Protein production and purification in structural genomics

Martin Hammarström

School of Biotechnology
Royal Institute of Technology

KTH Biotechnology

Stockholm 2006
Martin Hammarström (2006): Protein production and purification in structural genomics. School of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden.

Abstract

The number of gene products available for structural and functional study is increasing at an unprecedented rate as a result of the successful whole genome sequencing projects. Systematic structure determination of proteins on a genomic scale, called structural genomics, can significantly contribute to the field of protein science and to functional annotation of newly identified genes.

This thesis covers different aspects of protein production in *Escherichia coli* for structural studies in the context of structural genomics. Protocols have been downscaled and standardized to allow for a rapid assessment of the production characteristics for multiple proteins in parallel under a number of different conditions. Foremost, the ability of different proteins and peptide tags to affect the solubility of the recombinant protein when produced as fusion proteins has been systematically studied. Large differences in the success-rate for production of soluble protein in *E. coli* were found depending on the fusion partner used, with more than a twofold increase in the number of proteins produced as soluble when comparing the best and the poorest fusion tags. For different constructs with a histidine tag, commonly used to facilitate protein purification, large differences in yield depending on the design of the expression vector were found. When comparing different fusion proteins produced from identical expression vectors, fusions to the GB1 domain were found to result in the highest yield of purified target protein, on average 25 % higher than any of the other fusions.

The suitability for further structural studies was tested at an intermediate scale for proteins that were identified as soluble in the expression screening. For this purpose, protocols for rapid purification and biophysical characterization using nuclear magnetic resonance and circular dichroism spectroscopy were developed and tested on 19 proteins, of which four were structured.

Keywords: Recombinational cloning, *Escherichia coli*, gene expression, fusion proteins, solubility, circular dichroism, nuclear magnetic resonance.

©Martin Hammarström, 2006
Contents

1 Background ............................................................................................................. 1
2 General introduction ............................................................................................... 3
  2.1 Genes, proteins and their structures ................................................................. 3
  2.2 Genomics ........................................................................................................... 6
  2.3 Structural genomics ............................................................................................ 7
3 Methods in protein production ............................................................................... 11
  3.1 Recombinant protein production in E. coli ....................................................... 11
  3.2 Cloning strategies ............................................................................................ 11
  3.3 Transcription ..................................................................................................... 13
  3.4 Translation ......................................................................................................... 16
  3.5 Fusion proteins and tags ................................................................................... 17
  3.6 Protein folding ................................................................................................... 18
  3.7 Alternative hosts ............................................................................................... 24
  3.8 Protein purification ............................................................................................ 25
  3.9 Biophysical characterization ............................................................................. 27
4 Present Investigation .............................................................................................. 33
  4.1 Experimental set-up .......................................................................................... 33
  4.2 Expression screening ....................................................................................... 38
  4.3 Rapid purification and biophysical characterization ......................................... 48
5 Conclusions and perspectives ............................................................................... 53
6 References .............................................................................................................. 57
Main References
This thesis is based on the following papers, referred to in the text by their roman numerals.


All papers are reproduced with permission from the copyright holders: © Cold Spring Harbor Laboratory Press A, © Springer B, © Elsevier science C.
Acknowledgements

I would like to acknowledge the following persons for their support during the course of this project:

My supervisors; Torleif for letting me work on the “high throughput” projects and sharing your expertise, Helena for your full support, especially during the last few years – I fully appreciate it, and Per-Åke for making this thesis possible after the departure of the structural biochemistry group.

Past and no longer present members of the structural biochemistry group; Mange for your encouraging and inspiring supervision of my diploma work, Per-Åke for being such an excellent diploma worker and great room-mate, Niklas “the musician” for all the discussions – fruitful or not, Esme for fruitful collaborations, Susanne for teaching me all good lab stuff and Henke, Magnus, Anja, Peter, Christofer, Elisabet, Jakob, Anders H and Ö, Inger, Vildan, Alex, and John for all being around and making work fun. The lunchroom team at CSB and all the nice people at the SBNnet and SNAR meetings.

Claes Gustafsson for an inspiring collaboration until the mediator project fell through and also all the nice people in the group of Pär Nordlund that have participated in the “high throughput” projects, especially Pär, Martin A, Marie, Benita, Ulrika and Tobias.

Jag vill även passa på att tacka mina föräldrar för det stöd de har gett mig under hela min uppväxt. Ni har gett mig en trygghet i mig själv som jag hoppas att jag kan överföra till Emil.

Slutligen vill jag rikta all min tacksamhet mot Anna, vars tålmod har prövats åtskilliga gånger av mitt avhandlings-arbete. Under den här krävande perioden har jag lärt mig att verkligen uppskatta min familj och jag hoppas att den uppskattningen blir mer påtaglig framöver.

Emil – varje dag med dig är unik!
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANS</td>
<td>1,8-Anilino naphthalene sulfonic acid</td>
</tr>
<tr>
<td>CD</td>
<td>Circular dichroism</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>DCW</td>
<td>Dry cell weight</td>
</tr>
<tr>
<td>GB1</td>
<td>The B domain of streptococcal protein G</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione S-transferase</td>
</tr>
<tr>
<td>HSQC</td>
<td>Heteronuclear single quantum coherence</td>
</tr>
<tr>
<td>IgG</td>
<td>Immunoglobulin, subtype G</td>
</tr>
<tr>
<td>IMAC</td>
<td>Immobilized metal ion chromatography</td>
</tr>
<tr>
<td>IPKA</td>
<td>Inhibitor of protein kinase alpha</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl-β-D-thiogalactopyranoside</td>
</tr>
<tr>
<td>MBP</td>
<td>Maltose binding protein</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>OD&lt;sub&gt;600&lt;/sub&gt;</td>
<td>Optical density at 600 nm wavelength</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>SG</td>
<td>Structural genomics</td>
</tr>
<tr>
<td>TCIP</td>
<td>Translationally controlled tumor protein</td>
</tr>
<tr>
<td>TEV</td>
<td>Tobacco etch virus</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
</tbody>
</table>
1 Background

This is an era of genomic exploration. The large number of completed genomes, accentuated by the high profile announcement of the completed sequence of the human genome in 2001 [Venter et al., 2001, Lander et al., 2001], has lead to an explosion in the number of genes available for study. The completed genome sequences define the framework for the molecular mechanisms of cells and organisms. As we continue to fill in the gaps in the gene function map this will ultimately lead to a more comprehensive understanding of the functional and sometimes malfunctioning mechanisms of life. Functional information for a large proportion of the genes in each new genome is currently lacking. This leads to an increased demand on biochemical studies to generate functional data at a higher pace.

Traditional biochemical approaches for functional studies of proteins are usually highly adapted to the individual nature of proteins. Large efforts are spent on extensive trial-and-error based optimization for individual proteins, and success is far from granted. These approaches are, and will continue to be, indispensable for the elaborate work of understanding complex mechanisms among highly diverse proteins. However, it has become clear that biochemical studies of proteins can also be performed in bulk, with standardized protocols and a high degree of parallelization and automation. Proteomics, the study of all proteins on the scale of organisms, is thus picking up the pace set by genomics. This increase in throughput also allows for more systematic and generalized comparisons of the efficiency of different protein production and purification protocols.

This thesis focuses on methodology in parallel protein production, primarily for structural studies. Particularly, the capability of different fusion tags as a means of increasing solubility of fusion proteins produced in *Escherichia coli* has been explored. Downscaling of cell cultivations to a multi-well plate format allows for several different production conditions to be screened for a large number of proteins and to systematically compare their efficiency. This methodology has been applied to a large set of human genes involved in cancer or disease. Proteins that are produced as soluble in the expression screening are purified at an intermediate scale to enable structural characterization and assessment of suitability for further structural studies. Based on the optimal fusion tag as seen in the expression screen, different
purification protocols were adapted to allow for very rapid purification and thorough structural characterization by the combination of NMR and CD spectroscopy.
2 General introduction

2.1 Genes, proteins and their structures

Proteins carry out most of the active processes of the cell. They are the active machinery that delivers the function encoded in the genes. The central dogma of life describes the flow of information from DNA to protein. Transcription is the first step, in which the original information of the DNA gene is copied into the blueprint molecule called mRNA. In the translation step, the coding sequence of the mRNA is used as template to build the amino acid sequence in the protein polymer chain. A protein complex called RNA polymerase performs transcription whereas translation occurs at the ribosome, a large rRNA-protein complex. The DNA polymer is built up by four different monomers, called nucleotides, and their linear combination can be experimentally determined by DNA sequencing. The protein polymer is built up from 20 different amino acid monomers and the genetic code (Fig. 2.1) tells us how the gene sequence is translated into the amino acid sequence of the protein.

<table>
<thead>
<tr>
<th>TTT = Phe</th>
<th>TCT = Ser</th>
<th>TAT = Tyr</th>
<th>TGT = Cys</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTC = Phe</td>
<td>TCC = Ser</td>
<td>TAC = Tyr</td>
<td>TGC = Cys</td>
</tr>
<tr>
<td>TTA = Leu</td>
<td>TCA = Ser</td>
<td>TAA = Stop</td>
<td>TGA = Stop</td>
</tr>
<tr>
<td>TTG = Leu</td>
<td>TCG = Ser</td>
<td>TAG = Stop</td>
<td>TGG = Trp</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CTT = Leu</th>
<th>CCT = Pro</th>
<th>CAT = His</th>
<th>CGT = Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTC = Leu</td>
<td>CCC = Pro</td>
<td>CAC = His</td>
<td>CGC = Arg</td>
</tr>
<tr>
<td>CTA = Leu</td>
<td>CCA = Pro</td>
<td>CAA = Gln</td>
<td>CGA = Arg</td>
</tr>
<tr>
<td>CTG = Leu</td>
<td>CCG = Pro</td>
<td>CAG = Gln</td>
<td>CGG = Arg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ATT = Ile</th>
<th>ACT = Thr</th>
<th>AAT = Asn</th>
<th>AGT = Ser</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATC = Ile</td>
<td>ACC = Thr</td>
<td>AAC = Asn</td>
<td>AGC = Ser</td>
</tr>
<tr>
<td>ATA = Ile</td>
<td>ACA = Thr</td>
<td>AAA = Lys</td>
<td>AGA = Arg</td>
</tr>
<tr>
<td>ATG = Met</td>
<td>ACG = Thr</td>
<td>AAG = Lys</td>
<td>AGG = Arg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GTT = Val</th>
<th>GCT = Ala</th>
<th>GCT = Asp</th>
<th>GGT = Gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTC = Val</td>
<td>GCC = Ala</td>
<td>GCC = Asp</td>
<td>GGC = Gly</td>
</tr>
<tr>
<td>GTA = Val</td>
<td>GCA = Ala</td>
<td>GCA = Glu</td>
<td>GGA = Gly</td>
</tr>
<tr>
<td>GTG = Val</td>
<td>GCG = Ala</td>
<td>GCG = Glu</td>
<td>GGG = Gly</td>
</tr>
</tbody>
</table>

Figure 2.1: The genetic code. Nucleotides are shown by their single letter code and amino acids by their three-letter code. The nucleotide sequence is read in triplets, called codons. The translation machinery recognizes the codons on the mRNA and incorporates the corresponding amino acid to the peptide chain. The genetic code is highly conserved among species, but the relative frequencies at which the codons occur in the genes vary considerably.

The amino acid sequence uniquely identifies individual proteins. But the function of proteins depends critically on how the linear chain of amino acids folds into the
three-dimensional structure of the protein. DNA in the cell almost always forms the structure of an elongated double stranded helix, independently of the sequence. Proteins on the other hand, are very diverse in their structures. The amino acid sequence is called the primary structure. Short-range interactions favor the formation of low energy conformations called secondary structure elements. The two most common secondary structure elements are the spiral $\alpha$-helix and the elongated $\beta$-strand. Secondary structure elements are connected by flexible loops or unstructured elements called random coils. The spatial conformations of the different secondary structure elements define the structure of the proteins and are formally referred to as tertiary structure. A specific arrangement of secondary structure elements is called a fold (Fig. 2.2)

![Figure 2.2: Different representations of the structure for the human TCTP protein (PDB entry 1Y41). To the left, a cartoon representation showing the conformation of the peptide backbone with the different secondary structure elements: (A) $\alpha$-helices as thick helices, (B) $\beta$-strands as arrows, and (C) extended regions of flexible random coils. In the middle, a ball and stick representation showing all atoms and covalent bonds and to the right a molecular surface representation. Based on the gene sequence it is possible to predict the primary structure of the encoded protein using the genetic code (Fig. 2.1). The prediction of secondary structure is not as reliable but can be made based on the amino acid sequence, as the amino acids have different propensities to form secondary structure elements. Specific combinations of different amino acids favor the formation of $\alpha$-helices, $\beta$-strands, or random coils, although other factors also contribute. Thus in rare cases, an]
identical amino acid sequence can be α-helical in one context but β-strand in another. It is currently not possible to predict the structure of proteins based on their sequence alone. However, one powerful way to enable prediction of both structure and function is to make comparisons to closely related sequences. Proteins that are homologous, i.e. they have evolved from a common ancestral protein; retain a similar function, structure and usually also a high degree of sequence identity. For closely related proteins it is usually possible to correctly attribute homology based on sequence alone. An empirically derived threshold for sequences sharing identities above 30 %, or sometimes lower if several related sequences are known, have been defined for attributing homology based on sequence. An important exception is the artificial protein Janus [Dalal et al., 1997], which was created by mutating half of the amino acids of the GB1 domain and thus creating a protein with a different structure. This illustrates that sequence identities above 30 % do not necessarily imply structural homology, although in this case it was an artificial protein. Several different algorithms are used to group proteins into families based on homology. One well-known example is the Pfam database [Bateman et al., 2004]. The power of evolutionary selection ensures that the function and structure of a protein is conserved to a higher degree than the sequence. That is, mutations in the amino acid sequence can be accumulated as long as they do not drastically disturb the function or the structure of the protein. This has the important consequence that is easier to detect homology based on structure than on sequence alone. Hence structural information can contribute to functional annotation of proteins with unknown function by detecting distant relationships to proteins of known function.

There are two methods to experimentally determine the structure of proteins to atomic resolution: nuclear magnetic resonance and X-ray crystallography. Both require large amounts (milligrams) at high concentrations of highly pure protein in a soluble form. Additionally, the combination of two-dimensional crystals and electron microscopy has reached near atomic resolution in favorable cases. The high-resolution structure of a protein is very useful for understanding the molecular mechanisms underlying the function of the protein. From this knowledge it is possible to predict how the function can be modulated, for instance in the rational design of a drug against a certain protein or engineering commercially interesting enzymes for better performance. Based on the experimentally determined structure of a protein, it is also possible to predict the structure of homologous proteins – to build
a structure model. Critical assessment of such models [Kryshtafovych et al., 2005] has shown that they are fairly reliable for proteins that share at least 30% sequence identity and that the quality increases with an increased homology [Baker and Sali, 2001] although the level of detail can not match that of experimentally determined structures.

2.2 Genomics

Traditional biochemical studies of proteins are usually highly protein specific with a trial and error approach to optimization of production, purification, and functional conditions for the protein being studied. There are several logical reasons for this: Biochemical properties like stability, solubility, and surface chemistry of proteins are much more individual than for example for DNA and thus it is not as easy to find conditions that are suitable for all proteins. Proteins can be highly sensitive to variations in their environment like pH, temperature, and buffer composition, especially at the high concentrations usually required for many of the functional and structural studies. Historically, the number of proteins available for study has to some extent been limited and thus justified extensive input of effort for individual proteins. As much of the work has been highly protein specific it has also been difficult to make comparisons and extrapolate general solutions.

In the field of DNA science, on the other hand, it has been easier to find general methods for production and purification of DNA. For example the polymerase chain reaction that allows for in vitro overproduction of specific DNA fragments and the restriction enzyme based methods for cloning are both essential contributions to all fields depending on genetic engineering. The development of the shotgun sequencing method opened the door for large scale sequencing of entire genomes and established the field of genomic studies. To date, the sequences of 250 (181) prokaryotic, 24 (20) archaeal and 39 (31) eukaryotic genomes have been completed and some additional 1000 organisms are ongoing or planned for the near future\textsuperscript{A}. The large number of proteins that has become available for study through these genomic sequencing projects and their success have created a surge of related studies of proteins on the scale of entire organisms, termed functional genomics or functional proteomics.

\textsuperscript{A} \url{www.genomesonline.org} as of October 30\textsuperscript{th} 2005 (November 30\textsuperscript{th} 2004 in brackets)
Structural studies of proteins on this scale are referred to as structural genomics or structural proteomics [Shapiro and Lima, 1998].

2.3 Structural genomics

The ultimate goal of a SG project would be to provide structures for all of the proteins encoded in the genome of an organism. There are examples of SG projects that are targeting the entire proteome of small organisms, like the 1877 genes of *Thermotoga maritima* [Lesley et al., 2002b, DiDonato et al., 2004, Lesley and Wilson, 2005] or the approximately 2200 genes of *Pyrococcus furiosus* [Wang et al., 2005]. However, SG projects are usually restricted by practical limitations, for instance in what proteins that are possible or reasonable to produce, and hence they usually only target a subset of the proteins encoded in the genome of the organism [Linial and Yona, 2000]. For instance trans-membrane proteins, constituting some 20-40 % of the proteome, are often excluded due to that they are inherently difficult to produce in large quantities and that they require highly specialized protocols [Grishammer et al., 2005], although progress towards more generalized methods are ongoing [Eshaghi et al., 2005]. Proteins that are predicted to be unstructured or for which the structure of a homologous protein is known are usually also excluded. Some of the organisms for which a substantial subset of the proteome is targeted include *Methanobacterium thermoautotrophicum* [Christendat et al., 2000], *Mycoplasma genitalium* and *M. pneumoniae* [Kim et al., 2005], *Thermus thermophilus* [Yokoyama et al., 2000], *Saccharomyces cerevisiae* (baker’s yeast) [Quevillon-Cheruel et al., 2003, Quevillon-Cheruel et al., 2004, Yee et al., 2002], *E. coli* [Yee et al., 2002, Vincentelli et al., 2003], *Caenorhabditis elegans* [Chance et al., 2002, Wang et al., 2005], *Arabidopsis thaliana* [Thao et al., 2004, Jeon et al., 2005], and *Homo sapiens* [Heinemann et al., 2003, Ding et al., 2002]. Other criteria for target selection involve families of proteins [Hurley et al., 2002, Chambers et al., 2004], metabolic pathways [Bonanno et al., 2001], or proteins of special interest for pharmaceutical development, either from pathogens [Abergel et al., 2003, Segelke et al., 2004] or human genes involved in cancer or disease [Gong et al., 2003]. The latter criterion has also been included in the target selection for paper I-III. Additional aspects of target selection are to discover all possible protein folds and to generate structure models for all proteins [Linial and Yona, 2000, Bonanno et al., 2005]. The current threshold for accurate homology detection and for generating meaningful structure models is 30 % sequence identity. By clustering proteins into families based
on at least 30% sequence identity and solving the structure for at least one protein in each cluster, it would be possible to make structure models for almost all proteins. It has been estimated that this would require at least 16000 carefully selected proteins to be targeted for structure determination in a coordinated effort [Vitkup et al., 2001]. This does not include proteins without detectable sequence homology to any other protein, however.

SG projects can significantly contribute to the annotation of function for newly discovered genes. Based on sequence alone it is often not possible to deduce the function of genes as they lack detectable sequence homology or are homologous only to other proteins of unknown function. In each newly sequenced genome, as many as 30-50% of genes are of unknown function. Structural genomics have proven useful for rapid identification of protein function for a large number of previously uncharacterized proteins by a number of different means. Structures are more conserved than sequence and hence distant homology can often be detected. Sometimes it is possible to deduce function from identified active sites. In some cases, the proteins contain cofactors or substrates that facilitate functional identification [Yakunin et al., 2004, Norin and Sundström, 2002, Skolnick et al., 2000, Laskowski et al., 2003, Kim et al., 2003]. A recent evaluation of 316 structures deposited by different SG consortia generated over 9000 homology models and showed that 30% of the determined structures revealed distant homologies not apparent from sequence alone [Todd et al., 2005].

Besides structure determination and target selection strategies, an additional major goal of most SG projects involves method development to enable that as many as possible of the selected proteins can be produced in quantities enough for structural studies (Table 1). The method development also concerns making as much as possible of the protein production and structure determination processes automated and rationalized to allow a high degree of parallel handling. This usually means decreasing the number of protocols and making them as general and independent of intrinsic protein properties as possible. Automation has been implemented to a high degree with the use of standard pipetting robots, mainly for cloning, small-scale expression screening, and crystallization set-up, and in some cases custom built industrial robots [Lesley, 2001]. The two major driving forces for method development can be in conflict to each other; on one hand there is a need to expand the repertoire of methods to be able to produce as many proteins as possible,
but on the other hand there is also a need to limit the number of methods to decrease
time and cost and thus be able to include as many proteins as possible. There is a
balance between the number of proteins to be included and the number of protocols
to use on each protein. This balance varies among different SG projects depending
on available resources and specific goals. Careful assessments of the utility of
different approaches to protein production in combination with an increased ability
to process many samples in parallel are key components in the method development
for SG projects. This offers the opportunity to make large-scale comparison of
different strategies regarding cloning, expression, and purification in a way that has
not been possible before. The unprecedented mass of experimental data generated by
the SG projects allows for extensive evaluations of parameters critical for protein
production [Goh et al., 2004]. Most of the currently ongoing SG projects are
presented in Table 2.

**Table 1.** Combined statistics for a number of large SG projects\(^A\)
showing progress through the protein production and structure
determination pipeline.

<table>
<thead>
<tr>
<th>Status</th>
<th>Number</th>
<th>Percent(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloned</td>
<td>56705</td>
<td>100</td>
</tr>
<tr>
<td>Expressed</td>
<td>32839</td>
<td>58</td>
</tr>
<tr>
<td>Soluble</td>
<td>14339</td>
<td>25</td>
</tr>
<tr>
<td>Purified</td>
<td>12064</td>
<td>21</td>
</tr>
<tr>
<td>Crystallized</td>
<td>4692</td>
<td>8.3</td>
</tr>
<tr>
<td>Diffraction</td>
<td>2063</td>
<td>3.6</td>
</tr>
<tr>
<td>Crystal Structure</td>
<td>2034</td>
<td>3.6</td>
</tr>
<tr>
<td>HSQC</td>
<td>1657</td>
<td>2.9</td>
</tr>
<tr>
<td>NMR Structure</td>
<td>769</td>
<td>1.4</td>
</tr>
<tr>
<td>In PDB</td>
<td>2517</td>
<td>4.4</td>
</tr>
<tr>
<td>Work Stopped</td>
<td>13714</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)Percent relative to number of cloned genes.

<table>
<thead>
<tr>
<th>Project/ Organisation</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-genomic Proteomics</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Structural Genomics Cons</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Proteome Institute, Lyon (FR)</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Protein Structure Facility (US)</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>PDB</td>
<td>2000</td>
<td>215</td>
</tr>
<tr>
<td>PDB-Select</td>
<td>235</td>
<td>241</td>
</tr>
<tr>
<td>The Jackson Laboratory (US)</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>The Rockefeller University Proteomics Center (US)</td>
<td>96</td>
<td>78</td>
</tr>
<tr>
<td>The Broad Institute of MIT and Harvard Proteomics Center (US)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Short description and number of solved structures for most active SG projects. Numbers are from v 4.0.
3 Methods in protein production

3.1 Recombinant protein production in \textit{E. coli}

\textit{E. coli} is one of the, if not the, most widely used host for recombinant protein production. It has been a model organism for molecular and cellular biology and as such, several non-pathogenic laboratory strains have been isolated and engineered to allow for easy and convenient handling. Nowadays, there are numerous and well-established protocols for most of the different methodologies involved in using \textit{E. coli} as a host for protein production [Makrides, 1996, Baneyx, 1999, Hunt, 2005]. Different strategies for cloning and factors that influence expression are discussed in this chapter. One of the major disadvantages of \textit{E. coli} is the relatively high risk of not obtaining the recombinant protein in a functional form, especially for eukaryotic proteins. The reasons for lack of function can be several but usually involve different aspects of improper protein folding. The protein might depend on disulphide bridges that cannot form in the reducing environment of the cells, unless it is secreted to the periplasm. It might need a stabilizing co-factor that is not available or it could be part of a multi-protein complex were the other components are missing. The folding could be depending on assistance in the form of chaperones, which are different in prokaryotic and eukaryotic cells. Some proteins require post-translational modifications or polypeptide chain processing to reach their functional form. There is also a coupling between the over-production of protein and improper folding. Even native \textit{E. coli} proteins that are normally soluble can be difficult to over-produce in \textit{E. coli} in a soluble form due to the high expression levels that are achieved with strong promoters. Improper folding of the protein usually leads to protein aggregation and the production of the protein in an insoluble form, which is frequently seen in \textit{E. coli} as dense inclusion bodies [Carrió, 2002, Fahnert, 2004, Villaverde, 2003].

3.2 Cloning strategies

The first step in the protein production chain is to clone the gene of interest into an expression vector. This usually involves PCR amplification of the coding sequence with primers that include elements that allow for a general cloning approach. Traditionally this is done by including cleavage sites for specific restriction enzymes
in the primers, cleaving the PCR fragment with the restriction enzymes, and subsequent ligation of the fragment into a vector containing the corresponding sites that has been similarly prepared. This has been the dominating strategy in molecular biology and numerous and highly specific restriction enzymes are available. However, this strategy is frequently hampered by a low efficiency (see for instance the initial cloning step in reference [Holz et al., 2003]), a fact that is seldom reported in literature as was stated by Aslanidis and de Jong “Our own results as well as anecdotal reports from other scientists indicate that this procedure is often inefficient.” [Aslanidis and de Jong, 1990]. Additionally there is a risk that the gene to be cloned contains the cleavage sites intended to be used for cloning, which require alternative enzymes to be used. It was calculated that 7% of the Yersinia pestis (plague) genes would be unavailable for cloning due to such interferences [Segelke et al., 2004]. Despite these potential disadvantages, some of the ongoing and successful SG projects rely on the traditional ligation depending cloning strategy [Biissow et al., 2000, Christendat et al., 2000, Lesley et al., 2002b]. However, a number of alternative cloning strategies have been developed that do not depend on ligation for cloning, collectively referred to as ligation independent cloning (LIC) [Aslanidis and de Jong, 1990, Heyman et al., 1999, Liu et al., 1998, Walhout et al., 2000, Hartley et al., 2000].

For the cloning step, all the studies in this thesis have used the Gateway system [Walhout et al., 2000, Hartley et al., 2000] that is commercialized by Invitrogen. The principle is based on the site-specific recombination system used by phage λ to integrate into, and excise out of, the genome of E. coli. It has been modified to accommodate two slightly different recognition sequences, termed att-sites, which allows for directional transfer of the DNA sequence located between the att-sites. The expression clones are created in two separate reactions termed BP and LR reactions based on the att-sites constituting substrates of corresponding reaction (Fig. 3.1). Advantages of this system include the simple reactions and the very low ratio of false positives. An additional advantage is that once an entry clone has been verified as correct, no additional PCR amplification is required and it can be sub-cloned into a large number of different destination vectors without the need for further sequence verification. This is especially beneficial for applications involving screening of large number of different expression vectors. A potential disadvantage lies in the inclusion of the attB-sites in the coding sequences and translated as part of any N- or C-
terminal fusion protein. Including a protease cleavage site in the construct can in part circumvent this. The cost of the enzymes can also be considered as a disadvantage.

![Gateway System Diagram](image)

**Figure 3.1:** Principles of the Gateway system. (A) An entry clone is created in the first reaction by the recombination between the attB sites of the PCR fragment and the attP sites of the donor vector. (B) The expression clone is similarly created in the LR reaction between entry clone and destination vector. The lengths of the att-sites are drawn schematically; the attB sites are actually 25 bp long, the attL and attR sites are approximately 125 bp long, and the attP sites are approximately 200 bp long. The donor and destination vectors contain an element for negative selection that is lethal for normal *E. coli* strains, and thus by-products (not shown in figure) or un-reacted vector will not give rise to colonies after transformation of the reaction into *E. coli*.

### 3.3 Transcription

Promoters are the DNA element from which RNA polymerase initiates transcription. In order to over-produce the recombinant protein, a strong promoter that will support synthesis of high levels of mRNA is frequently required. In extreme cases, this can result in the recombinant protein constituting up to 50% of total cellular protein. Besides supporting high expression levels, it is important to be able to induce transcription of the gene in an easy and reproducible way. Specifically, it is important that the level of transcription is as close to zero as possible before induction. This is because overproduction of recombinant protein represents a metabolic burden to the
host cell, resulting in slower growth rates for cells that are actively producing recombinant protein. When cells are cultivated for longer periods, this can result in cells losing the ability to produce the recombinant protein, for instance by accumulation of mutations in the plasmid or loss of plasmid. Loss of plasmid can occur if the antibiotic marker is unable to maintain the selective pressure throughout the cultivation period, as is often the case with the frequently used ampicillin. These non-producing cells will rapidly outgrow cells overproducing protein, which are growing at slower rates and this will result in lower yields. Additionally, some proteins are highly toxic to the host cells and pre-induction expression will be lethal to the cells. Examples and discussion about these effects can be found in the following references [Miroux and Walker, 1996, Studier et al., 1990, Grossman et al., 1998].

There are numerous different promoters that are used for recombinant protein production in *E. coli* [Baneyx, 1999, Makrides, 1996, Hannig and Makrides, 1998], and some of the most common include the *lac* promoter including the derivatives *lac* [de Boer et al., 1983] and *tre* [Brosius et al., 1985], the T7 promoter [Studier et al., 1990, Tabor and Richardson, 1985], the araBAD promoter [Guzman et al., 1995], and the *tetA* promoter [de la Torre et al., 1984]. Specific proteins that bind to promoter proximal sequences often regulate promoter activity. The *lac* operator (*lacO*) sequence is such an element from the *lac* promoter, and it is often used also in conjunction to other promoters like the T7 promoter. The *lac* repressor protein, which is encoded by the *lacI* gene, recognizes this sequence and binds in the absence of lactose. This binding efficiently inhibits transcription from the promoter. In the presence of lactose, the repressor protein dissociates and the promoter can be induced. As an alternative to lactose, which is metabolized by the cells, IPTG is usually used for induction. Also the araBAD and *tetA* promoters have similar regulation and there are examples when they all have been used in a modular design [Lutz and Bujard, 1997] to allow a wide range of transcriptional regulation.

The promoter of gene 10 of the T7 phage is commonly referred to as the T7 promoter. It is only recognized by the phage T7 RNA polymerase and not by the normal *E. coli* RNA polymerase. Thus, T7 RNA polymerase must be supplied in order to initiate transcription from the T7 promoter. This is usually done in special *E. coli* strains carrying a chromosomal copy of the gene encoding T7 RNA polymerase under the control of the *lacUV5* promoter. Hence, transcription from the T7 promoter is indirectly induced with IPTG as with the *lac* promoter. An alternative for
induction is the use of auto-induction media, in which expression is induced when cultures are grown to high density by balancing the availability of lactose and glucose in the media [Studier, 2005]. The advantage of the T7 promoter system is that the T7 RNA polymerase, which is highly effective, only recognizes the T7 promoter of the expression vector and hence very high levels of transcription from this promoter can be achieved. The disadvantage is that already very low levels of pre-induction transcription of the highly active T7 RNA polymerase can result in high levels of the recombinant protein. Two systems are usually used to repress pre-induction expression from the T7 promoter: (1) The lac operator-repressor system described above, in this case the promoter is called T7/lac promoter and (2) co-expression of the protein T7 lysozyme that inhibits T7 RNA polymerase activity.

In some cases it is useful to regulate the expression also to very low levels, for instance when the produced protein aggregates at higher concentrations or if it is toxic to the cells in high concentrations. For this purpose, the promoter of the araBAD operon is usually used together with the regulatory gene araC. This system allows for the expression level to be regulated in each cell over several orders of magnitude by controlling the concentration of arabinose that is used to induce expression. However, this system must be used together with a constitutive arabinose transporter [Morgan-Kiss et al., 2002, Khlebnikov et al., 2000]. Otherwise an all-or-nothing phenomenon occurs in which the apparent expression level of the cell culture is regulated by modifying the proportions of partially induced cells and uninduced cells rather than regulating the expression level in each cell [Siegel and Hu, 1997]. The all-or-nothing phenomenon occurs also for the lac family of promoters, but only when lactose is used as inducing agent and thus not when IPTG is used [Khlebnikov and Keasling, 2002]. As expression at lower temperatures is becoming increasingly popular as a method to avoid aggregation of the produced protein, the cold shock promoter of the capA gene can be an interesting alternative. Lowering the temperature in this case induces expression [Vasina and Baneyx, 1997, Qing et al., 2004].

Another important aspect of transcription to be considered is the termination. To stabilize plasmids and to avoid unnecessary synthesis of mRNA, effective transcriptional terminators should be included in the expression vector after the gene to be expressed. If this is not the case, expression of other plasmid-encoded proteins can also be elevated upon induction, which can disturb plasmid stability. If there are
other promoters on the plasmid that are on the same strand as the promoter used for expression, terminators could also be included in front of the promoter used for expression to avoid background expression from read-through transcripts. The turnover of mRNA is rapid in E. coli, with half-times around 20 minutes. It has been shown that 3' and 5' secondary structures in the mRNA can drastically improve stability and increase protein production [Baneyx, 1999].

3.4 Translation

The basic elements needed to initiate translation in E. coli are few and simple. A short element called the Shine-Dalgarno sequence, which is complementary to a part of the 16S ribosome subunit, constitutes the central part of the ribosome-binding site. If there is a start codon situated approximately seven bases downstream of the Shine-Dalgarno sequence, translation can be initiated at this site. However, the efficiency of translational initiation varies depending on the sequence in the ribosome-binding site. Upon binding to the Shine-Dalgarno element, the ribosome covers and can potentially interact with an additional 30-35 bases, including also parts of the coding sequence to be translated. This can pose a problem for production of eukaryotic proteins since their sequences are not optimized for prokaryotic translation initiation.

Using 5' gene fusions of prokaryotic origin, either full-length, short leader sequences or mini-cistrons [Schoner et al., 1986], can circumvent this problem. If gene fusions are to be avoided or if even higher rates of translation are intended, there are specific sequence elements that are known to positively influence translational initiation. These include the upstream enhancer element [Olins and Rangwala, 1989] that is present in the popular pET vectors and the downstream box [Etchegaray and Inouye, 1999], both from phage T7 gene 10. The latter, however, is part of the coding sequence and may not always be used. Additionally, it has been shown that the codon following the start codon can have a large influence on protein production [Looman et al., 1987, Stenström et al., 2001] and thus mutation of the second amino acid or introduction of an extra amino acid can be advantageous. In some cases, also the following codons can influence translational efficiency [Nishikubo et al., 2005].

The production of heterologous proteins can be hampered during the translational elongation. This is because organisms have evolved differently and are using codons with different frequencies. Some codons that are frequent in human genes are very rare in E. coli. The occurrence of such rare codons can slow down the
translational process or even lead to ribosome halting, premature termination, and protein degradation [Kapust et al., 2002]. The solution to this problem is to co-express the genes coding for tRNAs corresponding to rare codons. Several auxiliary plasmids or special *E. coli* strains are available for this purpose, including the CodonPlus™ and Rosetta™ strains.

### 3.5 Fusion proteins and tags

The cloning techniques allow for practically any gene to be modified by adding additional genes, part of genes, or synthetic sequences to the coding sequence. The expressed products of such artificial gene constructs are called fusion proteins. The protein that is added is called fusion partner or fusion tag, however the latter is often specifically used for short adducts. The major advantage of fusion proteins is that they offer a general approach to problems that are otherwise highly protein specific and they have been explored for a number of different applications [Nilsson et al., 1997]: (1) Purification based on the affinity of the fusion partner for a specific ligand or using the fusion as a specific ligand [Terpe, 2003]. (2) Detection based either directly on the enzymatic or immunogenic activity of the fusion partner or coupled via an affinity interaction of the fusion. (3) Selection or screening applications. (4) The influence on several different aspects of recombinant protein production like cellular localization, improved expression levels, regulation of solubility, protection from degradation, and avoiding toxicity effects [Waugh, 2005]. (5) Immobilization on solid support, for instance protein chips. It is obviously advantageous if a single fusion tag allows for a combination of different applications like high production levels in a soluble form, quantitative detection, and availability of affinity purification. Alternatively, multiple tags can be used in combination [Nilsson et al., 1996, Waugh, 2005]].

This section presents some examples of useful applications of fusion proteins in high throughput projects. The applications of fusion tags for solubilization and purification are covered elsewhere. The factorial effect of screening the expression of multiple genes under several different conditions requires analysis of a large number of samples. Immuno-based detection of a fusion tag for samples spotted on filters have been developed as an alternative to traditional SDS-PAGE protein gels. These methods, called dot-blots, offer the advantages of easy and automated sample preparations in combination with the possibility to analyze a very high number of
samples in parallel. It is usually based on vacuum-driven filtration through 96-well filter plates and subsequent spotting onto blotting paper [Knaust and Nordlund, 2001, Vincentelli et al., 2003, Doyle et al., 2002, Nguyen et al., 2004] although detection can also be performed directly in the clarified lysate [Coleman et al., 2004, Vincentelli et al., 2005]. Similar approaches that allow for even higher throughput have also been used for expression screening based on high density spotting of colonies onto in situ filters [Büssow et al., 1998] or by direct filtering of colonies from agar plates [Cornvik et al., 2005]. Other methods for in vivo screening of large libraries for soluble expression, for instance as part of a directed evolution project, uses C-terminal fusions of GFP [Pédelacq et al., 2002, Kawasaki and Inagaki, 2001, Waldo et al., 1999], split GFP [Cabantous et al., 2005b, Cabantous et al., 2005a] or α-fragment of β-galactosidase [Wigley et al., 2001] used for structural complementation, or chloramphenicol acetyl transferase giving chloramphenicol resistance [Maxwell et al., 1999]. Fusion tags are also used extensively for large scale mapping of protein-protein interactions as in the yeast-two-hybrid system and similar systems based on by structural complementation [Michnick, 2003, Ito et al., 2001, Fields and Song, 1989]. Chong and co-workers have used a modified intein protein as C-terminal fusion for auto-catalytic removal of itself and an N-terminal fusion tag used for affinity purification, thus allowing purification and fusion protein removal in a single step [Chong et al., 1997].

3.6 Protein folding

3.6.1 In vivo folding

The folding process for a protein is often a complex process depending both on the intrinsic properties of the amino acid sequence as well as environmental factors. Additionally, the energy difference between the folded and the unfolded states is in general small so proteins are comparably unstable and once in their native state they can easily be unfolded again. Due to differences in size and folding mechanisms, some proteins easily and rapidly fold into their native structure whereas others still are slow folding proteins or inherently difficult to maintain in their native state. The alternative path to proper folding is often unspecific aggregation. As newly synthesized proteins protrude from the ribosome, they rapidly collapse into an intermediate state as the hydrophobic side-chains condense to reduce their solvent
exposed surface area and as rapidly folding secondary structure elements are formed. The newly synthesized protein can either fold from this intermediate state into its native structure, by protein specific pathways, or aggregate with other proteins into large and insoluble assemblies [Fahnert et al., 2004, Carrió and Villaverde, 2002, Villaverde and Carrió, 2003]. It is thought that the competition between the folding rate and aggregation rate is one of the determining factors for the solubility of the newly produced protein [Agashe et al., 2004]. However, unspecific aggregation can occur also between folded proteins or between folded proteins and cellular components [Frankel et al., 1991, Frangioni and Neel, 1993] depending on the surface properties of the proteins.

**Figure 3.2:** Simplified and schematic representation of the *in vivo* folding pathways for newly synthesized proteins in *E. coli*. For most proteins folding begins immediately as the pep-tide is synthesized and starts to protrude from the ribosome. The complex folding mechanism to reach the native state is represented by a single rate-limiting step. Non-specific associations lead to small soluble aggregates, which can associate further to form insoluble aggregates and eventually inclusion bodies.

There are several different strategies to influence the chance of proper folding or the risk of aggregation (Fig. 3.2) [Sorensen and Mortensen, 2005]. The rate of protein synthesis can be lowered either at the transcriptional level by lowering the concentration of transcriptional inducing agent, using a weaker promoter [Mayer et al., 2004], or a plasmid with low copy number [Lutz and Bujard, 1997, Keasling,
1999] and hence lowering the amounts of available mRNA's or at the translational level by using a sub-optimal ribosome-binding site. It has also been shown that silent mutations thought to influence the translation rate of the protein can affect the solubility of the produced protein [Cortazzo et al., 2002]. Cultivation at a lower temperature will lower both transcription and translation level and will also reduce the strength of hydrophobic interactions involved in aggregation. Lowering the synthesis rate (Fig. 3.2) reduces the risk of overloading the folding assisting systems of the cell and also lowers the concentration of folding intermediates and hence the rate of aggregation, which is concentration dependant. The drawback of most of these strategies is of course the risk of lowering the overall protein yield.

3.6.2 Chaperones

The folding rate (Fig. 3.2) can be influenced by the presence of chaperones and folding assisting proteins that facilitate the proper folding of the protein. There are several different chaperone systems present in the E. coli cell that promote different rate-limiting steps in the folding process and which more or less specifically recognizes different sets of proteins [Baneyx and Mujacic, 2004]. Since the transcription of most chaperones is up-regulated under stress-conditions that cause endogenous E. coli proteins to aggregate, like elevated temperatures, many of the chaperones were identified as heat shock proteins. Lesley and co-workers showed that specific heat shock proteins are also up-regulated in response to production of insoluble recombinant proteins. They coupled this regulation to the transcription of a reporter gene that allowed a rapid colorimetric in vivo detection of aggregated proteins [Lesley et al., 2002a].

One class of folding assisting proteins is the peptidyl prolyl isomerases that assist in the cis-trans isomerization of the proline peptide bond, which is a slow folding event. This class of chaperones includes the trigger factor protein that binds to the ribosome near the exit tunnel for nascent peptide chains and thus probably is the first chaperone to interact with newly synthesized proteins [Agashe et al., 2004]. For trigger factor and the peptidyl prolyl isomerases PpiA and FkpA, it has been shown that co-expression could increase the yield of soluble recombinant protein [Nishihara et al., 2000, Zhang et al., 1998, Zhang et al., 2003].

Another class of folding assisting proteins is the disulphide bond isomerases that assist in the formation and isomerization of correct disulphide bonds. The cytoplasm
of *E. coli* is a reducing environment and hence disulphide bonds are not stable. The oxidizing environment of the periplasm, however, can maintain disulphide bonds and this compartment contains several different proteins for this purpose. It has been shown that co-expression of, or fusions to, different disulphide isomerases can increase yields for proteins containing a large number of disulphides [Zhang et al., 2002]. Alternatively, special *E. coli* strains like the Origami™ strains have been developed with deficient thioredoxin and glutaredoxin pathways. This makes the cytoplasm permissive to disulphide bond formation [Jurado et al., 2002, Bessette et al., 1999]. However, it should be stressed that the disulphide isomerases still are located in the periplasm unless variants lacking their leader sequences are co-expressed as described in the references above.

The chaperones assist in the folding of proteins by binding to exposed hydrophobic sequences and actively contribute to the formation of native structure, probably by promoting partial unfolding of the misfolded protein, although the exact mechanism for this is not known. Two major energy-dependant systems are present in *E. coli*: the DnaK, DnaJ, and GrpE system and the GroEL and GroES chaperonin system [Hemmingsen et al., 1988]. The DnaK system is thought to work upstream of the GroEL-system so that proteins that are unable to fold with the aid of the DnaK system will eventually be passed on to the GroEL system. Several studies have shown that co-expression of the two systems can work independently or in concert to improve solubility of over-produced proteins, but that it is difficult to foresee for individual proteins which of the systems will be active and that it is probably most beneficial to co-express them both [Nishihara et al., 1998, Nishihara et al., 2000].

Another type of chaperones, including the family of small heat shock proteins, act to reduce the aggregation rate rather than promoting proper folding and these are referred to as holding chaperones. In this case the partially folded protein binds to the surface of the holding chaperone and is thus prevented from aggregation with other proteins [Baneyx and Mujacic, 2004].

### 3.6.3 Solubility enhancing fusion partners

Solubility enhancing fusion partners are thought to promote solubility by a similar mechanism as the holding chaperones acting primarily by intra-molecular sequestering of aggregation prone target proteins on the surface of the fusion partner, thus preventing inter-molecular aggregation. There are several different fusion tags
that have been exploited to improve solubility of the fusion protein, including maltose binding protein [di Guan et al., 1988, Bedouelle and Duplay, 1988, Kapust and Waugh, 1999], glutathione S-transferase [Smith and Johnson, 1988], thioredoxin [LaVallie et al., 1993, Lu et al., 1996], NusA [Davis et al., 1999], the Z domain and the albumin binding domain of staphylococcal protein A [Uhlén et al., 1983, Nilsson et al., 1985], the GB1 domain of streptococcal protein G [Huth et al., 1997], mutant DsbA [Zhang et al., 1998] or DsbB/C [Zhang et al., 2002], the head protein D of phage λ [Forrer and Jaussi, 1998], and ubiquitin [Butt et al., 1989] with the close homologue SUMO [Chance et al., 2002, Malakhov et al., 2004]. Interestingly, the majority of these proteins are normally involved in protein-protein interactions as part of their function. Despite their frequent use, few large-scale comparisons of different solubilising fusion tags have been reported. The study in paper I represents the first such comparison for a larger number of genes and subsequently others have followed [Braun et al., 2002, Dyson et al., 2004, Korf et al., 2005, Shih et al., 2002].

Besides the likely mechanism suggested above, a number of other mechanisms can be involved in the solubilising effect of fusion partners, including secondary effects such as modulation of protein production rate for individual proteins, overall expression level, or protection against degradation [Waugh, 2005]. For MBP it has been suggested that the beneficial effect only arises when the MBP fusion tag is N-terminal of the recombinant protein and not when it is C-terminal or co-expressed [Sachdev and Chirgwin, 1998a, Kapust and Waugh, 1999]. However, although most work has been with N-terminal fusions, there are studies that argue against the exclusive requirement of the solubilising fusion being N-terminal [Dyson et al., 2004, Tucker and Grisshammer, 1996]. Mutational studies of MBP have shown that neither disruption of the known protein binding-site nor of the maltose binding-site impaired solubilization. Instead it was found that MBP mutants with destabilized folding, although still correctly folded, lost their solubilising capability [Fox et al., 2001]. Thus it is likely that rapid folding into a stable conformation is a key characteristic of effective solubilising fusion partners, in addition to specific surface properties. Not all soluble E. coli proteins can act as a solubilising fusion partner [Davis et al., 1999]. In several cases it has been shown that the soluble fusion protein represent an aggregated but still soluble population or that the recombinant protein is soluble but not in an active conformation [Ashraf et al., 2004, Nominé et al., 2001, Sachdev and
Chirgwin, 1999]. In these cases the solubilising fusion partner could be positioned as a hydrophilic shell around the hydrophobic aggregates analogous to micelles. Solubilising fusion partners can also exert their effect under *in vitro* refolding [Sachdev and Chirgwin, 1998b, Samuelsson et al., 1994].

The functional mechanisms of solubilising fusion partners are elusive and can very well be a combination of the suggested mechanisms. It is also likely that the actual mechanisms vary from fusion partner to fusion partner and also with the target proteins being solubilized. This could be an explanation for the differences in ability to solubilize recombinant proteins as has been detected upon comparisons [Vincentelli et al., 2003, Braun et al., 2002, Shih et al., 2002, Kapust and Waugh, 1999, Wang et al., 1999, Berthold et al., 2003, Korf et al, 2005, Dyson et al, 2004].

### 3.6.4 *In vitro* refolding

When it is impossible or undesirable [Murby et al., 1996] to obtain the recombinant protein in a soluble form, the insoluble protein can be solubilized by high concentrations of denaturing agents like 8 M urea or 6 M guanidinium hydrochloride. Upon removal of the denaturant by dialysis or dilution, the proteins can in some cases refold from their denatured form to reach their native structure. However, this method is often associated with extensive protein precipitation and optimizing the buffer composition to obtain the protein in a soluble form can be a painstaking procedure. Protocols have been developed to screen buffer conditions and allow a more rapid evaluation of the achievability of a refolding strategy [Vincentelli et al., 2004, Trésaugues et al., 2004, Armstrong et al., 1999, DiDonato et al., 2004, Voziyan et al., 2005]. Also a simplified standard approach has proven useful in allowing a substantial fraction of insoluble proteins to be recovered in a soluble form by refolding [Maxwell et al., 2003, Altamirano et al., 1997, Oganesyan et al., 2005], at least for proteins in the 20 kDa range.

### 3.6.5 Natively disordered proteins

As a final comment on protein folding, it should be stressed that the eventual successful folding of a given protein ultimately depends on the energy difference between the folded and the unfolded states. Thus some proteins may be inherently difficult to maintain in a folded state and it is easy to imagine that proteins that have evolved to be unstable even in their physiological environment are not easily
produced in a heterologous host, especially not in large amounts and under stressful conditions.

Additionally, it has been shown for a large number of proteins that they may not at all be structured in their native form, either in part or in whole [Uversky, 2002, Wright and Dyson, 1999, Dunker et al., 2001, Dyson and Wright, 2002]. It is estimated that in the order of 30 % of all eukaryotic proteins contain sequences longer than 50 amino acids that are disordered [Dunker et al., 2001]. These disordered regions may be involved structural transitions, either from structured to unstructured or vice versa, can be part of the functional mechanism of proteins, for instance in protein-protein interactions or substrate binding. In the absence of the correct binding partner, a substantial fraction of the protein may be unfolded [Dyson and Wright, 2002]. These partly structured proteins are often challenging both in terms of protein production, as they are prone to aggregate [Dyson et al., 2004], and of course also for structure determination [Gao et al., 2005, Pantazatos et al., 2004], as it can be difficult or even impossible to reach the fully structured state.

Proteins that have evolved to be entirely unstructured on the other hand, are often easy to produce in a soluble form due to their unusual amino acid composition [Uversky, 2002], and this was also seen in paper IV.

### 3.7 Alternative hosts

Despite the popularity of *E. coli* as an expression host, there are limitations. As mentioned in the previous section, incorrect folding and subsequent aggregation is frequently encountered in *E. coli*, especially for larger proteins with multiple domains. Results from the large scale expression tests performed as part of the structural genomics projects have shown that approximately 80-90 % of the cloned genes can be expressed in *E. coli* and that approximately 40 % of the proteins are soluble [Lesley et al., 2002b, Yokoyama et al., 2000, Yee et al., 2002]. Typically, these results were achieved for thermophilic or bacterial proteins produced in the cytoplasm using an expression vector that adds a His tag. As eukaryotic proteins are expressed, the fraction of proteins that are soluble can drop below 10 % [Chance et al., 2002, Gong et al., 2003]. However, other studies including paper I, III, and IV have shown that the success rate can be improved further by screening additional expression conditions [Abergel et al., 2003, Folkers et al., 2004, Quevillon-Chenuel et al., 2003]. Despite this potential increase, it is clear that a number of proteins require alternative
approaches for the production of properly folded proteins. Another potential drawback of *E. coli* as a production host is the lack of most of the post-translational modification systems found in eukaryotes. Post-translational modifications, such as N- and O-linked glycosylations or serine/threonine and tyrosine phosphorylations, are common among eukaryotic proteins. Although genetic engineering of *E. coli* has allowed the introduction of some of these modifications to recombinant proteins, this approach is still in its infancy [Yue et al., 2000, Zhang et al., 2004].

The most popular eukaryotic alternatives to *E. coli* include cell cultures of insect cells transfected with *Baculo-virus* [Albala et al., 2000, Chambers et al., 2004, Ikonomou et al., 2003, Loomis et al., 2005] and the yeast strains *S. cerevisiae* [Holz et al., 2002, Holz et al., 2003] and *Pichia pastoris* [Boetner et al., 2002, Cereghino and Cregg, 2000, Lueking et al., 2000]. Potential drawbacks of eukaryotic cells as expression host include that they are not as extensively characterized for this purpose as *E. coli*. Culture media is often more expensive, especially when it comes to isotopic labeling, and handling is more delicate than for *E. coli* cells. It has also been shown, at least for yeast, that they contain a higher degree of histidine rich proteins than *E. coli* and that this results in higher background in IMAC purification [Prinz et al., 2004].

Yet another interesting alternative is *in vitro* translation systems. In these systems, the necessary components for translation are reconstituted from fractioned cellular extracts of for instance *E. coli* or wheat germ cells and protein over-production can take place without the potential interference of other cellular proteins or organelles such as proteases. They are usually also coupled to transcription systems. High amounts of protein (mg/mL) can rapidly be produced, even compared to *E. coli*, from small volumes with low levels of contaminating proteins. Recent technology developments have allowed *in vitro* translation systems to be used both for expression screening [Busso et al., 2003, Busso et al., 2004, Endo and Sawasaki, 2004, Kigawa et al., 2004, Tyler et al., 2005] and sample preparation for structural studies [Kigawa et al., 1999, Yokoyama, 2003, Vinarov et al., 2004, Tyler et al., 2005].

### 3.8 Protein purification

To enable functional and structural studies it is required to obtain the protein in a highly pure form, usually at least 90%. Even minute amounts of contaminants can severely disturb the structure determination process or destabilize the sample; especially during the long time periods usually needed both for crystallization and
NMR experiments. To obtain such a high purity it is necessary to use a combination of different purification techniques. In the capture step, the bulk of contaminating proteins are removed and this is usually performed as an initial purification step. Removal of as much as possible of the remaining contaminants usually requires at least one or two additional polishing steps.

As mentioned earlier, fusion tags are often added to a target protein to facilitate purification [Terpe, 2003, Waugh, 2005]. This has the advantage of making the purification protocol highly insensitive to the specific properties of target proteins and thus protein dependant optimization is generally not needed. The affinity binding between ligand and protein is usually highly specific, allowing the removal of most of the contaminating proteins without extensive loss of target protein. Therefore affinity purification is very commonly used as a capture step. However, the drawback of this is that chemically modified or partially degraded proteins are not removed. Several different affinity purification systems using short fusion peptides (His tag [Gentz et al., 1988, Hochuli et al., 1988, Smith et al., 1988], FLAG tag [Hopp et al., 1988], strep tag [Skerra and Schmidt, 1999], or biotinylation tag [Schatz, 1993]), longer peptides and domains (Streptavidin binding peptide [Keefe et al., 2001], biotinylation tag [Cronan, 1990], calmodulin binding peptide [Zheng et al., 1997], S-peptide [Kim and Raines, 1993], GB1 [Huth et al., 1997], Z-domain [Nilsson et al., 1987], and the cellulose binding domain [Ong et al., 1991]) or full-length proteins (GST [Smith and Johnson, 1988], MBP [Bedouelle and Duplay, 1988, di Guan et al., 1988], and protein A [Uhlén et al., 1983]) have been described.

The combination of a His tag and immobilized metal ion affinity chromatography (IMAC) is probably the most commonly used affinity purification approach, both in SG projects and in general. It offers several advantages: It can be used under both native and denaturing conditions, elution conditions are mild, the His tag is short and inert enough to not adversely disturb protein function and allow structure determination without tag removal, the resin is comparably cheap and stable enough to allow regeneration, and depending on the application different chelating resins or metal ions can be used to optimize yield or separation.

The polishing steps usually involve at least one of the classical chromatographic separation techniques that are based on the intrinsic properties of the protein: Ion exchange chromatography separates proteins based on their interaction with a ionized resin, hydrophobic interaction chromatography is based on the interaction of surface
exposed hydrophobic patches with the resin whereas gel filtration, or size exclusion chromatography, separates proteins based on their size. Gel filtration is a popular last step because it offers an almost free choice of elution buffer. Both ion exchange and hydrophobic interaction chromatography require more protein specific optimization of binding and elution conditions than gel filtration and hence the latter is more readily incorporated into SG protocols. Methods for highly automated protein purification at the scale required for SG are available [Sigrell et al., 2003, Kim et al., 2004b]. As an alternative polishing step it is possible to use a second affinity tag fusion at the other terminus of the protein [Mueller et al., 2003, Hammarberg et al., 1989]. All degradation products are efficiently removed by such an approach. Modifications leading to an inhomogeneous sample can still pose a problem, for instance the observed incomplete initial methionine removal seen in Paper III or glyconylation of his-tagged proteins [Yan et al., 1999, Geoghegan et al., 1999], but they are not frequently encountered in *E. coli*.

Structural studies often require the removal of any fusion tags in order to minimize the risk for interference with crystallization by flexible tags or the crowding of NMR spectra with peaks arising from unwanted amino acids. Viral proteases like TEV and PreScission™ offer a good combination of specificity and activity and are available as recombinant fusion proteins, which allow for convenient removal of the protease after the cleavage reaction. A combination of IMAC purification followed by proteolytic removal of the His tag and a final negative IMAC, in which uncleaved protein, his-tagged protease and most of the impurities remaining after the first IMAC step bind to the affinity resin whereas the target protein does not, have been successfully applied [Kim et al., 2004b]. However, it is not always necessary to remove the fusion tag as there are now hundreds of structures that have been solved with the his-tag remaining. In some cases fusion tags are required for successful structure determination [Center et al., 1998].

### 3.9 Biophysical characterization

#### 3.9.1 Objectives

The main objective of biophysical characterization of proteins in the context of structural studies is to evaluate the suitability of the sample for structure determination, i.e. secondary structure content, if the protein is in a folded, partially
folded or unfolded state, the homogeneity or potentially aggregated state of the sample, the purity, and the stability of the sample. There are several different methods that can be used individually or complementarily to obtain information regarding the state of the protein sample, some of which are presented below. Ideally these evaluations can be performed beforehand structure determination to screen for suitable samples or buffer conditions to stabilize a sample, but perhaps more commonly they are performed alongside or in retrospect to support a successful outcome or to evaluate the cause for an unsuccessful outcome and the potential for improvement.

3.9.2 Methods

There are several spectroscopic methods that can be used to obtain structural information or to examine potentially multimeric or aggregated states of a protein. FarUV CD spectroscopy is very suitable to study secondary structure conformations of proteins and NMR is very suitable to study the folded state of proteins. Both these methods are discussed in more detail below. Dynamic light scattering [Ferré-D’Amaré and Burley, 1994] or fluorescence correlation spectroscopy [Thompson et al., 2002], in case a suitable fluorophore is available, both measure the time dependant fluctuation in signal for a small volume and give the diffusion constant for the protein and hence a size approximation that can be used for detection of aggregated or multimeric populations. Also static light scattering or straightforward turbidity measurements can be used to detect large aggregates [Trésaugues et al., 2004, Vincentelli et al., 2004]. Infrared spectroscopy can be used to study the backbone conformation and hence secondary structure content of proteins. Fluorescence spectroscopy or nearUV CD spectroscopy of tryptophan sidechains can give information on solvent exposure and hence degree of folding. Similarly, fluorophoric dyes like ANS can be used to detect proteins with partially folded conformations, e.g. molten globules [Semisomov et al., 1991]. In this case there is an increase in fluorescence when the ANS molecule binds to hydrophobic pockets that are present in the partially folded state but absent in the completely folded state. Many of the mentioned methods can be used also to measure protein stability by thermal or denaturant induced melting from the folded to the unfolded state. This can be very useful to identify specific buffer compositions or additives that stabilize the target protein [Pantoliano et al., 2001].
Non-spectroscopic methods to detect aggregated or multimeric populations include analytical gel filtration and analytical ultracentrifugation [Laue and Stafford, 1999], which measure the migration property of proteins through either porous media or density gradients, respectively. Native PAGE is a more rapid but less sensitive alternative on the same mechanism. Additionally, calorimetry can be used to study melting of proteins and measure thermodynamic parameters.

Mass spectrometry plays a fundamental role in traditional proteomics where it is used for protein identification. In SG projects, mass spectrometry is primarily used as a quality control step to confirm the identity of the produced proteins [Huang et al., 2003], but is also very useful for detection of modifications, degradations or even small molecules bound to the target protein that can be very difficult to detect by other methods. Recently, mass spectrometry methods have also been applied in high throughput applications to detect disordered regions in proteins [Pantazatos et al., 2004, Gao et al., 2005]. This information can be used to make new expression constructs in order to obtain samples more suitable for structure determination.

**3.9.3 Structural characterization**

**3.9.3.1 By NMR**

In NMR spectroscopy, the resonance frequencies for transitions in the nuclear spin state are detected via interactions with electromagnetic radiation in the radio frequency range when the sample are placed in a strong external magnetic field. Only isotopes with nuclear spin 1/2 are regularly used in NMR, including $^1$H, $^{15}$N and $^{13}$C, but not $^{14}$N or $^{12}$C. Each atom type, i.e. hydrogen, carbon, nitrogen etc., has a specific resonance frequency but for each individual atom in a molecule that frequency is slightly shifted depending on its unique chemical surrounding, giving rise to individual peaks in the recorded spectrum. The hydrogen atom ($^1$H) is the preferred atom for recording NMR spectra given its high sensitivity. For small molecules the peaks corresponding to different nuclei can usually be identified. For an unknown molecule, this can be used to deduce how the atoms are connected to each other and hence to determine its structure. However, for larger molecules like proteins the spectra will be very complex with extensive overlapping of peaks. The very fast $^1$H spectrum can still be useful as an indicator of the structural state of a protein [Scheich
et al., 2004, Peti et al., 2004] and offers the advantage of not requiring isotopical labeling of the sample.

Multidimensional NMR experiments have been introduced to reduce complexity of spectra. A large number of different types of multidimensional experiments have been designed and are used to identify individual nuclei and also to detect if and how different nuclei interact with each other in a molecule. One such experiment that has proven to be useful for rapid characterization of proteins is the $^{15}$N-HSQC experiment. In the resulting spectrum, one peak is obtained for each hydrogen-nitrogen bond, which corresponds to one peak per non-proline amino acid (the peptide bond) plus additional peaks from amide and amine containing side-chains. The key characteristic of the $^{15}$N-HSQC spectrum of a structured protein is the visualization of the large chemical shift differences, seen as a dispersion of the peaks over a broader range along the $^1$H axis. In a structured protein, the amides of the peptide bonds experience different chemical surroundings due to their strict positioning in the structure. They are influenced not only of the covalently bonded neighboring atoms but also non-covalent atoms brought into close proximity by the

![Figure 3.3: Examples of $^{15}$N-HSQC spectra for a structured (A) and an unstructured (B) protein. Typical side-chain peaks are encircled and the 7.5 to 8.5 ppm $^1$H chemical shift region is indicated by an arrow.](image)

The spectrum in (A) was recorded directly on a clarified lysate from a 5 mL expression culture of the GB1-TCTP fusion protein; only the peaks corresponding to the GB1 domain are visible however. The spectrum in (B) was recorded on a fragment of the yeast protein Rox3 encompassing amino-acids 100-220. The low dispersion and the broad peaks indicate aggregation of unfolded protein. 

In an unstructured protein on the other hand, the conformation of the backbone is more flexible and the amides of the peptide bonds are mainly influenced only by their covalent neighbors resulting in a similar chemical environment and a low dispersion of peaks in the spectrum (Fig. 3.3). However, small
α-helical proteins often have a limited dispersion even though they are structured and their spectra can thus be miss-interpreted. The number of peaks, not including the easily identified side-chain peaks, should correspond to the number of residues in the protein. If this is not the case it can indicate problems with protein degradation or unstructured regions in the case of too few peaks or multiple stable conformations in the case of too many peaks. Additionally, peaks should be sharp and of equal intensity for structured and monomeric proteins, whereas peak broadening or loss of intensity can indicate multimerization or aggregation.

An advantage of this exploratory experiment is the comparably high sensitivity; allowing spectra to be recorded at lower sample concentrations or for larger proteins than are usually used for structural studies by NMR. The utility of 15N-HSQC as part of large-scale SG projects has been highlighted in a number of studies [Folkers, 2004, Maxwell, 2003, Yee, 2002, Tyler, 2005]. It should be pointed out, however, that a poor 15N-HSQC spectrum does not necessarily mean that the structure of the protein cannot be solved by X-ray crystallography. Likewise, proteins that are unable to form crystals can still be suitable for structure determination by NMR [Savchenko et al., 2003].

3.9.3.2 By CD

In circular dichroism spectroscopy, the difference in absorbance for left- and right-handed circular polarized light is measured. Only electronic transitions in optically active molecules give rise to this phenomenon. The principal chromophore for CD measurements on proteins is the peptide bond, which has a strong absorbance at 195 nm. The peptide bond, however, is symmetric and the CD effect can only be observed because the interaction of the transition dipole of the chromophore with other neighboring transition dipoles constituting an asymmetric unit. Thus the strength of the CD signal will depend on the relative orientation of the peptide bonds and this will give structural information.

For α-helical structure elements, the peptide absorption at 195 nm is split into different components with characteristic minima in CD signal at 222 and 208 nm and a maximum at 190 nm, random coil structures have a strong negative CD signal at 195 nm while β-strands have a minimum at 215 and a maximum at 198 nm (Fig 3.4). Due to the characteristic CD signals for the different secondary structure elements, their relative contribution to the CD spectrum for a given protein can be calculated.
Hence the CD spectrum of a protein can be used to calculate the content of the different secondary structure elements with good accuracy. Additionally, CD can be used to obtain melting curves of proteins to assess their stability and the cooperativity of unfolding. This is done by monitoring the change in CD signal as a function of temperature at a wavelength with a strong CD signal corresponding to secondary structure.
4 Present Investigation

4.1 Experimental set-up

4.1.1 High throughput methods for cloning, expression, and purification.

A major objective of this thesis has been to develop and establish parallel and semi-automated methods for most of the steps in the protein production chain ranging from initial cloning of the gene to purification of the produced protein. These methods have been used in the publications included in this thesis (I-IV). They are based on the 96-well plate format, which have been used for decades in different fields of molecular biology but only recently have been adapted for protein production purposes. All cloning of genes have been done by site-specific recombination using the Gateway system. Cell cultures have been grown in 96- or 24-well plates. Purification has been performed in 96 well filter plates. The exception is the study in paper IV that uses predominantly intermediate-scale protocols for cell cultivation and purification in gravity flow columns. A second objective has been to use these methods to evaluate different parameters affecting the production yield of recombinant protein in E. coli. The aim has been to use these evaluated parameters to build a sparse matrix of conditions that would maximize the chance of obtaining recombinant protein in a soluble form while minimizing the number of experiments.

A majority of the experimental steps have been adapted to the 96-well plate format and readily accommodate a high degree of parallel handling. However, some bottlenecks remain that still slow down the process. Running gels are still a time and labor consuming step, especially when very large number of samples are analyzed, e.g., the study in Paper IV involved approximately 2000 gel samples. Recent development of equipment, like multi-pipettes with adjustable channel spacing, pre-cast agarose and SDS-PAGE gels with extended shelf lives, electrophoresis equipment for multiple gels in parallel, and gels in the 96-well format have allowed for a higher degree of parallelism and made sample transfer easier. Plating of cell cultures on agar plates is another step that is not readily automated, although there are solutions that make it faster than the traditional plating techniques [Mehlin et al., 2004]. The use of
glass beads instead of loops or glass-rods allow multiple plates to be processed in parallel. Selection without plating on agar as described in paper I or by Dieckman et al. [Dieckman et al., 2002] is also an alternative.

### 4.1.2 Presentation of projects

In all of the studies included in this thesis, human genes encoding proteins of unknown structure involved in cancer or disease have been used as target proteins. The study described in paper I, included 32 genes as a test set to evaluate the success rates for different steps in the protein production chain. All steps were adapted to the 96-well plate format. A screen of seven different expression vectors was also used to evaluate different fusion partners and the overall utility of small-scale expression screening at this step. Six of the expression vectors encode different N-terminal fusion proteins (GST, thioredoxin, ZZ, MBP, GB1 and NusA) and one encodes only an N-terminal His tag. The different expression vectors and the fusions they encode are presented in Table 3. The expression levels and solubility of the produced proteins were evaluated using SDS-PAGE by comparing soluble and insoluble fractions after cell lysis.

Paper II describes a study of different expression vectors used to produce His-tagged proteins and how they affect the expression levels and solubility of the produced proteins. It involved 20 genes in five different constructs produced from four different expression vectors (Table 3). Two of the vectors produce N-terminal His tag fusions, one produces a C-terminal His tag fusion or protein without C-terminal fusion when a stop codon is included in the gene sequence, and one produces an N-terminal solubilising fusion protein followed by a His tag. Expression levels and solubility were quantified using digitalized gel-pictures and yields were compared between the different vectors. Native and denaturing IMAC purifications were also performed to compare purification yields for the different His-tag constructs.

Paper III describes a study of how solubilising fusion partners affect the yield of purified target protein, i.e. after removal of the fusion partner by proteolysis. In this study a set of 45 genes that were cloned into four different expression vectors (Table 3) were used. Three of the vectors encode different N-terminal fusion partners (GST, thioredoxin, and GB1) followed by a protease cleavage site and a His tag whilst the fourth vector encodes only the N-terminal His tag. The vectors are otherwise
identical and they leave the same amino acids attached to the target protein after removal of the fusion protein as when produced with only the N-terminal His tag. After removal of the fusion partner and IMAC purification of the target proteins, yields of purified protein were compared between the different expression vectors. Expression levels and solubility before purification were quantified from digitalized protein gel pictures and compared to the amounts of purified protein.

Table 3: GATEWAY-adapted expression vectors included in this thesis. All encoded fusions are N-terminal except for the pTH24 vector.

<table>
<thead>
<tr>
<th>Vector</th>
<th>Promoter</th>
<th>Fusion</th>
<th>Parental vector</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>pTH1</td>
<td>tac</td>
<td>MBP</td>
<td>pMAL-c2</td>
<td>I</td>
</tr>
<tr>
<td>pTH2</td>
<td>T7lac</td>
<td>NusA</td>
<td>pET-43a</td>
<td>I</td>
</tr>
<tr>
<td>pTH3</td>
<td>T7lac</td>
<td>GB1</td>
<td>pGB1 (pET-21aVIII)</td>
<td>I, II</td>
</tr>
<tr>
<td>pTH5</td>
<td>T7lac</td>
<td>ZZ II</td>
<td>pT7-ZZa (pET-21aVIII)</td>
<td>I</td>
</tr>
<tr>
<td>pDEST15</td>
<td>T7</td>
<td>GST</td>
<td>pET-3aVIII</td>
<td>I</td>
</tr>
<tr>
<td>pDEST16</td>
<td>T7</td>
<td>Thioredoxin</td>
<td>pET-3aVIII</td>
<td>I</td>
</tr>
<tr>
<td>pDEST17</td>
<td>T7</td>
<td>His tag (MSYY...) III</td>
<td>pET-3aVIII</td>
<td>I, III</td>
</tr>
<tr>
<td>pTH6</td>
<td>T7</td>
<td>GST</td>
<td>pDESt15</td>
<td>II</td>
</tr>
<tr>
<td>pTH10</td>
<td>T7lac</td>
<td>Z II</td>
<td>pT7-ZZa (pET-21aVIII)</td>
<td>II</td>
</tr>
<tr>
<td>pTH19</td>
<td>T7lac</td>
<td>His tag (MGSS...) III</td>
<td>pET-15b</td>
<td>III</td>
</tr>
<tr>
<td>pTH24</td>
<td>T7lac</td>
<td>His tag (C-terminal)IV</td>
<td>pET-DEST42a</td>
<td>III</td>
</tr>
<tr>
<td>pTH26</td>
<td>T7</td>
<td>434repressor + His tag</td>
<td>pT7-7 V, VI, VII</td>
<td>III</td>
</tr>
<tr>
<td>pTH27</td>
<td>T7lac</td>
<td>His tag (MGP...) III</td>
<td>pGB1 (pET-21aVIII)</td>
<td>IV</td>
</tr>
<tr>
<td>pTH28</td>
<td>T7lac</td>
<td>Thioredoxin + His tag</td>
<td>pGB1 (pET-21aVIII)</td>
<td>IV</td>
</tr>
<tr>
<td>pTH29</td>
<td>T7lac</td>
<td>GST + His tag</td>
<td>pGB1 (pET-21aVIII)</td>
<td>IV</td>
</tr>
<tr>
<td>pTH34</td>
<td>T7lac</td>
<td>GB1 + His tag</td>
<td>pGB1 (pET-21aVIII)</td>
<td>IV</td>
</tr>
</tbody>
</table>

1 This vector has a frameshift in the lacI gene. II Also includes the T7tag in front of the Z domain. III Amino-acid residues preceding the His tag. IV Also includes the V5 epitope in front of the His tag. V Ampicillin resistance gene in opposite direction relative to expression cassette. VI No T7 terminator. VII pUC-vector origin of replication. VIII The vector backbone of pET-21a but with other fusion coding sequences inserted as indicated.
The study presented in paper IV utilized a protocol for rapid production and purification of proteins at an intermediate scale (Fig. 4.1) and showed that this was a useful and feasible way to evaluate if proteins that were found to be soluble in an expression screen also were suitable for further structural studies. Standardized protocols were applied for three different N-terminal fusions that enable affinity purification. Most of the protocols are not performed in a multi-well plate format, but they still allow for multiple proteins to be produced in parallel. However, it was also demonstrated that purification and elution by proteolytic release in a 96 well format was possible. The protocols were used to evaluate 19 different human proteins by biophysical characterization.

**Figure 4.1:** Schematic representation of the intermediate scale purification. 50 ml cultures in $^{15}$N-enriched minimal medium were grown in shake-flasks at 37° C. Cultures were harvested by centrifugation after two hours of expression. The resuspended bacterial cells were disrupted by sonication and cleared by centrifugation. The clarified lysate was applied to affinity resin, either in gravity flow columns or in 96 deep-well plates depending on the number of samples. After washing, protease was added in a suitable buffer, incubated for 1-2 hours and the released target proteins were collected by centrifugation. If necessary, the eluted protein was concentrated prior to biophysical characterization.
4.1.3 Note on statistical analysis

Paper I, II, and III all involve comparisons of protein yields for large sets of proteins when produced under different conditions. Typically, there were large variations in yield among the different proteins, i.e. some proteins were consistently produced at very low levels whilst others were produced at high levels. Thus the standard deviation, describing the variation within the set of proteins produced under a given condition, was usually high and a direct comparison of only average yield and standard deviation between different conditions was not always informative (Fig. 4.2 A and B).

![Graphs A, B, C, D]

**Figure 4.2:** Illustration of pair-wise comparisons. Eight different proteins are tested under two different production conditions, light gray, and dark gray. Averages and standard deviation are given to the right by columns and error bars, respectively. (A) An example where there is a consistent difference between the two conditions. Due to the large variations in yield among the different proteins within each set, the consistent difference is not evident when averages and standard deviation are compared for the two sets. (B) An example where the difference between the two conditions is not consistent. The averages and the standard deviations are identical to those of the example in (A). (C) and (D) The differences between the two conditions in (A) and (B), respectively. In these cases, the averages and standard deviation are highly informative.

Instead pair-wise comparisons between the different conditions were used (Fig. 4.2 C and D). In this case the difference in yield for each protein when produced under the
different conditions was calculated. The average and standard deviation for the differences for all proteins within the sets were then used to test if there was a difference and if it was statistically significant at a given level of confidence. This method is referred to as Student’s t-test, named after W. S. Gosset who used Student as a pseudonym. However, as the number of tested conditions increases, the number of pair-wise comparisons increases even faster. To keep figures or tables lucid, it is often nevertheless necessary to include averages rather than average differences. In that case, it should be clearly stated which differences are statistically significant. As an example, it would take 21 entries in a table or figure to describe the differences among the seven expression vectors tested in paper I.

4.2 Expression screening

4.2.1 Screening for soluble expression

4.2.1.1 Paper I

Expression screening has been a central component in several of the included studies. In the study presented in paper I seven expression vectors (Table 3) were screened, all encoding different N-terminal fusions partners: His tag, GST, NusA, ZZ, Gb1, MBP and thioredoxin. All but the His tag are regularly used as solubilising fusion partners to increase the yield of soluble protein, but at the time no comprehensive comparison of the effect of the different fusions existed. 188 fusion constructs were expressed and a qualitative annotation of expression levels and fraction soluble protein based on protein gels was made. Large variations were found between the different vectors, both in the number of proteins that were produced and the levels of produced proteins (Fig. 4.3). Generally, the His tag and the GST fusion vectors performed poorly in comparison to the GB1, thioredoxin, and MBP fusion vectors. Pair-wise comparisons between the different vectors based on the amount and solubility of each produced protein allowed for a ranking of the fusion proteins for their effect on solubility to be made. From the results, a ranking in decreasing order would be: Thioredoxin, MBP or GB1, NusA or ZZ, GST, and finally the His tag which had the poorest performance of the tested fusions. One important conclusion was that no single expression vector could account for all the successfully produced
Figure 4.3: Results of expression screening (paper I). The vectors are designated by the N-terminal fusion they encode. The number of proteins that are at least partly soluble (light gray) and totally insoluble (dark gray) is shown. In addition, the number of proteins that were soluble in any of the seven different vectors and all of the seven vectors are included.

proteins, only by combining the result from at least three different expression vectors could the high overall success-rate for the expression step be reached (Fig. 4.3). Thus it is advantageous to screen different expression vectors in order to increase the number of proteins that can be produced. It should also be noted that there were variations in the amounts of soluble protein depending on the fusion-target combination, thus it was not always the same vector that produced the highest amount of soluble protein. This implies that screening can be beneficial also in order to maximize the yield for a specific protein.

4.2.1.2 Paper II

The study presented in paper II focused on screening different His tag constructs. The His tag is probably the most widely used affinity tag and there are numerous different expression vectors that include this tag for purification. Most, if not all, SG projects rely on the His tag for initial purification. The principle in this study was to make a comparison between different fusion constructs with the His tag at different
locations relative to the target proteins to see if this would affect the yield. This time a quantitative analysis of the expression levels based on band intensities on SDS-PAGE protein gels were used. Large variations in both expression levels and solubility of the produced proteins were found also for these constructs (Fig. 4.4).

**Figure 4.4:** Average amounts and fraction soluble protein produced from the different expression vectors. (A) Total amount (B) Fraction soluble protein (C) Total amount divided by dry cell weight (D) Soluble amount divided by dry cell weight. pDEST17 and pTH19 encode N-terminal His tags, pTH26 encodes the 434-repressor protein as an N-terminal fusion partner followed by a His tag and pTH24 encodes a C-terminal His tag. pTH24 (Nat) represents genes with stop codons, thus without the C-terminal His-tag.

Comparing total amounts of produced protein, the pTH24 vector (C-terminal His tag) was found to produce the highest amounts, followed by the pTH19 vector (N-terminal His tag), the pTH26 vector (434-repressor fusion) and finally the pDEST17 vector (N-terminal His tag) with almost four times lower amounts than the pTH24 vector. Comparing the fractions of soluble protein, a large difference between the two variants produced from the pTH24 vector was found, with the C-terminally His-tagged proteins having a lower solubility. There was also a large difference between the proteins that were produced with only an N-terminal His tag; In this case, the proteins expressed from the pTH19 vector had a higher solubility than the proteins expressed from the pDEST17 vector. The 434-repressor fusion proteins that have the His tag included in the linker also had a high solubility, comparable to the fusion
proteins produced from the pTH24Nat and pTH19 vectors. Importantly, the different vectors were found to strongly influence growth rates for the cell cultures (Fig 4.5A). When this was taken into account and the total amounts of produced protein per dry cell weight were compared, no major differences between the different vectors were observed (Fig 4.4 C). However, comparing the amounts of soluble protein per dry cell weight, the pDEST17 vector gave yields lower than the pTH19, -26 and -24 vectors (Fig 4.4 D). Thus, two factors that strongly influenced the yields of protein for the different vectors used in this study were found: Firstly, the effect on growth rates which directly influences total amount of protein and secondly, the differences in solubility of the produced proteins which may depend on the location of the His tag, the additional fusion components encoded by the vectors, or differences in vector design.

**Figure 4.5:** The effect of (A) vector and (B) target proteins on growth rate. (A) The average OD$_{600}$ for all targets cloned into respective vector plotted against time. Expression was induced at the time-point 2h. Cultures harboring the pTH24 vector reach at least twice as high OD$_{600}$ as the other vectors at harvest (=5h). (B) For each target protein, the relative deviation from the average of all targets in that vector was calculated. This was averaged over the four different vectors. Four target proteins have consistently lower than average final OD$_{600}$ and two proteins have consistently higher than average final OD$_{600}$.

This study clearly demonstrated the importance of making appropriate comparisons, in this case to compare actual levels of soluble protein per cell weight rather than just fraction soluble protein. It also emphasized that expression results for His tagged proteins differ considerably depending on the expression vectors being used. The pDEST17 vector gave the poorest results of the tested vectors, which may explain
the poor results for the His-tagged proteins produced from this vector in Paper I and also in some other studies [Braun et al., 2002, Huang et al., 2003].

4.2.1.3 Paper III

Although the focus of paper III is on comparing final purification yields for proteins produced with different N-terminal fusion partners that are cleaved of before purification, the study also includes a quantitative analysis of expression levels and solubility of the produced proteins before purification. As all the vectors in this study have the same vector backbone, no differences in growth rates were observed between the vectors. Large differences were observed when the expression levels of the different fusion constructs were compared, with the thioredoxin and GB1 domain fusions having higher expression levels than the GST fusions or the His tagged proteins (Fig 4.6 A).

![Figure 4.6](image)

**Figure 4.6**: (A) Average amount, (B) fraction soluble protein, and (C) amount of soluble protein produced from the different vectors in paper III. The different vectors are denoted by the fusion they encode. The light gray bars in (C) indicate the contributions of the fusion partners to the measured gel band intensities. The calculated amounts of soluble target protein, excluding the fusion partners, are indicated by the same colors as in (A) and (B).
This can probably be accounted for by differences in translational efficiency of the fusion coding sequences, given that they are the only differing components between the vectors. Comparing the fraction soluble protein produced from the different vectors, no large differences as seen in the earlier studies were observed. Only the GB1 fusions had an average solubility that was higher than the other fusions (Fig 4.6 B). The His tag constructs produced from the vector pTH24 had a higher solubility than the constructs produced from the pDEST17 vector in paper I and paper II but similar to that of the construct produced from the pTH19 vector in paper II. Comparing the amounts of soluble target protein, not taking into account the contribution of the fusion partners, the GB1 fusion was found to give the highest yield, followed by the thioredoxin fusion and the His tag, and finally the GST fusion with the lowest yield (Fig 4.6 C).

As seen in all these examples were an expression screen have been applied, there can be large variations in expression and solubility for individual proteins depending on the expression constructs used. In none of the screens can a single expression vector account for all the successfully produced proteins. From the results of paper II and III it is clear that a short affinity tag, in this case a His tag, can be used to produce a large number of proteins in a soluble form, provided the expression vector supports production of high levels of soluble protein as was the case for the pTH19 and pTH27 vectors but not for the pDEST17 vector. This is advantageous since it allows structural characterization, or even structure determination, without removal of the fusion. It is even likely that the production yields can be improved further by optimization of the coding and non-coding sequence of the fusion partner to improve translational efficiency as discussed in the introduction. An example could be the introduction of the first six codons from the E. coli thioredoxin gene in front of the N-terminal his tag as used by Lesley et. al. [Lesley, 2001]. Further optimization can also be done by directed evolution [Zhelyabovskaya et al., 2004]. The benefit of such optimization of a general N-terminal tag that can be put in front of all different solubilising fusion partners or directly in front of the target proteins could be substantial.

However, not all proteins can be produced by in a soluble form with only a short N-terminal tag (paper I and III) and thus it is beneficial to include a set of solubilizing fusion tags as alternatives. The GB1 domain is a suitable choice, given its small size, good solubility, and the possibility to use it for affinity purification. It is probably
beneficial to include also a larger fusion protein, given that there can be a size
dependence of the solubilising effect (paper III) and for instance GST, MBP or NusA
could be suitable choices in conjugation with a small tag for affinity purification
[Pryor and Leiting, 1997]. The performance of the GST fusion in paper III was not
electing but this could perhaps be improved by optimization of the fusion coding
sequence. However, there are examples where the GST protein has been successfully
used as a fusion partner, for instance by Braun et al. [Braun et al., 2002].

It is clear from papers II and III that it is more informative to compare actual
levels of soluble protein rather than fraction soluble, as there can be large variations
in total protein production levels depending on other vector elements (see below).
The size of the solubilising fusion partner should also be taken into account when
comparing the yields of soluble protein, as large fusion partner would be removed
before structural studies and they can constitute a large proportion of the size of the
fusion protein.

4.2.2 Effect on yields of purified protein

In the study presented in paper III the use of solubilising fusion partners and if they
affect the yield of purified target protein after removal of the fusion partner was
studied. Although it has been shown that fusion partners can have a beneficial
influence on solubility of the target protein when produced in E. coli, there are
numerous examples of proteins that become insoluble when the fusion partner is
removed. Thus a larger set of proteins was used to test whether the solubilising effect
of fusion partners is a transient effect or if it will positively influence the yields of
purified target protein. However, as stated in the previous section, the solubilising
effect in E. coli was found to be less pronounced than in the study presented in paper
I. Only the GB1 domain fusions had an average solubility that was higher than any of
the other fusion partners included in the study. Importantly, this was also reflected in
the yields of purified target protein (Fig. 4.7) were the GB1 fusion constructs gave an
average 25 % higher yield than the thioredoxin or His tag fusion constructs. It was
concluded that the solubilising effect of fusion partners in general have a positive
influence on the yield of purified protein. The solubilising effect was crucial for
successful purification of some proteins, whereas in most cases it was manifested as a
mere difference in yield. In some additional cases the solubilising effect is only
apparent and the target protein could not be purified after removal of the solubilising fusion tag.

**Figure 4.7**: Comparison of (A) expression data and (B) purification yield. (A) The average amount of soluble target protein, excluding the contribution of the fusion partner, calculated from expression data. (B) Average amount of purified target protein. The scales are not comparable between (A) and (B) due to differences in sample preparation and normalization, but the relative intensities can be compared.

Also the study presented in paper II included purification of target proteins under both native and denaturing conditions. In both cases the average purification yields for the different expression constructs correlated well with the expression data. When the data sets were normalized with respect to the differences in cell density resulting from the expression vectors that were used, as discussed in section 4.2.1.2, it was clearly seen that solubility of the produced proteins was the key determinant of purification yields.

### 4.2.3 Vector design

Although much of the focus of this thesis is on the use of different fusion proteins or tags, there are several other elements of the expression vector that can influence the yield of protein. Different promoters are included to a limited extent in the studies (Table 3), but mainly the T7 promoters [Tabor and Richardson, 1985, Studier et al., 1990] with or without the lac operator have been used. The primary effect of the promoter is the rate of transcription after induction, but also the suppression of transcription before induction. The T7 promoter allows for very high transcription rates and the fraction of recombinant protein can reach up to 50% of total protein. Without the proper suppression of pre-induction transcription this can give severe problems with differentiation of the cell culture into expressing and non-expressing populations where the latter rapidly will outgrow the former as discussed in the introduction. Thus it is advantageous to include additional elements that will suppress
“leaky” expression, like the lac operator in conjunction with lac repressor protein or co-expression of T7 lysozyme [Studier et al., 1990].

In both study II and III, co-expression of T7 lysozyme was used to minimize leaky expression. The effect of the lac repressor was clearly demonstrated in paper II, where growth curves were measured for all expression cultures. The target proteins that were expressed had a clear influence on the growth rate of the expression culture that was observable regardless of the expression vector used (Fig 4.5 B). This effect was observable already before induction for the vectors lacking the lac repressor system, despite presence of the T7 lysozyme repression system (Fig 4.8). From the growth curves it was also possible to predict that the lac repressor system for the pTH19 vector was not functional, and this was subsequently confirmed as sequencing of the lacI gene in that vector revealed a frame-shift that destroyed the tetramerization motif of the lac repressor protein.

![Graphs showing growth curves for individual vector-target combinations.](image)

**Figure 4.8:** Growth curves for individual vector-target combinations. Expression was induced at the time-point 2 hours. There is no target dependant variation before induction for the pTH24 vector, whereas this is clearly observable for the three other vectors.

It is clear that there can be severe negative effects on growth rates depending on the expressed gene and also that if the pre-induction expression is not suppressed this will have a large influence on the final yields of protein. It was demonstrated that T7 lysozyme alone is unable to suppress pre-induction expression from the T7 promoter.
without lac operator. The problem with pre-induction expression have probably also influenced the result in paper I, in which some of the vectors lacked lac operator and co-expression of T7 lysozyme was not used. However, the influence of this pre-induction expression is unclear as the three vectors lacking the lac repressor performed dissimilarly with the pDEST-15 (GST) and pDEST-17 (His tag) performing poorly, whereas the pDEST-16 (thioredoxin) vector was the best performing vector in that study.

As almost all of the expression vectors used in these studies are based on the backbones of the commonly used series of pET-vectors that in turn are based on the pBR322 plasmid, they are identical with respect to origin of replication. There are two vectors that are not pET-based, namely pTH1 and pTH26. Of these, pTH1 is also based on the pBR322 plasmid, whereas the pTH26 vector is based on the closely related pUC plasmid, which has a higher copy number. From the data it is not possible to see any influence of this.

### 4.2.4 Quality control

One important aspect of the downscaled protocols and increased throughput that cannot be neglected is the demands on quality control. Proteomics projects tend to lend the high throughput terminology from the pharmaceutical industry, although working at a pace and scale that is far from that of the original. After an era of ever increasing library sizes in the high throughput screening labs of the pharmaceutical companies, critical evaluation of sample quality indicated that a lot of effort was spent on screening incorrect compounds. An important lesson to be incorporated into the SG projects is to ensure thorough quality control throughout the screening and production processes; for example DNA sequencing of expression constructs, confirmation of dot-blot results by ordinary SDS-PAGE and homogeneity control of purified protein by mass spectrometry and other biophysical methods. Detection of the mutated lac repressor in paper II by evaluation of growth curves is a good example of an experimental error that would probably have went unnoticed without proper quality control.

A high throughput approach does not justify high rates of false negatives or false positives. In expression screening experiments, false negatives can be the result of mutations in the expression constructs, plasmid instability, impaired expression caused by leaky expression or variations in expression depending on uncontrolled
Experimental parameters like aeration of cultures. Variations in expression results can often be encountered and the reproducibility should be confirmed by replicate experiments and taken into account when evaluating the results. Moderate variations were observed in the results when performing the experiments in triplicate (papers II and III), probably due to the combined experimental uncertainties of all the experimental procedures from expression cultures to measured result. Similar variations were reported in a comparison of small-scale purification protocols [Scheich et al., 2003]. False positives in expression screening experiments can be tolerated to a certain degree, as long as they do not result in too much effort being spent on scaling up of unsuccessful target proteins. Examples of false positives could be the apparent solubility of some fusion proteins or differences arising from different conditions used in the small scale screening experiment and the up-scaled experiments. In an example of comparison between small-scale and large-scale results, it was concluded that 80% of the positive in small-scale experiments could also be scaled up. However, it was also shown that as much as 40% of the negative results in small-scale experiments was actually positive in large-scale experiments [Sugar et al., 2005].

4.3 Rapid purification and biophysical characterization

To fully benefit from the results of screening different fusion protein constructs, it is desirable to directly be able to scale up and purify the fusion protein that has the highest amount of soluble protein without any further sub-cloning. Thus it is advantageous if the solubilising fusion partner also can be used for affinity purification. Three different solubilising fusion partners that also could be used for affinity purification were used to establish a protocol for rapid purification and biophysical characterization of different proteins (paper IV) at an intermediate scale (Fig. 4.1). Proteolysis was used to separate the target protein from its fusion partner and elute it from the affinity resin. Separating the target protein from its fusion partner was necessary for the subsequent characterization and it also allowed for more freedom in the choice of elution buffer. Especially the interactions between the GB1 or the Z domain and IgG-resin are strong and normally require harsh elution conditions that can be avoided using proteolysis. The purification protocol was also adapted to the microtiter plate format, which had previously not been done for the proteolysis step.
As structure determination of proteins can be a time consuming process it is desirable to get an early indication of how suitable a protein is for structural studies. As more and more proteins have been studied it has become increasingly clear that not all proteins are fully structured when isolated or even in their native form. In fact it is predicted that substantial fractions of the eukaryotic proteomes are at least in part unstructured [Dunker et al., 2001]. As a consequence of this, it is informative to get structural data as early as possible. In the study presented in paper IV the combination of far-UV CD spectroscopy and \(^{15}\text{N}-\text{HSQC}\) NMR spectroscopy was used to assess the secondary and tertiary structure content of the target proteins (Fig. 4.9). The dispersion of peaks along the \(^{1}\text{H}\) axis is an excellent indication of foldedness of a protein in the \(^{15}\text{N}-\text{HSQC}\) spectrum. Also the number and uniformity of intensity of the peaks in this spectrum give information on the state of the protein.

The far-UV CD spectrum gives information on the relative amounts of different secondary structure elements present in the protein. As a complement to the CD and NMR spectroscopy methods, ANS fluorescence was also used to further characterize proteins that have secondary structure but low tertiary structure content, key characteristics of proteins in the molten globule state. It has been shown that ANS binds to proteins in the molten globule state, which induces an increase in ANS fluorescence intensity and also a shift in wavelength for the emission maximum [Semisomov et al., 1991].

Examples of spectra are shown in Figure 4.9. The TCTP protein was classified as folded based on the HSQC spectrum and the CD spectrum indicated a mixed \(\alpha\) and \(\beta\) structure, with a large proportion of random coil. This was subsequently confirmed by the solution structure of this protein (Fig. 2.2). The IPKA protein had a HSQC spectrum with strong uniform peaks with low dispersion and a CD spectrum indicating lack of secondary structure. This protein was classified as unstructured. The HSQC spectrum for the HBPI protein showed low dispersion but variations in intensity. The CD spectrum indicated almost exclusively \(\alpha\)-helix secondary structure. The ANS fluorescence spectrum showed high fluorescence intensity indicating presence of hydrophobic pockets. The combination of low dispersion in the HSQC spectrum, strong secondary structure signal, and ANS binding classified this protein as a molten globule. However, this classification can be uncertain in some cases as was highlighted for this protein. Further biophysical studies by Tai et al. indicated
that the HBP1 protein in fact is a coiled-coil trimer with slow motions on the millisecond scale [Tai et al., 2002].

**Figure 4.9:** Examples of how proteins can be classified based on biophysical characterization. The first column shows purity and concentration of the proteins as seen by SDS-PAGE protein gels. In the second column far-UV spectra are shown, in the third column the $^{15}$N HSQC spectra, and in the last column ANS fluorescence emission spectra; gray spectra for ANS only and black for ANS and protein. The proteins shown, TCTP, IPKA, and HBP1, were classified as folded, unfolded, and molten globule based on their spectra as discussed in the text.

It was demonstrated that this was a feasible approach to get rapid information on the structural status of the proteins when produced on an intermediate scale. The results were reproducible independently of the different fusion and protease systems used. Of the 19 proteins included in this study, four were classified as structured; one as molten globule; twelve as unstructured or severely aggregated and two precipitated
during purification. These numbers clearly demonstrate the benefits of a rapid screen to identify proteins suitable for further structural studies.
5 Conclusions and perspectives

The work reported in this thesis focuses mainly on screening protein production parameters to increase the success-rate and yield of proteins intended for structural studies. Efficient protocols for multiple and parallel cloning of human genes have been designed using the Gateway cloning system. Similar approaches are now widely used in a number of different structural genomics projects [Albeck et al., 2005, Chance et al., 2002, Ding et al., 2002, Thao et al., 2004, Vincentelli et al., 2003, Busso et al., 2005b]. The protocols have been used to systematically evaluate different parameters influencing production of recombinant protein in *E. coli* based on large numbers of small-scale expression and purification experiments performed in a multi-well plate format.

The benefit of fusion proteins as a way of increasing the success-rate for production of soluble protein in *E. coli* has been demonstrated by comparing the results obtained with different fusion partners when applied to larger sets of target proteins. Several other studies have arrived at similar conclusions although there are some differences in the extent of solubilization obtained for the different fusion partners used [Braun et al., 2002, Dyson et al., 2004, Korf et al., 2005, Shih et al., 2002]. These differences can in part be contributed to differences in experimental design, the type of target proteins used, and how the results are evaluated. However, there seem to be a broad consensus on the beneficial effect on solubility when using the maltose binding protein as a fusion partner, and nearly as broad consensus for the NusA protein. Both these fusion partners have also been used in some large-scale SG projects [Vincentelli et al., 2003, Jeon et al., 2005]. Similarly, the GST protein is often found to be comparably ineffective in promoting solubility in these large systematic comparisons. For the thioredoxin protein the results are somewhat divergent, probably depending on the type of target protein being used in each study. This fusion partner seems better suited for smaller target proteins. Although not yet widely used, the work in this thesis supports the GB1 domain as a good alternative, especially for smaller target proteins given the small size of this fusion partner as compared to the large MBP and NusA fusion partners.

A major concern with solubilising fusion partners is that the solubility enhancement effect is only apparent and that the target protein will become insoluble
once the fusion partner is removed. This has been addressed in one of the studies by comparing the results from expression experiments to the yield of purified target protein after proteolytic removal of the fusion partner. On average, there was a good agreement between expression and purification data although some target proteins that showed low levels of soluble protein in the expression experiments failed to yield purified protein. The difference in yield of purified and isolated target protein between proteins produced with or without a solubilising fusion partner was substantial for a few specific target proteins but modest when averaged over the entire set. This is in agreement with the findings of several others; that although the use of fusion partners can increase the success-rate of obtaining specific target proteins, the use of only a his-tag to facilitate purification still allows a substantial number of target proteins to be produced in quantities required for structural studies. However, there are only a few reports of systematic use of larger fusion partners for large-scale protein production involving a larger number of target proteins whereas the use of only a his-tag is widespread, so the full benefit of solubilising fusion proteins remains yet to be seen.

The widespread use of the his-tag prompted for a comparison of different his-tag constructs to evaluate how parameters like fusion to the N- or C-terminal of the target proteins, use of different vector backbones or transcriptional control elements, and the addition of additional fusion elements affected protein production yields. As an example, a clear difference in yield was seen between proteins produced from the same vector with or without a C-terminal his-tag, with the latter showing lower solubility. Additionally, a difference in solubility of the produced proteins was seen for two different expression vectors both producing N-terminally his-tagged proteins. Others have also studied the impact of his tag location and found that it can indeed influence the yield of soluble protein [Busso et al., 2004, Dyson et al., 2004]. Large differences in yield of protein depending on differences in cell growth for cell cultures harboring different expression vectors were also seen and could be correlated with the presence or absence of tight regulation of pre-induction expression. This study highlighted the importance of having identical vector backbones when comparing different fusions or tags, which is now becoming standard in more recent publications [Korf et al., 2005, Busso et al., 2005a, Dyson et al., 2004], including paper III (above).
It is clear from all of the current large-scale SG projects that observing soluble expression in *E. coli* is no guarantee for obtaining purified protein in quantities and concentrations enough to enable structural studies and even when that is achieved there is no guarantee that the protein sample will enable structure determination (Table 1). Thus it can be beneficial to include a biophysical screening step at an intermediate production scale before passing the protein on to the large-scale production pipeline in order to optimize usage of production capacity. Paper IV describes rapid small-scale purification protocols from 50 mL cultures in order to obtain $^{15}$N labeled samples for $^{15}$N-HSQC NMR experiments, allowing a fast structural characterization of the proteins. A quality assessment of the protein based on the $^{13}$N-HSQC experiment is a natural checkpoint before deciding whether or not to produce the more expensive $^{13}$C labeled samples usually required for full structure determination by NMR. This experimental checkpoint is usually used, and reported, by most SG projects employing NMR methodology as a measure of success-rate on the route to structure determination [Tyler et al., 2005, Yee et al., 2002, Folkers et al., 2004], although production is usually on the scale of 1L cultures. To better utilize the results from expression screening, different affinity purification protocols were used depending on the N-terminal fusion partner and a proteolysis step was included to remove the fusion partner. As many as twelve of the 19 proteins included in this study were found to be unstructured. The relatively high proportion of unstructured proteins among the target proteins used can in part be due to the selection criteria of small, disease related proteins that do not belong to large Pfam families but still highlights the fact that a substantial subset of especially eukaryotic proteins contain a high degree of disordered regions or are unstructured [Uversky et al., 2000, Dyson and Wright, 2002, Dunker et al., 2001].

Finding the right balance between number of expression conditions to screen and the number of proteins to include is an elaborate task. It is clear from the large-scale projects reported so far, that a fairly high success-rate can be achieved with a single condition approach, typically expression in *E. coli* with a His tag. However, there is a risk for redundancy in the type of proteins that can be produced in this way, e.g. intracellular enzymes. It is also clear that as focus is shifting towards eukaryotic proteins, the success-rate will be reduced and hence it is necessary to include alternative approaches. This thesis covers the utility of solubilising fusion tags as an alternative to increase the number of proteins that can be successfully produced in *E.
coli. Others have explored the possibilities offered by alternative production hosts, co-
expression of folding mediators in E. coli or re-solubilization from inclusion bodies.

I think that several developments are foreseeable and already under way. (1) Production approaches adapted to specific sets of proteins are re-introduced in
combination with pre-expression selection of production system or tier-based
approaches; i.e. disulphide-containing proteins are expressed in E. coli strains
engineered to have an oxidizing cytoplasm, membrane associated proteins in strains
engineered to overproduce membranes etc. The problem with such an approach, of
course, is that for many proteins we lack the information to be able to predict what
will a suitable choice of production condition. Results from the ongoing SG projects
may help in clustering proteins for optimized production conditions. (2) A shift in
focus from high throughput towards high output with increased emphasis on
improved quality control and salvaging of promising target proteins [Lesley and
Wilson, 2005, Liu et al., 2005]. (3) Screening of several different construct boundaries
for each target protein to exclude flexible regions that might interfere with proper
folding or structure determination [Dyson et al., 2004, Gao et al., 2005, Pantazatos et
al., 2004]. (4) Method development to be able to address challenging proteins in a
systematic way, i.e. membrane proteins and protein complexes [Eshaghi et al., 2005,
Kim et al., 2004a]. With the progress being made in methods and technology
developments, the SG-projects are now positioned to contribute 1000 new structures
annually.
6 References


de la Torre, J. C., Ortín, J., Domingo, E., Delamarter, J., Allet, B., Davies, J.,
Bertrand, K. P., Wray, L. V., Jr. and Reznikoff, W. S. (1984) Plasmid, 12, 103-
49.

Struct Funct Genomics, 5, 133-46.


Ding, H. T., Ren, H., Chen, Q., Fang, G., Li, L. F., Li, R., Wang, Z., Jia, X. Y., Liang,

Doyle, S. A., Murphy, M. B., Massi, J. M. and Richardson, P. M. (2002) J Proteome Res,
1, 531-6.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S.,
Oldfield, C. J., Campen, A. M., Radliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D.,
Chiu, W., Garner, E. C. and Obradovic, Z. (2001) J Mol Graph Model, 19, 26-
59.


BMC Biotechnol, 4, 32.


Eshaghi, S., Hedrén, M., Nasser, M. I., Hammarberg, T., Thornell, A. and Nordlund,


119-31.


Gao, X., Bäin, K., Bonanno, J. B., Buchanan, M., Hendersom, D., Lorimer, D., Marsh,
J Struct Funct Genomics, 6, 129-34.

Gentz, R., Certa, U., Takacs, B., Matile, H., Döbeli, H., Pink, R., Mackay, M., Bone,

Geoghegan, K. F., Dixon, H. B., Rosner, P. J., Hoth, L. R., Lanzetti, A. J., Borzilleri,
K. A., Marr, E. S., Pezzullo, L. H., Martin, L. B., LeMotte, P. K., McColl, A.


Jeon, W. B., Aceti, D. J., Bingman, C. A., Vojtik, F. C., Olson, A. C., Ellefson, J. M.,
McCombs, J. E., Sreenath, H. K., Blommel, P. G., Seder, K. D., Burns, B. T.,
Geetha, H. V., Harms, A. C., Sabat, G., Sussman, M. R., Fox, B. G. and

Biol, 320*, 1-10.


440-6.


Bacteriol, 182*, 7029-34.

Kigawa, T., Yabuki, T., Matsuda, N., Matsuda, T., Nakajima, R., Tanaka, A. and

Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T. and Yokoyama,


Kim, K. J., Kim, H. E., Lee, K. H., Han, W., Yi, M. J., Jeong, J. and Oh, B. H. (2004a)


Oganesyan, N., Degiovanni, A., Lou, Y., Henriquez, M., Huang, C., Jancarik,
J., Pufan, R., Choi, I. G., Chandonia, J. M., Hou, J., Gold, B., Yokota, H.,

Kim, Y., Dementieva, I., Zhou, M., Wu, R., Lezondra, L., Quartey, P., Joachimiak,
111-8.


Korf, U., Kohl, T., van der Zandt, H., Zahn, R., Schleger, S., Ueberle, B.,
Wandschneider, S., Bechtel, S., Schnölzer, M., Ottleben, H., Wiemann, S. and


Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J.,
Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D.,
Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R.,
McEwan, P., McKean, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris,
W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez,
C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D.,

References 61


References


