widely used approach; a relevant example is the EphB2 protein, from which the kinase and the SAM domains were expressed and purified, and
the structure of each was determined individually [82-84]. There are several methods available for domain border definition. Pfam [85] is a widely used tool for initial domain composition analysis. The domain border can also be experimentally defined by employing limited proteolysis and mass spectrometry (MS) analysis [86, 87]. As an example, Leulliot et al. could obtain soluble protein but failed to crystallize hPTPA, Ypa1, and Ypa2, therefore, they decided instead to design constructs of truncated versions of these proteins. Non-structured regions identified by bioinformatics analysis and limited proteolysis were removed in the expressed construct, and thereby they could obtain high resolution structures [88]. There are however limitations to this approach when the bioinformatics analysis is linked with a significant degree of uncertainty, in particular with poorly characterized protein families. Experimental domain border definition using partial proteolysis requires that the full-length protein can be expressed and purified. Fragments resulting from proteolysis do not necessary implicate an intact domain since flexible loops within a domain also are subjected to proteolysis. Another disadvantage of both approaches is that they do not guarantee soluble expression even though the domain is correctly predicted.

Our group has developed a novel approach that does not require preceding knowledge about the domain border, and does not have the disadvantages mentioned earlier. In this approach a unidirectional deletion method based on the commercially available Erase-A-Base™ kit (Promega) [89] is used to generate a set of truncated versions of the protein of interest. The rationale behind this approach is that constructs with different lengths will have different solubility characteristics. In addition to defining the domain border of the protein, the strategy will potentially also improve the chances of obtaining protein crystals when multiple variants of the proteins are tested. The method basically takes advantage of the two features specific of Exonuclease III. First its uniform and controlled rate of digestion, and second its inability to initiate digestion at 3’ overhangs. Therefore, for the purpose of producing unidirectional cleavage, two restriction enzymes are used, one that leaves a protective 3’ overhang, and a second that leaves the unprotected 5’ overhang. Time aliquots are collected from the reaction mix containing the cleaved plasmid and Exonuclease III. The resulting ssDNA overhang from Exonuclease III treatment is removed by the action of S1 nuclease. Klenow DNA polymerase is subsequently added for blunt end formation. The plasmid is circularized, and can be used for soluble protein expression screening.
This idea was first put to test on the cytosolic part of the human EphB1 protein, that is composed of two domains and showed poor solubility in initial expression trials. Indeed constructs with varying lengths and improved solubility profiles were obtained as indicated in Figure 2. Constructs shown in lane 3-6 are of varying sizes, but with improved solubility compared to the initial construct. These results provided the proof of principle of the deletion screen method.

![Figure 2: Analysis of constructs obtained from the deletion screening. Constructs 1-6 are positive clones from the Erase-A-Base reaction (EAB). The Ec is the cytosolic construct before the EAB. Q20, M19, and RGS-His are antibodies directed to the SAM domain, the kinase domain, and to the His-tag respectively. These results represent the proof of concept, and an effort to demonstrate the usefulness of the method for a large set of proteins has been carried out by Cornvik et al. [90] in combination with the CoFi-blot.

One of the positive constructs contained both the kinase and the SAM domains (Cyt EphB1) and was selected for crystallographic studies. A purification scheme was worked out, and the protein could be purified to homogeneity in four steps: affinity chromatography, desalting, anion exchange chromatography, and gel filtration. The results are presented in Figure 3. In
addition the gel filtration chromatogram established that the Cyt EphB1 is a monomer, and that no higher aggregation forms were present. The pure protein was concentrated and used in extensive crystallization trials, including several commercially available HTP crystal screens at two different temperatures, using both native protein and cocrystallization with ATP, ADP, and AMP-PNP. Later on a dephosphorylated variant of the Cyt EphB1 was also employed in crystallization trials. The investigation of the phosphorylation state of the protein is detailed in the following section.

**Figure 3:** Purification of Cyt EphB1. The optimized purification protocol included Ni-chelating affinity purification (A), followed by a desalting step before an ion exchange chromatography step (B), and finally gel filtration chromatography (C). Molecular mass markers (kDa) are indicated to the left.

**Dephosphorylation of EphB1 construct**

Analysis of the purified protein showed that the EphB1 construct is produced in a phosphorylated state. A previous study on the phosphorylation sites on the EphB2 and EphB5 led to the identification multiple phosphorylation sites [91]. The purified protein was auto-phosphorylated, which was evident from the signal of an anti-phosphotyrosine antibody, as shown in Figure 4. Autophosphorylation of EphB1 in its native environment depends on an extracellular signaling, but can apparently also occur under overexpression conditions in *E. coli*. The observation of autophosphorylation in our protein raised further questions about its origin and how it might affect the structural studies. There are multiple possible phosphorylation sites on the Cyt-EphB1 [91, 92], including the juxtamembrane tyrosines (Tyr604 and Tyr610), Tyr750, and Tyr788 [93, 94] using EpB2 numbering. Phosphorylation, and other modifications can often add significant microheterogeneity, which is a
major disadvantage in crystallization trials. To investigate if the phosphorylation could be reversed, LAR [95], and YOP [96] tyrosine phosphatases were tested. This experiment demonstrated that Cyt EphB1 could indeed be dephosphorylated by the two enzymes, and that the reaction went to completion as judged by the complete disappearance of the signal. This is presented in Figure 4. The desphosphorylated protein could be prepared in sufficient amount for crystallization trials. Another question raised was to what extent that protein was phosphorylated and whether the phosphorylation levels could be further increased. The outcome indicated that the signal was not further increased upon prolonged incubation of the protein with additional Mg-ATP.

To sum up this study was a proof of concept for the deletion screen method. The originally insoluble construct turned soluble by subjecting it to the deletion approach. The soluble protein was purified to homogeneity, and subsequently the phosphorylation could be reversed. Homogeneous dephosphorylated Cyt EphB1 was also subjected to crystallization trials.

<table>
<thead>
<tr>
<th></th>
<th>ATP</th>
<th>LAR</th>
<th>YOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

**Figure 4:** The phosphorylation state of the purified Cyt EphB1 obtained from the deletion screen method. Lane C represents a control sample containing the purified Cyt EphB1. The pure protein was positive towards the antiphosphotyrosine. Prolonged incubation with Mg-ATP for 30 and 60 minutes did not result in an increase in the signal. Upon treatment with LAR and YOP, the Cyt EphB1 phosphotyrosine signal disappeared after prolonged incubation.
Protein kinases, RTK, and the Eph family

Phosphorylation is probably the most common protein modification event in the cell and probably the most studied. For the cell to maintain its integrity and communicate events that are essential for its survival, it needs devices to regulate certain responses such as that to growth factors. The first direct evidence for the occurrence of enzymatic phosphorylation of a protein substrate, and an attempt to isolate it was provided in 1954 [97], and in the following year the role of phosphorylation became apparent as Fischer, Krebs [98], Wosilait and Sutherland [99] found that the interconversion of phosphorylase A involved phosphorylation and dephosphorylation, suggesting that reversible phosphorylation could control enzyme activity. This discovery was awarded the Nobel price in 1992 because it established that reversible phosphorylation works as a molecular switch to regulate a multitude of processes in the living cell. Protein kinases constitute a large group of proteins that are involved in regulating the activity of other proteins by transferring a phosphate group to certain residue(s) on the target protein. It is estimated that around 30% of the proteomes has covalently linked phosphate [100]. Kinases catalyze the reaction by binding and orienting the phosphate donor group as well as positioning the substrate for the phosphate transfer. This modification ultimately leads to an alteration in the activity of the target proteins. The outcome can either be an activating or an inhibiting effect if the target protein is an enzyme. Phosphorylation can however have other effects such as priming the modified substrate for degradation [101, 102], provide a binding site for modulator proteins containing domains such as SH2 and PTB [103]. These domains specifically bind to phosphorylated proteins and act as a platform for mediating effects of the signaling pathways.

The catalytic domain

Early on it was realized that the kinases have striking sequence similarity in the catalytic domain. This was first noted when the sequence of the catalytic domain of cAMP dependent protein kinase was compared to the sequence of pp60<sup>c-src</sup> (the cellular protein kinase from which retroviral protein pp60<sup>v-src</sup> was derived) [104]. The substrate was different but the two enzymes showed high sequence similarity in their catalytic domain [104]. In a review from 1987 [105], the author hypothesized the existence of as many as a thousand and one genes encoding kinases in the human genome. The sequencing of the human
genome now allows a comprehensive overview of the protein kinases in human. Based on the analysis of the human genome 518 protein kinases have been identified [106], about half that were predicted earlier. This means that approximately 2% of the human genome encodes for kinases. Furthermore chromosomal mapping revealed that 244 of the kinase genes are located within disease loci. Eukaryotic protein kinases are composed of Ser/Thr kinases, tyrosine kinases, and dual specificity kinases [107]. The catalytic domains are highly conserved, but are distinguished by specific signature motifs [108]. Purified protein-tyrosine kinases are highly selective and do not phosphorylate the substrate when the tyrosine is mutated to serine or threonine residues [109]. The structural basis for this strict specificity was explained when the first reported crystal structures of PKA [110-113] were compared with the insulin receptor protein kinase [114, 115].

**Receptor tyrosine kinases**

Tyrosine kinases were first discovered in 1997 [116]. Today some 90 tyrosine kinases are identified in the human genome [117]. Of these 58 are receptor tyrosine kinases (RTK) that constitute a large and diverse group of membrane spanning cell surface receptors. They are anchored in the membrane by a single pass α-helix, and have a basic architecture with a ligand interaction domain on the extracellular side of the membrane and minimally a kinase domain in the cytosolic compartment. Their activities are modulated by ligand binding. Upon activation they are capable of mediating different elementary responses such as control of metabolism and cell division. RTKs contain a tyrosine kinase domain on the cytosolic part of the receptor. Binding of ligand to RTK results in the transfer of the ATP γ-phosphate to tyrosines on target proteins. RTKs are monomeric single polypeptide chains in the absence of the ligand. Exceptions are the insulin receptor which is an α2β2 heterotetramer [118] and Met and its family members, that are composed of an α-chain linked by a disulfide to the membrane spanning β-chain [119]. RTKs include receptors for many growth factors such as epidermal growth factor (EGF) and platelet-derived growth factor (PDGF). RTK dimerization is, in the majority of cases, the activating mechanism by which the receptors mediate and transmit the signal to the intracellular compartment [120, 121]. However ligand induced dimerization is not a universal device of activation [122, 123], as revealed in the EphB2 ephrinB2 complex structure [83] where the receptor-ligand complex is a tetramer. However, regardless of the activation
mechanism, the active RTK needs to be counterbalanced to prevent constitutive activation. Down-regulation is mediated by different processes; receptor mediated endocytosis [124] and ubiquitin-directed proteolysis [125] are typically employed. Furthermore the action of a protein phosphatase is broadly utilized [126-128]. The central role that kinases and phosphatases play in cells makes them a hot target for drug development, particularly against diseases affected by oncogenic signaling [129].

**Eph receptors, the largest family of RTK**

Eph receptors are members of the RTK family. The first Eph receptor was identified as an orphan receptor in a search for tyrosine kinases involved in cancer [130]. The new receptor was named Eph since its cDNA was isolated from the Erythropoietin Producing Hepatocellular carcinoma cell line. Their functions emerged from studies showing that almost all of the receptors are expressed in the developing or adult nervous system. The Eph family of receptor tyrosine kinases is now the largest known family of receptor tyrosine kinases identified in mammals, consisting of 14 receptors (Eph) that bind specific cell bound ligands called ephrins. There are 8 ligands (ephrins) identified [131]. Based on the manner they are associated to the membrane the ephrin ligands are divided into two classes [132], namely ephrinA, and ephrinB. ephrinA ligands are membrane anchored by a glycosylphosphatidylinositol (GPI) linkage, whereas ephrinB ligands are attached through a transmembrane domain, and possess an additional cytoplasmic domain. They share sequence identities ranging between 23 and 56%. The initial view that ephrinA binds and activates EphA and ephrinB binds and activates EphB, with the exception of EphA4, has been further refined by the discovery that ephrinA5 binds and activates EphB2 [133]. The Eph receptor family has multiple roles in the cell. In the nervous system, the Eph receptors and Ephrin are involved in, for example, patterning the developing hindbrain rhombomeres, axon pathfinding, and guiding neural crest cell migration [134]. A series of studies has demonstrated the essential role of Eph receptors/ephrin in vascular development during embryogenesis and in adult angiogenesis [135]. Overexpression of Eph has been linked to oncogenesis [136, 137] and in a recent report it was shown that EphA3 is critical in heart development [138].
**Eph receptors and cancer**

Eph receptors and their cognate ligands possess a wide functional diversity in development. It is now well established that Eph receptors and ligands also play a diverse role in carcinogenesis [139]. For instance, it has been reported that the Eph receptors are differentially expressed in benign human tissues and cancers [140]. The mRNA levels of ephrin-A3 were upregulated 26-fold in a lung cancer tissue sample, whereas in hepatocellular carcinoma the mRNA level of EphB2 was upregulated 9-fold. The involvement of Eph receptors in colon cancer has been indicated; EphB2 has enhanced expression in colon carcinoma specimens [141] and in small cell lung carcinomas [142]. Eph receptors have also been found to regulate angiogenic processes associated with tumor growth [143-147]. In a recent report it was shown that ephrinA5 expression promotes invasion and transformation of murine fibroblasts [148]. In a study of the level of expression of Eph and Ephrin in ovarian cancer, increased levels of ephrinA1 and ephrinA5 in the presence of high levels of Eph receptors A1 and A2 led to more aggressive tumor phenotypes and shortened survival [149]. Because of their role in malignancy Eph receptors have attracted attention as potential drug targets [150, 151].

**Unique features of the Eph receptor**

The Eph receptor family is unique among RTK family in three aspects. First the ligand is cell attached and not soluble as is the case for other receptor kinases such as the insulin or growth factor receptors. Second the ligand needs to be membrane bound [152] or clustered. The clustering effect was demonstrated in a study showing that ligands were unable to act as conventional soluble factors. However, they did function when presented in membrane-bound form, suggesting that cell-to-cell contact is required to activate the receptors. Furthermore antibody-mediated clustering activated previously inactive soluble forms of these ligands [153]. The third characteristic is that the signaling is bidirectional. The signal is transduced in Eph as well as ephrin expressing cells. The reverse signaling capability of the ligand was first observed as an increase in transforming capability of a ligand lacking the entire C-terminal domain compared to the wild type ligand [154]. The effect of receptor ligand association was manifested for ephrinB by phosphorylation of conserved tyrosines in the cytoplasmic domain [92, 155, 156].
Previous structural work on Eph family

Because of its importance in various biological processes, the Eph receptor family has been subjected to intense structural investigations by several laboratories which has resulted in the elucidation of several components of the Eph receptor family. The ligand binding domain of EphB2 was the first structure to be determined [157]. Shortly afterwards, four structures of the SAM domain were solved [82, 158-160]. The kinase domain of murine EphB2 was subsequently determined in an autoinhibited state [84], followed by the kinase domain of human EphA2 [161]. The ectodomain of ephrinB2 [162], and the ligand binding domains of EphB2 [157] were solved separately, and later in complex with each other [83]. The ephrinB2 cytoplasmic domain was also structurally characterized [163]. Furthermore the EphB2 ligand binding domain in complex with ephrinA5 has been solved [133]. Recently the EphB4 receptor in complex with a highly specific antagonistic peptide was also determined [164], and that of an EphB4 ephrinB2 complex [165]. These concentrated efforts of structural studies along with biochemical research have considerably increased and detailed the understanding of this member of RTK family. However more studies are needed to add to our current perceptions.

Motivation of the Eph study

The soluble construct that we have obtained by the deletion screening method is composed of part of the juxtamembrane, the tyrosine kinase and the SAM domain. The SAM domain is frequently found as part of other proteins involved in regulating cell signaling and transcription. At first the SAM domain was recognized as a conserved domain found in 14 eukaryotic proteins [166]. The SAM domain is now known as an abundant motif found in more than 1300 proteins across different genomes; in the human genome it is found in 206 proteins. It plays a versatile role as an interacting module [167-169], mediating protein-protein interactions, and capable of forming different oligomerization states. It can bind to other related SAM domains [170-173] and to non SAM domain containing proteins [174]. The SAM domain has additionally emerged as an RNA binding molecule, and numerous structures of RNA binding SAM domains have appeared [175-180]. It has also been shown that the SAM domain can interact with lipid membranes [181]. The value of solving the structure of the EphB1 kinase/SAM domain construct was to gain structural insights into how the SAM domain and the tyrosine kinase
domains are arranged, and consequently the involvement of the SAM domain in functional aspects.

**Human EphB1 kinase domain structure (Paper II)**

The deletion library screening resulted in the production of several constructs with improved solubility profiles. The construct chosen was composed of the two domains. The presence of the SAM domain was demonstrated by antibody directed against an epitope in the SAM domain (Figure 2). The function of the juxtamembrane tyrosines was established by mutation of the two conserved tyrosines to glutamic acid residues. These mutations resulted in the abrogation of SH2 binding, but did not reduce the kinase activity [182]. The inhibitory effect on kinase activity was demonstrated when these two residues were instead changed to phenylalanines. Structural evidence was provided by the autoinhibited structure of the murine EphB2 receptor [84], where it was found that the juxtamembrane region was folding on the kinase domain leading to the distortion in the kinase domain which inhibits the phosphoryl transfer reaction. Upon comparison of the EphB2 structure to insulin receptor kinase (IRK) [183], it became apparent that the conformation of the αC helix was distorted. Based on this observation a mechanism was proposed in which phosphorylation of the two conserved tyrosines of the juxtamembrane region releases these residues from their initial position, leading to a conformational change in the αC helix and the activation of the kinase domain [84].

**The overall structure of the kinase domain of EphB1**

Two structures of the kinase domain of the human EphB1 were solved: The native structure was solved to 1.7 Å, and a binary complex with an ATP analogue AMP-PNP was determined to 2.4 Å resolution. In the final model only the kinase domain was present. SDS-PAGE and antibody analysis of the purified Cyt EphB1 supported the presence of the two domains in the purified protein, as shown in Figure 3. However, it is possible that the crystals obtained may be a result of proteolysis during the crystallization experiment. This unfortunate event was confirmed by analyzing a drop from the crystal optimization experiment using SDS-PAGE (data not shown), which revealed partial proteolysis. As anticipated, the EphB1 structures possessed large overall structural similarity to the earlier reported EphB2 kinase structure. The overall structure encompasses the classical kinase domain architecture [184]. There are two domains: a smaller N-terminal lobe composed of 5 β stands and
one α-helix (αC), and a C-terminal lobe that is larger and predominantly α-helical. The two domains contribute to the cleft that forms the nucleotide binding site as observed in the complex structure with AMP-PNP. However comparison of the native and the complex structure revealed conformational changes upon nucleotide binding, as shown in Figure 5.

The purified protein used for crystallization trials was phosphorylated as indicated by Western blot analysis with antiphosphotyrosine antibodies. However we could not model any phosphotyrosines in the structure, indicating that they may have been located in the flexible regions, or that only the nonphosphorylated portion of the pure protein crystallized. The activation loop could not be modelled, which is similar to other Eph receptor structures. Its possible that the activation loop only become ordered when the loop tyrosine is phosphorylated. Another possibility is that substrate binding is required for an ordered loop structure. This hypothesis remains to be experimentally proven or discarded.

**The conformational change of Tyr740**

While comparing the structure of the kinase EphB1 to the autoinhibited EphB2 structure an interesting observation was made. Tyr740 (Tyr750 in EphB2) was found in a conformation different from that of the corresponding residue reported in the autoinhibited EphB2 structure (Tyr750) implying a role in the activation mechanism (see Figure 5). The basis for this suggestion is that Tyr740 possessed the same orientation as that observed in the activated form of IRK (Phe1128). Noteworthy is that the orientation Tyr740 would make a clash with Tyr604 of the juxtamembrane region. Additionally, comparing the structure of EphB1 with that of the active form of IRK reveals that the observed conformation of Tyr740 would allow the activation loop to attain a productive orientation similar to that observed in IRK structure. This is unlike its position in the autoinhibited structure where it would be hindered. This is depicted in Figure 5.

**The αC helix**

The crystal structure of the autoinhibited EphB2 revealed that the αC helix was distorted, and it is believed that the binding of the juxtamembrane region is responsible for inducing this conformation [84]. However, the crystal structure of the kinase domain of EphA2, which does not include the juxtamembrane domain and this represents an active kinase form, shows a
similar αC helix conformation as the autoinhibited structure [161]. The distortion in the αC helix is apparent when comparing the EphB1 kinase domain to the active form of IRK, see Figure 5. The soluble construct obtained from the deletion screen included Tyr610 of the juxtamembrane region. However the juxtamembrane region could not be modeled, indicating disorder in this region. The EphB1 structures reported are of the active form, nevertheless they contained the αC helix in a similar conformation as reported in the structures of EphB2.

**Activation loop**

The activation loop was not visible in our final EphB1 kinase model indicating that it was disordered in the crystal. However the importance of the activation loop has been demonstrated for other protein tyrosine kinases. Very informative are the structures of the unliganded inactive form, and in the activated form of the kinase domain of the insulin receptor. In the unliganded IRK structure [114] the activation loop was found to traverse the cleft between the N- and C-terminal lobes where the ATP and the substrate bind. In the active structure the activation loop was displaced some 30 Å, permitting unhindered access of the substrate and ATP [183]. An autoinhibitory mechanism displayed by Tyr1162 of the activation loop was suggested, because this residue competed with the substrate for binding to the active site. Furthermore the cis-autophosphorylation of Tyr1162 is hindered by residues in the beginning of the activation loop, are hindering the access of ATP to the active site. The activation loop in EphB1 kinase structures potentially has the possibility to adopt the same conformation as observed in IRK structure.

Interestingly, the native and complex structures of EphB1 differ in the conformation in the starting region of the activation loop. In the AMP-PNP bound structure Asp762 of the conserved DFG motif adopts the same orientation as observed in the active form of other kinases such as IRK [183] and FGF [185]. However in the native structure of EphB1 Asp762 is adopting a different conformation. This residue is pointing in an opposite direction towards the hydrophobic region normally occupied by the conserved Phe763. Furthermore, the unusual conformation of the DFG motive was observed in the Tie2 kinase structure [186].

Based on these results we concluded that Eph has a multistep activation mechanism, and that our structures represent intermediate active state conformations.
Figure 5: (A) Overall view of the structure of the human EphB1 kinase domain in complex with AMP-PNP. (B) The structure of IRK in approximately the same orientation as EphB1, highlighting the similar overall fold of the two proteins. (C) The two structures have large structural agreement, however the kinked αC helix in EphB1 is however visible, in contrast to the straight helix observed in IRK. The EphB1 αC helix is shown in red and the corresponding helix in IRK is colored blue. (D) Comparison of EphB1 to the autoinhibited EphB2 kinase domain structure revealed that Tyr740 (red) has a different orientation than the equivalent residue in the inhibited state (Tyr750). Tyr740, as observed in our structure, would hinder the juxtamembrane region to achieve the inhibitory conformation because of the clash it would make with the juxtamembrane residue Tyr604 (Phe604 in the mutated structure).
More on the activation mechanism

While our structural work on EphB1 was being concluded, two interesting structures appeared, one representing the structure of the kinase domain of murine EphB2 (mEphB2) in complex with ADP and a Mg\(^{2+}\), and the second structure was of EphA4 including the juxtamembrane domain and a gain in function mutation Y750A EphA4JMS-Tyr750Ala [187]. These two structures along with biochemical and NMR data were revealing and gave support to the features found in our refined structures. Initially we proposed that our structures represented intermediary active state conformations since the kink that was observed in the \(\alpha C\) helix, and also due to the subtle conformation change that was more similar to the active state of the insulin receptor kinase structure. The recent work demonstrated that the juxtamembrane region of EphA4JMS-Tyr750Ala was disordered, strengthening the notion that the juxtamembrane region is actually the main regulatory element in the regulation of Eph family, and when it is repositioned the enzyme activity increases. Additional evidence was also given by NMR studies on EphB2 where the juxtamembrane tyrosines were phosphorylated [187]. The role of Tyr740 (Tyr750 in EphB2) was also investigated. The biochemical evidence was in form of mutational analysis including Tyr750Phe and Tyr750Ala. The Tyr750Ala mutant resulted in gain of function. It was suggested that the gain of function was achieved by dissociation of the juxtamembrane segment, which ordered the activation loop. However that was not observed in the crystal structure of EphA4, which included this gain of function mutation. Although the juxtamembrane was disordered, the activation loop did not become ordered. The complex nature of the regulation of Eph family urges further detailed structural studies. Most urgently is the complex structure of a fully phosphorylated kinase domain in a ternary complex with a substrate peptide and a nucleotide.
PTPA a regulator of PP2A

PP2A
Protein phosphatases, are classified into three structurally distinct families [188]. PP2A belongs to the PPP family and is among the most conserved proteins known [189]. It is essential in regulation of multiple essential processes in the cell including cell growth and cell cycle regulation [190, 191]. Certain pathogenic organisms disrupt the level of cellular protein phosphorylation by either encoding protein kinases and phosphatases or by adjusting the endogenous activity of these enzymes [100]. PP2A is a tumor suppressor [192] and a number of toxins and tumor promoter exert their effect by inhibiting PP2A. For instance microcystein, and okadaic acid are potent inhibitors of protein phosphatases [193, 194]. PP2A has a complex composition and regulation. The 36 kDa catalytic domain (C), and the 65 kDa scaffolding subunit (A) comprise the core enzyme. PP2A of this type represents at least one third of the cellular PP2A [195]. The core enzyme can further interact with the variable regulatory B subunits dictating the subcellular localization, and substrate specificity. These regulatory subunits are diverse and constitute three families: B, B’, and B’’. In total there are at least 16 members [191, 196]. Besides the regulation of activity by composition, PP2A is further subjected to regulation by phosphorylation and methylation [197-200]. Structural studies have been carried out, and the structures of protein phosphatase 2A core enzyme bound to tumor-inducing toxins okadaic acid and microcystein have recently been determined [201]. Predictably the catalytic domain of PP2A had the same overall fold as previously determined for related serine/threonine phosphatases [202-208]. It adopts an α/β fold, and contains two metal ions at the active site as found in related structures. Significant differences were however observed at the solvent exposed surface loops. Furthermore, the scaffolding subunit PR65/A in the core complex showed the same architecture as the same protein solved separately [209]. It forms an elongated structure made up of 15 HEAT repeats consisting of antiparallel double layeres of α-helices. The catalytic subunit was found bound to HEAT repeats 11-15 of the scaffolding subunit as predicted in a previous study [210]. Comparison of the free and bound scaffolding subunit indicated major conformational changes in HEAT repeat 13-15 induced by the binding of the
catalytic subunit. The scaffolding subunits recognize and bind only the catalytic subunit of PP2A despite the significant sequence similarity of the latter to other serine/threonine phosphatases. This observation was explained by the PP2A core structure. Interface residues between the catalytic and the scaffolding subunits were found substituted by non-conserved residues in related phosphatases. In addition, two tumor derived missense mutations (Arg418Trp in the α-isoform and Val545Ala in the β-isoform) map to the interface. These mutations disrupted the binding of the catalytic and the scaffolding subunit preventing the normal function of PP2A.

More information about PP2A evolved as a consequence of the PP2A heterotrimeric holoenzyme structure recently reported by Cho and Xu [211]. The structure of the PP2A catalytic (C-α) domain in complex with the scaffolding (A-α) and the regulatory B’ (B56-γ1) showed that the scaffolding A subunit forms a horseshoe-shaped fold holding the catalytic C and regulatory B’ subunits on the same side. Surprisingly the B’ subunit had HEAT repeats despite its lack of homology to HEAT repeat proteins. Residues identified to be important for A-B’ interactions [212] were mapped to the interface shaped by the two domains.

The C-terminal tail of the PP2A catalytic subunit is subject to modification by reversible methylation by the specific methyltransferase (LCMT) and methylesterase (PME-1) [213-216] at the C-terminus of the catalytic subunit. In the heterotrimeric complex structure, the catalytic subdomain was in a methylated state as a result of protein production, as noted in an earlier study [217]. The regulatory role of this modification could be explained at the molecular level. First, the C-terminal tail of the catalytic domain was found between the A and B’ subdomains reinforcing their weak interaction. Second, the methylated residue was positioned inside a highly charged environment created by acidic residues from the A subunit. Because the C-terminal tail also contains Asp306, which may contribute to charge-charge repulsion between the A and C subunits, it was believed that methylation of Leu309 would decrease the charge-charge repulsion and stabilize the AC dimer, thereby promoting the binding of the B’ subunit.

Not only could the effect of methylation be explained by the crystal structure, but also the inhibitory effect of phosphorylation [197]. Phosphorylation prevents complex formation between the C-terminal domain of PP2A and B56γ. The C-terminal Tyr307 of PP2A forms a hydrogen bond with the main chain carbonyl of Val257 in the B’ subunit. Phosphorylation of Tyr307 would
break this interaction, and as a consequence would affect the interaction between the two domains. Importantly, the structures also explain the genetic variations found in melanoma cancer (Arg418Trp PP2A Aα) [218] and colon adenocarcinoma (Val545Ala in PP2A Aβ corresponding to Val533Ala in Aα) [219]. These two residues are both located in the AC dimer. Our understanding of the PP2A function, and the complex nature of regulation, have increased in recent years, and our contribution was of the first crystal structure of the PP2A regulatory protein PTPA. A brief introduction on PTPA is given in the next section followed by discussion of the structural features of PTPA.

**PTPA**

PTPA is protein phosphatase two A phosphatase activator. PTPA has been shown to work as an activator of inactivated PP2A [220]. It has also been shown that PTPA can stimulate the phosphotyrosyl phosphatase activity of the PP2A A-C dimer [221-223]. The activation of PP2A has been shown to be dependent on ATP/Mg²⁺ in vitro [221, 222, 224]. However, PTPA lacked phosphotransfer activity when tested for autophosphorylation and the ability to phosphorylate PP2A or exogenous substrates [221]. Upon cloning and analysis of the primary structure, it was realized that PTPA lacked the canonical GXGXXG (where X represents a variable amino acid) typically found in most kinases, however a potential putative ATP binding site was detected [224]. Nevertheless, incubation of PTPA with the PP2A A-C dimer stimulates ATP hydrolysis but the reaction is blocked by okadaic acid, an inhibitor of the PP2A activity [225]. Furthermore, the binding of PTPA to PP2A is independent of ATP or Mg²⁺ [211]. However, the purified separate components did not result in detectable ATPase activity when incubated with ATP [222]. The importance of PTPA was demonstrated by showing that a PTPA deletion is lethal in *Saccharomyces cerevisiae* [226]. A recent study reported PTPA to be a peptidyl prolyl cis-trans isomerase of the Pro190 in the catalytic subunit of PP2A [227]. Previous studies have also indicated that PTPA can assist in the generation of a catalytic active di-metal centre [226].

**PTPA structure (Paper III)**

The structure of the human PTPA was solved to 1.6 Å resolution. The structure revealed a novel fold of a bilobal architecture that contains an all α-helical structure [228]. The large core domain is composed of 5 long helices
(α3-7) and a smaller four helix domain (α-9-12) connected by α-8. The highly conserved loop could not be modeled, indicating structural flexibility. The structure indicates that the majority of the conserved residues are found on the surface. Moreover two clefts were also prominent: cleft one is made from the edges of the large and the small domain, whereas cleft two is situated in close proximity to the loops connecting helices 5 and 6 (see Figure 6). Three loops connecting helices 3/4, 5/6, and 7/8 are found on the same side of the protein, pointing towards cleft 2 and harboring a glycerol molecule. This region of PTPA structure is rich in conserved residues implying a functional role. A hydrophobic pocket was identified at the end of the cleft where part of the TEV linker from a neighboring molecule in the crystal lattice interacts. This hydrophobic pocket was suggested to play a functional role based on the peptide binding and its degree of conservation.

In order to identify the ATP/Mg\(^{2+}\) binding site, binding studies with ATP and an analogue were performed both in solution and crystal form. No density was found after soaking ATP into the crystal. Furthermore, no detectable interaction was observed using ATP and ATP analogues in a thermal shift assay [229], implying weak or no binding. Soaking experiments with the tri- and penta-peptide (VPH, and EVPHE), and cocrystallization with the longer peptide (LNEVPHEGPMCAL) were not successful in capturing a complex structure. This can possibly be explained by poor accessibility in the crystal lattice or alternatively by direct competition of the TEV linker peptide. Jordens et al. reported that PTPA has a peptidyl-isomerase activity, and established that PTPA belongs to a novel peptidyl-prolyl cis/trans-isomerase family based on the failure of other PPIases to reactivate the inactive form of PP2A [227]. We therefore set out to investigate this by thermoflour binding studies using the peptide corresponding to the region in PP2A containing the highly conserved Pro190, but the results were not conclusive probably because of the interaction with the flouro-probe.

The PTPA structure is unique and does not show resemblance to any of the reported PPIase structures. Shortly after our structure was published two other human PTPA structures were published [88, 211, 230]. Significant insight into the function of PTPA was gained by the mutational study of Chao et al. [211]. Eighteen surface residues of PTPA were mutated and their effects on binding to PP2A and on ATP hydrolysis were investigated. The pocket on PTPA predicted from our structure to be a possible interaction surface with other molecules was indeed confirmed by this mutagenic study. A structure with a
A non-hydrolysable ATP analog established that the nucleotide binds to cleft 2 with the adenine and ribose moieties almost in the same position where we found a glycerol, as shown in Figure 6.

Mutated variants tested for ATPase activity resulted in the identification of residues in this cleft that are essential for activity. In addition, mutations affecting the binding to PP2A had a negative effect on the ATPase activity. Together, these results demonstrated that the interaction of PTPA and PP2A is a prerequisite for ATPase activity.

**Figure 6**: The overall view of PTPA structure and a structural summary of the current knowledge. Domains 1 and 2 are colored red and green respectively. (A) The overall view of human PTPA and cleft 1. (B) Cleft 2 with the glycerol molecules placed at its center. (C) The complex structure of PTPA (PDB code, 2HV7) in complex with ATPγS shown in yellow is superimposed on the ribbon diagram of our PTPA. The nucleotide was found bound to cleft2 region. (D) Surface representation of the structure of yeast PTPA superimposed on the nucleotide complex structure of PTPA. The peptide substrate succ-AAPK-pNa (yellow), and the nucleotide (blue) are represented in stick. Residues affecting binding to PP2A are shown in stick.
The ATPase activity was illustrated by the observation that PTPA exhibited an almost 6-fold activation of the PP2A A-C dimer activity towards p-nitrophenylphosphate (pNPP), and that this activity strictly depends on the presence of ATP-Mg$^{2+}$. A nonhydrolyzable ATP analogue, as well as ATPase deficient mutated forms, did not support the activation reaction. In an effort to identify the mechanism of PTPA action at the molecular level, Leulliot et al. determined the crystal structure of yeast Ypa1 and Ypa2 as well as the human PTPA [88]. This group was successful in co-crystallizing Ypa1 with the peptide substrate of the prolyl-isomerase reaction (succ-AAPK-pNa), shown in Figure 6. The model represents a head to head dimer with two-fold symmetry. The dimer induced a structural rearrangement of the central $\alpha_7$-$\alpha_8$ loop that was unstructured in the native structure. The proline residue in the peptide was found stacking with the strictly conserved Trp202 residue. To highlight the importance of this residue, a Trp202Gly mutant was tested for its ability to reactivate PP2A, and to perform PPIase activity on $^\text{186}\text{LQEVPHGPMCDL}^{\text{198}}$ containing the conserved Pro190 of PP2A. The Trp202Gly protein showed low specific activity for activating PP2A, and lacked PPIase activity. Preliminarly these results suggest that PTPA may act as an activator of PP2A by peptidyl isomerization of the conserved Pro190 that serves as a molecular switch [231].

Together our work and others have contributed to an increased understanding of PTPA and provided a tool for further investigations. A complex structure of PTPA and PP2A would contribute to a more detailed understanding of the interaction at the molecular level, in addition the role of PPIase activity and ATPase activity should be studied in more detail.
The YegS a remarkable enzyme (Paper IV and V)

Functional assignment and structure of an archetypical lipid kinase

Genomic projects have resulted in a myriad of protein sequences. However the majority of these sequences represent proteins with unknown function. Analysis of the human genome indicated that around 60% of the predicted proteins could be assigned a tentative function based on similarity to proteins with known function [14]. Decoding the biological function of the proteins encoded by the genomes remains a main challenge in the post-genomic era. It is widely accepted that annotation of function based on structural similarity is often more advantageous than sequence based similarity alone, since the structure is often more conserved than the sequence [232], therefore structural genomics was suggested as a tool to assign function from structural information [233]. The YegS project described here, suggests that biochemical characterization is necessary for an accurate refinement of such functional hypotheses.

Biological membranes and lipid kinases

Biological membranes are a prerequisite of life. In the absence of membrane chemical barriers, the cell would rapidly equilibrate with its surroundings. Beside defining a physical barrier between that cell and its surrounding, biological membranes mediate many important processes such as transport events and energy conversion. Some membrane lipids are also important for signaling when they serve as specific anchors for signaling proteins and as a source for the generation of second messengers. In particular phosphorylation of membranes lipids is central in signaling.

In eukaryotic cells, diacylglycerol (DAG) is a lipid second messenger generated as an effect of receptor activation and has important functions in the regulation of cell differentiation, proliferation and carcinogenesis [234, 235]. DAG kinases constitute a large family of enzymes catalyzing the conversion of diacylglycerol (DAG) to phosphatidic acid (PA). The major cellular function of DAG is to activate protein kinase C (PKC) [236]. Through the action of DGKs this effect is reversed by the conversion of DAG to PA. Other lipid kinases such as the mammalian sphingosine and ceramide kinases, as well as the recently identified Multi-lipid kinase (MulK) [237] with broader
substrate specificity, contain conserved motifs with a large family of bacterial proteins, including the *E. coli* protein YegS. Despite their importance there has been no structures reported for this class of lipid kinases.

**YegS a putative diacylglycerol kinase**

YegS is a 32 kDa cytosolic *E. coli* protein with sequence similarity to mammalian lipid kinases such as diacylglycerol, ceramide and sphingosine kinases [238, 239]. It was annotated as a putative diacylglycerol kinase (DGK) because of the shared sequence similarity to the first reported mammalian DGK-α [240]. The function of YegS was not known, and in fact, nothing was reported on this protein in the literature before this work. Gaining structural information about any member of the lipid kinase family would reduce our current lack of knowledge about lipid kinases, and this motivated our studies of YegS. In *E. coli* there is, however, already a well characterized membrane bound DGK [241, 242]. This DGK is a homotrimeric integral membrane protein in which each subunit has three predicted transmembrane helices [243], and there is no sequence similarity between the two proteins. However, YegS has a multitude of close relatives in the sequence databases. The close relatives are found in human pathogens such as *Shigella flexner*, *Salmonella typhimurium LT2*, and *Yersinia pestis*.

We set out to solve the crystal structure of YegS, and to investigate the function of this large group of putative prokaryotic diacylglycerol kinases. In the present study we have shown that YegS is a prototype lipid kinase with phosphatidylglycerol kinase activity, a previously unidentified lipid kinase activity. The structure of YegS reveals a two-domain protein with significant similarities to the fold and active site of NAD-kinases. An early mutagenesis study on porcine DAG α kinase identified six functionally important aspartic acid residues. These residues are shared across the ceramide and sphingosine kinases, as well as in some prokaryotic kinases with putative related functions to YegS. Based on the structure of YegS solved in the present work, we can now, to some extent, rationalize the functional role played by these conserved residues. Surprisingly the structure reveals a novel metal binding site where two of the conserved residues, Asp125, and Asp218, are involved in metal coordination. Activity studies indicate that YegS activity is sensitive to the presence of Ca\(^{2+}\) and Mg\(^{2+}\) and the novel metal site might in fact be a Ca\(^{2+}\) regulatory site.
Protein production and crystallization

The *E. coli* YegS gene was subcloned into expression vector PT73.3 \cite{244} using the Gateway® system. The procedures for scale-up purification were modified to optimize yield. Since YegS is supposed to act on a substrate located in the membrane, an ultra centrifugation step was included to capture the soluble fraction only and avoid the membrane bound fraction. For purification, two steps were employed; affinity and gel filtration chromatography. This was sufficient to produce homogeneous protein as determined by SDS-PAGE.

Based on the gel filtration elution chromatograms we could conclude that YegS was a monomeric protein, and that no other aggregated forms were present. The pure protein was subjected to crystallization trials. Eventually, crystals could be grown and a 1.9 Å native dataset could be obtained from one of these crystals. However, these well diffracting crystals could not be reproduced and we decided to change the strategy by changing to a construct with a cleavable N-terminal His-tag. New crystals were produced and selenomethionine labeled protein was used for phasing. The purification protocol was essentially the same as for the previous construct, but also included subtractive affinity purification after cleavage with a TEV-protease containing a His-tag \cite{68}. Eventually, two structures of YegS were solved, one native and one with YegS in complex with pyrophosphate moiety of ADP. These structures are discussed below.

From sequence based assignment to structural and biochemical characterization

As mentioned earlier YegS was proposed to harbor a diacylglycerol kinase activity based on its resemblance to the catalytic subunit of DGKs. YegS also shares conserved motives with other lipid kinases as is highlighted in the alignment shown in Figure 8. Most notable is the presence of a highly conserved GGDG motif. This motif is shared not exclusively within the lipid kinase family but also with the NAD kinases, 6-phosphofructokinases (PFK), as well as with several putative proteins encoded in prokaryotic genomes. The presence of a common signature sequence (GGDG) in these enzymes led to the suggestion that these proteins possess similar three-dimensional structures and phosphorylation mechanisms \cite{238}. 
**Figure 7**: (A) Structure of YegS showing the two domain composition. The N-terminal Rossmann like fold, and the larger C-terminal domain containing the novel metal binding site. Secondary structure elements different from the structurally related NAD kinase are shown in green. (B) By combining the structures of NAD kinase bound to pyrophosphate and another in complex with the substrate NAD, this picture depicts the likely nucleotide and substrate binding sites. The differences in the C-terminal domain are more apparent at the substrate binding region. (C) Comparison of YegS sequence to related prokaryotic and eukaryotic sequences. Representative of prokaryotic sequences is YPO2856; this protein is assigned as putative diacylglycerol kinase catalytic domain-containing protein, and originates from *Yersinia pestis*. Human DGKα represents related sequences from the diacylglycerol kinase family; MulK is the human sequence of the recently characterized multilipid kinase. CERK, is the human ceramide kinase, and SK-1, is the human sphingosine kinase isoform 1. The secondary structure of YegS is shown. The α helices represented in red, and β-strands in blue arrows.

This motif has been reported to be important for the function of these enzymes in several studies [239, 245-249]. Some mutational studies were based on the notion that the GGDG motif is involved in the ATP binding, since the classical ATP binding motifs in other kinases are rich in glycine residues and is often followed by a basic residue [250].

To evaluate the proposed function of YegS, we set up experiments where YegS was tested for its action on DAG, and other potential lipid substrates in combination with [γ-33P] ATP. Surprisingly, the data indicated that YegS lacked activity towards the presumed substrate DAG. YegS instead reacted with phosphatidylglycerol resulting in the formation of radiolabeled phosphatidylglycerol phosphate (PGP). When employing *E. coli* total lipid extract a product corresponding to PGP was detected, suggesting that this reaction can occur with an *in vivo* like substrate. The results of the radiolabeling experiments are represented in Figure 8. Additional experiments showed that YegS is unusual in catalyzing the phosphorylation of both membrane bound and soluble substrates, when an analog of PG with a shorter carbon chain was confirmed by mass spectrometry analysis to be a substrate of YegS. Despite the fact that no experiments were done to compare the efficiency of YegS phosphorylation of the soluble PG and the membrane bound form, one can interpret this result to indicate that the YegS catalytic site is already arranged to perform the reaction, and that the membrane is not a prerequisite for the catalysis to occur.
**The metal effect**

The crystal structure of YegS revealed a metal binding site as shown in Figure 7. A putative magnesium ion was found coordinated by the carbonyl oxygens of Leu215 and Leu220, the side chain of Asp218, and three water molecules that completed the octahedral coordination. One of these water molecules was coordinated by the conserved Asp125. In an attempt to explore the relevance of this metal site, experiments were designed to study the effect of Mg$^{2+}$ and Ca$^{2+}$ on the activity of YegS. Before testing the metal effect, the protein was treated with EGTA to remove contaminating metals that could potentially interfere during measurements. YegS showed strict metal dependency, and no activity could be detected in the absence of both Mg$^{2+}$ and Ca$^{2+}$. In the presence of 1 mM Ca$^{2+}$, some activation could be detected. The same amount of Mg$^{2+}$ resulted in lower activation than with Ca$^{2+}$. A boost in activity was however observed with 5 mM Mg$^{2+}$, and activation could also be seen with 1 nM of calcium only. When 5mM Mg$^{2+}$ and 1 mM Ca$^{2+}$ was employed, an inhibiting effect on the activity was observed, as compared to using 5 mM Mg$^{2+}$ only. Together these results point to a complex behavior of YegS towards the two metals. The behavior is, however, similar to the human MulK kinase and it might be that related enzymes would have a similar mechanism for metal dependent regulation [237].

**Figure 8:** Elucidation of YegS catalytic activity. (A) YegS activity tested towards diacylglycerol, phosphatidylglycerol, and total *E. coli* lipid extract shown in lanes 1-3 respectively. The lack of activity of YegS toward diacylglycerol, its proposed substrate, is evident in lane 1 by the lack of radio labeled product. (B) The metal dependency of YegS at varying concentrations of Mg$^{2+}$ and Ca$^{2+}$.
Structural features of YegS

The structure of YegS established that the enzyme has a two domain composition as depicted in Figure 7. A small N-terminal domain containing a Rossmann like fold [251], and a larger β-strand dominated C-terminal domain, are the main structural components. Furthermore, the cleft between the lobes harbors the nucleotide binding site as revealed by the ADP complex structure. Even though YegS is only similar to PFK and NAD kinases at the GGDG motif, these three enzymes are structurally related. The three enzymes share the N-terminal domain, whereas the C-terminal domain is uniquely found in YegS, NAD kinases and in the Vp4 a sialic acid binding domain of Rhesus Rotavirus [252]. It is also worth noting that the Vp4 protein does not contain a Rossmann fold, therefore we conclude that the overall fold of YegS is uniquely shared with NAD kinases. In a recent classification of kinases, NAD kinases, and PFK structures were grouped into the group 2 Rossmann-like family and PFK-like subfamily, and 1998 sequences were predicted to be part of this subfamily [253]. Our current results suggest that a revision of the classification should be made, and we propose YegS and NAD kinases to be included into a distinctive subgroup of the Rossmann fold kinases along with the distantly related human lipid kinases, as well as several other hypothetical protein with close sequence similarity to YegS.

A considerable effort was carried out to capture a complex structure of YegS including a nucleotide and substrate. This effort was unfortunately fruitless except for experiments including ADP. In the complex structure with ADP locate the binding site for the phosphate donor, however the entire ADP molecule could not be modeled since the electron density allowed only the modeling of the pyrophosphate moiety of ADP. This was located in the cleft formed in between the N- and C-terminal domains, and in close proximity to the conserved GGDG motif. A glycine rich cluster is a conserved feature of nucleotide binding proteins such as kinases. This consensus is somewhat different for different kinases but a notable feature is an overrepresentation of glycines close to the phosphates. They form a so-called G-loop that is involved in the binding of the nucleotide, and generally side chains facing the nucleotide are avoided [250]. The GGDG motif in YegS seems to perform a similar role. Furthermore, the crystal structure allowed the positioning of other conserved elements in the primary sequence. These are located in the surrounding region of the nucleotide binding site as demonstrated in Figure 9. Based on structural comparisons of the monomeric structure of NAD kinases
to PFK, the modeled pyrophosphate in YegS is found in an equivalent position as the phosphate donors in these proteins, supporting a common role for the GGDG motif.

The YegS substrate PG is situated in the membrane, therefore in order for the reaction to take place, membrane translocation is a prerequisite. However, it should be noted that the protein was classified as a cytosolic protein, and it shows the characteristics of a soluble protein i.e. it remained soluble without utilization of detergents during protein extraction and purification procedures. In addition the soluble extract preparation included an ultra centrifugation step. However it is not perceptible from the surface charge distribution which surface region might be involved in membrane interaction. Interestingly, the Membrane Protein Explorer (MPEx) server [254], proposed the region 157-175, including α6 which contains the conserved Tyr164, as a potential membrane interaction region, as shown in Figure 9. Tyr164 is conserved as a Tyr or Phe in most lipid kinases. It is well documented that tyrosine along with tryptophan residues are overrepresented at the membrane interface [255, 256]. Furthermore, the fact that the substrate binding site in NAD kinases is found close to this region strengthens the possibility that this is in fact the region involved in membrane interaction. It is of course plausible that the membrane interaction of YegS is only transient and does not require an extensive interaction surface with the membrane. The two structures of YegS along with biochemical data have increased the level of our understanding, but many unresolved questions remain to be tackled. Most important, efforts towards the elucidation of the biological relevance of the novel activity of YegS should be undertaken. The structural details are a prime starting point to develop testable hypotheses about YegS function.
Figure 9: The active site of YegS resides in the cleft region between the two domains. The Pyrophosphate (POP) of ADP is surrounded by highly conserved motifs. Asn12 and Lys14 belong to the NXXK motif, see Figure 7 for alignment. Asp67 denotes the GGDG motif here. Thr95 represents the T-motif. Together these residues, because of their degree of conservation and their location, are probably involved in the reaction mechanism, and are hot targets for mutational studies. The α-helix containing Tyr164 is proposed to take part in membrane association. Experimental evidence is of course needed to confirm this.
Concluding remarks

Evidently, some interesting aspects of the studied proteins have been exposed. Furthermore, some questions have been answered and others clarified. More importantly specific questions can now be asked and tested based on the reported results. Efficient protein production strategies will continue to be critical for structural biology. The deletion screening method together with the FiDo or the more recently developed CoFi method [80] are useful tools to address the challenging “high hanging fruits” in structural genomics, such as mammalian multidomain proteins. In our laboratory a parallel fermentation system allowing for automated screening of external and environmental cultivation conditions have been established and is expected to be very useful in the future [58]. Addition of compact experiments that can screen buffers, salt concentration, and different additives that are included in the lysis and/or gel filtration buffer will be profitable for difficult targets. In the EphB1 project, the native and the complex structures of the kinase domain have been solved. However, a structure of the entire cytosolic part of the Eph receptor will be a necessity to allow the understanding of the complex nature of the regulation of the receptor. This knowledge will be useful in putting forward a model for how the SAM domain may function in this system. The PTPA structure is suggestive and interesting mechanistic possibilities have been outlined. The binding site of ATP has been mapped, and the location of the substrate binding has been suggested. However, more in-depth studies on the role of the ATPase and the PPlase activities are needed. These and other novel aspects will, at the end of the day, be comprehensible in the light of a complex structure. Further efforts should be focused on capturing the complex structure of YegS with its substrate or product. Along with being related to mammalian lipid kinases, YegS has a multitude of homologues in other prokaryotic organisms, many of them pathogenic. The occurrence of YegS homologues in pathogenic organisms and the fact that they are still annotated as putative diacylglycerol kinases motivates further biochemical and structural studies. Characterization of these proteins might now be more feasible as a new candidate substrate has been added. Structural and biochemical studies of other lipid kinases will provide novel insights into the nature of this large group of enzymes in terms of reaction mechanism, and variation in regulation.
All of this together will inevitably increase our understanding of mammalian lipid kinases.
Finally, the results of these studies have been rewarding both at a personal and at an intellectual level. The results communicated here have added fine details, expanded, and modified the field. I certainly hope that the generated knowledge will be beneficial.
Acknowledgments

This part is doubtless the toughest part to write, if I would name each and every person that I am grateful for, and state the reason for that, it would result in another thesis work, and believe me I do not want to do that. In advance I am sorry for the people that are not mentioned by name.

I would like to thank first of all my beloved mother, father, sisters, brothers, the rest of the family, and all my dear friends. Oops! Many names are already left out. I am sure that you all and with no exception have been the most important people in my life. You were all highly appreciated before and after the thesis, and will remain so as long as I am metabolically active. However, I do not want this part to be nameless and should mention at least a few names.

From SU: Arne, klaas Jan, Jean-Benoit, Iwona, Stanislaw Karpinski, Linda, Annelie, Nina V, Ulrika P, Stefan Nilsson, Louise, Shashi, Nyosha, Thomas, Anders, Jeanette, David, Bogos, Kicki, Anki, Ann, Stefan, Elzbieta, Buck, Joe, Jan Willem, Gunnar, Peter, Pia, Inger, Dan, Joy, Agneta, Birgitta, Amelie, Åke, and Astrid. The KTH group at Albanova: Susanne, Martin, Esmeralda, Helena, and Torleif. KI collaborators: Jorge, Teresa, Lorenz, A great collaboration, Highly pure protein was delivered, diffraction to 5Å. It was a protein from hell, however very interesting discoveries were made. Thank you all. The PN group: Albert, Andreas, Anna Karin, Benita, Damian, Daniel G, Daniel MM, Deb, Ditte, Florian, Heidi, Hervig, Jessica, Karin, Ken, Lina, Lola, Maria, Marie, Marina, the Four Martins: Martin Andersson, Martin Moche, Martin Högbom, and Martin Hällberg, Monika, Pål, Rosi, Said, Sue-Li, Susanne, Therese, Tove, Ulrika, and Viktoria. I wish you good luck.

Arne Elofsson ett Mega tack det var mycket trevligt att göra examensarbete hos dig, och extra kul att det kom till användning som undervisnings material. Deb, I am going to remember you as a brilliant scientist with solid integrity. Anders Thornell, det var ett nöje att samarbeta, och lära känna dig. Calamares och vetenskapliga diskussioner, godare luncher får man leta efter. Karl-Magnus Larsson, du ska ha tack för bland annat alla vetenskapliga diskussioner och Matrix filmerna, för att inte nämnraalla synchrotron resor. Albert, thank you for more than just the revision of the manuscript. It has been great to have you around. Heidi, hyggelig å lære kjenne deg. Tobias, jag uppskattar verkligen din välvilja och jag ser framemot att CoFi metoden slår igenom. Marina, It has always been a pleasure talking to you. Audur, tack för
bland annat PTPA diskussionerna. Sue-Li, tackar för genomgången av manuscriptet. Said, tusen tack för allt du gör för labet, och för din positiva attityd. Benita och Marie, ni har varit en stor tillgång för labet. Jag vill poängtera min uppskattning för det arbete ni gör, och jag önskar er lycka till. Pelle, det var trevligt att lära känna dig i DBB, och det är extra trevligt att ha dig ombord i våran grupp nu. Monika, Viktoria, och Jessica, det var mycket angenämt att lära känna er. Florian, It has been great to get to know you. A major breakthrough is on the way. Andreas, thanks for all synchrotron trips and discussions. Lina, du var ett mycket positivt tillskott i våran grupp. Martin Högbom, våra sena samtal på DBB, biografbesöken, och en delad syn på vetenskapen, lycka till. Martin Moche, din personlighet är högt uppskattad. Tara, an impressive achievement-simply the best, keep on the good work. Medlemmarna av det killiga rummet: Damian, the triumph rider, motocykeldräkten kan med fördel användas vid protein reningen också-det har jag bevittnat. Excel! Daniel G, du livade upp rummet tack vare dina ständiga inlägg. Karin, tack för alla synchrotron resor, resetiden gick väldigt fort. Anna Karin, välkommen ombord! Jag önskar er alla lycka till. Stefan Nordlund, det var roligt att vara din student, samt att undervisa med dig. Rosi, thank you for setting the ground for the Eph project. Martin Hällberg, tack för natt patrullen i Labet och för samarbetet. The SBnet community, your relaxed and interesting conferences were the major events. I enjoyed them all, and I wish that other students would be able to do so in the future. Keep the same spirit. Younes, sommarjobbet var ingen dans på rosor, men nu har du respekt för en single band on a gel. Jag uppskattar verkligen din insats och prestation. Tack brorsan och kollegan, well done! Finally: Pär, our job is to rewrite the textbooks and to increase our knowledge. I think that I have made a small contribution towards that direction. A precondition beside my personal effort and commitment was yours in setting up the PN-lab, I am grateful for that among other things. This part would not even be close to complete if not EphB1, PTPA, and YegS were acknowledged. EphB1 you were my Mount Everest project. PTPA, your unique fold was beautiful, as well as suggestive. I am looking forward to your complex structure to be revealed. YegS, my kinder egg project in you I found a toy-my first structure, chocolate-the completely novel function, and a surprise-the unexpected metal site. YegS only Beethoven could have made justice to your beauty.
References


215. Leulliot, N., et al., *Structure of protein phosphatase methyltransferase 1 (PPM1), a leucine carboxyl methyltransferase involved in the*


