Chapter 2

System and Methods

The System and Methods chapter describes implementation and calculation details of the Common Denominator Prodecure (CDP). It contains information about:

1. The used infrastructure.
2. The necessary data sources.
3. The individual steps of the data mining procedure.
4. The setup of the HUVEC proliferation high throughput assay that was used for experimental validation.

2.1 Infrastructure

The data was stored in an Oracle® 10g database running on a Red Hat® Advanced Server™ 3.0 Platform with two 2.8 GHz Intel Xeon™ CPUs and 4 GB RAM. The underlying data structure is illustrated in an entity relationship diagram (ERD) for the generalized version of the CPD. The ERD is shown in the Discussion in Figure B.1. The Java™ application was developed using the Java Development Kit 1.4.2 from Sun Microsystems®. A class diagram for the generalized version of the CDP is shown in the Discussion in Figure B.2. The Java Genetics Algorithms Package (JGAP) was used to implement the genetic algorithm for the gaCDP. It provides basic genetic mechanisms that can be easily used to apply evolutionary principles to problem solutions. JGAP 1.0 was downloaded from http://jgap.sourceforge.net. Calculations were computed on a Rocks™ 3.2.0 Cluster which is based on Red Hat Linux®.

The cluster consisted of a frontend with two 2.4 GHz Intel Xeon CPUs and 2 GB RAM as well as eight nodes with two 2.6 GHz Intel Xeon and 512 MB RAM each.

2.2 Data Sources

CGAP Expression Data (as of June 2004) was downloaded and used for all subsequent analyses (ftp://ftp1.nci.nih.gov/pub/CGAP/Hs_ExprData.dat). First, CGAP Expression Data had to be adapted to the needs of the CDP. Second, phenotype- or pathway-specific input data needed to be selected. To this end, known angiogenic factors were used to define a set of angiogenesis-specific indicator genes (IndicatorGeneSet) and an angiogenesis-specific test group (AngioTestGroup).

2.2.1 Adaptation to CGAP Expression Data

Many sequencing efforts do not routinely extend beyond a limited redundancy in the ESTs. Therefore, they do not comprehensively reflect the distribution of mRNAs. Which is nescesarry to obtain information about the co-expression of genes, a prerequisite for the CDP. This representation of mRNAs in cDNA libraries is reflected by the distribution of corresponding ESTs, which was assessed by the average ratio of the number of ESTs per expressed gene observed in that library. To ensure reasonable mRNA representation, only libraries having a ratio greater than three were processed further. The set was further restricted to libraries which represented at least 1000 different UniGene clusters, but fewer than 5000 clusters. These requirements were met by 75 libraries (see Table $A.1$), which represented 28 different tissues. 30 libraries were from normal tissue, 43 neoplastic and two from uncharacterized tissue. 54 of these libraries were generated using a non-normalized protocol, five were normalized and 16 had multiple or uncharacterized treatment. In order to estimate the coverage $\hat{C}$ (completeness of EST representation) of the selected 75 cDNA libraries equation 2.1 was used [Susko and Roger, 2004]. Here the coverage is estimated by one minus the percentage of UniGene clusters that occur only once in all reads of a given cDNA library. For those libraries the estimated coverage was between 77.5 % and 99.4 % with a standard error of less than 0.01.

$$
\begin{aligned}
\hat{C} &:= \text{estimate of coverage;} \\
n &:= \#\text{ reads of a given library;} \\
n_1 &:= \#\text{ genes represented once in } n;
\end{aligned}
$$

$$
\hat{C} = 1 - \frac{n_1}{n} \tag{2.1}
$$

2.2.2 Definition of the IndicatorGeneSet

The iGCDP is based on the determination of presence or absence of a user-defined set of known angiogenesis-associated genes. The chosen indicator genes were all known pro-angiogenetic factors and were termed IndicatorGeneSet: HIF1A (Hs.412416), DDIT4 (Hs.111244), VEGF (Hs.73793), IGFR1 (Hs.239176), ECGF1 (Hs.435067) and EPAS1 (Hs.8136). Combinations of three to six of these indicator genes were used to generate 42 LibraryProfiles of sets of cDNA libraries which best resemble the combined expression pattern of these genes (see 2.3.1).

2.2.3 Definition of the AngioTestGroup

For selection of angiogenesis-specific LibraryProfiles and as internal control, an additional set of genes associated with angiogenesis, termed AngioTestGroup, was selected in a semiautomatic manner, based on publicly available gene annotation. To this end, three consecutive steps were applied:

- All genes annotated in Gene Ontology [Camon et al., 2004; Harris et al., 2004] with angiogenesis (GO:0001525) or response to hypoxia (GO:0001666) themselves or any of their children in the Gene Ontology hierarchy were selected.
- All genes with a gene reference into function (GRIF, LocusLink) [Pruitt and Maglott, 2001] description containing the phrases 'hypoxia', 'hypoxic', 'angiogenic' or 'angiogenesis' were added.

- Negative modulators of angiogenesis were manually eliminated.

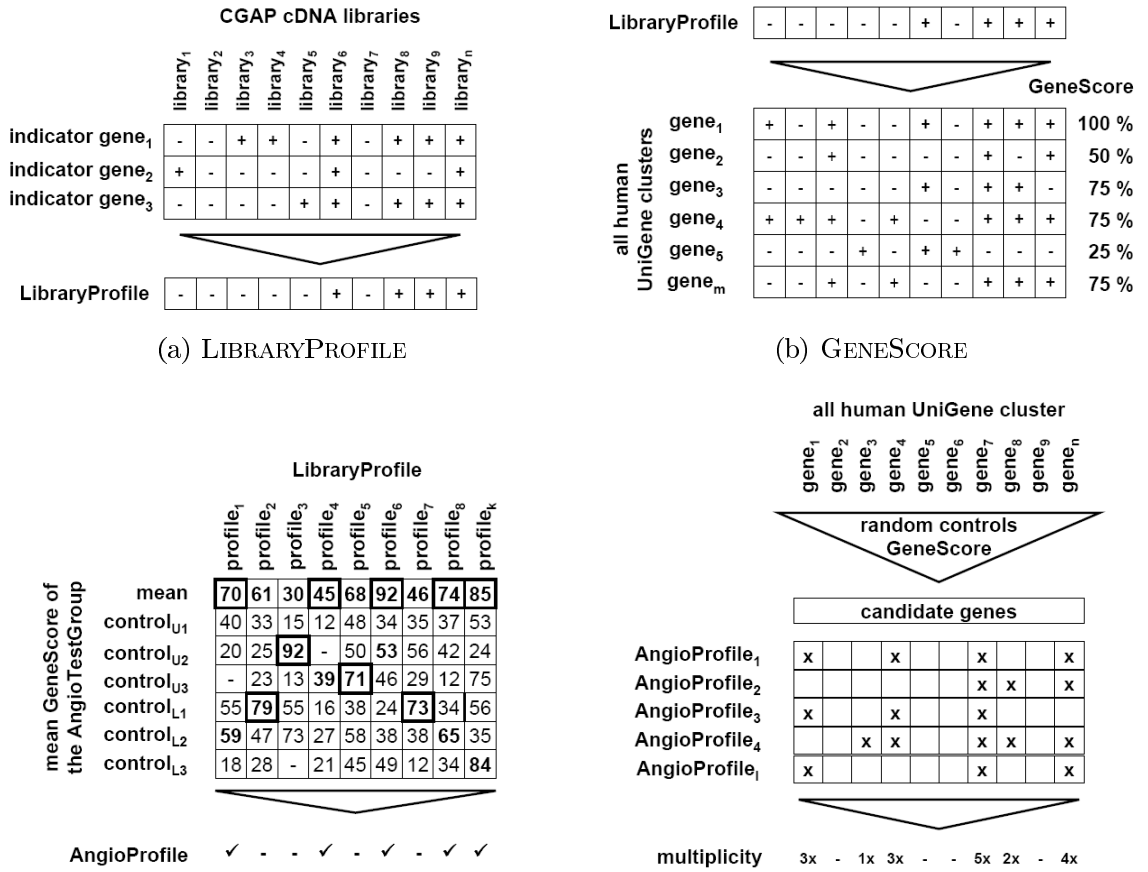
These criteria were met by 170 human UniGene clusters which comprised the AngioTestGroup and used for the bCDP and the gaCDP. The igCDP used a subset of these UniGene clusters. Here, the AngioTestGroup was used to assess the quality of a given LibraryProfile which needs to be composed of at least eight libraries. Therefore, UniGene clusters present in less than eight of the 75 usable cDNA libraries were removed from the AngioTestGroup. Likewise, UniGene clusters present in more than 50 % of the 75 usable cDNA libraries were removed with the rationale that genes expressed in most libraries are inadequate for assessing LibraryProfiles, due to lack of specificity. With these additional restrictions, a subgroup of 73 UniGene clusters was selected and used as AngioTestGroup for the igCDP to asses LibraryProfiles and as internal control.

This concludes the adaptation of CGAP Expression Data to the CDP and the manual selection of pathway- or phenotype-specific input data. Subsequently, the fully automatic data mining procedure is described.

2.3 Common Denominator Procedure (CDP)

The procedure for *in silico* identification of phenotype- or pathway-associated genes is exemplary described for the phenotype angiogenesis in the results chapter of this thesis. There, workflow diagrams of the different Common Denominator Procedures are provided (see Figure 3.1, 3.2 and 3.3). Here, implementation details for the individual steps of the Common Denominator Procedures are described. A detailed schematic representation of key steps of the igCDP is shown in Figure 2.1. Generally all three Common Denominator Procedures are composed of the following three key steps:

- A LibraryProfile was determined as a set of angiogenesis-specific cDNA libraries.
- For each human UniGene cluster a GeneScore was calculated by percentage of libraries from the LibraryProfile containing it.
- Candidate genes were selected according to their GeneScore.



(a) LIBRARYPROFILE (b) GENESCORE

(c) ANGIOPROFILES (d) Candidate Genes

Figure 2.1: Key Steps of the COMMON DENOMINATOR PROCEDURE. (a) The presence of selected UniGene clusters from the INDICATORGENESET in CGAP libraries is used to compile the LIBRARYPROFILE. (b) Subsequently, the LIBRARYPROFILE is used to calculate the GENESCORE for every human UniGene cluster based on the number of LIBRARYPROFILE libraries containing the particular UniGene cluster. (c) To identify those LIBRARYPROFILES that are most suited for the identification of angiogenesis genes random control profiles were used. For each LIBRARYPROFILE three random UniGene cluster control profiles ($control_U$) and three random library control profiles ($control_L$) were calculated. The mean GENESCORE of genes from the ANGIOTESTGROUP is used to assess the angiogenesis-specificity of the LIBRARYPROFILE compared to the control profiles. Those LIBRARYPROFILES with a higher mean value compared to all controls are considered most angiogenesis-specific and termed ANGIOPROFILES. (d) For each ANGIOPROFILE candidate genes were selected based upon the GENESCORE compared to random control profiles. Candidate genes occurring in multiple ANGIOPROFILES were considered more reliable hits. Therefore, the final candidate gene list was rank according to the multiplicity of candidate genes in the ANGIOPROFILES.

In addition to the key steps, do the gaCDP and the igCDP use a more sophisticated candidate gene selection, where random control profiles are needed (see 2.3.3). Furthermore, does the igCDP generate a multitude (42 in this case) of LibraryProfiles, of wich the most angiogenesis-specific (AngioProfiles) are selected.

2.3.1 Generation of the LibraryProfile

A LibraryProfile is a small set of angiogenesis-specific cDNA libraries. To obtain those angiogenesis-specific cDNA libraries presence or absence of angiogenesis-associated genes in individual cDNA libraries is determined. All three CDP differ in the way in which they determine those angiogenesis-specific cDNA libraries.

basic Common Denominator Procedure

The bCDP selects its pathway- or phenotype-specific libraries from those that contain most of the genes from a particular pathway- or phenotype-specific test group. To this end, all 75 selected CGAP libraries were ranked according to occurrence of different UniGene clusters from the AngioTestGroup. The eight cDNA libraries containing the highest number of AngioTestGroup genes compose the LibraryProfile for the bCDP (see Table 2.1).

genetic algorithm based Common Denominator Procedure

For a more sophisticated automatic selection of angiogenesis-specific libraries a genetic algorithm (GA) was used. GAs are evolutionary algorithms that evolve a sample set of solutions (individuals) toward an optimum solution through application of Darwin's principle of natural selection [Darwin, 1859; Goldberg, 1989; Holland, 1975; Rechenberg, 1973; Schwefel, 1977]. In the GA individuals are represented by chromosomes. Each chromosome represents a possible solution to the problem of interest. It is composed of discrete parts (genes) that represent the individual features of the solution. The key component of the GA is the fitness function. It calculates the 'fitness' of an individual chromosome, i. e. how 'good' this individual solution is compared to the other potential solutions. Fitness determines the chance of a particular chromosome to survive to the next generation, simulating reproduction. Additionally, a small fraction of the chromosomes that were selected for the next generation undergo mating or mutation. Mating is simulated by swapping gene values

# of AngioTestGroup genes	CGAP cDNA library id
5	41585
6	34284
7	37694, 41618, 41171
8	41049, 41620
9	41586, 41617
10	33318, 33405, 35023
11	40024
12	34184, 35614
13	39336, 39982
14	35645
15	39339, 41603
16	33628, 34333, 41612
18	33286, 34334, 41591, 35646
19	33313, 39895, 37946, 34182
20	34620, 39287, 35029
21	34187, 39276, 39340, 35615, 34619
22	39995, 41607, 41609, 41614, 40063, 41614, 40063
23	34347, 35639, 40023, 39927, 41619
25	34186, 35026, 35031, 34317
26	41615
27	39275
28	33664, 37458
29	33401
30	37900, 39925
31	34188, 35623, 37948, 37949
33	34300, 41605
34	34185
37	41613
39	37853
42	33320
47	39928
50	35627
51	39951
52	35629
76	41631

Table 2.1: Ranking of the 75 selected CGAP libraries according to the number of AngioTestGroup genes contained in that libraries.

between two chromosomes. Mutation is simulated by randomly altering values of a gene in a chromosome (see Figure 2.2).

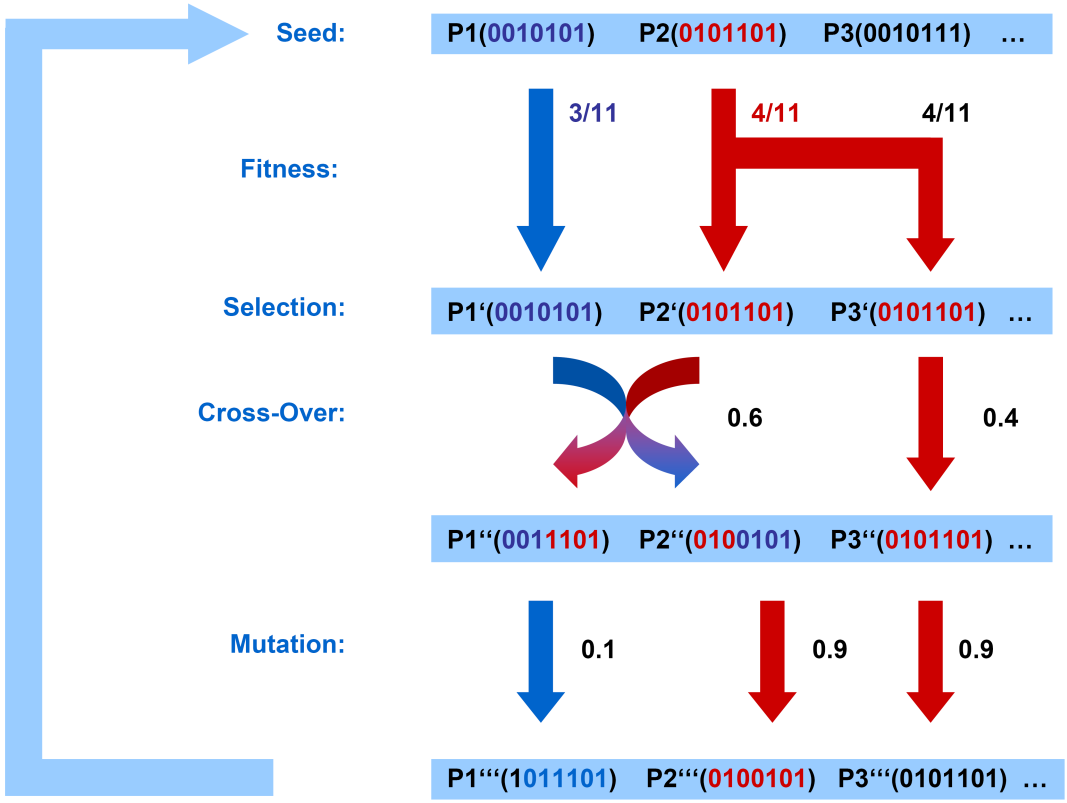


Figure 2.2: Example of a Genetic Algorithm. The initial population consists of three chromosomes (P1, P2, P3). Each of those chromosomes contains seven discrete genes which are active (1) or inactive (0). Fitness is determined by the number of active genes of a given chromosome. Based on the fitness of individual chromosome is its chance to be selected for the next generation. Note that P2 is selected twice and P1 was selected although its fitness is smaller than that of P3. From the selected chromosomes (P'1, P'2, P'3) P'1 mates with P'2. During the so called cross-over a fraction of the genes is exchanged between both chromosomes. Finally a small part of the chromosomes mutates by randomly setting the value if single genes.

The GA was used to select a set of at least eight angiogenesis-associated cDNA libraries composing the LIBRARYPROFILE. Therefore, the mean GENESCORE of all UniGene clusters from the ANGIOTESTGROUP present in the active cDNA libraries was maximized. To this end, the JGAP 1.0 application program interface (API) was used. A chromosome represents a LIBRARYPROFILE and was defined as a set of exactly 75 Genes. Each Gene represents the usage of a particular cDNA library. Therefore, it has two discrete states: active (1) or not active (0). The fitness function determines the mean value of the GENESCORES of all UniGene clusters from the ANGIOTESTGROUP that occurred in the active cDNA

libraries of a given Chromosome. As an additional constraint, a penalty was introduced for Chromosomes with less than eight active Genes. Here, the chance to survive was set to 0. After 2000 steps of evolution the fittest Chromosome was chosen. Its active libraries compose the LIBRARYPROFILE for the GACDP.

INDICATOR GENES BASED COMMON DENOMINATOR PROCEDURE

In contrast to the above two procedures, the IGCDP did not use the ANGIOTESTGROUP to determine its LIBRARYPROFILES. Instead, it ranks libraries according to the number of chosen UniGene clusters from the INDICATORGENESET contained in that particular library. First, all libraries containing every selected indicator gene were added to the nascent LIBRARYPROFILE. Second, if the nascent LIBRARYPROFILE consisted of less than eight libraries, all libraries containing one less indicator gene were added, as long as the libraries contained at least two indicator genes (see Algorithm 1).

```
select a subset ≥ 3 UniGene clusters from the IndicatorGeneSet;
forall used 75 cDNA libraries do
    if library contains all selected indicator genes then
        LibraryProfile ← add library;
    end
end
t ← 0;
while LibraryProfile contains ≤ 8 UniGene cluster do
    t = t + 1 ;
    forall used 75 cDNA libraries do
        x ← number of selected indicator genes minus t;
        if x ≤ 2 then
            break while loop;
        end
        if library contains x selected indicator genes then
            LibraryProfile ← add library;
        end
    end
end
```

Algorithm 1: search profile - library similarity score

2.3.2 Calculation of the GENESCORE

For each UniGene cluster similar expression according to the LIBRARYPROFILE was measured. Therefore, the percentage of libraries from the LIBRARYPROFILE containing a par-

ticular UniGene cluster was calculated and termed GeneScore (see Algorithm 2).

forall *UniGene clusters* u **do**
  GeneScore$(u) = 100 \frac{\#\text{ libraries from the LibraryProfile containing } u}{\#\text{ libraries from the LibraryProfile}}$;
end

Algorithm 2: GeneScore calculation

Ubiquitously expressed genes have a high probability to achieve high GeneScore independently of the chosen LibraryProfile. This drawback was handled in the genetic algorithm based and in the igCDP by subsequently eliminating all UniGene clusters from the analyses which had a probability of more than 36.8 % ($s_{imp} \leq 1.0$, see below) to have reached at least their GeneScore by chance, and thus are in all likelihood false positives. This probability reflects the differences in distribution of UniGene clusters and the different size of the LibraryProfile (number of libraries) within the 75 selected cDNA libraries and was represented by the discrete score improbability s_{imp}. The score improbability is the negative natural logarithm of the probability to reach at least GeneScore s, rounded to the first decimal place. It was calculated by the following equation with N representing the number of selected libraries, 75 in this case; n number of libraries from N containing the observed UniGene cluster; k number of libraries from the LibraryProfile; x number of libraries from n contained in k; s GeneScore of the observed UniGene cluster; $s_i := \frac{i}{k}$ for $i = 0 \ldots k$. Be $s > 0$, $n \geq k$, $N - k \geq n$, then:

$$s_{imp} = -\ln \sum_{s_i \geq s} P(s_i) \tag{2.2}$$

$$P(s) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \tag{2.3}$$

$$\begin{aligned} \ln P(s) &= \ln \frac{k!n!(N-k)!(N-n)!}{N!x!(k-x)!(n-x)!(N+x-k-n)} \\ &= \ln k! - \ln x! - \ln(k-x)! + \ln(N-k)! - \ln(n-x)! \\ &\quad - \ln(N-k-(n-x))! - \ln N! + \ln n! + \ln(N-n)! \end{aligned} \tag{2.4}$$

$$\ln a! = \sum_{i=1\ldots a} \ln i \tag{2.5}$$

$$\ln(a+b) = \ln b + \ln(e^{\ln a - \ln b} + 1) \text{ for } a > 0,\ b > 0 \tag{2.6}$$

For UniGene clusters which did not match the constraints above ($s > 0$, $n \geq k$, $N - k \geq n$), no score improbabilities were calculated. Those UniGene clusters were excluded from further analysis.

2.3.3 Control Profiles

Inspired by Monte Carlo sampling [Metropolis et al., 1953], random control profiles were used during the candidate gene selection of the gaCDP or the igCDP. The latter additionally used random control profiles to distinguish between non-specific LibraryProfiles and angiogenesis-associated LibraryProfiles (AngioProfiles), without making any assumptions on the underlying distribution of the LibraryProfiles in regard to GeneScores, score improbabilities and the number of LibraryProfile libraries. For each control profile GeneScores and probability of the GeneScores were calculated as described above.

Random Library Control Profiles

Control profiles were generated using a random sampling from the 75 selected libraries. In case of the basic and the gaCDP eight random libraries were selected for each control profile. Under the rationale that LibraryProfiles of the igCDP were composed of different numbers of libraries for each combination of UniGene clusters from the IndicatorGeneSet, matching numbers of random libraries were used to generate control profiles for each LibraryProfile.

Random UniGene Control Profiles

As the LibraryProfiles of the igCDP were selected with help of a small set of indicator genes (see above) an additional type of control profiles was introduced. This control profile uses random UniGene clusters instead of the angiogenesis-specific indicator genes to generate its LibraryProfiles. To generate random UniGene control profiles, initially 50 sets each of three, four, five and six random UniGene clusters were drawn from all UniGene clusters within CGAP resulting in 200 random control profiles. The nascent LibraryProfile of 179 of those control profiles (89.5 %) did not contain a single library in their LibraryProfiles

that met the selection criteria (occuring in at least two of the selected random UniGene clusters). Only three of the 200 control profiles (1.5 %) were composed of eight or more libraries, which was the minimum size requirement to define a LIBRARYPROFILE. Most UniGene clusters are represented by ESTs in only one or two of the selected 75 cDNA libraries. Thus, a selection of random UniGene clusters from all UniGene clusters within the CGAP Expression Data does not generate reasonable LIBRARYPROFILES. To generate suitable random control profiles, the library distribution of the indicator genes was additionally considered. To this end, the number of libraries from the selected 75 libraries in which a given indicator gene is present was calculated. Then, for each indicator gene used for the generation of the LIBRARYPROFILE, one random UniGene cluster present in the same number of libraries was selected. To get random UniGene clusters with as little association with angiogenesis as possible the INDICATORGENESET and the ANGIOTESTGROUP were excluded from this selection. These random UniGene clusters were subsequently used to generate the random UniGene control profiles for each LIBRARYPROFILE.

For the LIBRARYPROFILE of the BCDP and the GACDP six random library control profiles were generated. For each of the 42 LIBRARYPROFILES of the IGCDP three random Library and three random UniGene control profiles were generated.

2.3.4 Definition of ANGIOPROFILES

For the selection of angiogenesis-associated profiles, termed ANGIOPROFILES, from all 42 LIBRARYPROFILES of the the IGCDP, performance of UniGene clusters from the ANGIOTESTGROUP was determined for each LIBRARYPROFILES and its six related control profiles was calculated. Following, the mean GENESCORES of genes from the ANGIOTESTGROUP with a $s_{imp} \geq 1$ was calculated. LIBRARYPROFILES with a higher mean GENESCORE than the highest mean GENESCORE of its six corresponding control profiles (16 ANGIOPROFILES) were considered to be suitable for enrichment and identification of angiogenesis-associated genes.

2.3.5 Selection of Candidate Genes

For each of the three COMMON DENOMINATOR PROCEDURES one final candidate gene list was created. The candidate gene list for the BCDP was composed of all UniGene

clusters with a GeneScore of 100 (i. e. UniGene clusters present in all libraries of the LibraryProfile). A less stringent candidate gene list may be created by using all UniGene clusters of at least a particular GeneScore. As mentioned above, ubiquitously expressed genes have a high probability to achieve high GeneScore independently of the chosen LibraryProfile. To this end, the genetic algorithm and the igCDP use a more sophisticated candidate gene selection, exploiting the probability of obtaining at least the achieved GeneScore as well as random control profiles. In case of the gaCDP, candidate genes were selected for the LibraryProfile as described below (see Algorithm 3). In the final candidate gene list those candidate genes were ranked according to their GeneScore. In case of the igCDP candidate genes were selected for each of the 16 AngioProfile as described below (see Algorithm 3). The final candidate gene list was generated by ranking all those candidate genes according to their multiplicity of occurrence as candidate gene in the 16 AngioProfiles.

```
candidates := list of candidate genes;
s_imp := score improbability;
s_imp_t := score improbability threshold;
s_imp_t ← 1;
forall GeneScores > 34 from highest to lowest do
    forall s_imp > s_imp_t from highest to lowest do
        hits ← UniGene clusters having exactly that GeneScore and s_imp;
        h ← # of hits;
        p ←% of h with at least that GeneScore and s_imp in at least one corresponding
        control profile;
        if (h > 10 and p ≥ 33) or h > 50 then
            s_imp_t ← s_imp;
            continue with next GeneScore;
        else
            forall hits do
                if hit has at most that GeneScore and s_imp in all corresponding control
                profiles then
                    add hit to candidates;
                end
            end
            continue with next s_imp;
        end
    end
end
```

Algorithm 3: candidate gene selection

Candidate Gene Selection

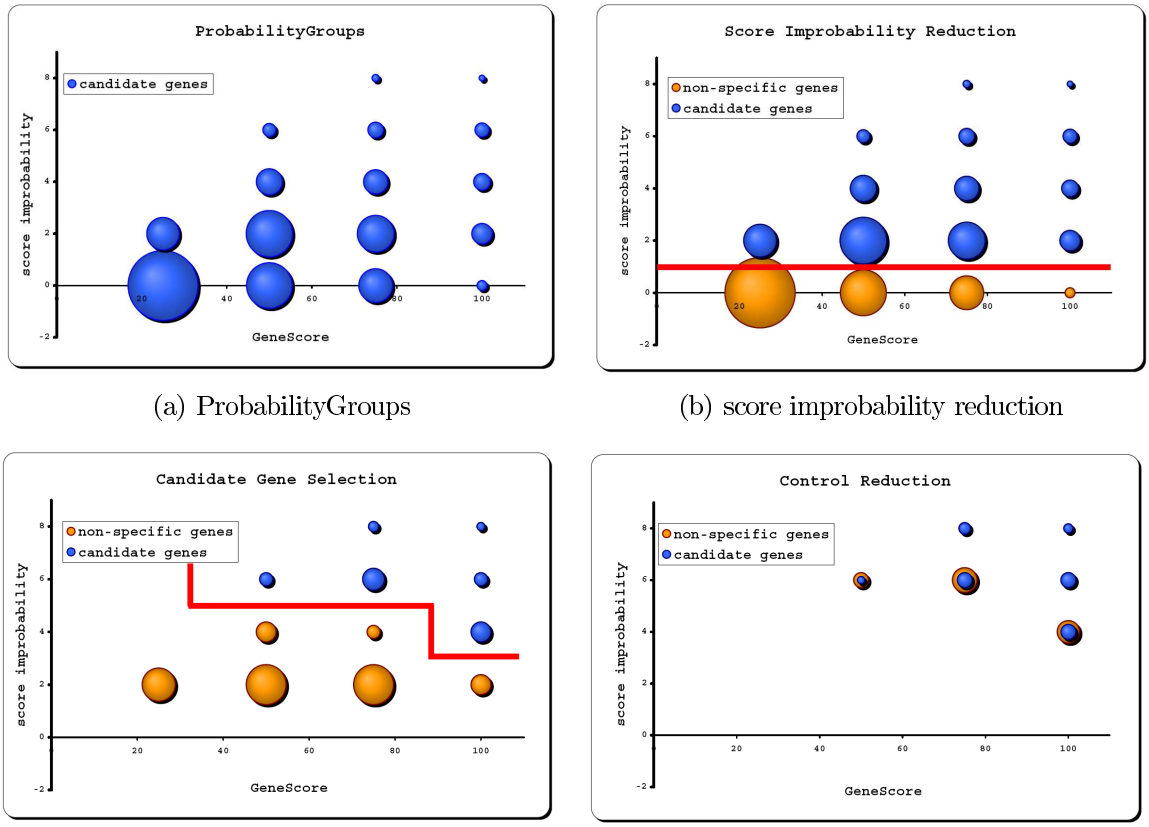
To select candidate genes for a LIBRARYPROFILE, all UniGene clusters present in the libraries of the libraries of the LIBRARYPROFILE were analyzed. Generally, UniGene clusters with a high GENESCORE and a low probability for reaching at least that score by chance (high s_{imp}, see System and Methods) were prefered. For a given GENESCORE all UniGene clusters were grouped into PROBABILITYGROUPS according to their discrete GENESCORE and score improbability (s_{imp}). First, genes with a GENESCORE of less than 34 % were removed, because their similarity to the LIBRARYPROFILE was too low (see Figure 2.3(b)). Next, a score improbability cut off criterion was determined for each GENESCORE as described below. All UniGene clusters having a particular GENESCORE with a score improbability above that threshold were considered candidate genes (see Figure 2.3(c)). The threshold for each GENESCORE was determined by the highest score improbability meeting the following constraints:

- A threshold (of a lower GENESCORE) must not be lower than one of a higher GENESCORE.
- To cap the overall number of candidate genes, PROBABILITYGROUPS must contain less than 50 UniGene clusters.
- The percentage of UniGene clusters with equal or higher score and score improbability in at least one matching control profile must be below 33 %, ignoring PROBABILITYGROUPS containing less than ten genes for statistical reasons.

After completing this procedure independently for all GENESCORES, the remaining list of genes was further pruned by removing genes with better or equal score and score improbability in at least one of the profile-specific controls (see Figure 2.3(d)).

2.4 XantoScreen™

Xantos Biomedicine AG is a functional biology and drug discovery company that is developing biopharmaceuticals in the areas of cancer, inflammatory, metabolic and degenerative diseases. Studies on disease relevant biological functions such as angiogenesis, cell proliferation, cell differentiation and apoptosis are carried out rapidly by screening for phenotypic



(a) ProbabilityGroups

(b) score improbability reduction

(c) hit selection

(d) control subtraction

Figure 2.3: Schematic Candidate Gene Selection. The volume of each sphere correlates to the amount of UniGene clusters having that particular score/score improbability values. (a) UniGene clusters were grouped into PROBABILITYGROUPS according to their discrete GENESCORE and score improbability (s_{imp}). (b) Only those transcripts satisfying a minimum score improbability are considered candidate genes. (c) Based on performance compared to the phenotype control sets, score and the score improbability hits are selected from each profile. (d) From the remaining genes those with better or equal scores in at least one of the profile-specific controls were subtracted.

changes caused by increase of gene function in disease relevant human cells. For this Xantos has established a proprietary high-throughput, fully automated cellular gene-transfection and assay system. The core principle of Xantos' technology is the detection of cellular phenotypes such as alterations of biochemical markers or growth properties, as a consequence of recombinant over-expression (gain of function) of human genes. An experimental high-throughput screen for pro-angiogenic factors was previously performed by Xantos, using a HUVEC proliferation HTS assay (see below). It was set up with the objective to find secreted factors via a functional genomics approach [Grimm and Kachel, 2002; Koenig-Hoffmann et al., 2005; Zitzler et al., 2004]. For that, proliferation of human umbilical vein endothelial cell (HUVEC) was used to identify candidate genes with functionality in angiogenesis [Denekamp, 1982] (see below). This XantoScreen™ identified well known angiogenesis-factors such as VEGF or FGF, as well as a list of 466 novel target candidates. The candidate gene list was compared to experimental results from Xantos' high-throughput screen for pro-angiogenic factors. Therefore, a BLAST [Altschul et al., 1990] of clone sequences from the screen hits against all UniGene clusters (identity $\geq 98\,\%$, ≥ 250 nulceotides) was performed. It identified 611 UniGene clusters that were not member of the AngioTestGroup or the IndicatorGeneSet.

2.4.1 HUVEC Proliferation High Throughput Screening Assay

Human embryonal kidney (HEK293) [Graham et al., 1977] cells were seeded in 100 µl DMEM / 5 % FCS (Invitrogene) on 96-well plates (Costar) at $2.2 \times 10^4\,{}^{\text{cells}}/_{\text{well}}$. After 24 h the cells were transfected using calcium phosphate co-precipitation. 4 h later cells were switched to DMEM with 1,5 % FCS, 1 % Napyruvate, 1 % glutamine and $100\,{}^{\mu\text{g}}/_{\mu\text{l}}$ amphotericin B. Human umbilical cord vein endothelial cells (HUVEC) [Jaffe et al., 1973] were plated in ECGM with supplements (Promocell Heidelberg, single quots) containing 1 % serum, $50\,{}^{\mu\text{g}}/_{\mu\text{l}}$ gentamycin, $0.4\,{}^{\mu\text{g}}/_{\mu\text{l}}$ amphotericine B and $50\,{}^{\text{U}}/_{\mu\text{l}}$ nystatin at $2.5 \times 10^3\,{}^{\text{cells}}/_{\text{well}}$. To test the influence of supernatants derived from transfected HEK293 cells on the proliferation of HUVEC cells, four days after seeding, 90 µl of medium was removed from the HUVECs, which were washed with 200 µl of PBS. Then 75 µl nutrient deficient medium (ECBM, with supplements, Promocell Heidelberg) containing $1\,{}^{\mu\text{g}}/_{\mu\text{l}}$ hydrocortisol, $50\,{}^{\text{U}}/_{\mu\text{l}}$ gentamycil, $0.4\,{}^{\mu\text{g}}/_{\mu\text{l}}$ amphotericin B and $50\,{}^{\text{U}}/_{\mu\text{l}}$ nystatin and 25 µl of supernatants from the transfected 293 cells was added. After four days of incubation of the supernatants, amount and viability of HU-

VEC was determined with the Alamar Blue™ assay (Biosource, California USA). For each well, 11 µl Alamar Blue reagent were mixed with 9 µl of ECBM and added directly to the cells without medium removal. After 4 h incubation at 37 °C fluorescence was measured at 530 nm excitation and 590 nm emission using a Fluoroskan Ascent FL (Labsystems). As positive controls for HUVEC proliferation supernatant containing VEGF was used. Negative controls were supernatants from vector-transfected and PDGF-transfected HEK293 cells.