4. Results

4.1. Parallel Protein Production Scheme

High throughput experimentation is characterized by parallel and often automated processing of many samples. Thus, the first challenge in developing a HT protein production method is the identification of conditions that enable the standardized expression of proteins.

4.1.1. Standardization of Protein Expression

4.1.1.1. Culture Conditions

In order to explore the growth rates of parallel bacterial cultures, the growth rates bacterial cultures containing different plasmids in the same genetic background were investigated. These experiments indicated that a leaky protein expression interferes with consistent growth. In the tac/lac promoter system, leaky protein expression can be repressed by addition of glucose into the growth media (Weickert et al., 1996). As shown in the figure, bacterial growth was even when bacterial cultures were grown in the presence of 2% glucose.

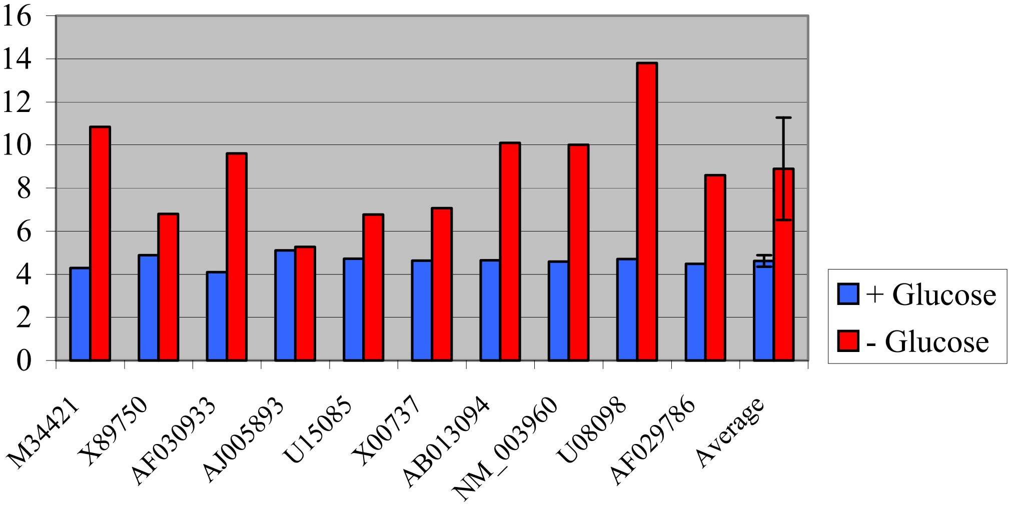
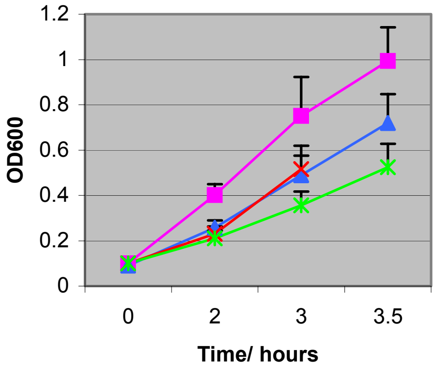


Figure 9: Effect of 2% Glucose on Growth consistency of different Bacterial Cultures. The OD_{600} of ten different cultures after 14 hours of growth in the presence or absence of glucose. All cDNAs were in a pDEST17 background and transformed into BL21pLys^{s} cells. The identity of each cDNA is indicated below each bar (Genbank accession number).

In early experiments it was noticed that the growth rates of bacterial cultures after dilution from the starter culture were very variable. Therefore the effect of the length of the starter culture on variation of the growth rates in the expression culture was investigated. 16 bacterial clones containing plasmids encoding different cDNAs were grown for 12, 14, 16 hours or 18 hours, diluted to a final OD_{600} of 0.1 and the growth curves recorded. Figure 3 shows that starter cultures grown for 12 hours pick up growth most quickly. However, at the same it can be noted that the cultures diluted from this starter cultures show the greatest variation in growth rate. In contrast, the expression cultures started after 14 hours or 16 hours exhibit more even growth. The cultures inoculated with an 18 hour old starter culture grow extremely slow. In response to these data the starter culture was always grown for 14 hours.



	0	2	3	3.5
■ 12 hrs	0.1	0.401	0.751 ± 0.17	0.993 ± 0.15
▲ 14 hrs	0.09	0.257	0.489 ± 0.09	0.72 ± 0.13
× 16 hrs	0.1	0.231	0.516 ± 0.1	n.d.
✳ 18 hrs	0.1	0.211	0.356 ± 0.06	0.525 0.1

Figure 10: Growth curves of cultures diluted from starter cultures of different age.
16 starter cultures were grown over night in a 96-deep well block. At the indicated time points (12 hrs, 14 hrs, 16 hrs, 18 hrs) the OD_{600} of the 16 cultures was measured and fresh cultures were inoculated so that their average OD600 at time=0 was 0.1. The growth of the 4x 16 cultures was monitored by OD600 measurements over the course of 3.5hrs.

4.1.1.2. Bacterial Strain

Genetically modified protein expression strains have been reported to have a beneficial impact on the expression of individual genes. Therefore, the effects of two different bacterial strains, BL21-DE3*pLys*^{S} and BL21-DE3 CodonPlus, on protein yield and on growth characteristics of the bacterial cultures were evaluated. BL21-DE3*pLys*^{S} cells carry a plasmid encoding T7 Lysozyme, which represses protein expression during bacterial growth and assists protein lysis, once the cell membrane has been perforated.

The BL21-DE3 CodonPlus strain has been modified to express additional copies of the tRNAs for AGA-Arg and CCC-Pro and may thus assist the expression of proteins whose coding regions are rich is these codons. First the growth characteristics of both strains were assessed. The results show that both populations of cultures grow with similar rates, however the tighter repressed bacteria grow more evenly.

As a second criterion the effect of the two bacterial strains on the expression of eight different proteins was analyzed. Because statistically long cDNAs contain more rare codons the eight selected proteins were of a size range between 40kDa and 110kDa and thus ‘larger’. Furthermore, in order to be able to measure a wide spectrum of effects the selected proteins encompassed very good and very low expressers. Figure 4 shows no detectable difference of expression in the two strains. Thus, because the BL21-DE3*pLys*^{S} cells grew somewhat more evenly and because we anticipated a beneficial effect of the T7 Lysozyme for lysis under non-denaturing conditions, this bacterial strain was selected for further experiments.

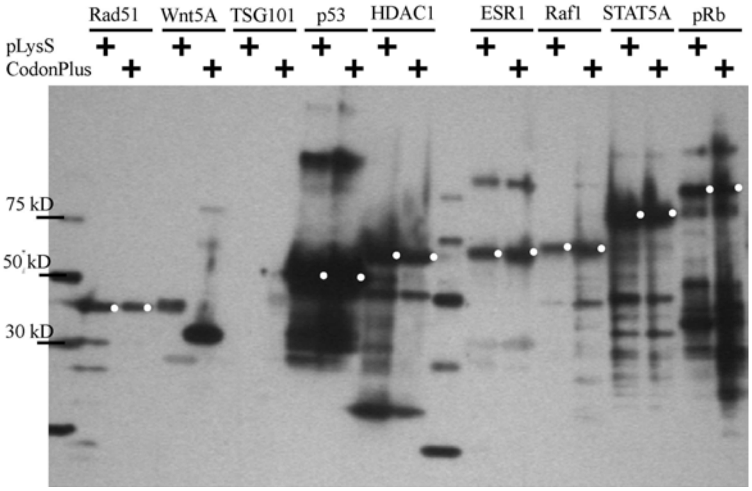


Figure 11: Comparison of two bacterial strains expressing 8 different proteins.
The indicated proteins were expressed in BL21pLyss and BL21CodonPlus cells. After protein induction for 2 hrs at 30°C protein levels were analyzed by western blot analysis on whole cells. White dots indicate bands of the correct size.

4.1.2. Parallel Protein Purification under Denaturing Conditions

4.1.2.1. Bacterial Lysis

The first step in developing a 96-well protein purification procedure was the development of 96-well-format compatible lysis conditions. Under denaturing conditions, bacteria can be easily lysed by resuspension in lysis buffer that contains 8M Guanidine hydrochloride. Previous experiments in the lab had demonstrated that the use of an inverted 96-well pin device is required to achieve proper mixing of 2 liquids in the wells of a 96 deep-well block (A. Halleck – unpublished results). When this device was used in combination with denaturing lysis condition a homogeneous lysis was achieved as judged by successful purification (see end of this chapter). These lysis conditions were used to develop the remaining procedure under denaturing purification conditions.

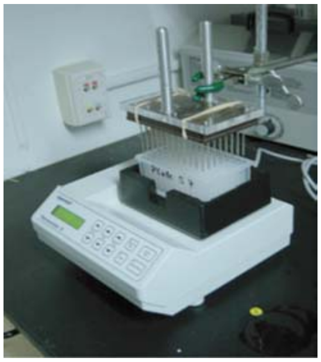


Figure 12: The Swizzler. An inverted 96-pin device assists resuspension of bacteria and mixing of solutions.

4.1.2.2. Lysate clearing

In order to clear the lysate and remove cellular debris that may contaminate the purified proteins, removal of the debris by centrifugation or by filtration using either vacuum or centrifugation were considered. The most powerful centrifuge that accommodates 96-well plate holders, achieves a relative g-force of 7000 x g, which was not capable of removing the cellular debris. When using a 96-well filter plate with glassfiber filterplates and a vacuum manifold to clear the lysate, the high protein concentration of the lysate caused foaming, which resulted in contamination of neighboring wells. Filtration of the lysate through a glassfiber filterplate resulted in successful clearing of the lysate. In the final process, the cleared lysate was collected in a second identical filterplate, into which the affinity matrix had been dispensed. Glassfiber was selected as a filter material, because glass frits have been successfully used as inert material to retain protein purification columns.

4.1.2.3. Washing and Elution

After the binding reaction the proteins were purified by repeated centrifugation at 4°C followed by addition of the respective buffers (see methods). Under denaturing protein purifications, a relative g-force of 50 x g gave satisfactory results. After four washes the OD_{280} of the wash was indistinguishable from buffer.

Applying the developed method for denaturing conditions to the 32 His_6-tagged test-set proteins, all proteins, except the 110kDa pRb, which were detectable *in vivo,* could be purified successfully. (A visible band of the correct size on a GelCode® Blue stained SDS-PAGE was considered a 'successful purification' throughout the experiments. Such a band corresponds to a total yield of at least 300ng) Denatured proteins are a useful product in some application such as antigens to make antibodies, as substrates in some enzymatic assays or in diagnostic applications. In addition, enzymatic activity can frequently be recovered from proteins purified under denaturing conditions after using a refolding step (Lilie et al., 1998). Given the success of denaturing purification conditions, it may be worthwhile to consider the development of HT renaturation methods.

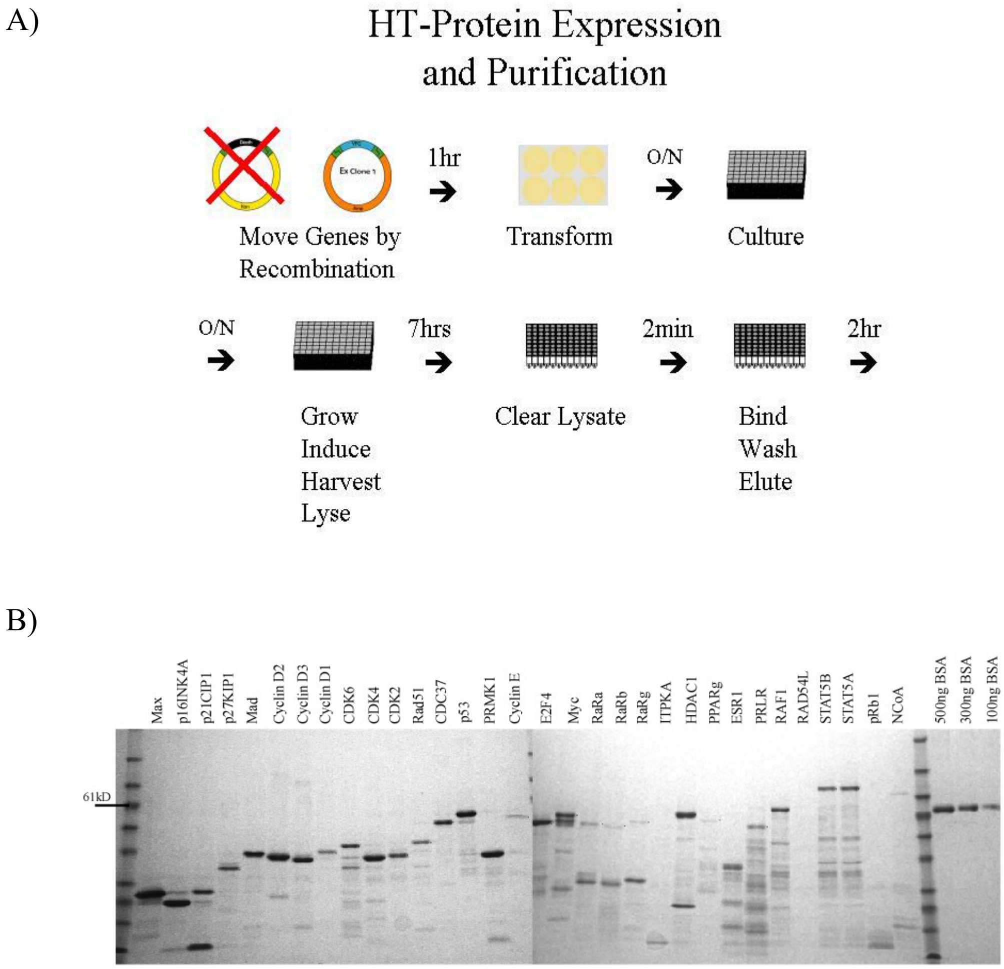


Figure 13: HT Protein Purification Process.

A) Schematic overview. Proteins are subcloned into expression vectors by recombinational cloning and retransformed into a bacterial protein expression strain. Overnight starter cultures are inoculated in 96-well deep well blocks. Fresh expression cultures are grown in an identical but different deep well block, in which proteins are expressed and the bacteria lysed. After lysis, the lysates are transferred into a glass fiber filter plate, which is placed in top of second, identical filterplate. The second plate contains the affinity matrix and is sealed at the bottom. The lysate is cleared and transferred into the second filterplate by centrifugation. After transfer the plate is sealed on top, the proteins bound to the matrix in batch mode. Washing and eluted are done by repeated centrifugation and addition of respective buffers.

B) Example of the purification of 32 His-tagged proteins under denaturing conditions in HT format. 15% of purifications from 1ml culture were loaded on a 4-20% gradient SDS-PAGE. The proteins were visualized by colloidal Coomassie Blue staining. In the very last three lanes different amounts

4.2. Purification Chemistry for Functional Protein Production

4.2.1. Expression Constructs

In bacteria, protein affinity tags can profoundly influence stability, solubility and expression levels of human proteins. Thus, an important aspect of the development of a HT protein purification method is the identification of a polypeptide purification tag with robust chemistry, which has favorable effects on yield and purity of many different proteins when used in parallel protein purification.

The biochemical and biophysical properties of proteins may vary significantly from one protein to another. To define conditions and identify a purification chemistry that works well when fused to a variety of proteins a test set of 32 full-length human genes was chosen. As it has been observed that protein expression in bacteria may be dependent on the size of the recombinant protein (Smith, 2000), the test set included proteins with a broad range of molecular weights. In addition, the set contained proteins that localize to different subcellular compartments and which have different biochemical activities. Integral membrane and secreted proteins were excluded because these classes of proteins require separate optimization and purification methods (Cornelis and Van Gijsegem, 2000). The test set was assembled in the Gateway™ recombinational cloning system to enable the rapid creation of the required expression constructs (Walhout et al., 2000b). The sequence-confirmed cDNAs were then transferred into each of the expression vectors to create 128 expression constructs. All transfer-reactions were confirmed by analytical PCR.

4.2.2. Protein Expression

The previously established protein expression conditions were used to evaluate the four protein affinity tags. It was found that nearly all of the 128 fusion proteins were expressed as determined by western blot analysis on whole cells using the respective antibodies (Figure 7). Within each set of proteins that carry one polypeptide-tag a decrease in expression levels was observed with increasing protein size. Of the four tags, the GST fusion proteins were most prone to degradation *in vivo*. In some cases, such as p21^{Cip1} and Rad51, only the full-length fusion protein and the GST band were detectable, suggesting that the GST moiety was cleaved off in some molecules. In other instances, like E2F4, multiple degradation products were visible, indicating that the fusion partner has been degraded in a step-wise fashion, with GST as a stable end-product.

For some proteins, ITPKA, RAD54L and NCoA, only degradation products were detectable regardless of the fusion tag. Since neither of the corresponding cDNAs has any mutation, this behavior suggests an inherent instability of these proteins when expressed in bacteria. Even though the retinoblastoma protein (pRb – 110kD) was expressed as a His$_6$-fusion protein, it could not be observed as a fusion with any of the larger tags.

Before the HT protein expression experiment described in the result section C, the protein expression conditions were re-optimized for GST-tagged proteins. In this experiment the influence of 2% Ethanol at the time of induction and the effect of induction with 0.3mM IPTG were explored for multiple induction times and growth temperatures (18°C, 25°C, and 37°C).

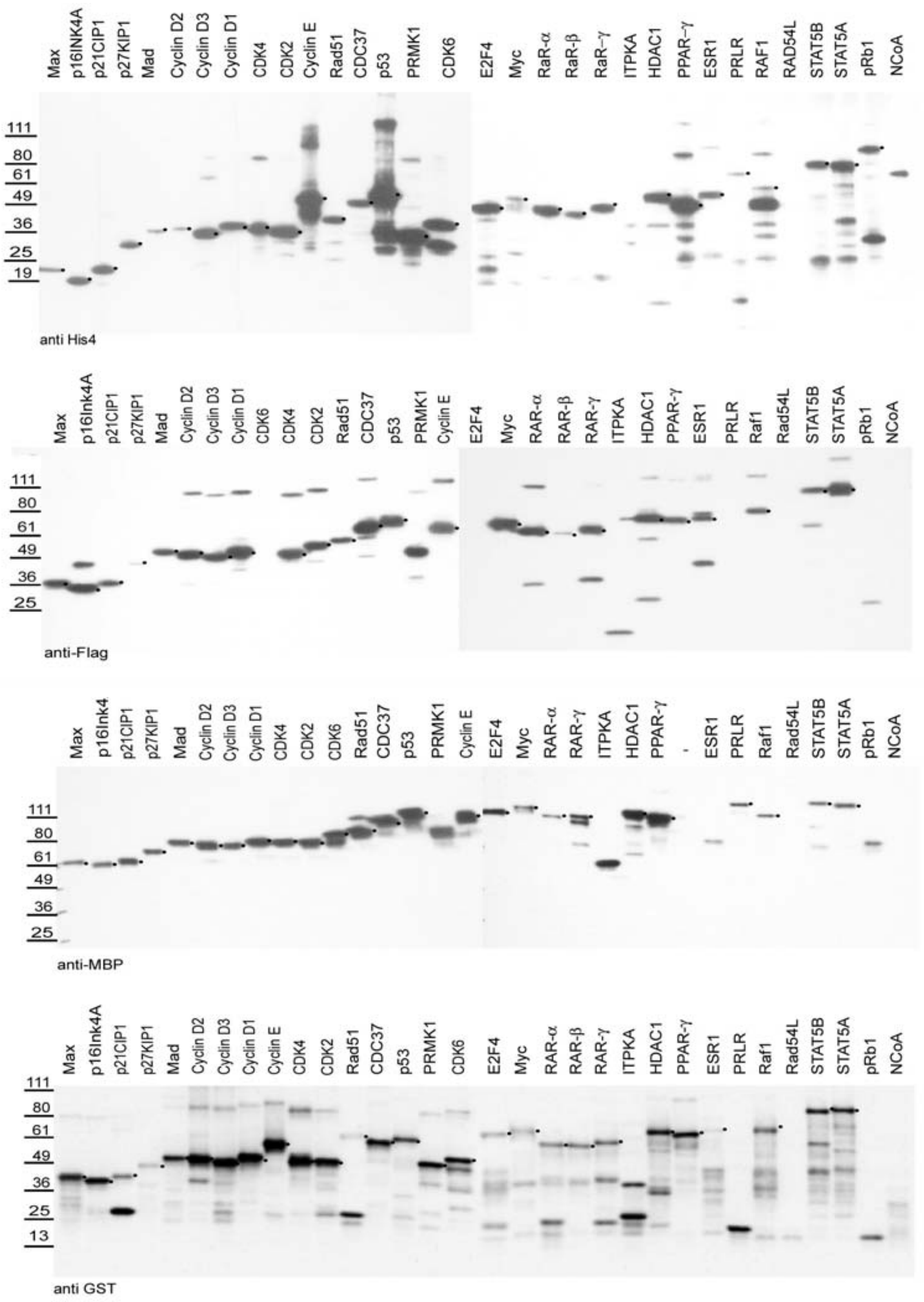


Figure 14: Expression of 32 test set proteins fused to four different purification tags.
The 32 genes in the test set were transferred into each of the four expression vectors, transformed into BL21pLyss cells, and cultured and induced as described. 10µl (~1%) of each culture were lysed directly in Laemmli buffer and analyzed by western blot using antibodies against the peptide tags as indicated. Bands of the correct size are indicated by a dot on the right side of the band.

4.2.3. Evaluation of Protein Affinity Tags

4.2.3.1. Process Development Control

It was expected that several modifications need to be made to adapt the previously developed purification process to non-denaturing conditions. In order to evaluate each process in detail, fractions were collected throughout the purification process and the losses in each step were analyzed by western blot analysis against the purification tag. One exemplary analysis of eight His$_6$-tagged proteins is shown in the figure below.

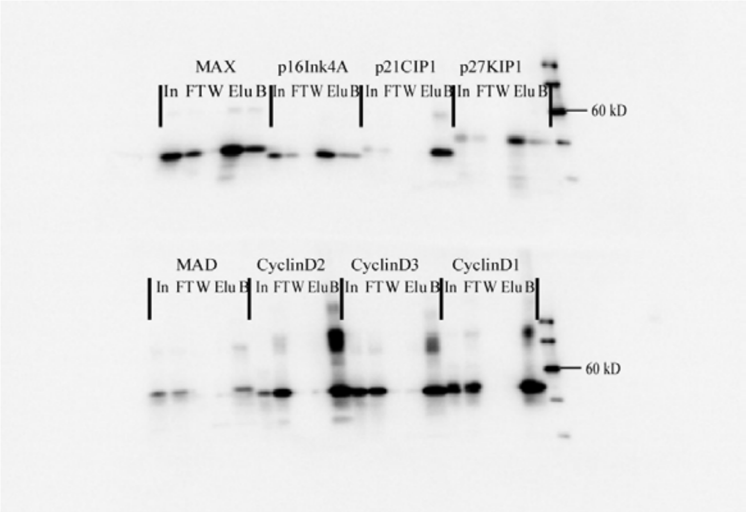


Figure 15: Process Monitoring by Fraction Analysis.
During the process development all fractions of protein purification were collected and analyzed for recombinant protein by western blot analysis. Of each fraction the following amounts were loaded on each gel: (Cleared lysate) Input In: 4%; Flow through (FT): 4%; wash (W): 10%; Eluate (Elu): 15% and Beads (B): 15%.

4.2.3.2. Parallel Bacterial Lysis in Non-denaturing Conditions

In order to lyse bacteria under non-denaturing conditions, several methods were evaluated including commercial lysis reagents, freeze/thaw cycles and chemical methods. A decisive drawback of the commercial lysis reagents was that no reagent was compatible with all four purification chemistries. All tested reagents had a high concentration of detergents, which interfered with the MBP purifications and the Bugbuster II reagent contained Tris-buffer at a concentration incompatible with the His$_6$-purifications. The freeze/thaw cycle was incompatible with 96-well format, because of a poor temperature conductivity of the deep-well blocks which were used for both protein expression and bacterial lysis. A combination of Lysozyme/ Triton X-100, followed by DNase treatment worked well. As shown in figure 9 most expressed proteins could be found in the soluble fraction. For all 128 proteins five fold more material was loaded of the soluble fraction (S) than of the total cells (T). The performance of the individual tags will be discussed below.

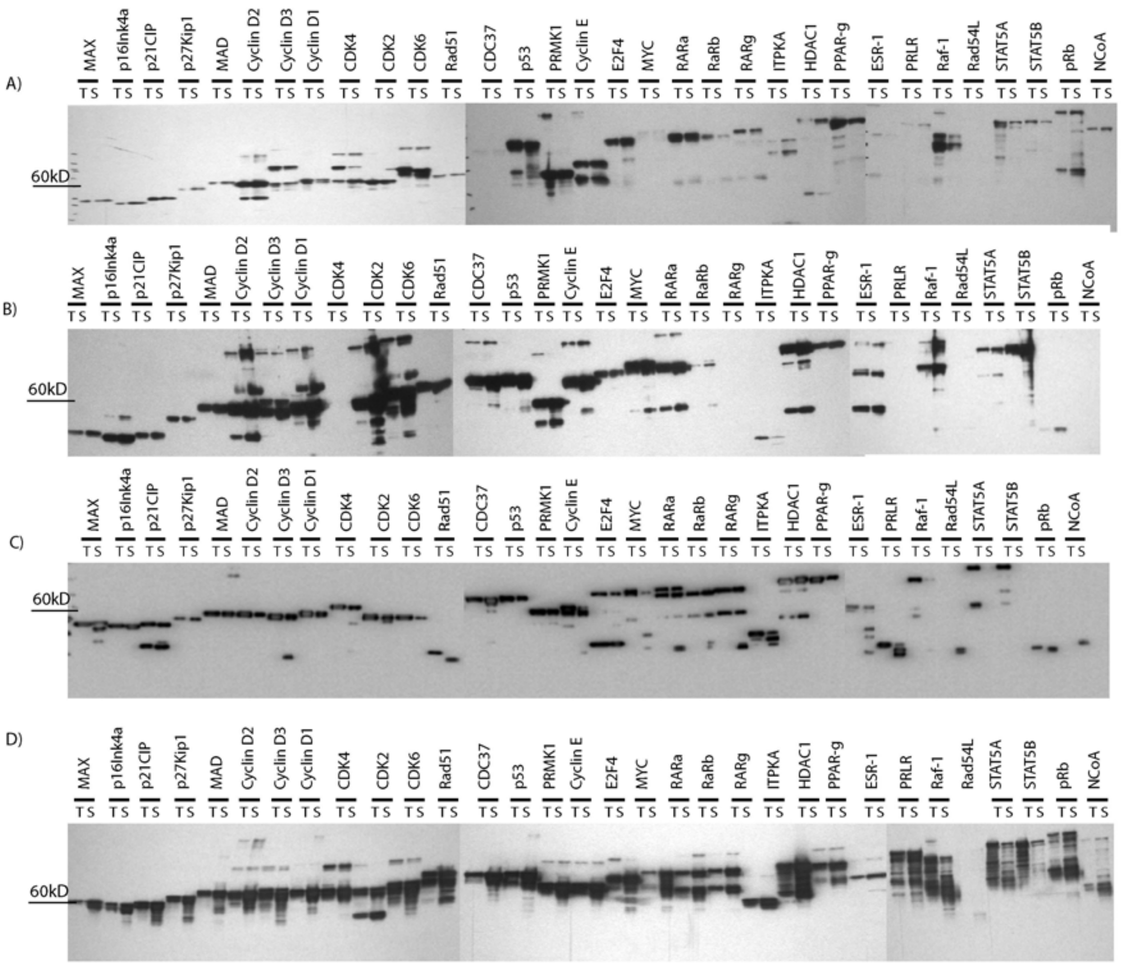


Figure 16: The developed lysis method transfers recombinant proteins into the soluble fraction. The equivalent of 0.1% of a 1 ml culture were loaded of total cells (T) and the equivalent of 0.5% of a 1 ml culture was loaded of the soluble fraction (S). The different proteins are indicated by name. A) His$_6$-tag; B) CBP-tag, C) GST-tag, D) MBP-tag

4.2.3.3. Binding is Dependent on Lysis Volume

During the first purification attempts it was observed that a surprisingly large fraction of all proteins was detected in the flow-through and did not bind to the affinity matrix. Under denaturing conditions lysis was done in 300µl volume. However, the equilibrium equation for binary complex formation shows that the total amount of formed complex is inversely proportional to the reaction volume ([Complex] $\propto$1/Vol). Therefore, the effect of relative lysate concentration on binding efficiency was evaluated.

Four different proteins were expressed in 50ml cultures and after harvesting the bacterial pellets were lysed in 1ml lysis buffer (50x concentration factor). This lysate was divided into five fractions and the aliquots were respectively diluted to final concentration factors of 50x, 25x, 10x, 2x and 1x. The protein in all lysates allowed binding to the affinity matrix under standard conditions, the matrix was washed and all bound protein was eluted by direct addition of Laemmli buffer to the beads. Corresponding amounts of eluate and flow through were loaded next to each other on a 10% SDS-PAGE. Figure 10 shows for two his-tagged proteins under non-denaturing conditions that low concentration factors (1x and 2x), which mimic the previously used conditions, results in very poor binding efficiency. In contrast, proteins in more concentrated lysates bound better to the matrix. A similar, however less pronounced effect was observed when the same experiment was repeated with MBP and CBP-tagged proteins.

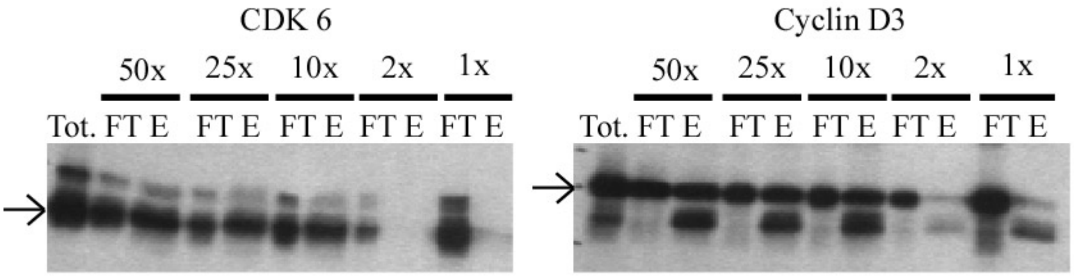


Figure 17: Concentration dependence of binding
His6-tagged CDK6 and Cyclin D3 were expressed in 50ml cultures and lysed in 1ml volume (50x). This lysate was diluted to 25x, 10x, 2x and 1x and equivalent amounts were added to affinity matrix. After binding the lysate (FT) was removed from the matrix, the latter was washed and the proteins were eluted by addition of Laemmli buffer. Equal amounts of FT and eluate were analyzed by western blot analysis.

In response to these results, the lysis volume was decreased to 100µl, which corresponds to a concentration factor of 10x (starting from 1ml culture volume). Further decrease of the lysis volume resulted in unacceptable losses during subsequent transfer reactions.

In addition to these general adaptations, the binding and buffer conditions were optimized for each affinity tag. Using the optimized conditions, the behavior of the 128 fusion proteins in parallel protein purification was characterized with respect to yield and purity.

4.2.3.4 Evaluation of Four Affinity Tags in a Parallel Protein Purifications

4.2.3.4.1. His$_6$

Despite the efficiency with which the His$_6$-tag functioned under denaturing conditions, only five proteins were detected by Coomassie staining after purification under non-denaturing conditions, although many more proteins (15/32) could be detected by western blot. Of these four, only MAX and p16^{INK4a} were reasonably pure (70%). All His-tagged proteins had a solubility of ~20% and the total levels were comparable for most proteins. However in analyzing the different fraction it was noticed that His-tagged proteins were lost in the FT and/or could not be eluted from the matrix (complete data in appendix). This was true for both Ni^{2+} and Co^{2+} matrices and with a broad range of imidazole concentrations (200mM–500mM) or 5mM EDTA used for elution in the presence of 500mM NaCl and/or 0.1% NP-40.

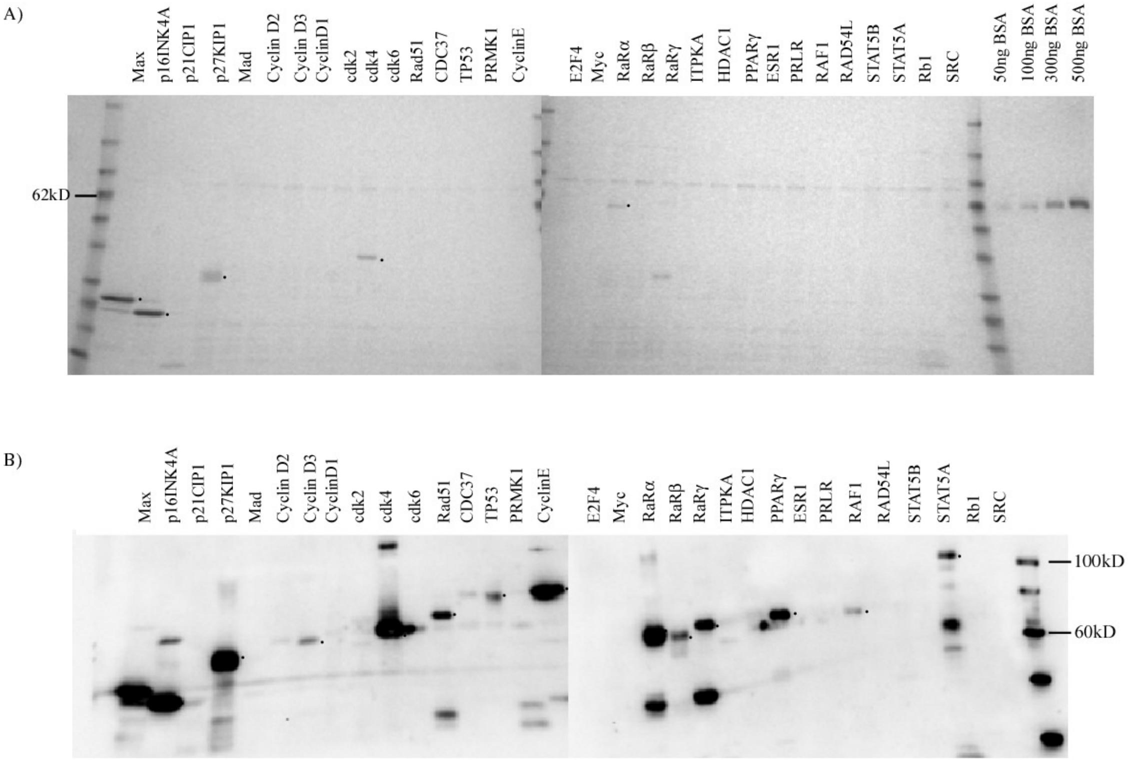


Figure 18: Protein Purification under non-denaturing conditions using the His$_6$-tag.
32 His$_6$-tagged test set proteins were expressed and purified as described in Materials and Methods. 15% of a protein purification from 1ml starting culture was analyzed by SDS-PAGE and A) Coomassie Blue staining and B) western blot analysis. Bands of the correct size are labeled with a black dot to the right of the band.

4.2.3.4.2. Calmodulin Binding Peptide

As with the His_6-tag, only five proteins of 32 proteins that had been purified with the CBP-tag could be detected on a coomassie stained gel and seven more by western blot. Of the CBP-tagged proteins CDC37, a chaperone for cdk4 (Lamphere et al., 1997), had the highest yield and purity. Like the His-tagged proteins all CBP-tagged proteins were soluble to some extend. A further analysis of the different fractions demonstrated that matrix binding for CBP-fused proteins was more efficient – only 10/32 proteins were detected at all in the flow-through. However, 22 of 32 matrix-bound proteins did not elute using 5mM EDTA even in the presence of 1M NaCl. Difficulties eluting CBP-tagged proteins have been mentioned in the literature previously (Vaillancourt et al., 2000).

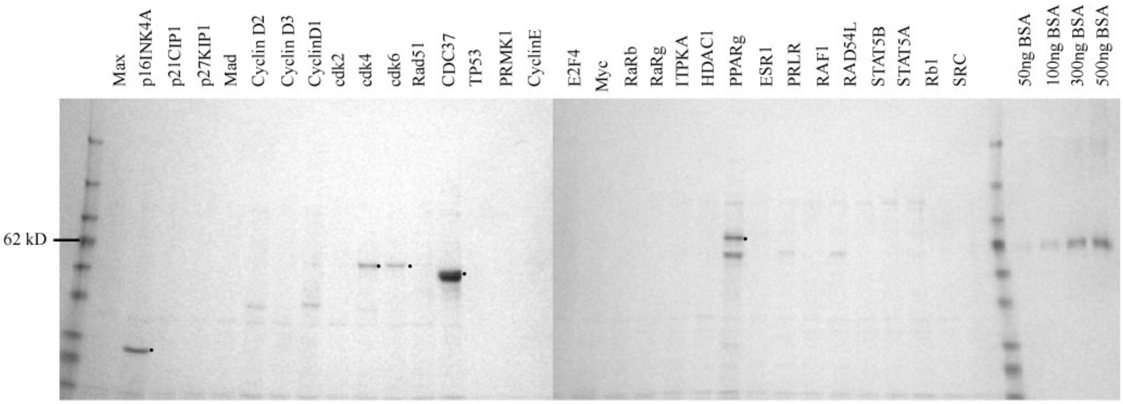


Figure 19: Protein Purification under non-denaturing conditions using the CBP-tag.
32 CBP-tagged test set proteins were expressed and purified as described in Materials and Methods. 15% of a protein purification from 1ml starting culture was analyzed by SDS-PAGE and Coomassie Blue staining. Bands of the correct size are labeled with a black dot to the right of the band.

4.2.3.4.3. Glutathione-S-Transferase

Of the GST-tagged proteins 26/32 proteins were purified with a yield of at least 300ng protein/ml culture and 22 of these with a yield of >1µg/ml culture based on a comparison with a quantity standard. Six of the 32 proteins could not be purified as GST constructs because the full-length protein could not be detected or the GST moiety had been lost *in vivo* (Figure 9). The total purity of most proteins was in the range of 30-70%. It is likely that many impurities are degradation products of the recombinant protein, because a similar pattern of bands was observed by an anti-GST antibody on a western blot. In addition, many of these bands were already detectable *in vivo* (Figure 9), indicating that the degradation occurred prior to cell lysis.

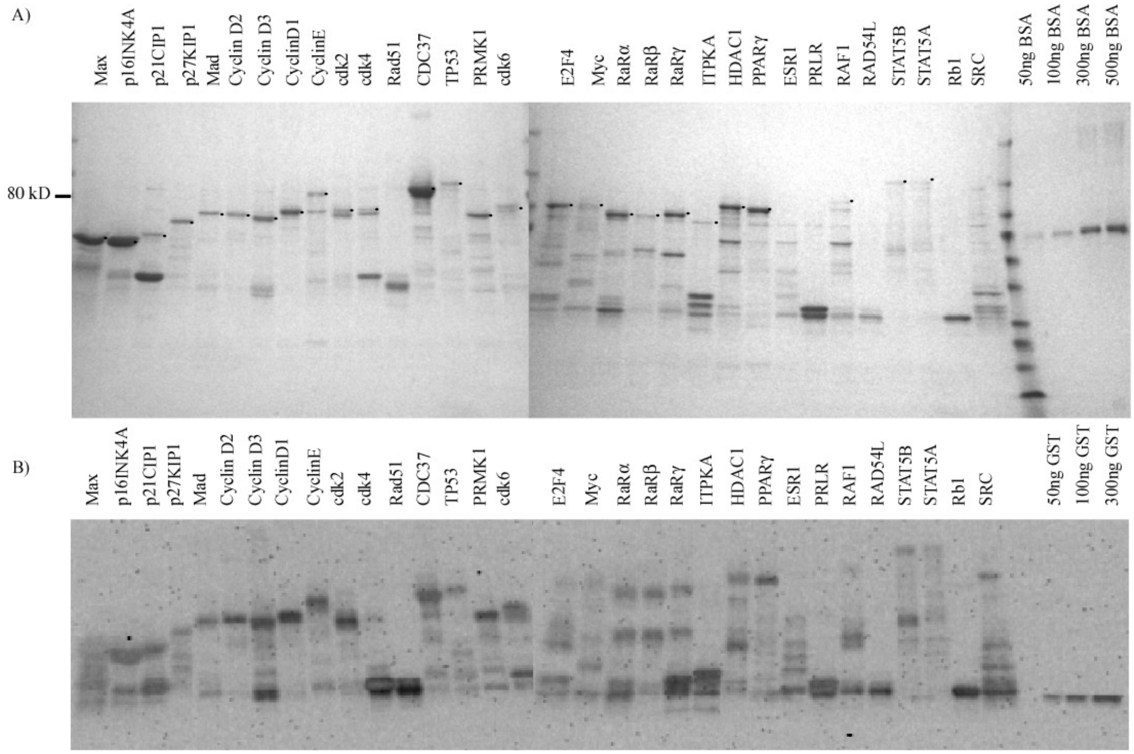


Figure 20: Protein Purification under non-denaturing conditions using the GST-tag.
32 GST-tagged test set proteins were expressed and purified as described in Materials and Methods. 15% of a protein purification from 1ml starting culture was analyzed by SDS-PAGE and A) Coomassie Blue staining and B) western blot analysis. Bands of the correct size are labeled with a black dot to the right of the band.

4.2.3.4.4. Maltose Binding Protein

The maltose binding protein purified 26/32 proteins to yields of at least 300ng protein per 1ml original culture, and 18 of these with yields of >1µg/ml (Figure 14). The purity of most MBP purified proteins ranged from 20% to 70%. As reported in the literature the MBP-tag proteins produced the greatest fraction of soluble protein. Most MBP-tagged proteins were primarily lost in the FT. A low binding efficiency of MBP-tagged proteins has been reported previously and is a result of the low affinity of MBP for its matrix (Pryor and Leiting, 1997).

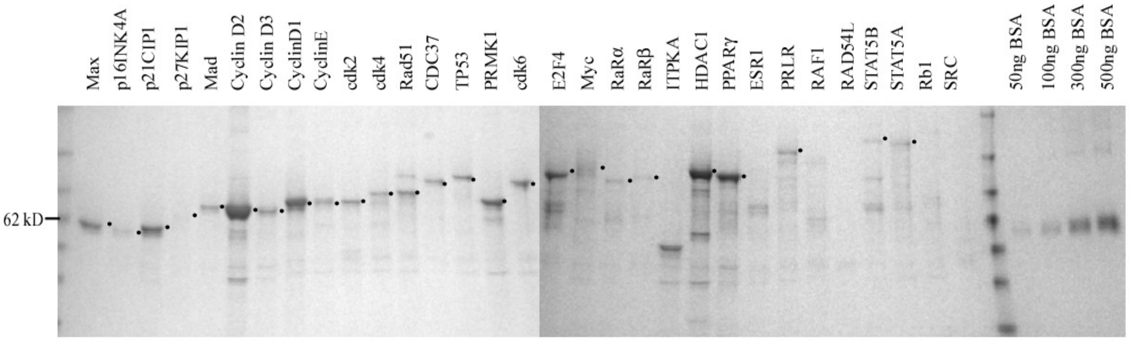


Figure 21: Protein Purification under non-denaturing conditions using the His$_6$-tag.
32 MBP-tagged test set proteins were expressed and purified as described in Materials and Methods. 15% of a protein purification from 1ml starting culture, were analyzed by SDS-PAGE and Coomassie Blue staining

4.2.3.4.5. Functional Integrity of GST- and MBP-constructs

To test whether the applied purification conditions produced biochemically active proteins, proteins tagged with GST and MBP were tested in two different biochemical assays. In the first assay, GST- or MBP-tagged Cyclin E, purified from bacteria in 96-well format, was combined in standard kinase reactions with recombinant cdk2 purified from insect cells using Histone H1 as a substrate. Cyclin E/ cdk2 complex purified from insect cells served as a positive control. Figure 15a shows that both fusion proteins activated histone phosphorylation.

As a second functional test, it was examined if GST-p16$^{\mathrm{INK4a}}$ specifically inhibited Cyclin D1/cdk4 kinase activity against C-terminal fragment of the retinoblastoma protein (LaBaer et al., 1997). Kinase reactions were incubated with 16 different GST-tagged proteins purified in HT format. As expected, GST-p16$^{\mathrm{INK4a}}$ selectively inhibited the kinase activity (Figure 15b). In the same experiment with MBP-tagged proteins no inhibition of the kinase could be observed. Analysis of relative levels indicated that the MBP-tagged p16Ink4a was about 100-fold dilute compared to the GST-tagged protein.

The results from both of these experiments demonstrate that our HT purification method is consistent with obtaining protein that is active in both enzymatic and inhibition assays.

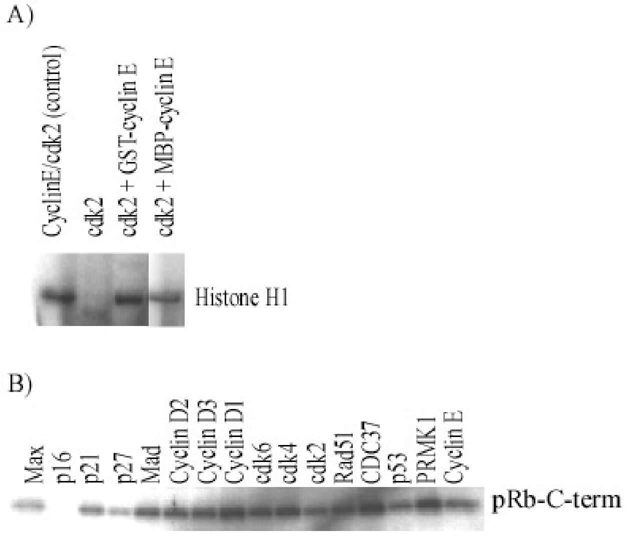


Figure 22: GST- and MBP- tagged proteins are functionally intact.
A) Equal amounts of bacterially purified GST- and MBP-tagged Cyclin E were added to GST-cdk2 purified from insect cells in Histone H1 kinase reactions. Both constructs activate cdk2 kinase activity.
B) GST-p16$^{\mathrm{Ink4a}}$ specifically inhibits CyclinD1/cdk4 kinase activity. Equal volumes of 16 GST-tagged test set proteins were added to kinase reactions using Cyclin D1/cdk4 purified from insect cells and C-terminal fragment of pRb as a substrate.

4.2.4. Summary

To develop methods for the high-throughput purification of human proteins for use in biochemical assays, a test set of 32 sequence-verified human genes of varying sizes and activities was employed. Using recombinational cloning, these 32 proteins were attached to four different affinity-purification tags: hexa-histidine, calmodulin-binding peptide, glutathione-S-transferase, and maltose-binding protein. By means of an automatable 2hr protein purification procedure, all 128 proteins were purified and subsequently characterized for yield, purity and losses. Under denaturing conditions using the His_6-tag, 84% of samples could be purified successfully as judged by a band of the correct size on an SDS-PAGE.

Under non-denaturing conditions, both the GST- or MBP- tags were successful in 81% of samples. In contrast, with the smaller His_6-tag and CBP-tag only 5 proteins purified, respectively. Analysis of the losses revealed that most of the His_6-tagged proteins were either lost in the flow through or could not be eluted. The CBP-tagged proteins bound very efficiently to the affinity matrix, but could generally not be eluted. The most prominent losses of the MBP tag were in the flow-through, whereas the losses of GST-tagged proteins were evenly distributed in all categories. Both purifications using the GST- and the MBP-tag are compatible with functional proteins.

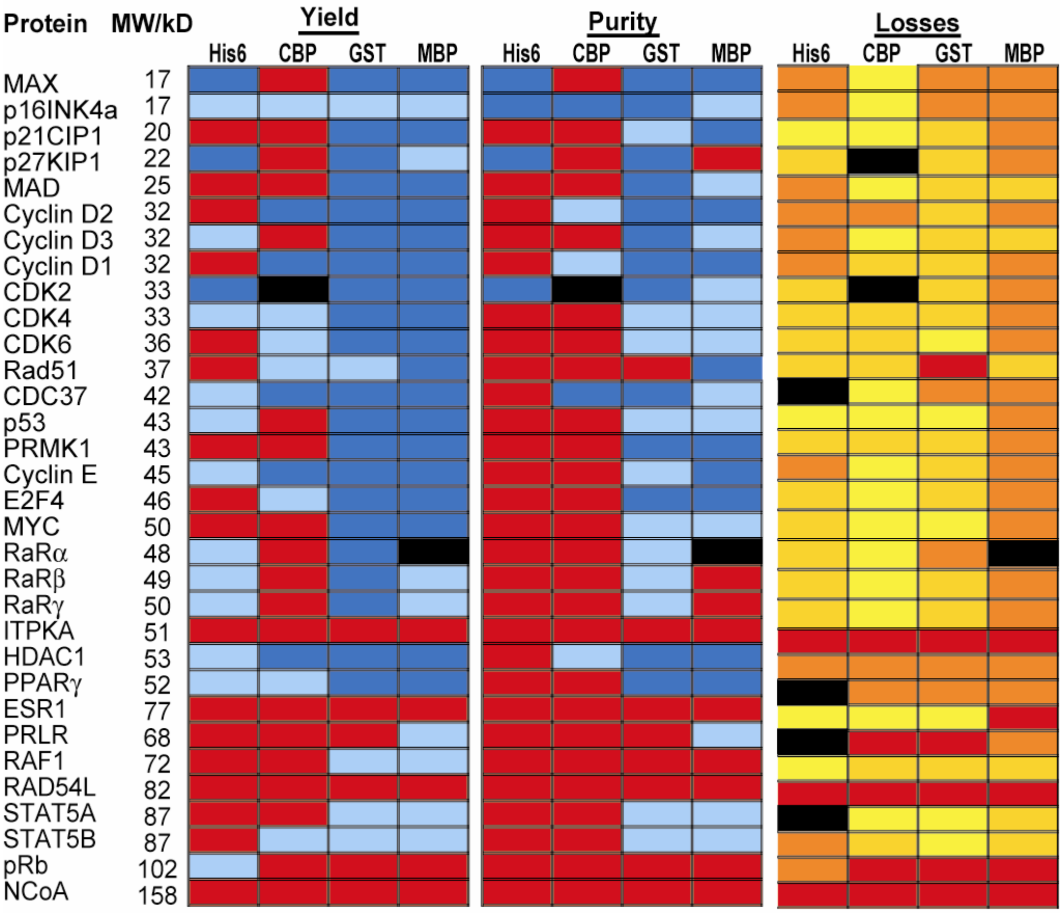


Figure 23: Summary of all test set protein purifications.

All 128 proteins were purified using the respective affinity tag and total yield and purity of the purified proteins were analyzed by GelCode® staining and image analysis. Losses were characterized by western blot analysis of five key fractions:

Yield: red: <300ng, light blue 300ng - 1µg; dark blue: >1µg.

Purity: red: <10% purity or no detectable band; light blue: 10% - 30% purity; dark blue: >30% purity.

Losses: red: protein degraded *in vivo*; orange: >60% of lost protein in the FT; dark yellow: losses evenly distributed between FT and matrix; bright yellow: >60% of lost protein was found on the matrix.

4.3. High Throughput Protein Expression and Purification

In order to evaluate the developed method in a true HT setting, we obtained Entry-clones for ~1000 different cDNAs. The cDNA clones came from different sources, which are listed in the following table.

Source	# Different cDNAs	Sequence Verification	His_6	GST
HIP – Set 1	337	PARTIALLY (in process)	337	189
DKFZ	484	YES	407	388
HIP – Set 2	85	YES	85	68
Total (non redundant)	854		773	428

Table 3: Starting cDNAs for the HT protein purification experiment.
The first column describes the source of the clones, column two the number if different cDNAs, column three indicates whether the clones were sequence verified and the last two columns indicate how many different cDNAs were processed with each protein affinity tag.

The first set of 500 clones was not sequenced verified, when the clones were obtained. Therefore 4 isolates were processed for each gene in context of the His_6-tag and purified under denaturing conditions (20 Plates). In addition sequence end-reads and analytical PCR were performed to confirm the identity of the clones. This analysis revealed that 1004 correct isolates had been processed, which corresponded to 337 different genes. 204 of these could be purified under denaturing conditions. The successfully expressed and purified clones are currently being sequence verified in collaboration with the German Resource Center. Most of the failed cDNAs have been sequence confirmed. Of the successfully purified proteins, one isolate for each gene, was processed using the GST-tag. After protein expression and purification under non-denaturing conditions, 153/198 cDNAs gave rise to a protein band of the expected size.

In addition, six plates of Gateway entry clones containing sequence verified human cDNAs were obtained from the German Resource Center (RZPD). One additional plate of sequence verified entry clones was obtained from the cloning pipeline at the Harvard Institute of Proteomics. These clones were processed under both denaturing and non-denaturing purification conditions. All transfer reactions were verified by analytical PCR. The results are summarized in the following table and the complete dataset for the German clones, as well as the successful purifications for the initial experiment are displayed in the appendix.

	#verified cDNAs	**# Successful Purifications**	**Fraction of Total**
His$_6$-tag (denaturing)	**773**	**517**	**67%**
GST-tag (non-denaturing)	**428**	**212**	**49%**
Total (non-redundant)	**854**	**598**	**70%**

Table 4: Success Rates of HT Protein Purification using two different affinity tags

The results show that 67% of all tested proteins could be purified from bacteria under denaturing purification conditions. Under non-denaturing conditions ~50% of all proteins were purified using the GST tag.

4.3.1. Denaturing versus Non-denaturing Conditions

In the test set of 32 proteins we had noticed that none of the proteins that failed under denaturing conditions could be purified with any tag under non-denaturing conditions. Thus we hypothesized that denaturing conditions reflect whether proteins can be expressed in bacteria at all. According to this hypothesis all proteins purified under non-denaturing conditions with the GST-tag should have scored positive under denaturing conditions with the His$_6$-tag as well. In addition, the relationship between non-denaturing and denaturing purification conditions was important to assess the magnitude of the bias, which had been introduced by selecting 200 proteins for GST purifications, which had been successfully purified under denaturing conditions using the His$_6$-tag (HIP set I).

To address the relationship of protein purification success between GST and His6-tagged proteins, those proteins which had been purified using both tags, were analyzed. The figure shows that the success rate for GST-tagged proteins was ~60% of those proteins, which were successful with the His6-tag and ~40% for those proteins, which had failed using the His6-tag. This result clearly demonstrates that the hypothesis, that failure of protein purification under denaturing conditions is predictive of purification failure with the GST-tag under non-denaturing conditions, is false. In addition, the data suggests that the bias, introduced by selecting for GST-purifications only proteins out of the first set, which had been successful under denaturing conditions, has introduced an comparatively small error.

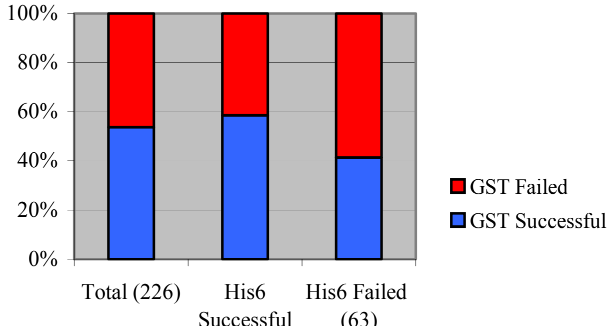


Figure 24: Purification success of GST-tagged proteins which were successful purified with the His6-tag or had failed with this tag.

4.4. Parameters Influencing Protein Purification Success

As the success rate for protein purification was, not unexpectedly, below 100%, we wondered whether it was possible to identify parameters that influence the purification success with either investigated tag. Such parameters would be helpful into ways. Firstly, they may assist the development of optimization algorithms that assign proteins to specific expression and purification systems before the protein production is started. Secondly, for sub-proteomic approaches it would be sufficient if biologically meaningful protein groups could be produced in bacteria.

In order to address the relationship between purification success and several protein purification parameters, a relational database was developed in collaboration with Dr. Y. Hu – a bioinformaticist at the Institute of Proteomics. The development of the data structure was a collaborative effort with significant input from the author of this study. Implementation of the database on a Microsoft Access platform was completely done by Dr. Hu.

4.4.1. Primary Structure

4.4.1.1. Protein Size

The great majority of bacterial proteins are smaller than 100kDa (>95%). Based on individual cases and anecdotal evidence it has been suggested that larger proteins are more difficult to express in bacteria than smaller proteins. To investigate this question with our dataset, the purification success for proteins in different size ranges was investigated for both tags.

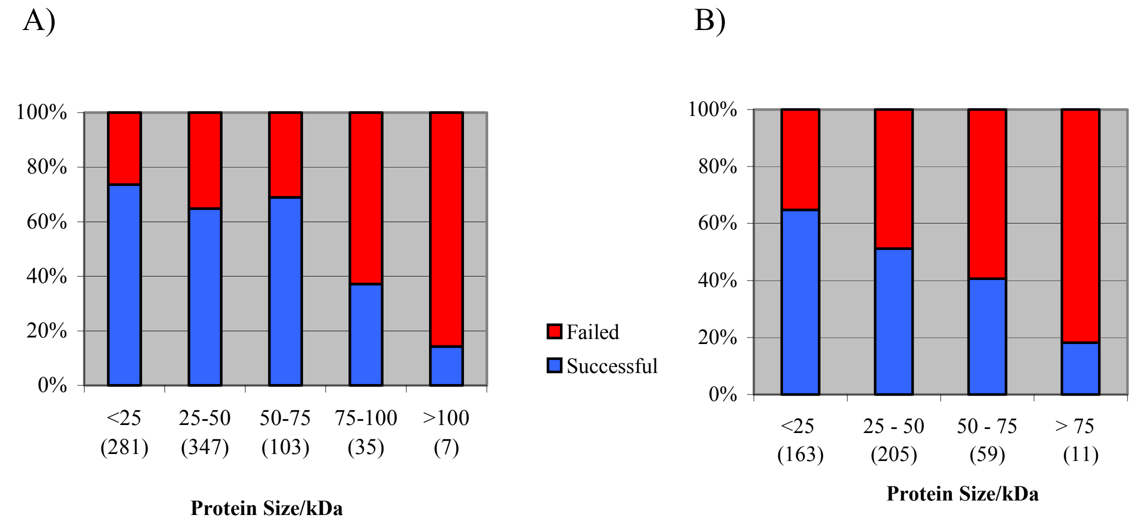


Figure 25: Size Dependence of Purification Success.
Plotted are the fraction of proteins of a given size range for A) Denaturing purification conditions and B) non-denaturing purification conditions using the GST-tag. Numbers in brackets indicate the number of proteins in this group.

The figure shows that proteins the molecular weight of the proteins affects their success rate. For His$_6$-tagged proteins the success rate of protein up to 75kDa is constant at approximately 65-70%. For larger His$_6$-tagged proteins the success rate drops rapidly. For small GST-tagged proteins (<25kDa) the success rate is 60%. In contrast to His$_6$-tagged proteins, however, the success rate is lower for proteins of 25-50kDa and drops further for larger proteins.

4.4.1.2. Amino Acid Composition, Isoelectric Point and Hydrophobicity

It seemed plausible that protein behavior in bacterial cells is influenced by the composition and biochemical properties of its constituent amino acids. Therefore, it was tested, whether the frequency of any amino acid in the target protein is indicative of the likelihood that the target can be successfully purified from bacteria. Some charged amino acids, Aspartate, Glutamate and Lysine had a positive influence on the purification success under denaturing conditions. Other charged amino acids, Arginine and Glutamine had a moderate to no influence in this experiment. Neither of the other amino acids exhibited a detectable effect on the purification success under denaturing conditions. Under non-denaturing conditions, no relationship between amino acid frequency and purification success was found.

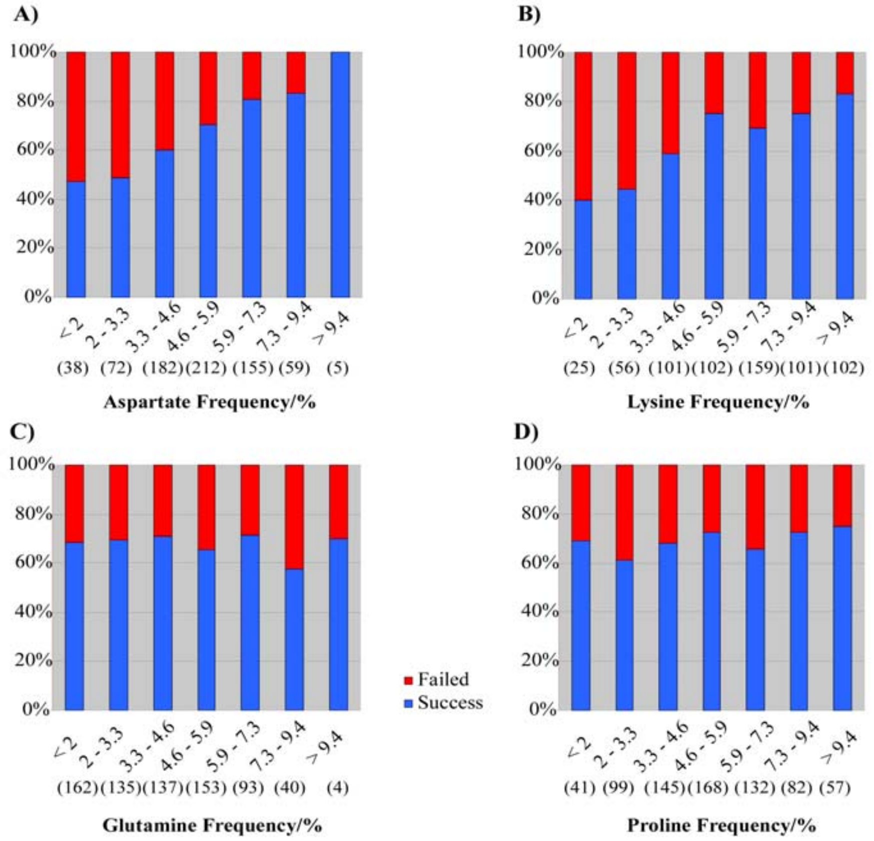
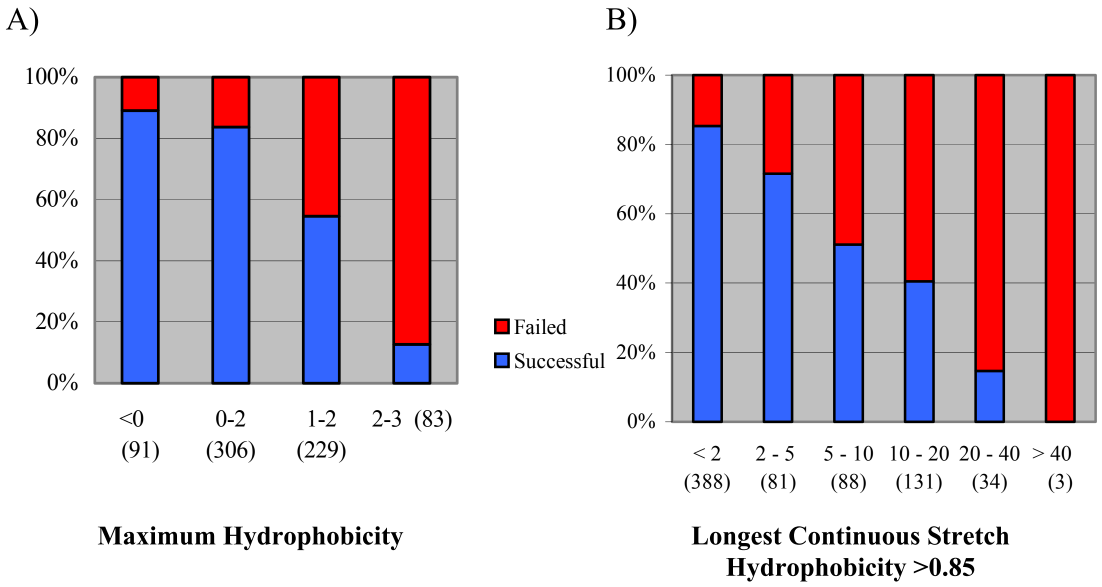


Figure 26: Amino acid frequency and purification success under denaturing conditions.
A) Aspartate B) Proline C) Glutamine and D) Proline frequencies. On the x-axis, proteins are grouped by frequency of the respective amino acids. The numbers in brackets indicate number of proteins in the group.

The isoelectric point of the protein did not have an impact on purification success under either denaturing or non-denaturing conditions.

Christendat et al. reported the purification of 500 archaeal proteins from *Methanobacterium thermoautotrophicum* (Christendat et al., 2000). After data mining they designed a decision tree, which allowed the identification of presumably soluble proteins by analyzing properties of their primary structure. At the first node in the decision tree, proteins are grouped into proteins, which have a hydrophobic stretch of 20 amino acids, which have a GES scale hydrophobicity > 0.85 (Engelman et al., 1986). Thus, the relationship between the hydrophobicity and purification success was investigated.



Maximum Hydrophobicity

Longest Continuous Stretch Hydrophobicity >0.85

Figure 27: Hydrophobicity and Purification Success under Denaturing Conditions. The hydrophobicity for every position of all proteins was calculated according to the GES scale using a window size of 20aa (ref). Number in brackets indicates the number of proteins in every group. A) Proteins were grouped by maximum hydrophobicity. B) Proteins were grouped by the longest stretch of a hydrophobicity >0.85. The numbers indicate the number of amino acids in the longest stretch.

For the denaturing purification conditions a clear relationship could be detected firstly between the maximum hydrophobicity and purification success and secondly between the length of the longest continuous stretch of amino acids with a hydrophobicity >0.85. This relationship was related, but not identical to the finding that membrane proteins were difficult to purify, an analysis omitting all membrane proteins gave identical results.

When the same analysis was performed for GST-tagged proteins, no such relationship could be found.

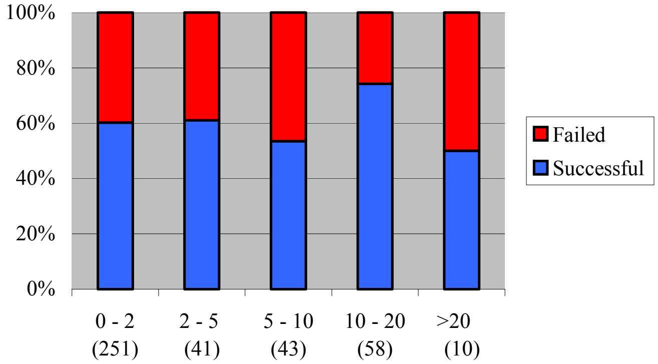


Figure 28: Effect of Hydrophobicity Purifications Using the GST-tag.
Relationship between longest continuous stretch with a hydrophobicity >0.85, and purification success under non-denaturing conditions. Number in brackets indicates the number of proteins in this group.

4.4.2. Structural Parameters

We investigated whether the three dimensional structure of the target protein influences, whether a protein can be made in the prokaryotic environment. Structural information was retrieved for the purified proteins from the SCOP database (Lo Conte et al., 2002; Murzin et al., 1995). 71 proteins in our set had an associated structure entry. These entries populated 57 different folds, so that no meaningful analysis could be performed.

4.4.2.1. Protein Domains

Proteins are composed of protein domains, which can frequently be discovered by sequence alignment. In order to investigate whether sequence defined protein domains have a detectable effect of the protein purification success, protein domain information from the Pfam and Smart databases were retrieved through LocusLink and the relationship to purification success was investigated (Bateman et al., 2002; Letunic et al., 2002; Sonnhammer et al., 1997). The data are displayed in the following two figures.

Under denaturing conditions targets proteins containing any of several protein domains had a very high or low chance of purification success. Expectedly, proteins with long hydrophobic stretches, such as transmembrane proteins failed to purify. However, kinases, small GTPases and several other protein domains correlated well with a good purification success under denaturing conditions.

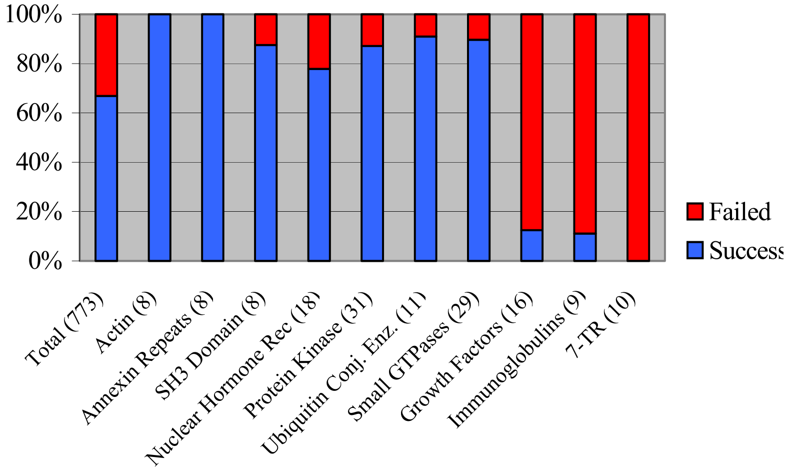


Figure 29: Correlation of protein purification success under denaturing conditions and proteins domains in the pfam and smart databases. Numbers in brackets indicate the proteins in these groups.

Under non-denaturing purification conditions, enrichments in some groups were detectable, although generally the data were either less clear than for denaturing purification conditions or supported by few data points.

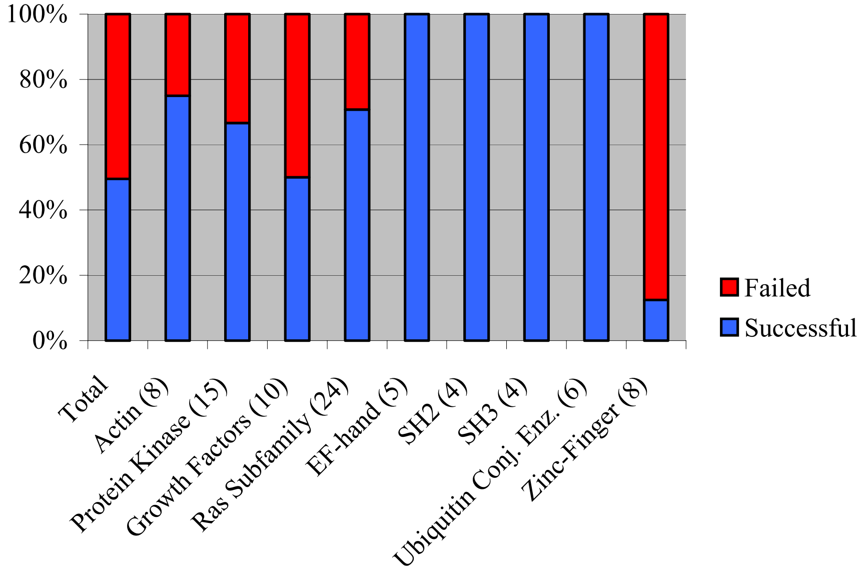


Figure 30: Correlation of protein purification success under non-denaturing conditions and proteins domains in the pfam and smart databases. Numbers in brackets indicate the proteins in these groups.

4.4.3. Parameters Related to Physiological Functions

The correlation of purification success and protein domains indicated that proteins, which fulfill certain functions may have a higher chance to be successfully purified from bacteria, than others. To test this hypothesis directly, the protein purification data were correlated to gene ontology annotation available from the Gene Ontology homepage and the LocusLink database (Maglott et al., 2000). Gene ontology annotation provides a standardized vocabulary for the functional annotation of genes (2001). The purification success with both purification tags was correlated to the biochemical function, cellular functions and physiological localizations of the expressed proteins. The analysis of GST-tagged proteins was complicated by the fact that only 270 of the 428 purified proteins were annotated by GO annotation. Consequently, the GST data gave less clear distinctions in any of the analyzed categories than the data from purifications of His_6-tagged proteins.

When the physiological roles of the purified proteins were analyzed, proteins, which fulfill elementary functions on the cellular level such as energy generation and metabolism of fundamental building blocks, tend to have higher success rates (than average), whereas proteins which fulfill functions that evolved relatively late, such as apoptosis or cell-to-cell signaling tend the have lower than average success rates.

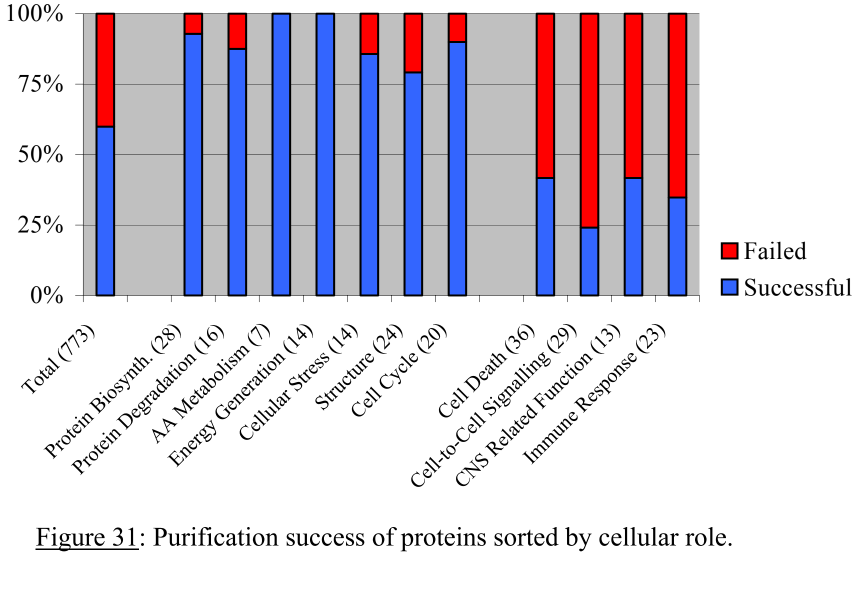


Figure 31: Purification success of proteins sorted by cellular role.

4.4.3.1. Biochemical Function

The strong positive correlation between purification success and protein domains prompted us to investigate whether a similar correlation would be detectable between the biochemical function of proteins and their purification success. Such a correlation would be of use for focused proteomics approaches.

In order to further investigate the relation of the function and purification success the relationship to biochemical functions was investigated. The figure shows that many investigated enzyme groups, such as hydrolases or DNA and RNA binding proteins have a higher have a higher than average likelihood of successful purification under denaturing conditions. The heterogeneous group of receptors includes transmembrane receptors which have already found to be hard to purify.

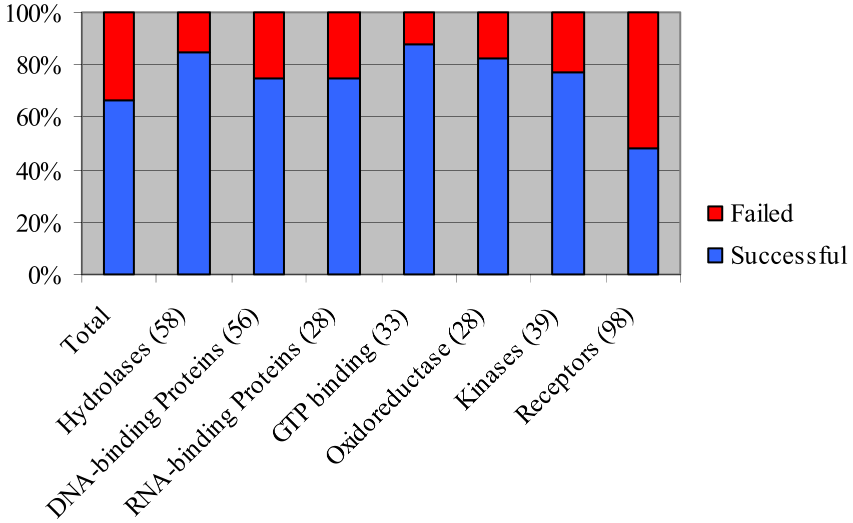


Figure 32: Protein purification success under denaturing conditions of proteins with common biochemical functions.

The analysis of GST-tagged proteins did not yield clear results, because most groups contained only 5 or less proteins.

4.4.3.2. Physiological Localization

In the analysis of the primary structure of the proteins, it had been noticed that the length of hydrophobic stretches in the amino acid sequence can bias against a purification success under denaturing conditions. Such stretches are more likely to occur in transmembrane proteins. To investigate whether further relationships can be found, the physiological localization of all purified proteins was analyzed and correlated to the purification success. Figure 25 illustrates that proteins localizing to mitochondria, the cytoplasm nucleus and ribosomal proteins have a very high chance of getting purified with the His_6-tag. In contrast, only a small fraction of proteins that are associated with or integrated into membranes can be purified with the His_6-tag.

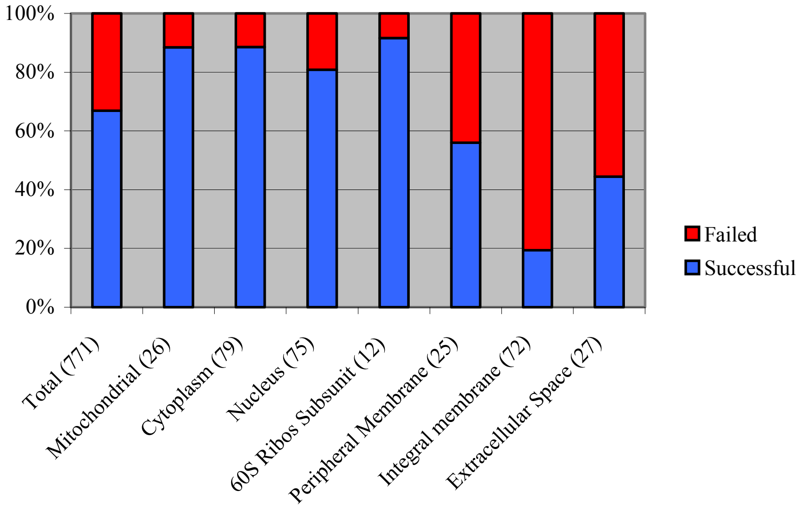


Figure 33: Subcellular localization of protein purified using the His_6-tag.

When GST-tagged proteins were analyzed the results were generally supported by fewer data points. A large fraction of cytoplasmic proteins could be purified (5/6), however only two of the eight nuclear proteins could be purified. In addition four out five proteins associated with the cytoskeleton were successful. Furthermore, two of the seven transmembrane proteins were successfully purified.

4.4.3.3. Codon Composition of the cDNA

Even though the genetic code is fundamentally identical in all kingdoms of life, some subtle differences exist in the way species are utilizing different codons that encode the same amino acid. This phenomenon is called ‘differentia codon usage’ and some reports indicate that the differential codon usage between *Homo sapiens* and *E. coli* hamper the high level expression of some cDNAs, which are rich in codons that are only infrequently used in the bacterium (Ivanov et al., 1997; Zahn, 1996).

To test whether differential codon usage affects the success rate of HT protein purifications, we related the purification success of all proteins to the fraction of rare codons (AGA, AGG, CCC) in the respective coding sequences. The figure shows that there was only a moderate influence when the total number of all codons or of each was plotted for purification with either tag. The analysis of the frequency of any of the three individual codons showed no effect on protein purification success. When the proteins were grouped by the frequency of any of the rare codons, all groups contained 60% - 70% successfully purified proteins. The variation was randomly distributed and no trend was detectable.

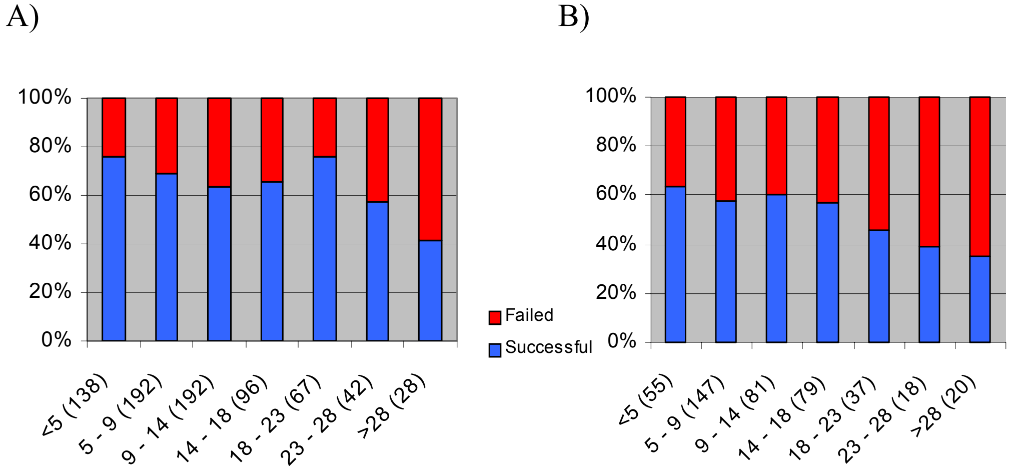


Figure 34: Effect of the Number of Rare Codons on the Purification Success.
Proteins were grouped by the total number of the codons CCC_{Pro}, AGG_{Arg}, and AGA_{Arg} and the fraction of proteins that were successfully purified indicated for A) denaturing (His_6) and B) nondenaturing conditions (GST).

4.4.4. Summary

In order to evaluate the success rate of the developed HT protein purification method with a large number of samples, 770 and 428 different cDNAs were processed with the His_6-tag and GST-tag, respectively. After purification under denaturing conditions 67% of proteins were purified and after purification under non-denaturing conditions approximately 50% were purified successfully.

In order to identify parameters that influence the purification success of target proteins from bacteria, we correlated the purification data to biochemical, biophysical and biological data about the purified proteins. This analysis revealed that several parameters have an impact of the purification success using the His_6-tag and denaturing conditions. These parameters include the size of the target protein, its hydrophobicity, it content in selected amino acids and its physiologic function. For purifications using the GST-tag, only the size of the target proteins could be identified as a parameter that influences purification success.