5. Discussion

The emerging field of functional proteomics requires HT methods to express and purify proteins. The availability of such methods will enable the application of biochemical methods to genome wide characterization of protein function and thereby fill a significant gap in the repertoire of proteomic technologies. *E. coli* is the most widely used, easiest and the most straightforward expression system. Therefore we aimed to develop HT protein expression and protein purification methods for Escherichia coli.

5.1. Technical Development

5.1.1. Protein Expression

The first step in developing a HT protein production using *Escherichia coli* as an expression system was the definition of robust protein expression and purification conditions that can be applied to many targets in parallel. After initial tests an expression temperature of 25°C was chosen, because this gave the highest success rate. This result is consistent with previously described advantages of lowering the expression temperature for selected proteins to increase their solubility and stability (Liao, 1991; Lin et al., 1990; Pradel and Schnaitman, 1991; Steczko et al., 1991). It is likely that these conditions decrease the yield for several proteins, which would profit from a higher expression temperature. However, this must be accepted as the goal of HT experiments is the identification of conditions that give best results for the majority of proteins. Other authors investigating HT protein expression have been growing their cultures at 37°C (Hammarstrom et al., 2002). We found, however, that at this temperature many GST-tagged proteins are getting degraded and consequently the success rate dropped significantly. The discrepancy may be explained by the small size of test set proteins in the other study. In addition, the ideal HT expression condition may vary for different fusion-tags and bacterial strains.

We found that tight repression of protein expression during expansion of the bacterial cultures is crucial for consistency of the protein production process. Leaky protein expression altered the growth rates of many cultures to a degree that growth phase

dependent induction of protein expression was impossible to time properly for all cultures.

5.1.2. Protein Purification under Non-Denaturing Conditions

The next step in developing HT protein expression was the development of a HT protein purification method for the expressed proteins. In order to avoid complications from a limited solubility of the expressed proteins, we set up the method using His_6-tagged proteins and denaturing purification conditions.

In early experiments we noticed a significant decrease in binding efficiency when the proteins were bound to the matrix in a single centrifugation step in which the total protein is loaded onto the column and unbound protein removed from the column. Others have suggested this approach, however, their process starts with 50-fold more material and is therefore hard to compare to the process developed in this study (Felleisen et al., 1996).

The total volume of the binding reaction was found to strongly affect binding efficiency in microscale protein purifications. This can be explained by simple mass action for the equilibrium, whereby the total amount of complex is inversely proportional to the total volume of the reaction ($[Cplx] \propto 1/vol$). Consequently, the smaller the volume the more complex can be recovered. However, in the developed application, a further decrease in lysis volume caused a disproportional increase of losses in subsequent transfer steps. Therefore we decided to do the lysis and binding reactions in 100µl total volume.

We found that a low g-force for the wash- and elution steps was important to successful purification under non-denaturing conditions, but optional under denaturing conditions. We speculate that the higher g-force caused dehydration of the beads, which denatured the proteins and consequently caused non-specific binding.

Using the developed protein purification method, we found that all proteins, which were detected in the whole bacteria by western blot, could be detected by SDS-PAGE followed

by coomassie blue staining after purification under denaturing conditions. This finding validates the basic purification method.

In order to evaluate the purification method on a protein set of proteomic scale we expressed and purified 728 different proteins from sequence verified cDNAs. We found that ~70% of these proteins could be purified under denaturing conditions using the His_6-tag. Denatured proteins can be used for antibody production, for the manufacture of analytical protein arrays or submitted to preparative refolding (Lilie et al., 1998).

5.1.3. Purification under Non-denaturing Conditions

In order to analyze the functions of proteins, it is important that the purified proteins are in their native conformations. Therefore, a crucial aspect of developing a HT purification method is the identification of conditions that enable the purification of a maximum number of proteins under non-denaturing conditions. Because protein affinity purification tags have been shown to exhibit a major influence on ability to purify individual proteins from E. coli, we decided to evaluate four different purification tags using a test set of 32 proteins.

Under non-denaturing conditions, we found that relatively few proteins could be purified using the His_6- and CBP-tags, which means that neither of the small tags was suited to parallel protein purification under non-denaturing conditions. Many early screens for soluble proteins in structural proteomics efforts have relied solely on the very small hexa-Histidine-tag (His_6-). Table 1 summarizes the cumulative data. Although most of these approaches did not involve high-throughput microscale purifications, the data are nonetheless instructive with regard to the success rates. For prokaryotic proteins the success rate seems to be around 50%, which most likely reflects the close phylogenetic relationship between the target organisms and the expression system. For eukaryotic proteins, however, the success-rates are significantly lower (see also Table 2). An exception is the set of 93 proteins from S. cerevisiae of which Yee et al. can purify 63% (Yee et al., 2002). This is most likely due to the small size of the proteins. All reports note that proteins become progressively harder to purify as their molecular weight increases. Thus, the use of the His_6-tag in the purification of eukaryotic proteins under

non-denaturing conditions is obviously limited and this method is likely to be of use in special applications only.

Organism	# Prot.	Protein Properties	Purification Scale	% Purified (% Soluble)	Ref
M. thermo-autotrophicum	424	no TM	50ml	41% (47%)	Christendat et al. (2000)
T. maritima	1376	random	96-well, after 65ml fermentation	47%	Lesley et al.(2002b)
T. maritima	21	all < 23kDa no TM	1 liter	93%	Yee et al. (2002)
E. coli	130			61%	
M. thermo-autotrophicum	250			51%	
Myxoma Virus	25			46%	
S. cerevisiae	93			63%	
C. elegans	167	10 – 140kDa no TM	96-well, 1ml	6%	Chance et al. (2002)
H. sapiens	32	16 – 110kDa no TM	96-well, 1ml	13%	Braun et al. (2002)

Table 5: Purification success of His_6-tagged proteins under non-denaturing conditions in different studies.

Given the poor performance of the His_6-tag for eukaryotic proteins, many alternative purification tags have been developed. Several of these tags advantageously increase the solubility of the recombinant proteins, often outweighing the increased risk for steric interference of protein function by these larger tags. Therefore we had included the GST- and the MBP-tag in our evaluations.

Using either of the GST- or the MBP-tag we purified 26/32 proteins (82%). Recently, Hammerstroem et al. used 27 small human proteins (6-19 kD) to assess the solubilizing properties of 7 different tags. Unfortunately, the results from this study cannot be generalized without further verification because of the strong bias towards smaller proteins, which are generally easier to solubilize than larger proteins. In contrast, Shih et al. (2002) investigated the effects of eight different tags on the solubility of 40 different proteins from five different organisms. The proteins covered a size range of 9 – 140 kD. The results of the three studies are summarized and compared in Table2.

Tag	Size	Hammarstrom et al., (2002).		Shih et al., (2002)		Braun et al., (2002).	
		27 Proteins 6-19 kDa		40 Proteins 9-140 kDa		32 Proteins 16 – 110 kDa	
		Soluble	Purif'd	Soluble	Purif'd	Soluble*	Purified
His_6	1 kDa	29%	n/a	n/a	n/a	>80%	13%
CBP	4 kDa	n/a	n/a	< 38%	n/a	>80%	15%
GB1	6 kDa	**68%**	n/a	n/a	n/a	n/a	n/a
Thioredoxin	11 kDa	**74%**	n/a	< 38%	n/a	n/a	n/a
ZZ	14 kDa	55%	n/a	n/a	n/a	n/a	n/a
CBP	17 kDa	n/a	n/a	< 38%	n/a	n/a	n/a
GST	26 kDa	48%	n/a	38%	n/a	>80%	**80%**
MBP	42 kDa	**70%**	n/a	**60%**	n/a	>80%	**80%**
NusA	54 kDa	52%	n/a	**60%**	n/a	n/a	n/a
Intein-CN	55 kDa	n/a	n/a	< 38%	n/a	n/a	n/a

Table 6: Summary of 3 Studies addressing the effect of several purification tags on solubility and purification of test proteins. Hammerstroem et al and Shih et al. score solubility based coomassie blue analysis of total vs. cleared lysates on SDS-PAGE. In contrast Braun et al. scored for the presence of the target proteins by western blot analysis on cleared lysates.
Abbreviations: His_6: Hexa-Histidine-tag, CBP: Calmodulin Binding Peptide; GB1: B1 domain of *Staphylococcus* Protein G; ZZ: 2 sequential Z-domains of *Staphylococcus* Protein A; GST: Glutathione-S-transferase; MBP: Maltose Binding Protein

All three studies confirm the result obtained on a smaller scale that MBP is the best tag with regard to solubility. Also NusA performs well in the two studies addressing solubility; however, the large size of both of these tags may increase the chance of sterical interference in downstream applications. Both Thioredoxin and the B1 domain of *Staphylococcus* aureus' Protein G (GB1) appear to combine good solubilizing properties with a small size. However, Thioredoxin performed less well when larger proteins were tested. However, NusA and Thioredoxin do not have a natural high affinity interactor and these tags would have to be combined with a second small tag like His_6- that can contribute an affinity moiety for subsequent purifications.

In our purification experiments, both the GST- and the MBP-tag enabled the purification of 80% of the test set proteins with yields of >0.3 mg/ml. The better performance of the GST tag in our study compared may be partly attributable to a better expressing vector construct that we used in comparison to the other studies. Furthermore, in our study the final protein yields obtained with the MBP tag were slightly lower than those obtained with the GST-tag, despite a higher solubility of the proteins carrying MBP.
Analysis of the different purification fractions revealed that MBP tagged proteins bind less efficiently to their affinity matrix, and the lower yields are partially caused by inefficient binding of MBP to its affinity matrix. This finding underscores the importance of evaluating candidate tags in the context of the final application because focusing on isolated aspects of the process, i.e. solubility, may give misleading results. Problems with high losses of MBP-tagged proteins have been described before and may be due to weak affinity of MBP for amylose which was used as immobilized affinity matrix (Pryor and Leiting, 1997). It has been suggested that a His_6-tag N-terminal to MBP may alleviate the problem of large losses to the flow through by exploiting the high-affinity of His_6 for its ligand with the solubilizing properties of MBP. Likewise, a double tagging strategy has been suggested to increase the purity of preparations of GST-tagged proteins (Strugnell et al., 1997). It may be worthwhile to evaluate some double-tags for HT applications.

In protein purifications under non-denaturing conditions, we found that medium sized proteins (45kDa – 100kDa) were generally more difficult to purify than small proteins (<45kDa). Paradoxically, however, large affinity tags, which increase the size of the fusion proteins significantly, enabled us to purify much larger proteins than the small tags. The fact that proteins fused to the small tags cannot be specifically eluted is a sign of unspecific binding to the affinity matrix and indicates that the proteins may be denatured. Because some proteins, including the positive control provided by the manufacturer of the CBP-system, purified well in our method, and because no better results were obtained when the proteins were purified in individual tubes, it is likely that the low success rate is not an artifact of the developed methodology. Because the formation of inclusion bodies is time and temperature dependent, it is possible that the proteins, although recovered in the soluble fraction, were not properly folded but in an

early state of aggregation (Klein and Dhurjati, 1995). The co-expression of heat shock proteins might help to improve these results (Wall and Pluckthun, 1995). In contrast, the solubilizing and stabilizing effect of the large GST and MBP affinity tags on human proteins is well documented (Kapust and Waugh, 1999; Smith, 2000). However, while this explains the positive effect of these tags on protein expression, it does not address the increased difficulty of purifying medium sized human proteins (45kDa – 100kDa) compared to small proteins (<45kDa). Obviously, it is not impossible *per se* to express and purify medium sized proteins. Instead, some medium sized or large proteins, such as HDAC1 (51kDa) or STAT5A (87kDa), express and purify very well, whereas others, like ITPKA (52kDa) or Rad54L (81kDa), degrade very easily under the same conditions. This finding prompted us, to investigate the effect of protein domains on the ability to express and purify proteins (see below).

5.2. Heterologous Protein Expression in Bacteria

5.2.1. Functional Integrity of Bacterially Produced Proteins

An important question of recombinant protein production regards the functional integrity of the produced proteins. We have demonstrated that the developed HT purification method is compatible with obtaining functional protein, by assaying selected proteins for previously reported activities. Unfortunately, however, this does not enable us to make global statements about the functional integrity of proteins expressed in bacteria. Others have attempted to address this question. Based on the quality of NMR spectra, Christendat et al. estimated that 57/100 soluble proteins may be in a state of aggregation and thus potentially non functional (Christendat et al., 2000). This finding is supported by Yee at al, who found that the NMR spectra of 27% to 55% of soluble proteins, depending on the organism, indicated aggregation or conformational instability of the protein (Yee et al., 2002). At this point the significance of these findings with respect to the function of the target proteins is unclear. Our own data indicate that His_6-tagged proteins even though they can be found in the soluble fraction have a tendency to bind and elute very inefficiently and thus to "misbehave". The same was found for CBP-tagged, but to much lesser extend for GST and MBP-tagged proteins indicating the possibility that proteins fused to the latter tags may be more stably folded. Together these findings underscore the requirement for approaches that can assess the functional integrity of large numbers of purified proteins. Generic assays for specific groups of proteins, e.g. phosphatases, may be a step in this direction. However, despite justified skepticism it must also not be forgotten that thousands of functional proteins have been produced in bacteria over the last few decades and despite numerous alternatives *E. coli* is still the most widely used protein expression system.

5.2.2. Parameters Influencing Protein Purification Success in *E. coli*

The goal of HT protein expression is the production of a maximum number of full length, often functional, proteins in practical yields with the simplest and cheapest approach. Our and others data indicate, however, that it will be impossible to express all human proteins with any single method in *E. coli*. Historically, purification success has often depended on a tedious trial and error process of trying to express the target protein in the context of different purification tags and in different expression systems. While this approach is possible to imitate on a larger scale, it would be cost and time consuming and largely impractical. In order to make HT protein production more efficient, it would be of great advantage to identify parameters that influence the purification success of individual proteins. Proteins that have low success chances in E. coli could be identified early in the process and could be immediately expressed in systems where the success chance is higher. We aimed to identify protein parameters that have such prognostic value.

5.2.2.1. Protein Size

When the expression levels and final yield of the 128 fusion proteins were analyzed, it was found that large proteins within one set were generally expressed at lower levels than smaller proteins. This effect is particularly prominent within the MBP-fusion set. Pryor et al. mention the expression of more than 120 proteins between 5kDa and 75kDa as MBP fusion proteins without noticing a correlation between size and expression levels (Pryor and Leiting, 1997). It is difficult to compare their results to ours, because their purifications were not done in parallel, because their protein set included bacterial proteins and because their data were not shown.

Analysis of our large data sets for the GST and His_6-tagged proteins revealed a clearer picture of the relationship between protein size and purification success. Interestingly, the relationship was different for GST- and His_6-tagged proteins. The purification success for His6-tagged proteins up to 75kDa was constant at approximately 70% and dropped rapidly for larger proteins, The slightly lower success rate of proteins in the range 25-50kDa may either be caused by an accidental bias in the protein selection

or in this size range are protein families, which are difficult to purify for other reasons than their size.

For GST-tagged proteins a linear relationship between protein size and purification success was observed. For small proteins up to 25kDa the success rate was 65% - remarkably similar to that of His_6-tagged proteins. In contrast to His_6-tagged proteins, however, the success rate drops progressively as the proteins become larger, and is below 20% for proteins larger than 100kDa. It is possible that attachment of a fusion tag to the target protein creates a multi-domain protein. Potentially, a multi-domain structure is more sensitive to protein degradation or misfolding than proteins consisting of one domain.

5.2.2.2. Amino Acid Composition

The content of selected amino acids influenced the purification success of His_6-tagged proteins. All of the amino acids which had a positive impact were charged amino acids. In contrast, the content of polar but uncharged or hydrophobic amino acids did not influence the success rate. However, the length of a continuous stretch of hydrophobicity, depressed the success rate of His_6-tagged proteins in a linear fashion. Importantly, the latter finding is not caused by transmembrane proteins in the dataset, as omission of these proteins from the analysis yielded the same result. The same parameters have been identified previously by Christendat et al as important for the production of soluble His_6-tagged proteins from *Methanobacterium thermoautotrophicum* in *E. coli* (Christendat et al., 2000). However, the authors use these parameters to design a decision tree, which predicts which proteins can be made soluble and which ones not. Using our dataset, we were not able to confirm their specific cut-off values. This discrepancy can be caused by many factors including the facts that proteins from different organisms were expressed, that the proteins in the two sets spanned very different size ranges, that different expression conditions were used or that we purified the proteins under denaturing conditions whereas the other study used non-denaturing purification conditions. In addition it is possible that the decision tree and the used cut-off values are merely adequate to describe relationships in the specific dataset but do not posses universal

validity. However, even though the decision tree could not be reproduced, it is very encouraging that the same parameters have been identified in two different analyses and on datasets from two different organisms. Next, it will be interesting to analyze whether the more detailed relationships of our analysis can be detected in the second data set. If so, it may be possible to develop a refined algorithm to predict protein purification success.

The same analyses for the data of GST-tagged protein showed no effect of charged amino acids or of a continuous stretch of hydrophobicity on the purification success. While the absence of a positive effect of charged amino acids may be dismissible as a consequence of different purification methods (i.e. denaturing vs. non-denaturing), the fact that proteins with an extended stretch of hydrophobicity purify better with the GST-tag and under non-denaturing conditions is noteworthy. Thus, different or additional criteria must influence the purification success of fusion proteins and potentially more generally multi domain proteins. It is possible that a hydrophilic N-terminal tag such as GST can prevent the aggregation of hydrophobic stretches close to the N-terminus of the target protein. According to this hypothesis, only hydrophobic stretches in the middle or close to the C-terminus of the recombinant protein will prevent its purification. A second hypothesis is that the position of the N-terminus of the target protein in its 3-dimensional structure affects whether a large N-terminal fusion tag is compatible with correct folding or not. Thus, when the N-terminus is in a fixed in an essential position in the structure, the addition of a GST-fusion tag may destabilize the structure and cause denaturation and degradation. In contrast, when the N-terminus of the target is loosely incorporated into its structure, the addition of a tag may be compatible with correct folding. If this hypothesis is true, the success rate for GST-tagged proteins can potentially be increased, by increasing the length of the linker between GST and the target protein.

5.2.2.3. Differential Codon Usage

Several authors have reported a benefit for some proteins when these are expressed in genetically modified bacterial strains, which carry extra copies of tRNAs for codons that are only rarely used in coding regions of *E.coli* (Kane, 1995). Using a test set of eight

different proteins we did not see such an effect. In order to address the possibility that differential codon usage did influence purification success on a larger scale, the purification success of several hundred proteins was correlated to total amount and frequency of rare codons in their cDNAs. In both conditions, a weak correlation between the total amount of codons in a cDNA and the purification success was found. In neither case did the frequency of the analyzed codons play a role. This suggests that the employed expression conditions do lead to a depletion of the intracellular pool of minor tRNAs. Only when the number of rare codons becomes overwhelming do these interfere with protein synthesis and the ability to detect the purified proteins.

5.3. Protein Production for Functional Proteomics

In this study, we demonstrate the utility of *E.coli* as an expression system for functional proteomics enterprises. Of 776 different proteins, a product of the expected length could be purified for 517 proteins using the His_6-tag and denaturing purification. This corresponds to a success rate of 67%. Although no special selection methods were used to assemble the set of ~800 proteins, this number represents a reasonable estimate for the success rate of high-throughput expression of medium-sized human proteins in bacteria. The proteins purified under denaturing conditions can be used to manufacture diagnostic protein arrays or to make affinity reagents. In addition, enzymatic activity can frequently be recovered from proteins purified under denaturing conditions after using a refolding step. It would be useful to adapt refolding to a HT setting (Lilie et al., 1998)

Under non-denaturing conditions 49% of 428 attempted proteins could be purified as full-length proteins using the GST-tag under non-denaturing conditions. The protein yield from each successful purification (>300ng) is sufficient for the construction of several hundred-protein arrays, which require a few nanogram of proteins per spot (MacBeath and Schreiber, 2000; Zhu et al., 2001).

The numbers represent the best estimate of the success rate for the purification of human proteins from bacteria. The evaluated set of proteins is not representative for the human proteome, though. A bias towards smaller proteins has been introduced by the cloning operations from which the analyzed cDNAs were obtained. These operations are currently biased toward smaller cDNAs (<2000bp) and proteins were consequently enriched in proteins of 70kDa or smaller. A subset of the GST-tagged proteins was enriched in proteins that can be purified using the His_6-tag. However, analysis of the success rates of GST-tagged proteins, which had failed or had been successful with the His_6-tag, suggests that this bias only marginally affects the overall estimate of the success rate for purifications using the GST-tag.

5.3.1. Coverage of Proteomes

The presented study indicates that it may not be possible to produce whole proteome representations in bacteria. When such protein sets are required there are two alternative strategies to obtaining them: i) a staggered approach in which several complementary expression systems are used to produce all proteins or ii) switching to an expression system, in which all required proteins can be expressed. Currently, no datasets comparable to the here presented are available for other expression systems and thus it is impossible to evaluate whether any single expression system will be capable of producing most proteins for any application. Zhu et al used *S. cerevisiae* to express 5800 proteins from the same organism, which were used to build protein arrays. Based on western-blot analysis of 60 samples, the authors estimated that approximately 80% of all proteins produced detectable amounts of proteins of the correct size. This number was confirmed by immunodetection of the spotted proteins on the slides using the same antibody. One caveat of this experiment is that the antibody recognized the N-terminal GST moiety, which can give a false positive readout in cases where the GST tag has been lost from the target protein during expression or lysis. This phenomenon has been frequently observed for GST-tagged proteins expressed in bacteria. Furthermore, homologous expression of proteins is likely to produce higher success rates that the heterologous expression of recombinant proteins.

A recent report by Sawasaki explores wheat germ lysate for proteomic microscale protein production. The authors report that 50/54 both human and *Arabidopsis thaliana* proteins could be detected by colloidal

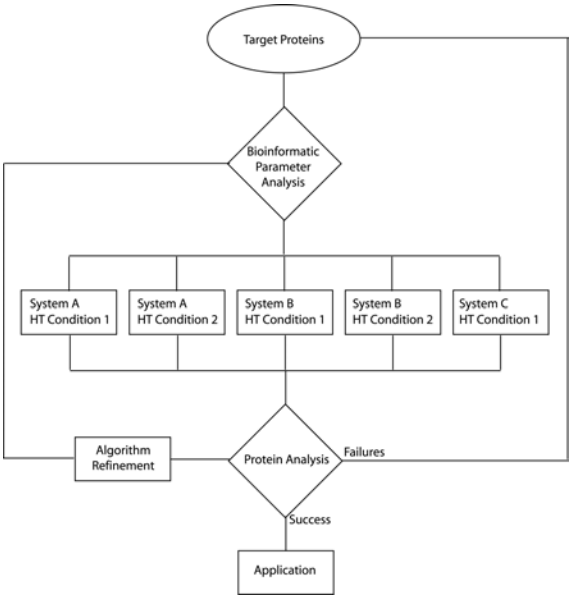

Figure 35: Protein Production Triage for Functional proteomics enterprises

coomassie staining after HT PCR cloning, transcription and HT cell-free protein production. The obtained protein yields ranged from 0.1mg/ml to 2.3mg/ml. The authors show that their method is compatible with functional protein by demonstrating autophosphorylation activity of 4/5 Arabidopsis kinases and recording an NMR spectrum of one additional protein (Sawasaki et al., 2002).Chambers suggested analyzing all targets before any purification to identify proteins that have a high probability of expression success in any particular system (Chambers, 2002). In order to investigate whether we could identify parameters that carry prognostic information, we related the purification success of all purified proteins to features of the coding sequences, the proteins on several structural levels, and to function and location of the physiologically expressed proteins. This analysis revealed several parameters, which correlated well with an increased or decreased likelihood of successful purification from bacteria (discussed in greater detail above). A crucial next step will be the development of algorithms that can use the information to make actual predictions about the purification success of individual proteins. The presented data will contribute to a better understanding of the factors that influence protein stability and solubility in bacteria and this improved understanding will assist development of predictions algorithms.

5.3.2. Proteins for Focused Proteomics Approaches

A global analysis of many biological and biochemical relations is often unpractical or unnecessary. Certain types of experiments are only sensible for a subclass of proteins, e.g. those which are likely to exhibit an enzyme activity of interest. Current data indicate that bacteria will not be able to make all proteins expressed by humans and many other organisms of interest. However, the analysis of proteins containing certain domains or mediating certain cellular functions indicates that proteins in some of these subgroups have success rates of up to 90% in *E. coli*. Obviously, it is important to verify the functionality of the produced proteins. The presented data indicate, however, that E. coli is a very useful expression systems to produce (near-) comprehensive proteins sets for some protein subgroups.