Chapter 1

Introduction

1.1 Knowledge Discovery in Databases

To facilitate analysis of huge amounts of data the interdisciplinary field of Knowledge Discovery in Databases (KDD) emerged. KDD is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in large data collections [Fayyad et al., 1996]. It provides techniques which extract interesting patterns in a reasonable amount of time at the cross point of machine learning, statistics and database systems. Figure 1.1 shows a schematic representation of the different steps that need to be accomplished for KDD:

1. **Data Selection:** Integrate *a priori* knowledge about the data to select reasonable data as input source.
2. **Data Preprocessing:** Minimize background noise and transform data into a computationally manageable format.
3. **Data Mining:** Recognize interesting patterns and regularities within the data.
4. **Data Interpretation and Evaluation:** Visualize data, reduce redundant information and draw appropriate and sensible conclusions.

The core step within the process of KDD is data mining. Here, machine learning or statistical approaches like Neural Networks, Evolutionary Algorithms, Support Vector Machines, Bayesian Networks or Hidden Markov Models are used to accomplish different goals. The most common among them are:

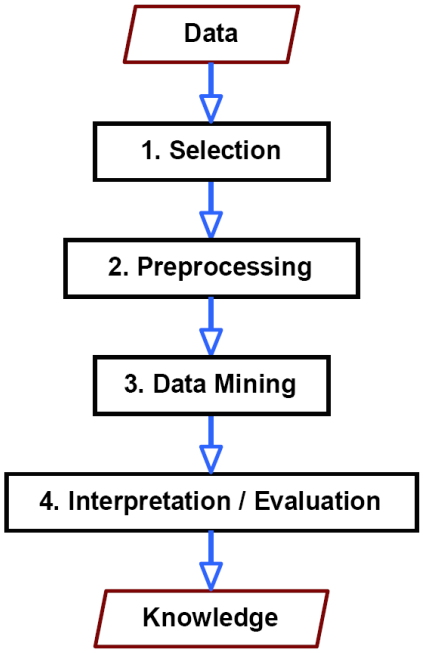


Figure 1.1: Different Steps of Knowledge Discovery in Databases. 1. Integrate *a priori* knowledge about the data to select reasonable data as input source. 2. Preprocess data to minimize background noise and transform data into a computationally manageable format. 3. Mine data to recognize interesting patterns and regularities. 4. A good visualization and reduction of redundant information helps interpretation and evaluation of the data.

- **Classification:** which separates data objects into predefined classes, e. g. automatic recognition of handwritten postal codes.
- **Clustering:** which finds groups of items that are similar based on their features, e. g. encoding data in a compressed form (image compression).
- **Association Rules:** which identify co-occurrence of data objects in the same transaction with a certain probability, e g. identifying groups of items commonly purchased together.

Most approaches can be separated into supervised or unsupervised learning. Supervised learning requires a training set of samples along with the desired classification of each of these samples. Unsupervised learning does not require information about the classification of these training samples. This lack of control may result in the cognition that there is no interesting knowledge within the selected features. However, this unbiased approach may also be of advantage, because it may recognize previously unconsidered patterns.

One form of supervised learning is classification, the process of separating data objects into predefined classes. Therefore, a classifier is trained with a labeled set of trainings objects specifying each class. The main goal of this supervised learning approach is to find a good general mapping that can predict the class for unknown data objects with high accuracy. It also tries to find a compact understandable class model for each of those classes. A recent example of classification is analyzing mass spectra data to improve diagnosis and biomarker discovery [Prados et al., 2004].

A form of unsupervised learning is clustering, which identifies groups of similar items without previous knowledge about the classification of the training data. It separates data objects into previously unspecified groups in a way that maximizes similarity of features of data objects within the group and minimizes similarity between groups. Clustering can help to categorize vast amounts of data and thereby provide a good overview of the data. It may also reveal previously unconsidered patterns. A prominent clustering method for text-mining is WEBSOM. It orders a collection of textual items according to their contents, and maps them onto a regular two-dimensional array of map units (Kohonen Map). Documents that are similar on the basis of their whole contents will be mapped to the same or neighboring map units [Lagus et al., 2004].

1.2 Biological Background

Advances in genomics and proteomics research led to improved molecular biology analytical and computational technologies. Accordingly, an exceptional amount of new bioinformatic databases has arisen. More than 700 key databases containing relevant biological, medical and gene related information have been listed in the NAR database issue of 2005 [Galperin, 2005]. Among others, those databases contain deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and protein sequences, information about protein structure, metabolic and signalling pathways, gene expression, disease related information as well as other genomics and proteomics data. The completion of the deciphering of the human genome in June 2000 was considered a 'milestone in history of mankind' [Lander et al., 2001; Venter et al., 2001]. It contains the 'blueprint of life' in its most condensed form and therefore forms an essential key to understanding complex molecular processes underlying biological systems. This sequence information is now readily available in nucleic and protein sequence databases like EMBL, GenBank®, RefSeq or UniProt [Bairoch et al., 2005; Benson et al., 2005; Kanz et al., 2005; Pruitt et al., 2005].

1.2.1 Central Dogma of Molecular Biology

The central dogma of molecular biology is the flow of genetic information in all living cells from DNA via RNA to protein (see Figure 1.2). Protein synthesis is composed of two different steps: DNA is transcribed to RNA (gene expression) and RNA is translated to protein (protein expression) (see Figure 1.3(a)). Those proteins interact with genes, small molecules and each other forming complex interwoven webs termed molecular pathways. Combination of those pathways leads to destinct phenotypes, e. g. eye color.

Transcription

In the first step of protein synthesis an enzyme called RNA polymerase transcribes one strand of a DNA double helix into messenger RNA (mRNA). In contrast to prokaryotes (unicellular organisms without a nucleus) the more complex eukaryotes (animals, plants, fungi and protists) contain a membrane-bound nucleus, which separates the location of transcription and translation. This enables a more complex regulation of protein synthesis. Here, DNA is not translated directly into mRNA but into a precursor mRNA (pre-mRNA).

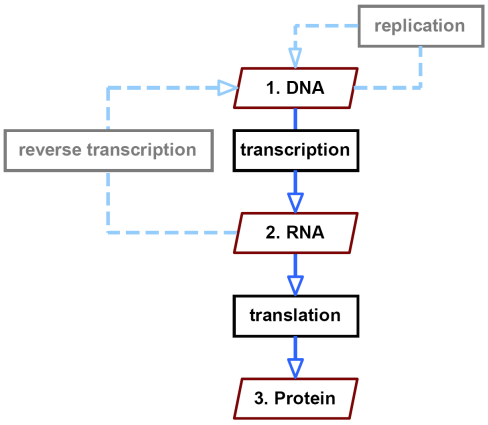


Figure 1.2: Central Dogma of Molecular Biology. Genetic information flows from DNA via RNA to protein.

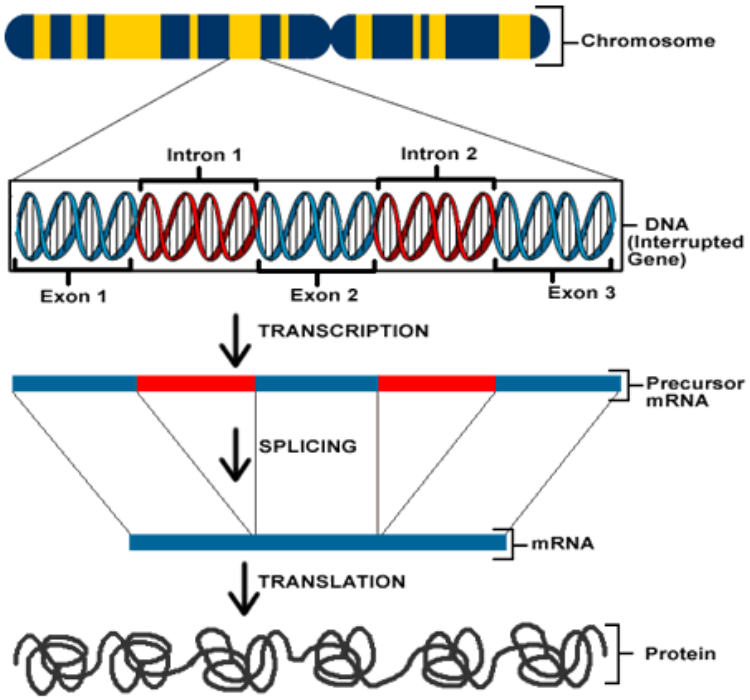
The pre-mRNA includes small coding regions (exons) that are separated by non-coding regions (introns). In a process termed splicing the introns are removed from the pre-mRNA, merging remaining exons to the final mRNA, which is transported out of the nucleus. With help of alternative splicing (different combinations of exons) many different mRNAs and finally many different unique proteins can be created from one gene.

Translation

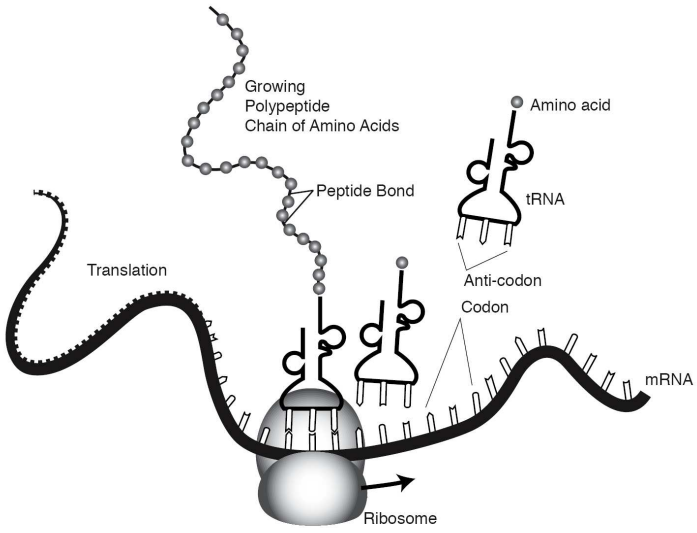
In the second step of protein synthesis, an organelle called ribosome translates mRNA into protein. The mRNA is composed of four different nucleotides (adenine, cytosine, guanine, uracil). Transfer-RNA (tRNA) matches to exactly three consecutive nucleotides, called a codon, on the mRNA. Each tRNA has attached to it a particular amino acid that corresponds to the particular codon on the mRNA. The ribosome fits the matching tRNA to the mRNA, thereby attaching the amino acid to the nascent polypeptide chain. Then, it releases the tRNA and steps forward to the next codon. Step by step the codons of the mRNA are translated into its corresponding amino acids, translating it into protein (see Figure 1.3(b)).

Regulation

Depending on the specific needs of different cell types, only a fraction of the approximately 20,000-25,000 protein-coding human genes are expressed [International Human Genome Sequencing Consortium, 2004]. Changes in qualitative (on/off) as well as quantitative (increases/decreases) gene expression are the means by which the cell responds to different



(a) protein synthesis



(b) translation

Figure 1.3: Different Steps of Protein Synthesis. Protein synthesis is composed of transcription and translation. (a) In the first step mRNA polymerase transcribes DNA into mRNA. During this step, (eukaryotic) mRNA is spliced, i. e. non-coding sequences are eliminated. In the next step, the mRNA is translated into protein. Image taken from from the Science Primer (http://www.ncbi.nlm.nih.gov/About/Primer), a work of the NCBI, part of the NIH. (b) During translation of the mRNA into Protein, tRNAs align at the ribosome adding their attached amino acid to the nascent polypeptide chain. Image taken from the Talking Glossary of Genetic Terms (http://www.genome.gov/10002096), a work of the NHGRI.

stimuli. It gives the cell the broadest control over structure and function. Disruption of this regulation is responsible for many diseases.

1.2.2 Gene Expression

Several methods have been established to access qualitative and quantitative information about expressed genes within the cell, 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).

Expressed Sequence Tags

ESTs are small fractions of DNA sequences generated by sequencing genes expressed in tissue samples [Boguski et al., 1993]. As mRNA is unstable, all mRNAs present in the tissue sample at biopsy are reversely transcribed into stable complementary DNA (cDNA). Those cDNAs are sequenced from both ends to obtain the ESTs. ESTs of various tissues are clustered to approximately 115,000 unique gene (UniGene) clusters [Wheeler et al., 2004], representing different splice variants of the approximately 20,000 to 25,000 protein-coding human genes. As of June 2004, more than five million ESTs from nearly eight thousand cDNA libraries have been compiled in the database of NCI's Cancer Genome Anatomy Project (CGAP) [Strausberg et al., 2002] for *Homo sapiens* alone.

1.3 Biological Challenge

Despite the progress made by various data mining procedures [Huminiecki and Bicknell, 2000; Vasmatzis et al., 1998], there is still a large gap between the amount of available expression data and the availability of data related to the function of those genes. For many genes, expression data can be easily extracted from databases; however, their biological function is not defined, which is indicated by different sources of genomic information like the Gene Ontology Annotation from the European Institute of Bioinformatics [Camon et al., 2004]. Recently, novel technologies have been set up, which speed up the functional assignment of genes. Examples of these techniques are the functional genomics screens that

have been devised by Human Genome Sciences, Inc., [Fiscella et al., 2003] or by Xantos Biomedicine AG [Grimm and Kachel, 2002; Koenig-Hoffmann et al., 2005; Zitzler et al., 2004]. However, improved *in silico* analyses are desired to complement the existing *in silico* and experimental approaches for discovery of disease relevant target genes.

Most recently gene expression data was mainly investigated in a quantitative manner. Research was focused on comparison of differentially expressed genes in pathological or healthy tissue samples. Unfortunately, homogeneous tissue samples of high quality are necessary for this approach. Furthermore, the influence of post-translational regulation on the expression level of biologically active proteins, like degradation or protein folding, is neglected by these approaches.

1.3.1 Common Denominator Concept

The key concept of this thesis is to address these problems by exploiting qualitative instead of quantitative gene expression information on system level, i.e. investigating expression data of as much different and diverse tissue samples as possible instead of using selected samples. The idea is mainly influenced by a novel interdisciplinary field called systems biology which attempts to understand biology on system level. The core concept of systems biology is to utilize all data retrieved from analysis of individual components to understand the whole system.

The novel *in silico* approach presented here extracts phenotype-associated genes from gene expression centered databases. Underlying the approach is the observation that proteins participating in a molecular pathway linked to a particular phenotype, generally are expressed in the same place or close proximity at about the same time. Likewise, post-translationally regulated proteins of common pathways should display similar qualitative expression profiles. Even secreted factors that trigger specific pathways are frequently produced in close proximity to their corresponding effector molecules [Alberts et al., 2002]. Therefore, soluble proteins are also likely to be co-expressed in a defined tissue sample (as long as the samples were not microdissected). The method for identifying phenotype- or pathway-associated genes via data mining utilizes the fact that any given tissue sample should contain most of the mRNAs encoding proteins participating in active pathway of that tissue. Provided with data from a sufficient number of different tissue samples with the same pathway activated,

it should be possible to identify proteins participating in that pathway. Their participating components are the common denominator of those samples as they should be present in each of them.

1.3.2 Phenotype Angiogenesis

Particular combinations of pathways manifest themselves in corresponding phenotypes. Therefore, it should be possible to detect not only pathway-related but also phenotype-related genes due to their co-expression. A particular interesting phenotype is angiogenesis. Angiogenesis, together with vasculogenesis participates in the development of the new blood vessels. The Gene Ontology [Harris et al., 2004], a controlled vocabulary in which each term is related to one another in a polyhierarchical manner, defines vasculogenesis (Gene Ontology identifier: GO:0001570) as differentiation of endothelial cells from progenitor cells during blood vessel development. It occurs primarily during embryogenesis and is responsible for the formation of the primary vasculatory net. Angiogenesis (GO:0001525) is defined as blood vessel formation when new vessels emerge from the proliferation of pre-existing blood vessels. Angiogenesis also plays a major role during embryonic development, and additionally also during postnatal organ growth [Beck and D'Amore, 1997; Risau, 1997; Risau and Flamme, 1995]. Physiological angiogenesis is almost completely down-regulated in the adult with the exception of the female reproduction system [Augustin, 2000; Reynolds et al., 1992]. Therefore, therapeutic stimulation as well as inhibition of angiogenesis in angiogenesis related diseases should be disease-specific and thus lead to minimal side effects. Pathological angiogenesis arises during cancer and various ischaemic and inflammatory diseases [Carmeliet and Jain, 2000]. Pathological angiogenesis may be desirable (e. g. in the case of wound healing) or undesirable (e.g in the case of tumor growth). In the case of tumor-angiogenesis it was already shown in 1971 that tumors generally have to initiate angiogenesis to enable a growth above a size of 1-2 mm^3 [Folkman, 1971]. As angiogenesis plays an important role in tumor development, progression and formation [Folkman, 1990] the inhibition of tumor growth by attacking the tumor's vascular supply has become a primary target for an antiangiogenic therapy [Barinaga, 1997; Ellis, 2003; Gastl et al., 1997; Harris, 1997]. Classical cancer therapies that target the tumor itself have to suffer from newly developed resistance of the tumor. As angiogenesis is a physiological mechanism of the host development of resistance is not expected [Boehm et al., 1997]. Another advantage of a angiogenesis related treatment of

tumors is its direct contact to the circulatory network, enabling good access to therapeutic agents. Finally, each tumor capillary potentially supplies hundreds of tumor cells, indicating a possible potentiation of the antitumorigenic effect through its destruction.

Avastin™ as angiogenesis suppressing drug

On February 26th 2004 the U. S. Food and Drug Administration (FDA) approved the angiogenesis inhibitor Avastin™ (bevacizumab; Genentech®/Roche®) in combination with intravenous 5-FU-based chemotherapy for the indication of first-line treatment of metastatic colorectal cancer (FDA press release P04-23). Thus, Avastin™ is the first antiangiogenic cancer treatment. The monoclonal antibody targets and inhibits the function of the pro-angiogenic growth factor VEGF [Ferrara et al., 2004; Leung et al., 1989; Willett et al., 2004]. The Avastin FDA approval was based on data from a phase III clinical trial where a statistically significant and clinically meaningful improvement in survival (20.3 months versus 15.6 months) among patients with metastatic colorectal cancer was shown [Hurwitz et al., 2004]. Other indications for the usage of Avastin™ like metastatic non-small cell lung, kidney and breast cancers are already in late-stage clinical trials [Johnson et al., 2004; Ramaswamy and Shapiro, 2003; Rini et al., 2004].

1.4 Outline

In this thesis a novel Java™ application is introduced, which uses the CGAP Expression Data to rank UniGene clusters by their co-occurrence with pre-defined phenotype- or pathway-specific genes. This novel data mining procedure is called Common Denominator Procedure (CDP). The following Systems and Methods chapter contains implementation and calculation details of the CDP. The first time reader is recommended to skip this chapter and proceede with the Results chapter. There, three different versions of the CDP are described and explained on the phenotype angiogenesis:

- basic Common Denominator Procedure (bCDP)
- genetic algorithm based Common Denominator Procedure (gaCDP) and
- indicator genes based Common Denominator Procedure (igCDP).

A generalized version of the CDP is debated in the Discussion, extending it beyond the usage of CGAP Expression Data.