1. Introduction

1.1. High Throughput Experimentation

1.1.1. Foundations of Biological High Throughput Experimentation

Undoubtedly the human genome project has been the crib for HT biology. The idea for the human genome project was born at the so called Alta Summit 1984 where participants had gathered to address the problem of a too low sensitivity of existing technologies to detect mutations in survivors of the Atomic bombs dropped on Hiroshima and Nagasaki (Cook-Deegan, 1989). The consensus among the attendees was that a colossal project was necessary to address this question: the sequencing of the entire 3 billion bases of the human genome. The Human Genome Project (HGP) was eventually launched in 1990 by the American Department of Energy and the National Institutes of Health, and was intentionally conducted as an international collaboration to express the universally human inheritance of the genome. Celera Genomics, a Rockville (Md) based Biotech Company, announced an independent, private genome sequencing project in 1998.

In the beginning the Human Genome Project faced heavy criticism for being too expensive, technically impossible and of very limited scientific value (Roberts, 2001). Today it has become obvious that the Human Genome Project is revolutionizing and transforming biological and biomedical research. The basic genomic sequence data of humans and a rapidly increasing number of organisms have been supplemented by data about human genetic variations, so called single nucleotide polymorphisms (SNPs), and by information about all protein coding regions (Holden, 2002; Strausberg et al., 1999) .

The fall-out from the collection of these data has been immense and can be broadly categorized as giving birth to the fields of genomics, functional genomics and functional proteomics and transforming the possibilities in analytical proteomics. For this development the comprehensive knowledge of gene inventories and the introduction of automation and massive bioinformatic analysis to biological research have been key

innovations. The following paragraphs will discuss gene inventories and their impact on biological experimentation as well as the necessity for technology development.

1.1.1.1. Gene Catalogues – Boundaries of the Biological Universe

Of central importance to understanding biology is the question of how many and which proteins are encoded in any given genome. At the start of the human genome project estimates about the actual number of human genes ranged from 25.000 to 150.000 genes (Dunham et al., 1999; Fields et al., 1994). After assembly of the human genomic sequence drafts in 2001, the sequence was analyzed to detect protein coding portions and regulatory elements. The actual criteria used to identify protein-coding regions included known translation and splice signals, EST information, codon usage, CpG islands and homology to known proteins (Burge and Karlin, 1997; Fujibuchi and Kanehisa, 1997; Guigo et al., 1992; Kleffe et al., 1998; Makalowska et al., 2001; Milanesi et al., 1999; Xu and Uberbacher, 1996). These studies predict that *S. cerevisiae* has about 6200 genes (Goffeau et al., 1996), *D. melanogaster* 14.000 (Myers et al., 2000), *C. elegans* about 19.000 (1998), and humans 30.000 - 40.000 genes (Lander et al., 2001; Venter et al., 1998). Recently, the alignment of the human genome with the complete mouse genomic sequence, in conjunction with a data from cDNA sequencing projects has supported the estimates of approximately 30.000 human genes and has refined predictions of genomic structure (Strausberg et al., 1999; Wasserman et al., 2000).

A study in *C. elegans* indicated that the weakness of gene prediction algorithms for the worm is not the detection of ORFs. Rather the prediction of the exon-intron structure of genes is difficult if the corresponding cDNA has not previously been characterized (Reboul et al., 2001). This result is supported by Venter et al. who note that known genes have an average of 7.8 exons, whereas newly predicted genes only contain 3.7 exons and many novel genes contain only one (Venter et al., 1998). As a consequence of inaccurately predicted gene structure two exons that have been attributed to different genes may actually belong to one gene and vice versa. Related to the issue of gene structure is the almost complete ignorance about differentially spliced transcripts originating from a single gene. This dimension of biological complexity is currently

difficult to access and a presents a significant gap in our knowledge about biological systems.

The predicted gene catalogues can be translated *in silico* into putative protein inventories of the respective organism. When these predicted protein inventories were analyzed for homology to previously studied proteins, one central finding was that only ~50% of all genes have been studied, or have significant homology to a gene that has been studied in any model organisms (Costanzo et al., 2000). Thus, a major goal of the post-genomic era is the elucidation of the function of all the unknown gene products. With help of the predicted gene catalogues it is possible to make predictions, based on homology, about the function of the encoded gene. Figure 1 shows a comparison of mouse and human (predicted) proteome functionalities by gene ontology annotation (Waterston et al., 2002).

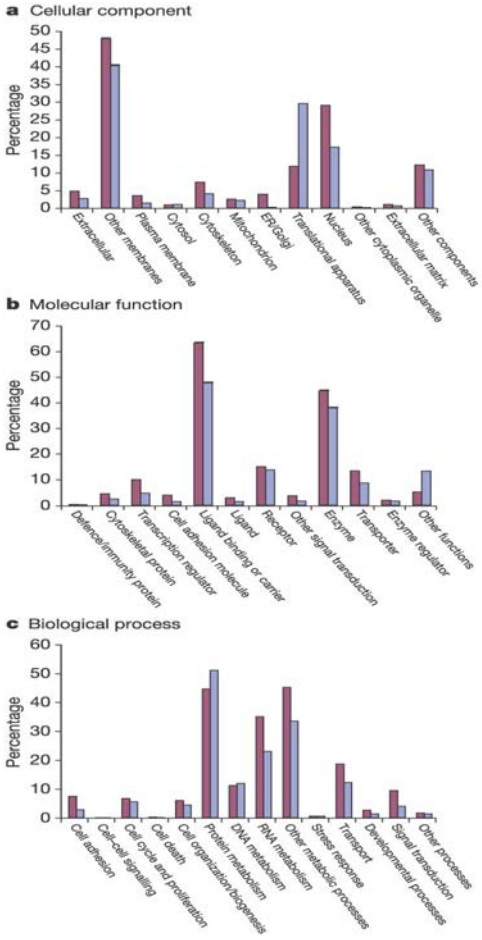

Figure 1: Gene ontology (GO) annotations for mouse and human proteins. The GO terms assigned to mouse (blue) and human (red) proteins based on sequence matches to InterPro domains are grouped into approximately a dozen categories. These categories fell within each of the larger ontologies of cellular component (**a**) molecular function (**b**) and biological process (**c**) (D. Hill, personal communication). In general, mouse has a similar percentage of proteins compared with human in most categories.

Thus, although of preliminary nature and subject to revisions and additions, complete gene catalogues indicate the boundaries of the biological universe[1] for every organism and have been a starting point for comparative analyses and for a new dimension in the possibilities of genome-wide experimentation (Brown and Botstein, 1999; Gopal et al., 2001).

[1] The expression 'biological universe' has been coined by Pat Brown and David Botstein (1999) to picture the multidimensional space, which accommodates all genes, proteins and all interactions and functions, which are now accessible to investigation.

1.1.1.2.1. Genome-wide Functional Experiments

Genome-wide experiments are not a novelty of recent years. However, the knowledge of complete gene inventories has transformed them from simple pre-genomic screens to global mapping experiments and system-wide determination of transcriptional dynamics.

In pre-genomic screens, pools of genomic or cDNA libraries were used in many biologic screens. As a result of a compositional bias in the utilized libraries a multiple oversampling was necessary to ensure adequate coverage of the genome. This oversampling consumed valuable resources and posed a daunting technical problem, which limited the complexity of available assays and made a comprehensive data collection impossible. In 2001 Stevenson et al. used both a genomic library and a normalized yeast cDNA library to assay 30.000 plus 150.000 samples from the two libraries respectively to approach genome-wide coverage in a screen for gene products that cause cell cycle arrest when overexpressed. Importantly, even this extensive screening did not result in the repeated identification of all positives – an indication of saturation and a common criterion for comprehensiveness (Stevenson et al., 2001). For human studies the problem is orders of magnitude more severe. Stevenson's experiments indicate that a 25-fold oversampling is insufficient for saturation. Thus, for the human genome, assuming 30,000 different genes, at least 750.000 events will have to be scored to obtain a similar coverage. Furthermore, whereas only few percent of *S. cerevisiae* genes contain introns and thus are spliced, humans use extensive splicing to create numerous different proteins from one gene. As indicated above, the average (known) human gene contains approximately 7.8 exons. It can easily be seen that the numbers required to perform a comprehensive screen for *Homo sapiens*, will be impossible if traditional cDNA libraries are employed.

The knowledge of gene inventories has enabled the assembly of indexed collections of cDNAs or deletion strains, which are one-to-one representations of all genes in a genome (Reboul et al., 2001; Tong et al., 2001a; Uetz et al., 2000). Such genomic reagents allow comprehensive assays to be carried out at the least possible cost and effort. Often the samples in these collections are kept separate and this setup enables a comprehensive data collection and the acquisition of global datasets. Thus, as a result of

indexing and automated data collection, more detailed phenotypes may be recorded in a given assay (see below).

For proteomics, the knowledge of all genes has provided common reference, which has greatly enhanced the utility of proteomic profiling. Without the availability of gene catalogues, peptides that are identified in a proteomic profiling experiment are often unconnected data points, for which it is difficult to identify the protein from which they originate. In contrast today, the identification of proteins can most often be done by automatable database searches against translated gene catalogues (Cash, 2002).

1.1.1.2. Technology Development

Both critics and advocates of the Human Genome Project realized that a requirement for a success of the HGP was the development of sequencing technology, automation and information technology. Thus, a major objective of the HGP was the development of technologies that would allow fast and efficient sequencing (*NIH Publication No. 90-1590*) and throughout the project a significant part of the budget was devoted to develop sequencing, automation and information technology. Besides enabling the timely completion of the Human Genome Project, the massive investment in technology development spurred a new enthusiasm for novel technologies in biology. Two factors have contributed to this trend. Firstly, the advances and byproducts of sequencing technology development (e.g. DNA microarray technology (Eickhoff et al., 1996; Khrapko et al., 1989; Parinov et al., 1996)) have started to revolutionize biology by providing a radically new perspective on biological phenomena (described in the next section). It is widely expected that more technologies will expand this revolution to other aspects of molecular biology and transform our understanding of biological regulation. Secondly, the knowledge of complete gene inventories has spurred the desire to elucidate their function and to do so quickly. Traditional technologies, which are practical for one or a handful of samples, are obviously not adequate for this task. Instead it is widely accepted that novel, highly multiplexed, high throughput technologies are required for the post-genomic era (Fields et al., 1999). General advantages and disadvantages of high throughput experimentation are discussed in the immediately following sections. The

particular applications and difficulties of high throughput protein purification will be discussed further below in the respective chapter.

1.1.2. The Potential of High Throughput Experimentation

One of the first technologies for the parallel study of a molecular parameter has been the hybridization of nucleic acids on glass slides or 'DNA microarray technology', which enables the simultaneous analysis of the abundance of thousands of different mRNA transcripts. In conjunction with knowledge of complete gene catalogues this technology for the first time enabled the global analysis of molecular events underlying physiological and pathologic processes. Because of the pioneering role of this technology, it will serve here as an example to illustrate the power of high throughput experimentation.

The impact of transcriptional profiling can be categorized as the rapid generation of hypothesis about the function of individual gene products, the analysis of global molecular events underlying physiological phenotypes and the analysis of signaling networks. First, transcriptional profiling technology will be introduced briefly.

1.1.2.1. DNA Microarray Technology

In DNA microarray technology, fluorescently labeled experimental nucleic acid samples are hybridized to either oligonucleotides or cDNAs which are immobilized on glass slides. In cDNA based microarrays, nucleic acids from 2 different experimental samples are labeled with two different fluorescent probes, mixed, and hybridized to cDNAs on the glass slides. The ratio of the two different fluorescent probes at any spot indicates the relative abundance of this nucleic acid species in the two samples (Duggan et al., 1999).

Oligonucleotide based microarrays employ up to 20 different oligonucleotides (and 20 respective controls) to measure the abundance of a particular cDNA. For each oligonucleotide the fluorescence is measured independently and transformed into a numerical value. The numerical values from all 20 oligos are averaged and normalized, and thereby transformed into a numerical value for transcript abundance. The normalized numerical values from different arrays, i.e. different experimental samples, can be

compared to extract information about the dynamic behavior of the measured cDNA (Lipshutz et al., 1999).

The main application for DNA microarray technology has been transcriptional profiling. However, significant insights have been gained by alternative microarrays, which are used technically identically but have been constructed by spotting yeast intragenic genomic DNA (Ren et al., 2000), human BACs for comparative genome hybridization (Pinkel et al., 1998) or oligonucleotides that are suited to detect human SNPs (Cutler et al., 2001). The following section describes the insights and progress made possible by DNA microarray technology.

1.1.2.2. Gene Annotation and Hypothesis Generation

In all sequenced eukaryotic genomes up to fifty percent of all genes are unknown and have no homology to any studied gene (Costanzo et al., 2001). Thus, one important goal of HT experimentation is the generation of hypothesis about the function of the analyzed samples. Transcriptional profiling data have been utilized to elucidate functions of genes by two broadly distinguishable mechanisms, manual inspection and automated information extraction.

Many transcriptional profiling data sets have been manually inspected to identify transcripts whose products may execute a long sought-after function or at least a function that ‘makes sense’ in a certain biological setting (Aitman et al., 1999; Clark et al., 2000). Clark et al. utilize microarrays to investigate physiological changes during metastasis formation of tumors. From their data they hypothesize, that rhoC could play a central role in the pathogenesis of metastases. They support this hypothesis by showing that overexpressed rhoC enhances metastatic behavior of tumor cells, whereas dominant negative rhoC inhibits metastasis formation without affecting proliferation (Clark et al., 2000).

The process of investigating genes one at a time yields very accurate results, but it is also a laborious process. Analysis of transcriptional data from previously characterized genes indicated that eukaryotic genes, analogous to genes in bacterial operons, are often co-regulated when they functionally act together (Eisen et al., 1998; Wu et al., 2002). It was suggested that these relationships could be exploited to imply a physiologic function

for previously uncharacterized genes (Lander, 1996). To this end Wu et al. exploited a dataset consisting of 300 microarray experiments to build a database of clusters of co-regulated genes. Next, gene annotations from hand-curated databases were imported. Based on existing functional annotation, all clusters were assigned a p-value, which indicated the confidence that this cluster contains meaningful biological information. This database was then utilized to make functional predictions for unknown genes. Using various evaluations and experimental tests the authors estimate that between 30% and 50% of their predictions are correct (Wu et al., 2002). Thus the HT analysis of gene function can greatly accelerate the creation of hypotheses about the functions of individual genes.

1.1.2.3. Global Molecular Phenotypes

Until recently the analytical possibilities of molecular biology have only allowed the study and investigation of a limited number of genes and proteins simultaneously. Consequently, the data obtained in this process only informed about small aspects of the investigated phenomenon and analysis of wide-range effects was difficult or impossible. In contrast, DNA microarray technology enables the analysis of all transcripts expressed in a cell or tissue[2] at any given time and thus provides a global view onto molecular events in any natural, experimental or pathologic process. In several instances the greater detail has lead to novel insights about physiological or pathological processes.

One of the first genome-wide profiling experiments investigated the effects of glucose depletion, a diauxic shift, in *S. cerevisiae* (DeRisi et al., 1996). *S. cerevisiae* prefers glucose as its energy source. However, upon depletion of glucose in the environment it can utilize other carbon sources such ethanol. Such a change, which is called a diauxic shift, is accompanied by a slower metabolic rate. The authors find that most proteins of the well characterized glucose metabolic pathways (Pentose-phosphate pathway, Glycolysis, Krebs-cycle, Glyoxylate cycle) respond to diauxic shift and that the up- and down-regulation of the respective genes can be well explained by the changing

[2] Current DNA microarray technology only enables the characterization of all transcripts for organisms with small gene catalogues like S. cerevisiae or C. elegans. While this limited coverage is sufficient for some applications, it the acquisition of global pictures and consequently the confidence that all relevant aspects have been analyzed.

demands of energy source switching (Figure 2). In parallel, many proteins of the cytosolic biosynthetic machinery were down regulated as expected by a slower rate of biomass production under these growth conditions. This study illustrates how transcriptional profiling allows insights into the global changes in the cellular repertoires of enzymes, structural components and signaling molecules, which collectively define physiological states. In addition, transcriptional profiling experiments identify groups of genes, which may be of importance for a biological process of interest and thus suited for further focused high throughput approaches (see below).

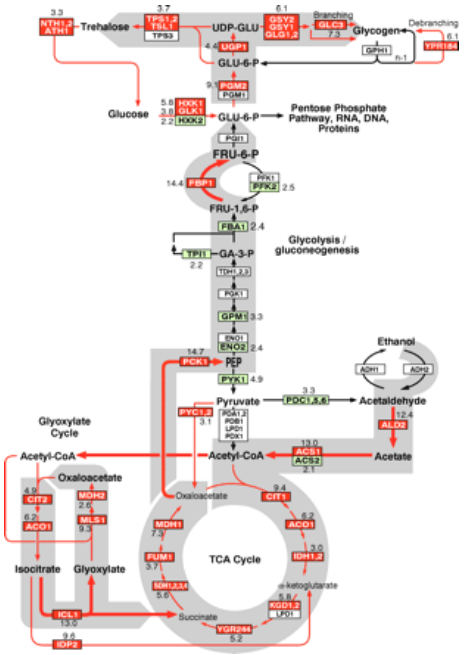

Figure 2 (from DeRisi et al, 1997): Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

In medicine, the enhanced phenotypic resolution of DNA microarrays has allowed the separation of previously indistinguishable tumor types (Alizadeh et al., 2000; Bittner et al., 2000; Perou et al., 2000; Ross et al., 2000). The classification based on the molecular data often correlates well with a differential prognosis of the patients. In a well known study by Armstrong et al. the transcriptional profiles of two leukemia subtypes were compared. The two investigated subtypes, acute lymphoblastic leukemia (ALL) and acute myelogenous leukemia (AML) are clinically difficult to distinguish and require significantly different treatments. The authors find that the transcriptional profiles of ALL can be differentiated unambiguously and robustly from those of AML. Further

analysis even revealed that the morphologically defined class of ALL contains a second leukemia subtype (MLL), which can be similarly well distinguished by its transcriptional profile and which correlates with particularly bad prognosis (Armstrong et al., 2002). Currently efforts are underway to identify small sets of genes that can serve as clinical markers for the different leukemia types. Furthermore, the profiles or a set of selected markers are currently used in profile-driven drug screening.

1.1.2.4. Network Analysis

Biological responses to stimuli are mediated by complex regulatory networks, which integrate physiological and external conditions to form a response to signals and changes in external or internal conditions. While some properties of regulatory networks have been uncovered by traditional techniques, global approaches hold particular promise for this line of research. Global approaches have the chance to systemic and combinatorial effects, which have not been anticipated and are difficult to detect with traditional technologies. DNA microarray data have been exploited to shed light on the combinatorial mechanisms regulating transcription. The networks most accessible by DNA microarray technology are transcriptional regulatory networks, although, signal transduction cascades have also been analyzed (Fambrough et al., 1999; Roberts et al., 2000).

Virtually all of the early yeast profiling data have been used to identify regulatory motifs in upstream regions of genes involved in processes like diauxic shift (DeRisi et al., 1996) and cell cycle regulation (Banerjee and Zhang, 2002; Bussemaker et al., 2001; Eisen et al., 1998). This approach has been extended to analyze more complex relationships between regulatory motifs and attempt to detect combinatorial regulation and thus delineate regulatory networks (Banerjee and Zhang, 2002; Pilpel et al., 2001). Pilpel et al. design a database of known and putative regulatory motifs and search the yeast genome sequence for genes, whose upstream regions contain one or more of these motifs. Analysis of several microarray data sets for synergistic and antagonistic effects of motif combinations reveals both known and novel synergistic motif combinations. The authors can draw a map of motif-synergies and identify nodes, which correspond to binding sites for transcription factors with known pleiotropic functions. Additionally,

synergistic-motif combinations cluster together and these clusters correspond to transcription factors, which all function in the context of one physiologic process, e.g. cell cycle control or stress response (Pilpel et al., 2001). Thus, using transcriptional profiling, potential large range and combinatorial interactions between transcriptional regulators have been identified.

Measuring global transcriptional responses can be used to analyze system-wide consequences of an experimental manipulation and to investigate the complexity of upstream signaling networks. A study by Leroy Hood and co-workers illustrates this point. Ideker et al. aim to investigate the very well-characterized Gal4 metabolic pathway in yeast using a set of high-throughput technologies, including transcriptional profiling. The authors employ a panel of yeast strains deleted for different elements on the Gal4 metabolic pathway and analyze the transcriptional profiles. Analyzing the various data sets, the authors realize that various observations cannot be explained by the current model. One such observation is that *Gal7* and *Gal10* deletion strains, in the presence of Galactose, show reduced expression of several Gal genes. The effect can be explained by a model in which Galactose-1-phosphate (Gal-1-P), exhibits a regulatory effect on the core regulatory unit (see Figure 3). The authors test this hypothesis by analyzing the *Gal1Gal10* double deletion strain, which should not show the reduced expression levels, because Gal-1-P is not made and thus cannot accumulate. This hypothesis could be confirmed (Ideker et al., 2001). This example illustrates the power of global experimentation to further our insight even in the most extensively studied metabolic pathways.

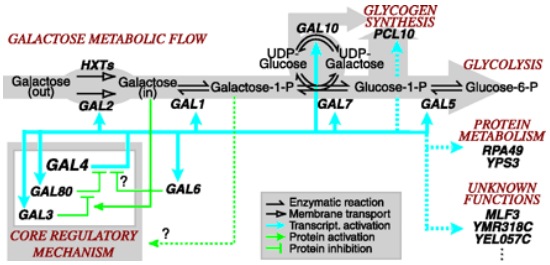

Figure 3 (from Ideker et al, 2001): Model of galactose utilization. Yeast metabolize galactose through a series of steps involving the GAL2 transporter and enzymes produced by GAL1, GAL7, GAL10, and GAL5. These genes are transcriptionally regulated by a mechanism consisting primarily of GAL4, GAL80, and GAL3. GAL6 produces another regulatory factor thought to repress the GAL enzymes in a manner similar to GAL80. Dotted interactions denote model refinements supported by this study.

1.1.3. Limitations of High Throughput Experimentation

Global experimentation has opened the doors for the investigation of large scale effects and faster annotation of gene function. However, despite the significant enrichment of the possibilities of biological experimentation, it is important to keep the limitations of high throughput experimentation in mind: the presence of false data points, which will be illustrated at the example of protein-interaction mapping and the loss of detail about any single gene product.

1.1.3.1. False Data Points

Most biological assays yield a certain fraction of false datapoints, for example readouts that indicate a biological relationship when there is none. This phenomenon can be worse when experiments are executed in parallel, like in screens or HT assays, because the conditions for many samples are sub-optimal, each sample is less well controlled in comparison to small-scale experiments and because an assay may be inappropriate for a subset of gene products. For the yeast-two-hybrid assay Walhout et al have estimated that the rate of false positives and false negatives is ~50% (Walhout et al., 2000a; Walhout and Vidal, 2001). In order to obtain a more reliable picture of biological relationships, traditional experimentation exploits several complementary approaches to obtain mutually enforcing data (LaBaer et al., 1997). The same approach has been started in HT experimentation. The best example is protein-interaction mapping in *S. cerevisiae*, where four subsequent HT analyses confirmed many known and discovered numerous novel protein interactions. Uetz and co-workers published the first 'comprehensive' interaction network in 2000. In their paper, the authors report 957 interactions involving 1004 different proteins (Uetz et al., 2000). A second 2-HA study conducted by Ito et al identifies an additional 4549 interaction among 3278 different proteins. Interestingly, these interactions only marginally overlap with the first data set (Ito et al., 2001). Gavin et al and Ho et al. both use an MS based approach to analyze several thousand protein interactions in yeast and discover thousands of novel interactions (Gavin et al., 2002; Ho et al., 2002). This example illustrates the need for complementary technologies and the insufficiency of any one technology to give final answers even about one parameter.

In order to improve the overall data quality the combination of several data sets, which provide both similar as well as different types of information, may be a promising approach. Ge et al. have shown that proteins, found to be co-regulated in transcriptional profiling experiments, have a higher likelihood of physically interacting (Ge et al., 2001). Thus, protein interactions found in clusters of co-regulated genes have a higher probability of being 'real' than proteins whose transcripts are not co-expressed.

1.1.3.2. Perspective versus Detail

The global pictures obtained by high throughput technologies have made a lasting impression on biologists. Undoubtedly, the methods for parallel information collection must be applied on a proteome-wide scale to catalogue the subcellular locations of all proteins (Simpson et al., 2000) to catalogue all enzymes and their specificities, and to delineate interaction networks (Gavin et al., 2002; Ho et al., 2002; Uetz et al., 2000; Walhout et al., 2000a). The resulting maps of biological functionalities will provide a crucial groundwork that will greatly advance our understanding of biology and diseases.

At the same time it must be kept in mind that the increase in perspective comes at the expense of only a very small increase in information about any single gene or gene product (see figure 4). Naturally, the function of genes or proteins cannot be deduced from a single data point in any single method. Thus, because global experimentation cannot collect all relevant parameters of all proteins in all relevant experimental designs, traditional experimentation and importantly focused high throughput studies will be crucial aspects of biomedical research.

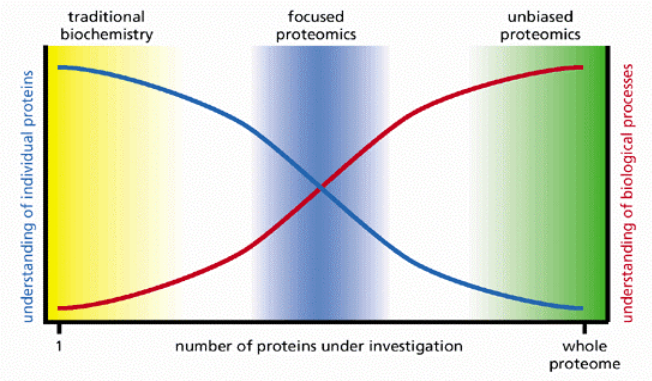

Figure 4 (from Macbeath, 2002): The spectrum of protein analysis, ranging from traditional biochemistry to unbiased proteomics. Focused proteomics occupies a middle ground, where one seeks to maximize both the depth and the breadth of biochemical investigation.

Several examples of focused high throughput experimentation have been generated by Vidal and colleagues.

Boulton et al. analyze the DNA-damage response (DDR) module of the nematode *C. elegans* using different high throughput approaches. First the authors selected 75 gene products, which have either been directly involved in DDR or are homologues of proteins in other organisms, which have been involved in this function. The authors test all pair wise combinations in a protein interaction matrix. In this experiment they can find 17 of 33 expected interactions (51%) and the authors point out that this false negative rate is close to previous estimates. Subsequently, 67 of the 75 selected proteins are used for 2-hybrid screening against a cDNA library. In this experiment the authors find 165 novel protein interaction partners, including previously identified interaction and novel ones. On a protein interaction map, the authors find that proteins involved in similar functions like damage sensing tend to be connected to each other by more protein interactions than to proteins that fulfill other functions. Recognizing the potential for false data, the authors then start to verify the novel protein interaction partners in genetic experiments using RNA mediated interference. Using 'RNAi by feeding' the authors interfere with the expression of all identified genes and score four different phenotypes. 23 of the tested genes confer a DDR phenotype and 11 of these genes have not been reported previously to be involved in DDR. Addressing the other 90% of genes, which did not show a DRR phenotype, the authors speculate that the used assay may only be suited to detect a subset of genes and is inert to defects in base excision or nucleotide excision (Boulton et al., 2002). This study clearly demonstrates the power of focused high throughput approaches in delivering large data sets, which provide a birds-eye perspective, deliver partially verified data of higher quality and increase the depth of the study.

1.1.3.3. Limitations of Genomic Approaches

The most prominent high throughput technologies investigate nucleic acids. Transcriptional profiling and other genomic approaches, like systematic gene inactivation by deletion or RNA mediated interference (Ashrafi et al., 2003; Fairhead et al., 1998; Tong et al., 2001b), have opened the doors for a system-wide analysis of biological phenomena and a deeper understanding of physiological and pathological processes. However, at the same time, it became clear that the molecular understanding of biological complexity requires a better understanding of molecular events and interactions on the

protein level. The results obtained with DNA microarray technology indicate how much more there is to learn about the complex interactions that regulate gene expression and physiologic responses. Most importantly, in order fully capitalize on the analytic capabilities of DNA microarray technology it is necessary to know the biological and biochemical functions of the up- and down-regulated genes and their mutual interactions. This knowledge is not accessible from transcriptional profiling data or any other type of genomic data, even though hypotheses about potential functions can be deduced from co-expression data. However, any co-expression can be incidental or functionally insignificant and thus complementary technologies are required to verify or falsify such hypotheses. Furthermore the involvement of a protein in a particular cellular function does not inform about its biochemical function and regulatory impact. Modern biology therefore exploits a large panel of different technologies to investigate many different aspects of a protein of interest.

1.1.4. Summary

The completion of the human genome has revealed tens of thousands putative proteins with unknown function. A central aspect of the post-genomic era is the rapid functional characterization of these gene products. The availability of indexed, non-redundant collections of cDNAs or knockout strains enables a great variety of genome-wide functional assays and the association of both positive and negative results to every interrogated sample.

High throughput experimentation holds great promise for the study of biological relationships. One of the earliest HT technologies, DNA microarray technology, has enabled the rapid formulation of hypothesis about the function of many individual gene products and proteins and thereby guided the formulation of numerous hypotheses. Furthermore, the ability to observe global transcriptional changes has deepened our understanding of the relation between the molecular composition of a cell and morphological phenotypes. Moreover, the speed and small scale of microarray experiments has provided a tool for the analysis of system wide transcriptional dynamics and the underlying regulatory mechanisms.

The impressive progress from transcriptional profiling and the simultaneous notion of the limitations of any single HT technology has spurred the desire to expand the repertoire of HT technologies. Because proteins are the acting agents in cells, the focus of my interest has turned to proteomics.

1.2. Proteomics

Proteins convey information inside and outside of cells, metabolize nutrients, give structure to cells and are a major component of the machinery that uses genomic information to synthesize new proteins. Thus, proteins are of primary importance for the understanding of cellular behavior and diseases and the perspective to study protein abundance, dynamics and activities on a global scale promises phenomenal advances in the understanding of biological systems.

The term 'proteome' was first used in 1994 to describe all proteins expressed by a cell or tissue at a given time (Huber, 2003). 'Proteomics' evolved in recent years to encompass the characterization of protein inventories, analyzing global changes in protein levels, investigating the biochemical functions of all proteins, analyzing global patterns of post-translational modifications, delineation of protein-protein interaction networks and more. Within proteomics it is possible to distinguish analytical or quantitative proteomics and functional proteomics. The following section will give a brief overview over the different approaches. The second section will discuss functional proteomics in greater detail, introducing different experimental approaches and their role in understanding protein function as well as the challenges functional proteomics is facing. It will become clear that cDNA collections are essential to nearly all functional proteomics enterprises. In addition, it will become obvious that the entire field of *in vitro* biochemistry is inaccessible to high throughput experimentation.

1.2.1. Approaches in Proteomics

Proteins are extremely complex biomolecules and the analysis of proteomes is a multi-facetted area of research. In this wide area it is possible to distinguish 'analytical proteomics', which aims at analyzing proteomes and 'functional proteomics', which aims at unraveling the functional roles of proteins including the impact of posttranslational modifications their subcellular localizations, regulatory function etc. Naturally, both areas are closely intertwined and may not always be clearly separated.

Analytical proteomics is currently using mass spectrometry as its main technology. Specific questions of this area of proteome research ask about the protein

composition of cells, subcellular structures or body-fluids at a certain time or physiologic state and about the posttranslational modifications these proteins are carrying. In the context of disease marker and drug target identification, the difference in these parameters between pathologic and healthy cells or organisms is of interest. Analytical proteomics is using protein separation by 2-dimensional gel electrophoresis (2-DE) and 2-dimensional liquid chromatography (2D-LC) and a plethora of fractionation techniques to break down proteomes so that they are amenable for the identification of proteins and their modifications by mass spectrometry. Unfortunately, analytical proteomics is held-back by numerous fundamental challenges that prevent a global perspective onto proteomes. These challenges include the wide dynamic range of protein concentrations in cells, the biochemical heterogeneity of both proteins and their peptides and the dynamic nature of the proteome (Anderson and Anderson, 2002; Huber, 2003). Thus, a major focus of analytical proteomics is the development of technologies that enable the acquisition of more global pictures.

In one of the most comprehensive proteome analysis to date, Koller et al analyzed the proteomes of three different tissues of the rice *Oryza sativa*: leaf, stem and root. Using a combined 2D-LC and 2DE approach, the authors identified a total of 2,528 different proteins in the three tissues. Analyzing functional annotation of the identified genes, the authors find that proteins involved in energy flow and carotenoid synthesis are predominantly expressed in leaf, whereas proteins involved in starch degradation were mainly detected in roots (Koller et al., 2002). However, the functions of several hundred proteins were unknown and consequently uninformative. This example illustrates the use and limits of analytical proteomics. Analysis, by definition, provides insight in the protein composition, identity of posttranslational modifications and potentially of the dynamics of proteomes *in vivo*. Thus, it is a crucial discipline for discovery and verification of biological relationships. However, to unravel the task of the identified proteins or to investigate the impact of posttranslational modifications, manipulative and thus 'functional' approaches to proteome analysis are essential.

Functional proteomics uses *in vitro* and *in vivo* technologies as well as mass spectrometry to elucidate the functions of proteins. Aspects of interest include protein-protein

interaction networks (Gavin et al., 2002; Ho et al., 2002; Uetz et al., 2000; Walhout et al., 2000a), protein localization (Kumar et al., 2002; Simpson et al., 2000), protein structure (Burley, 2000), enzymatic activities (Zhu et al., 2001) as well as posttranslational modifications and their effect on protein interactions, localization and enzymatic activity. Several traditional technologies have been adapted to high throughput experimentation, including 2-hybrid analysis, subcellular localization mapping and protein crystallography. In addition, several novel technologies have been developed for example protein arrays. Common to all mentioned types of experiments is the dependence on large, ideally comprehensive collections of cDNAs.

1.2.2. Functional Proteomics

The elucidation of the function of a given gene product or protein requires the combination of many different technological approaches to investigate many parameters of the protein. Many of the technologies that have been employed in the past to investigate the function of individual proteins are currently adapted to HT experimentation to investigate the function of thousands of proteins in single experiments. The first part of this section will discuss different approaches that are currently being pursued to analyze the function of proteins in HT format. The second part will discuss challenges that are encountered in HT functional studies. One major problem is the lack of cDNA and protein reagents. The last section will introduce efforts to build comprehensive cDNA collections that are compatible with an infinite selection of experimental designs.

1.2.2.1. Approaches in Functional Proteomics

The function of a protein is characterized by many different parameters. Very important factors include the protein-protein interaction partners, the localization of the protein, its enzymatic activity as well as its regulation by posttranslational modifications and protein interactions. The most advanced HT enterprises are the analysis of protein interaction networks and analysis of subcellular localizations.

1.2.2.1.1. Protein Interaction Mapping

One of the most instructive data-types for learning about the function of unknown proteins is the identity of protein-interaction partners. Because all proteins execute their function in concert with other proteins, the identification of (previously characterized) protein interactors allows the formulation of hypotheses about the function of the protein under investigation. In order to collect information of protein interactions on a large scale several global protein interaction maps have been acquired for *S. cerevisiae*. The methods used were HT 2 hybrid analysis (Ito et al., 2001; Uetz et al., 2000) and HT immunoprecipitation of protein complexes followed by mass spectrometry (Gavin et al., 2002; Ho et al., 2002). Combined these analyses provide interaction data for nearly 75% of the proteins encoded in the yeast genome (Bader and Hogue, 2002). The data provide hypotheses for the function of hundreds of unknown proteins and enable the analysis of the global topology of protein interaction networks.

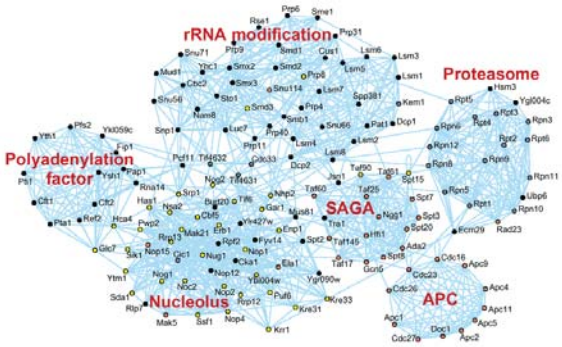

Figure 5 (from Bader and Hogue (2001)): Four integrated yeast data sets after addition. The complex connectivity surrounding the nucleolus is clearer and more complete in the fully integrated data set, indicating that data integration is necessary for better understanding of a biological system. APC, Anaphase-promoting complex; SAGA, Spt-Ada-Gcn5-acetyltransferase transcriptional activator–histone acetyltransferase complex; DDR, DNA damage response; TRAPP, transport protein particle complex; 19S regulatory subunit of the proteasome labeled "proteasome".

In analyzing the large scale patterns of protein interaction networks, one important discovery was the notion of functional modules; groups of proteins that act in concert to fulfill a cellular function and which are connected to each other by many more protein interactions than to proteins in other such modules. Among others, this finding has enabled the discovery of a nucleolar protein module, which contains proteins involved in several steps involved in RNA processing (Bader and Hogue, 2002). These data were supported by mass spectrometric data describing the composition of human nucleoli (Andersen et al., 2002).

A second discovery in the complete interaction network of *S. cerevisiae* was the notion that proteins can be distinguished based on the number of interactions they are undergoing. Proteins undergoing five or less interaction constitute the majority of yeast proteins (93%). In contrast only 0.7% of yeast proteins undergo 15 protein interactions or more. Interestingly, when deleted, 62% of genes coding for proteins in this second group have a lethal phenotype, whereas only 21% of genes coding for proteins in the first group are essential (Jeong et al., 2001). Thus, highly connected proteins have central functions that are less dispensable than proteins, which are more sparsely connected.

Protein interaction data keep the promises of HT experimentation by providing hypotheses about the function of hundreds of unknown proteins and enabling the discovery of higher order patterns and relationships that are not accessible by traditional experimentation.

1.2.2.1.2. Localization Mapping

Protein localization may reveal important insights about the function of a protein, especially when other types of data are available. In addition, the localization of a protein may help to shed light on the context dependence or simple possibility of a protein interaction. Currently, the most widely used methods to analyze the subcellular localization of a large number of proteins involve expression of a green fluorescent protein fusion construct and detection of the subcellular localization by confocal microscopy (Kumar et al., 2002; Simpson et al., 2000) or purification of subcellular structures and analysis of their composition by mass spectrometry.

The most comprehensive data set determining protein localization by expression of fusion proteins and microscopic analysis has been collected by Kumar et al. The authors analyze the localization of 2744 yeast protein and extrapolate the relative numbers to the complete yeast proteome. Thus, they authors conclude that about 47% of all yeast proteins are cytosolic 27% are nuclear, 13% mitochondrial and so on (Kumar et al., 2002). More importantly, the localization data of any gene product can be combined with other data types to gain a better insight into the functional relationships of proteins in the context of cellular environments (see below).

The proteomic analysis of subcellular structures has become an integral part of analytical proteomics, which simultaneously provides important data about the subcellular localization of proteins (Dreger, 2003; Jung et al., 2000)). Andersson et al analyze the composition of the human nucleolus. The authors purify human nucleoli; separate the proteins by 1D- and 2D-PAGE and analyze proteins by MALDI-MS or LC-tandem-MS, and identify proteins by searching peptide fingerprints and peptide sequences against protein databases and the translated human genome sequence, respectively. The authors identify 271 nucleolar proteins, 80 of which have not been described previously. In addition to recording a comprehensive gene catalogue, the authors confirm the nucleolar localization for a subset of novel proteins, identify a novel nucleolar compartment and demonstrate dynamic rebuilding of the nucleolus in response to Actinomycin D exposure of the cells (Andersen et al., 2002). The data from this study have been used to support the protein interaction network identified by Bader and Hogue (2002).

Protein localization is an important parameter that allows insight into the function of a gene product or protein. Often localization data can be combined with other types of data for example about protein interactions and biochemical activities to yield valuable insights

1.2.2.1.3. Protein Arrays: A Platform for HT Biochemistry

Biochemistry is a crucial discipline for the elucidation of the function of proteins. Recently, protein arrays were introduced as a platform for carrying out biochemical experiments in HT format. Protein arrays are manufactured by spotting or arraying recombinant proteins on various surfaces. On this surface the proteins are available for biochemical experiments like binding studies, enzymatic manipulations or others. In 1999, Macbeath and Schreiber demonstrated that functional protein arrays can be manufactured with standard arraying equipment. Furthermore the authors demonstrate that functional enzyme assays, protein, protein interaction assays and protein-compound interaction assays can be carried out with proteins that have been immobilized on glass slides (MacBeath and Schreiber, 2000). At about the same time, Zhu et al. analyzed 122

of the 134 kinases of *S. cerevisiae* for their activity towards a set of 20 substrates. One surprising outcome of this study was the large number of kinases that exhibit tyrosine kinase activity (Zhu et al., 2001). Later the same authors manufactured the first proteome scale protein arrays, which consisted of 5800 yeast proteins, spotted onto glass slides. The authors employ these protein arrays to analyze protein interaction partners of biotinylated calmodulin, which they detect by using a Cy3-labeled streptavidin. In this experiment the authors identify six known targets of calmodulin and 33 novel interactors. A sequence analysis of the identified proteins reveals a novel calmodulin binding motif (Figure 5B). In addition, the authors probe the protein arrays with micelles into which different physiological lipids have been incorporated. In this experiment the authors identify potential targets that selectively bind to either of the investigated lipids.

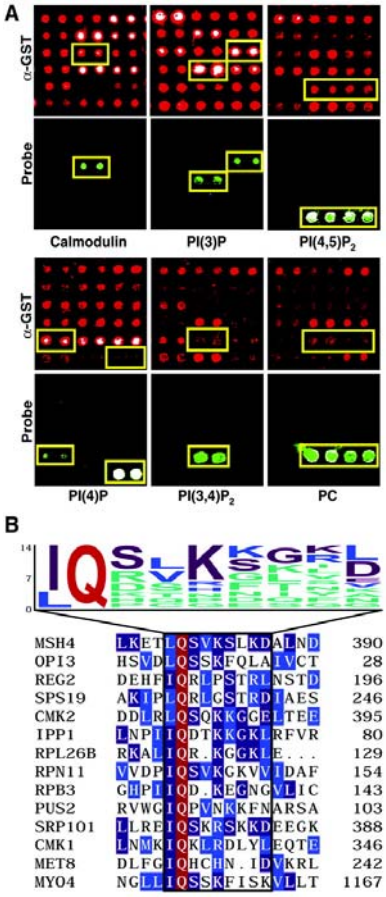

Figure 6 (from Zhu et al., 2002): (**A**) Examples of different assays on proteome chips. Proteome chips containing 6566 yeast proteins were spotted in duplicate and incubated with the biotinylated probes indicated. The positive signals in duplicate (green) are in the bottom row of each panel; the top row of each panel shows the same yeast protein preparations of a control proteome chip probed with anti-GST (red). The upper panel shows the amounts of GST fusion proteins as detected by the anti-GST (red). (**B**) A putative calmodulin-binding motif (32) is shown, which was identified by searching for amino acid sequences that are shared by the different calmodulin targets (10). Fourteen of 39 positive proteins share a motif whose consensus is (I/L)QXK(K/X)GB, where X is any residue and B is a basic residue. The size of the letter indicates the relative frequency of the amino acid indicated.

This example shows the tremendous potential of protein arrays for the biochemical investigation of proteins in high throughput format. However, one fundamental and unsolved aspect of protein array technology is the production of proteins that are required to manufacture protein arrays.

1.2.2.2. Challenges of Functional Proteomics

Functional proteomics is facing many technological and conceptual challenges including the incomplete knowledge and dynamic character of the proteome to unavailability of necessary reagents.

1.2.2.2.1. Protein Diversity

One fundamental problem of proteome research is its immense complexity which is several orders of magnitude greater than that of the genome. It has been argued that up to 10 – 20 million protein species are expressed by the human body (Huber, 2003). Among others, this complexity is generated by differential splicing and between 100 and 200 posttranslational modifications (Huber, 2003). Thus, taken the extreme stand that all or at least a significant fraction of these proteins ought to be investigated to achieve the objective of proteomics of analyzing all proteins in a cell, the goal of proteomics becomes unrealistic.

One response to this challenge is the use of genes as the common reference for proteins. This practical approach, which is neglecting the variety introduced by differential splicing and posttranslational modifications, is made by both analytical and functional proteomics. In analytical experiments, proteins are commonly identified by gene names. This has motivated Rappsilber & Mann to suggest the wording that a ‘gene product’ as opposed to a ‘protein’ has been identified (Rappsilber and Mann, 2002). In functional proteomics, because of the lacking knowledge of splice variants of a gene and the inability to isolate all such variants in a time and cost efficient manner, cloning projects are currently focusing on acquisition of one representative cDNA isolate per gene. It can be anticipated, though, that ongoing cDNA sequencing and improvements in both DNA microarray technology and analytical proteomics will provide a more detailed picture of the different splice variants, which will then be incorporated into cDNA collections. The identification of splice variants by proteomic profiling is less trivial, because the necessary near complete peptide-sequence coverage of proteins has not been achieved.

1.2.2.2.2. Dynamic Character of the Proteome

The proteome, in contrast to the genome, is a very dynamic entity, which is different in every cell of the body and is constantly remodeling itself. Thus it is not sufficient to identify the biochemical relationships between proteins (e.g. a physical interaction), but it is equally important to find the conditions, which are required for this relation to occur. This aspect also complicates the verification of *in vitro* studies, which bear the possibility of false data. In this verification process any not-verified relation may be false or just not occurring in the cell type or physiological state that was analyzed.

1.2.2.2.3. Bioinformatic Representation

Given the numerous proteomic approaches and the plethora of important biochemical parameters, one critical aspect of proteomics will be the well-designed integration of these data. Currently, all data about protein function are stored in relational data bases, which enable the retrieval of information about proteins (SGD, WormBase, FlyBase, LocusLink). While this approach serves well the requirements of the study of single proteins, it is falls short when functional correlations of proteomic scale are to be investigated. Given the impossibility of analyzing modification, localization, activity status and protein interactions of even a small number of proteins in any single experiment, it will be crucial to create a bioinformatic environment in which the respective data can be overlaid with each other and conveniently analyzed for global.

1.2.2.2.4. Reagent Requirements

Most functional assays depend on the availability of cDNAs, purified proteins or antibodies for the protein under investigation. Consequently, the HT analysis of protein function requires the availability of large, ideally comprehensive reagent collections. As the production of antibodies usually requires purified proteins for the immunization and screening process and the production of purified protein requires cDNAs for the protein of interest, it is clear that comprehensive cDNA collections are indispensable reagents for functional proteomics. However, many genes have only been uncovered by analysis of genomic sequence data and no physical copies for the corresponding cDNAs are

available. Efforts to create cDNA collections for functional proteomics and functional genomics experiments are described in the following section.

1.2.3. Reagents for Functional Proteomics

Most modern functional assays require a cDNA of the protein of interest at one stage of experimentation. Consequently, the HT analysis of protein function requires the availability of large, ideally comprehensive cDNA collections, which must fulfill certain criteria to be useful. The first part of this section will introduce these criteria. One central requirement Is the compatibility with many different experimental set ups and this can be achieved by using recombinational cloning technology, which will be introduced in the second part. The last part of this section will introduce the FLEXGene repository, which provided most cDNAs used in this study.

1.2.3.1. Criteria for Post-Genomic cDNA Collections

In order to capitalize on the knowledge of complete gene catalogues, post-genomic comprehensive and indexed cDNA collections have to be assembled. Ideally, every mRNA in the organism should be represented. For complex animals, like mammals, isolating every splice variant may be an enormous task. A more proximal goal is the collection of one representative for every gene. Furthermore, the cDNAs need to be indexed, i.e. the identity of every gene in every location must be known. Only then can information can be connected to every gene. Moreover, indexing will avoid wasting effort to determine the identity of positives post facto and to screen multiple copies of the same gene to ensure sampling of the rare genes. Lastly, the collections must be compatible with fusion tags. Many modern functional assays require the expression of fusion proteins, Gal4-AD-fusion in yeast, GFP in mammalian cell and a GST-fusion protein in bacteria. No tag can be added if the untranslated regions are still present and their removal is thus mandatory. In addition, all cDNAs should be arranged in the same reading frame.

Several groups have begun to assemble such cDNA collections for various applications (Martzen et al., 1999; Uetz et al., 2000). However, as a result of the great variety of biological assays, expression systems and expression constructs, technologies

are required that enable the reliable transfer of these collections from one vector into another.

1.2.3.2. Recombinational Cloning

Recombinational cloning is a cloning technique that exploits site-specific recombination to shuttle inserts from one vector into another while maintaining reading frame and orientation of the insert (Hartley et al., 2000). A typical subcloning procedure consists of mixing two vectors with recombination enzymes in a reaction tube followed by incubation and a subsequent bacterial transformation. The cloning strategy yields the desired product with high efficiency (95% - >99%) via a combination of selectable markers (antibiotics) and negative selection (e.g. ccdB gene or sucB gene) (Afif et al., 2001; Walhout et al., 2000b). Because the selection is very tight, recombinational cloning does not require the identification of successful recombinants. In addition, unlike traditional cloning with restriction enzymes, recombinational cloning is independent of the sequence of the insert that is to be cloned. As a result, this approach does not need to be individualized for each gene but instead provides a universal strategy that can be applied to all CDSs encoded in a genome. Once a complete high-quality collection has been constructed in such a system, virtually any cDNA dependent experiment can be scaled to genomic dimension.

The two vectors in the reaction are a **master**- or **donor**-vector and an **expression** vector. The **master** vector is an inert vector that is used to maintain and amplify the insert. It is made up of the insert, flanked by one or two recombination sites, and a few essential functionalities that enable selection and amplification of the plasmid in bacteria. In contrast, the **expression** vector carries all sequences that are required for vector maintenance in other systems (yeast, insect cells, mammalian cells, etc.) and for protein expression, such as promoters, regulatory sequences and coding sequences for fusion tags like GFP, His-tag, HA-tag, Gal4-AD etc. The expression vectors must also have recombination sequences that allow them to receive the inserts. Presently, two recombinational cloning systems are commercially available: The Gateway™ system, which is based on the site-specific integration of bacteriophage Lambda into the genome

of *Escherichia coli* and the Creator™ system, which is based on the Cre/lox site-specific recombination by which phage P1 integrates into the *E.coli* genome. The reactions of the Gateway system, which has been used in this study, are shown in the figure.

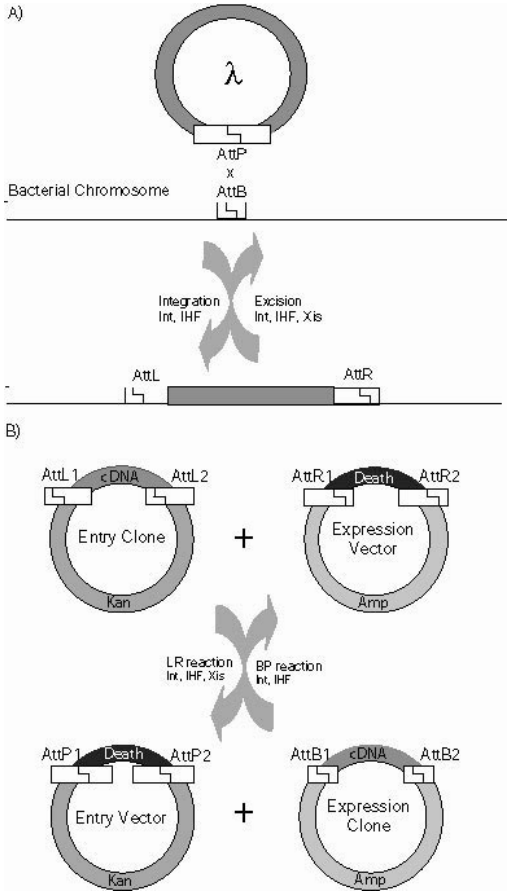

Figure 7: Gateway System. A) Phage Lambda integration into the *E. coli* genome is the basis for the Gateway system. In order integrate into the *E. coli* genome, Lambda uses Integrase (Int), which is encoded by the phage, and the Integration host-factor (IHF), which is provided by *E. coli*. For the excision, excisionase (Xis) is required in addition to the other two enzymes.
B) Schematic of transfer reactions in the Gateway system. The cDNA is maintained and propagated as an Entry clone. In the transfer reaction (LR), this plasmid is mixed with the expression vector, which carries all functionalities (promoter etc.) for the downstream experiments. By addition of the LR-enzyme mix, the AttL and AttP sites form an intermediate (not shown), which is resolved to yield the Entry-vector and the expression clone, which carries the gene of interest in the desired expression vector backbone. All undesired plasmids are eliminated via a combination of positive (Amp, Kan) and negative selection (death gene).

1.2.3.3. The FLEXGene Repository and Related Resources

Many groups have assembled comprehensive cDNA collections, mainly for yeast, in traditional cloning systems (Martzen et al., 1999; Uetz et al., 2000). Such collections, however are specially tailored to one experimental set up (e.g. 2-hybrid assay), and are consequently incompatible with the majority of functional assays. Using traditional cloning systems every assay requires a new cDNA collection to be built. Because this approach is time-consuming and expensive for small organisms and impossible for mammalian genomes, efforts are underway for the most popular model systems to

construct a representative cDNA collection in a recombinational cloning system. For *Homo sapiens*, an international effort lead by the Institute of Proteomics is underway to build such a collection, called FLEXGene repository (Full-Length EXpression). This repository will contain sequence-verified physical copies of full-length coding sequences (CDSs) in a recombinational cloning system (Brizuela et al., 2001). Similar efforts are underway for other model organisms, including *Saccharomyces cerevisiae* (Marsischky, HIP), *Caenorhabditis elegans* (Vidal, DFCI), *Arabidopsis thaliana* (CERN), *Pseudomonas aeruginosa* (Brizuela, HIP) and *Plasmodium falciparum* (Sullivan, NIH).

1.2.4. Summary

Proteomics aims at analyzing the protein composition and dynamic of cells and tissues and to learn the functional roles of all proteins. Functional proteomics is using both, traditional technologies that have been adapted to high throughput experimentation as well as novel proteomic technologies to investigate the function of gene products. Most, if not all, functional proteomics experiments depend on the availability of recombinant cDNA collections at one stage of experimentation. The Institute of Proteomics and other institutions are building cDNA collections in recombinational cloning systems that are an indispensable resource for functional proteomics. A subset of experiments, like 2-hybrid protein interaction mapping and determination of the subcellular localization of recombinant proteins, can be done once the respective cDNAs are available. However, an important subset of proteomic experiments requires the availability of purified proteins.

1.3. High Throughput Protein Purification

Purified proteins are a key reagent for numerous biochemical assays addressing important questions about enzyme/substrate relationships or investigating the effects of posttranslational modifications on activity and interactions of modified proteins. Potential application of high throughput protein purification methods will be discussed in the first section of this chapter. The second section will discuss problems associated with the parallel purification of hundreds and thousands of proteins. The last section will introduce potential protein expression systems and discuss their respective advantages and disadvantages.

1.3.1. Applications for HT Protein Purification Methods

In proteomic biochemistry recombinant proteins are required for many different applications. Within the large spectrum of experiments two different types of applications can be distinguished (Figure 8). On the other side of the spectrum one can find high throughput biochemical experiments like studies of enzymatic function and substrate specificity of modifying enzymes. Such biochemical studies can be done on protein arrays or similar high-throughput platform and usually require only microgram amounts of protein. Thus, for these applications microscale protein purifications may provide sufficient protein for a few hundred assays. Other applications, like structural studies, require milligram amounts of protein, and for these application, parallel

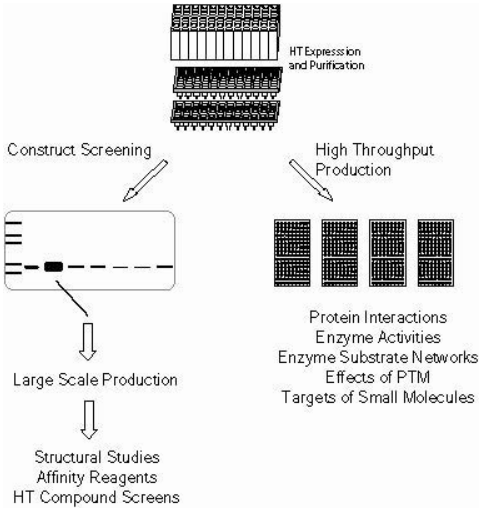

Figure 8: Application for microscale HT protein purification methods. One application is the production of many different proteins for HT biochemical experiments (right branch). In a different set of application, HT protein purification may be used a screening to tool, to identify well expressing soluble constructs for which production can be scaled up immediately.

protein expression and purification provides an important platform for the identification of well expressing constructs and expression conditions, for which production can be scaled up. The application types are displayed in the figure and will be discussed in more detail in the following sections.

1.3.1.1. High Throughput Biochemistry

1.3.1.1.1. Enzymatic Activities

The study of enzymatic function is as old as biochemistry itself, dating back nearly a century. Enzymes catalyze a wide variety of physiological reactions including metabolism, posttranslational modification of proteins, regulatory modification of lipids and nucleic acids, detoxification and others. Analysis of *in silico* translated gene catalogues suggest that approximately 40% of all human genes code for proteins with enzymatic activity (Waterston et al., 2002) and the task of proteomics is the determination of the catalyzed reactions, their regulation and physiological significance in cell and body. In order to fulfill this task, proteomics depends on methods to express and purify hundred and thousands of proteins for both global and focused approaches.

One method to identify biochemical activities is the expression of all cDNAs of an organism in an expression system of choice, and screen individual clones or pools for the sought after enzymatic activity. Traditionally, this expression cloning approach has been done with cDNA libraries in bacteria. However, with the availability of comprehensive cDNA collections, it became possible to screen more efficiently and systematically. Martzen et al have expressed all proteins encoded in 6144 yeast genes in the same organism, pooled 64 x 94 strains and purified proteins in 64 pools. These pools were screened for various biochemical activities and the authors identify three novel activities, a cyclic diesterase, which acts on adenosine diphosphate ribose 1”-2” cyclic phosphate (Appr>p), an Appr-1”-p-processing activity and a cytochrome c methyltransferase (Martzen et al., 1999).

A potential pharmaceutical application of HT biochemistry is the identification of enzymes or enzyme isoforms that are lowering the efficiency of an otherwise functional drug. If such enzymes can be identified, it may be possible to develop combinatorial therapies, which consist of the actual drug and a second drug that prevents its untimely

metabolism. In addition, enzymes play increasing roles in technical applications ranging from laundry detergents to environmental protection. However, a limiting factor for the identification of novel enzymes is the requirement for an assay that requires a prior knowledge about the catalyzed reaction.

1.3.1.1.2. Enzyme/Substrate Relationships

Proteins can be posttranslationally modified by approximately 100 - 200 different modifications and many of these modifications have a regulatory function that affects numerous aspects of protein function. Countless examples illustrate the effect of phosphorylations, arguably the most important regulatory PTM, on enzyme activities (Solomon and Kaldis, 1998), protein-protein interactions (Shen et al., 1998), protein half-life etc. (Blanco et al., 2000; Tintignac et al., 2000). Likewise, acetylation, proteolytic cleavage, methylation, ubiquitination to name just a few, may affect the subcellular localization, enzymatic activity and half-life of the modified proteins. According to a recent analysis, the human genome encodes approximately 1000 different kinases (Manning et al., 2002). Furthermore, a recent analysis of the functions of the proteins encoded in the human genome concludes that approximately 40% of all genes code for proteins with enzymatic activities and many of these catalyze the posttranslational modification of proteins. Thus, beside protein-protein interaction networks, enzyme-substrate networks are crucial layers of biological regulation, whose understanding is quintessential for an understanding of biological regulation and biological systems.

Before completion of the human genome, biomedical research relied on pre-genomic screen to identify potential substrates of kinases (Zhao et al., 1998), or on a hypothesis based on prior knowledge (Akiyama et al., 1992; Knudsen and Wang, 1996). In order to transform the pre-genomic screens in postgenomic mapping experiments, high-throughput protein expression and purification methods are required.

1.3.1.1.3. Compound Binding

Small molecule compounds are important signaling molecules and the major class of pharmaceutical drugs. In addition, a new field of chemical genomics is aiming at making significant impact on biomedical research. In this new line of research, the physiological

processes are disrupted by application of small compounds. The resulting phenotypes can be analyzed and classified using genetic methods like epistasis experiments. Analogous to the genetic method of identifying the gene carrying the mutation responsible for an observed phenotype, in chemical genetics it is important to identify the mode of action of the small molecules, i.e. identify their target proteins (Alaoui-Ismaili et al., 2002; Zheng and Chan, 2002). For this identification, large numbers of proteins would be of great use.

Small compounds can act through many mechanisms, including the prevention or nucleation of protein-protein interaction or the inhibition of enzyme activities. Independent of the particular mechanism of action, however, the compound has to physically interaction with the target. Large, ideally complete, sets of proteins have the potential of greatly accelerating this target identification process.

1.3.1.1.4. Protein-Protein Interactions and Effects of PTMs and Compounds

Many posttranslational modifications have a regulatory impact on the modified protein and most small molecules are interesting to biologists because they interact and interfere with proteins or other molecules in the cell. While protein-interaction mapping by 2-hybrid analysis or mass spectrometry are very powerful tools to delineate protein interaction networks, these assays do not provide the opportunity of manipulating either of the involved interaction partners to investigate the consequences of a PTM or add a small molecule compound to the interaction assay. Such experimental manipulations can be best controlled in an in-vitro setting. Thus, in order to elucidate the underlying mechanisms of proteome plasticity, *in vitro* methods are of preeminent significance. Once large numbers of proteins are available, the effect of posttranslational modification on protein interactions can be investigated. For example an experiment similar to the calmodulin experiment on protein arrays described above could be conducted using protein arrays, which have previously been phosphorylated by a kinase of interest or all kinases encoded in the *S. cerevisiae* genome. It is possible, that this approach may identify proteins, which depending on the phophorylation status bind or release calmodulin.

1.3.1.2. Screening Tool for Large Scale Production

High throughput protein expression can be used as a preliminary step to screen for optimal expression conditions or gene constructs before scaling up for high yield protein expression. Experimental approaches such as protein crystallization and the production of protein affinity reagents often require milligram quantities of protein. Obtaining protein for these studies can be challenging because many proteins express poorly or fold improperly when produced in heterologous systems such as *E. coli* (Chance et al., 2002). Historically, success has often depended on a tedious trial and error process of trying to express different versions of the target protein, such as orthologues or modified coding sequences with trimmed amino- and carboxyl-termini until a well-behaved construct can be identified. Recently, Savchenko et al. formally demonstrated the merit of this approach. Using a set of 62 pairs of orthologues proteins from *Thermotoga maritima* and *Escherichia coli* the authors demonstrate that the inclusion of a single orthologue of every target increases the over all success rate for obtaining soluble proteins by 50% (Savchenko et al., 2003). The ability to screen many constructs simultaneously in a multi-well format could speed up this screening process considerably and would facilitate the testing of more constructs.

1.3.2.2.1. Structural Proteomics

Knowledge of the three dimensional structure of a protein can greatly assist the understanding its biochemical function, its regulation by posttranslational modifications and the effects of small molecule compounds. In addition, uncharacterized proteins may have structural similarity with other proteins of known function and thus a hypothesis about their function can be deduced. Knowledge of structures can also assist the design of drugs by rational design or combinatorial chemistry. The ‘pragmatic’ goal of structural proteomics is the solution of all structures, so that the structures of all other proteins can be modeled onto these. Less than 42% of the yeast genes can be modeled onto structures in the Protein Data Base (PDB). Early estimates indicate that approximately 16.000 models to cover ~90% of the proteins space (Vitkup et al., 2001).

Two main methodologies that are currently used to investigate protein structures in HT format are nuclear magnetic resonance (NMR) and x-ray crystallography.

Common to both methods is a requirement for substantial amounts of proteins. Producing these quantities for thousands of different proteins constitutes one fundamental problem in structural proteomics. The problems of protein production in different expression systems will be discussed below, however, a high throughput method of screening different expression constructs and expression conditions efficiently, may enable a rapid and efficient identification of well expressing, soluble constructs.

1.3.2. Challenges of Parallel Protein Purification

Many biochemical assays require the availability of purified enzymes and in the modern era this is nearly always recombinant protein. The section will discuss the difficulties of parallel protein expression and purification, which include the biochemical heterogeneity of proteins to the necessity of a universal purification strategy.

1.3.2.1. Biochemical Heterogeneity

In contrast to nucleic acids, proteins are a chemically very heterogeneous class of molecules. In the study of single proteins the respective buffer conditions are usually adjusted for every new protein. The purification of large protein populations with any single approach requires the identification of conditions which work for most proteins. However, this implies that any set of conditions will fail to work for proteins, for which these conditions are inadequate because they require different salt, pH or detergent conditions. Traditionally, protein purification conditions are by and large different for every protein. The problems can be alleviated by the employment of recombinant protein expression technology and protein affinity tags; however, a different set of complications is introduced by this approach.

1.3.2.2. Purification Strategy

Protein affinity tags are proteins or peptides, which are genetically fused to the protein of interest and which can be captured by a high affinity-ligand, which itself is immobilized on a solid matrix. This technology becomes compatible with high-throughput processes by the construction of cDNA collections in recombinational cloning systems (see above).

1.3.2.2.1. Advantages of Protein Purification Tags

Any protein which has been genetically engineered to express an affinity fusion tag can be captured by the respective complementary affinity partner. Thus, affinity tags provide a universal purification strategy, which allows, theoretically, the parallel purification of protein in HT format.

Secondly, because the employed affinity tags are usually high affinity binders, the captured protein can be washed extensively without a significant loss of total protein yield. Thus frequently, single step purifications provide proteins of sufficient purity for many downstream applications. One particular advantage of affinity tags is an increase in protein solubility. In fact, this aspect was one of the first motivations to genetically engineer fusion proteins in 1986 (Smith and Johnson, 1988). Lastly, especially in bacteria it is often difficult or impossible to express small proteins and peptides. The fusion of the respective peptide to a protein fusion tag may allow the peptide to be expressed, because the total size of the protein has grown to a size that is compatible with protein expression.

1.3.2.2.2. Potential Problems with Purification Tags

One of the most significant problems of affinity tags is the potential of sterical interference of the tags with function or folding of the target protein. Halliwell et al have demonstrated that the introduction of the small His_6-tag at the C-terminus of L-lactate dehydrogenase induces conformational changes that lead to loss of enzymatic activity. The authors show that homogeneous active populations of the enzyme are obtained by fusing the tag to the other end of the protein (Rumlova et al., 2001). In HT application, this type of tailoring is obviously not possible, and thus there is likely to be a fraction of recombinant proteins that is sterically inhibited by the affinity tag.

In one of the most popular protein expression systems, *Escherichia coli*, it has been observed that large proteins (>100kDa) are more difficult to express than medium sized proteins. The addition of a fusion tag, which ranges in size from 6 amino acids to 58 kDa proteins, may thus have a negative effect on the yield or stability of recombinant proteins.

Lastly, there is significant variation between affinity tags and between recombinant proteins and not all combinations work equally well. Thus, while several

proteins can be easily and well purified with one particular tag, another set of proteins require a different tag for optimal purification. An important goal of high throughput protein purification must therefore be the identification of protein purification tags that enable a maximal fraction of protein to be purified.

1.3.2.3. Quality Control

The goal of high throughput protein production is the investigation of the function of the produced proteins. Thus, in order for these assays to be meaningful it is important that the respective proteins are properly folded and functional. While this is true for both traditional as well as high throughput protein production, in traditional experimentation researchers were usually familiar with the biochemistry of the produced proteins and could devise assays to control for proper function. In contrast, in high throughput protein production, the diversity of the produced proteins makes it difficult to design any single assay (or even a handful of assays) that could control for the functional integrity of all proteins. The most unsatisfactory way to deal with this problem is evoking the notion that high throughput assays always have an inherent failure rate and that the true relevance of HT assays is the identification of novel candidates that may be involved in a process.

A sounder, but still suboptimal, approach to the problem of quality control (QC) is the evaluation of a random sample of proteins. This approach will fulfill a minimal function of quality control in detecting fundamental flaws in the protein purification scheme. However, even this approach will not be able to detect problems of protein folding, which may be important only for a subset of proteins. One approach is the use of certain expression systems only for the study of protein families that have been shown to be compatible with a particular expression system, purification tag and/or condition.

An approach of potentially general applicability has been devised by Lesley et al. (2002) and was initially developed to screen for soluble proteins in *Escherichia Coli*. The authors placed β-gal under control of a promoter, which is activated by misfolded proteins. This construct was used to measure the cellular response to protein expression. To increase robustness, the assay was complemented by an ELISA-like method, which determines the amount of soluble recombinant protein in cell lysate. In a test-set of 186 proteins from *Thermotoga maritima* the combined assays identified almost 90% of the 62

soluble proteins correctly (Lesley et al., 2002a). It needs to be shown, though, whether this assay can predict functional integrity of the expressed proteins.

The choice of the expression system is likely to have an immense impact on the functionality of the recombinant proteins.

1.3.3. Expression Systems

Over the past several decades, a small number of protein expression systems have been employed for a variety of applications and novel systems have been explored in recent years. Obviously, each system has its strengths and weaknesses ranging from protein quantity and quality in terms of posttranslational modifications and folding, to cost, speed and ease of use.

1.3.3.1. Escherichia coli

E. coli is the simplest and by far the most widely used protein expression system. The most appreciated advantages of this system are its speed, ease of use and low cost, although difficulties can arise from low protein solubility or from the lack of posttranslational modifications[3]. To date, the largest data sets using E. coli as an expression system have been generated by structural genomic enterprises. These have been taking a low-hanging-fruit approach by purifying prokaryotic and archaeal proteins using the small His_6-tag. Christendat et al aimed to explore the feasibility of HT structure determination and analyzed 424 non-membrane proteins of *Methanobacterium thermoautotrophicum*, of which they found ~200 (47%) to be soluble and 175 (41%) of these could be purified using the His_6-tag (Christendat et al., 2000). An important question of recombinant protein production regards the functional integrity of the produced proteins. Based on NMR spectra quality the authors estimate that 57/100 analyzed soluble proteins may be in a state of aggregation and thus potentially non functional. Yee et al selected 513 non-membrane proteins smaller than 23 kDa from 5 different organisms and expressed and purified these using the His_6-tag. Subsequently the

[3] In some cases the absence of posttranslational modifications may be an advantage, e.g. if PTMs are to be introduced in a controlled manner or if very homogeneous protein populations are desired.

protein structures were determined by NMR analysis. Of all proteins the authors find that 68% are soluble (Yee et al., 2002). Importantly, however, the percentage of soluble protein varies significantly for different organisms, ranging from 95% for 21 proteins of *Thermotoga maritima*, over 63% of *Saccharomyces cerevisiae* (total ~90 proteins), 61% of *E.coli* (total ~130 proteins), 51% of *T. thermoautotrophicum* (total ~250 proteins) to 46% of the 25 Myxoma virus proteins. As in the previous study, the NMR spectra of a large fraction of soluble proteins indicated some state of aggregation or conformational instability of the protein (between 27% and 55%) (Yee et al., 2002). In a third study, which targeted the complete proteome of *Thermotoga maritima* 542/1.376 proteins were found soluble and could be purified (40%). The discrepancy to the previous study (95% vs. 40%) is probably caused by the low number of proteins analyzed in the first study and the fact that the latter study included transmembrane proteins and proteins of all sizes. The authors of all three studies report that smaller proteins are more likely to be soluble than larger proteins. Together these three studies provide good evidence that approximately 50% of prokaryotic and archaeal proteins can be produced in soluble form in *E.coli* and can subsequently be purified using the His_6-tag. For the remaining proteins, alternative expression conditions or systems will have to be identified.

In the past year, several reports have addressed the ability of different protein fusion tags to produce soluble protein for diverse eukaryotic proteins. Hammerstroem et al investigated the effect of seven fusion tags on the solubility of 27 small human proteins (6kDa – 19kDa). Based on their data the authors suggest that Thioredoxin (Trx), Maltose Binding Protein (MBP) and the Gb1-domain of protein G from *Streptococcus* are the best fusion tags with regard to improving protein solubility. These tags produced 20 (74%), 19 (70%) and 18 (67%) soluble proteins, respectively. In contrast, only 8 His_6-taged proteins, 13 GST-tagged, 14 NusA-tagged and 15 proteins tagged with the ZZ domain of Protein A from *Staphylococcus aureus* were soluble. Not unexpectedly, the best fusion tag varies for different proteins and several tags are required to obtain soluble protein for a maximum number of soluble proteins (Hammarstrom et al., 2002). In a similar study, Shih et al. expressed 40 different proteins ranging from 9kDa to 150 kDa from 5 different organisms with eight different protein fusion tags. In their experiments, 60% of NusA tagged proteins, 60% of MBP-tagged proteins and 38% of GST-tagged protein are high-

expressing and soluble (Shih et al., 2002). The other analyzed tags, for which no numbers were presented, are: Trx, Calmodulin-Binding-Peptide (CBP), Intein, and cellulose associated protein. The authors find that 80% of the tested proteins are soluble with at least one protein tag. In addition the authors find that 80% of the tested proteins are soluble with at least one of eight different fusion tags (Shih et al., 2002).

TAG	SIZE	LIGAND	AFFINITY	ELUANT
His6	**1kDa/ 6aa**	**Ni^{2+}/ Co^{2+}**	**$K_D=10^{-13}$ M**	**Imidazole/ (EDTA)**
Streptag	1kD/ 8aa	Avidin/ Streptavidin		(Biotin)
Avi-Tag	2kDa/14 aa	Avidin/ Streptavidin	$K_D=10^{-15}$ M	(Biotin)
CBP	**4kDa/ 26aa**	**Calmodulin**	**$K_D=10^{-9}$ M**	**EGTA**
ProtA Z-domains	13kDa	IgG - Fc		pH/Fc-peptide
Thioredoxin	13kDa	-	n.a.	
CBindD	17kDa	Cellulose		Ethylene glycol
GST	**29kDa**	**Glutathione**	**$K_D=10^{-4}$ M**	**Glutathione**
ProtA	42kDa	IgG		pH/Protein A/G
MBP	**42kDa**	**Maltose**	**$K_D=10^{-6}$ M**	**Maltose**
NusA	55kDa	-	n.a.	
Intein-CBD	55kDa	Chitin	n.d.	self cleavage

Table 1: Properties of Selected Protein Tags.

For several proteins the formation of insoluble protein aggregates has been shown to be dependent on the growth temperature of the bacteria and a lower expression temperature often favors the production of soluble protein. During optimization of the expression conditions for individual proteins, many researchers test a higher temperature first and decrease the temperature when this approach is unsuccessful. Hammerstroem et al. did their experiments at 37°C. The best expression condition for large protein sets may vary for fusion-tags and bacterial strains.

Genetically engineered bacterial expression strains are routinely used by structural protein production groups. *E. Coli* strains are available that alternatively supply tRNAs against codons rare in bacteria, exhibit very tight gene regulation to prevent expression of toxic recombinant genes or have a different intracellular redox potential and thus permit disulfide bond formation (Kane, 1995; Wall and Pluckthun, 1995).

1.3.3.1.2. Screening for Soluble Targets

Several applications require milligram quantities of proteins. However, only a small fraction of expression constructs is producing sufficient quantities in soluble form and often minor modifications in the primary structure or in the fusion tag may transform an insoluble protein into a soluble one. In this area, technologies that enable the rapid indentification of soluble constructs are needed.

In recent years, several technologies have been developed to identify such well expressing soluble constructs. Some assays use lysis in 96-well format, sometimes followed by microscale protein purifications and quantification of soluble protein by dot-blot (Doyle et al., 2002), ELISA (Lesley et al., 2002a), mass spectrometry (Chance et al., 2002).) or SDS-PAGE (Hammarstrom et al., 2002). A promising *in vivo* assay has been developed by Lesley et al, who placed β-gal under control of a promoter, which is activated by misfolded proteins. This construct was used to measure the cellular response to protein expression. To increase robustness, the assay was complemented by an ELISA-like method, which determines the amount of soluble recombinant protein in cell lysate. In a test-set of 186 proteins from *Thermotoga maritima* the combined assays identified almost 90% of the 62 soluble proteins correctly (Lesley et al., 2002a). A different *in vivo* screen based on structural complementation was developed by Wigley et al. (Wigley et al., 2001). In their application, the target protein is fused to one domain of β-gal, whereas the second domain of the reporter protein is expressed independently. When the target protein is soluble β-gal is reconstituted and activated.

1.3.3.2. Cell-Free Expression

Cell-free expression systems (cell lysates) have become more popular in recent years as the protein yields obtained with these systems have reached preparative levels. For HT

applications cell-free expression systems are particularly attractive, because the lack of a cell membrane eliminates process steps associated with the introduction of DNA into cells, cell lysis and lysate clearing. These properties have inspired technologies for structural studies, that does not require protein purifications (Guignard et al., 2002; Kigawa and Yokoyama, 2002) and novel concepts for protein arrays, for protein arrays, in which proteins are simultaneously produced and captured on a slide without separate protein purifications (He and Taussig, 2001; Tabuchi et al., 2002).

Historically, low protein yields obtained from cell lysates, high cost associated with these expression systems and the dependence on traditional, highly redundant cDNA libraries have limited their use to protein-protein interaction studies and few expression cloning experiments. However, the latter already indicated the tremendous potential of cell free expression, when combined with genomic cDNA collections. This was widely recognized after Phyzicky and colleagues employed a similar cloning strategy using the complete set of cDNAs of *S. cerevisiae*. In addition, several developments in recent years have significantly improved the yields so that yields up to 6mg/ml have been reported for individual proteins (Kigawa et al., 1999). Important changes were the concentration of the lysate (Jermutus et al., 1998), the introduction of semi-continuous and continuous reaction by dialysis (Hino et al., 2002), the addition of energy regeneration systems (Kim and Swartz, 1999) and in case of wheat-germ lysate, the removal of a suicide system on the outside of the seeds (Madin et al., 2000). The most widely used open expression systems are bacterial, wheat germ and reticulocyte lysates, although lysates from other cell types have been made and used. All major cell lysates are commercially available as optimized kits. Unfortunately, these are quite expensive and of proprietary composition with respect to salt and buffer concentrations, which can be a disadvantage if these conditions need to be adjusted.

The Riken Structural Genomics Initiative in Japan produces its target proteins almost exclusively in cell-free expression systems. In an early review article, Yokoyama et al. indicated that about a quarter of randomly chosen mouse cDNA clones could be produced with yields of 0.1 mg/ml or higher (Yokoyama et al., 2000). As these reactions were done in batch format, which is inferior to semi-continuous and continuous reactions, this number is likely to have increased since.

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 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 kinases and recording an NMR spectrum of one further protein (Sawasaki et al., 2002).

1.3.3.3. Other Systems

Despite its undeniable advantages and widespread use, *E. coli* suffers from significant disadvantages when it comes to correctly modified and active protein. Mammalian cells, which are capable of executing all posttranslational modifications important for human proteins, require very delicate growth conditions and are at best suited for very specialized applications. Thus, interest in *Saccharomyces cerevisiae* as protein expression systems has increased recently beyond the 2-hybrid assay as this organism combines the advantages of an inexpensive, fast growing unicellular organism with the physiological properties of a eukaryotic cell. In an impressive and pioneering work by Zhu et al, *S. cerevisiae* has been used to express 5800 proteins from the same organism, which were used to build protein arrays of the same size (Zhu et al., 2001). These protein arrays were used to characterize novel interaction partners of calmodulin, identify lipid binding proteins as well as nucleic acid binding proteins. Based on western-blot analysis of 60 samples using an antibody against an N-terminal GST, the authors estimate that approximately 80% of all proteins produce detectable amounts of proteins of the correct size. This number was confirmed by immunodetection of all 5800 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. Staring from a cDNA library in yeast expression vectors Holz et al. described a process to express and purify proteins in *S. cerevisiae* in 96-well format. Using a set of seven human proteins (~20kDa – 40kD), the authors demonstrate all seven proteins can

be expressed and purified and that yields up to 10μg/ml can be obtained (Holz et al., 2002).

Protein production in insect cells is usually done in context of the baculovirus system and this is a very popular protein expression system for large scale protein expression. Appreciated advantages of insect cells are the robust and relatively inexpensive cell culture and the fact that most eukaryotic posttranslational modifications are executed properly (Possee et al., 1999). Albala et al. have described a HT compatible strategy to express proteins in insect cells. In initial tests, 34 out of 81 wells produced soluble protein (Albala et al., 2000). For HT operations, however, the need to make the viruses, which is a multistep process, and the need to maintain high virus titers are significant disadvantages of the baculovirus system.

1.3.4. Summary

Purified proteins are an indispensable reagent for biochemical studies, and such studies are elementary for the elucidation of several central aspects of protein function like enzyme activity, effect of posttranslational modification and HT structural studies. Unfortunately, HT protein production faces several severe hurdles including the biochemical diversity of proteins and the choice of the best purification strategy. Despite its known disadvantages, *Escherichia coli* is one of the most popular protein expression systems. It would be of great value to have HT methods for the parallel expression and purification of large numbers of human proteins in *E. coli*.