Chapter 4

Discussion

Pathway/phenotype-associated genes are frequently expressed in the same or nearby places and at identical or similar time points. This novel data mining approach for extraction of phenotype- or pathway-associated genes from gene expression databases is based on that observation. CGAP Expression Data contains information of UniGene clusters in different cDNA libraries. Most of these cDNA libraries are individually derived from defined tissues (e. g. from biopsies). In accordance with the rationale mentioned above, these libraries represent gene sets that are expressed at the same place (in the same sample) at the same time (time of biopsy). If a phenotype or pathway of interest has been active at the time the sample was taken, a multitude of genes associated with the pathway or phenotype should have been expressed in the sample and thus be present in the library. On the other hand, such samples and derived cDNA libraries not only represent the expressed genes or pathway of interest, but in fact of many pathways. The challenge is the enrichment and extraction of pathway- and phenotype-associated genes from the multitude of genes that are additionally expressed in these samples. This is addressed by the Common Denominator Procedure (CDP).

The Common Denominator Procedures relies on the definition of as many cDNA libraries that are as diverse as possible with the desired common phenotype. To demonstrate the feasibility of the procedure, three different variations of the CDP have been applied to the phenotype angiogenesis. The three variations differ in the way they identify those phenotype-specific (angiogenesis) cDNA libraries. The basic Common Denominator Procedure (bCDP) and the genetic algorithm based Common Denominator Procedure (gaCDP) achieve this task by detecting presence or absence of a set of phenotype-specific

genes (AngioTestGroup) that were selected in a semiautomatic manner, based on publicly available gene annotations. The indicator genes based Common Denominator Procedures (igCDP) detects the presence or absence of a small set of manually selected well defined phenotype-specific indicator genes (IndicatorGeneSet).

All phenotype-specific libraries identified by the three approaches will individually carry angiogenesis-genes and many other genes. For example, libraries may in addition harbor proliferation genes (e. g. tumor cells), differentiation genes (normal tissue), disease-specific genes (disease tissue) and housekeeping genes (all tissues). However, a common denominator is the presence of genes that are associated with angiogenesis. Provided a sufficient number of libraries with the common phenotype can be defined, extraction of these angiogenesis genes from the whole set of genes should be possible because these genes are the common denominator. The more libraries are used as input and the more diverse these libraries are (e. g. from different tissues, different diseases and different stages/grades) the smaller and hence more specific should be the common denominator.

4.1 Comparison of the Procedures

The LibraryProfiles of the basic and the genetic algorithm based Common Denominator Procedures differ only in one out of the eight identified cDNA libraries. Therefore, it is not surprising that 20 of the 22 candidates from the gaCDP were also contained in the candidates from the bCDP (GeneScore = 100). Due to the more sophisticated candidate gene selection of the gaCDP its final candidate gene list was pruned of genes that had a high probability of reaching their GeneScore by chance, and was therefore smaller. Due to the one different library two novel genes were contained in the candidate gene list that were not contained in the candidate gene list of the bCDP. A more prominent difference is observable between the indicator gene based Common Denominator Procedures and the former two. Here, only a fraction of the original candidate genes from the bCDP (9 of 55) or the gaCDP (7 of 55) CDP qualified for the stringency setting of occurring in at least eight AngioProfiles (see Table 4.1).

To compare the performance of the different Common Denominator Procedures, the total number of verified angiogenesis-associated genes within the candidate genes was utilized. Candidate genes from the AngioTestGroup or the IndicatorGeneSet were

	bCDP	gaCDP	igCDP
bCDP	42	20	9
gaCDP	-	22	7
igCDP	-	-	55

Table 4.1: Comparison of the Common Denominator Procedures. Intersection of verified angiogenesis-associated genes within the candidate genes of the different versions of the Common Denominator Procedures. To this end the 42 candidate genes of the basic Common Denominator Procedures (bCDP) with a GeneScore of 100, the 22 candidate genes of the genetic algorithm based Common Denominator Procedures (gaCDP) with a GeneScore of 100 and the 55 candidate genes of the indicator gene based Common Denominator Procedures (igCDP) occurring in at least eight AngioProfile (ig) were compared to each other.

excluded, due to the lack of independence of the Common Denominator Procedures from those genes. In summary, 22 % of the 41 candidate genes with a GeneScore of 100 from the bCDP were either experimentally verified by XantoScreen™ or previously known to modulate angiogenesis. For the gaCDP this percentage increased to 32 % of the 22 candidate genes with a GeneScore of 100. Most likely, the introduction of a more sophisticated candidate gene selection improved the performance of the gaCDP compared to the bCDP. 25 % of the 52 candidate genes that occurred in at least eight AngioProfiles of the igCDP were verified in the same way. This percentage increased further to 57 % for the seven candidate genes that occurred in at least twelve of the 16 AngioProfiles. The introduction of a subjective IndicatorGeneSet in the igCDP allows to influence the procedure toward co-expression with the desired genes of interest. Ranking the candidate genes according to their occurrence in AngioProfiles led to an improved scalability of the number of desired candidate genes at different stringency settings. This may be particular interesting for adaption of the number of candidate genes to the needs of the scientist, e. g. higher stringency settings for a primary *in silico* screen for target discovery, or lower stringency settings for quickly assessing preliminary high-throughput screening results.

In summary, comparison of the resulting candidate genes lists with the results of an experimental high-throughput screen for pro-angiogenic factors as well as already known but not yet annotated (in Gene Ontology and GRIF) angiogenesis modulators, showed the feasibility of the approach applied to the phenotype angiogenesis. All three variations of the Common Denominator Procedure were able to accumulate angiogenesis-associated genes. The simplest variation, the bCDP, achieved the lowest total enrichment. Nevertheless, a significant enrichment of angiogenesis-associated genes was possible. Better performance comes with increased complexity and higher computational demands, which are mainly necessary

for the computation of the score improbability (s_{imp}) and the more sophisticate candidate gene selection. Additionally, control profiles are necessary which *per se* multiply the computational expense. The GACDP should be preferred if no additional set of subjective indicator genes can be provided. This may even be desired, if the process should be as unbiased as possible. On the other hand, this bias may be particular useful, if a specific group of genes should be involved in the desired phenotype or pathway. In such cases the IGCDP is the procedure of choice. Another advantage of the latter is the scalability in regard to the number of candidate genes, where increased stringency correlates with increased enrichment of phenotype-associated genes.

4.2 Comparison to Established Procedures

More diverse libraries within the input data will lead to increased specificity of the COMMON DENOMINATOR PROCEDURE. Here, the novel data mining procedure deviates from (and in some aspects is superior to) other approaches. Already established procedures include e. g. application of clustering, differentiation and filtering steps for identification of disease-specific genes from EST, SAGE or proteomics data [Becquet et al., 2002; Cai et al., 2004; Krieg et al., 2004; Vasmatzis et al., 1998], as well as from expression data generated via hybridization analyses (e. g. Affymetrix-Chips, [Nishizuka et al., 2003; Wei et al., 2004]). However, all these approaches benefit from or even require the use of data from well defined samples. A general rule for these procedures is that a homogeneous and more defined sample will lead to more useful data for subsequent bioinformatics processing. Any presence of tissues with divergent identity (e. g. not of disease origin) within the samples to be analyzed will taint *in silico* analyses. Accordingly, elaborate sample description (e. g. grade, stage of disease, gender, age) is beneficial and needed for these approaches [Brazma et al., 2001]. In addition to sample description great efforts have been invested on the technical side. To ensure sample homogeneity elaborate procedures such as tissue microdissection were applied to remove all non-disease components from library sources. This technique separates tumor (or other desired) cells from infiltrating lymphocytes, matrix/stroma components, nerves or vasculature, hence leading to library sources of excellently defined origins [Burgemeister et al., 2003; Fend et al., 1999].

In contrast to that, the above described selection of phenotype-specific cDNA libraries is

done completely independent of any pathological characterization or classification just by detecting phenotype-associated genes either from the AngioTestGroup or from the IndicatorGeneSet. The Common Denominator Procedure will, at least in theory, perform the better the more diverse the original source of the data, i. e. the biological samples, are. This includes diversity within the samples from which libraries were made. For example, it is known that the generation of many phenotypes relies not only on pathways within one defined cell type, but on complex interactions between cells and adjacent tissues. The phenotype tumor-angiogenesis provides a good example as it involves not only the induction of classical intracellular signaling pathways like hypoxia regulation but also feedback loops within cells, as well as crosstalk between cancer cells, endothelial cells, neighboring tissues and adjacent stroma. These crosstalks can be assigned to transacting factors, e. g. secreted proteins, which modulate the interplay between matrix, endothelium, tumor tissue, and external stimuli like hypoxia. Because of this complex modulation of tumor angiogenesis, it is of advantage to include into the analysis not only one specific tissue type, but instead many cells, such as the before mentioned cancer cells, stroma, endothelial cells, and possibly even infiltrating lymphocytes. Obviously, this approach is the opposite to the generation and utilization of data sets generated from microdissected sources.

Although the procedure is in some aspects superior to the above mentioned alternatives, it has its limitations. One of these is the need for reliable information about the expression of genes in tissues. To ensure this, a good representation of the mRNA expression by the EST composition of the library prepared from the tissue sample is essential. However, the mRNA representation of a cDNA library from the CGAP Expression Data can be quite low. More than 80 % of the libraries contained fewer than 100 different UniGene clusters (see Figure 4.1). Therefore, the input data had to be restricted to a subset with good mRNA representation. This obviously carries the risk of losing valuable information contained in the other libraries.

4.3 Extensibility

The CDP profiling concept can easily be extended beyond the usage of CGAP Expression Data as input source. To this end, adaptation of the new data source may be necessary. Generally, qualitative expression information of a given transcript in a particular tissue sample

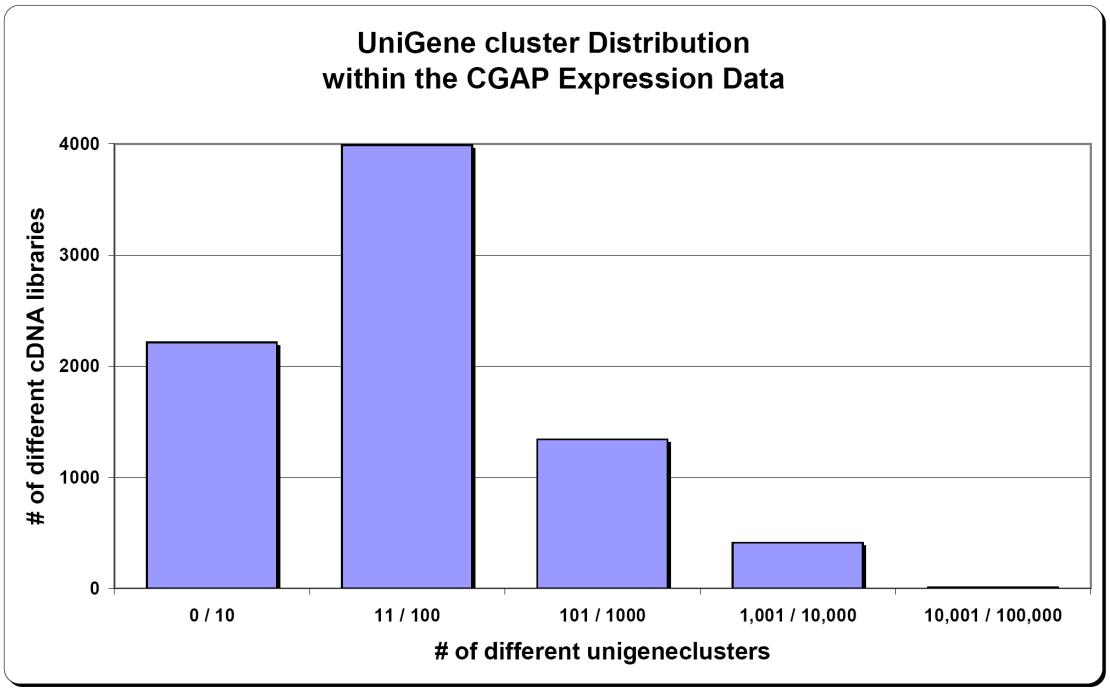

Figure 4.1: Distribution of UniGene Clusters within CGAP Expression Data. Shown is the number of different cDNA libraries (y-axis) containing a particular number of different UniGene clusters (x-axis).

is needed, i. e. information if the observed transcript is expressed in a particular tissue sample. Provided with this information, the CDP qualifies for any form of expression analysis, e. g. Reverse Transcription Polymerase Chain Reaction (RT-PCR), Quantitative Polymerase Chain Reaction (QPCR), Northern Blot, Expressed Sequence Tags (EST), Serial Analysis of Gene Expression (SAGE) and Microarrays (e. g. Affymetrix® GeneChip® Arrays). An entity relationship diagram (see Figure $B.1$) of the data structure as well as a class diagram (see Figure $B.2$) showing the relations between classes of a generalized version of the IGCDP are provided in the appendix.

Affymetrix chip data are especially suited for this approach, as it routinely provides detection (presence or absence) calls for each gene represented on the chip (by a set of perfect match and mismatch oligonucleotides). Ideally, all available expression information of one particular Affymetrics chip, e. g. HG-U133A, will be used as data source. This data comprises a more homogeneous and statistically confirmed data set compared to the CGAP Expression Data. As mentioned above only a small subset of the cDNA libraries and sequence sets within the CGAP Expression Data provided reasonable mRNA (EST) numbers, complexity and distribution which was needed for selection as data source for the procedure.

This restriction was necessary since many sequencing efforts do not routinely extend beyond a limited redundancy in the ESTs, and therefore do not comprehensively reflect the distribution of mRNAs. A better representation of the mRNA expression will not only improve the current procedures, but allow us to expand the procedures to include information about the lack of gene expression in defined tissue samples. Such NEGATIVE LIBRARYPROFILES could further increase the specificity of the candidate genes. For example, different sample RNA populations on the same type of GeneChip (e. g. HG-U133A) will provide us with presence and absence information for every gene on the GeneChip for each of the tissue sample even for more sparsely expressed genes. This is in contrast to EST data were prominent expressed genes increase the effort to identify more sparsely expressed genes. To enable reliable statements about absent of particular genes a very high sequencing effort would be necessary. Using the same type of Affymetrix GeneChip adaptation to the CDP due to mRNA representation is not necessary. The only prerequisite is a reasonable amount of diverse samples of RNA populations for this particular type of GeneChip.

At least in theory every gene present in the particular tissue sample, may be detected by EST analyses. In contrast to that, GeneChips can only detect those genes spotted on the used microarray, i. e. even if a particular gene is present in the observed tissue sample a microarray will not detect it unless it looks for it. Therefore, adaptation of chip expression data sources to the CDP is necessary for the use of different types of Affymetrix GeneChips or even from different microarray platforms. The CDP may only be applied to a subset of common genes on all used chips. To this end, those microarrays should be selected that maximize the amount of common detectable genes, i. e. common spotted genes, as well as the amount of different tissue samples, i. e. sample RNA populations.

4.4 Future Perspective

In conclusion, the literature-based and experimental validation of the candidate genes shows that the COMMON DENOMINATOR PROCEDURES can be applied to identify angiogenesis-associated genes. For all three variations, a significant enrichment was shown even at the lowest stringency settings. Since its phenotype-specificity is solely based upon a small set of angiogenesis-associated genes (ANGIOTESTGROUP and INDICATORGENESET), the procedures can easily be extended to other phenotypes by definition of different phenotype-

associated IndicatorGeneSets and AngioTestGroups.

For comparison of differential expression data between different microarray platforms poor correlation was shown [Kothapalli et al., 2002; Kuo et al., 2002; Li et al., 2002; Tan et al., 2003]. Better results were achieved by comparison of differential expression analyses on two different short-oligonucleotide based microarray platforms (Affymetrix and GE Healthcare) [Shippy et al., 2004]. Shippy *et al.* proposed that different platforms have a strong similarity in calls and blame differences in sensitivity and levels of noise to be responsible for low correlation. Due to the high demands in data quality and comparability as well as different noise levels, a quantitative comparison across different gene expression data platforms is problematic. The Common Denominator Procedures exploits only qualitative expression information, i. e. presence or absence of particular gene in a particular tissue sample. Therefore, noise and sensitivity play a minor role compared to e. g. differential expression analyses. It looks for joint presence of a group of phenotype-associated genes (AngioTestGroup or IndicatorGeneSet) for each individual tissue sample. All genes present in that particular tissue sample suffer from the very same experimental setup, resulting in negligible influence on their joint detection. Therefore, expression data across different sources can be combined, thus minimizing the bias within results of the single platforms. Another advantage of a qualitative view is the ability to reflect post-translational regulation of protein expression. Even post-translationally regulated proteins of common pathways should display similar qualitative expression profiles. Additionally, the Common Denominator Procedures is more frugal in regard to sample homogeneity. It even desires a holistic view on as many diverse data sources as possible. Accordingly, complicate and expensive techniques for creation of homogeneous tissue samples, like micro-dissection, are not necessary, in fact even undesired.

Adaptation of other expression data sources like chip data to the procedure is possible. Fortunately, an increasing amount of microarray data is available in public databases [Diehn et al., 2003; Edgar et al., 2002; Gollub et al., 2003; Praz et al., 2004; Rocca-Serra et al., 2003]. The NCBIs' Gene Expression Omnibus (GEO) for example, provides 2055 different tissue samples for the Affymetrix GeneChip Human Genome U133 Array Set HG-U133A (GPL96) including present and absent information for each of the more than 39,000 transcript variants represented on the chip (as of 01/2005). A possible follow-up would be to use this data as input for the Common Denominator Procedure. A next step would be to investigate the predicted applicability to cross-platform expression data. As mentioned

above, many different phenotype could be evaluated. Investigation of a single well defined pathway instead of a phenotype, which may be composed of several pathways, would also be of interest. Due to its high specificity, especially the IGCDP is suitable as primary screen for target discovery. All three COMMON DENOMINATOR PROCEDURES can be combined with functional genomics techniques for identification of target genes for the diagnosis and therapy of human diseases.